Machine learning to support physicians in endoscopic examinations with a focus on automatic polyp detection in images and videos

vorgelegt von

Adrian Krenzer

Würzburg, 2023

15 83

Kumulative Dissertation zur Erlangung des naturwissenschaftlichen Doktorgrades der
Bayerischen Julius-Maximilians-Universität Würzburg

Abstract

Deep learning enables enormous progress in many computer vision-related tasks. Artificial Intelligence (AI) steadily yields new state-of-the-art results in the field of detection and classification. Thereby AI performance equals or exceeds human performance. Those achievements impacted many domains, including medical applications.

One particular field of medical applications is gastroenterology. In gastroenterology, machine learning algorithms are used to assist examiners during interventions. One of the most critical concerns for gastroenterologists is the development of Colorectal Cancer (CRC), which is one of the leading causes of cancer-related deaths worldwide. Detecting polyps in screening colonoscopies is the essential procedure to prevent CRC. Thereby, the gastroenterologist uses an endoscope to screen the whole colon to find polyps during a colonoscopy. Polyps are mucosal growths that can vary in severity.

This thesis supports gastroenterologists in their examinations with automated detection and classification systems for polyps. The main contribution is a real-time polyp detection system. This system is ready to be installed in any gastroenterology practice worldwide using open-source software. The system achieves state-of-the-art detection results and is currently evaluated in a clinical trial in four different centers in Germany.

The thesis presents two additional key contributions: One is a polyp detection system with extended vision tested in an animal trial. Polyps often hide behind folds or in uninvestigated areas. Therefore, the polyp detection system with extended vision uses an endoscope assisted by two additional cameras to see behind those folds. If a polyp is detected, the endoscopist receives a visual signal. While the detection system handles the additional two camera inputs, the endoscopist focuses on the main camera as usual.

The second one are two polyp classification models, one for the classification based on shape (Paris) and the other on surface and texture (NBI International Colorectal Endoscopic (NICE) classification). Both classifications help the endoscopist with the treatment of and the decisions about the detected polyp.

The key algorithms of the thesis achieve state-of-the-art performance. Outstandingly, the polyp detection system tested on a highly demanding video data set shows an F1 score of 90.25 % while working in real-time. The results exceed all real-time systems in the literature. Furthermore, the first preliminary results of the clinical trial of the polyp detection system suggest a high Adenoma Detection Rate (ADR). In the preliminary study, all polyps were detected by the polyp detection system, and the system achieved a high usability score of 96.3 (max 100). The Paris classification model achieved an F1 score of 89.35 % which is state-of-the-art. The NICE classification model achieved an F1 score of 81.13 %.

Furthermore, a large data set for polyp detection and classification was created during this thesis. Therefore a fast and robust annotation system called Fast Colonoscopy Annotation Tool (FastCAT) was developed. The system simplifies the annotation process for gastroenterologists. Thereby the gastroenterologists only annotate key parts of the endoscopic video. Afterward, those video parts are pre-labeled by a polyp detection AI to speed up the process. After the AI has pre-labeled the frames, non-experts correct and finish the annotation. This annotation process is fast and ensures high quality. FastCAT reduces the overall workload of the gastroenterologist on average by a factor of 20 compared to an open-source state-of-art annotation tool.

Zusammenfassung

Deep Learning ermöglicht enorme Fortschritte bei vielen Aufgaben im Bereich der Computer Vision. Künstliche Intelligenz (KI) liefert ständig neue Spitzenergebnisse im Bereich der Erkennung und Klassifizierung. Dabei erreicht oder übertrifft die Leistung von KI teilweise die menschliche Leistung. Diese Errungenschaften wirken sich auf viele Bereiche aus, darunter auch auf medizinische Anwendungen.

Ein besonderer Bereich der medizinischen Anwendungen ist die Gastroenterologie. In der Gastroenterologie werden Algorithmen des maschinellen Lernens eingesetzt, um den Untersucher bei medizinischen Eingriffen zu unterstützen. Eines der größten Probleme für Gastroenterologen ist die Entwicklung von Darmkrebs, die weltweit eine der häufigsten krebsbedingten Todesursachen ist. Die Erkennung von Polypen bei Darmspiegelungen ist das wichtigste Verfahren zur Vorbeugung von Darmkrebs. Dabei untersucht der Gastroenterologe den Dickdarm im Rahmen einer Koloskopie, um z.B. Polypen zu finden. Polypen sind Schleimhautwucherungen, die unterschiedlich stark ausgeprägt sein können.

Diese Arbeit unterstützt Gastroenterologen bei ihren Untersuchungen mit automatischen Erkennungssystemen und Klassifizierungssystemen für Polypen. Der Hauptbeitrag ist ein Echtzeitpolypenerkennungssystem. Dieses System kann in jeder gastroenterologischen Praxis weltweit mit Open-Source-Software installiert werden. Das System erzielt Erkennungsergebnisse auf dem neusten Stand der Technik und wird derzeit in einer klinischen Studie in vier verschiedenen Praxen in Deutschland evaluiert.

In dieser Arbeit werden zwei weitere wichtige Beiträge vorgestellt: Zum einen ein Polypenerkennungssystem mit erweiterter Sicht, das in einem Tierversuch getestet wurde. Polypen verstecken sich oft hinter Falten oder in nicht untersuchten Bereichen. Daher verwendet das Polypenerkennungssystem mit erweiterter Sicht ein Endoskop, das von zwei zusätzlichen Kameras unterstützt wird, um hinter diese Falten zu sehen. Wenn ein Polyp entdeckt wird, erhält der Endoskopiker ein visuelles Signal. Während das Erkennungssystem die beiden zusätzlichen Kameraeingaben verarbeitet, konzentriert sich der Endoskopiker wie gewohnt auf die Hauptkamera.

Das zweite sind zwei Polypenklassifizierungsmodelle, eines für die Klassifizierung anhand der Form (Paris) und das andere anhand der Oberfläche und Textur (NICE-Klassifizierung). Beide Klassifizierungen helfen dem Endoskopiker bei der Behandlung und Entscheidung über den erkannten Polypen.

Die Schlüsselalgorithmen der Dissertation erreichen eine Leistung, die dem neuesten Stand der Technik entspricht. Herausragend ist, dass das auf einem anspruchsvollen Videodatensatz getestete Polypenerkennungssystem einen F1-Wert von 90,25 % aufweist, während es in Echtzeit arbeitet. Die Ergebnisse übertreffen alle Echtzeitsysteme für Polypenerkennung in der Literatur. Darüber hinaus deuten die ersten vorläufigen Ergebnisse einer klinischen Studie des Polypenerkennungssystems auf eine hohe Adenomdetektionsrate ADR hin. In dieser Studie wurden alle Polypen durch das Polypenerkennungssystem erkannt, und das System erreichte einen hohe Nutzerfreundlichkeit von 96,3 (maximal 100). Bei der automatischen Klassifikation von Polypen basierend auf der Paris Klassifikations erreichte das in dieser Arbeit entwickelte System einen F1-Wert von 89,35 %, was dem neuesten Stand der Technik entspricht. Das NICE-Klassifikationsmodell erreichte eine F1-Wert von 81,13 %.

Darüber hinaus wurde im Rahmen dieser Arbeit ein großer Datensatz zur Polypenerkennung und -klassifizierung erstellt. Dafür wurde ein schnelles und robustes Annotationssystem namens FastCAT entwickelt. Das System vereinfacht den Annotationsprozess für Gastroenterologen. Die Gastroenterologen annotieren dabei nur die wichtigsten Teile des endoskopischen Videos. Anschließend werden diese Videoteile von einer Polypenerkennungs-KI vorverarbeitet, um den Prozess zu beschleunigen. Nachdem die KI die Bilder vorbeschriftet hat, korrigieren und vervollständigen Nicht-Experten die Annotationen. Dieser Annotationsprozess ist schnell und gewährleistet eine hohe Qualität. FastCAT reduziert die Gesamtarbeitsbelastung des Gastroenterologen im Durchschnitt um den Faktor 20 im Vergleich zu einem Open-Source-Annotationstool auf dem neuesten Stand der Technik.

Acknowledgements

I am very thankful for all the people who supported me while writing this thesis. Without these people, this thesis would not have been possible. First, I would like to thank my supervisor, Frank Puppe. He was always there to support and help me with difficult decisions and feedback.

Furthermore, I acknowledge Alexander Hann and Alexander Meining for their help and encouragement to venture into gastroenterology. They broadened my horizon and gave me insight into the daily clinic routine that I would not have gotten otherwise.

I want to thank all my colleagues at the chair of computer science VI and in the group InExEn. Special thanks to Michael Banck, Amar Hekalo, Joel Troya, Thomas Lux, Boban Sudarevic and Daniel Fitting.

I appreciate my mother, Christiana, my father, Stefan, and my siblings, Julian and Katharina, for their constant support during all my efforts in the last years. I want to thank my close friends Jona Bödeker and Tobias Friedetzki for their mental support during the demanding phases of this work. Finally, I want to thank my girlfriend, Johanna Odenwald, for her continued help, encouragement and allegiance throughout the development of this thesis.

Würzburg, February 2023

Contents

Abstract i

Abstract iii

Acknowledgements v

Contents vii

List of Abbreviations xi

1 Introduction 1
- 1.1 Overview of creating machine learning based medical assistance systems 2
- 1.2 Challenges of AI assistance systems for medical applications in gastroenterology . 4
- 1.3 Research questions and contributions 6
- 1.4 Nomenclature 8

2 Overview of the data used for this thesis 11

3 Related work with regard to the publications 15
- 3.1 Semi-automated polyp annotation 15
  - 3.1.1 Related work 15
    - 3.1.1.1 Machine learning annotation tools 16
    - 3.1.1.2 Medical annotation tools for machine learning 17
  - 3.1.2 Contribution and conclusion 17
- 3.2 Polyp detection in still images 18
  - 3.2.1 Related work 18
    - 3.2.1.1 Automated detection with handcrafted features 19
    - 3.2.1.2 Automated detection involving machine learning 19
  - 3.2.2 Contribution and conclusion 21
- 3.3 Polyp detection in videos 22
  - 3.3.1 Related work 22
    - 3.3.1.1 3D convolutions 23
    - 3.3.1.2 3D ResNet 23
    - 3.3.1.3 Optical flow 23
    - 3.3.1.4 Structural similarity 24
    - 3.3.1.5 Post-processing 24
    - 3.3.1.6 Object tracking 24

3.3.2 Contribution and conclusion 25
3.3.3 Future work 26
3.4 Polyp detection with extended vision 27
3.4.1 Related work 27
3.4.2 Contribution and conclusion 27
3.4.3 Future work 28
3.5 Automated polyp classification 28
3.5.1 Related work 28
3.5.2 Contribution and conclusion 30
3.5.3 Future work 31

4 Conclusion 33

A Lead author publications 35
A.1 Fast machine learning annotation in the medical domain: a semi-automated video annotation tool for gastroenterologists 35
A.2 Semi-Automated Machine Learning Video Annotation for Gastroenterologists . . 59
A.3 Endoscopic Detection And Segmentation Of Gastroenterological Diseases With Deep Convolutional Neural Networks 61
A.4 Bigger Networks are not Always Better: Deep Convolutional Neural Networks for Automated Polyp Segmentation 67
A.5 A Real-Time Polyp Detection System with Clinical Application in Colonoscopy Using Deep Convolutional Neural Networks 70
A.6 Deep Learning using temporal information for automatic polyp detection in videos 108
A.7 A User Interface for Automatic Polyp Detection Based on Deep Learning with Extended Vision 114
A.8 Automated classification of polyps using deep learning architectures and few-shot learning 132

B Co-author publications 167
B.1 Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy 167
B.2 Pilot study of a new freely available computer-aided polyp detection system in clinical practice 192
B.3 New concept for colonoscopy including side optics and artificial intelligence . . . 198
B.4 A video based benchmark data set (ENDOTEST) to evaluate computer-aided polyp detection systems 203
B.5 Development and evaluation of a deep learning model to improve the usability of polyp detection systems during interventions 210
B.6 Frame-by-Frame Analysis of a Commercially Available Artificial Intelligence Polyp Detection System in Full-Length Colonoscopies 218

C Declaration of own Contributions 227

Bibliography **231**

List of Abbreviations

ADR	Adenoma Detection Rate
AI	Artificial Intelligence
BE	Barret's Esophagus
CNN	Convolutional Neural Network
COCO	Common Objects in Context
CRC	Colorectal Cancer
CRISP-DM	Cross Industry Standard Process for Data Mining
CSS	Cascading Style Sheets
CVAT	Computer Vision Annotation Tool
DICOM	Digital Imaging and Communications in Medicine
EAD	Endoscopy Artifact Detection
EndoCV	Endoscopy Computer Vision Challenge
FastCAT	Fast Colonoscopy Annotation Tool
Fuse	Full-spectrum endoscopy
FPS	Frames Per Second
GAN	Generative Adversarial Network
GPU	Graphics Processing Unit
HGD	High-Grade Dysplasia
HTML	HyperText Markup Language
IOU	Intersection over Union
ISBI	IEEE International Symposium on Biomedical Imaging
IZKF	Interdisziplinäres Zentrum für Klinische Forschung
JSON	JavaScript Object Notation
KI	Künstliche Intelligenz
MICCAI	Medical Image Computing and Computer Assisted Interventions
MIUA	Medical Image Understanding and Analysis
MRI	Magnetic Resonance Imaging
NBI	Narrow Band Imaging

NICE	NBI International Colorectal Endoscopic
REPP	Robust and Efficient Post-Processing
ResNet	Residual Neural Network
ROI	Region of Interest
SSD	Single Shot Detection
SSIM	Structural Similarity
SVM	Support Vector Machine
TER	Third Eye Retroscope
VIA	VGG Image Annotator
VoTT	Visual Object Tagging Tool
WCE	Wireless Capsule Endoscopy
XML	Extensible Markup Language
YOLO	You Only Look Once

1 Introduction

Over the last decade, the field of computer vision increased with the continued development of object classification and detection systems [15, 91, 92]. Especially great advances in the field of AI created new opportunities in many domains [19, 73, 82]. One particular domain is the medical domain of gastroenterology. In this domain, visual computer assistance systems are used to advance the detection and diagnosis of pathologies further [7, 24].

In gastroenterology, the development of CRC is one of the most critical concerns. CRC is one of the main causes of cancer-related deaths worldwide [14, 74]. Detecting polyps in screening colonoscopies is one of the most important procedures to prevent CRC. A gastroenterologist screens the colon for different pathologies, e.g., polyps. Polyps are mucosal growths that can vary in severity. They develop due to increased cell division on an organ's mucosa, often due to inflammation. Colonoscopies for polyp detection can be assisted by machine learning algorithms analyzing the incoming stream of images from the endoscope. These detection systems increase gastroenterologists' performance and enhance the quality of the colonoscopy [34, 36].

Three projects at the University of Würzburg in cooperation with the gastrology department of the university clinic Würzburg are concerned with the detection, documentation, and diagnosis of polyps during colonoscopies. The first project aims to develop a machine learning system for AI polyp detection and applies and tests it in a clinical trial in different centers in Germany [59]. This AI for polyp detection is one of the main contributions of this thesis. The first preliminary results of the clinical trial are already published [70] and show an increase in ADR and high usability.

The second project is funded by the Interdisziplinäres Zentrum für Klinische Forschung (IZKF) and involves adding additional cameras to create a further advanced AI based automated polyp detection system. This system was developed and tested in an animal trial [58].

The EndoAssist project is the third project, which develops a tool for faster documentation of gastroenterological processes. The EndoAssist AI system interprets endoscopic images and videos by combining deep learning techniques and medical domain knowledge. The result is used as a second opinion system for the gastroenterologists during the examination. Additionally, it can be directly transferred into routine documentation. These documentation processes also involve the classification of polyps.

The main focus of the thesis is the development of medical assistance systems and in particular automated polyp detection and classification systems to assist gastroenterologists during colonoscopies. Therefore, the next chapter is an overview of the creation process of general machine learning-based assistance systems. Afterward, challenges of AI-assisted medical systems in gastroenterology are presented. This involves technical as well as ethical concerns. Lastly, the research questions are introduced and the contributions of the work are presented.

1.1 Overview of creating machine learning based medical assistance systems

This chapter presents eight steps for creating a machine learning-based medical assistance system. The steps are based on a well-known framework in data mining called the Cross Industry Standard Process for Data Mining (CRISP-DM) model [121]. Some steps are changed, renamed, or added, but the essential structure is kept. An overview is shown in figure 1.1. All machine learning-based assistance systems developed during this thesis follow this workflow. Here, only supervised machine learning-based medical assistance systems are considered, as all of the developed systems in this thesis are based on supervised learning. Data quality is one of the most critical aspects of any machine learning system. All of the model's predictions are based on the input data. Therefore, data quality must be consistently maintained. Additionally, the acquisition of new medical data is protected by law. Thus, data anonymization methods with Generative Adversarial Network (GAN) have to be used to anonymize the data while still maintaining high-quality data [100].

In the following, the steps are briefly explained:

- **Medical understanding:** First, the medical problems have to be clearly defined with the medical staff, in our case, gastroenterologists. Next, the objectives and requirements are discussed. Then a clear outline of how to detect the object with the available machine learning methods is established.

- **Data understanding:** As machine learning methods are always based on data, getting suitable data for the selected machine learning problem is essential. The selected medical problem and solution may have to be adjusted based on the amount and accessibility of suitable data. Therefore, medical understanding and data understanding are interconnected.

- **Data preparation:** After the data and the medical background is understood properly, the data needs to be prepared. Data preparation for medical assistance systems consists of cleaning the data, formatting the data, rescaling the image data and augmenting the data. The data preparation precedes the closely related data annotation.

- **Data annotation:** Medical experts must create high-quality annotations to achieve a reliable machine learning system. However, medical experts have very limited time for annotation. Thus, medical experts can be assisted by active learning pipelines to reduce the experts' time per annotation. The data annotation step lays the ground truth for the machine learning problem, which is then subdivided into training, validation, and test data.

- **Modeling:** In this step, different machine learning architectures are developed to solve the medical problem. In the second step, the hyperparameters of the models are optimized to achieve better performance. The modeling step is interconnected with data annotation and preparation. Data annotation can be assisted by a first model pre-labeling or pre-selecting data to improve the annotation process [56]. Additionally, further preprocessing steps must be added or removed from the data preparation step if the models are not performing as intended.

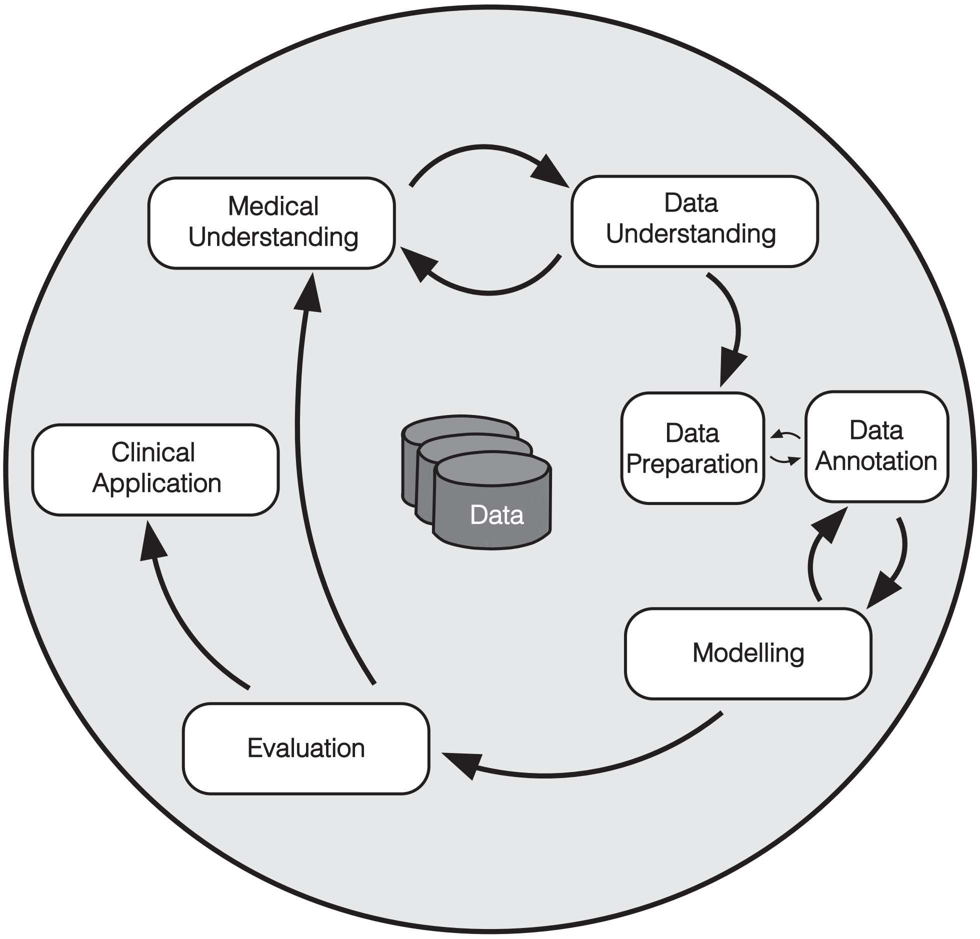


Figure 1.1: The figure depicts the essential steps for creating a machine learning-based medical assistance system. Data is the most critical component in supervised machine learning-based assistance systems. So it is placed in the figure's center. Adopted from Wirth et al. [121]

- **Evaluation:** In this phase, the models are evaluated. The best-performing model is selected via data splits for testing and validation. In addition, the model is evaluated using benchmark data sets to compare the results to other models in the research domain. Afterward, the model is evaluated during medical examinations by medical experts. The feedback from the medical experts is used for further optimization of the model.

- **Clinical application:** Before applying the model broadly, there has to be an ethical committee approving the system and procedure for general use. Afterward, the machine learning system can be used in clinical practice e.g., a university clinic using AI assistance. The main purpose of the application step is to adapt the system to the different conditions, which are unique for every clinical practice. Thereby the desired outcome of the system can be achieved and the chosen objectives can be obtained.

1.2 Challenges of AI assistance systems for medical applications in gastroenterology

AI assistance systems for medical applications in gastroenterology encounter two major challenges. The first challenge is to adapt the polyp detection system for real-time support of gastroenterologists. The second challenge is the acquisition of large high-quality data sets. Especially, video artifacts in data sets are a constant problem in computer vision. Filtering the incoming image stream to remove video artifacts is mostly not feasible for real-time applications.

In regards to the clinical setting, the acceptance of the AI system from the examiner and the patient is another challenge. Lastly, ethical concerns about the usage of AI systems in the medical domain must be considered.

Real-time application Real-time performance is a crucial requirement for AI assistance systems in gastroenterology, as the detection of polyps during endoscopic examinations must be timely to provide effective treatment. To be considered real-time capable, a model must be able to process images at a speed of approximately 25 Frames Per Second (FPS). The incoming stream of most endoscopes during colonoscopies is typically in the range of 25-30 FPS, although some newer systems can process images at higher rates of 50-60 FPS. However, using high-performance hardware to achieve faster processing speeds can also increase the cost of the system, which may not be feasible for many gastroenterologists and clinics. Therefore, it is important to design assistance systems that are robust, efficient, and affordable, while still meeting real-time performance requirements. To meet those criteria, the assistance system's entire workflow must be adjusted and optimized from the incoming image to the detection display on the examiner's screen.

One potential approach to optimizing the performance of AI assistance systems in gastroenterology is to focus on reducing the computational complexity of the algorithms used. This could involve techniques such as simplifying the model's architecture, reducing the number of parameters or layers, or using more efficient data structures or optimization techniques. Another approach is to leverage specialized hardware or acceleration technologies, such as Graphics Processing Unit (GPU)s, to speed up the processing of images and reduce the overall latency of the system.

It may also be possible to optimize the data preprocessing and postprocessing steps or to reduce the number of images that need to be processed by applying filtering or sampling techniques.

Training and testing data sets The biggest problem for creating supervised machine learning-based assistance systems is the low availability of usable open-source data sets. This applies not only to the amount of data available but also to the class imbalance in polyp data sets. There is a significant difference between the number of healthy tissue frames and frames with polyps [43, 46]. Polyps that rarely occur, like the Paris type IIc polyp, are also underrepresented in all video and image training data sets. This lack of data is a common problem in the medical field.

Nevertheless, there are different techniques to deal with this problem. The first one is transfer learning. Transfer learning uses a model that has already been pre-trained on a similar data set and is then re-trained with e.g. a polyp data set [46]. A second approach is data augmentation. Data augmentation for images transforms the original image data with the help of various methods. Examples of such transformations for polyp detection image data are rotating, mirroring, blurring, and color adjusting [124]. Additionally, there are approaches for creating additional training material artificially. E.g., Thomaz et al. [20] train a Convolutional Neural Network (CNN) to insert polyps into images of healthy mucosa to increase the available training data.

Another challenge in the data is the effect of video artifacts on the detection performance of the models. Light reflections and blurring often occur in endoscopic videos. In 2019 the Endoscopy Artifact Detection (EAD) challenge was conducted to confront this problem. During the EAD challenge different algorithms for the detection and filtering of blurry or unusable images are presented. Additionally, the EAD offers artifact-specific methods to restore unusable frames. Soberanis-Mukul et al. [104], show the effect of blurring, light reflections and artifacts on the detection rate of automated polyp detection systems. The authors create a multi-class model to detect different artifact types. They pair their artifact detection algorithm with a polyp detection system to make it more reliable.

Acceptance of the examiner and patient Adopting computer-aided technologies in medical practice can be challenging due to several factors. One issue is the reluctance of healthcare professionals to incorporate new AI assistance systems into their established routines. Additionally, some healthcare providers may view AI technology as competition rather than assistance and fear being replaced by it. There is also a general skepticism about technology among some individuals, including healthcare providers and patients, which can hinder the adoption of AI-based systems even when they have been shown to be effective in scientific studies and clinical trials. Patients may also be hesitant to rely on AI for diagnoses due to a preference for human interaction and trust in a healthcare provider as an individual rather than a machine. However, AI-based diagnostic systems have the potential to provide accurate and efficient diagnoses and may be used as a complementary tool to support clinical decision-making [48, 116].

To further increase the adoption of AI-based technologies in medical practice, it may be helpful to address the challenges and concerns mentioned above. For example, efforts should educate healthcare professionals about the capabilities and limitations of AI systems and to demonstrate the benefits of incorporating them into clinical practice. It may also be helpful to establish protocols and guidelines for the use of AI in medical settings, including measures to ensure the transparency, accountability, and ethical use of these systems. Additionally, efforts should improve communication and collaboration between AI researchers and medical professionals and to involve medical experts in the development and testing of AI systems to ensure that they are tailored to the needs of clinical practice [41, 94]. Finally, it may be necessary to address the broader social and cultural factors that contribute to skepticism about AI and other technological advances and to engage in dialogue and outreach with stakeholders to build trust and understanding about the role of these technologies in healthcare.

Ethical concerns of medical AI application There are several ethical concerns that arise when using AI assistance systems in the medical field. One such concern is the potential for variability in the predictions made by these systems, which could lead to incorrect diagnoses with potentially serious consequences for patients. In order to mitigate these risks, it is necessary to incorporate human oversight and control in the use of AI in medical settings. This helps to ensure that the benefits of these systems outweigh the potential risks and that they are used in a fair, transparent, and accountable manner. It is also essential to carefully consider the ethical implications of using AI in medical settings and to establish appropriate protocols and safeguards to ensure that these systems are used in a responsible and ethical manner [49, 97, 114].

In addition to the issues of accuracy and reliability, the ethical considerations surrounding the use of AI in medicine also include issues of fairness and inclusivity. It is important to ensure that AI systems do not perpetuate or exacerbate existing biases or inequalities and that they are accessible and fair to all patients regardless of their background or circumstances. To achieve this, it is important to carefully evaluate and address potential sources of bias in the data and algorithms used to develop AI systems and to ensure that these systems are designed and implemented in a way that is transparent and accountable. In addition, it is necessary to establish protocols and guidelines for the use of AI in medical settings, including measures to ensure that these systems are used in a responsible and ethical manner. These measures should include provisions for human oversight and control, as well as mechanisms for monitoring and evaluating the performance and impact of these systems on patient care [41, 94].

1.3 Research questions and contributions

In recent years, machine learning has shown great promise in assisting physicians with interpreting and analyzing medical images, including those obtained during colonoscopic examinations. To further investigate the advancements and transformation in this domain, this thesis is guided by the following research question:

Guiding research question:

> ***Can machine learning assist physicians in endoscopic examinations with the treatment of colon cancer?***

The literature presents several machine learning assistance systems for gastroenterologists. Nevertheless, most of these systems are not implemented for clinical application. To answer the guiding research question, the thesis involves four main contributions. The most important contribution is the automated polyp detection system. This detection system is designed to detect polys in real-time with clinical application and thereby assist gastroenterologists during the examination. Developing this detection system involves different algorithms and steps to create a final real-time polyp detection system. This AI is applied for the automated detection polyps in four different centers in Germany. The second contribution is the data set and the annotation tool to create the data set for training the polyp detection and classification systems. The third contribution is an extension of the polyp detection system with additional micro cameras to create an extended vision for polyp detection. The last contribution is the development of two polyp classification systems. The guiding research question (RQ) has been divided into the following four sub-research questions:

RQ1: *Can automated polyp detection assist gastroenterologists in their daily clinical practice?*

Scientific publications provide numerous methods for automated polyp detection systems (see sections 3.2 and 3.3). These detection systems use different techniques like 3D architectures, optical flow, structural similarity, post-processing, and object tracking. However, the majority of these systems have not been implemented for clinical application. Therefore, the main contribution of this thesis is the first open-source real-time polyp detection system for clinical application. This contribution allowed for several publications (see sections A.5, [59]; A.3, [53]; B.1, [3]; A.4, [52]; A.6, [57]). Several approaches to image and video detection algorithms were developed and examined while developing a sufficient polyp detection system.

These publications involve creating detection systems for challenges like the Endoscopy Computer Vision Challenge (EndoCV) 2020 challenge [3]. EndoCV is an international challenge about the use of AI algorithms for endoscopic imaging. Mostly detection, classification and segmentation architectures are tested. The created detection system is the winner of the EndoCV 2020 detection challenge (see sections A.3, [53]; B.1, [3]). These efforts led to the final publication [59] in which the real-time detection system is illustrated.

The system can be used directly in the clinical setting and works in real-time. It includes a novel post-processing step that further increases detection accuracy. In addition, the system's performance exceeds state-of-the-art results on the public CVC VideoClinicDB [29] data set.

The developed polyp detection system is now being tested in a clinical trial. Preliminary results of this trial indicate a high ADR of 41.5 % and high system usability (see Sections B.2, [70]). The results consider 41 colonoscopies. The automated polyp detection system detected 66 of 66 polyps and 29 of 29 adenomas. Additionally, all examiners were questioned about the system's usability. The system's usability is very high, with a score of 96.3 (max 100).

RQ2: *How does a semi-automated annotation tool impact the workload of gastroenterologists and the quality of annotated data in endoscopic imaging for polyp detection?*

The literature addressing semi-automated annotation tools (see section 3.1) underscores the necessity for developing specialized medical annotation tools, given the significant expense associated with medical annotations and the unique demands posed by medical formats.

Therefore, a fast and accurate annotation tool for endoscopic videos has been developed (see sections A.1, [56]; A.2 [54]). This tool reduces the workload of the domain expert by a factor of 20 while retaining very high annotation quality. Advanced pre-selection and AI assistance are implemented to reduce the workload.

RQ3: *How does the integration of additional cameras with AI technology enhance the detection of polyps and improve the examination process compared to traditional methods?*

The literature (see section 3.4) presents various approaches to increase the view of the endoscopist. These approaches comprise the Third Eye Retroscope (TER), and the Full-spectrum endoscopy (Fuse). This use of additional cameras in colonoscopies increases the examiner's view and allows him to explore intricate areas further and find polyps otherwise missed. However, multiple views are very complex to analyze at once. The publication presents a new approach with additional cameras and AI assistance (see section A.7 [58]). The views of the additional cameras are managed by the AI and do not interfere with the endoscopist. Still, the endoscopist gets an alert if the AI detects a polyp.

RQ4: *Can deep learning methods achieve high accuracy on the classification of polyps in gastroenterology and does few-shot learning improve the efficiency of the classification process?*

The scientific literature of automated polyp classification (see section 3.5) details a variety of techniques that employ distinct CNN architectures and classification systems, including Kudo's pit-pattern classification, the Paris classification, and the NICE classification. Nonetheless, the assessment of these methods relied on privately acquired datasets, which hinders the possibility of replicating and benchmarking the findings. Therefore, two novel automated classification systems assisting gastroenterologists in classifying polyps based on the NICE and Paris classification are presented (see section A.8 [55]) and evaluated on a public benchmark data set. The paper shows a two-step process for the Paris classification: first, detecting the polyp on the image, then cropping the detected area, and then classifying the polyp based on the cropped area with a transformer network.

For the NICE classification model a deep metric learning-based approach has been developed for classifying polyps. The algorithm creates an embedding space for classifying polyps and utilizes few-shot learning to address the limited availability of annotated images. Overall, the Paris classification model achieves state-of-the-art results on a publicly available data set and the NICE classification model shows the viability of the few-shot learning paradigm for polyp classification in data-scarce environments.

1.4 Nomenclature

Throughout this thesis, a consistent terminology is used:

- **Adenoma:** Adenomas are a specific form of intestinal polyps that develop from normal tissue structure. Adenomas tend to change at the cellular level and can turn into cancer.
- **Adenoma Detection Rate (ADR):** The ADR is calculated by dividing the number of screening colonoscopies with adenomas by the total number of screening colonoscopies with

and without adenomas of a gastroenterologist. Highly experienced gastroenterologists can reach an average ADR of over 40 % [75].

- **Colon:** The colon is an organ located at the end of the digestive tract and outweighs the small intestine in thickness. In addition, the colon has some special anatomical features that distinguish it from other parts of the intestine and make it susceptible to certain diseases.
- **Endoscope:** An endoscope is an optical instrument equipped with an electric light source and mirrors for the examination of body cavities (here the colon) and for the targeted removal of tissue samples.
- **Narrow Band Imaging (NBI):** Narrow band imaging, or NBI , is a variation of endoscopy that uses blue and green light to enhance surface imaging of the mucosa.
- **NBI-International-Colorectal-Endoscopic NICE:** The NICE classification is a classification for grading colon polyps based on the aspect of polyps in NBI.
- **Mucosa:** Intestinal mucosa is the innermost of the four layers of the intestinal wall. It lines the open lumen of the intestine. Mucosa refers to healthy tissue.
- **Paris classification:** The Paris classification is a classification of polyps based on shape. It forms an important basis for targeted indication and underpins the importance of morphological correlation for integrative diagnosis of polyps.
- **Polyp:** In the context of this thesis a polyp is referred to as an intestinal polyp as all of our experiments are done on colonoscopy data. Intestinal polyps are mucosal protrusions that protrude from the intestinal mucosa into the interior of the intestine - the intestinal lumen. They develop because more cells than normal grow in one or more places in the mucosa.

2 Overview of the data used for this thesis

Creating high-quality data is one of the essential aspects of machine learning systems. This is also valid for colonoscopic images for polyp detection and classification. The performance of supervised machine learning systems is mainly based on their training data. Nevertheless, the creation of high-quality data is a cumbersome process. For polyp detection and classification, the annotation has to be done by domain experts, and this makes the process even more expensive.

The data used in the thesis consists of open-source and manually created data sets. Those data sets are merged into a comprehensive data corpus for polyp detection and classification. The overall data set incorporates 506,338 annotated images. The annotations done on the images are bounding box annotations, the size measurement of the polyp, the Paris classification, and the NICE classification. In the following, an overview of the data is given. This overview shows all of the open source data, the data created by the University Clinic of Würzburg and is also presented in our publication A.5:

- CVC-ColonDB [8] 2012: The CVC-ColonDB consists of 300 individual polyp images that were extracted from 15 colonoscopy procedures. Each image represents a random sample of 20 frames per sequence and has a size of 48×288 pixels. These images are available upon request from the CVC-Colon repository[1].

- ETIS-Larib [103] 2014: This data set includes 44 different polyps from 34 videos with 196 polyp images. It was created for the *Medical Image Computing and Computer Assisted Interventions (MICCAI) 2015 Endoscopic Vision Challenge* and used in the challenge. All annotations are done with segmentation masks. Therefore the bounding box coordinates had to be calculated from the segmentation masks. The size of the images is 348×288 pixels. The data is available on request from the CVC-Colon repository[1].

- CVC-VideoClinicDB [11] 2017: The GIANA sub-challenge, which was part of the *MICCAI 2017 Endoscopic Vision Challenge*, published the CVC-VideoClinicDB [4] data set. The data set consists of 18 videos with polyps and 18 videos without polyps. It includes a total of 11,954 frames with polyps and 18,733 frames without polyps, and the images have a size of 574×500 pixels. The ground truth masks for the polyps in this data set were created by approximating the shape of the polyps with ellipses, which were then converted into bounding box coordinates. The data is available in the CVC-Colon repository [1].

---

[1] http://www.cvc.uab.es/CVC-Colon/index.php/databases/

- CVC-Segementation-HD [117] 2017: The CVC-Segementation-HD data was created through the GIANA Polyp Segmentation sub-challenge at the *MICCAI 2017 Endoscopic Vision Challenge*. The data set includes 56 high-resolution images with a size of 1920×1080 pixels, each of which has a corresponding binary mask that was converted into bounding box coordinates. The data is available in the CVC-Colon repository[2].

- CVC-EndoSceneStill [115] 2017: The CVC-EndoSceneStill [117] consists of *CVC-ColonDB* [10] and *CVC-ClinicDB* [12, 26] with 912 polyp images and 44 videos of 36 patients. Each image received a border, mirror, lumen, and segmentation mask for both data sets. The mask in the data set marks the black border around each image, the mirror mask indicates reflections of endoscope light, and the lumen mask marks the intestinal lumen, or the space within the intestine. The segmentation mask includes polyp markers that identify visible polyps in an image. The CVC-ColonDB [10, 117] data set contains 300 selected images from 13 polyp video sequences with a size of 574×500 pixels. The CVC-Clinic-DB [12, 26, 117] contains 612 images from 31 polyp video sequences with a size of 348×288 pixels. The data is available on request from the CVC-Colon repository.

- Kvasir-SEG [44] 2020: The Kvasir-SEG data set includes 1000 polyp images, 1071 masks, and bounding boxes. The dimensions of the images range from 332×487 to 1920×1072 pixels. All images are verified by gastroenterologists from *Vestre Viken Health Trust* in Norway. The images include general information, which is displayed on the left side. The data is available in the Kvasir-SEG repository[3].

- Endoscopy Disease Detection Challenge 2020 (EDD2020) [2]: The EDD2020 data set is a collection of images, with associated masks and bounding boxes, that was released as part of the Endoscopy Disease Detection Challenge in 2020. It consists of five different classes and includes both masks and bounding boxes for each image and instance of a polyp. All of the images featuring polyps are stored in JSON format. The data consists of 127 images with a size of 720×576 pixels. The data is available on request in the ENDOCV repository[4].

- SUN Colonoscopy Video Database [79] 2021: The SUN Colonoscopy Video Database was created by the Mori Laboratory at Nagoya University's Graduate School of Informatics. It consists of 49,136 frames of fully annotated polyps from 100 different polyps, each with its own Paris classification. These images were collected at Showa University Northern Yokohama and annotated by expert endoscopists at Showa University. Also, 109,554 non-polyp frames are included. The size of the images is 1240×1080 pixels. The data is available in the SUN Colonoscopy Video repository[5].

---

[2] http://www.cvc.uab.es/CVC-Colon/index.php/databases/

[3] https://datasets.simula.no/kvasir-seg/

[4] https://endocv2022.grand-challenge.org/Data/

[5] http://sundatabase.org/

- EndoData [59] 2022: The EndoData data set was created with the University and the University Clinic of Würzburg. The team creating the data involved advanced gastroenterologists and medical assistants. The data set contains 346,165 images with 361 polyp sequences and 312 non-polyp sequences. The essential first 1-3 seconds of polyp appearance are used for the polyp sequence, which is critical for detecting polyps in real clinical scenarios. The data combines images from six centers involving three different endoscope manufacturers. The annotations are bounding boxes, the size measurement of the polyp, the Paris classification, and the NICE classification.

3 Related work with regard to the publications

In this section, the methods and results of the relevant publications (see Table 3.1) are briefly summarised and complemented by an overview of the related work. As most of the papers already incorporate a related work chapter, the sections below reflect and extend the context of those chapters.

The contributions show the results and creation process of endoscopic assistance AI models. Those models assist endoscopists during their screening colonoscopy routines. The first two contributions show the tool and the framework for the data annotation workflow. This tool is used to create the annotated data for the classification and detection systems. Afterward, machine learning-based detection systems for the detection of polyps in images and videos are shown. The next contribution is a polyp detection system using extended vision in an animal trial. Last, there is an overview of two polyp classification approaches.

Table 3.1: Overview of the lead author publications of this thesis.

Category	Section	Page	Publication
Polyp annotation	A.2, A.1	59, 35	[54, 56, 71]
Polyp detection in images	A.3, A.4	61, 67	[52, 53]
Polyp detection in videos	A.5, A.6	70, 108	[57, 59]
Polyp detection with extended vision	A.7	114	[58]
Polyp classification	A.8	132	[55]

3.1 Semi-automated polyp annotation

3.1.1 Related work

The created labeling tool FastCAT can be classified into two categories: general annotation tools and medical annotation tools. It is a general annotation tool as it can be used to do general annotation of videos in any domain. Furthermore, the paper compares the tool to a general annotation tool called Computer Vision Annotation Tool (CVAT). CVAT is sophisticated and well-known for fast and accurate annotation in different domains.

Furthermore, FastCAT is viewed in the field of medical annotation tools. Medical annotation tools are customized for a particular medical annotation task and mostly have special requirements that only apply to specific categories of data, e.g., Magnetic Resonance Imaging (MRI) scans.

The tool is especially fast when using gastroenterologists' video data. To allocate the annotation tool in the literature, the following overview shows a brief history of general annotation tools and annotation tools specialized in the medical field.

3.1.1.1 Machine learning annotation tools

The first methods to collect large data sets of labeled images were developed in the 1990s [13]. E.g., "The Open Mind Initiative", a web-based framework, was published in 1999. The aim was to collect data annotated by web users that intelligent models could use [107].

Through the years, several ways have been developed to obtain annotated data. E.g., the online game ESP was created to generate labeled images. In this game, two randomly selected players are given the same image and, without communicating, must guess the other player's thoughts to find a common term for the image as fast as possible. [1, 13]. As a result, several million player-annotated images have been collected. The first and foremost classic annotation tool called Labelme was developed in 2007 and is still one of the most popular open-source online annotation tools for creating computer vision data sets. Labelme enables the labeling of objects in images using specific shapes, as well as other features [96].

Since 2012 the rise of deep learning in computer vision has been followed by a rise in the creation of annotation tools. Thus, one of the most popular annotation tools, LabelImg, was released in 2015. LabelImg is an image annotation tool based on Python, which uses bounding boxes for image annotation. The annotations are stored in Extensible Markup Language (XML) files and saved in either PASCAL VOC or You Only Look Once (YOLO) format.

Also, in 2015, the annotation platform Playment was published. Playment creates training data sets for computer vision by labeling images and videos using different 2D or 3D boxes, polygons, points, or semantic segmentation. In addition, Playment provides automatic labeling for support. Two years later, the paid labeling tool Rectlabel was released, but only available on macOS. Rectlabel used the classic annotation options like bounding boxes and automatic labeling of images. It also uses the PASCAL VOC XML format and can export the annotations to various formats (e.g., YOLO or Common Objects in Context (COCO) JavaScript Object Notation (JSON)).

Following Rectlabel, Labelbox was introduced. Labelbox is a commercial training data platform for machine learning. Additionally, it provides an annotation tool for images, videos, texts, or audio and the management of labeled data.

More recent approaches are the following three annotation tools. Released in 2016, the VGG Image Annotator (VIA) [23] is a tool that runs in a web browser. It can be run without any installations and utilizes JavaScript, Cascading Style Sheets (CSS) and HyperText Markup Language (HTML). There are numerous annotation shapes available, including points, polylines, lines, rectangles, ellipses, and polygons.

In 2019 Microsoft released an open-source tool for the annotation of images and videos called Visual Object Tagging Tool (VoTT) [77]. The tool can be used in any web browser as it is written in TypeScript and leverages the React framework for its implementation. It is also possible to run VoTT locally on a personal computer as a native application.

Developed by Intel, the tool CVAT [99] is an open-source application for annotating videos and images. It was released in 2019. A user management system enables the ability to collaborate with screen-annotated data. It has a connection through a remote source-mounted file system to upload images to the server. CVAT is one of the most promising open-source annotation tools as the system is user-friendly and intuitive.

3.1.1.2 Medical annotation tools for machine learning

With the considerable increase in interest and progress in AI, machine learning models are entering various fields, including medicine. In the medical field, machine learning models can assist professionals in their daily routines [88, 102, 119]. Like all machine learning models, the models used in the medical field need labeled data for training. As a result, the need for labeled medical images and videos is a major issue for medical professionals. Standard annotation tools such as those already described above can be used. However, since medical annotations are always very expensive, special tools have been developed to support specific medical requirements. One specific medical requirement is annotating layered images, e.g., computed tomography and MRI. Other requirements are ultra-high-resolution images primarily used in pathology images, high data protection regulations, the processing of 3D images, or video stream data. The standard format for storing and transmitting medical images is the Digital Imaging and Communications in Medicine (DICOM) format.

A well-known example from 2004 is the “ITK-Snap”, an annotation tool for navigation and segmenting three-dimensional medical image data [126]. Another prevalent open-source tool widely used in the medical domain is the 3D slicer [25]. 3D slicer is a desktop software for solving advanced image processing tasks in the field of medical applications. It visualizes special medical formats like DICOM in the application and allows editing of the images with the 3D slicer software. Furthermore, the 3D slicer uses AI via a AI-assisted segmentation extension in the 3D slicer application. Thereby, 3D slicer allows automatic segmentation and editing, e.g., CT scans of brains.

Another well-known annotation tool was published in 2015 and is called TrainingData [30, 31]. TrainingData is a traditional annotation tool for labeling AI (computer vision) training images and videos. The annotation tool provides many features, including a labeling support system using AI models. TrainingData also supports the DICOM format.

In 2016 the Radiology Informatics Laboratory Contour (RIL-Contour) was published [86]. RIL-Contour is an annotation tool for medical images that uses deep learning models to label images.

Nowadays, the range of medical segmentation tools has become very broad, as they are usually specialized for many different areas of medicine. For example, ePAD is an open-source platform for segmenting 2D and 3D image data, focusing on radiological images [95].

Another tool is Endometriosis Annotation Tool [63]. A group of developers and gynecologists developed this web-based annotation tool for endoscopy videos. In addition to the traditional features, such as video controls, screenshots, or manual labeling of the images, the option of selecting between different endometriosis types is also offered. The Endometriosis Annotation Tool focuses on specific annotations for surgery.

3.1.2 Contribution and conclusion

The literature on semi-automated annotation tools highlights the need for specialized medical annotation tools due to the high cost of medical annotations and the specific requirements of medical images and videos. Considering the related work FastCAT extends the related work by offering an alternative semi-automated annotation tool that can be utilized for machine learning annotation tasks in a general sense and for specialized gastroenterological applications. FastCAT is the main output of two publications about the optimization of annotation processes in the field of medical video annotation for detection and classification models of this thesis. The paper “Semi-Automated Machine Learning Video Annotation for Gastroenterologists” published in the proceedings of Medical Informatics Europe in 2021 (see Section A.2, [54]) shows the preliminary results for a semi-automated workflow for fast and accurate annotation of polyps in images and videos.

The second publication “Fast machine learning annotation in the medical domain: a semi-automated video annotation tool for gastroenterologists” (see Section A.1, [56]) was published in the journal BioMedical Engineering OnLine in 2021. It consists of a data annotation workflow and framework to create fast and accurate annotations of polyps. The semi-automated annotation process is thereby specially designed to increase the annotation speed for the annotating gastroenterologist. Thereby, it extends the paper “Semi-Automated Machine Learning Video Annotation for Gastroenterologists” by enhancing the annotation tool and further evaluating the annotation workflow.

Gastroenterological data sets are mostly comprised of endoscopic videos, which are tiresome to annotate. Therefore, a framework was implemented in those papers to support the domain experts during this time-consuming process. This framework allows the expert to perform key annotations at the beginning and the end of sequences with pathologies, e.g., visible polyps, instead of annotating every frame in the video sequence.

Afterward, relevant frames will be selected and passed on to an AI model. The AI model uses the annotations to detect and mark the desired object on all following and preceding frames with an annotation. Afterward, a non-expert can adjust and modify the AI predictions and export the results. The annotation speed can be further optimized using a state-of-the-art semi-automated AI model. A prospective study with ten participants showed that semi-automated annotation using the framework doubles the annotation speed of non-expert annotators compared to a traditional annotation approach. The software and framework are open-source[1]. Furthermore, the annotation tool FastCAT allowed the creation of a data set including over 500,000 annotated training images for polyp detection. Those results were crucial for training the polyp detection and classification models shown in this thesis. The following sections will use the created data with our annotation tool to train the automated polyp detection and classification models.

3.2 Polyp detection in still images

3.2.1 Related work

The detection of polyps has a long history of different techniques to identify polyps automatically. The approaches can be separated into two categories. First, the detection models before the rise of machine learning, which use handcrafted features. Machines or gastroenterologists manually created those handcrafted features [39, 47, 60] and then used them as input for a classifier for

---

[1] https://github.com/fastcatai/fastcat

polyp detection. Second, the machine learning detection approaches. Today most object detection tasks in images and videos are done using machine learning models.

Mainly those machine learning models involve a CNN or convolutional steps in their feature creation process. The following two sections illustrate the detection with handcrafted features and the polyp detection using machine learning models in still images. Subsection 3.3 shows the most recent approaches using detection models on video data.

3.2.1.1 Automated detection with handcrafted features

In the late 1990s, the first approaches to computer-assisted detection of polyps were already explored. For example, Krishnan et al. suggested using curvature analysis to identify polyps by their shape [60]. In 2003, Karkanis et al. used the wavelet transform to identify polyps by color and texture [47]. Afterward, Hwang et al. compared polyp features based on curvature, intensity, curve direction, and distance from the edge to distinguish the elliptical shapes of polyps from non-polyp regions [39].

Bernal et al. [9] proposed a different method in which images of polyps were converted to greyscale so that the elevations of the polyps could be detected. Subsequently, they highlighted the outlines of the polyp, which they termed valleys. Based on the intensity of the valleys, polyps could be extracted and localized [9]. In addition, with the help of expert knowledge, rules for recognizing polyps based on certain characteristics such as size, shape, and color were implemented. Newer examples of similar approaches can be found in [40] and [89], both of which use a Support Vector Machine (SVM).

In 2019, real-time detection of polyps with handcrafted features was tested in a clinical setting [50]. The authors used a weighted mix of color, structure, textures, and motion information to identify areas of the image where a polyp might be located. The detection rate was 73 %. Nevertheless, the rise of CNN-based methods in the field of image processing has replaced all these techniques, as CNN methods have been proven to produce better features for automated detection tasks.

3.2.1.2 Automated detection involving machine learning

During the last decade, computer-aided polyp detection has been shaped in particular by various deep-learning methods. An overview of the essential models on still image data sets is available in table 3.2. In particular, research interest has developed in the object recognition capabilities of CNNs.

For example, in 2015, Zhu et al. introduced a seven-layer CNN as a feature extractor with a SVM as a classifier to detect anomalies in endoscopy images [133]. The method was trained on custom data. Other approaches used an existing CNN architecture for polyp localization, called AlexNet [61][111][125]. AlexNet was developed for general image classification and not specifically for medical data. Tajbakhsh et al. [111] suggest not fully training the AlexNet, i.e., starting from random weights but using the already pre-trained weights for polyp detection. To improve the performance of the AlexNet, Yuan et al. [125] first extracted an image section via edge-finding algorithms and used those sections as input to the AlexNet [61]. This resulted in a high recall of 91.76 %, which exceeded other state-of-the-art approaches.

Furthermore, it is shown that transfer learning is a practical approach in the presence of limited data, as is generally the case in the medical field.

In 2018, Mo et al. [80] were the first to use an unmodified Faster R-CNN [92] architecture for polyp detection. This model was trained on the CVC-ClinicDB data. Unlike previous approaches, the Faster R-CNN can detect polyps that are mostly obscured or close to the camera. The model is robust to exposure changes or bubbles, but it misses smaller polyps.

Shin et al. [101] were the first to apply the Inception-Residual Neural Network (ResNet) [108] architecture unmodified for polyp detection. This model was trained on the ASU-Mayo-Video-DB and included two post-processing methods, false positive learning and offline learning, to further improve the performance. Rather than a patch extraction step, this model can use an entire frame for training. However, as with Mo et al., the model has many false positives triggered by polyp-like shapes.

In 2018, Zheng et al. [132] introduced the unmodified YOLO architecture [90] for polyp detection. The advantage of this architecture is that only a single processing step is required, i.e., there is no previous step to extract an Region of Interest (ROI). As a result, the model is faster than the two-step approaches but does not reach real-time capability (16 FPS). The CNN features of white light and narrow-band images differ greatly and should therefore be considered separately. The model was trained on the CVC-CLinicDB, CVC-ColonDB, and custom data.

In 2019, Liu et al. [65] compared different backend models as feature extractors for the Single Shot Detection (SSD) architecture [66]. The considered backend models were ResNet50 [35], VGG16 [129], and InceptionV3 [109], with InceptionV3 scoring the best balanced result. The advantages of the models are their robustness in terms of size and shape, as well as their speed, which is real-time capable at 30 FPS. All considered models were trained on the CVC-ClinicDB, CVC-ColonDB, and ETIS-Larib data. In the future, other improved backend models could further boost the models' performance during polyp detection.

Also in 2019, Zhang et al. [130] applied the SSD-GPNet for polyp detection. The SSD-GPNet is based on the SSD architecture [66] but incorporates information normally lost by the standard pooling layers into the result through various customized pooling methods. Since it is based on the SSD architecture, the model is fast and real-time capable at 50 FPS. Additionally, it has a good recall, particularly for small polyps. Zhang et al. presented another deep learning method for polyp detection and localization [131]. They proposed a special single-shot multibox detector-based CNN model, which reuses displaced information through max-pooling layers to achieve higher accuracy. The model works in real-time at 50 FPS while maintaining a F1-score of 84.24 %. The model was trained on custom data. Bagheri et al. suggested converting input images into three color channels first and then passing them to the neural network. This allows the network to learn correlated information using the preprocessed information to locate and segment polyps [6]. A similar approach by Sornapudi et al. used region-based CNNs to localize polyps in colonoscopy images and Wireless Capsule Endoscopy (WCE) images. During localization, images were segmented and detected based on polyp-like pixels [106].

In addition to CNN's, research also includes other deep learning methods for polyp detection. In 2017, a special sparse autoencoder method called stacked sparse autoencoder was used by Yuan and Meng [123] to detect polyps in WCE images. A sparse autoencoder is an artificial neural network with image manifold constraint, commonly used for unsupervised learning [81]. The described

Table 3.2: This table shows the polyp detection models on still image data sets for polyp detection.

Author	Year	Method	Test data set	F1-score	Speed
Yuan et al. [123]	2017	SSAEIM	Custom Dataset	N/A	N/A
Mo et al.[80]	2018	Faster R-CNN	CVC-ClinicDB	91.7%	17 FPS
Zheng et al. [132]	2018	YOLO	ETIS-Larib	75.7%	N/A
Shin et al. [101]	2018	Faster R-CNN	ASU-Mayo	83.3%	25.2 FPS
Zhang et al. [131]	2019	SSD	ETIS-Larib	84.24%	50 FPS
Zhang et al. [127]	2019	SSD	ETIS-Larib	79.2%	29.8 FPS
Liu et al.[65]	2019	SSD	CVC-ClinicDB	78.9%	30 FPS
Wang et al.[118]	2019	CenterNet	CVC-ClinicDB	97.88%	52 FPS
Liu et al.[67]	2020	ADGAN	Custom	72.96%	N/A
Yuan et al.[124]	2020	DenseNet	Custom	81.83%	N/A
Own contrib. A.3	2020	Ensemble	EDD2020 data	86.34%	30 FPS

sparse autoencoder achieved 98 % accuracy in polyp detection [123]. The system was trained and tested on the ASU-Mayo-Video-DB.

In 2019, Wang et al. [118] were the first to use the AFP-Net architecture for polyp detection. Unlike an SSD model, the anchor free AFP-Net model does not require predefined anchor boxes. Through a context enhancement module (CEM), a cosine ground-truth projection, and a customized loss function, the speed was increased to real-time capable 52.6 FPS. The model was trained on the CVC-ClinicVideoDB.

In 2020, Liu et al. [67] introduced an anomaly detection generative adversarial network (ADGAN) for polyp detection. The ADGAN architecture is based on the WGAN [5]. ADGAN was trained to reconstruct healthy images without polyps. If the model receives an image with a polyp as input, the model cannot reconstruct it. This causes a noticeably large difference between the input and output, which can be easily detected. The problem of connecting the input to the GAN's latency space was solved by implementing a second GAN. In addition, a new loss function was added to improve performance further. The model was trained on custom data. Also in 2020, Yuan et al. [124] established the DenseNet-UDCS architecture for frame classification of WCE images. For this task, the loss function of the DenseNets is adapted, while the overall architecture stays the same [38]. As part of the adaptation, weights are added to compensate for the large imbalance in class size (without or with polyps). Further, the loss function is modified to be class sensitive, i.e., within a class, similar features are learned, while the difference to the features of the other class is maximized. These changes improve the model's performance and can easily transfer to other applications.

3.2.2 Contribution and conclusion

The related work section of polyp detection on still images overviews various deep learning models over the last decade. Several approaches have utilized pre-existing CNN architectures, such as AlexNet and YOLO, while others have introduced specialized CNN architectures, like AFP-

Net and DenseNet-UDCS. Our contributions are alternative polyp detection methods using ensemble techniques and domain knowledge. The article “Endoscopic Detection And Segmentation Of Gastroenterological Diseases With Deep Convolutional Neural Networks” (see A.3, [53]) was published in the workshop proceedings of the EndoCV at the IEEE International Symposium on Biomedical Imaging (ISBI) conference. It deals with detecting diseases, especially polyps, in still images taken from a colonoscopy. The paper resulted from the EndoCV challenge, in which we could achieve the first place by presenting the best detection model.

Previous research on endoscopic computer vision has focused primarily on detecting a single disease, such as polyps. The EndoCV challenge extends this classification task by providing data for different diseases in different organs. The EndoCV has two sub-tasks: Multi-class disease detection which includes the localization of bounding boxes and class labels for five disease classes: polyp, Barret’s Esophagus (BE), suspicious, High-Grade Dysplasia (HGD) and cancer. The second task is region segmentation which includes drawing boundary delineation of detected diseases automatically.

The paper addresses those tasks using deep CNNs. The performance of two general state-of-the-art object detection approaches is evaluated for multi-class disease detection. The first is SSD, and the second is a two-step region proposal-based CNN. The architecture was an ensemble of a YOLOv3 object detector [90] and Faster R-CNN [92] combined with a post-processing step involving domain knowledge. It achieved an F1 score of 86.34 % on the polyp detection task.

A state-of-the-art Cascade Mask R-CNN is used for the region segmentation task. Different backbones of the Cascade Mask R-CNN are evaluated to determine the most efficient one. Data augmentation is used to minimize generalization errors. As the last step, post-processing for specific classes is used to refine the model further. The model achieved a dice score of 69.07 %. The dice score is comparable to the F1-score but is suited for the segmentation task.

The second publication “Bigger Networks are not Always Better: Deep Convolutional Neural Networks for Automated Polyp Segmentation” (see A.4, [52]) was published in the CEUR proceedings of the Medico automatic polyp segmentation challenge. The challenge focuses on the segmentation and detection of polyps in still images.

A deep CNN was well suited for segmentation and detecting polyps in still images. To determine the best-suited architecture, state-of-the-art backbones and two different heads were tested and compared. The final model achieved a dice score of 83.10 % on the challenge’s test set. Furthermore, it was demonstrated that growing network size always increases computational complexity, but more extensive networks do not guarantee increased performance.

3.3 Polyp detection in videos

3.3.1 Related work

Older publications were evaluated on image benchmark data sets, like CVC-ClinicDB [8] discussed in section 3.2. Nowadays, better and more realistic video data sets like the CVC-VideoClinicDB [11] are available and should be used for a state-of-the-art comparison of models. All frames are extracted and annotated in those video data sets with bounding boxes surrounding the polyp. Table 3.3 gives an overview of essential models on video data sets. The approach of Misawa et al. [79]

uses the YOLOv3 architecture evaluated on the SUN-Colonsoly dataset and achieves an F1-score of 87.05 % with a speed of 30 FPS. The methods of Tajbakhsh et al.[111], Yuan et al.[125], Shin et al.[101] and Yuan et al.[124] are already illustrated in the previous section and are additionally evaluated on video data by processing every frame of the video with an image detection algorithm. Therefore those approaches do not include temporal information. The literature shows different ideas for including temporal information in the automated polyp detection task. In the following section, different approaches for this task are discussed in detail.

3.3.1.1 3D convolutions

One approach to include temporal information in polyp detection is the expansion from 2D to 3D CNNs. The literature offers many approaches using 3D CNNs for the task of action recognition [18, 45]. In a 2D CNN, the activity of each neuron is calculated via a discrete convolution. A comparatively small convolution matrix (filter kernel) is moved gradually over the input. A neuron's input in the convolutional layer is calculated as the inner product of the filter kernel with the currently underlying image section. A 3D CNN works similarly but uses a 3D kernel and a 3D input. Thereby moving the filter in three directions (x, y, z). In our application, the third dimension is the time dimension.

Some of the 3D CNN approaches are also considered for automated polyp detection in videos. The approach of Misawa et al. [78] uses a 3D-CNN for automated polyp detection. They report a sensitivity of the AI system of 86 % and a false positive rate of 26 %. Training and evaluation are done on a custom private data set. 3D CNNs are highly computationally demanding, so they currently have no real-time capability.

3.3.1.2 3D ResNet

3D CNNs have been widely adopted for action recognition tasks, with various approaches developed to extend classic 2D CNN architectures such as ResNet to work with 3D data [18, 32]. For example, Itoh et al. [42] used a 3D ResNet architecture to create an automated polyp detection system for videos. The 3D ResNet was trained using a class-balanced loss function with weighted cross-entropy. Additionally, they created a GAN to subsample underrepresented cases. This approach achieved an accuracy of 80 %. However, the large size of the network and the limited amount of data available can lead to overfitting, and the 3D ResNet architecture is not suitable for real-time applications due to its computational demands.

3.3.1.3 Optical flow

Optical flow is the impression of motion conveyed by shifting patterns in successive images of a sequence of images. Therefore optical flow can represent the motion of objects to extract short-term temporal information, and this temporal information can then track and detect an object in video streams.

Zhang et al. [127] used optical flow for real-time detection of polyps. They trained a SSD detector for object detection in their paper. As the structure of the SSD is computationally less expensive than other architecture, it allows real-time detection. The SSD is paired with an optical flow feature generator through a fusion module. The optical flow can easily be computed, allowing the general architecture to achieve real-time capability. Using this architecture, Zhang et al. scored an accuracy of 84.24 % on their custom private data set with a speed of 50 FPS.

3.3.1.4 Structural similarity

Structural Similarity (SSIM) measures or estimates image quality based on an uncompressed or noise-free source image as a reference. SSIM is, e.g., used for measuring the perceived quality of digital television and cinema images.

Xu et al. [122] used SSIM for their real-time polyp detection approach. They are modeling SSIM of consecutive frames using three dimensions: luminance, contrast, and structure. Thereby, consecutive frames with a high correlation in luminance, contrast and structure are considered similar. For the collection of similar frames, Xu et al. remove bounding boxes that only appear in one or two similar frames and do not intersect with previous bounding boxes of similar frames.

They trained a YOLOv3 object detector to detect polyps. Afterward, they combined the YOLOv3 outputs and the SSIM through an inter-frame similarity correlation unit to make final decisions about the polyp detection. The model is trained on custom data and tested on the CVC-VideoClinicDB data set with an F1 score of 75.86 %.

3.3.1.5 Post-processing

Post-processing is a common step in real-time applications to enhance detection results without losing significant application speed. In the paper of Qadir et al. [87], the authors utilize two 2D localization networks SSD [66] and a Faster-R-CNN [92]. Afterward, a false positive reduction unit further processes the network's output. The false positive reduction unit considers an incoming frame's seven preceding and following frames to detect and correct outliers. This outlier removal step results in fewer false positives. As future frames are used for the calculation, the post-processing slightly delays the real-time application pipeline.

Another promising method by Qadir et al. [87] is a model consisting of a two-step process. The first step generates several ROIs, similar to classical detection architectures. Afterward, the ROIs are compared based on the ROIs previous frames and classified as true or false positive frames. The authors assume that consecutive frames in a video are similar to each other. They trained their model on the private ASU-Mayo-Video DB [110] data set and added additional data.

3.3.1.6 Object tracking

Like the post-processing approaches, object tracking is used after detecting a neural network to filter the detection outputs in real-time. Nogueira-Rodríguez et al. [83] used YOLOv3, a 2D convolutional neural network (CNN) object detector, for polyp detection. YOLOv3 was trained on a custom data set of 28,576 images containing 941 polyps. The output of YOLOv3 was then combined with a post-processing step based on an object tracking algorithm to reduce false-positive predictions. The object tracker used the Intersection over Union (IOU) of predicted bounding boxes in preceding frames to filter new predictions, considering only consistent bounding boxes in similar

Table 3.3: Overview of polyp detection models on video data sets.

Author	Year	Method	Test data set	F1-score	Speed
Tajbakhsh et al.[111]	2016	AlexNet	Custom	N/A	N/A
Yuan et al.[125]	2017	AlexNet	ASU-Mayo-Video-DB	N/A	N/A
Shin et al.[101]	2018	Inception ResNet	ASU-Mayo-Video-DB	86.9%	2.5 FPS
Itoh et al.[42]	2019	3D-ResNet	Custom	N/A	N/A
Misawa et al.[78]	2019	3D-CNN	Custom	N/A	N/A
Zhang et al.[127]	2019	SSD-GPNet	CVC-VideoClinicDB	69.8%	40 FPS
Yuan et al.[124]	2020	DenseNet	Custom	81.83%	N/A
Qadir et al.[87]	2020	Faster R-CNN SSD	CVC-VideoClinicDB CVC-VideoClinicDB	84.44% 71.82%	15 FPS 33 FPS
Xu et al.[122]	2021	CNN + SSIM	CVC-VideoClinicDB	75.86%	N/A
Misawa et al. [79]	2021	YOLOv3	SUN-Colonoscopy	87.05%	30 FPS
Rodríguez et al. [83]	2022	YOLOv3 + Object tracking	Custom	88.10%	N/A
Own contrib. A.5	2023	YOLOv5 + RT-REPP	CVC-VideoClinicDB	90.24%	43 FPS

positions with an IOU exceeding a selected threshold. With this approach, the authors achieved an F1-score of 0.88 on a custom test data set.

3.3.2 Contribution and conclusion

Researchers in the field of automated polyp detection in videos use multiple approaches to incorporate the time dimension as detection input, like 3D architecture, optical flow, structural similarity, post-processing and object tracking. Our contribution proposes an alternative system for polyp detection that can operate in real-time and also considers the temporal dimension by incorporating post-processing.

The article “A Real-Time Polyp Detection System with Clinical Application in Colonoscopy Using Deep Convolutional Neural Networks” published in the MDPI journal of imaging in January 2023 (see Section A.5, [59]), presents a fully automated polyp detection device to assist physicians during colonoscopy in real-time.

This work discusses several techniques for training a CNN on polyp detection, which includes preprocessing, data augmentation and hyperparameter optimization. A post-processing step based on video detection was developed to work with a stream of images in real-time. This approach incorporates the endoscope's incoming stream context while maintaining real-time performance. Moreover, the presented polyp detection system is integrated into a physical prototype ready for clinical applications. The polyp detection system was evaluated on the CVC-VideoClinicDB benchmark with an F1-score of 90.24 %, considered state-of-the-art. Addtionally, a new performance metric called "first detection time" is introduced. The first detection time measures the time between the appearance of a polyp and the first detection by the system. Furthermore, it is shown that the trade-off of a higher number of FPS in return for a better recall is more important for clinical application. Therefore, the first detection time is a more accurate metric to measure model performance in clinical applications.

Another paper of this section focuses on the detection of polyps in videos without considering the real-time requirement for clinical application and is called "Deep Learning using temporal information for automatic polyp detection in videos" (see Section A.6, [57]). It was published in the workshop proceedings of the Endoscopic Computer Vision Challenge 2.0 (EndoCV 2.0) 2022 at the ISBI conference.

The field of endoscopic computer vision has mainly focused on polyp detection in single images but not in videos or streams of images. This was the reason for starting EndoCV 2.0. The goal of EndoCV 2.0 is to use streams of image sequences to detect polyps accurately. One approach to solving this challenge presented in this work is based on Gong et al. [27]. The architecture leverages the power of deep CNNs combined with temporal information to improve existing solutions for polyp detection.

The presented detection system combines matching ROI features across multiple frames with temporal attention to predict the polyp detection for the consecutive frame. For evaluation, the shown approach is compared to two traditional image detection models on a validation set based on training data provided by the challenge. The first tested model is a SSD called YOLOv3, and the second model is a two-step region proposal-based CNN called Faster R-CNN. Data augmentation was done to minimize the generalization error, and additional open-source training data was added.

3.3.3 Future work

Currently, real-time detection is only possible with limited neural network size. Nevertheless, the advances in GPU development, also illustrated in Moore's law [98], suggests that the capacity for computational power will increase exponentially in the coming years. Therefore, different architectures with higher computational power, like 3D convolutions or big transformer architectures, will soon find application in real-time medical assistance systems. Hence, the structure of 3D convolutions can be used further to attain more information about the incoming stream of images. E.g., extensions of the ideas of Carreira et al. [18] may be used to exploit the advantages of 3D architecture further. In this paper, 3D convolutional networks are applied to small videos to classify the action performed in the video. Through the 3D networks, Carreira et al. allow the network to incorporate all of the temporal information of the video to create the classification.

One potential approach for improving the performance of the polyp detection system in the future is to utilize large transformer architectures, which have shown strong results in object detection tasks (not real-time) [22, 68]. This approach could be particularly promising if it is feasible to train and integrate a transformer network into the existing polyp detection system. Combining the benefits of the transformer architecture's high detection accuracy with the box correction capabilities of Robust and Efficient Post-Processing (REPP) may create a more powerful and effective polyp detection system. However, it should be noted that this approach would likely require a significant amount of computational power and is currently not feasible for real-time applications.

Another approach for future work is to design a detection framework classifying different pathologies. The most reasonable in the field of medical assistance systems in colonoscopy is the detection of diverticulosis, as it is the second most diagnosed disease by gastroenterologists. Nevertheless, there are no open-source data resources for diverticulosis. To acquire a sufficient amount of training data would require a lot of work. The annotations system presented in this thesis is especially useful to annotate this data set as diverticulosis has the same annotation and preparation requirements as polyps.

3.4 Polyp detection with extended vision

3.4.1 Related work

As there is mostly no similar literature to our contribution, the contribution is placed between automated polyp detection systems and approaches to increase the vision of the gastroenterologist in endoscopy. For a comprehensive overview of the field of automated polyp detection read sections 3.2 and 3.3. Several publications in the literature extend the vision of the gastroenterologist [21, 28, 113]. The first one emerged in 2008 by Triada et al. [113]. The authors referred to the back view as a “Third Eye retrograde auxiliary imaging system” in the study. The authors consider the system safe, technically feasible, and clinically promising. They showed that the system achieves an 11.8 % increase in diagnostic yield. Afterward, similar approaches use mostly the same technology [64, 72, 120]. Primarily this procedure is referred to as TER [21, 28, 64, 72, 113, 120].

Then, more cameras are added to the endoscope resulting in a method called Fuse [28, 84, 105]. Fuse involves adding cameras on both sides of the endoscope and creating a 360-degree view around the endoscope. Gralnek et al. showed the feasibility of Fuse in colonoscopies [28]. Furthermore, Song et al. and Nulsen et al. showed an increase in ADR using Fuse [84, 105]. However, all of these approaches to full-spectrum colonoscopy involve the creation of additional views on additional monitors for the examiner. Installing such monitors may be cumbersome or even unfeasible because of space restrictions. Additionally, showing the examiner two additional screens and the main monitor may overwhelm the examiner and result in additional missed polyps [33].

3.4.2 Contribution and conclusion

The related work section discusses various approaches to increase the view of the endoscopist. These approaches include TER, and Fuse. Nevertheless, these methods may be cumbersome or overwhelming for the examiner. Therefore our contribution introduces a system, which keeps the classic view of the endoscope for the gastroenterologist but adds two additional cameras to the endoscope that are just viewed by an AI trained for polyp detection.

“A User Interface for Automatic Polyp Detection Based on Deep Learning with Extended Vision” (see Section A.7, [58]) is an article published in the conference proceedings of the Medical Image Understanding and Analysis (MIUA) in Cambridge (England) 2022. The publication shows a new animal-tested automated polyp detection approach using two additional cameras. The two additional cameras use the automated detection of polyps to extend the examiner's view.

In the colon, polyps may hide behind folds or uninvestigated areas. Those polyps have a higher chance of being missed by the gastroenterologist. Therefore, as described above, researchers suggest expanding the examiner's view by adding additional cameras to the endoscopes [21, 28, 113]. Nevertheless, these additional views may be overwhelming. The paper, therefore, suggests keeping the classic view of the endoscope for the gastroenterologist but adding additional two views to the endoscope that are just viewed by an AI. If a polyp is found, the examiner can focus on his classic routine but is alarmed by the AI. This prototype is tested in an animal trial using gene-targeted pigs. The results indicated that the AI system could find additional polyps missed on the main endoscope cameras. Nevertheless, the system has limitations. Light conditions significantly impact the detection results, and false detections might occur if the side cameras of the endoscope are too close to the mucosa. The polyp detection system may also falsely detect light reflections, bubbles, or feces.

3.4.3 Future work

Currently, the polyp detection system with extended vision was only applied in animal trials. As the results suggest high usability of the approach, the next step would be to rebuild the endoscope for usage in a human. This involves a long testing period to achieve ethical and medical approvals. The system must be more narrow and tested sufficiently to receive approval for the clinical trial in humans.

Also, the additional cameras used in the system currently require many cables, which could be inconvenient for both the endoscopist and the patient. To address this issue, it would be evident to implement wireless communication between the cameras and the polyp detection system. This could potentially improve the usability and convenience of the system for both the endoscopist and the patient.

Another potential area for further development is the design and user experience of the system. By considering the needs and preferences of endoscopists and patients, it may be possible to create a system that is more intuitive and easy to use, which could improve the adoption and acceptance of the system in the medical community. Finally, it may also be beneficial to explore the potential for integrating the polyp detection system with other medical technologies or software platforms, in order to create a more comprehensive and seamless experience for endoscopists and patients.

3.5 Automated polyp classification

3.5.1 Related work

In the literature, the automated classification of polyps started in 2016. An overview of all of the methods is shown in table 3.4.

In the paper of Ribeiro et al. [93], a CNN is used to classify polyps into healthy and abnormal classes utilizing Kudo's pit-pattern classification. Pit-pattern classification uses the surface structure [62] to classify polyps. The classification system achieves an accuracy of 90.96 %. Another approach using the pit-pattern classification is the paper of Tanwar et al. [112]. This paper uses

Table 3.4: Overview of the related work for the polyp classification task.

Author	Year	Method	Data	Classification	Accuracy
Ribeiro et al. [93]	2016	custom CNN	private	healthy abnormal	90.96 %
Zhang et al. [128]	2016	CaffeNet	private and [76]	hyperplastic adenoma	85.9 %
Bryne et al. [17]	2017	InceptionNet	private	hyperplastic adenoma	94 %
Komeda et al. [51]	2017	custom CNN	private	adenoma non-adenoma	75.1 %
Lui et al. [69]	2019	custom CNN	private	curable non-curable	85.5 %
Bour et al. [16]	2019	ResNet-50	private	not dangerous dangerous cancer	87.1 %
Tanwar et al. [112]	2020	VGG-16	private	Benign Malignant Nonmalignant	84 %
Ozawa et al. [85]	2020	SSD	private	hyperplastic adenoma	83 %
Hsu et al. [37]	2021	custom CNN	private	hyperplastic neoplastic	72.2 % 82.8 % (NBI)
Own contrib. A.8	2022	Transformer	SUN-Colonoscopy	Paris	89.35 %
Own contrib. A.8	2022	Few shot learning	private	NICE	81.13 %

three classes: nonmalignant, malignant, and benign. The authors trained their model on a private data set and yielded an accuracy of 84 %.

Another approach using a CNN for polyp classification is presented by Zhang et al. [128]. The author uses the NICE classification in this paper. The polyps were additionally classified by color and structure as polyp type one or two. Then they are categorized as hyperplastic or adenoma tumors. Zhang et al. pre-trained the neural network on non-medical data. The accuracy of their classification system is 86 %.

Bryne et al. [17] show another approach for the NICE classification. Bryne et al. used only NBI video frames for training and validation of the system. The system is based on a real-time capable CNN model. The accuracy of the system is 94 %, validated on 125 polyps. Another approach with classification utilizing CNN architectures is Komeda et al. [51]. They classify polyps into two categories (adenoma and non-adenoma) using a CNN. They used 10-fold cross-validation and scored an accuracy of 75.1 %.

Lui et al. categorized polyps into curable and noncurable lesions using NBI and white-lighted images [69]. Their classification system classified polyps with an accuracy of 85.5 %. The model performed better on NBI images. Using the Paris classification, Bour et al. classified polyps "Not Dangerous", "Dangerous" and "Cancer". The authors used several well-known CNN architectures. Their system classified polyps with an accuracy of 87.1 % [16] with ResNet50.

Ozawa et al. [85] used a CNN based on a single-shot detector to classify and detect polyps. The model was trained and validated with a private data set, achieving a true-positive rate of 92 %. The system's accuracy was 83 %. Last, Hsu et al. [37] used gray-scaled images to classify polyps. They used a custom-designed classification network embedded, and the network incorporates a detection and classification step. They classified polyps as neoplastic or hyperplastic polyps. The accuracy was 82.8 % for NBI images and 72.2 % for white light images.

3.5.2 Contribution and conclusion

The literature shows different approaches to automated polyp classification using various CNN architectures and classification systems such as Kudo's pit-pattern classification, the Paris classification and the NICE classification. However, these methods have been assessed using only privately collected datasets, which poses a challenge to reproducing and comparing the obtained results. Our contribution is the first evaluation of the Paris classification on an open-source data set with state-of-the-art results. The article "Automated classification of polyps using deep learning architectures and few-shot learning" (see A.8, [55]), which is currently under minor revision from the Journal BMC Medical Imaging, presents two novel approaches to polyp classification. The first approach automatically characterizes polyps based on shape (Paris classification). The second approach characterizes polyps based on texture and surface patterns (NICE classification).

Gastroenterologists classify polyps using different classification systems. Further treatment is based on the classification of those polyps. The classification of polyps is not easy and misclassifying polyps may lead to additional difficulties with further treatments. Therefore in this paper, we presented two novel automated polyp classification systems. One system for Paris and one for NICE classification.

The Paris classification model involves a two steps process. First, the polyp is detected and cropped through a polyp detection system, and the cropped image is input for a transformer network in the second step. By achieving state-of-the-art results, the Paris classification system yields an accuracy of 89.35 % on a public benchmark data set. The NICE classification system utilizes a model based on deep metric learning. Thereby an embedding space for polyps is created. This embedding space allows the classification using a small amount of data. The NICE classification system produces a competitive accuracy of 81.13 %. Thereby the approach demonstrates the utility of few-shot learning for polyp classification with a low amount of data. The study shows different ablations of both systems and further elaborates on the explainability of the Paris classification system.

3.5.3 Future work

In order to evaluate the effectiveness of the polyp classification system, it is beneficial to conduct a clinical trial in which the system is applied in a test environment involving gastroenterologists with varying levels of expertise. By measuring the accuracy of the polyp classification both with and without the use of the AI assistant system, it is possible to determine the value of the system in improving diagnostic accuracy.

Additionally, as the classification system is refined and improved, it has the potential to increase the speed of documentation and free up time for gastroenterologists to perform other tasks. Thereby, the integration of the system into an automated documentation system could further streamline gastroenterological examinations.

Another approach to foster research in automated NICE and Paris classification is to collect and release additional annotated data sets for the research community. By providing open-source benchmark data sets, similar to those that already exist in the polyp detection field, researchers would have access to a larger pool of data to use in evaluating and comparing different architectures and models. This would facilitate more comprehensive and fair evaluations of different approaches and could accelerate the development of new and improved polyp classification systems. Additionally, the release of open-source data sets might foster collaboration and facilitate the sharing of knowledge and resources among researchers in the field.

4 Conclusion

The use of deep learning in computer vision has led to significant progress in various tasks, including medical imaging applications in gastroenterology. This thesis aims to show how automated polyp detection and classification systems can assist gastroenterologists in their endoscopic examinations of colon cancer. Thereby, the objective is guided by the following research question introduced in chapter 1:

Can machine learning assist physicians in endoscopic examinations with the treatment of colon cancer?

Yes, as the literature, the developed polyp detection and classification systems of this thesis and the results of the publications (A.5, B.2) suggest machine learning does assist physicians with the treatment of colon cancer. To answer this question in detail, the guiding research question was divided into four sub-questions (RQ1 - RQ4):

RQ1 *Can automated polyp detection assist gastroenterologists in their daily clinical practice?*

Yes, an automated polyp detection system assisting gastroenterologists in real-time is developed and presented. The system can be downloaded[1] and applied in a clinical setting (A.5). Two publications laid the groundwork for the final system (see A.3, A.4) and another is an extension of the system (A.6). The presented system consists of a fast and accurate object detector paired with a novel post-processing technique. It achieves state-of-the-art performance on the open-source CVC VideoClinicDB data set [29] with an F1 score of 90.25 %.

Preliminary clinical trial results show an increased ADR when using the automated polyp detection system (see B.2). Furthermore, the system shows a high usability score of 96.3 (max. 100).

RQ2 *How does a semi-automated annotation tool impact the workload of gastroenterologists and the quality of annotated data in endoscopic imaging for polyp detection?*

An essential component for any machine learning system is high quality data (see Section A.1 and A.2). Therefore, an annotation tool specialized in gastroenterological annotations is created. This system increases the annotation speed of a domain expert annotator by a factor of 20. This is done with a specialized data filtering method and semi-automated AI-assisted annotation. The annotation tool is used to create a polyp detection and classification data set with over 500.000 annotated images.

---

[1] https://fex.ukw.de/public/download-shares/XllQRkZUhWVZcJqSrMIndkSfq07afWBV

RQ3 *How does the integration of additional cameras with AI technology enhance the detection of polyps and improve the examination process compared to traditional methods?*

Extending the endoscopist with additional cameras to achieve a full-spectrum colonoscopy helps the endoscopist find additional polyps. Assisting this full-spectrum colonoscopy with AI enables easier control and more focus for the endoscopist. An AI assistant endoscope prototype with extended vision is tested with gene-targeted pigs (see Section A.7). In the animal trial, 13 additional polyps were just found through the system's assistance. Additionally, the system achieves a F1 score of 72.13 % on the side camera data.

RQ4 *Can deep learning methods achieve high accuracy on the classification of polyps in gastroenterology and does few-shot learning improve the efficiency of the classification process?*

Yes, this thesis also presented two novel algorithms for the automatic classification of polyps (see Section A.8): First, a system for classifying polyps based on shape (Paris classification) is presented. A two-step process is required for the classification. This involves a polyp detection step that uses this thesis's main polyp detection architecture. Next, the image is cropped and afterward fed into a transformer network. The transformer is then classifying the cropped image. The Paris classification model exceeds state-of-the-art results on a publicly available data set.

Second, a polyp classification system based on texture and surface (NICE). The NICE classification system uses deep metric and few-shot learning to classify polyps. The model achieves an F1 score of 81.13 % on a publicly available data set.

A Lead author publications

Krenzer *et al. BioMedical Engineering OnLine* (2022) 21:33
https://doi.org/10.1186/s12938-022-01001-x

BioMedical Engineering OnLine

RESEARCH **Open Access**

Fast machine learning annotation in the medical domain: a semi-automated video annotation tool for gastroenterologists

Adrian Krenzer[1*], Kevin Makowski[1], Amar Hekalo[1], Daniel Fitting[2], Joel Troya[2], Wolfram G. Zoller[3], Alexander Hann[2] and Frank Puppe[1]

*Correspondence: adrian.krenzer@uni-wuerzburg.de

[1] Department of Artificial Intelligence and Knowledge Systems, Sanderring 2, 97070 Würzburg, Germany
Full list of author information is available at the end of the article

Abstract

Background: Machine learning, especially deep learning, is becoming more and more relevant in research and development in the medical domain. For all the supervised deep learning applications, data is the most critical factor in securing successful implementation and sustaining the progress of the machine learning model. Especially gastroenterological data, which often involves endoscopic videos, are cumbersome to annotate. Domain experts are needed to interpret and annotate the videos. To support those domain experts, we generated a framework. With this framework, instead of annotating every frame in the video sequence, experts are just performing key annotations at the beginning and the end of sequences with pathologies, e.g., visible polyps. Subsequently, non-expert annotators supported by machine learning add the missing annotations for the frames in-between.

Methods: In our framework, an expert reviews the video and annotates a few video frames to verify the object's annotations for the non-expert. In a second step, a non-expert has visual confirmation of the given object and can annotate all following and preceding frames with AI assistance. After the expert has finished, relevant frames will be selected and passed on to an AI model. This information allows the AI model to detect and mark the desired object on all following and preceding frames with an annotation. Therefore, the non-expert can adjust and modify the AI predictions and export the results, which can then be used to train the AI model.

Results: Using this framework, we were able to reduce workload of domain experts on average by a factor of 20 on our data. This is primarily due to the structure of the framework, which is designed to minimize the workload of the domain expert. Pairing this framework with a state-of-the-art semi-automated AI model enhances the annotation speed further. Through a prospective study with 10 participants, we show that semi-automated annotation using our tool doubles the annotation speed of non-expert annotators compared to a well-known state-of-the-art annotation tool.

Conclusion: In summary, we introduce a framework for fast expert annotation for gastroenterologists, which reduces the workload of the domain expert considerably while maintaining a very high annotation quality. The framework incorporates

© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

a semi-automated annotation system utilizing trained object detection models. The software and framework are open-source.

Keywords: Machine learning, Deep learning, Annotation, Endoscopy, Gastroenterology, Automation, Object detection

Background

Machine learning especially deep learning is becoming more and more relevant in research and development in the medical domain [1, 2]. For all of the supervised deep learning applications, data is the most critical factor in securing successful implementation and sustaining progress. Numerous studies have shown that access to data and data quality are crucial to enable successful machine learning of medical diagnosis, providing real assistance to physicians [3–7]. Exceptionally high-quality annotated data can improve deep learning detection results to great extent [8–10]. E.g., Webb et al. show that higher data quality improves detection results more than using larger amounts of lower quality data [11]. This is especially important to keep in mind while operating in the medical domain, as mistakes may have fatal consequences.

Nevertheless, acquiring such data is very costly particularly if domain experts are involved. On the one hand domain, experts have minimal time resources for data annotation, while on the other hand, data annotation is a highly time-consuming process. The best way to tackle this problem is by reducing the annotation time spend by the actual domain expert as much as possible while using non-experts to finish the process. Therefore, in this paper, we designed a framework that utilizes a two-step process involving a small expert annotation part and a large non-expert annotation part. This shifts most of the workload from the expert to a non-expert while still maintaining proficient high-quality data. Both of the tasks are combined with AI to enhance the annotation process efficiency further. To handle the entirety of this annotation process, we introduce the software Fast Colonoscopy Annotation Tool (FastCat). This tool assists in the annotation process in endoscopic videos but can easily be extended to any other medical domain. In the domain of endoscopic imaging, the main issue of clinical experts is to find and characterize pathologies, e.g., polyps in a screening colonoscopy. Thereby, the endoscopist examines the large intestine (colon) with a long flexible tube that is inserted into the rectum. A small camera is mounted at the end of the tube, enabling the physician to look inside the colon. The images from this camera can be captured and annotated to enable automatic real-time detection and characterization of pathologies [12, 13]. This process and other applications all need annotated data to enable high-quality results.

The main contributions of our paper are:

1. *We introduce a framework for fast expert annotation, which reduces the workload of the domain expert while maintaining very high annotation quality.*
2. *We publish an open-source software for annotation in the gastroenterological domain and beyond, including two views, one for expert annotation and one for non-expert annotation.*[1]

---

[1] https://github.com/fastcatai/fastcat.

(3) *We incorporate a semi-automated annotation process in the software, which reduces the annotation time of the annotators and further enhances the annotation process's quality.*

To overview existing work and properly allocate our paper in the literature, we describe a brief history reaching from general annotation tools for images and videos to annotation specialized for medical use.

A brief history of annotation tools

As early as the 1990s, the first methods were conceived to collect large datasets of labeled images [14]. E.g., "The Open Mind Initiative", a web-based framework, was developed in 1999. Its goal was to collect annotated data by web users to be utilized by intelligent algorithms [15]. Over the years, various ways to obtain annotated data have been developed. E.g., an online game called ESP was developed to generate labeled images. Here, two random online players are given the same image and, without communication, must guess the thoughts of the other player about the image and provide a common term for the target image as quickly as possible [14, 16]. As a result, several million images have been collected. The first and foremost classic annotation tool called labelme was developed in 2007 and is still one of the most popular open-source online annotation tools to create datasets for computer vision. Labelme provides the ability to label objects in an image by specific shapes, as well as other features [17]. From 2012 to today, with the rise of deep learning in computer vision, the number of annotation tools expanded rapidly. One of the most known and contributing annotation tools is LabelImg, published in 2015. LabelImg is an image annotation tool based on Python which utilizes bounding boxes to annotate images. The annotations are stored in XML files that are saved in either PASCAL VOC or YOLO format. Additionally, in 2015 Playment was introduced. Playment is an annotation platform to create training datasets for computer vision. It offers labeling for images and videos using different 2D or 3D boxes, polygons, points, or semantic segmentation. Besides, automatic labeling is provided for support. In 2017, Rectlabel entered the field. RectLabel is a paid labeling tool that is only available on macOS. It allows the usual annotation options like bounding boxes as well as automatic labeling of images. It also supports the PASCAL VOC XML format and exports the annotations to different formats (e.g., YOLO or COCO JSON). Next, Labelbox, a commercial training data platform for machine learning, was introduced. Among other things, it offers an annotation tool for images, videos, texts, or audios and data management of the labeled data.

Nowadays, a variety of image and video annotation tools can be found. Some have basic functionalities, and others are designed for particular tasks. We picked five freely available state-of-the-art annotation tools and compared them more in-depth. In Table 1, we shortly describe these tools and compare them.

Computer Vision Annotation Tool (CVAT)

CVAT [18] was developed by Intel and is a free and open-source annotation tool for images and videos. It is based on a client-server model, where images and videos are organized as tasks and can be split up between users to enable a collaborative working

Table 1 Comparison between video and image annotation tools

	Tool	CVAT	LabelImg	labelme	VoTT	VIA
	Image	●	●	●	●	●
	Video	●	-	-	●	●
	Usability	Easy	Easy	Medium	Medium	Hard
Formats	VOC	●	●	●	●	-
	COCO	●	-	●	-	●
	YOLO	●	●	-	-	-
	TFRecord	●	-	-	●	-
	Others	-	-	●	●	●

process. Files can be inserted onto the server through a remote source, mounted file system, or uploading from the local computer. Before a video can be annotated, it must be partitioned into its frames, which then can be annotated. Several annotation formats are supported, including the most common formats such as VOC, COCO, YOLO and TFRecord. Available annotation shapes and types are labeling, bounding boxes, polygons, polylines, dots, and cuboids. CVAT also includes features for a faster annotation process in videos. The disadvantages of this tool are that it currently only supports the Google Chrome browser, and due to the Chrome Sandbox, performance issues could appear.

LabelImg

LabelImg [19] is an image annotation tool that is written in Python and uses the Qt framework as a graphical user interface. It can load a bulk of images but only supports bounding box annotations and saves it as a XML file in VOC or YOLO format. The functionalities are minimal but sufficient for manual annotation of images. Furthermore, it does not contain any automatic or semi-automatic features which could speed up the process.

labelme

The annotation tool labelme [20] is written in Python, uses Qt as its graphical interface and only supports image annotation. It is advertised that videos could be annotated with this tool, but no video annotation function was found and the user must manually extract all frames from the video beforehand. Also, there are no automatic or semi-automatic features available and uses basic shapes like polygons, rectangles, circles, points, lines and polylines to annotate images.

Visual Object Tagging Tool (VoTT)

Microsoft's tool VoTT [21] is open-source and can be used for images and videos. Since it is written in TypeScript and uses the React framework as a user interface, it is possible to use it as a web application that can run in any web browser. Alternatively, it can also run locally as a native application with access to the local file system. Images and videos are introduced to the program via a connected entity. This can be a path on the local file system, a *Bing* image search query via an API key, or secure access to an *Azure Blob*

Storage resource. Available annotation shapes are rectangles and polygons that can be tagged. These can then be exported for the *Azure Custom Vision Service* and *Microsoft Cognitive Toolkit* (CNTK).

VGG Image Annotator (VIA)

VIA [22, 23] is a tool that runs in a web browser without further installation and is only build from HTML, JavaScript, and CSS. It can import and export annotations from COCO and a VIA-specific CSV and JSON. The available annotation shapes are polygons, rectangles, ellipses, lines, polylines, and points. Video annotation features the annotation of temporal segments to mark, e.g., a particular activity within the video. Defined segments of the track can also annotate an audio file. VIA does not contain any automatic functionalities within the tool itself; these are relatively independent steps. These steps can be broken down to: Model predicts on frames, save predictions so that they can be imported into VIA, and lastly, check and update annotations if necessary.

Medical annotation tools

With the considerable increase in interest and progress in machine learning in our society the need for machine learning models shifts in different domains including medicine. Thus, artificial intelligence can be used to assist medical professionals in their daily routines [24–26]. As a result, the need for labeled medical images and videos is also a major issue for medical professionals. While it is possible to use common annotation tools such as those already described above, some annotation tools have already been adapted to medical conditions. A well-known example from 2004 is "ITK-Snap", a software for navigating and segmenting three-dimensional medical image data [27].

Another example is an open-source tool widely used in the medical domain called 3D slicer [28]. 3D slicer is a desktop software to solve advanced image computing challenges in the domain of medical applications. Thereby, it is possible to visualize special medical formats like DICOM (Digital Imaging and Communications in Medicine) in the tool and edit it with the 3D slicer software. Additionally, 3D Slicer incorporates Artificial Intelligence (AI) via AI-assisted segmentation extension in the 3D slicer software (DeepInfer, TOMAAT, SlicerCIP, Nvidia Clara). Thereby, automatic segmentations can be created and edited for, e.g., CT scans of brains.

"ePAD" is an open-source platform for segmentation of 2D and 3D radiological images [29]. The range of medical segmentation tools has become very broad nowadays, as they are usually specialized for many different areas of medicine.

Another annotation tool published in 2015 is TrainingData [30, 31]. TrainingData is a typical annotation tool for labeling AI (computer vision) training images and videos. This product offers good features, including labeling support through built-in AI models. TrainingData also supports DICOM, a widespread format in the medical domain.

In 2016 Radiology Informatics Laboratory Contour (RIL-Contour) was published [32]. RIL-Contour is an annotation tool for medical image datasets. Deep Learning algorithms support it to label images for Deep Learning research.

The tool most similar to ours is Endometriosis Annotation Tool [33]. The software, developed by a group of developers and gynecologists, is a web-based annotation tool for endoscopy videos. In addition to the classic functions such as video controls, screenshots, or manual labeling of the images, the option of selecting between different endometriosis types is also offered here.

Nevertheless, most of these medical annotation tools are not suitable for our comparison as they only work with images or are too specialized. The most suitable would be Endometriosis Annotation Tool, but the application is focused on specific annotations for surgery and those do not allow the creation of bounding box annotations which are crucial for our gastroenterological annotations. Therefore, we choose a common, well-known state-of-the-art tool CVAT, for our comparison.

Results

This section presents the results of our introduced tool FastCAT and compares it to the well-known state-of-the-art annotation tool CVAT. We start by introducing our data acquisition and experimental setup. We show our results of the non-expert annotators, which suggests that our tool outperforms the state-of-the-art tool CVAT. We further show how the semi-automated AI annotation affects the annotation speed. Finally, we show our results of the expert annotator, which underline the time advantage using our tool.

Data acquisition and experimental setup

For our evaluation, we used two data sets: The GIANA data set and our data set created at a German clinic called “University Hospital Würzburg”[2]. The GIANA dataset is openly accessible[3] [34]. It is the first polyp dataset published, which includes videos. Former open-source datasets like CVC clinic database [35] or ETIS-LaribPolypDB [36] only provide single images. The GIANA dataset consists of 18 annotated polyp sequences. It is a standard dataset that has been used before for model benchmarking in different publications [37–39]. Therefore, we can reliably use it for evaluating the quality of our results. On average, the data set has 714 frames per video. According to their references, all annotations are done by expert gastroenterologists. We randomly selected two videos from the 18 available ones in GIANA for our evaluation, which turned out to be videos number 8 and 16.

Our data set is composed of an additional 8 videos. These videos include full colonoscopies and therefore have to be filtered first. For the filtering process, we used the method introduced in this paper. Furthermore, we contacted an expert gastroenterologist from the University Hospital Würzburg for the expert annotation. Since the expert annotation time of gastroenterologists is very costly and difficult to obtain, we could only manage to receive the work of two experts. In a second process, the expert annotators select the part of the video, including polyps, as explained in section Methods. However, since this annotation process is not yet completed, we can only evaluate the improvement in annotation speed and not the annotation quality with our dataset.

For the prospective study, all participants receive ten videos for polyp annotation. The videos are randomly selected and then given to the participants. For our preliminary evaluation, ten non-expert annotators are instructed to use our annotation

---

[2] https://www.ukw.de/en.

[3] https://endovissub2017-giana.grand-challenge.org.

Table 2 Comparison of FastCAT and CVAT by video. This table shows our comparison of the well-known CVAT annotation tool to our new annotation tool FastCAT in terms of annotation speed. Videos 1 and 2 are open source and annotated. Videos 3–10 are from the University Hospital Würzburg

	Speed (SPF)		**Total time (min)**		**Video information**		
	CVAT	**FastCat**	**CVAT**	**FastCat**	**Frames**	**Polyps**	**Framesize**
Video 1	3.79	1.75	23.43	10.82	371	1	384x288
Video 2	4.39	2.49	32.85	18.63	449	1	384x288
Video 3	2.82	1.42	60.11	30.27	1279	1	898x720
Video 4	4.09	2.00	56.85	27.80	834	1	898x720
Video 5	4.57	2.39	53.24	27.84	699	2	898x720
Video 6	1.66	0.61	18.01	6.62	651	1	898x720
Video 7	1.70	0.64	11.22	4.22	396	1	898x720
Video 8	1.55	0.76	34.13	16.73	1321	2	898x720
Video 9	1.87	0.88	34.91	16.43	1120	1	898x720
Video 10	2.74	0.92	77.68	26.08	1701	4	898x720
Mean	2.92	1.39	40.24	18.54	882	1.5	795x633

Table 3 Comparison of FastCAT and CVAT by user. This table shows our comparison of the well-known CVAT annotation tool to our new annotation tool FastCAT in terms of quality of annotation and annotation speed. The quality metric is the F1-score. We count a TP if the drawn box matches the ground truth box more than 70 %

	Quality (%)		**Speed (SPF)**		**Total time (min)**		**Medical Experience**
	CVAT	**FastCat**	**CVAT**	**FastCat**	**CVAT**	**FastCat**	
User 1	99.30	99.50	7.33	3.71	48.78	25.30	Low
User 2	98.85	98.90	3.47	1.88	23.38	13.70	Low
User 3	97.97	98.51	4.59	1.53	31.28	11.17	Low
User 4	98.93	99.75	5.12	2.57	33.96	16.53	Middle
User 5	98.53	98.83	5.41	2.49	37.00	18.10	Middle
User 6	98.52	99.23	4.04	3.24	27.90	24.95	Low
User 7	99.45	99.30	5.20	2.70	35.01	21.28	Middle
User 8	99.35	99.08	5.25	2.86	33.90	19.57	Low
User 9	99.12	98.54	4.12	2.25	27.12	14.99	Low
User 10	98.93	99.48	5.63	2.76	37.53	19.89	Low
Mean	98.98	99.03	5.79	2.93	33.59	18.55	Low

tool and the state-of-the-art annotation tool CVAT. Finally, all non-expert annotators receive our software FastCAT and a java tool for measuring the time. The expert annotator starts with annotation, as explained in “Methods”. He annotates Paris classification [40], the size of the polyp, and its location. Additionally, the expert annotates the start and end frame of the polyp and one box for the non-expert annotators. Afterwards, the AI calculates predictions on these frames. The results of the AI are given to the non-expert annotators, who then only correct the predicted boxes. The non-expert annotators in this experiment are students from computer science, medical assistance, and medical secretary. All non-expert annotators are instructed to annotate the polyp frames as fast and as accurately as they can.

Results of the non-expert annotators

We evaluated the tool with 10 different gastroenterological videos containing full colonoscopies. The results are shown in Table 2 and in Table 3. As mentioned previously, we only evaluate the quality of the annotation in two videos from the openly accessible GIANA dataset. The accuracy of the annotations is thereby calculated by comparing the ground truth box of the already annotated open-source GIANA dataset with our newly created annotations. The quality evaluation is done via the F1-score. The F1-score describes the harmonic mean of precision and recall as show in following equations:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$
$$F_1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 * \text{TP}}{2 * \text{TP} + \text{FP} + \text{FN}}.$$

We count an annotation as true positive (TP) if the boxes of our annotators and the boxes from the GIANA dataset have an overlap of at least 70%. Our experiments showed high variability between individual experts. We, therefore, concluded that a higher overlap is not attainable. Hence, to ensure reasonable accuracy, we choose an overlap of 70% which has been used in previous studies [41–43]. To determine annotation speed, we first measure the speed of the non-expert annotators in seconds per frame (SPF). On average, our annotators take 2.93 s for annotating one image while maintaining a slight advantage in annotation quality. Overall, our semi-automated tool's annotation speed is almost 2x faster than the CVAT annotation tool, with 5.79 s per image. In addition, we evaluate the average time non-expert annotators spend annotating an entire video. The average video takes 18.55 min to annotate. In comparison, using the CVAT tool takes 40.24 min on average per video. Due to some faulty prediction results of the AI, the annotators sometimes delete boxes and draw new boxes as some polyps may be hard to find for the AI. This leads to higher annotation time in the case where polyps are mispredicted. Nevertheless, our tool is self-learning, and increasing amounts of high-quality annotations improve the prediction quality of the AI. This, in turn, speeds up the annotation process further. We elaborate on this in detail in the following subsection. To include more information concerning the video data, we include the number of frames per video, the number of polyps per video, and each video's frame size. The videos provided by our clinic (Videos 3-10) have a higher resolution and a higher frame rate than videos gathered from different institutes. Overall the quality evaluation results show that almost similar annotation results to those of gastroenterology experts are achieved. For speed, our tool outperforms the CVAT tool in any video. In two videos, our tool is more than twice as fast as the CVAT tool.

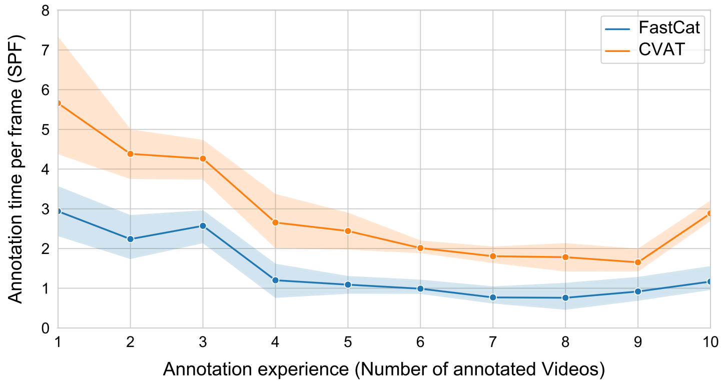


Fig. 1 Learning process of the non-expert annotators. The figure shows the speed of the annotator in seconds per frame (SPF) over the annotation experience measured by the total number of annotated videos by that point for both our tool and CVAT

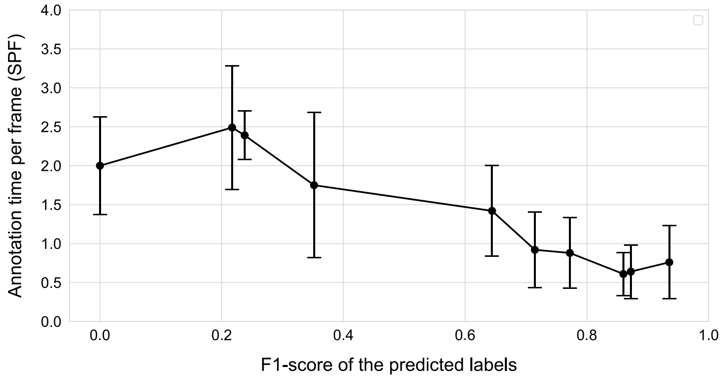


Fig. 2 Effect of AI performance on annotation speed. Plotted are the speed of the annotators in seconds per frame over the AI performance given by its F1-score on a video-by-video basis, where the AI used for prediction is the same for each video. Every point is computed as the average over all annotators

Learning process of the non-expert annotators

Figure 1 shows the learning process of the non-expert annotators, in blue using our tool and in orange using CVAT. The figure shows that the annotation of the first videos takes longer than annotating the subsequent ones since the non-expert annotator has to get to know the software and needs to adjust the software to his preferences. Therefore, annotation speed using both tools improves by further usage, and both tools feature a similar learning curve. However, this learning process slows down after the annotation of about 4 to 5 videos. After this amount of videos, annotators are well accustomed to the software and can competently use most features. In addition, Fig. 1 shows that this learning process is faster using our tool in comparison to the CVAT tool. This may be due to the information provided before use, the calculation we built directly into the software, and our user-friendly environment. Besides all, the CVAT software also shows excellent progress in learning

Table 4 Comparison of CVAT and FastCAT. The tables show the reduction of annotation time of the domain experts. Tgca stands for the time gained compared to annotation with CVAT and is the reduction of workload in %. Video 1 and video 2 are not used for this analysis as the open-source data do not provide full colonoscopies, but just polyp sequences and therefore it is not possible to perform an appropriate comparison

	Total time (min)		**Tgca (%)**	**Video information**		
	FastCat	**CVAT**		**Length (min)**	**Freezes**	**Polyps**
Video 3	0.50	60.11	99.15	15.76	2	1
Video 4	0.67	56.85	98.82	17.70	6	1
Video 5	1.09	53.24	97.95	23.12	4	2
Video 6	0.77	18.01	95.72	6.30	2	1
Video 7	0.70	11.22	93.79	13.05	5	1
Video 8	1.78	34.13	94.76	27.67	13	2
Video 9	1.50	34.91	95.70	20.53	4	1
Video 10	2.92	77.68	96.24	24.36	15	4
Mean	1.24	43.26	96.52	18.56	6.38	1.62

worth mentioning. We can even see annotators who use any of the two tools more frequently further improve their annotation speed up to 9 videos. However, after 8 to 9 videos, the annotation speed decreases. This may be due to two repetitions of the same process that may bore the non-expert annotator and, therefore, decrease annotation speed. Our data show that this effect is more prominent for CVAT than for our tool.

Impact of polyp pre-annotations

To further analyze the improvements in our framework, we investigate the impact of polyp detection on the annotation speed. We compare the final annotated videos with the predictions done during the investigated videos. For ten videos, we calculated the F1-score based on the analysis above. A higher F1-score implicates more detected polyps with less false positive detection. Then, we rank the videos according to their F1-score and display the annotation speed in seconds per frame (SPF), shown in Fig. 2. Overall, a high F1-score leads to a faster annotation speed. Nevertheless, as seen in Fig. 2 if the F1-score is low, the annotation speed at times is faster without any predictions, e.g., from 0.2 to 0.4. Furthermore, low F1-scores show a higher standard deviation in the labeling speed. This means that with a higher F1-score, the variance of the non-expert annotators' labeling speed decreases and therefore the overall performance is increased. Furthermore, we emphasize that continuing the annotation process and retraining the system detection results will increase, and therefore, the annotation speed will increase.

Results of the expert annotators

This subsection demonstrates the value of the tool for domain expert annotation. As domain experts are very costly, we only had two experts available for our study. Therefore, our evaluation between domain experts could not be done quantitatively. Nevertheless, we can qualitatively compare the amount of time the domain experts took to annotate our collected colonoscopies. This is shown in Table 4. On average, our gastroenterologists spend 1.24 min on a colonoscopy. Our final results show that we achieved qualitatively similar results to the GIANA dataset annotation. The expert annotators only take 0.5 to 1 minutes per video using our method, while taking at least 10-80 minutes per video using the CVAT software. Therefore, we can reduce the amount of time a domain expert has to spend on annotation by 96.79 % or by a factor of 20 on our data. This reduction is primarily due to expert and non-expert annotation structure, which reduces the expert's effort tremendously.

Discussion with limitations

By implementing a novel workflow consisting of both algorithmic and manual annotation steps, we developed a tool that significantly reduces the workload of expert annotators and improves overall annotation speed compared to existing tools. In this section, we highlight and discuss the impacts of our study, show the limitation of our presented work and propose new approaches to advance our study further.

Key features and findings

Our results show that by pre-selecting relevant frames using a combination of our freeze-frame detection algorithm and further, low-demand expert annotations and by using AI predictions for bounding box suggestions, we significantly increase the annotation speed while maintaining and even increasing annotation accuracy (see Tables 2 and 3). It is important to note that this improvement is not due to more annotation experience with one tool over the other since the annotators used the tools in an alternating fashion with random video order. Figure 1 further stresses this fact by showing a similar learning curve for both tools, with our tool being shifted down to shorter annotation times. In both cases, the annotation experience (i.e., adjustment to the tool) increases up to around seven videos or 10,000 annotated frames. The annotation speed first saturates and then increases again, possibly due to a human exhaustion effect of doing the same task for an extended duration [44].

Additionally, we inspected the effect of the prediction performance on the annotation speed. As shown in Fig. 2, there is a clear trend towards faster annotation time with better AI performance. The annotator works faster if the suggested bounding boxes are already in the correct location or only need to be adjusted slightly by drag and drop. If the predictions are wrong, the annotator needs to move the boxes further, perhaps readjust the size more, or even delete boxes or create new ones. However, the AI improvement saturates at an F1-score of around 0.8, where better AI performance does not equate to faster annotation speed. Additionally, the range of error is much more significant for the worse performing videos, so this point warrants further inspection in future studies. Nevertheless, it is apparent here that an AI only needs to be good enough instead of perfect to improve annotation speed significantly.

Finally, the results in Table 3 suggest that medical experience does not affect either the annotation speed or performance. The frame detection algorithm combined with the expert frame annotations and our AI's pre-detection provides enough feasibility for the non-experts to adjust the suggested annotations fast and accurately regardless of experience. However, it should be noted that the range of speeds across our non-expert annotators is more stable for middle experience annotators than low experience ones.

All in all, our tool significantly improves the annotation workflow, specifically in the domain of gastroenterology, where specialized tools are scarce. The annotation speed is more than doubled while keeping the same accuracy as other state-of-the-art tools and keeping the cost for expert annotators low.

Limitations of the study

In this subsection, we will shortly discuss the limitations of our analysis and provide an outlook for future studies.

First of all, we did not consider the difficulty of the video when analyzing annotation time. Some videos contain more and harder to detect polyps and thus provide a bigger challenge for both the AI and the annotator. The effect of video difficulty directly correlates to the AI performance in Fig. 2, where the standard error for low-F1 videos is much higher compared to the better ones. Some annotators can efficiently deal with false predictions, while others have more difficulties with those. Additionally, the total annotation time was measured from beginning to end for a video. While the applet we provided for the annotators includes a pause button, minor deviations, like checking their phone, are not removed from our total time measured. These statistical deviations could be removed by dividing the videos into difficulty categories and analyzing each category separately. We need more data or more annotators, where small statistical outliers should be averaged out.

Additionally, with only three medical assistants and seven non-experts, we need further tests to see if medical experience significantly affects annotation time and quality. As discussed above, Table 3 suggests that medium experience annotators work more consistently, whereas low experience ones can be both faster and slower than the medical assistants. These findings can be examined further in future studies with more annotators from various backgrounds, especially those with high medical experience.

Finally, we only indirectly measured the effect of bounding box pre-detection, where our non-expert annotators had no pre-detection for CVAT and suggestions with our tool. Thus, the improvement in annotation speed could also be due to our tool simply being easier to use and having a better user interface (UI) than CVAT. For future analysis, we intend to have the non-expert annotators annotate videos twice, once with bounding box suggestions and once without. However, both times they will use our tool. This way, we will be able to analyze the effect of the pre-detection directly.

Limitations of the tool and future improvements

While our freeze-frame detection algorithm is specific to the domain of gastroenterology, the specific method for detecting relevant frames can be exchanged for a function more suited to the annotators' domain. Additionally, while we only utilized the tool for polyp detection, it can be easily extended to feature more than one pathology, like diverticulum or inflammation. Since frame-wide annotations are separate from bounding

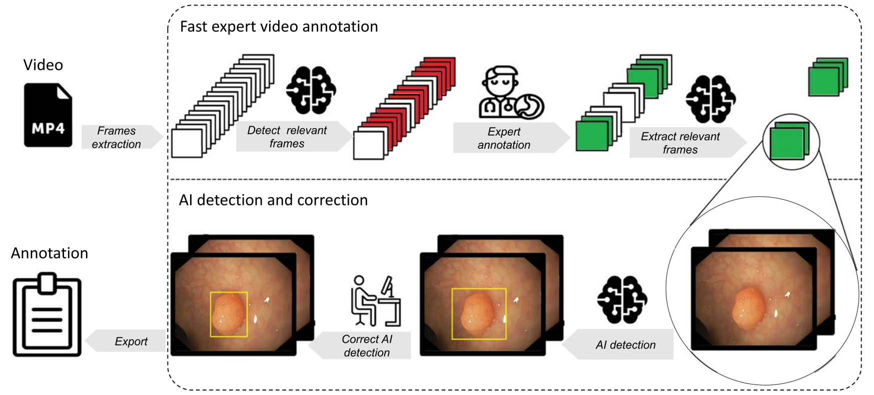


Fig. 3 Annotation framework for fast domain expert labeling supported by an automated AI prelabeling

boxes, this can also be used for standard image classification tasks and pathologies that are hard to confine to a bounding box area.

Additionally, within the medical domain, we plan to implement a feature for automatically detecting gastroenterological tools. When the acting doctor detects a suspicious polyp or other, they often remove them during the examination. The tools will then be visible on screen and are an indicator of pathology. Hence, the tool detection can be used as an algorithm to detect relevant frames within the videos.The pre-detection algorithm itself is also not limited to our deep learning AI trained on polyps but can be exchanged easily for a AI more suited to the user's task.

The algorithm used for tracking objects across several frames is currently limited by the implemented standard object trackers above. These trackers are standard tools that often lose the object and have much room for improvement. While we provide an option for resetting the trackers, we intend to implement state-of-the-art video detection algorithms in the future to fully utilize this feature [45, 46].

Conclusion

In this paper, we introduce a framework for fast expert annotation, which reduces the working amount of the domain experts by a factor of 20 on our data while retaining very high annotation quality. We publish open-source software for annotation in the gastroenterological domain and beyond. This includes two views, one for expert annotation and one for non-expert annotation. We incorporate a semi-automated annotation process in the software, which reduces time spent on annotation and further enhances the annotation quality. Our results suggest that our tool enhances the medical especially endoscopic image and video annotation, tremendously. We not only reduce the time spend on annotation by the domain expert, but also the overall effort.

Methods

In this section, we explain our framework and software for fast semi-automated AI video annotation. The whole framework is illustrated in Fig. 3. The annotation process is split between at least two people. At first, an expert reviews the video and annotates a few video frames to verify the object's annotations for the non-expert. In a second step, a non-expert has visual confirmation of the given object and can annotate all following and preceding frames with AI assistance. To annotate individual frames, all frames of the video must be extracted. Relevant scenes can be selected by saving individual frames. This prevents the expert from reviewing the entire video every single time. After the expert has finished, relevant frames will be selected and passed on to an AI model. This information allows the AI model to detect and mark the desired object on all following and preceding frames with an annotation. Therefore, the non-expert can adjust and modify the AI predictions and export the results, which can then be used to train the AI model.

Input

To annotate individual video frames, the program must have access to all frames of the video. If annotated frames already exist, the program can recognize this; otherwise, it will extract all frames from the video and save them into a separate folder. Relevant frames can be annotated manually or inferred automatically. To mark the frames manually, frame numbers or timestamps are entered in the program. In the context of our polyp detection task, we created a script that detects when the recording freezes and marks these frames as relevant. A video freeze is caused by photos taken of suspicious tissue or polyps that are taken during the examination. The endoscope is stabilized mechanically if the examiner is pushing a button to take the photo. Therefore, these parts of the video are most relevant for the expert. This reduces the expert's workload since he does not have to review the entire video, but can quickly jump to the relevant parts of the video. The extraction is done by using the OpenCV framework.

Detect relevant frames

We denote all frames that assist the expert in finding critical parts of the video as *freeze frames*. Such frames can be detected automatically or entered manually by a frame number or timestamp. During a colonoscopic or gastroscopic examination, when the acting doctor detects a polyp (or similar), they freeze the video feed for a second and capture a photo of the polyp. Hence, for our task (annotation in gastroenterology), we automatically detect all positions in which a video shows the same frames for a short time, i.e., where the video is frozen for a few frames. Overall, within our implementation, we call such a position a "freeze frame". The detailed explanation for detecting those freeze frames is shown in Algorithm 1.

In order to discover those freezes automatically, we extract all frames from the video using OpenCV [47]. OpenCV is one of the most famous computer science libraries for image processing. Afterwards, we compare each frame to its next frame. This is done by computing the difference in pixel values of both frames, converting it into the HSV color space, and calculating an average norm by using the saturation and value dimension of the HSV color model. A low average norm means that both frames are almost identical; hence a freeze could have happened. We save a batch of ten comparisons for a higher certainty and take an average of the ten last comparisons (similar to a moving average). If the average value falls below a certain threshold, we define the current frame as the start of a freeze. The end of a freezing phase is determined if the average value exceeds another defined threshold. This algorithm has high robustness and consistency as it rarely misses a freeze or creates a false detection.

Algorithm 1 Freeze Detection

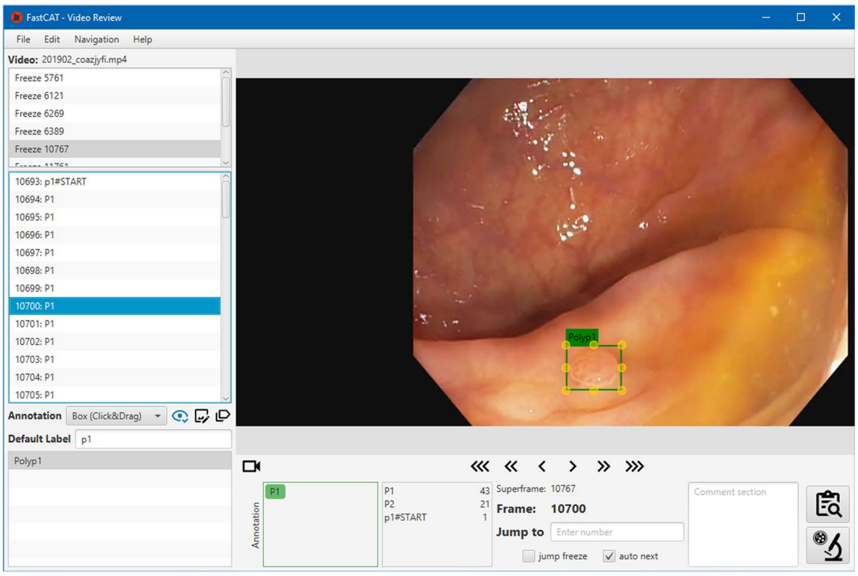
```
function FreezeDetection(video, windowSize)
    averages ← [ ], freezes ← [ ]                           ▷ List of averages (window) and freezes
    detected ← $False$                                      ▷ Flag if freeze detected
    while not end of video do
        frame, num ← nextFrame(video)
        diffFrame ← frame - prevFrame                       ▷ Calculate difference of each pixel
        diffFrame ← convertToHSV(diffFrame)                 ▷ Convert to HSV space
        h, s, v ← sumElements(diffFrame) / pixelCount       ▷ Average of each channel
        avg ← $\sqrt{s^2 + v^2}$                            ▷ Norm of s-/v-channel
        averages.add(avg)
        if len(averages) $\geq$ windowSize then
            $w$ ← sum(averages) / len(averages)
            if $w \leq 50$ and not detected then            ▷ Start of freeze phase
                freezes.add(num)
                detected ← $True$
            if $w > 75$ and detected then                   ▷ End of freeze phase
                detected ← $False$
            averages.removeAtIndex(0)
        prevFrame ← frame
    return freezes
```

Expert view

We refer to this part of the program *Video Review*, as the expert reviews the video to find polyps. For the expert to perform their task, they require the examination video, all individual video frames, and a set of relevant frame numbers, e.g., freeze frames. The video allows the expert to review the performed examination and get an overview of the presented situation to diagnose polyps correctly. All extracted video frames are necessary to be able to access and annotate individual frames. Lastly, a set of relevant frame numbers is given to the expert to jump to relevant video parts quickly. This led to a solution that provides the expert with two different viewpoints: (1) video player and (2) frame viewer. To enable fast and smooth transition between both viewpoints, it is possible to switch at any point in time from the current video time stamp t to the corresponding video frame f and vice versa. This is done by a simple calculation based on the frames per second (FPS) of the video and the current timestamp in milliseconds: $f = \frac{t[\text{ms}] \cdot \text{FPS}[1/\text{s}]}{1000}$.

It is possible to look at individual video frames within the frame viewer, assign classes to these frames, and annotate polyps within those frames. The class assignment is done through freeze frames, where each frame to which a class is assigned will be associated with a previously selected freeze frame. The second task, frame annotation, is independent of a class assignment and annotates the polyps within a frame with a bounding box that encloses the polyp. This primarily serves as an indication for non-experts to get visual information about the polyp that can be seen in the following/subsequent frames.

We use classes to mark frames if there is a polyp in the picture; we use these classes to mark relevant frames for the following annotation process by a non-expert. Two different approaches can be used to assign classes to frames. A range of frames is defined in the first approach by assigning start and end classes to two different frames. Consequentially, all frames in-between belong to the same class. The tool is also capable of assigning classes to each frame individually. The changes within video frames are small;



Video: 201902_coazjyfi.mp4

Freeze 5761
Freeze 6121
Freeze 6269
Freeze 6389
Freeze 10767

10693: p1#START
10694: P1
10695: P1
10696: P1
10697: P1
10698: P1
10699: P1
10700: P1
10701: P1
10702: P1
10703: P1
10704: P1
10705: P1

Annotation Box (Click&Drag)

Default Label p1

Polyp1

P1

P1	43
P2	21
p1#START	1

Superframe: 10767
Frame: 10700
Jump to Enter number
☐ jump freeze ☑ auto next

Fig. 4 Video Review UI. The figure shows the list of freeze frames, the corresponding child frames, and annotations within the image on the right side. In the bottom part of the view, the user can insert comments, open reports, delete classes, and see all individual classes. The diseased tissue is delineated via bounding boxes

therefore, many consecutive frames must be annotated with the same class. To make this process less time-consuming, the program allows the expert to go through a sequence of frames quickly and smoothly while classifying them by keeping a key pressed on the keyboard. However, mostly the assignment of start and end classes is faster and preferred.

Because all frames are mostly stored on an HDD/SSD, the loading latency is a performance bottleneck. We implemented a pre-loading queue that loads and stores the upcoming frames into the RAM to achieve fast loading times. This allows to display and assign frames with low latency. To prevent the queue from emptying rapidly, which causes high loading latency, we need to control the queue access times between two frames. Therefore, we use a capacity-dependent polynomial function to calculate a pausing time between frames: $\text{ms} = 50 \cdot (1 - \text{capacity})^{2.75}$. A full queue shortens the waiting time to 0 ms, while an empty queue leads to a 50-ms waiting time. This method combines fluent viewing and class assigning while providing enough time in the background to load new frames continuously.

Since the basic information about the presence of a polyp on an image is not sufficient for non-experts, and we want to ensure high-quality annotations, the expert has to annotate samples of all discovered polyps. This will provide visual information of the polyp to non-experts, allowing them to identify these polyps in all following and preceding frames correctly. Scenes in which polyps are difficult to identify due to perspective changes and other impairments should also be exemplary annotated by experts to provide as much information as possible to non-experts.

As we can see in Fig. 4 on the left side, the program lists all detected freeze frames. The list below shows all frames that belong to the selected freeze-frame and were annotated with specific classes, e.g., polyp type. Independent from the hierarchical structure above, we display all annotations that belong to the current frame in a list and on top of the image. In the lower part of the view, navigation controls skip a certain amount of frames or jump directly to a specific frame. The annotator can also leave a note to each frame if necessary or delete certain classes from the frame.

Semi-automated polyp prelabeling

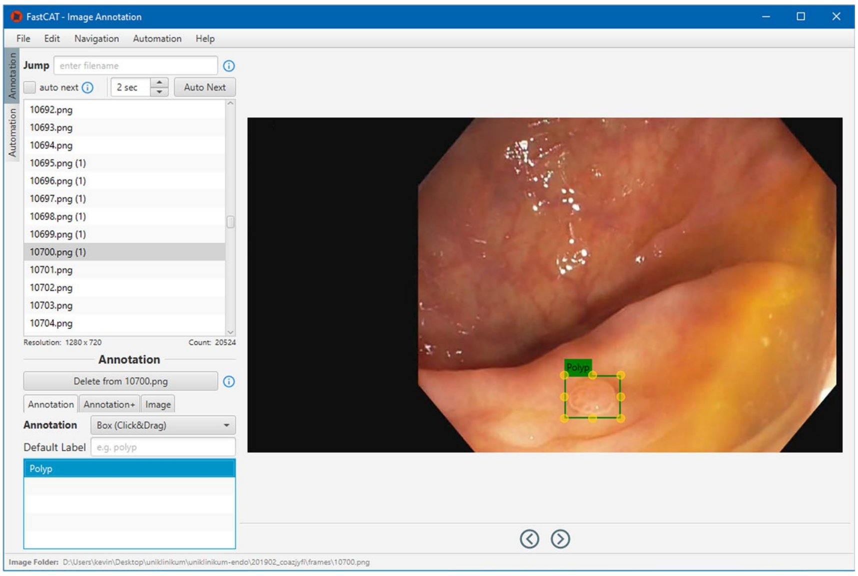
The prediction of polyps is made by an object detection model that was trained to detect polyps. The model we used is called EfficientDet [48]. EfficientDet is an object detection network that builds upon EfficientNet [49] and uses it as its backbone network. A feature extraction network is added on top of the backbone, which was named bidirectional feature pyramid network (BiFPN), and extracts the features of multiple layers. It is based on the idea of FPN and PANet [50] and combines multiple features of different sizes. This is called feature fusion and can be done by resizing or upsampling all feature resolutions to the same size and is combined by summing up. While previous methods did not consider the influence of a feature, BiFPN uses a weighted feature fusion that decides which features have the most influence. These features are then used for class and bounding box prediction. We adapted this network and trained it for polyp detection. The task of polyp detection is a combination of localizing and classifying an identified polyp. With this method, we aim for a fast AI-assisted annotation process for non-experts. Since every team has a different application, we distinguish between offline and online polyp prediction.

With an offline polyp prediction approach, we eliminate the need for high-end hardware for each user who uses AI assistance for fast annotation. The prediction is made by an external machine that is capable of running an AI model. With this approach, the extracted relevant frames are passed to this machine, generating a tool-specific JSON file that is then passed to the non-expert for further inspection.

As online polyp prediction, we define the performance of polyp detection locally on the machine of the annotator. Therefore, the machine on which our tool is executed must have the necessary hardware and software installed to run the detection AI. As there are different frameworks and deep learning networks, we need a unified interface to address all these different requirements. We decided to use Docker[4] for this task. Docker uses isolated environments called containers. These containers only carry the necessary libraries and frameworks to execute a program. By creating special containers for each model, we can run a prediction independent of our tool and its environment. Containers are built from templates called images, which can be published and shared between users. Therefore, it is possible to create a repository of different models and prediction objectives. Because a container shuts down after every prediction, it must reload the model for the next prediction. To counteract this, we run a web server inside the container and communicate to the model via HTTP. This ensures that a model does not have to reload after every prediction and provides a universal and model-independent communication interface. With this setup, the user can trigger a single prediction or run a series of predictions in the background.

---

[4] https://docker.com.



2 sec

10692.png
10693.png
10694.png
10695.png (1)
10696.png (1)
10697.png (1)
10698.png (1)
10699.png (1)
10700.png (1)
10701.png
10702.png
10703.png
10704.png

Resolution: 1280 x 720
Count: 20524

Delete from 10700.png

Annotation: Box (Click&Drag)

Polyp

Image Folder: D:\Users\kevin\Desktop\uniklinikum\uniklinikum-endo\201902_coazjyfi\frames\10700.png

Fig. 5 Image annotation UI. The figure shows a list of all available frames on the left with labeling functionality for a specific annotation and the whole image. The image to be annotated is displayed on the right. The diseased tissue is delineated via bounding boxes

As we have already stated, we use HTTP for our communication. This gives room for a hybrid solution, allowing predictions on an external server while retaining the user's control. This combines the advantages of the external and local approaches, where the user is not required to have expensive hardware, nor is it necessary to have a separate, time-consuming prediction step. The docker container is now running during the annotation process and AI is running in the container while using the program. Therefore, the diseased tissue delineating bounding box is directly drawn as an annotation on the image. This annotation can then be corrected or redrawn in the process.

Non-expert annotation

With the help of AI, it is possible to annotate a large number of frames quickly and easily. The AI is predicting the annotations directly to the image. However, this method does not ensure the correctness of the predicted annotations. For this reason, these annotations must be checked and modified if necessary. Non-experts can check these predictions or create new annotations with the help of verified example annotations from the expert and the indication in which frame a polyp is visible. Besides, the AI-assisted support of our tool provides annotation duplication across several frames and object tracking functionality which speeds up the annotation process. Figure 5 illustrates the UI of the non-experts view.

As mentioned in section *Semi-automated polyp prelabeling* our tool supports the integration of AI detection. It can trigger a single prediction or make predictions on the following frames in the background. This enables the user to immediately annotate the remaining frames without waiting for the external prediction process to finish.

Another helpful feature is the duplication of annotations. Sometimes, only subtle movements occur in polyp examination videos, causing a series of frames to only show minuscule changes. This feature allows the non-expert to use the bounding boxes of the previous frame and only make minor adjustments while navigating through the frames. Re-positioning an existing bounding box requires less time than creating an entirely new box with a click and drag motion.

Our last feature uses object tracking to track polyps throughout consecutive frames. This avoids the manual creation of bounding boxes for each video frame, especially in sequences where an object's visual and spatial transition between two frames is non-disruptive. For this task, we used trackers available in the OpenCV framework. Within the intestine, special conditions are usually present. First, the nature of colonoscopies leads to unsteady camera movement. Second, the color of polyps is often similar to the surrounding intestinal wall, which can make them hard to recognize. This can compromise the performance of the tracker and deteriorate polyp tracking. Given the fact that the annotation process requires a user to operate the tool and, therefore, the tracker does not need to track polyps fully automatically, we added two options to reset the tracker. This is described in more detail in the next section.

Object trackers

As described in section *Non-expert annotation* our tool has object tracking functionality. It assists in tracking an object across multiple frames. For our tool, we implement six of the available trackers in the OpenCV framework [47]. In the following, we give a short description of the available trackers:

- *Boosting.* It is using an online version of AdaBoost to train the classifier. Therefore, the tracking is viewed as a binary classification problem, and negative samples of the same size are extracted from the surrounding background. It can update features of the classifier during tracking to adjust to appearance changes [51].
- *MIL.* Multiple Instance Learning uses a similar approach as Boosting and extracts positive samples from the immediate neighborhood of the object. The set of samples is put into a bag. A bag is positive when it contains at least one positive example, and the learning algorithm has to the inference which is the correct sample within a positive bag [52].
- *KCF.* Kernelized Correlation Filter uses the same basic idea as MIL, but instead of sampling a handful of random samples, it trains a classifier with all samples. It exploits the mathematical properties of circulant matrices to make tracking faster and better [53].
- *CSRT* CSRT uses discriminative correlation filters (CDF) with channel and spatial reliability concepts. The correlation filter finds similarities between the two frames. The spatial reliability map restricts the filter to suitable parts of the image. Scores estimate the channel reliability to weight features [54]. In addition, it is worth mentioning that rapid movements are not handled well by trackers that use CDF [55].
- *Median flow.* Median flow tracks points of the object forward and backward in time. Thereby, two trajectories are measured, and an error between both trajectories is

estimated. By filtering out high error points, the algorithm tracks the object with all remaining points [56], It is best applicable for smooth and predictable movements [57].

- *MOSSE.* Minimum Output Sum of Squared Error is an adaptive correlation filter robust to light variation, scale, post, and deformations. It applies a correlation filter to detect the object in new frames. It works only with grayscale images, and colored images will be converted internally [58].
- *TLD.* TLD decomposes a long-term tracking task into tracking, learning, and detection. The tracker is responsible for tracking the object across the frames. The detector finds the object within a frame and corrects the tracker if necessary, and the learning part of the algorithm estimates the error of the detector and adjusts it accordingly [59].

An object tracker is designed to follow an object over a sequence of frames by locating its position in every frame. Each tracker uses different strategies and methods to perform its task. Therefore, trackers have to be switched and tested when tracking different pathologies. It can collect information such as orientation, area, or the shape of an object. However, also many potential distractions can occur during tracking that can make it hard to track the object. Distraction causes are, e.g., noisy images, unpredictable motion, changes in illumination, or complex shapes. As a result, the performance of different trackers can vary between different domains and datasets. For this reason, our tool allows the user to choose the best tracker for their task and dataset. Because trackers are primarily designed to track objects across many frames automatically, the tracker may generate less accurate bounding boxes over time or entirely lose track of the object. Since the tracking conditions for polyp detection are complex and our tool uses a semi-automated solution, we implemented two additional options for the annotation task.

By default, the tracker is initialized by placing a bounding box around an object that should be tracked. Consequently, the tracker will find the object on one consecutive frame and place a bounding box around it. We found that the tracker loses track of the initialized polyp with a high number of consecutive frames. Therefore, we implemented options to reinitialize the tracker automatically. The first option reinitializes the tracker after every frame, giving the tracker the latest visual information of the polyp. The second option only initializes the tracker if the user changed the bounding box size. Both options ensure that the tracker has the latest visual information of the polyp since the user corrects misaligned bounding boxes.

Output and conversion

We use JSON as our standard data format. The JSON prepared by the expert stores detected freeze frames with all corresponding frames that contain at least one class. Additionally, annotated frames are stored in the same file but independently from the class assignments. The resulting JSON from the expert annotation process serves as an intermediate output for further annotations. All annotations that are done automatically are annotated so they can be distinguished from the annotations done manually.

The non-expert produces the final output with all video annotations. This file contains a list of all frames with at least one annotation. The tool produces a JSON with a structure designated to fit our needs. However, since different models require different data formats, we created a *python* script that converts our format into a delimiter-separated values (DSV) file format. Via a configuration file, the user can adjust the DSV file to its need, e.g., convert it into YOLO format. It is also possible to convert the DSV file back to our format. This enables seamless integration of different formats. In the future, further predefined formats can be added.

Abbreviations

COCO: Common Objects in Context; CSV: Comma-separated values; JSON: JavaScript Object Notation; YOLO: You Only Look Once; XML: Extensible Markup Language; CVAT: Computer Vision Annotation Tool; TFRecord: Tensor Flow Record; FER: Frame extraction rate; HTML: HyperText Markup Language; DICOM: Digital Imaging and Communications in Medicine; AI: Artificial intelligence; GIANA: Gastrointestinal image analysis; SPF: Seconds per frame; Tgca: Time gained compared to annotation; UI: User interface; HSV: Hue, saturation, lightness; RAM: Random-access memory; HTTP: Hypertext Transfer Protocol; DSV: Delimiter-separated values.

Acknowledgements

We kindly thank the University Hospital of Würzburg, the Interdisziplinäres Zentrum für Klinische Forschung (IZKF) and the Forum Gesundheitsstandort Baden-Württemberg for supporting the research. This publication was supported by the Open Access Publication Fund of the University of Würzburg.

Author contributions

AK implemented and coordinated the study, drafted the manuscript, and interpreted the data. KM and AK designed and implemented the software. AH1 and KM contributed to complete the manuscript. DF helped with the creation of the data. JT helped with the data preprocessing. FP, AH2 and WZ provided funding and reviewed the manuscript. All authors read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. This research is supported using public funding from the state government of Baden-Württemberg, Germany (Funding cluster Forum Gesundheitsstandort Baden-Württemberg) to research and develop artificial intelligence applications for polyp detection in screening colonoscopy and by Interdisziplinäres Zentrum für Klinische Forschung (IZKF) from the University of Würzburg.

Availability of data and materials

The first dataset used for the analysis of this article is available in the GIANA challenge repository (https://endovissub2017-giana.grand-challenge.org/). The second dataset used during the analysis is available from the corresponding author on reasonable request.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Artificial Intelligence and Knowledge Systems, Sanderring 2, 97070 Würzburg, Germany. [2]Interventional and Experimental Endoscopy (InExEn), Department of Internal Medicine II, University Hospital Würzburg, Oberdürrbacher Straße 6, 97080 Würzburg, Germany. [3]Department of Internal Medicine and Gastroenterology, Katharinenhospital, Kriegsbergstrasse 60, 70174 Stuttgart, Germany.

Received: 4 August 2021 Accepted: 25 April 2022
Published online: 25 May 2022

References

1. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform. 2018;19(6):1236–46.
2. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, Van Der Laak JA, Van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88.
3. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine learning for medical imaging. Radiographics. 2017;37(2):505–15.
4. Gunčar G, Kukar M, Notar M, Brvar M, Černelč P, Notar M, Notar M. An application of machine learning to haematological diagnosis. Sci Rep. 2018;8(1):1–12.
5. Halama N. Machine learning for tissue diagnostics in oncology: brave new world. Br J Cancer. 2019;121(6):431–3. https://doi.org/10.1038/s41416-019-0535-1.
6. Kim K-J, Tagkopoulos I. Application of machine learning in rheumatic disease research. Korean J Intern Med. 2019;34(4):708.

7. Zerka F, Barakat S, Walsh S, Bogowicz M, Leijenaar RT, Jochems A, Miraglio B, Townend D, Lambin P. Systematic review of privacy-preserving distributed machine learning from federated databases in health care. JCO clinical cancer informatics. 2020;4:184–200.
8. Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E. Deep learning applications and challenges in big data analytics. J Big Data. 2015;2(1):1–21.
9. Chang JC, Amershi S, Kamar E. Revolt: collaborative crowdsourcing for labeling machine learning datasets. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems; 2017. p. 2334–2346.
10. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, Li B, Madabhushi A, Shah P, Spitzer M, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463–77.
11. Webb S. Deep learning for biology. Nature. 2018;554(7693):555–8.
12. Hoerter N, Gross SA, Liang PS. Artificial intelligence and polyp detection. Curr Treat Options Gastroenterol. 2020;18(1):120–36.
13. Krenzer A, Hekalo A, Puppe F. Endoscopic detection and segmentation of gastroenterological diseases with deep convolutional neural networks. In: EndoCV@ ISBI; 2020. p. 58–63.
14. Bhagat PK, Choudhary P. Image annotation: then and now. Image Vis Comput. 2018. https://doi.org/10.1016/j.imavis.2018.09.017.
15. Stork DG. Character and document research in the open mind initiative. In: Proceedings of the fifth international conference on document analysis and recognition. ICDAR '99 (Cat. No.PR00318); 1999. p. 1–12. https://doi.org/10.1109/ICDAR.1999.791712.
16. Ahn Lv, Dabbish L. Labeling images with a computer game. 2004;319–26. https://doi.org/10.1145/985692.985733.
17. Russell BC, Torralba A, Murphy KP, Freeman WT. Labelme: a database and web-based tool for image annotation. Int J Comput Vis. 2008. https://doi.org/10.1007/s11263-007-0090-8.
18. Sekachev B, Manovich N, Zhavoronkov A. Computer vision annotation tool: a universal approach to data annotation. https://software.intel.com/content/www/us/en/develop/articles/computer-vision-annotation-tool-a-universal-approach-to-data-annotation.html Accessed 01 Jun 2021
19. Tzutalin: LabelImg. https://github.com/tzutalin/labelImg Accessed 01 Jun 2021
20. Wada K. labelme: image polygonal annotation with Python; 2016. https://github.com/wkentaro/labelme
21. Microsoft: Visual Object Tagging Tool. https://github.com/microsoft/VoTT Accessed 01 Jul 2021
22. Dutta A, Zisserman A. The VIA annotation software for images, audio and video. In: Proceedings of the 27th ACM international conference on multimedia. MM '19. ACM, New York, NY, USA; 2019. https://doi.org/10.1145/3343031.3350535.
23. Dutta A, Gupta A, Zissermann A. VGG image annotator (VIA); 2016. http://www.robots.ox.ac.uk/~vgg/software/via/ Accessed 09 Jun 2021
24. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380(14):1347–58.
25. Sidey-Gibbons JA, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. BMC Med Res Methodol. 2019;19(1):1–18.
26. Wang F, Casalino LP, Khullar D. Deep learning in medicine-promise, progress, and challenges. JAMA Intern Med. 2019;179(3):293–4.
27. Yushkevich PA, Gao Y, Gerig G. Itk-snap: An interactive tool for semi-automatic segmentation of multi-modality biomedical images. In: 2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC); 2016. p. 3342–3345. https://doi.org/10.1109/EMBC.2016.7591443.
28. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin J-C, Pujol S, Bauer C, Jennings D, Fennessy F, Sonka M, et al. 3d slicer as an image computing platform for the quantitative imaging network. Magn Reson Imaging. 2012;30(9):1323–41.
29. Rubin-Lab: ePAD: web-based platform for quantitative imaging in the clinical workflow. [Online; Stand 13.05.2021]; 2014. https://epad.stanford.edu/
30. Gupta G, Gupta A. TrainingData.io. [Online; Stand 13.05.2021]; 2019. https://docs.trainingdata.io/
31. Gupta G. TrainingData.io: AI assisted image & video training data labeling scale. [Online; Stand 13.05.2021]; 2019. https://github.com/trainingdata/AIAssistedImageVideoLabelling/
32. Philbrick K, Weston A, Akkus Z, Kline T, Korfiatis P, Sakinis T, Kostandy P, Boonrod A, Zeinoddini A, Takahashi N, Erickson B. Ril-contour: a medical imaging dataset annotation tool for and with deep learning. J Digit Imaging. 2019. https://doi.org/10.1007/s10278-019-00232-0.
33. Leibetseder A, Münzer B, Schoeffmann K, Keckstein J. Endometriosis annotation in endoscopic videos. In: 2017 IEEE international symposium on multimedia (ISM); 2017. p. 364–365. https://doi.org/10.1109/ISM.2017.69
34. Guo YB, Matuszewski BJ. Giana polyp segmentation with fully convolutional dilation neural networks. In: VISIGRAPP; 2019. p. 632–641.
35. Mahony NO, Campbell S, Carvalho A, Harapanahalli S, Velasco-Hernandez G, Krpalkova L, Riordan D, Walsh J. Deep learning vs traditional computer vision. https://doi.org/10.1007/978-3-030-17795-9. arXiv:1910.13796.
36. Silva J, Histace A, Romain O, Dray X, Granado B. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. Int J Comput Assist Radiol Surg. 2014;9:283–93. https://doi.org/10.1007/s11548-013-0926-3.
37. Qadir HA, Balasingham I, Solhusvik J, Bergsland J, Aabakken L, Shin Y. Improving automatic polyp detection using cnn by exploiting temporal dependency in colonoscopy video. IEEE J Biomed Health Inform. 2019;24(1):180–93.
38. Hasan MM, Islam N, Rahman MM. Gastrointestinal polyp detection through a fusion of contourlet transform and neural features. J King Saud Univ Comput Inf Sci; 2020.
39. Sun X, Wang D, Zhang C, Zhang P, Xiong Z, Cao Y, Liu B, Liu X, Chen S. Colorectal polyp detection in real-world scenario: Design and experiment study. In: 2020 IEEE 32nd international conference on tools with artificial intelligence (ICTAI), IEEE; 2020. p. 706–713.
40. Lambert RF. Endoscopic classification review group. update on the paris classification of superficial neoplastic lesions in the digestive tract. Endoscopy. 2005;37(6):570–8.

41. Zhang X, Chen F, Yu T, An J, Huang Z, Liu J, Hu W, Wang L, Duan H, Si J. Real-time gastric polyp detection using convolutional neural networks. PLoS ONE. 2019;14(3):0214133.
42. Jha D, Ali S, Tomar NK, Johansen HD, Johansen D, Rittscher J, Riegler MA, Halvorsen P. Real-time polyp detection, localization and segmentation in colonoscopy using deep learning. IEEE Access. 2021;9:40496–510.
43. Bernal J, Tajkbaksh N, Sánchez FJ, Matuszewski BJ, Chen H, Yu L, Angermann Q, Romain O, Rustad B, Balasingham I, et al. Comparative validation of polyp detection methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge. IEEE Trans Med Imaging. 2017;36(6):1231–49.
44. Shackleton V. Boredom and repetitive work: a review. Personnel Review; 1981.
45. Pal SK, Pramanik A, Maiti J, Mitra P. Deep learning in multi-object detection and tracking: state of the art. Appl Intell. 2021;1–30.
46. Li Y, Zhang X, Li H, Zhou Q, Cao X, Xiao Z. Object detection and tracking under complex environment using deep learning-based lpm. IET Comput Vision. 2019;13(2):157–64.
47. Bradski G. The OpenCV Library. Dr Dobb's Journal of Software Tools. 2000.
48. Tan M, Pang R, Le QV. Efficientdet: Scalable and efficient object detection. arXiv:1911.09070v4. Accessed 2020-07-16
49. Tan M, Le QV. Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv:1905.11946v3. Accessed 16 Jul 2020
50. Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 2117–2125.
51. Grabner H, Grabner M, Bischof H. Real-time tracking via on-line boosting. In: BMVC; 2006.
52. Babenko B, Yang M, Belongie S. Visual tracking with online multiple instance learning. In: 2009 IEEE conference on computer vision and pattern recognition; 2009. p. 983–990.
53. Henriques J, Caseiro R, Martins P, Batista J. Exploiting the circulant structure of tracking-by-detection with kernels, vol 7575. 2012. p. 702–15.
54. Lukezic A, Vojir T, Cehovin Zajc L, Matas J, Kristan M. Discriminative correlation filter with channel and spatial reliability. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); 2017.
55. Gong F, Yue H, Yuan X, Gong W, Song T. Discriminative correlation filter for long-time tracking. Comput J. 2019;63(3):460–8. https://doi.org/10.1093/comjnl/bxz049.
56. Kalal Z, Mikolajczyk K, Matas J. Forward-backward error: Automatic detection of tracking failures. In: 2010 20th international conference on pattern recognition; 2010. p. 2756–2759. https://doi.org/10.1109/ICPR.2010.675
57. OpenCV: MedianFlow tracker class reference. https://docs.opencv.org/4.3.0/d7/d86/classcv_1_1TrackerMedianFlow.html#details Accessed 12 May 2021
58. Draper BA, Bolme DS, Beveridge J, Lui Y. Visual object tracking using adaptive correlation filters. In: 2010 IEEE conference on computer vision and pattern recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA; 2010. p. 2544–2550. https://doi.org/10.1109/CVPR.2010.5539960.
59. Kalal Z, Mikolajczyk K, Matas J. Tracking-learning-detection. IEEE Trans Pattern Anal Mach Intell. 2012;34(7):1409–22. https://doi.org/10.1109/TPAMI.2011.239.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC

Public Health and Informatics
J. Mantas et al. (Eds.)
© 2021 European Federation for Medical Informatics (EFMI) and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/SHTI210206

Semi-Automated Machine Learning Video Annotation for Gastroenterologists

Adrian KRENZER[a,1], Kevin MAKOWSKI[a], Amar HEKALO[a] and Frank PUPPE[a]
[a] *Julius-Maximilian University of Würzburg, Germany*

Abstract. A semi-automatic tool for fast and accurate annotation of endoscopic videos utilizing trained object detection models is presented. A novel workflow is implemented and the preliminary results suggest that the annotation process is nearly twice as fast with our novel tool compared to the current state of the art.

Keywords. machine learning, deep learning, video annotation tool, endoscopy

1. Introduction

Recently machine learning started to play an important role in the domain of medical analysis, classification and disease prevention [1]. Most supervised machine learning algorithms need lots of high-quality data. The annotation process to acquire this data is very costly and labor-intensive, especially if domain experts are involved. Therefore, a tool for an efficient annotation process is presented, which reuses routinely made pathologic snapshots from an endoscopy. It requires an expert just to mark the beginning and end of the pathology in the video based on the snapshot, automatically detects bounding boxes of the pathologies (in our case polyps) within all the frames of the marked video sequences and offers a non-expert annotator comfortable fine-tuning of the bounding boxes if necessary. Figure 1 summarizes the workflow of our tool.

2. Methods

First, the user inputs a video file to our annotation tool. All the frames of the video are extracted and stored as images in the referencing folder. Then, special frames are selected by an automated domain-specific process these frames are then marked by an expert.

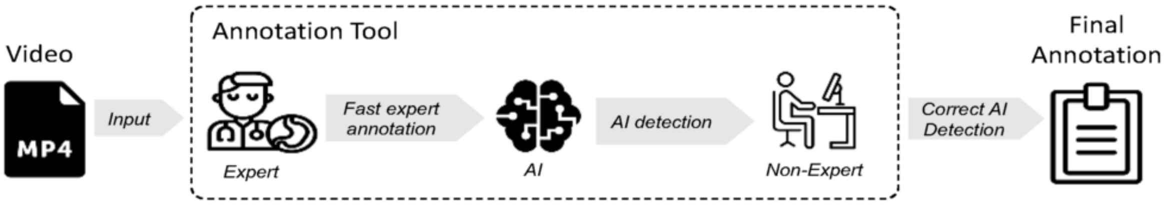


Figure 1. Workflow of our proposed annotation process.

---

[1] Corresponding Author, Adrian Krenzer, Julius-Maximilian University of Würzburg, Department of Artificial Intelligence and Knowledge Systems, Sanderring 2, 97070 Würzburg, Germany; E-mail: adrian.krenzer@uni-wuerzburg.de.

Table 1. Comparison of the well-known CVAT annotation tool to our new annotation tool for faster annotation for gastroenterologists. Videos 1 and 2 are open source and annotated. Video 3 is from a German clinic. For video 3 no quality evaluation is performed since there are no ground truth annotations available. The quality metric is the mean average precision (mAP70) when the drawn box matches the ground truth box to 70 %.

	Quality (%)		Speed (FPS)		Speed (min)		Videoinfomation		
	CVAT	**Ours**	CVAT	**Ours**	CVAT	**Ours**	Frames	Polyps	Framesize
Video 1	**99.16**	99.04	9.36	**3.64**	57.88	**22.5**	371	1	384x288
Video 2	99.44	**99.58**	5.31	**3.81**	39.74	**28.5**	449	1	384x288
Video 3	-	-	2.72	**1.35**	57.98	**28.78**	1279	1	898x720
Mean	99.3	**99.31**	5.79	**2.93**	51.86	**26.59**	700	1	555x432

The model is trained to predict polyp bounding boxes on the marked frames. This model can always be retrained with the newly annotated data. As the detection gets better, the time for annotation should decrease. Therefore, the annotation time will be reduced with annotation progress. For the detection of the polyps state of the art object detector YOLOv4 [2] is used. The detector is trained on openly available datasets for polyp detection. In the final step, a non-expert corrects the model's predictions. He gets all the model's predictions and one or more annotated boxes from the expert. Those annotations are still of high quality as the expert annotations assure the presence of a pathology and the non-experts only adjust the predicted bounding boxes.

3. Preliminary Evaluation

For our preliminary evaluation, two test subjects are instructed to use our annotation tool and the state of the art annotation tool CVAT [3]. The test subjects in this experiment are undergraduates from the field of computer science therefore just the non-expert part of our tool is evaluated. Both students are instructed to annotate the polyp frames as fast and as accurately as they can. The results are shown in Table 1. The quality evaluation results show that almost similar annotation results to those of gastroenterology experts are achieved. For speed, our tool outperforms the CVAT tool in any video. In two videos our tool is more than twice as fast as the CVAT tool. A total of three videos were annotated. We plan a full evaluation to further investigate the speed of the expert annotators.

4. Conclusion

All in all, a novel tool for machine learning video annotation in endoscopic recordings is presented. The method's annotation speed exceeds the classic computer science tool CVAT [3] while maintaining high-quality results.

References

[1] Hamet P, Tremblay J. Artificial intelligence in medicine. Metabolism. 2017 Apr 1;69:S36-40.

[2] Bochkovskiy A, Wang CY, Liao HY. YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv preprint arXiv:2004.10934. 2020 Apr 23.

[3] Sekachev B et al.: Opencv/cvat: v1.1.0. https://github.com/opencv/cvat 2020 May 25.

CEUR-WS.org/Vol-2595/endoCV2020_paper_id_6.pdf

ENDOSCOPIC DETECTION AND SEGMENTATION OF GASTROENTEROLOGICAL DISEASES WITH DEEP CONVOLUTIONAL NEURAL NETWORKS

Adrian Krenzer, Amar Hekalo, Frank Puppe

Department of Artificial Intelligence and Knowledge Systems, University of Würzburg, Germany

ABSTRACT

Previous endoscopic computer vision research focused mostly on the detection of a singular disease like, e.g. polyps. The endoscopic disease detection challenge (EDD2020) extends this classification task by providing data for different diseases in various organs. The EDD2020 includes two sub-tasks[1]: (1) Multi-class disease detection: localization of bounding boxes and class labels for the five disease classes: Polyp, Barret's Esophagus (BE), suspicious, High Grade Dysplasia (HGD) and cancer; (2) Region segmentation: boundary delineation of detected diseases. In this paper, we describe our approach by leveraging deep convolutional neural networks (CNNs). We highlight the comparison of two general state-of-the-art object detection approaches. The first one is Single Shot Detection (SSD), and the second one are two-step region proposal based CNNs. We, therefore, compare two different models: YOLOv3 (SSD) and Faster R-CNN with ResNet-101 backbone. For the second task, we leverage the state-of-the-art Cascade Mask R-CNN with various backbones and compare the results. In order to minimize generalization error, we apply data augmentation; finally, we use knowledge from the endoscopic domain to further refine our models during post-processing and compare the resulting performances.

1. INTRODUCTION

Endoscopic vision is a procedure which covers many different areas and organs of the human body, such as the bladder, the stomach or the colon, allowing gastroenterologists to potentially discover a wide array of diseases and abscesses, like polyps, cancer and Barrett's esophagus. Naturally, in order to assure detection of all diseases and to improve the workflow, application of real-time detection using Deep Learning is becoming more prevalent. There have been previous publications with good results on real-time detection of endoscopic polyps using Single Shot Detector [1] based CNNs [2] as well as an anchor free approach called AFP-Net [3]. Existing work usually focuses on one disease class, like polyp or cancer detection, mostly due to lack of annotated data. The Endoscopic Disease Detection Challenge 2020 [4] partially solves this issue by providing endoscopic images of three different organs, namely colon, esophagus and stomach, with five disease classes. Additionally they provide corresponding bounding boxes for object detection as well as polygonal masks for image segmentation. In this paper we apply and train state-of-the-art Deep Learning models for both tasks using various architectures and comparing their performance.

2. DATASETS AND DATA ANALYSIS

In order to choose and prepare the right deep CNN for the task, we start by analyzing the given training data in detail. The EDD2020 challenge [4] provides a training data set for multi-class disease detection, which contains 386 endoscopic images labeled with 684 bounding boxes and 502 segmentation masks. While analyzing the data, we recognize class imbalance. Therefore we counted the occurrences for each class throughout the dataset based on the bounding boxes. The dataset has more than 200 images with polyps and BE but less than 100 samples for the three remaining classes respectively. So, it might be challenging to learn the correct assessment of the classes HGD, suspicious and cancer. This unbalanced sample distribution is one difficulty of the dataset and is therefore considered while choosing our model and it's hyperparameters. The second difficulty we recognize is the variation in box sizes. We therefore calculated the area of all the boxes. Most of the boxes have nearly the same mean area while the variation of the areas differs enormously, especially for the polyp class, where the standard deviation is significantly larger than within other classes.

Finally, for the segmentation task, for every image there are given masks specifying which regions are of interest which is done separately for each class. While most of the images belong to a unique class, some of them have several masks with overlapping regions, which is especially apparent for the "suspicious" class. The latter is often only part of a region of an already existing class. Hence this is a multi-class multi-label segmentation task with independent classes. We randomly split the dataset into 90% training and 10% validation set, where the best model is chosen by minimum

---

[1] https://edd2020.grand-challenge.org

Copyright ©2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

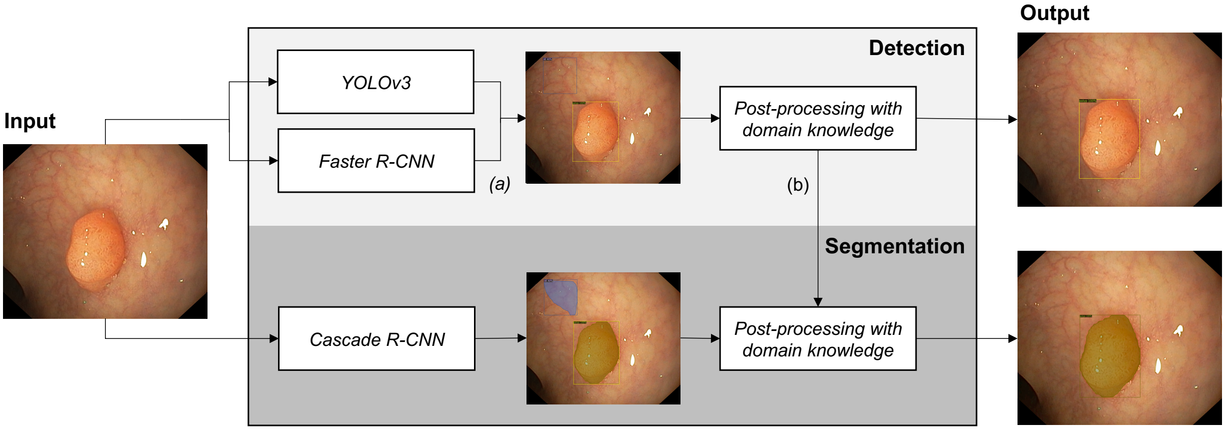

Input
Output
Detection
YOLOv3
Faster R-CNN
(a)
Post-processing with domain knowledge
(b)
Segmentation
Cascade R-CNN
Post-processing with domain knowledge

Fig. 1: This figure illustrates our final pipeline for the detection and segmentation task. At step (a) the predictions for polyps and HGD of the YOLOv3 algorithm and the predictions of BE, suspicious, and cancer of the Faster R-CNN are applied for the final result. At step (b) the box output of the detection architecture is utilized to filter the segmentation masks.

validation loss during training.

Additional data: In order to improve generalization, we extend the training dataset by including images from openly accessible databases. We include two datasets from a previous endoscopic vision challenge [5], namely the ETIS-Larib Polyp database [6], which consists of 196 polyp images, and the CVC-ClinicDB [7], which consists of 612 polyp images, as well as the dataset from the Gastrointestinal Image Analysis (GIANA) challenge [8], with 412 polyp images. All three datasets have corresponding segmentation masks. We add corresponding bounding boxes using the segmented masks ourselves. In addition we include the Kvasir-SEG dataset [9], which consists of 1000 polyp images with both segmentation masks and bounding boxes. Finally, we extract images annotated with esophagitis from the Kvasir2 dataset [10]. Esophagitis and Barret's esophagus occur at the same position in the esophagus, and some symptoms of esophagitis are very similar to Barret's esophagus symptoms. Therefore we add images with esophagitis symptoms which looked close to Barret's esophagus and test if those improve our results. We receive a light improvement in BE results and therefore include 103 additional images for a total of 2323 additional training images. Nevertheless, Barret's esophagus and esophagitis are different diseases and have to be distinguished in further research if more classes are included in the classification task.

3. METHODS

In this section, we illustrate our approaches for the two subtasks. All our models are trained on a Tesla P100 Nvidia GPU. After exploring the data, we decided to choose CNNs for the challenge as they have proven to be very stable in classic multi-class detection tasks like the COCO challenge [11]. In the domain of object detection, we consider two main concepts that have proven successful in multi-class object detection. First, a two-step method of region proposals and subsequent classification of the proposed regions like Faster R-CNN. Second single-shot detection (SSD), which is mostly applicable in real-time. We compare the results of the SSD model and Faster R-CNN. To improve our results further, we combine those two algorithms in our final architecture. For the second task, since both bounding boxes and segmentation masks are available, we choose the Cascade Mask R-CNN. Incorporating both types of annotations achieves the best results. For both of these tasks we add a post-processing with gastroenterological knowledge. Figure 1 depicts our final architecture for the detection and segmentation task. For training the Faster R-CNN we leverage the open source Detectron2 framework [12].

By including additional 2220 polyp images, we significantly increase the class imbalance of the training data. Class balance is crucial for training and inference of neural networks. To tackle this problem, we use class weights in the algorithms. Therefore the loss of an underrepresented class multiplies by a weight that balances the outcome of the total loss function. By adding those weights, we observe an enhancement in polyp detection while not losing the detection score in the other classes [13].

3.1. Task 1 multi-class bounding box detection:

As mentioned above, we want to compare two common object detection approaches, namely SSD and what we call a classic region proposal approach. Compared to classical approaches, SSD enables real-time detection. In practice, real-time detection is critical. Often, the gastroenterological diseases receive treatment directly (e.g., ablation of a polyp). Therefore a low inference time has to be considered to apply the models in real practice. On the contrary, larger architectures may perform better in tasks suited for procedures like detecting the stadium of the disease, which mostly has no real-time restrictions. Nevertheless, a larger architecture may perform well on our challenge task, too. Therefore, we leverage one model from each of these sub positions. The model for SSD we utilize is called the YOLOv3 algorithm [14], which is the third version of the well-known YOLO architecture [15] and has added residual blocks that allow training deeper networks while preventing the vanishing gradient problem. We use the YOLOv3 algorithm with initial weights pre-trained on the COCO dataset [11]. In the next step, we unfreeze the last two layers of the network and train them utilizing the adam optimizer [16]. We train for 50 epochs. In addition, we unfreeze the whole network and train until it stops through early stopping, resulting in an additional 33 epochs.

As a classic larger architecture, we use a Faster R-CNN [17] with a 104 depth Retinanet backbone. We use a batch size of 2 because of the computational expense of this large network. We initialize the network with weights pre-trained on the COCO dataset. We choose a learning rate of 0.00025 for the training.

Post-processing: The YOLOv3 architecture is more successful in classifying polyps and HGD whereas classic architecture is better in detecting BE, suspicious and cancer. We therefore assemble both networks to improve our detection results. Hence, the YOLOv3 predicts HGD and polyps while the Faster R-CNN algorithm predicts BE, suspicious and cancer. Both algorithms can predict all labels, but we only use the predictions of the specified classes from each algorithm respectively. To further improve our results we use gastroenterological knowledge and knowledge of the data set structure. As the probability is low that BE and polyp are predicted in the same image we implement a simple rule: If both polyps and BE are detected, we only produce boxes for the class with higher probability, i.e., if the probability for polyps is higher than for BE, no bounding boxes are predicted for BE.

3.2. Task 2 region segmentation:

For the image segmentation task, we train two similar architectures with various backbones, namely Mask R-CNN [18] and its successor, Cascade Mask R-CNN [19]. Both architectures are primarily two-stage object detection models based on Faster R-CNN, i.e. a region proposal network first proposes candidate bounding boxes (Regions of Interest, RoI) before the final prediction. Here, they add another branch used to predict segmentation masks, where the proposed RoIs are used to enhance the segmentation mask predictions in contrast to using fully convolutional networks only. Cascade Mask R-CNN is an extended framework using a cascade-like structure and is essentially an ensemble of several Mask R-CNNs with weight sharing on the backbones.

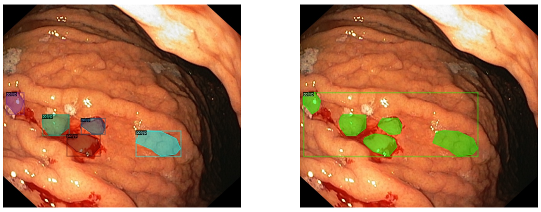


Fig. 2: In order to train Mask and Cascade Mask R-CNN for semantic segmentation, some bounding boxes had to be adjusted. We transform the boxes from including several instances (left) to be only one instance (right).

We choose these types of models for two reasons: First, since we have both bounding boxes and segmentation masks available as training data, we can utilize the Mask R-CNN approach, where RoI influences the segmentation, to the fullest. Second, since these networks are set to perform instance segmentation, each class is predicted independently from each other, which is a prefect fit for our multi-class multi-label problem. As this is a semantic task, we treat this as an instance segmentation with only one instance per occurrence per class. As such, we had to adjust some of the ground truth bounding boxes in our data, as shown in Fig. 2.

For Mask R-CNN we use the ResNeXt-101-32x8d [20] and for Cascade Mask R-CNN the ResNeXt-151-32x8d [20] models as backbones, both of which are CNN classifyers pre-trained on the ImageNet-1k dataset [21]. Additionally, both full architectures are pre-trained on the COCO dataset [11], hence we utilize transfer learning due to the small size of our training dataset.

The networks are trained using the Detectron2 framework [12] which provides a wide range of pre-trained object detection and segmentation models. As a pre-processing step, we convert our data to the COCO dataset format. Image pre-processing, i.e. padding, resizing, rescaling the pixel values etc., is then performed automatically within the framework. The total loss is the sum of classification, box-regression and mask loss $L = L_{\text{cls}} + L_{\text{box}} + L_{\text{mask}}$ [18], where L_{mask} is the binary cross-entropy for independent segmentation of all masks. The models are trained using stochastic gradient descent with a learning rate of 0.00025 and a batch size of 2. They are trained for up to 10000 iterations with checkpoints every 500 iterations. We then choose the checkpoint with the lowest validation loss as our final model. We also apply data augmentation in the form of random horizontal and vertical flipping as well as random resizing with retained aspect ratio in order to minimize the generalization error.

Post-processing: To further improve our results we use knowledge from gastroenterology and knowledge from the data set structure. As mentioned above, the probability that BE and polyps are present in the same image is very low. We apply the following procedure on the polyp/BE predictions:

- We utilize the predictions from object detection and only predict masks, where there are bounding boxes present from Yolov3 and Faster R-CNN.
- As an additional criterion, pixels within bounding boxes of probability < 0.2 are labeled with 0, i.e. no disease present.
- If both polyps and BE are detected, we only produce masks for the class with higher probability, as with the detection model.

4. RESULTS

In this section, we describe our results of the two subtasks. In both settings, we highlight the performance of the algorithms for every single disease. Therefore, we create a validation set. The validation set consists of 40 images randomly chosen from the provided data (no additional data is included). We test the detection as well as the segmentation on the created validation set.

4.1. Task 1

Table 1 shows our results on our created validation set for the detection task where YOLOv3 is the described SSD algorithm, Faster R-CNN is the FASTER R-CNN algorithm with ResNet-101 backbone and ensemble with pp (post-processing) is the ensemble of those two added with the hardcoded rule. We display the mean average precision with a minimum IoU of 0.5 (mAP) [11]. We highlight the performance of the algorithms split on the five diseases. All of the algorithms have an excellent performance in detecting polyps; this is mostly due to our additional polyp training data (see chapter 2). BE is better detected by the Faster R-CNN algorithm, which is why we used this algorithm for detecting BE in the ensembled version. Notably, suspicious is one of the harder classes to correctly classify as YOLOv3 is only showing a detection performance of 10 % mAP. As depicted in Table 1, cancer is detected quite well by all of the algorithms. All things considered, the ensemble with post-processing is the best algorithm in this task. The post-processing and combination of YOLOv3 and Faster R-CNN (Ensemble with pp) enhances the performance compared to the single YOLOv3 method by 7.95%. Figure 3 shows a detection result of the YOLOv3 algorithm and a segmentation result of the Cascade Mask R-CNN. Our detection score on the EDD2020 challenge [4] test set using the ensemble architecture produces a score of 0.3360 ± 0.0852.

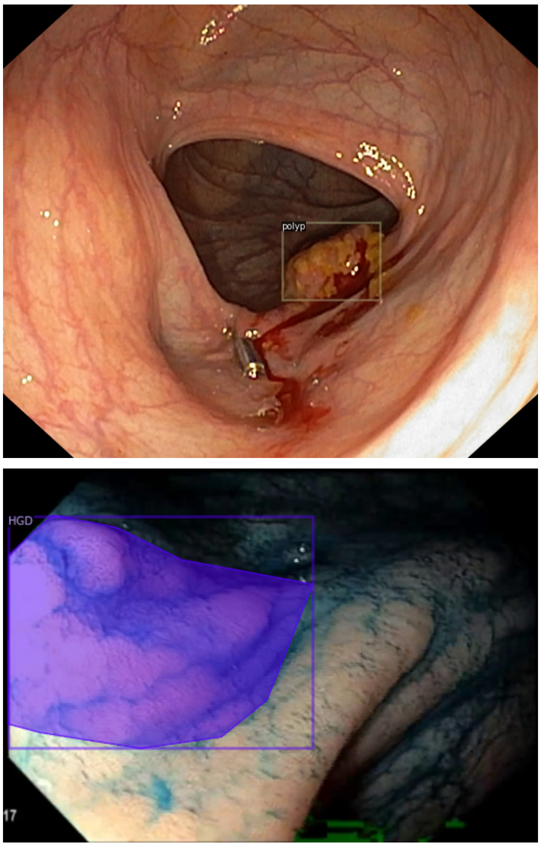


Fig. 3: Exemplary results for both detection with YOLOv3 (upper) and segmentation with Cascade Mask R-CNN (lower)

Table 1: Detection results on the validation data (mAP). MAP is the mean average precision over the five classes. Ensemble$_{pp}$ denotes the ensemble of YOLOv3 and Faster R-CNN with additional post-processing. All values are in %.

	YOLOv3	Faster R-CNN	Ensemble$_{pp}$
Polyp	84.19	73.50	84.46
BE	38.25	50.40	50.88
Suspic.	10.00	33.70	33.70
HGD	39.98	28.31	39.98
Cancer	49.99	53.20	53.20
mAP	44.49	37.29	**52.44**

4.2. Task 2

As in task 1, we evaluated our models on our validation set as a subset of the provided data on both Dice coefficient as well as intersection over union (IoU). Table 2 summarizes these results. While Mask R-CNN outperforms Cascade Mask R-CNN in both polyp and BE classes, Cascade Mask-RCNN provides better results overall, especially on the other three classes, which are comparatively underrepresented in our training data. Applying the post processing steps described in section 3 further improves the results of Cascade Mask R-CNN, but interestingly worsens the micro (μ) averaged score,

Table 2: Segmentation results on the validation data. R-CNN$_M$, R-CNN$_{CM}$ and R-CNN$_{CMpp}$ denote Mask R-CNN, Cascade Mask R-CNN and Cascade Mask R-CNN with post processing respectively. We also computed the micro averaged scores, denoted by μ mean, in contrast to mean, which is averaged over class scores. All values are in %.

	R-CNN$_M$		R-CNN$_{CM}$		R-CNN$_{CMpp}$	
	Dice	IoU	Dice	IoU	Dice	IoU
Polyp	69.41	67.03	61.57	60.08	69.07	67.58
BE	46.41	43.84	44.48	41.06	46.56	43.08
Suspic.	27.64	25.94	40.03	38.83	52.53	51.33
HGD	41.83	38.28	63.59	60.25	68.25	65.75
Cancer	53.77	52.14	55.86	54.96	57.24	57.00
mean	47.81	45.45	53.11	51.04	**58.73**	**56.95**
μ mean	36.57	27.05	**47.66**	**38.44**	45.36	37.17

which we discuss below. Our segmentation score on the EDD2020 challenge [4] test set using Cascade Mask R-CNN is then 0.6526 ± 0.3418.

5. DISCUSSION & CONCLUSION

All of our models in both tasks perform best on the polyp class and worst on the suspicious category. Since data on polyps is abundant in our training set, it is clear why the networks show good results in this area. The suspicious class, however, shows a similar amount of samples as HGD and cancer, yet, with the exception of Cascade Mask R-CNN, all models perform significantly worse on this class. This is most likely due to the unclear nature of this class as it often denotes regions belonging to different types of diseases, i.e. in some images it denotes possible cancer, whereas in others it signifies possible BE. Additionally, performing gastroenterologists often have differing opinions on what areas can be considered as suspicious, which adds further noise to our data. The performance of Cascade Mask R-CNN on suspicious and the other less represented classes can be attributed to its ensemble-like structure. The discrepancy of the micro-averaged scores can be explained as such: Our post processing severely reduces the amount of false positives, but also adds some false negatives. This improves the class-based score, since classes on one image with empty masks receive perfect scores this way. With micro-averaging, however, since precision and recall are the same, we essentially look at the per pixel accuracy of the entire mask, ultimately worsening this score.

Our model outperforms the best network from [2], namely SSD with a InceptionV3 backbone, which was partially trained using the same polyp databases and showed a precision of 73.6% on the MICCAI 2015 evaluation dataset, compared to our 84.19% with YOLOv3. AFP-net performs better than our model [3] with a precision of 88.89% on the ETIS-Larib dataset and 99.36% on the CVC-Clinic-train dataset. However, for both cases, direct comparison is difficult since both different training and different evaluation data are used. Additionally, we perform multi-class prediction, which can be a more difficult task to perform than binary prediction.

We applied state-of-the-art Deep Learning architectures for the detection and semantic segmentation of five different gastroenterological diseases. For detection, we evaluated three architectures, the YOLOv3 and the Faster R-CNN, and our combination of those algorithms. Furthermore, our ensemble includes domain knowledge-based post-processing, which further enhances our results in the challenge. For segmentation, we evaluate three models: Cascade Mask R-CNN, its predecessor Mask R-CNN, and the Cascade Mask R-CNN combined with post-processing. In the region segmentation task, the Cascade Mask R-CNN with additional post-processing reliably performs as good or better than the other networks. For future work we intend to improve our results by adding more training data, applying additional forms of data augmentation and further hyperparameter tuning. All in all, we present state-of-the-art results in the EDD challenge with our detection and segmentation applications.

6. REFERENCES

[1] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: single shot multibox detector. *CoRR*, abs/1512.02325, 2015.

[2] J. Jiang M. Liu and Z. Wang. Colonic polyp detection in endoscopic videos with single shot detection based deep convolutional neural network. *IEEE Access*, 7:75058–75066, 2019.

[3] Dechun Wang, Ning Zhang, Xinzi Sun, Pengfei Zhang, Chenxi Zhang, Yu Cao, and Benyuan Liu. Afp-net: Realtime anchor-free polyp detection in colonoscopy, 2019.

[4] Sharib Ali, Noha Ghatwary, Barbara Braden, Dominique Lamarque, Adam Bailey, Stefano Realdon, Renato Cannizzaro, Jens Rittscher, Christian Daul, and James East. Endoscopy disease detection challenge 2020. *arXiv preprint arXiv:2003.03376*, 2020.

[5] J. Bernal, N. Tajkbaksh, F. J. Snchez, B. J. Matuszewski, H. Chen, L. Yu, Q. Angermann, O. Romain, B. Rustad, I. Balasingham, K. Pogorelov, S. Choi, Q. Debard, L. Maier-Hein, S. Speidel, D. Stoyanov, P. Brandao, H. Crdova, C. Snchez-Montes, S. R. Gurudu, G. Fernndez-Esparrach, X. Dray, J. Liang, and A. Histace. Comparative validation of polyp detection methods in video colonoscopy: Results from the miccai 2015 endoscopic vision challenge. *IEEE Transactions on Medical Imaging*, 36(6):1231–1249, June 2017.

[6] J. Silva, A. Histace, O. Romain, et al. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *Int J CARS*, 9:283 – 293, 2014.

[7] Jorge Bernal, F. Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 43:99 – 111, 2015.

[8] Y. B. Guo and Bogdan J. Matuszewski. Giana polyp segmentation with fully convolutional dilation neural networks. In *VISIGRAPP*, 2019.

[9] Debesh Jha, Pia H. Smedsrud, Michael Riegler, Pål Halvorsen, Dag Johansen, Thomas de Lange, and Håvard D. Johansen. Kvasir-seg: A segmented polyp dataset. In *Proceedings of the International Conference on Multimedia Modeling (MMM)*. Springer, 2020.

[10] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, Michael Riegler, and Pål Halvorsen. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, MMSys'17, pages 164–169, New York, NY, USA, 2017. ACM.

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[12] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

[13] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5375–5384, 2016.

[14] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[15] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017.

[19] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: high quality object detection and instance segmentation. *CoRR*, abs/1906.09756, 2019.

[20] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. *CoRR*, abs/1611.05431, 2016.

[21] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014.

CEUR-WS.org/Vol-2882/paper12.pdf

Bigger Networks are not Always Better: Deep Convolutional Neural Networks for Automated Polyp Segmentation

Adrian Krenzer[1], Frank Puppe[1]
[1]Julius-Maximilian University of Würzburg, Germany
adrian.krenzer@uni-wuerzburg.de,frank.puppe@uni-wuerzburg.de

ABSTRACT

This paper presents our team's (AI-JMU) approach to the Medico automated polyp segmentation challenge. We consider deep convolutional neural networks to be well suited for this task. To determine the best architecture we test and compare state of the art backbones and two different heads. Finally, we achieve a Jaccard index of 73.74% on the challenge's test set. We further demonstrate that bigger networks do not always perform better. However, the growing network size always increases the computational complexity.

1 INTRODUCTION

Worldwide colorectal cancer (CRC) represents the third most commonly diagnosed cancer [6, 17]. According to Herszenyi and Tulassay [10] CRC attributes to 9% of all cancer incidence globally and is the fourth cause of cancer death worldwide [6, 17]. In order to detect potentially cancerous tissues early, physicians conduct a so-called colonoscopy. During this procedure, the physician searches for polyps inside the colon in order to remove them. Polyps are abnormally growing tissues that usually look like small, flat bumps or tiny fungal stems. Due to the aberrant cell growth, they can eventually become malignant or cancerous. Nevertheless, even the best physicians have a risk of overlooking a polyp. Missed polyps are not removed and can therefore have fatal consequences. Automated detecting and segmenting polyps is the task of the medico challenge [12]. This challenge is special because it is not allowed to include training data other than the 1000 provided polyp images of Jha et al. [13]. In this paper, we present our challenge results and explain how we select the networks for our final predictions.

2 RELATED WORK

In the domain of object segmentation with deep learning, there are two general state of the art approaches: Fully convolutional networks [7, 16, 21] and encoder-decoder architectures [1, 5, 24]. Some state of the art polyp segmentation methods include encoder-decoder architectures. However due to the high computational complexity of those models, polyp segmentation research focuses mostly on fully convolutional architectures to enable real-time segmentation systems [11, 28]. We consider our approaches to belong to the field of fully convolutional networks. The chosen models are based on our previous study [14], which we advance for this challenge by: Focusing exclusively on polyp segmentation, testing a new state of the art backbone in polyp segmentation [27] and comparing different architectures comprehensively.

3 APPROACH

This section focuses on our approaches for the Medico automated polyp segmentation tasks. We train all our models using a Tesla Turing RTX 8000 Nvidia GPU. For this challenge, Deep CNNs are best suited as they provide very stable outcomes in multi-class segmentation tasks like the COCO challenge [15]. Since both bounding boxes and segmentation masks are available in the dataset, we choose networks that can handle both inputs. Therefore we select the Mask R-CNN [8] and the Cascade Mask R-CNN [3]. We build both architectures based on two-stage object detection models using Faster R-CNN [19]. Therefore a region proposal network first suggests candidate bounding boxes (Regions of Interest, RoI) before making the final prediction. In this case, an additional branch is added designed to predict segmentation masks, where the suggested RoIs enhance the segmentation mask predictions. A Cascade Mask R-CNN uses an extended framework which is defined by a cascade-like composition utilizing several Mask R-CNNs with shared weight on the backbones. We train both the Cascade Mask R-CNN and Mask R-CNN with the open-source Detectron2 framework [23].

We select these types of models because of two rationales: First, the availability of both bounding boxes and segmentation masks for training purposes allows us to maximize the Mask R-CNN performance, because RoI and segmentation are closely related. Second, because the mask of polyps included in the KVASIR-SEG dataset [13] often vary significantly in size and shape we desire a network that is unaffected by those variations and determines a pixel-wise mask of the polyp. Because we use semantic segmentation, we deal with this as an instance segmentation defined by a single instance per incident per class. Therefore, we alter the ground truth bounding boxes in our data to include only one instance instead of multiple instances.

We test the Cascade Mask R-CNN and Mask R-CNN with ResNet [9] as well as the new ResNeSt [27] backbone. The latter adds a split attention block to the ResNet backbone and reconfigures the ResNet structure. This block and structure enable the network to share attention across feature-map groups. This might offer some benefits to the polyp segmentation task. Additionally, we vary the depth of both backbones, with depths of 50 and 101 for ResNet as well as 50, 101, and 200 for ResNeSt. The backbones we use consist of CNN classifiers pre-trained using the ImageNet-1k dataset [20]. The whole architecture is pre-trained on the COCO dataset [15]. Consequentially we use transfer learning to compensate for the small size of the training dataset. We train networks with the Detectron2 framework [23] and a fork of the Detectron2 framework published by Zhang et al. [27]. Both provide a wide range of pre-trained object detection and segmentation models. Prior to the actual processing, we convert our data to the COCO dataset format. Afterward, the required image preprocessing steps, i.e. padding,

---

Copyright 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
MediaEval'20, December 14-15 2020, Online

Table 1: Segmentation results on the validation data. R50 and R101 denote ResNet50 and ResNet100. Rt50, Rt101 and Rt200 denote ResNeSt50, ResNeSt101 and ResNeSt200. Cascade R-CNN denotes Cascade Mask R-CNN. All values excluding FPS are in %.

	Mask R-CNN				Cascade R-CNN			
	IoU	Dice	Acc	FPS	IoU	Dice	Acc	FPS
R50	71.0	78.9	90.9	**13**	73.2	81.4	93.8	9.8
R101	72.3	80.0	91.8	11.8	74.1	82.1	94.3	8.7
Rt50	72.8	78.7	90.4	10.9	75.2	81.9	94.2	8.2
Rt101	73.9	80.8	93.2	9.0	**75.9**	**83.1**	**95.7**	7.1
Rt200	-	-	-	-	73.3	81.6	93.4	2.9

resizing, rescaling the pixel values, etc., are automatically performed within the framework.

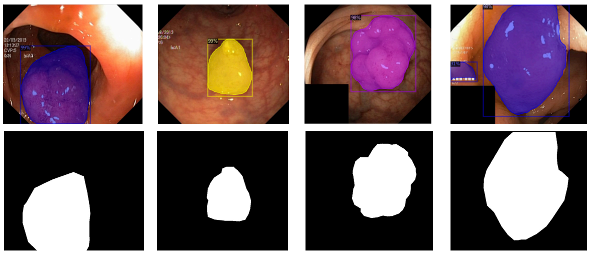
We define the total loss as the sum of classification, box-regression and mask loss $L = L_{\text{cls}} + L_{\text{box}} + L_{\text{mask}}$ [8], where L_{mask} is the binary cross-entropy for autonomous segmentation of all masks. The training of all models includes a stochastic gradient descent using a learning rate of 0.00025 and a batch size of 16. Every model trains for up to 80000 iterations, maintaining checkpoints every 300 iterations. Afterward, we adopt the checkpoint with the lowest validation loss for the final outcome. Additionally, we utilize random horizontal flipping, vertical flipping, and random resizing as data augmentation while retaining aspect ratio to diminish the generalization error.

4 RESULTS AND ANALYSIS

We evaluate the models on our validation dataset which is a subset of the Kvasir-SEG data [13]. For the evaluation we consider quality and speed. For quality we compute the dice coefficient, intersection over union (IoU), and accuracy (Acc). For speed we specify frames per second (FPS). All our validations are carried out using an Nvidia V100 GPU within the cloud solution of Google Colab [2]. Table 1 depicts our results. While Cascade Mask R-CNN outperforms Mask R-CNN in every quality metric, Mask R-CNN is faster with computation. However, the architecture's speed shows a clear pattern: the Mask R-CNN using the smallest backbone (lowest computational complexity) is the fastest, and Cascade Mask R-CNN (highest computational complexity) with the largest backbone is the slowest. Comparing the ResNet and RestNeSt backbone: Using the ResNeSt backbone results in higher scores in all metrics. Nevertheless, the RestNeSt backbone increases the computational complexity and therefore decreases FPS. Concerning the depth of the network: Changing the depth from 50 to 101 increases the quality of the results. This implies that a deeper backbone may always result in better quality. However, our results show that a larger backbone not always causes better quality, but always diminishes the speed due to higher computational complexity, in our case dropping FPS down to 2.9 for ResNeSt200. We evaluate ResNeSt200 backbone only with the Cascade Mask R-CNN because there are no pre-trained weights available for the Mask R-CNN version.

Overall, Cascade Mask R-CNN with a ReStNest101 backbone provides the best quality results. Therefore, we consider this backbone for the quality task of the Medico challenge. For the efficacy task of the challenge we choose the Cascade Mask R-CNN with ReStNest50 backbone. It is faster and less taxing on memory than ReStNest101 while still maintaining high-quality results. Our challenge scores for the quality task are an IoU of 0.737. For the efficacy task our results are an IoU of 0.721 while computing with 3.36 FPS on an Nvidia GTX 1080. To qualitatively demonstrate a set of our results, we depict the four best and worst classified images of our validation set in figure 1. The algorithm performs best on big, unconcealed polyps. Nevertheless, small polyps like shown in the first three images of figure 1_b are harder to segment. In addition, concealing the polyp with a tool like in the last image of figure 1_b prevents the algorithm from detecting the polyp.

a) Best performing



b) Worst performing

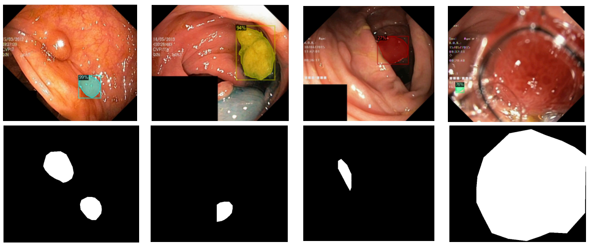


Figure 1: Qualitative results of the Cascade Mask R-CNN with ResNeSt101 backbone. Binary images are ground truth and rgb images with are predictions. Different colors are just highlighting the predictions.

5 CONCLUSION AND OUTLOOK

In summary, our results suggest that using a deeper neural network, extending it with another backbone, or adding a computationally more expensive architecture like Cascade Mask R-CNN leads to higher quality segmentations. Nevertheless, the increasing network size is not always beneficial. Moreover, we demonstrate that the ReStNeSt101 backbone combined with the Cascade Mask R-CNN structure is the best segmentation algorithm among our examples.

Further research could extend our architectures and compare them with other state of the art segmentation models like the DeepLabv3+ [18], HRNet [22], MRFM [26]. Those three architectures and the proposed architecture are currently the best performing architectures on object segmentation benchmarks [18, 22, 26, 27]. Especially promising is the speed and quality trade off using HarDNet [4] and BiSeNet [25] for further evaluations.

REFERENCES

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence* 39, 12 (2017), 2481–2495.

[2] Ekaba Bisong. 2019. Google Colaboratory. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform*. Springer, 59–64.

[3] Zhaowei Cai and Nuno Vasconcelos. 2019. Cascade R-CNN: High Quality Object Detection and Instance Segmentation. *CoRR* abs/1906.09756 (2019). arXiv:1906.09756 http://arxiv.org/abs/1906.09756

[4] Ping Chao, Chao-Yang Kao, Yu-Shan Ruan, Chien-Hsiang Huang, and Youn-Long Lin. 2019. HarDNet: A Low Memory Traffic Network. *CoRR* abs/1909.00948 (2019). arXiv:1909.00948 http://arxiv.org/abs/1909.00948

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*. 801–818.

[6] Pasqualino Favoriti, Gabriele Carbone, Marco Greco, Felice Pirozzi, Raffaele Emmanuele Maria Pirozzi, and Francesco Corcione. 2016. Worldwide burden of colorectal cancer: a review. *Updates in surgery* 68, 1 (2016), 7–11.

[7] Chaoyi Han, Yiping Duan, Xiaoming Tao, and Jianhua Lu. 2019. Dense convolutional networks for semantic segmentation. *IEEE Access* 7 (2019), 43369–43382.

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. 2017. Mask R-CNN. *CoRR* abs/1703.06870 (2017). arXiv:1703.06870 http://arxiv.org/abs/1703.06870

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *CoRR* abs/1512.03385 (2015). arXiv:1512.03385 http://arxiv.org/abs/1512.03385

[10] Laszlo Herszenyi and Zsolt Tulassay. 2010. Epidemiology of gastrointestinal and liver tumors. *Eur Rev Med Pharmacol Sci* 14, 4 (2010), 249–258.

[11] Debesh Jha, Sharib Ali, Håvard D Johansen, Dag D Johansen, Jens Rittscher, Michael A Riegler, and Pål Halvorsen. 2020. Real-Time Polyp Detection, Localisation and Segmentation in Colonoscopy Using Deep Learning. *arXiv preprint arXiv:2011.07631* (2020).

[12] Debesh Jha, Steven A. Hicks, Krister Emanuelsen, Håvard D. Johansen, Dag Johansen, Thomas de Lange, Michael A. Riegler, and Pål Halvorsen. Medico Multimedia Task at MediaEval 2020:Automatic Polyp Segmentation.

[13] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. 2020. Kvasir-SEG: A Segmented Polyp Dataset. In *Proc. of International Conference on Multimedia Modeling (MMM)*. 451–462.

[14] Adrian Krenzer, A. Hekalo, and F. Puppe. 2020. Endoscopic Detection And Segmentation Of Gastroenterological Diseases With Deep Convolutional Neural Networks. In *EndoCV@ISBI*.

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[16] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3431–3440.

[17] Michael Marmot, T Atinmo, T Byers, J Chen, T Hirohata, A Jackson, W James, L Kolonel, S Kumanyika, C Leitzmann, and others. 2007. Food, nutrition, physical activity, and the prevention of cancer: a global perspective. (2007).

[18] Heungmin Oh, Minjung Lee, Hyungtae Kim, and Joonki Paik. 2020. Metadata Extraction Using DeepLab V3 and Probabilistic Latent Semantic Analysis for Intelligent Visual Surveillance Systems. In *2020 IEEE International Conference on Consumer Electronics (ICCE)*. IEEE, 1–2.

[19] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *CoRR* abs/1506.01497 (2015). arXiv:1506.01497 http://arxiv.org/abs/1506.01497

[20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. 2014. ImageNet Large Scale Visual Recognition Challenge. *CoRR* abs/1409.0575 (2014). arXiv:1409.0575 http://arxiv.org/abs/1409.0575

[21] Evan Shelhamer, Jonathan Long, and Trevor Darrell. 2017. Fully convolutional networks for semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence* 39, 4 (2017), 640–651.

[22] Andrew Tao, Karan Sapra, and Bryan Catanzaro. 2020. Hierarchical Multi-Scale Attention for Semantic Segmentation. *arXiv preprint arXiv:2005.10821* (2020).

[23] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. https://github.com/facebookresearch/detectron2. (2019).

[24] Jimei Yang, Brian Price, Scott Cohen, Honglak Lee, and Ming-Hsuan Yang. 2016. Object contour detection with a fully convolutional encoder-decoder network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 193–202.

[25] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. 2020. BiSeNet V2: Bilateral Network with Guided Aggregation for Real-time Semantic Segmentation. *arXiv preprint arXiv:2004.02147* (2020).

[26] Jianlong Yuan, Zelu Deng, Shu Wang, and Zhenbo Luo. 2020. Multi Receptive Field Network for Semantic Segmentation. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1883–1892.

[27] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Mueller, R. Manmatha, Mu Li, and Alexander Smola. 2020. ResNeSt: Split-Attention Networks. (2020). arXiv:cs.CV/2004.08955

[28] Jiafu Zhong, Wei Wang, Huisi Wu, Zhenkun Wen, and Jing Qin. 2020. PolypSeg: An Efficient Context-Aware Network for Polyp Segmentation from Colonoscopy Videos. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 285–294.

Journal of
Imaging

Article

A Real-Time Polyp-Detection System with Clinical Application in Colonoscopy Using Deep Convolutional Neural Networks

Adrian Krenzer [1,2,*], Michael Banck [1,2], Kevin Makowski [1], Amar Hekalo [1], Daniel Fitting [2], Joel Troya [2], Boban Sudarevic [2,3], Wolfgang G. Zoller [2,3], Alexander Hann [2] and Frank Puppe [1]

[1] Department of Artificial Intelligence and Knowledge Systems, Julius-Maximilians University of Würzburg, Sanderring 2, 97070 Würzburg, Germany
[2] Interventional and Experimental Endoscopy (InExEn), Department of Internal Medicine II, University Hospital Würzburg, Oberdürrbacher Straße 6, 97080 Würzburg, Germany
[3] Department of Internal Medicine and Gastroenterology, Katharinenhospital, Kriegsbergstrasse 60, 70174 Stuttgart, Germany
* Correspondence: adrian.krenzer@uni-wuerzburg.de

Abstract: Colorectal cancer (CRC) is a leading cause of cancer-related deaths worldwide. The best method to prevent CRC is with a colonoscopy. During this procedure, the gastroenterologist searches for polyps. However, there is a potential risk of polyps being missed by the gastroenterologist. Automated detection of polyps helps to assist the gastroenterologist during a colonoscopy. There are already publications examining the problem of polyp detection in the literature. Nevertheless, most of these systems are only used in the research context and are not implemented for clinical application. Therefore, we introduce the first fully open-source automated polyp-detection system scoring best on current benchmark data and implementing it ready for clinical application. To create the polyp-detection system (ENDOMIND-Advanced), we combined our own collected data from different hospitals and practices in Germany with open-source datasets to create a dataset with over 500,000 annotated images. ENDOMIND-Advanced leverages a post-processing technique based on video detection to work in real-time with a stream of images. It is integrated into a prototype ready for application in clinical interventions. We achieve better performance compared to the best system in the literature and score a F1-score of 90.24% on the open-source CVC-VideoClinicDB benchmark.

Keywords: machine learning; deep learning; endoscopy; gastroenterology; automation; object detection; video object detection; real-time

check for updates

Citation: Krenzer, A.; Banck, M.; Makowski, K.; Hekalo, A.; Fitting, D.; Troya, J.; Sudarevic, B.; Zoller, W.G.; Hann, A.; Puppe, F. A Real-Time Polyp-Detection System with Clinical Application in Colonoscopy Using Deep Convolutional Neural Networks. *J. Imaging* **2023**, *9*, 26. https://doi.org/10.3390/jimaging9020026

Academic Editor: Mohamed Medhat Gaber

Received: 19 December 2022
Revised: 18 January 2023
Accepted: 19 January 2023
Published: 24 January 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Colorectal cancer (CRC) is the second leading cause of cancer-related deaths worldwide [1]. One of the best methods to avoid CRC is to perform a colonoscopy to detect the potential disease as early as possible. A colonoscopy examines the large intestine (colon) with a long flexible tube inserted into the rectum. A small camera is mounted at the end of the tube, enabling the physician to look inside the colon [2]. During this procedure, the colonoscopist searches for polyps and examines them closely. Polyps are protrusions of the mucosal surface of various shapes and sizes that can be benign or malignant and, thus, can develop into CRC. Polyps grow on the colon lining, which often does not cause symptoms. The two main types are non-neoplastic and neoplastic polyps. Non-neoplastic polyps usually do not become cancerous and polyps of type neoplastic might become cancerous [2]. Even if many polyps are not cancerous, some become colon cancer. Ideally, the colonoscopist detects every polyp during a colonoscopy and decides, on closer inspection, whether it needs to be removed. Nevertheless, there is still a potential risk of polyps being missed. It has been shown that up to 27% of diminutive polyps are overlooked by physicians [3,4], which happens due to lack of experience or fatigue. It has also been shown that even a general error rate of 20–24% leads to a high risk of patients dying from

CRC [5,6]. Two studies have shown that the missing rate is related to the size of the polyp. Kim et al., showed that polyps of size ≤5 mm, 5–10 mm and ≥10 mm had a missing rate of 35.4%, 18.8%, and 4.9%, respectively [7]. Ahn et al., demonstrated missing rates of 22.9%, 7.2%, and 5.8% for sizes of ≤5 mm, 5–10 mm, and ≥10 mm, respectively [8]. Both studies also found that the missing rate was higher when the patient had more than one polyp. Additionally, a systematic review calculated a similar value and received missing rates of 2%, 13%, and 26% for polyp sizes of ≥10 mm, 5–10 mm, and 1–5 mm, respectively [6]. This indicates that smaller polyps have a higher risk of being missed by the colonoscopist. Missed polyps can have fatal consequences for the patient. Thus, the colonoscopist must detect and remove all potential cancerous polyps to minimize the risk of CRC [8].

To avoid missing polyps, computer science research methods have been developed to assist physicians during the colonoscopy. The use of computers to detect polyps is called *computer-aided detection (CAD)*. The research field already has publications examining the problem of polyp detection. Nevertheless, most of these systems are only used in research context and are not developed to be ready for clinical application. There are commercial systems ready for clinical application; however, they are very expensive [9–12]. Therefore, we introduce the first fully open-source system scoring best on current benchmark data and implementing it for clinical-ready applications.

The main contributions of our paper are:

(1) *We introduce the first fully open-source (https://fex.ukw.de/public/download-shares/d8NVHA2noCiv7hXffGPDEaRfjG4vf0Tg, accessed on 18 December 2022), clinically ready, real-time polyp-detection system.*
(2) *We show that the system outperforms current systems on benchmark data with real-time performance.*
(3) *We introduce a novel post-processing method working in real-time based on REPP [13] and use a new metric for polyp detection, which has value for clinical usage.*

Additionally, the polyp-detection system was publicly funded and developed by computer engineers and endoscopists in the same workgroup to ensure high-quality polyp detection. Figure 1 shows the results of the polyp-detection system. To overview existing work and properly allocate our paper in the literature, we describe a brief history from general polyp detection with handcrafted features to state-of-the-art polyp detection with deep learning techniques.

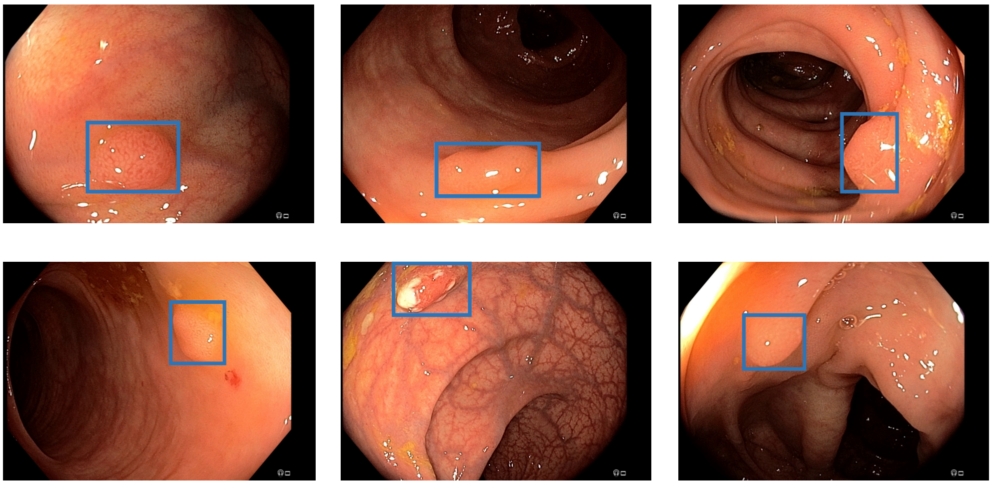


Figure 1. Detection examples. This figure illustrated some detection examples of the polyp-detection system on our own data (EndoData).

1.1. A Brief History of Computer-Aided Polyp Detection

The field of CAD is divided into two subfields, CADe and CADx. CADe deals with the detection and localization of polyps and CADx deals with the characterization of polyps.

This paper will focus exclusively on the CADe area. This section only considers methods that localize polyps by specifying a rectangular section of the screen, a *bounding box*.

1.1.1. Computer-Aided Detection with Handcrafted Features

The first approaches for computer-aided detection of polyps were explored as early as the late 1990s. For example, Krishnan et al., proposed using curvature analysis to detect polyps by shape [14]. Another method was presented in 2003 by Karkanis et al. They used wavelet transform to detect polyps based on their color and texture [15]. Hwang et al., used a new technique to distinguish the elliptical shape features of polyp regions from non-polyp regions. They compared the features based on curvature, intensity, curve direction, and distance from the edge [16]. Bernal et al. (2012) proposed another method by converting images of polyps to grayscale so that the elevations of the polyps could be seen. Subsequently, the authors illuminated the outlines of the polyps, which they termed valleys. Based on the intensity of the valleys, the polyps were extracted and localized [17].

Furthermore, expert knowledge was used to handcraft rules for detecting polyps based on specific properties, such as size, shape, and color. Newer examples can be found in [18,19], both of which use support vector machines (SVMs). Additionally, real-time detection with handcrafted features was tested in clinical applications [20]. The authors used a weighted combination of color, structure, textures, and motion information to detect image areas where a polyp is possibly located. The detection rate was 73%. Nevertheless, the rise of convolutional neural network (CNN)-based methods in image processing has superseded all of these techniques, as CNN methods have proven to show better results.

1.1.2. Methods Involving CNNs

Computer-aided polyp recognition was particularly shaped by various deep learning methods from the beginning of the last decade. We listed an overview of the essential models on still image datasets in Table 1. Specifically, a great deal of research interest has developed in the object recognition capabilities of CNNs. For example, in 2015, authors Zhu et al., presented a seven-layer CNN as a feature extractor with a SVM as a classifier to detect anomalies in endoscopy images [21]. The system was trained on custom data. The earlier approaches considered using an existing CNN architecture to localize polyps, the AlexNet [22–24]. This was developed for general image classification, i.e., not specifically for the medical field. The paper by Tajbakhsh et al. [23] states that the AlexNet [22] for polyp detection is better not trained thoroughly, i.e., starting from random weights, but the already pre-trained weights should be used. It is shown that *transfer learning* is a practical approach in the presence of limited data, as generally given in the medical field.

Table 1. Overview of polyp detection models on still image datasets. The table includes the following abbreviation: DenseNet-UDCS: densely connected neural network with unbalanced discriminant and category sensitive constraints; ADGAN: attribute-decomposed generative adversarial networks; CenterNet: center network; SSD: single shot detector; YOLO: you only look once; R-CNN: region-based convolutional neural network.

Author	**Year**	**Method**	**Test Dataset**	**F1-Score**	**Speed**
Yuan et al. [25]	2020	DenseNet-UDCS	Custom	81.83%	N/A
Liu et al. [26]	2020	ADGAN	Custom	72.96%	N/A
Wang et al. [27]	2019	CenterNet	CVC-ClinicDB	97.88%	52 FPS
Liu et al. [28]	2019	SSD	CVC-ClinicDB	78.9%	30 FPS
Zhang et al. [29]	2019	SSD	ETIS-Larib	69.8	24 FPS
Zheng et al. [30]	2018	YOLO	ETIS-Larib	75.7%	16 FPS
Mo et al. [31]	2018	Faster R-CNN	CVC-ClinicDB	91.7%	17 FPS

Yuan et al. [24] first extract an attractive image section via edge-finding algorithms as a preliminary step and use it as input to the AlexNet [22]. This resulted in a high recall of 91.76% compared to the state-of-the-art at that time. Mo et al. [31] are the first to use the unmodified *faster region-based convolutional neural network* (Faster R-CNN) [32] architecture for polyp detection. This allows the detection of polyps that are mostly obscured or very close to the camera, unlike previous models. The model is trained on the CVC-ClinicDB data. The model is robust to illumination changes or circular bubbles, but it misses some smaller polyps and sometimes becomes too guided by an oval shape, increasing the number of false positives (FP). The authors also wanted to focus on these problems in the future. Shin et al. [33] were the first to use the inception-residual neural network (ResNet) [34] architecture unmodified for polyp detection. The model is trained on the ASU-Mayo-Video-DB. They also added two post-processing methods, *false positive learning* and *offline learning*, to further improve the model's performance. The advantage of the model was that the entire frame could be used for training rather than a previous patch extraction step. As with Mo et al., one problem with the model is the high number of FP triggered by polyp-like structures. The authors plan to focus on improving the speed in the future, which was only 2.5 frames per second (FPS). Zheng et al. [30] use the unmodified you only look once (YOLO) architecture [35]. Again, the advantages are that there is only one processing step, so there is no preliminary step to extract an regions of interest (RoI). As a result, the model was faster than the two-step methods but only achieved 16 FPS. Further, the authors note that the CNN features of *white light* and *narrow-band* images differed greatly, thus, they should be considered separately. The model is trained on the CVC-CLinicDB, CVC-ColonDB, and custom data. Liu et al. [28] implemented and compared different back-end models as feature extractors for the *single shot detection* architecture (SSD) [36]. These were *ResNet50* [37], *Visual Geometry Group-16 (VGG-16)* [38], and *InceptionV3* [39], with InceptionV3 showing the best balanced result. The advantages of the models are robustness to size and shape, as well as speed, which is real-time capable at 30 FPS. The models are trained on the CVC-CLinicDB, CVC-ColonDB, and ETIS-Larib data. In the future, other back-end models could result in a further boost in performance.

Zhang et al. [40] used the *SSD-GPNet*, which is based on the SSD architecture [36], but tries to incorporate information that is normally lost by the standard pooling layers into the result through various customized pooling methods. Since it is based on the SSD architecture, this method is also fast and achieves real-time capability at 50 FPS; it also achieves good recall, especially for small polyps. In the future, the authors want to test their approaches for other diseases and find more ways to use as much image information as possible without increasing the complexity of the models. Furthermore, Zhang et al., presented another deep learning method for polyp detection and localization. They presented a special single-shot multi-box detector-based CNN model that reused displaced information through max-pooling layers to achieve higher accuracy. At 50 FPS, the method provided real-time polyp detection while achieving a mean average precision of up to 90.4% [40]. The model is trained on custom data. Authors Bagheri et al., staged a different idea in which they first converted the input images into three color channels and then passed them to the neural network. This allows the network to learn correlated information using the preprocessed information about the color channels to locate and segment polyps [41]. With the same goal, Sornapudi et al., in their paper, used region-based CNNs to localize polyps in colonoscopy images and in wireless capsule endoscopy (WCE) images. During localization, images were segmented and detected based on polyp-like pixels [42].

In addition to CNNs, research is also being conducted on other deep learning methods for polyp detection. For example, a special sparse autoencoder method called stacked sparse autoencoder with image manifold constraint was used by Yuan and Meng [43] to detect polyps in WCE images. A sparse autoencoder is an artificial neural network commonly used for unsupervised learning methods [44]. The sparse autoencoder achieved 98% accuracy in polyp detection [43]. The system is trained and tested on the ASU-Mayo-Video-DB. Wang et al. [27] used the *AFP-Net*. Unlike an SSD model, an AFP-Net model does not require predefined anchor boxes, it is *anchor free*. It was the first application of such an architecture for polyp detection. Through *context enhancement module* (CEM), a *cosine ground-truth projection* and a customized loss function, the speed was increased and 52.6 FPS was achieved, which is real-time capability. In the future, the authors still want to improve the detection of hard-to-detect small and flat polyps. The model is trained on the CVC-ClinicVideoDB. Liu et al. [26] used an *anomaly detection generative adversarial network* (ADGAN), which is based on the Wasserstein GAN [45]. ADGAN aims to learn only based on healthy images without polyps to reconstruct them. If this model receives an image with a polyp as input, the model cannot reconstruct it, so at this point in the output, there is a noticeably large difference from the input, which is easy to check. The problem of connecting the input to the latency space of the GAN was solved by a second GAN. In addition, a new loss function was added to improve performance even further. The model is trained on custom data.

The advantage of this approach is that no costly annotated datasets are needed and significantly larger amounts of data of normal intestinal mucosa are available. For the future, the authors want to ensure that frames with biological disturbances, such as stool residue or water, are also processed well since these are often sorted out beforehand from many datasets. Yuan et al. [25] use the *DenseNet-UDCS* architecture for frame classification, not localization, of WCE images. The DenseNets [46] structure is kept unchanged, but the loss function is adapted. On the one hand, weighting is introduced to compensate for the significant imbalance in the size of the classes (with or without polyps). On the other hand, the loss function is adapted to be class sensitive. It forces that similar features are learned for the same class and the features of the other class have as significant differences as possible. These adaptations improve performance and can be easily applied to other applications and models. In the future, the researchers still want to find a way to compensate for different illuminations by pre-processing and testing attention-based methods. Another method is to use transformers in combination with CNNs. Zhang et al., used in parallel the ability to view global information of the whole image through the attention layers of the transformers and the detection of the CNNs to efficiently segment polyps. In addition, a new fusion technique called BiFusion was used to fuse the features obtained by the transformers and the CNNs. The resulting method called TransFuse stood out mainly because of its segmentation time of 98.7 FPS [47]. The model is trained on custom data.

Furthermore, Jha et al. [48] proposed a segmentation, detection, and classification model. The model achieves a mean average precision (mAP) of 81.66 while being very fast, with 180 FPS on an NVIDIA Volta 100 GPU. The results are evaluated on the Kvasir-SEG dataset. Another approach is combining different networks in an ensemble to increase the polyp-detection performance further. While ensemble techniques significantly increase detection systems' performance, the downside is that those systems mostly have high latency. For example, if an ensemble combined five different neural networks, the computational complexity would be increased at least five times. Therefore, ensemble techniques are not implemented for real-time detection. The paper of [49] shows a polyp-detection system using an ensemble. The authors combined three different models by using majority voting to increase the performance of the polyp-detection system.

Additionally, Livovsky et al. [50] trained a large-scale AI called detection of elusive polyps (DEEP2). They used 3611 h of colonoscopy videos for training and 1393 h for testing. They trained and evaluated their AI on a custom dataset. As neural network architecture they used RetinaNet. Livovsky et al., achieved a recall of 88.5% with polyps visible longer than 5 s. Moreover, they showed recall and specificity results for different lengths of polyps visibility.

1.1.3. 3D and Temporal Methods

While older publications are evaluated on still-image benchmarks, such as the CVC-ClinicDB, the new state-of-the-art is evaluated on the more challenging and more realistic video dataset, such as CVC- VideoClinicDB. For example, Wang et al. [27] have a high score of 97.88% on the CVC-ClinicDB dataset. Nevertheless, this dataset only involves

612 still images. We reconstructed the algorithm of Wang et al. [27] but could not reproduce the results on the video datasets. For these video datasets, all frames are extracted and polyps in these frames are annotated with corresponding bounding boxes. We listed an overview of the essential models on video datasets in Table 2. Another approach is to use temporal information within the video. In the methods mentioned above, only single frames are considered. Thus, information that is given by the sequence of the frames is lost. In Itoh et al. [51], temporal information is included through a *3D-ResNet*. In addition, a weighted loss function and selection of so-called *hard negative frames* address the problem of training-data class imbalance. These lead to an improvement of 2% F1-score. However, one problem is that the model has a higher probability of being overfitted than its 2D counterpart because it has more parameters and is not applicable in real-time. Zhang et al. [29] combine the output of a conventional SSD model [36] via a *Fusion module* with a generated *optical flow*. This is similar to a heat map showing motion over short periods and is easy to compute. This approach is much less complex and, therefore, faster than other temporal systems that use 3D methods; still, it is not applicable for real-time polyp detection. Misawa et al. [52] use a *3D-CNN* to include temporal information. This allows many different types of polyps to be well detected. The model is trained on custom data.

Table 2. Overview of polyp detection models on video datasets.

Author	Year	Method	Test dataset	F1-Score	Speed
Nogueira et al. [53]	2022	YOLOv3	Custom	88.10%	30 FPS
Xu et al. [54]	2021	CNN + SSIM	CVC-VideoClinicDB	75.86%	N/A
Livovsky et al. [50]	2021	RetinaNet	Custom	N/A	30 FPS
Misawa et al. [11]	2021	YOLOv3	SUN-Colonoscopy	87.05%	30 FPS
Qadir et al. [55]	2020	Faster R-CNN	CVC-VideoClinicDB	84.44%	15 FPS
		SSD	CVC-VideoClinicDB	71.82%	33 FPS
Yuan et al. [25]	2020	DenseNet-UDCS	Custom	81.83%	N/A
Zhang et al. [40]	2019	SSD-GPNet	Custom	84.24%	50 FPS
Misawa et al. [52]	2019	3D-CNN	Custom	N/A	N/A
Itoh et al. [51]	2019	3D-ResNet	Custom	N/A	N/A
Shin et al. [33]	2018	Inception ResNet	ASU-Mayo-Video-DB	86.9%	2.5 FPS
Yuan et al. [24]	2017	AlexNet	ASU-Mayo-Video-DB	N/A	N/A
Tajbakhsh et al. [23]	2016	AlexNet	Custom	N/A	N/A

Additionally, Qadir et al. [55] use a conventional localization model, such as SSD [36], or Faster R-CNN [32], and further process the output of these through a *False Positive Reduction Unit*. This looks at the position of the generated bounding boxes over the seven preceding and following frames and tries to find and possibly correct outliers. Because future frames are used, there is a small delay, but the actual calculation of the *False Positive Reduction Unit* is fast. A different and promising method was provided by Qadir et al., in a two-step process. They used a CNN in the first step, which generated several RoIs for classification. Then, these proposed RoIs were compared based on the subsequent frames and their RoIs and classified into true positive (TP) and false positive (FP). This method assumes that the frame in a video should be similar to the next frame. It intends to reduce the percentage of false predictions [55]. Because CNNs are sensitive to noise in the data, they may produce a high count of FPs. Another approach is therefore using a two-stage method that first suggests multiple RoIs. Then, the current proposed RoIs are categorized as TPs and FPs by considering the RoIs of the following frames [55]. With this method, they are reducing the number of FPs and reaching state-of-the-art results. The model is trained on the ASU-Mayo-Video-DB and custom data.

Furthermore, Misawa et al., developed a real-time polyp-detection system based on YOLOv3. The system has a speed of 30 FPS and achieves a F1-score of 97.05% on their open-source dataset (SUN-Colonoscopy) [11]. Moreover, Xu et al. [54] designed a 2D CNN detector including spatiotemporal information involving a structural similarity (SSIM) to advance polyp detection further while maintaining real-time speed. The model is trained on custom data. In 2022 Nogueira et al., published a real-time polyp-detection system using the YOLOv3 model with an object-tracking algorithm. The system scores a F1-score of 88.10% on their custom dataset.

2. Materials and Methods

This section explains the software and hardware for our polyp-detection system. We call our polyp-detection system ENDOMIND-Advanced. An early, preliminary version of our detection system was experimentally tested and called ENDOMIND [56]. Nevertheless, ENDOMIND used an early version of YOLOv5 that did not involve our preprocessing, hyperparameter, optimization, and post-processing, and was trained with a smaller dataset. First, we introduce our datasets for training and testing the AI. Afterward, we illustrate typical challenges in the field of automatic polyp detection. We continue by showing our data preprocessing and data augmentation. We then show the full polyp-detection system and explain its components. The full polyp-detection system involves the CNN-based YOLOv5 model and our implemented post-processing solution real-time REPP (RT-REPP), which uses an algorithm called *Robust and Efficient Post-Processing* (REPP) [13]. We close this section by elaborating on the clinical application of our system.

2.1. Datasets

Obtaining qualitative data on an appropriate scale is often one of the biggest problems for applying deep learning methods. This is no different for colonoscopy videos/images for polyp detection. The difficulties in the acquisition are due to data protection issues on the one hand and the expensive and time-consuming but necessary annotation of the data by experienced medical experts. Therefore, for developing our model, we use our own data and all the publicly available data we could find on the internet and in the literature. For training our model, we combined the available online sources and our own data to forge a dataset of 506,338 images. Figure 2 shows an overview of the data material. The details about creating our own dataset will follow below. All data consist of images and bounding box coordinates of boxes referring to the image. For a listing of publicly available datasets we used for training, we show the following overview:

- ETIS-Larib [57] 2014: It contains 196 polyp images from 34 different videos and shows 44 different polyps. ETIS-LaribPolypDB [57] is from the *MICCAI 2015 Endoscopic Vision Challenge* and was used as the testing dataset in the challenge. Here, we include this dataset in our training dataset. It has 196 polyp images with the corresponding mask for boxes. For our training, we extracted the bounding boxes from the segmentation masks. The size of the images is 348 × 288 pixels. This dataset contains no healthy mucosa images. This dataset contains no healthy mucosa images. The data are available on request in the CVC-Colon repository (http://www.cvc.uab.es/CVC-Colon/index.php/databases/, accessed on 18 December 2022).
- CVC-VideoClinicDB [58] 2017: The CVC-VideoClinicDB [59] dataset was provided in the context of the GIANA sub-challenge that was part of the *MICCAI 2017 Endoscopic Vision Challenge*. This dataset contains 18,733 frames from 18 videos without ground truth and 11,954 frames from 18 videos with ground truth. We exclusively used these frames for final evaluation. It has to be noted that the ground truth masks that label a polyp are approximated by using ellipses. Furthermore, we filtered out all images with no polyps (empty mask) and only used frames with at least one polyp for training. The size of the images is 574 × 500 pixels. This dataset is only used for testing in this paper. The data are available upon request in the CVC-Colon repository (http://www.cvc.uab.es/CVC-Colon/index.php/databases/, accessed on 18 December 2022).
- CVC-EndoSceneStill [60] 2017: It combines *CVC-ColonDB* and *CVC-ClinicDB* and contains 912 polyp images from 44 videos of 36 patients. CVC-EndoSceneStill [60]

is a dataset that combines CVC-ColonDB [17] (CVC-300) and CVC-Clinic-DB [58,61] (CVC-612). Both datasets gave each image a border, specular, lumen, and segmentation mask. The border mask marks the black border around each image, the specular mask indicates the reflections that come from the endoscope light, and the lumen mask labels the intestinal lumen, which is the space within an intestine. The segmentation mask contains polyp markings that tag visible polyps within a picture. Because we needed the bounding box from a polyp, we only used the segmentation masks and extracted a bounding box by calculating a box that fits around a single blob. The dataset CVC-ColonDB [17,60] contains 300 selected images from 13 polyp video sequences with a resolution of 574 × 500 and CVC-Clinic-DB [58,60,61] holds 612 images from 31 polyp video sequences with a size of 348 × 288 pixels. This dataset contains no healthy mucosa images. The data are available on request in the CVC-Colon repository (http://www.cvc.uab.es/CVC-Colon/index.php/databases/, accessed on 18 December 2022).

- Kvasir-SEG [62] 2020: The dataset contains 1000 polyp images with corresponding 1071 masks and bounding boxes. Dimensions range from 332 × 487 to 1920 × 1072 pixels. Gastroenterologists verified the images from *Vestre Viken Health Trust* in Norway. Most images have general information displayed on the left side and some have a black box in the lower left corner, which covers information from the endoscope position marking probe created by ScopeGuide (Olympus). This dataset contains no healthy mucosa images. The data are available in the Kvasir-SEG repository (https://datasets.simula.no/kvasir-seg/, accessed on 18 December 2022).
- SUN Colonoscopy Video Database [11] 2021: The database was developed by Mori Laboratory, Graduate School of Informatics, Nagoya University. It contains 49,136 fully annotated polyp frames taken from 100 different polyps. These images were collected at the Showa University Northern Yokohama and annotated by expert endoscopists at Showa University. Additionally, 109,554 non-polyp frames are included. The size of the images is 1240 × 1080 pixels. The data are available in the SUN Colonoscopy Video repository (http://sundatabase.org/, accessed on 18 December 2022).
- CVC-Segementation-HD [60] 2017: This dataset was made available within the GIANA Polyp Segmentation sub-challenge that was part of the *MICCAI 2017 Endoscopic Vision Challenge*. It contains 56 high-resolution images with a size of 1920 × 1080 pixels. This dataset contains no healthy mucosa images. There is a binary mask from which we have extracted the bounding boxes for each image. The data are available on request in the CVC-Colon repository (http://www.cvc.uab.es/CVC-Colon/index.php/databases/, accessed on 18 December 2022).
- Endoscopy Disease Detection Challenge 2020 (EDD2020) [63]: The EDD2020 challenge released a dataset containing five different classes with masks and bounding boxes for each image and polyp instance. We extracted all images labeled as a polyp and stored the relevant bounding boxes into a custom JSON file for our task. These data contain 127 images, and the size of the images is 720 × 576 pixels. This dataset contains no healthy mucosa images. The data are available on request in the ENDOCV repository (https://endocv2022.grand-challenge.org/Data/, accessed on 18 December 2022).

Own Data Creation

Previously, we designed a framework that utilizes a two-step process involving a small expert annotation part and a large non-expert annotation part [64]. This shifts most of the workload from the expert to a non-expert while still maintaining proficient high-quality data. Both tasks are combined with artificial intelligence (AI) to enhance the annotation process efficiency further. Therefore, we used the software Fast Colonoscopy Annotation Tool (FastCat) to handle the entirety of this annotation process. This tool assists in the annotation process in endoscopic videos. The design of this tool lets us label coloscopic videos 20 times faster than traditional labeling. The annotation process is split between at least two people. At first, an expert reviews the video and annotates a few video frames to verify the object's annotations. In the second step, a non-expert has visual confirmation of the given object and can annotate all following and preceding images with AI assistance. To annotate individual frames, all frames of the video must be extracted. Relevant scenes can be pre-selected by an automated system, and this prevents the expert from reviewing the entire video every single time. After the expert has finished, relevant frames will be selected and passed on to an AI model. This allows the AI model to detect and mark the desired object on all following and preceding frames with an annotation. The non-expert can adjust and modify the AI predictions and export the results, which can then be used to train the AI model. Furthermore, the expert annotates the Paris classification [65], the size of the polyp, its location, the start and end frame of the polyp, and one box for the non-expert annotators.

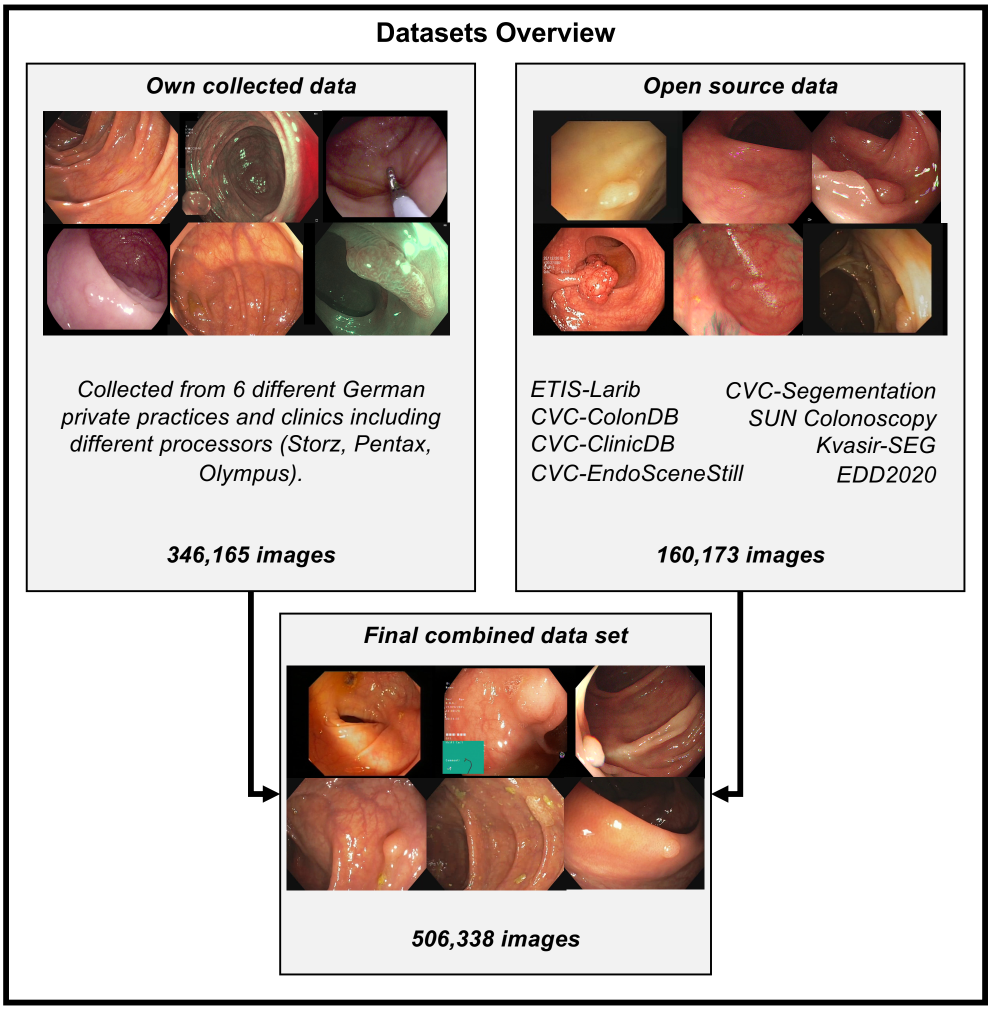


Figure 2. Training datasets overview. This figure illustrates all the data we combined and gathered for training the polyp-detection system. Open-source data are combined with our data collected from different German private practices to create one dataset with 506,338 images. Storz, Pentax, and Olympus are different endoscope manufacturing companies, and we collected the data using their endoscope processors. The different open source datasets have the following number of images: ETIS-Larib: 196, CVC-Segmentation: 56, SUN Colonoscopy: 157,882, Kvasir-SEG: 1000, EDD2020: 127, CVC-EndoSceneStill: consist of CVC-ColonDB: 300 and CVC-ClinicDB: 612. Overall this sums up to 160,173 open-source images.

We built a team of advanced gastroenterologists and medical assistants. We created a dataset of 506,338 images, including the open-source images listed above. Figure 2 shows an overview of the different datasets. Our dataset consists of 361 polyp sequences and

312 non-polyp sequences. The polyp sequence was selected in high quality as we were generally only annotating the first 1–3 s of the polyp's appearance, which is critical for detecting polyps in a real clinical scenario. We combined training material from six centers involving three different endoscope manufacturers, named Karl Storz GmbH und Co. KG (Storz), Ricoh Company Ltd. (Pentax), and Olympus K.K. (Olympus). Overall, 91% of the images are from Olympus, 5% are from Pentax, and 4% are from Storz processors. We create a dataset of 24 polyp sequences involving 12,161 images and 24 non-polyp sequences involving 10,695 images for the test data (EndoData). Therefore, the test data consist of an additional 22,856 images. We assured the independency of the test data as EndoData is created from a different clinic with different polyps and patients compared to the training data.

2.2. Current Challenges

There are still some challenges left. The most important of these can be divided into two categories: on the one hand, the hurdles to achieving the actual goal, real-time support of physicians, and on the other hand, problems arising from the acquisition or use of the datasets. The volume and quality of the data is a constant problem factor, and although there are various ways to deal with these problems, they still need to be solved.

2.2.1. Training and Testing Datasets

The biggest problem faced by most papers, e.g., [66] or [67], is the low availability of usable datasets. This refers not only to the number and size of datasets but also to the usually significant imbalance between the two classes *healthy frames* and *frames with polyps*. The need for more availability of pathological data is a common problem for medical deep learning applications. However, there are also various attempts to deal with it.

One widely used approach by Kang et al. [66] is *transfer learning*, which uses a model that has already been pre-trained on a non-medical dataset and is re-trained with a polyp dataset. The advantage is that sufficiently large non-medical datasets are readily available. With these, general problem-independent skills, such as edge detection, can already be learned well and only need to be fine-tuned to the specialized application.

Another method that almost all papers use is *data augmentation* of the training data. This involves slightly modifying the existing training data using various methods, thus increasing the available data and the system's robustness to the applied transformations. Examples of such transformations are rotation, mirroring, blur, and color adjustments [25].

An interesting approach by Guo et al. [68] is that the test data are also augmented at test time. More precisely, they are rotated and then passed to the model. To arrive at the result for the original image, all generated masks are rotated back accordingly and averaged. This makes the system more robust against rotation and leads to better accuracy.

Other ideas can be found in Thomaz et al. [69], where a CNN inserts polyps into healthy images to increase the available training data. In Qadir et al. [70], a framework is created to annotate data that can generate the rest of the segmentation masks with only a few ground truth masks.

2.2.2. Video Artifacts

A problem that still needs to be considered is the influence of video artifacts, such as reflections or blurring, on the detection rate of the methods. Attempts have been made to detect these and use artifact-specific methods to restore the frames; for example, the *Endoscopy artifact detection (EAD 2019) challenge* [71] was also conducted for this purpose.

The article by Soberanis-Mukul et al. [72] examines in detail the impact of artifacts on polyp detection rates. This allowed us to determine the artifact types with the most significant influence. With these, a multi-class model was developed to recognize the different artifact types and the polyps. Since the artifacts were known, regions affected by artifacts could be avoided being wrongly classified as polyps and polyps containing artifacts could be better classified.

2.2.3. Real-Time Application

To support a physician in a real-world scenario, models should be real-time capable, meaning they should achieve a processing speed of about 25 FPS, as colonoscopy videos usually run at 25–30 FPS, and newer systems may run with up to 50–60 FPS. Of course, some speed-up can be achieved by using appropriate hardware. However, concerning real-world use, speed measurement should be performed on hardware reasonable for a physician or hospital.

2.3. Data Preprocessing

To ensure high processing speed while maintaining high detection accuracy, we rescale the images to a size of 640 × 640 pixels. This rescaling allows the detection system to be efficient and high performing, maintaining a speed of 20 ms on an NVIDIA RTX 3080 GPU. In the clinical application subsection, we further explain the use of different GPUs and the GPU requirements for a system capable of real-time processing. We empirically tested different image sizes and found the best trade-off between speed and accuracy at a scale of 640 × 640 pixels. Additionally, we transfer the image and model to a half-precision binary floating-point (FP16). Typically, most machine learning models are in precision binary floating-point (FP32). With FP16, the model calculates faster but maintains high-performing results. We found no decrease in performance of the system by decreasing the network to FP16. The next step is image normalization. All image pixels are normalized in the following way: the min–max normalization function linearly scales each feature to the interval between 0 and 1. Rescaling to intervals 0 and 1 is completed by shifting the values of each feature so that the minimum value is 0. Then, a division by the new maximum value is performed (which gives the difference between the original maximum and minimum values).

The values are transformed element-wise using the following formula:

$$X_{sc} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

where X denotes the old pixel value, X_{sc} (X scaled) the new pixel value, $X_{\min}$ the minimum pixel value of the image, $X_{\max}$ the maximum pixel value of the image. After normalization, we apply data augmentation. In deep learning, augmenting image data means modifying the original image data by using various processes. We are applying the augmentation displayed in Figure 3. The most basic augmentation we apply is flip augmentation. This is well suited for polyps as the endoscope can easily rotate during colonoscopy. Here, the image is flipped horizontally, vertically, or both. We applied a probability of 0.3 for up and down flips and a vertical flipping probability of 0.5. We additionally apply rescaling to the image with a probability of 0.638. Rescaling creates polyps in different sizes, adding additional data to our dataset. The translation moves the image along the horizontal axis. Furthermore, we applied a low probability of 0.1 to rotate the image with a randomly created degree. For example, 20-degree rotation clockwise. As the last augmentation, we apply mosaic data augmentation. Mosaic data augmentation mixes up to four images into one image. In this implementation images can not overlap. Thereby, the image is rescaled, causing the images to appear in a different context. Mosaic data augmentation is applied with a probability of 0.944. These data augmentations are only applied to the training data. All of the hyperparameters for these augmentations are chosen by the parameter optimization of the genetic algorithm which is further illustrated in the hyperparameter optimization subsection below. All of the augmentations are combined to create new training images.

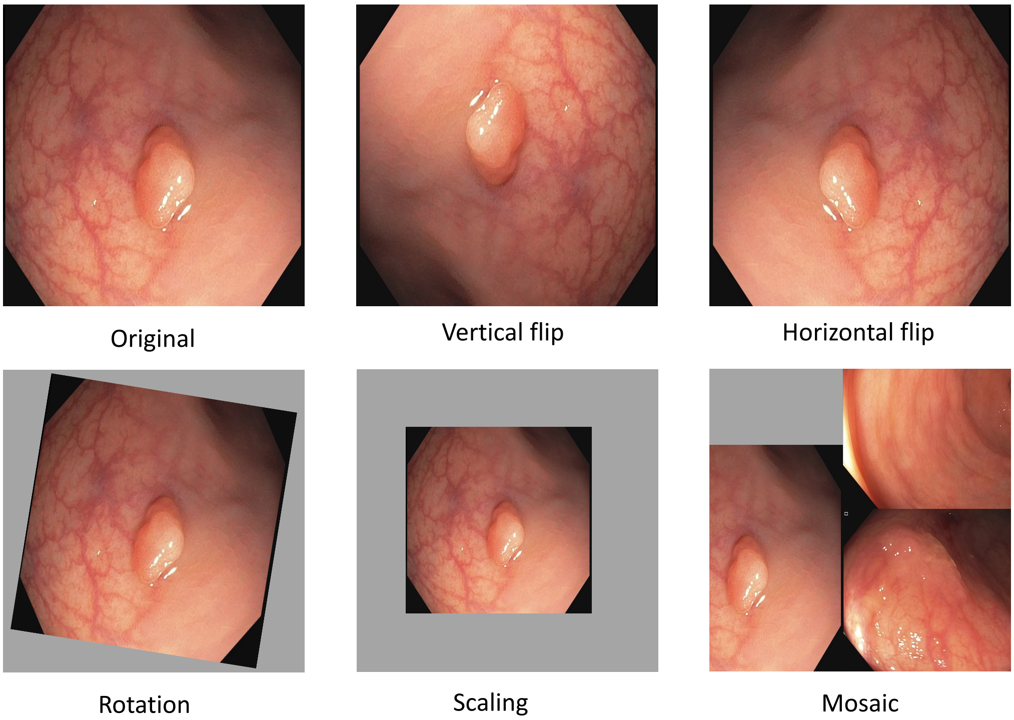


Figure 3. Data augmentation for polyp detection. This figure shows the isolated augmentation we perform to create new training samples. All of these are executed together with a certain probability in our implementation.

2.4. Polyp-Detection System

As illustrated in Figure 4, the polyp-detection system starts with an input of a polyp image sequence. A polyp image sequence consists of a stream of single images extracted from a grabber of the endoscope in real-time. t states the currently processed frame. $t-1$ denotes the frame before t, $t-2$ the frame before $t-1$, etc. The parameter ws denotes our new window size, which we introduce to apply real-time robust and efficient post-processing (RT-REPP). The polyp image sequence is now passed to the polyp-detection system. The polyp-detection system consists of two parts: the CNN detection architecture, here YOLOv5, and the post-processing, here real-time REPP (RT-REPP). The trained YOLOv5 model is now predicting boxes and passing those boxes to RT-REPP. RT-REPP consists of three main steps: first, boxes are linked across time steps, i.e., frames. This step is linking boxes according to the linking score. Details on the linking score are displayed in a subsection below. Second, unmatched boxes or boxes which do not meet specific linking and prediction thresholds are discarded through the system. Third, the boxes are adjusted using the predicted boxes from past detections. Finally, the filtered detections are calculated and displayed on the screen.

2.5. YOLOv5

In an end-to-end differentiable network, the YOLOv5 (https://github.com/ultralytics/yolov5, accessed on 18 December 2022) model was the first object detector to connect the technique of predicting bounding boxes with class labels. The YOLOv5 network consists of three main pieces. The neck is a construction of multiple layers that mix and combine image features to pass the prediction forward. The head takes the features from the neck and tries to predict boxes and classes. They use a CNN that aggregates and forms image features at different granularities for the backbone. To create the bounding boxes, YOLOv5 predicts them as deviations from several anchor box dimensions. In Figure 5, we illustrate the YOLOv5 architecture. The objective function of YOLOv5 is defined by minimizing the sum of three losses box-loss, cls-loss, and obj-loss. The box-loss measures how accurately the predicted BBs are drawn around the polyp. The cls-loss measures the correctness of the classification of each predicted bounding box. In our case, it is just one class (polyp). The objectiveness loss (obj-loss) penalizes the model for detecting the wrong bounding box.

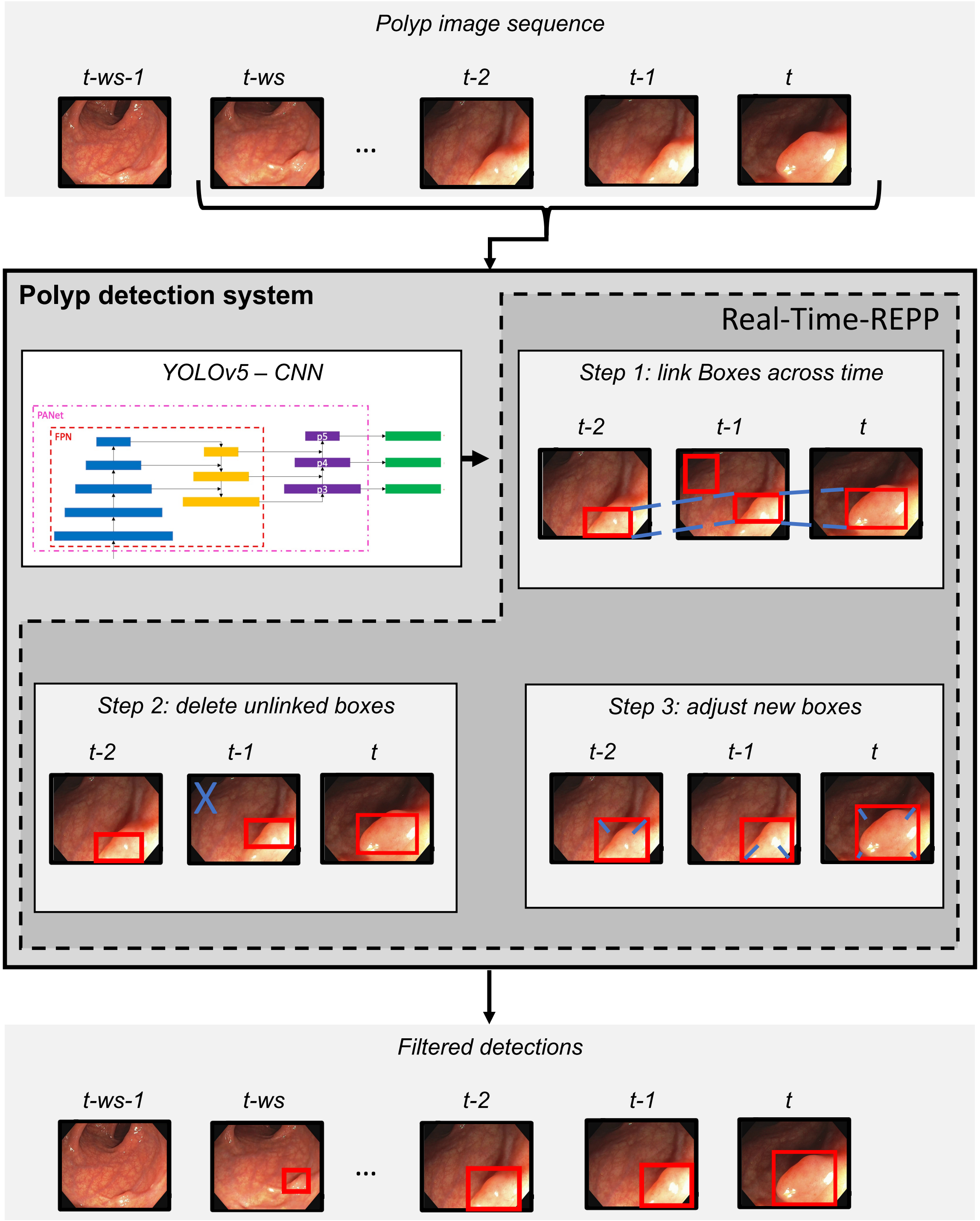


Figure 4. Overview of the polyp-detection system. This figure shows all the steps of the whole polyp-detection system. The start is an input of a polyp sequence ending with the last frame from the endoscope (t). From this sequence, ws frames are extracted and given to CNN architecture. Then detections are performed with YOLOv5, and the predicted boxes are post-processed by RT-REPP. Afterward, final filtered detections are calculated.

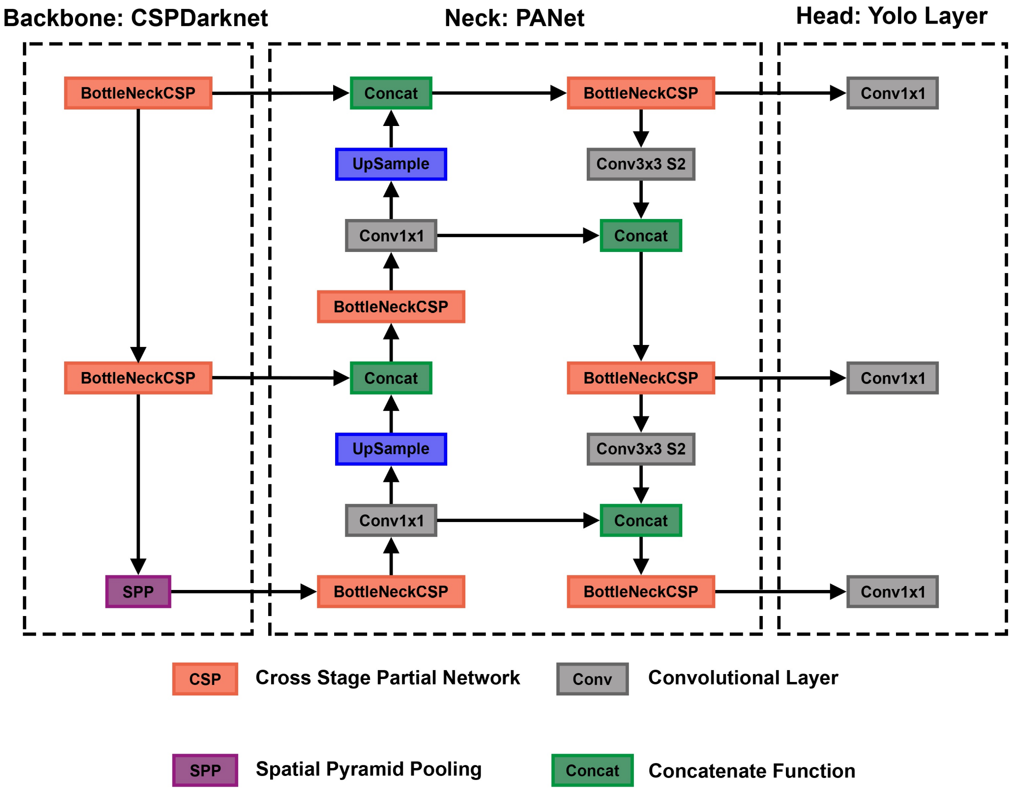


Figure 5. Detailed overview of the YOLOv5 architecture. This overview shows the whole architecture of YOLOv5. The starting point is the backbone CSPDarknet, the main feature extractor (the image is input for the BottleNeckCSP). These extracted features are then given to the PANet neck structure at three stages. Finally, in the head, three outputs are computed. These three outputs are specially designed for small, medium, and large objects and already contain the bounding box predictions. This figure is adapted from Xu et al. [73].

2.5.1. CSP as a Backbone

The cross stage partial network (CSPNet) model is based on DenseNet, which was created to connect layers in CNNs to build the backbone for the YOLOv5 model [74]. YOLOv5 created CSPDarknet by incorporating a CSPNet into Darknet. The most significant change of CSPDarknet is that the DenseNet has been reworked to divide the base layer's feature map by cloning it and sending one copy via the dense block, while sending the other directly to the next stage. Therefore, the CSPDarknet solves the problem of vanishing gradients in large-scale backbones. This is accomplished by splitting the base layer's feature map into two sections and combining them using a suggested cross-stage hierarchy. The fundamental idea is to separate the gradient flow to propagate over several network pathways. It was demonstrated by varying concatenation and transition phases that the propagated gradient information could have a considerable correlation difference. In addition, CSPNet may significantly minimize the amount of processing required and enhance inference speed and accuracy. CSPDarknet uses two BottleNeckCSPs and one spatial pyramid pooling (SPP) shown in Figure 5. SPP is a pooling layer that removes the network's fixed size limitation, allowing a CNN to operate with changing input sizes. It aggregates the features and provides fixed-length outputs that are then sent to the next layer or classifier. This works by pooling the results of each spatial bin (like max-pooling). The SSP produces kM-dimensional vectors, with M being the number of bins and k being the number of filters in the last convolutional layer. Therefore, the output is a fixed-dimensional vector. We chose CSP as a backbone for our models using VGG-16 or ResNet50 yields worse results on our validation data than the CSP backbone. Nevertheless, VGG-16 or ResNet50 could also be used as a backbone for this network, as those are valid options for polyp-feature extraction also shown in Tajbakhsh et al. [75] and Sharma et al. [76]. Still, we had the best results using the CSP backbone.

2.5.2. PANet as a Neck

The YOLOv5 architecture uses a path aggregation network (PANet) as its neck to improve the information flow [77]. Figure 6 illustrates the PANet and its connections to the architecture in more detail. PANet uses a novel feature pyramid network (FPN) topology with an improved bottom–up approach to improving low-level feature propagation. In the present architecture, the path starts with the output of the SPP from the backbone, which is passed to a CSP. This output is sent into a convolutional layer and is then upsampled. The result is then concatenated with the output from the second CSP in the backbone through a lateral connection and passed through the same combination again, which is then concatenated with the output from the first CSP of the backbone. Simultaneously, adaptive feature pooling is employed to restore broken information paths between each proposal and all feature levels. It is a fundamental component aggregating features from all feature levels for each proposal, preventing outcomes from being allocated randomly. Furthermore, PANet uses fully-connected fusion. These augments mask prediction with small fully-connected layers, which have complementary features to the fully-connected network (FCN) initially utilized by Mask R-CNN, to capture multiple perspectives on each proposal. Information diversity increases and higher-quality masks are generated by combining predictions from these two perspectives. Both object detection and instance segmentation share the first two components, resulting in a much-improved performance for both tasks.

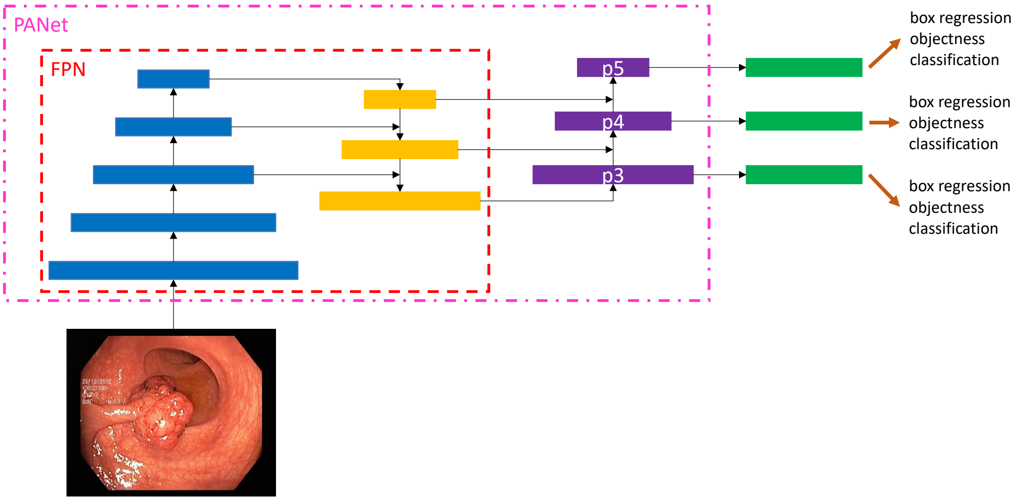


Figure 6. Overview of the PANet of YOLOv5. This overview shows a more detailed view of the PANet structure in YOLOv5. The starting point is a polyp input image. The FPN feature pyramid architecture is illustrated in interaction with the PANet. Finally, three outputs are given. These three outputs are specially designed for small (p5), medium (p4), and large (p3) objects.

The following steps are used for the adaptive feature pooling. First, the authors map each suggestion to distinct feature levels. Next, a function is utilized to pool feature grids from each level, following Mask R-CNN. The feature grids from different levels are fused using a fusion operation (element-wise max or sum). To integrate features into the network, pooled feature grids are passed through one parameter layer individually in the following sub-networks, followed by the fusion operation. For example, the box branch in FPN contains two fully-connected levels. Between the first and second convolutional layers, the two levels fuse together. For further prediction, such as classification, box regression, and mask prediction, the fused feature grid is utilized as the feature grid of each proposal.

The primary route is a tiny FCN with four convolutional layers in a row and one deconvolutional layer. Each convolutional layer has 256 × 3 × 3 filters, whereas the deconvolutional layer upsamples by two. Like Mask R-CNN, it predicts a binary pixel-wise mask for each class individually to decouple segmentation and classification. A short path from layer conv3 to a fully-connected layer is also created. The network is used with half-precision, cutting the computational cost by halving. A fully-connected layer is employed to forecast a class-agnostic foreground/background mask. It is efficient and allows the fully-connected layer's parameters to be trained with more data, resulting in improved generality. They employ a mask size of 28 × 28 such that the fully-connected layer generates a 784 × 1 × 1 vector. The mask predicted by the FCN is reshaped to the same spatial size as this vector. The final mask prediction is obtained by combining the masks of each class from the FCN with the foreground/background prediction from YOLOv5. Compressing the concealed spatial feature map into a short feature vector, which loses spatial information, is avoided using just one YOLOv5 layer for final prediction instead of several. Finally, the last YOLOv5 layer creates three different feature maps to provide multi-scale prediction, allowing the model to handle small, medium, and large objects.

Hyperparameter Optimization

We use a genetic algorithm to find optimal hyperparameters [78]. The genetic algorithm starts by using YOLOv5 default set of hyperparameters, training the entire YOLOv5 model until 15 epochs, then calculating the F1-score. After that, the hyperparameters are mutated and training is restarted. If the calculated F1-score is higher than the scores in the past, the best score and its hyperparameters are saved. After iterating over 10,000 genetic optimizations, we found our final parameters and retrained the algorithm for 54 epochs with the corresponding hyperparameters. At this point, the algorithm is stopped through early stopping.

Model Selection

We tested different architectures to select the best polyp detection model. We used a 5-fold cross-validation on our training data to determine the final model. We used 80% of the data for training and 20% of the data for validation. The cross-validation results are shown in Table 3. The best results are achieved in precision, recall, F1, and mAP using the YOLOv5 model. Still, the model keeps real-time capability while having an average number of parameters compared to the other models.

Table 3. Results of the 5-fold cross-validation for selecting the final model deep learning model. Values displayed in bold font indicate the highest or most optimal results. The abbreviation "adv." is an acronym for the term "advanced".

	Precision	Recall	F1	mAP	Speed	Parameter
Faster R-CNN [32]	81.79	85.58	83.64	79.43	15	91 M
YOLOv3 [35]	80.45	82.46	81.44	81.92	41	65 M
YOLOv4 [79]	83.04	83.68	82.36	83.54	**47**	81 M
YOLOv5 (adv.)	**88.02**	**89.38**	**88.70**	**86.44**	43	79 M
SSD [36]	75.52	76.19	75.85	78.69	30	**64 M**

2.6. Robust and Efficient Post-Processing (REPP) and Real-Time REPP (RT-REPP)

Image object detectors process each frame of a video individually. Each frame of an incoming stream of frames is viewed independently of the previous and subsequent frames. As a result, information is lost and the performance of such detectors can significantly differ between images and videos. Moreover, video data confronts the object detector with unique challenges, such as blur, occlusion, or rare object poses. To improve the results of the object detector for video data, the post-processing method REPP [13] relates detections to other detections among consecutive frames. Thus, the temporal dimension of a video is included. REPP links detections across consecutive frames by evaluating their similarities and refining their classification and location. This helps to suppress and minimize FP detections. The algorithm can be divided into three modules: (1) object detection, (2) detection linking, and (3) detection refinement. Figure 7 shows an overview of the REPP modules.

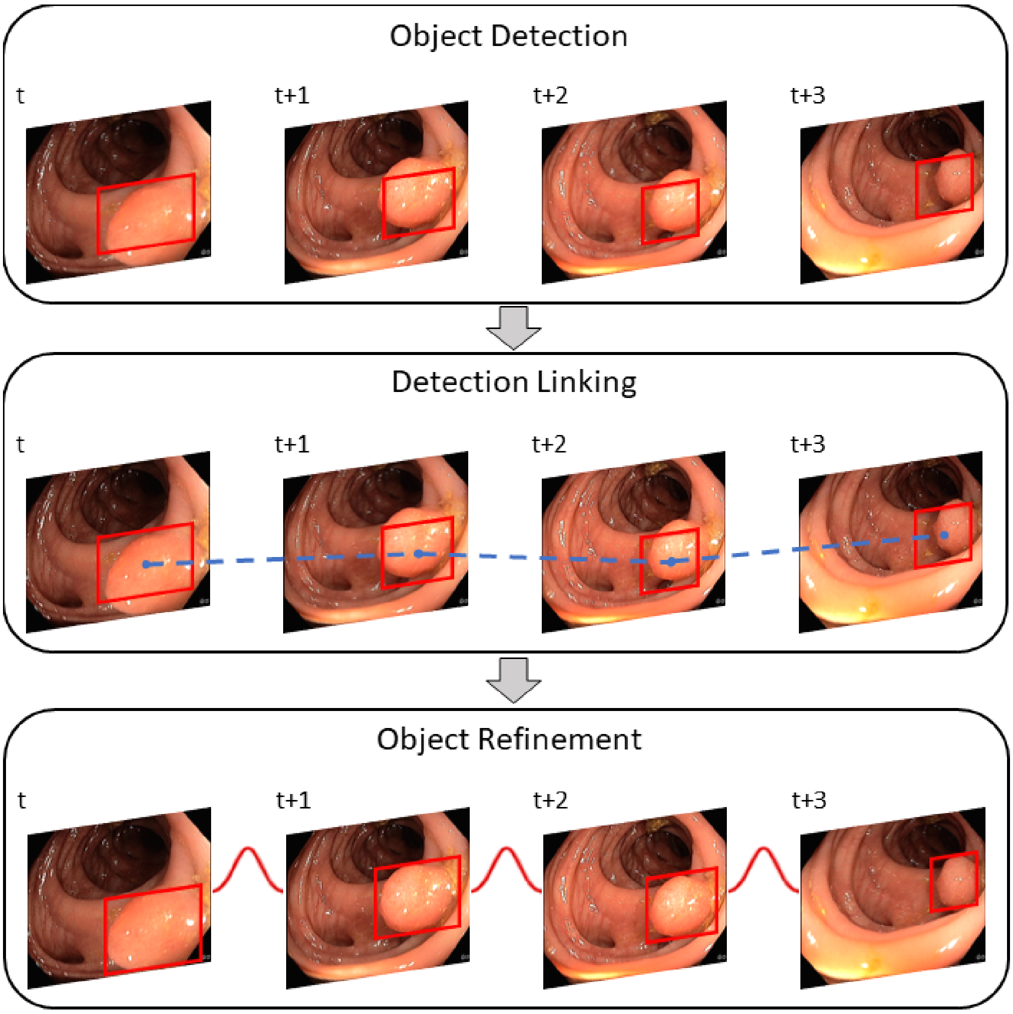


Figure 7. The REPP modules used for video object detection post-processing. The object detector predicts a polyp for a sequence of frames and links all bounding boxes across frames with the help of the defined similarity. Lastly, detections are refined to minimize FPs. This figure is adapted from Sabater et al. [13].

2.6.1. Object Detection

Object detection works on any object detector that provides bounding boxes and a class confidence score. For each frame t, the detector delivers a set of object detections. Each detection o_t^i is described by a bounding box (bb), semantic information and the appearance of the patch (small piece of an image). The bounding box is defined as $bb_t^i = \{x, y, w, h\}$, where x and y is the upper left corner, w the width, and h the height of the bounding box. Semantic information, such as the vector of class confidences, is defined as $cc_t^i \in \mathbb{R}^C$ with C for the number of classes and a L2-normalized embedding $app_t^i \in \mathbb{R}^{256}$ which represents the appearance of a patch.

2.6.2. Detections Linking

Linking detections along the video are created by a set of tubelets and continue as long as corresponding objects are found in the following frames. A similarity function is used to link two detections between two consecutive frames.

$$f_{\text{loc}} = \{\text{IoU}, d_{\text{centers}}\} \tag{1}$$

$$f_{\text{geo}} = \{\text{ratio}_w, \text{ratio}_h\} \tag{2}$$

$$f_{\text{app}} = d_{\text{app}} \tag{3}$$

$$f_{\text{sem}} = f^a_{\text{sem}} \cdot f^b_{\text{sem}} \tag{4}$$

where f_{loc} is the location which is specified through the Intersection over Union (IoU) and the relative euclidean distance between two bounding box center points (d_{center}). The IoU indicates the overlap between two boxes. The larger the overlap, the more likely it is that both boxes mark the same object.

In addition, the distance between the two center points is used. f_{geo} is the geometry of the bounding boxes, which is defined as the ratio of width (ratio_w) and height (ratio_h) between the two bounding boxes. This score is high if both boxes have a similar size. f_{app} is the appearance similarity in which the Euclidean distance between the appearance embeddings (d_{app}) are calculated. A more similar appearance results in a higher score. Lastly, f_{sem} is the dot product of the class confidence vectors cc^i_t. The greater the confidence vectors of both detections, the more likely it is that both boxes mark a polyp. Using these features, a link score (LS) is calculated between two detections.

$$LS(o^i_t, o^j_{t+1}) = f_{\text{sem}}\mathbf{X}(f_{\text{loc}}, f_{\text{geo}}, f_{\text{app}}) \tag{5}$$

Thereby, **X** is a function for logistic regression trained so that it can differentiate if two detections belong to the same object instance. In the following, the linking process is algorithmically explained.

Algorithm 1 shows a general description to obtain pairs of frames. The algorithm uses a list of predictions, processes each frame, calculates the distance between objects in both frames, and saves the value in a distance matrix. The objects with the lowest distance are then considered a pair, and a list of pairs is returned.

Algorithm 1 Get a list of pairs of frames that are linked across frames

```
function GETPAIRS(predictions)
    for $index \leftarrow 0$ to count of frames do
        predsFrame1 ← predictions[$index$]    ▷ Get frame predictions from current index
        predsFrame2 ← predictions[$index + 1$]    ▷ Predictions of next frame
        framePairs ← empty list
        if length(predsFrame1) $\neq 0$ and length(predsFrame2) $\neq 0$ then
            distances ← 2D-Array with 0 for each cell
            for $i \leftarrow 0$ to length(predsFrame1) do
                for $j \leftarrow 0$ to length(predsFrame2) do
                    distances[$i$][$j$] ← LOGREG(predsFrame1[$i$], predsFrame2[$j$])
                end for
            end for
            framePairs ← SOLVEDISTANCES(distances)
        end if
        pairs.append(framePairs)
    end for
    return pairs
end function
```

Next, tubelets are created (Algorithm 2) from a list of linked pairs. Tubelets link all bounding boxes that identify the same object across a series of frames.

Algorithm 2 Tubelets creation from list of linked pairs

```
function GETTUBELETS(predictions, pairs)
    tubelet ← empty list
    for each frame do
        for each pair in following frames do
            if frame has no pair then
                start new tubelet
            end if
            if frame has pairs then
                append box from pair to tubelet
            end if
        end for
    end for
    return tubelets
end function
```

2.6.3. Object Refinement

The use of tubelets improves the classification and location. The first step of tubelet creation is recalculating the detection classification scores. Therefore, all class confidence vectors are averaged and assigned to each detection within the tubelet (see Algorithm 3). This helps correct mislabeled detections and disambiguate those with low confidence.

Algorithm 3 Rescore tubelets

```
function RESCORETUBELETS(tubelets)
    for each $t \in$ tubelets do
        $s_{avg} = \frac{1}{|t|} \cdot \sum_{p \in t} s_p$        ▷ Average score $s$ of predictions $p$ of tubelets
        $\forall p \in t : s_p = s_{avg}$                          ▷ Assign average to all prediction scores
    end for
    return tubelets
end function
```

The next step is to improve the detection positions. Each coordinate of a linked object is treated as a noisy time series. Smoothing is used to alleviate the noise with the help of a one-dimensional Gaussian filter convolution along with each time series. The smoothed series are then used as the set of coordinates of the object in the tubelet (Algorithm 4, line 7).

Algorithm 4 REPP

```
 1: function REPP(objectDetectionPredictions)        ▷ Gets all predictions from detection network
 2:     videoPredictions ← FILTERPREDICTIONS(objectDetectionPredictions)
 3:     pairs ← GETPAIRS(videoPredictions)
 4:     tubelets ← GETTUBELETS(videoPredictions, pairs)
 5:     tubelets ← RESCORETUBELETS(tubelets)
 6:     if recoordinate == True then                  ▷ Tubelets reccordination is optional
 7:         tubelets ← RECOORDINATETUBELETS(tubelets)
 8:     end if
 9:     predictions ← TUBELETSTOPREDICTIONS(tubelets)    ▷ Convert to specific format
10:     return predictions
11: end function
```

The final REPP algorithm (Algorithm 4) is a combination of all aforementioned algorithms executed in order. First, filter detection predictions (line 2), then obtain all pairs (line 3) and afterward compute the tublets out of the pairs (line 4). Then rescore detections within tubelets (line 5) and recoordinate them for improved prediction results (line 7). Lastly, filter all predictions that do not reach a certain threshold and convert them to a specific prediction result format, such as the COCO format (line 9).

Since REPP is a post-processing method that only works on finished videos, REPP includes past and future frames. We modified the code for real-time application to only include past frames, calling the algorithm RT-REPP. To compare REPP and RT-REPP in our real-time application, we included a buffer of pre-defined length to run the original REPP. The size of the buffer is adjustable to fit the available resources. The greater the length, the longer it takes to execute REPP. Before REPP is executed, the buffer must be completed, which causes a size-dependent delay at the beginning of each video. To overcome this delay, REPP is run from the start frame and executed for every new frame until the buffer is completed. The completed buffer is passed, and the oldest frame is deleted as a new frame is inserted. Since the delay times for our application are relatively short, we accept this short delay. See Algorithm 5 to understand the basic workflow of RT-REPP. We define a window size ws, which determines the window length. A buffer size ws of 300 is sufficient for a real-time stream of 25 FPS. A ws of more than 300 does not improve the accuracy significantly.

Algorithm 5 RT-REPP

```
function RT-REPP(framePrediction)
    if buffer is full then
        delete oldest frame
    end if
    add prediction to buffer
    run REPP on buffer
end function
```

To implement RT-REPP in our real-time system, we combined C++ and python. We used the lightweight header-only library pybind11, which allows us to use C++ types and methods in python and vice versa. To our knowledge, REPP and RT-REPP have not been used before in the domain of polyp detection. In Figure 8, the workflow of real-time REPP is illustrated.

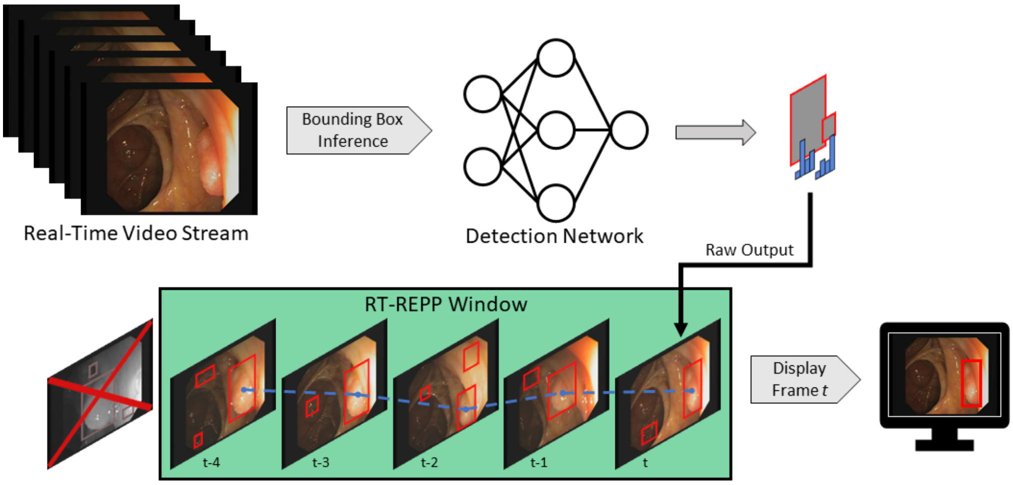


Figure 8. Real-time REPP. It obtains a stream of video frames, where each frame is forwarded into a detection network. The result of the current frame is stored into the buffer (green) and REPP is executed afterward. The improved result are then displayed.

We tested different linking-score thresholds and window sizes to choose the hyperparameters of RT-REPP. The boxes scoring below the linking-score threshold are removed from the final detection results. As described early, we set the window size to 300. We tested different linking-score thresholds. Our results determined a score of 0.2 to be the most effective.

2.7. Clinical Application

To develop a system for clinical trials, it is mandatory to understand the current settings of examination rooms. Endoscopic and other medical equipment are complex devices with intricate setups. Therefore, this is only a brief overview of these highly sophisticated components.

Figure 9 shows an example of medical devices used during endoscopic interventions. Figure 9a presents an endoscope composed of a flexible tube, controlled and operated by physicians during examination through several control buttons and physical force. A fish-eye camera is on the tip of this tube, combined with a light source to capture an RGB video stream. The endoscopy tower contains the entire endoscopic equipment and, most importantly, the camera's light source and an endoscopy processor (Figure 9b). The endoscopic processor captures the camera stream and processes it into a regular video signal. This signal can be displayed on a monitor, as shown in Figure 9c. These components provide physicians with real-time visual feedback during endoscopic interventions. Based on the given setting, we developed a prototype for clinical application. Instead of directly connecting the endoscopy processor to the monitor, our system stands between the processor and monitor, processing all frames before forwarding them to the monitor.

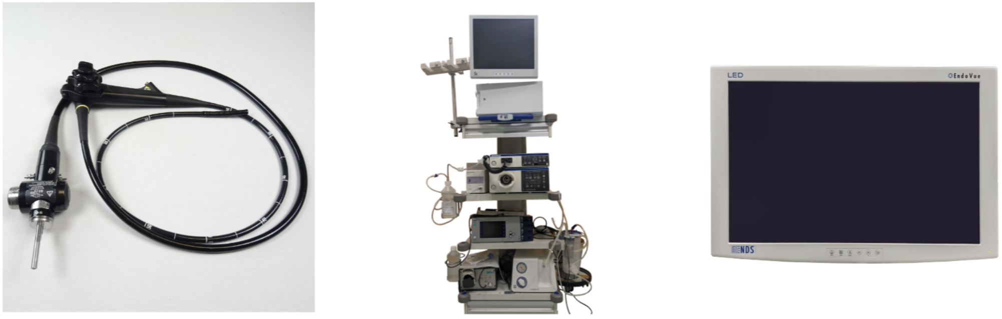

(a) Endoscope (b) Endoscopy Tower (c) Monitor

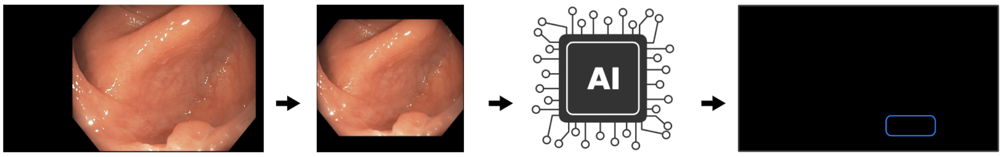
Figure 9. This figure illustrates the setting for the examination room.

Table 4 shows the main hardware components of our system, which allows for real-time image processing. However, these are just as important as a suitable software. In order to provide physicians with the best possible user experience, all incoming frames must be displayed as fast as possible to minimize latency. To this end, image capturing, displaying, and processing are running in separate threads. The first thread uses the Blackmagic SDK to capture frames, which depends on the frame rate. For instance, Olympus CV-190 provides 50 FPS, receiving a frame every 20 ms. Therefore, it is essential to distribute the additional workload on other threads. If only one thread is used, incoming frames are buffered, resulting in an overall delay across all related threads. Considering this, thread one only captures and transforms incoming data to an OpenCV matrix, passing it to subscribing pipelines.

One receiver is the AI pipeline, shown in Figure 10. In this thread, all incoming frames are cloned (Figure 10a) to ensure that all operations on those image matrices do not interfere with other processes. The clone shown in (Figure 10b) is preprocessed. Here, the frame matrices are transformed to fit the AI network. First of all, the black borders of the images are cropped. In Figure 10a to Figure 10b this is illustrated. The resulting 640 × 640 matrix is transformed from BGR to RGB and uploaded to GPU memory. Here, the matrix is processed through YOLOv5 (Figure 10c). Based on the input, it results in relative coordinates, classes, and scores for every detection. The last step is a transformation resulting in a vector of quadruples, containing xy-coordinates, width, and height of bounding boxes to suit the original matrix (Figure 10d). Under consideration of thresholds, detections with low confidence are removed, while the remaining detections are transformed and forwarded to the display pipeline.

Table 4. Prototype components.

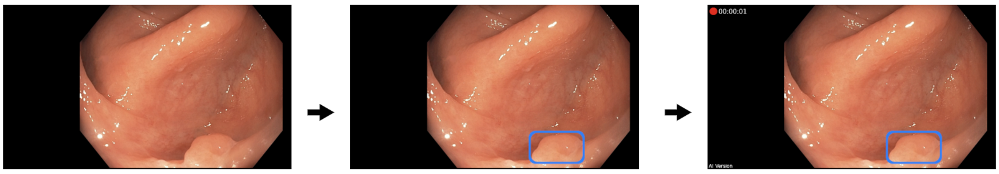
Component	Type	Info
CPU	AMD Ryzen 7 3800X	8 Cores, 3.9 GHz
GPU	MSI GeForce RTX 3080 Ti	12 GB GDDR6X
RAM	G.Skill RipJaws V DDR4-3200	2 × 8 GB
Disk	Samsung SSD 970 EVO Plus	500 GB
Mainboard	B550 Vision D	-
Frame Grabber	DeckLink Mini Recorder 4 K	-



(a) Clone (b) Preprocess (c) Inference (d) Postprocess

Figure 10. The AI pipeline. This figure depicts the AI pipeline used to apply the created polyp-detection system in a clinical environment.

The independent display pipeline thread is designed to display captured frame matrices as fast as possible. Like the AI pipeline, matrices are cloned at the beginning, shown in Figure 11. Consequently, no processing is applied on the original matrices; therefore, other pipelines remain unaffected. Then, based on the most recent detections of the AI, new boxes are drawn and old ones removed. The boxes remain on the screen until a new cycle of the AI pipeline has finished. Additionally, a few extra UI elements, such as a timer, indicating that the AI is running before frames are forwarded and displayed. This design, as mentioned earlier, decouples the AI and display pipeline. Hence, a slower AI does not directly result in higher display latency. Nevertheless, the performance of the AI pipeline remains an essential factor. Faster executions lead to more inferences and, therefore, more precise boxes, given that the displayed frame is closer to the AI pipeline frame.



(a) Clone (b) Draw Boxes (c) Draw UI Elements

Figure 11. The display pipeline. This figure depicts the display pipeline used to display the final detection results to the gastroenterologist.

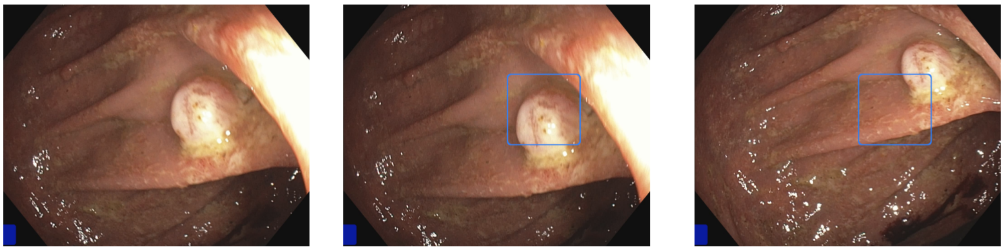
The prototype was tested on two different GPUs to show the performance differences during clinical application. The prototype components are listed in Table 4. A second computer streamed a colonoscopy video instead of an endoscopy processor, just like an endoscopy processor does. Meanwhile, the prototype captured the signal, as mentioned earlier. This ensures identical, reproducible conditions and guarantees occurring polyps during the experiment. The prototype is not able to distinguish this method from a live endoscopy. The streamed examination video is presented in 1920 pixels and 50 fps, equivalent to streams of Olympus CV-190. Our test used the MSI GeForce RTX 3080 Ti, an up-to-date high-end GPU released on 3 June 2021. The NVIDIA Geforce GTX 1050 Ti, a low-budget GPU two generations ago, was used for a second test run. This GPU was released on 27 May 2016. All other hardware components and software parts were constant throughout testing.

In the setting of Table 5 5000 frames are considered. Out of those 5000 frames, the RTX 3080 Ti executed the AI pipeline 2996 times. At the same time, the GTX 1050 Ti made 313 executions. This is based on the AI's average execution time (AI pipeline average execution time) 19.5 ms and 306.7 ms, respectively. During the usage of RTX 3080 Ti, there was a 15-fold performance gain. The AI pipeline was applied on every 1.7th frame on this GPU, while only every 16th frame was evaluated through the inferior GPU. Considering those results and a synchronized display pipeline, it takes two frames until bounding boxes are displayed. Furthermore, those two boxes remain displayed for two more frames until they are updated again, resulting in a total display time of four frames (80 ms) for the RTX 3080 Ti. In comparison, the GTX 1050 Ti accumulates 32 frames (640 ms), while 16 frames (320 ms) are needed to generate the first bounding box. This does not illustrate the worst or the best-case scenario.

Table 5. A 5000 frames system test. This table shows the speed of the detection system of two GPUs. Considering an image input with a speed of 50 FPS.

GPU	AI Exe. Count	AI Avg. Exe. Time	AI Evaluation Rate
RTX 3080 Ti	2996	19.5 ms	29.4 FPS
GTX 1050 Ti	313	306.7 ms	3.1 FPS

An example was created to show the delay in the appearance of a bounding box. Figure 12a shows a frame forwarded to the AI pipeline. Since the RTX 3080 Ti needs an average of 1.7 frames, bounding boxes appear in frame two. This is illustrated in Figure 12b. While the camera moves, frame two is slightly shifted to the bottom, but the polyp is still mainly boxed. The GTX 1050 Ti takes an average of 16 frames, shown in Figure 12c. The polyp is mainly outside the bounding box. A box might appear based on the speed at which the endoscope is moved, even if a polyp is no longer displayed. This is highly unlikely for the RTX 3080 Ti, which in the best case, shows a bounding box on the consecutive frame. This delay must be considered when using slower GPUs but can be neglected if the endoscope's withdrawal motion is made slowly.



(a) Frame 0 (b) Box on frame 2 (c) Box on frame 16

Figure 12. Detection shift through latency.

The test shown in Table 5 has been performed on an actual prototype. Therefore, the software has not been altered. In addition, a video was recorded simultaneously, and this is done for quality assurance and to retrieve additional test data. The recording pipeline is independent, but the GPU is used for H.264 video encoding, which causes an additional load and can affect the performance of the AI pipeline. In general, our prototype is not designed for a specific GPU, all NVIDIA GPUs with CUDA compatibility over the last five years can be used, but it will affect the user experience. In an actual examination, prototypes have been used with a MSI GeForce RTX 2080 SUPER Ventus XS OC with no significant change in user experience.

3. Results

For the evaluation, we use two datasets. The CVC-VideoClinicDB dataset is the first benchmark dataset for polyp detection in videos [59]. The previous benchmark datasets, e.g., ETIS-Larib and CVC-ColonDB, only allow a comparison based on still images. The CVC-VideoClinicDB dataset has the ability to evaluate models on video data, which is a more realistic scenario, as real polyp detection outputs from endoscopies are not images but a stream of images provided in real-time. As our architecture explained in methods only applies to videos or a stream of images, we chose the CVC-VideoClinicDB dataset as our main evaluation dataset. The second dataset is our own test dataset called EndoData. In the CVC-VideoClinicDB dataset, the polyp sequence begins with the polyp already in view in the first frame. Since our dataset contains the entire footage, the polyps appear further into the image sequence. Therefore, the EndoData dataset emulates the clinical practice more closely, which makes the evaluation even more realistic in the application. We can additionally calculate a metric measuring the time taken to detect a polyp. We published the code for the application and evaluation of our system on our webpage (https://fex.ukw.de/public/download-shares/d8NVHA2noCiv7hXffGPDEaRfjG4vf0Tg, accessed on 18 December 2022).

The F1-score evaluates the model's quality. The F1-score describes the harmonic mean of precision and recall as shown in following equations:

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{Recall} = \frac{TP}{TP + FN}$$

$$F_1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 * TP}{2 * TP + FP + FN}$$

We count an annotation as TP if the boxes of our prediction and the boxes from the CVC-VideoClinicDB dataset ground truth overlap at least 50%. Additionally, we choose the mAP, which is a standard metric in object detection [80]. The mAP is calculated by the integral of the area under the precision-recall curve. All predicted boxes are first ranked by their confidence value given by the polyp-detection system. Then we compute precision and recall for different thresholds of these confidence values. When reducing the confidence threshold, recall increases and precision decreases. This results in a precision–recall curve. Finally, the area under the precision–recall curve is measured. This measurement is called the mAP. Furthermore, our approach introduces new parameters to the polyp-detection system. One of the parameters is the width of the detection window ws.

We created the following evaluation on our dataset (EndoData). Our evaluation considers two baseline models: the YOLOv5 (base) model and the Faster R-CNN baseline. The YOLOv5 (base) model is the basic YOLOv5 model trained on the EndoData dataset without any hyperparameter optimization, data augmentation, post-processing, or other changes for polyp detection. The second baseline model is a Faster R-CNN with a ResNet-101 backbone. This involves training a Faster RCNN with default parameters using the Detectron2 framework [81].

Furthermore, we show three different stages of our polyp-detection system. First, YOLOv5 advanced (adv.), which is training the YOLOv5 model but with all our in section methods explained features and optimization to specialize it for the polyp detection task. Second, REPP is a trained YOLOv5 (adv.) model, including the REPP post-processing. This is not applicable in real-time, as the REPP algorithm only works on recorded videos. Afterward, we introduce the RT-REPP. The RT-REPP is our version of REPP, which works in real-time. Our polyp-detection system ENDOMIND-Advanced is in the following evaluation, referred to as RT-REPP. All models are trained on our training data using four Quadro RTX 8000 NVIDIA graphics cards, and the test application is made on an NVIDIA GeForce RTX 3080. The results of these models are shown in detail in Tables 6–12.

Table 6. Evaluation CVC-VideoClinicDB dataset. This table compares six different polyp detection approaches on the benchmarking data CVC-VideoClinicDB. The first two models are baseline models, and the third is the best model of the current literature. The last three models are different stages of our polyp-detection system. Precision, Recall, F1, and mAP are given in %, and the speed is given in FPS. Values displayed in bold font indicate the highest or most optimal results. The abbreviation "adv." is an acronym for the term "advanced".

	Precision	Recall	F1	mAP	Speed	RT Capable
YOLOv5 (base)	92.15	69.98	79.55	73.21	44	yes
Faster R-CNN	93.84	74.79	83.24	79.78	15	no
Qadir et al. [55]	87.51	81.58	84.44	-	15	no
YOLOv5 (adv.)	98.53	76.44	86.09	77.99	**44**	yes
REPP	**99.71**	**87.05**	**92.95**	**86.98**	42	no
RT-REPP	99.06	82.86	90.24	83.15	43	yes

Table 7. Detailed detection approaches on the benchmarking data CVC-VideoClinicDB. The first two models are baseline models, and the last three are different stages of our polyp-detection system. F1, and mAP are given in %. The abbreviation "adv." is an acronym for the term "advanced".

Video	YOLOv5 (Base)		F-RCNN		YOLOv5 (Adv.)		REPP		RT-REPP	
	mAP	F1	mAP	F1	mAP	F1	mAP	F1	mAP	F1
1	78.22	87.41	92.56	88.14	85.17	91.47	94.56	97.44	89.38	94.18
2	87.35	91.87	89.48	89.19	94.62	96.91	97.48	98.48	96.48	97.96
3	75.58	80.09	81.48	77.71	80.18	84.42	86.48	87.64	82.65	85.01
4	90.04	92.16	93.35	90.39	98.00	98.99	98.35	99.50	98.29	98.99
5	76.29	82.53	78.01	85.85	78.40	87.64	83.01	90.71	78.88	88.27
6	86.23	88.59	87.05	89.42	90.07	94.83	92.05	95.43	88.41	92.83
7	60.75	67.15	69.56	78.38	66.23	76.15	74.56	85.71	71.95	82.35
8	53.93	69.52	77.22	82.65	59.16	73.66	82.22	90.11	82.22	90.11
9	74.27	77.29	84.10	87.21	76.50	87.01	89.10	94.18	85.15	91.89
10	75.28	77.36	86.33	86.00	78.22	87.25	91.33	95.29	86.61	92.61
11	90.17	92.19	94.19	94.92	95.41	97.44	99.19	99.50	98.65	99.50
12	30.81	46.22	42.51	60.09	36.78	54.01	47.51	64.86	39.85	57.14
13	84.48	89.48	84.68	87.06	89.37	94.29	89.68	93.83	90.00	94.74
14	74.35	80.49	82.20	86.42	79.09	87.88	87.20	93.05	82.20	90.11
15	48.88	62.62	52.51	66.56	52.18	69.04	57.51	73.15	55.65	71.79
16	89.45	92.97	93.63	90.32	94.54	97.44	98.63	99.50	98.36	98.99
17	52.25	64.61	56.29	68.15	57.77	72.59	61.29	75.78	49.80	65.75
Mean	73.21	79.55	79.78	83.24	77.99	86.09	86.98	92.95	83.15	90.24

3.1. CVC-VideoClinicDB Data Evaluation

To compare our polyp-detection system to the published research, we use the publicly available CVC-VideoClinicDB dataset. To our knowledge, the best-performing algorithm on the dataset was published by Qadir et al. [55]. Therefore, Qadir et al. [55] is included in the evaluation in Table 6. In Table 6, different baseline and advanced stage models are compared. All values are calculated according to the CVC-VideoClinicDB challenge norm. The CVC-VideoClinicDB challenge norm defines the same calculations used for calculation in the GIANA challenge 2017 [60]. Therefore, the means are calculated by summing all the results for every image and dividing the sum by all images in the dataset (micro mean). We use this micro mean structure throughout this paper. All presented means are micro averages over all images. We excluded video 18 from the CVC-VideoClinicDB dataset, because 77 of 381 images are labeled incorrectly.

For the F1-score, REPP has the highest F1 score. However, REPP is not applicable in real-time as it is calculated by combining past, present, and future predicted boxes. Therefore, REPP can only be used on recorded videos. We like to include it in the comparison to show the enhancements using the full algorithm. RT-REPP achieves the second-best F1-score and functions in real-time. Using RT-REPP vs. YOLOv5 (adv.) improves the results by a F1-score of 4.15%. The baseline models Faster R-CNN and YOLOv5 (base) achieve lower F1-scores.

Overall our results show that using our hyperparameter, data augmentation, and training setup increases the F1 and mAP by 6.05% and 4.78%. By leveraging our implementation, RT-REPP results improve further by 4.15% and 3.14%. REPP and RT-REPP cause a minimal speed reduction, resulting in roughly a 1 FPS speed reduction for RT-REPP and a 2 FPS reduction in REPP. Therefore, those algorithms can easily be added to neural networks without losing much processing time.

For the detailed evaluation, we computed the mAP and F1-score for each of the 17 videos of the CVC-VideoClinicDB dataset. REPP-RT detects most videos with a F1-score of over 90%. Only videos 3, 5, 7, 12, 15, 17 have a score lower than 90%. These videos also have inferior test results. Negative examples are video 12, with a score of 57.14%; video 17, with a score of 65.72%; and video 15, with a score of 71.79%. We analyze those videos in more detail in the discussion section. The YOLOv5 baseline model also has inferior results with a value of 46.22% and a detection value lower than 50%. Comparing our approach to Jha et al. [48], we achieve better results on the CVC-VideoClinicDB data. However, the Jha et al., model is also capable of polyp segmentation and the system's speed is faster (180 FPS).

3.2. EndoData Evaluation

Our own validation set (EndoData) allows us to detect polyps more precisely and accurately. Table 8 shows an overview of the videos in the dataset and Figure 13 shows examples of the dataset. The EndoData dataset records sequences as the polyp appears in the scene. Therefore, polyps are marked precisely with their first appearance. In comparison, the polyp sequence of the CVC-VideoClinicDB dataset might not start when the polyp is already detected. Those early seconds are crucial as the gastroenterologist has to identify and not miss the polyp during this time. If the polyp is not detected in the early sequence, it increases the risk of missing it. As we like to focus on this early detection, we introduce a second metric that can just be evaluated with a dataset like ours. This metric marks the seconds from first seeing the polyp to first detecting the polyp. We call it first detection time (FDT). Additionally, we compute the FPs and the false positive rate (FPR) per video (Tables 10 and 11).

Table 8. Details of the EndoData. This table shows the details of our own evaluation data (EndoData). Width and height state the size of the used frames.

Video	1	2	3	4	5	6	7	8	9	10
Frames	14,947	18,026	1960	1923	9277	14,362	347	4627	6639	766
Polyps	2	5	1	1	2	5	1	2	4	1
Width	1920	1920	1920	1920	1920	1920	1920	1920	1920	1920
Height	1080	1080	1080	1080	1080	1080	1080	1080	1080	1080

Table 9. Evaluation of EndoData. This table compares five different polyp detection approaches on our EndoData dataset. The first two models are baseline models. The last three models are different stages of our polyp-detection system. F1, and mAP are given in %. Values displayed in bold font indicate the highest or most optimal results. The abbreviation "adv." is an acronym for the term "advanced".

	Precision	Recall	F1	mAP	Speed	RT Capable
YOLOv5 (base)	78.39	80.54	79.45	77.09	44	yes
Faster R-CNN	81.85	86.20	83.97	81.74	15	no
YOLOv5 (adv.)	86.21	86.43	86.32	82.28	**44**	yes
REPP	**90.63**	**89.32**	**89.97**	**87.24**	42	no
RT-REPP	88.11	87.83	87.97	84.29	43	yes

Table 10. Time to first detect on our own dataset (EndoData). This table compares five different polyp detection approaches on EndoData with our new metric time to first detection (FDT). The first two models are baseline models, and the last three are different stages of our polyp-detection system. FDT is measured in seconds. FP denotes the number of FPs in the video. Values displayed in bold font indicate the highest or most optimal results. The abbreviation "adv." is an acronym for the term "advanced".

Video	YOLOv5 (Base)		F-RCNN		YOLOv5 (Adv.)		REPP		RT-REPP	
	FDT	FP	FDT	FP	FDT	FP	FDT	FP	FDT	FP
1	0.07	201	0.00	159	0.00	155	0.00	109	0.00	150
2	0.68	13	0.62	11	0.51	4	0.51	8	0.51	5
3	0.10	21	0.00	17	0.00	30	0.00	12	0.00	13
4	0.00	234	0.00	198	0.00	145	0.00	135	0.00	123
5	1.33	663	1.07	572	0.93	425	0.93	379	0.93	352
6	0.13	35	0.07	31	0.03	127	0.03	22	0.03	68
7	5.00	50	3.40	33	2.60	51	2.67	22	2.63	28
8	0.20	99	0.08	83	0.05	152	0.05	58	0.05	50
9	0.68	41	0.32	35	0.32	83	0.32	25	0.32	115
10	0.03	22	0.00	19	0.00	15	0.00	13	0.00	9
Mean	0.82	137.9	0.56	118.7	**0.44**	113.5	0.45	**78.3**	**0.44**	91.3

Table 11. False positive rate (FPR) on our own dataset (EndoData). This table extentends Table 10 by providing the FPR for five different polyp detection approaches on EndoData. The first two models are baseline models, and the last three are different stages of our polyp-detection system. Values displayed in bold font indicate the highest or most optimal results. The abbreviation "adv." is an acronym for the term "advanced". The FPR is given in %.

Video	YOLOv5 (Base)	F-RCNN	YOLOv5 (Adv.)	REPP	RT-REPP
1	88.15	90.39	90.60	93.20	90.88
2	99.28	99.39	99.78	99.56	99.72
3	90.32	92.02	86.73	94.23	93.78
4	45.11	49.27	57.01	58.75	60.99
5	58.32	61.86	68.58	71.00	72.49
6	97.62	97.89	91.88	98.49	95.48
7	40.97	51.26	40.49	61.20	55.34
8	82.37	84.79	75.27	88.86	90.25
9	94.18	94.99	88.89	96.37	85.24
10	77.69	80.13	83.62	85.49	89.49
Mean	77.40	80.20	78.29	**84.72**	83.37

Table 12. Detailed evaluation of EndoData. This table shows a comparison of five different polyp-detection approaches on the our EndoData dataset. The first two models are baseline models, and the last three models are different stages of our polyp-detection system. F1 and mAP are given in %, and the speed is given in FPS. The abbreviation "adv." is an acronym for the term "advanced".

Video	YOLOv5 (Base)		F-RCNN		YOLOv5 (Adv.)		REPP		RT-REPP	
	mAP	F1	mAP	F1	mAP	F1	mAP	F1	mAP	F1
1	72.77	72.69	84.23	82.26	79.25	82.23	89.84	89.43	82.98	84.26
2	86.30	86.71	86.04	90.51	89.06	94.18	92.83	95.91	90.01	94.74
3	85.65	85.71	93.10	92.88	91.20	91.50	99.10	97.99	98.51	97.00
4	70.57	73.88	82.88	78.17	76.96	79.99	85.43	85.36	83.67	83.99
5	39.45	54.84	44.23	56.79	45.84	58.98	49.60	63.98	49.28	62.40
6	90.22	90.94	94.02	92.11	96.13	96.00	98.38	97.48	96.75	97.50
7	15.12	34.89	29.13	47.81	21.66	43.40	31.72	53.33	28.41	46.39
8	91.14	86.35	96.32	92.71	96.66	94.43	99.46	98.48	98.67	97.00
9	77.49	80.87	78.48	84.72	82.61	87.44	85.11	89.29	81.61	86.59
10	88.73	87.08	88.28	89.10	91.95	94.43	95.82	96.50	92.28	94.91
Mean	77.09	79.45	81.74	83.97	82.28	86.32	87.24	89.97	84.29	87.97

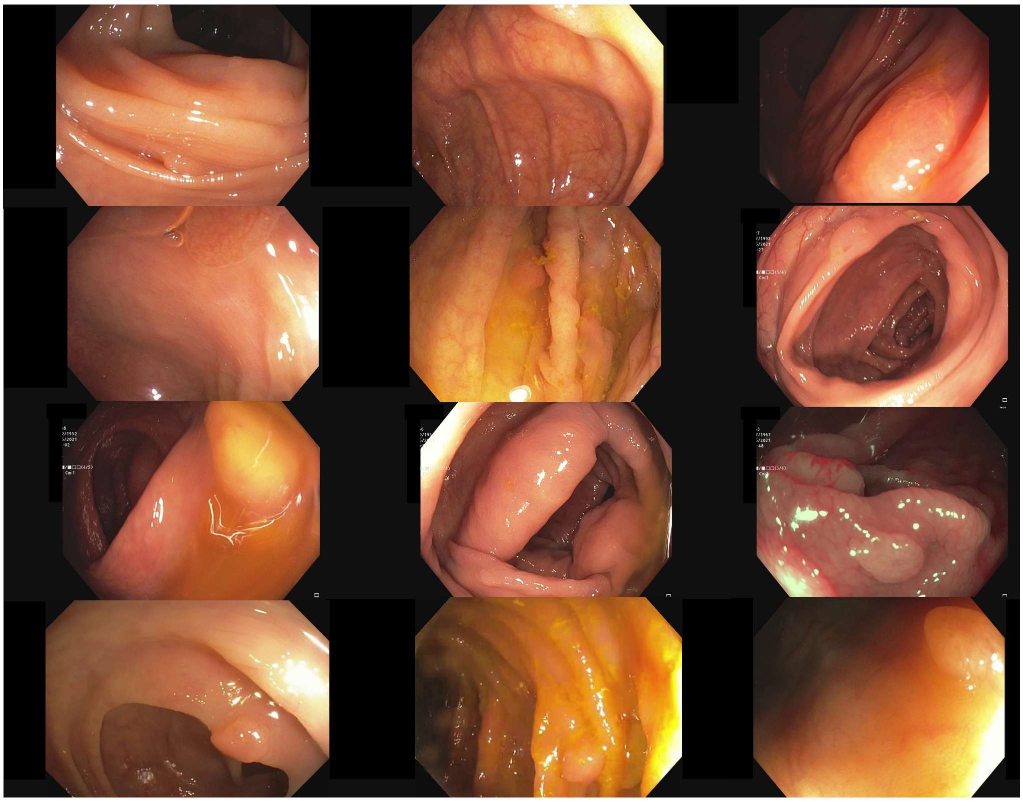


Figure 13. Example images of the Endodata dataset for evaluation.

The evaluation for FDT is shown in Table 10. For the YOLOv5 (base), only video 4 does not receive a delay in detection. Nevertheless, all polyps are detected at least once with every algorithm. The FDT of YOLOv5 (base) is inferior in all videos to the other models. The Faster R-CNN algorithm does recognize the polyp in the first frame in videos 1, 3, 4, and 10 for YOLOv5 (adv.), REPP, and RT-REPP. The FDT for these three models does not differ except for video 7. This difference is due to REPP and RT-REPP removing the detection in the post-processing process. Those three approaches also detect the polyps in the first frame for videos 1, 3, 4, and 10, like Faster R-CNN. For 9 out of the 10 videos, FDT is under 1 s; therefore, the polyp should be sufficiently detected to show the gastroenterologist its position. Nevertheless, in video 7 there is a FDT of 2.6 s. Such a late detection of a polyp may miss the polyp for the gastroenterologist. However, REPP and RT-REPP are reducing the number of FPs from an average of 113.5 to 78.3 and 91.3.

We evaluate the models on our dataset with the same metrics as the CVC-VideoClinicDB dataset. On the EndoData dataset, the results are equivalent to the predictions of the CVC-VideoClinicDB data. The mAP is, on average, consistently lower than the F1-score. Additionally, REPP is again the best scoring model. Again most values are over 90% F1 value for RT-REPP. The dataset appears to be more challenging than the CVC-VideoClinicDB dataset as there are just five videos with F1-scores over 90%.

Furthermore, we like to compare our results to the results of Livovsky et al. [50]. Livovsky et al., have only evaluated their approach on a closed custom dataset; therefore, we are unable to provide a qualitative comparison on the CVC-VideoClinicDB benchmark. Nevertheless, we qualitatively compare their results with our results on EndoData. Livovsky et al., achieved a recall of 88.5 % with polyps visible longer than 5 s on their custom test data. Our approach achieves a recall of 89.32 % on our custom dataset. As the two test datasets are different it is not possible to quantitatively show which system is better, nevertheless, both systems achieve similar recall values on their test sets.

3.3. Explaining the Model with Heatmaps

This paragraph presents a methodology to generate visual explanations for deriving insight into our polyp-detection systems decisions using the Grad-CAM algorithm [82]. We follow the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) [83]. Nevertheless, we changed the Grad-CAM algorithm to fit an object/polyp detection task instead of classification. The Grad-CAM algorithm receives the image of the model's prediction, the result, and the last two layers of the CNN YOLOv5. YOLOv5 outputs the detections for the three scales p3 (large), p4 (medium), and p5 (small). Each detection output with a shape of $[bsz; n_a; h; w; (5 + n_c)]$, where bsz is the batch size, n_a is the number of anchor boxes per grid cell, h is the height of the feature map, w is the width of the feature map, four box coordinates + objectness = 5, and n_c is the number of classes. Next, the three scales are concatenated and reshaped, resulting in an output of shape $[bsz; n_a \times h \times w; (5 + n_c)]$ followed by data augmentation. The augmentation identifies the scales from which the detections originate.

After that, the methodology employs a customized version of the non-max suppression algorithm (NMS) to reduce the number of detections to the most probable ones. For this purpose, the algorithm multiplies objectness probability p_o and class probability vector p_c and takes its maximum, $p_d^* = \max(p_d) = \max(p_o * p_c)$, which it subsequently uses as one decision criterion for reducing detections. This procedure ultimately results in significantly fewer and more confident detections. Furthermore, it associates each detection with a unique detection probability p_d^*, objectness probability p_o, and class probability $p_c^{(i)}, i = 1 \ldots n_c$. The presented methodology carries these values along and inserts them in the Grad-CAM algorithm for y^c.

The next step encompasses the execution of the Grad-CAM algorithm for each of the probabilities mentioned above. Here, the proposed methodology calculates for each probability the gradients $\frac{\partial y^c}{\partial A^k}$ for three feature map activations, namely for p3, p4, and p5.

Afterward, the presented approach transforms the emitted localization maps into heatmaps, which are significantly smaller than the original size of the input image. The proposed method upscales to the original image size by interpolation and then superimposes the heatmaps onto the original image. The resulting image shows highlighted image regions that contain pixels that positively influence the value of y^c. The method also draws a corresponding bounding box for each detection, its score, and the originating scale onto the superimposed image to increase the informational content. The final result is $|\#scores| \times |\#dets| \times |\#scales|$ superimposed images for each input image.

YOLOv5 was implemented in the python programming language with the PyTorch deep learning library. For this reason, this work also uses python and PyTorch to implement the Grad-CAM algorithm for YOLOv5 and necessary extensions. The most important feature of the PyTorch deep learning library is the concept of so-called hooks, which enable the extraction of the gradients obtained via backpropagation. Itoh et al. [84] showed that the classic Grad-CAM application may result in noisy heatmaps when using a YOLOv3 algorithm. We achieved less noisy heatmaps by recreating the Grad-CAM algorithm as described above. These heatmaps are similar to the results of Itoh et al., when using their application [84].

The first column of Figure 14 shows the five original frames on which the model should detect a single polyp. The second, third, and fourth columns in Figure 14 depict the resulting heatmaps of the approach when assigning p_c to y^c for backpropagation of the respective gradients.

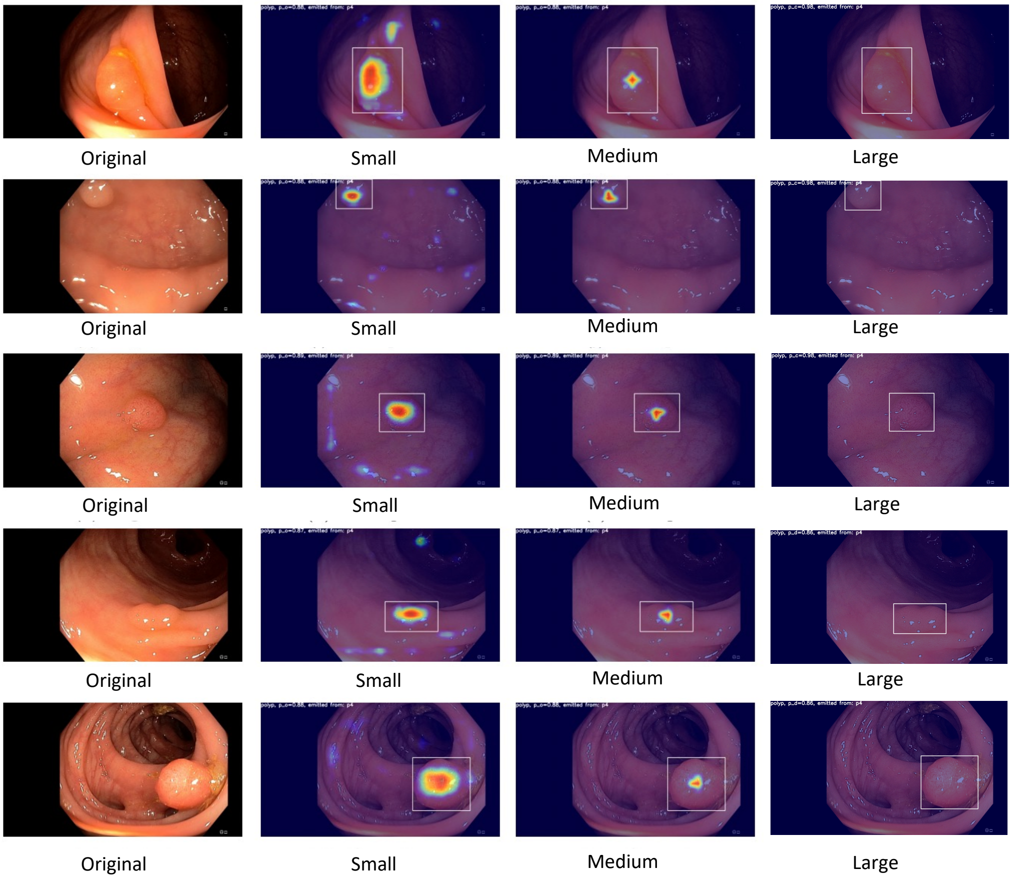


Figure 14. Heatmaps for polyp detection. This figure illustrates the detections of the model using the Grad-CAM algorithm. Thereby, the pixels most relevant for the detection are marked in warm colors such as red, and pixels less relevant for the detection in cold colors such as blue. The CNN has three detection outputs for small, medium, and large objects.

Figure 14 with a small and 14 medium polyp depicts the following behavior: the model focuses on the crucial image regions to classify and localize the present polyp while traversing from the output scale from small to medium. As expected, the model increases the pixel intensity important for the localization from p3 to p4. Furthermore, we notice that the highlighted red regions in Figure 14 encompass the center point of the respective bounding box. This shows that the model's focus in case of high polyp-detection confidence activates the necessary areas of the image.

Nevertheless, Figure 14 for large polyps displays the opposite behavior where the detected polyps are not highlighted in the heatmaps. The detected polyps are not large enough to activate the neurons for this part of the YOLOv5 architecture.

The above observations conclude that the model is functioning as expected. This shows the necessity of the proposed method to confirm or rebuke the assumption of the analyzed model's function and expected behavior.

4. Discussion

The system's limitations and clinical use are discussed in the following subsections. We especially focus on false polyp detections and discuss those system failures using frames of our CVC-VideoClinicDB and EndoData datasets. Additionally, we debate the clinical application of the system.

4.1. Limitations

We initiate the discussion of our limitations with a failure analysis of our model. First, we refer to Tables 7 and 12, and specifically to videos with significantly worse performance than the rest, i.e., videos 8, 12, 15, and 17 of the CVC-VideoClinicDB dataset and video 7 of EndoData. The videos differ in polyp detection difficulty; some videos only contain typical polyps with a straight angle and good lighting, while other videos have bad contrast, slanted angles and atypical polyps. Hence, multiple reasons can be attributed to the decreased performance on these videos:

Contrast and lighting are one of the main causes for missing or misidentifying a polyp. Figure 15 shows three frames taken from video 12 of the CVC-VideoClinicDB dataset. The image on the left shows a correct polyp detection and represents the exception. Most other frames either misidentify the size of the polyp, as seen in the middle image or do not detect the polyp at all, as seen in the right image. In this case, it is most likely an issue of contrast, as the polyp is oversaturated. As this applies to most video frames, the F1-score is negatively impacted.

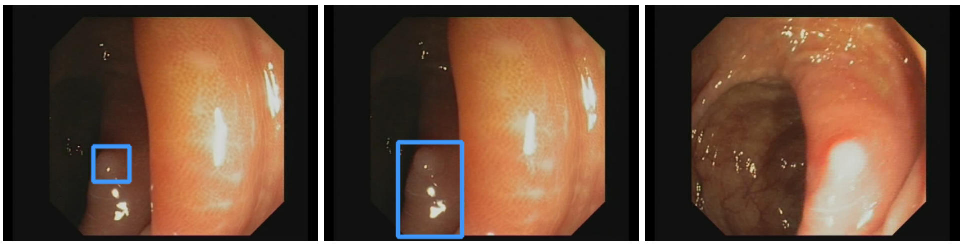


Figure 15. Examples of errors in video 12 of the CVC-VideoClinicDB dataset. The left image shows a correct polyp detection, the middle image misidentifies the size of the polyp and the right image shows no detection due to oversaturation.

Some polyps have uncommon shapes, an atypical surface texture, or a rare color and are underrepresented in the dataset. A notable example of such a polyp can be seen in video 15 of the CVC-VideoClinicDB dataset with some frames shown in Figure 16. Due to its peculiar appearance, this polyp is missed in most frames, especially in those frames containing another polyp, as seen in the right image. However, this is partly caused by the CVC-VideoClinicDB ground truth. The ground truth masks cover only one polyp at a time, even if both are visible in a frame. Rare polyps are a challenge for every supervised model, and this issue can only be improved by collecting more data.

As mentioned above, even typical polyps can be difficult to detect when obstructed by parts of the colon or due to bad lighting and angles. Figure 17 depicts these issues. Furthermore, small polyp size and color that is very similar to the surrounding colon lead to a general miss-rate of around 10% of the polyps frames.

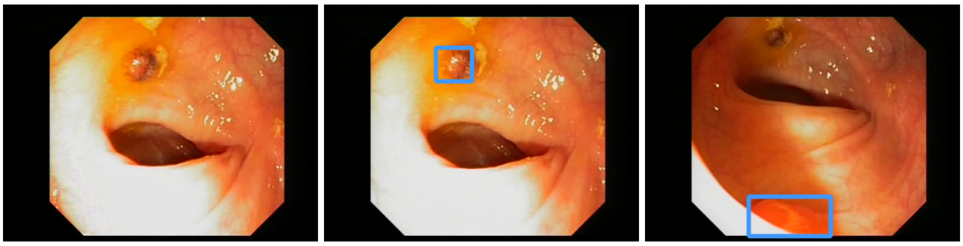


Figure 16. Examples of errors in video 15 of the CVC-VideoClinicDB dataset. The left image shows a missed polyp and the middle image a proper detection. On the right image, another polyp in the same frame is detected, while the other is missed.

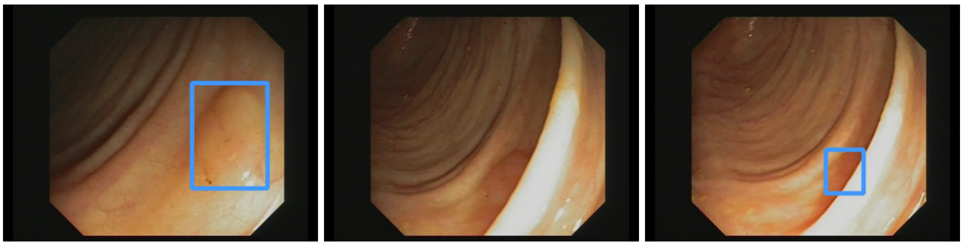


Figure 17. Examples of errors in video 17 of the CVC-VideoClinicDB dataset. The left image shows the detection of a flat polyp. The middle image shows the same polyp being missed because it is blocked by the colon wall. The right image shows a (short) re-detection.

FPs account for a large number of errors and, in turn, diminish the model's evaluation scores. Often, areas of the colon look similar to polyps due to lighting and contrast, leading to false bounding boxes and decreasing the mAP and F1-scores. The user can control the number of FPs by adjusting the probability threshold for discarding bounding boxes. A large threshold will reduce the number of FPs and increase the amount of missed polyps and vice versa. Therefore, our model tolerates more FPs and minimizes the amount of missed polyps in clinical practice. We discuss this point further in the following subsection.

Furthermore, our evaluation shows a significant advantage in using REPP and RT-REPP to reduce the number of FPs. In many cases, the FPs increase when using REPP or RT-REPP, e.g., in video 9 or 1 in Table 10. This happens if a false detection is highly significant. For example, the YOLOv5 architecture predicts a box on a bubble, which does not move and stays inside the frame. In this case, the detection score is high and REPP and RT-REPP include the FP. Nevertheless, REPP and RT-REPP reduce small FPs in several frames. In contrast, the YOLOv5 and Faster R-CNN architecture still include these small FPs. Therefore, in exceptional cases, FP can be increased. Nevertheless, longer-lasting FPs are less distracting than FPs with a short duration which might mislead the endoscopist and therefore increase the withdrawal time of the colonoscopy.

Finally, the usability of our system in a clinical setting depends on the financial cost. The system must operate in real-time during the examination, which a delay-prone server-based solution can not sustain. Therefore, every colonoscopy room needs its own system setup with one or more GPUs. Real-time detection needs both fast processing speed and enough VRAM for the video stream, especially while using RT-REPP. The current GPU with the best performance-to-cost ratio for these requirements is the NVIDIA Geforce RTX 3080, which cost around USD 800 in December 2021. Depending on the size of the clinic, the cost to equip the colonoscopy rooms will easily reach several thousand dollars. However, new GPUs are constantly developed, making current GPUs less expensive.

4.2. Clinical Use

A big advantage of our system is that it is already fully implemented as a complete package instead of having several conceptual parts. As described before, the system fits right between the video stream from an endoscopy camera, processes the input and displays the image on the clinical monitor. The direct video stream can still be displayed without our processing on a second monitor. Due to our multi-threaded implementation, the processed image is displayed essentially latency-free, which is a must in the clinical setting. Additionally, due to this implementation, in the future slower, more computationally heavy models can be used without having the disadvantage of higher latency. The system is also applicable to the most commonly used endoscopy processors, expecting a resolution of 1920 × 1080 pixels. Hence, the system can be set up easily in any common clinical setting.

As mentioned above, FPs are a topic of discussion for evaluation metrics in the context of clinical practice. An ideal model would only produce TPs, however a real trained model can not. In a clinical setting, false negatives are more dangerous to the patient than FPs. A FP box displayed by the model can be checked by the examiner and determined to be false, whereas a missed polyp may turn out to be fatal for the patient. As such, while common metrics essentially weight FPs and false negatives the same, clinical practice requires an increased weighting on false negatives in order to properly assess the models performance. We optimised the threshold value for the detection of polyps to rather show more FPs than missing a polyp. Nevertheless, the RT-REPP architecture is still achieving high precision values while also selecting a lower threshold. Therefore, our model does produce FPs rather than false negatives. Still, the amount of FPs is limited and does not disrupt the clinical workflow excessively. Nevertheless, the system has yet not been tested in a clinical trial. Therefore, we are planning to execute a clinical trial with the polyp-detection system.

Our code is open source and, as such, any information engineer can compile and install all necessary components by themselves. However, not every clinic has the necessary resources for this task. While remote support is possible in some cases, as of now, our dedicated software engineer needs to visit each clinic personally to solve more serious problems and to install software updates. We are working on a solution to make updates more dynamic and installable for any clinical environment.

5. Conclusions

In this study, we have implemented and tested a fully assembled real-time polyp-detection system that can be used directly in clinical applications. For this cause, we have developed and tested an object detection system, the core of our application, which consists of YOLOv5, an object detection CNN, and our novel post-processing step RT-REPP, a modified version REPP [13] for real-time detection. The system was tested on a public benchmark (CVC-VideoClinicDB) and our own newly collected and annotated dataset (EndoData) and surpassed state-of-the-art detectors with an F1-score of 90.25% one the CVC-VideoClinicDB data while still maintaining real-time speed.

Furthermore, we introduced a new performance metric “first detection time”, which measures the time between the first appearance of a polyp and the time of the first detection by the system. We discussed why the trade-off of a higher number of FPs in return for a better recall is more important for clinical application and, hence, why this metric is closer to measuring model performance in clinical application.

We have explained and discussed how our full system is assembled and implemented. The direct advantages are the flexibility derived from open-source installation and the out-of-the-box application placed between the endoscopy video stream and the clinic monitor for an almost latency-free bounding box detection display. While logistic disadvantages remain, such as the need for on-site visits for maintenance, we are working on finding solutions for these issues.

Author Contributions: A.K. implemented and coordinated the study, drafted the manuscript, interpreted the data and implemented the software. M.B. contributed to the creation of the prototype. K.M. helped with the implementation of the software. A.H. (Amar Hekalo), K.M., and M.B. contributed to the completion of the manuscript. D.F. helped with the creation of the data. J.T. helped with the data preprocessing. B.S. contributed to the installation of the software. F.P., A.H. (Alexander Hann), and W.G.Z. provided funding and reviewed the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: A.H. and W.Z. receive public funding from the state government of Baden-Württemberg, Germany (Funding cluster Forum Gesundheitsstandort Baden-Württemberg) to research and develop artificial intelligence applications for polyp detection in screening colonoscopy. F.P. receives funding from Interdisziplinäres Zentrum für Klinische Forschung (IZKF) from the University of Würzburg.

Institutional Review Board Statement: The study including retrospective and prospective collection of examination videos and reports was approved by the responsible institutional review board (Ethical committee Landesärztekammer Baden-Württemberg, 21 January 2021, F-2020-158). All methods were carried out in accordance with relevant guidelines and regulations.

Informed Consent Statement: Informed consent was obtained from all subjects and/or their legal guardian(s).

Data Availability Statement: The first dataset used for the analysis of this article is available in the GIANA challenge repository (https://endovissub2017-giana.grand-challenge.org/, accessed on 18 December 2022). The second dataset (EndoData) used during the analysis is available from the corresponding author on reasonable request.

Acknowledgments: We kindly thank the University Hospital of Würzburg, the Interdisziplinäres Zentrum für Klinische Forschung (IZKF) and the Forum Gesundheitsstandort Baden-Württemberg for supporting the research. Furthermore, the authors acknowledge the support by J.F. Riemann, "Stiftung Lebensblicke" the Foundation for early detection of colon cancer.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CRC	Colorectal cancer
CNN	Convolutional neural network
CAD	Computer-aided detection
CADx	Computer-aided diagnosis
SSD	Single-shot detector
REPP	Robust and efficient post-processing
RT-REPP	Real-time robust and efficient post-processing
COCO	Common objects in context
JSON	JavaScript object notation
YOLO	You only look once
YOLOv5	You only look once (version 5)
FDT	First detection time
AI	Artificial intelligence
GIANA	Gastrointestinal image analysis
WCE	Wireless capsule endoscopy
CEM	Context enhancement module
GAN	Generative adversarial network
FastCat	Fast colonoscopy annotation tool
FPS	Frames per second
GPU	Graphical processing unit
R-CNN	Region based convolutional neural network
SSIM	Structural similarity
ResNet	Residual neural network
SVM	Support vector machine
CSPNet	Cross stage partial network
SPP	Spatial pyramid pooling
VGG-16	Visual Geometry Group-16
FPN	Feature pyramid network
PANet	Path aggregation network
IoU	Intersection over union

References

1. Bray, F.; Ferlay, J.; Soerjomataram, I.; Siegel, R.L.; Torre, L.A.; Jemal, A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2018**, *68*, 394–424. [CrossRef] [PubMed]
2. Hazewinkel, Y.; Dekker, E. Colonoscopy: Basic principles and novel techniques. *Nat. Rev. Gastroenterol. Hepatol.* **2011**, *8*, 554–564. [CrossRef] [PubMed]
3. Rex, D.K.; Cutler, C.S.; Lemmel, G.T.; Rahmani, E.Y.; Clark, D.W.; Helper, D.J.; Lehman, G.A.; Mark, D.G. Colonoscopic miss rates of adenomas determined by back-to-back colonoscopies. *Gastroenterology* **1997**, *112*, 24–28. [CrossRef] [PubMed]
4. Heresbach, D.; Barrioz, T.; Lapalus, M.; Coumaros, D.; Bauret, P.; Potier, P.; Sautereau, D.; Boustière, C.; Grimaud, J.; Barthélémy, C.; et al. Miss rate for colorectal neoplastic polyps: A prospective multicenter study of back-to-back video colonoscopies. *Endoscopy* **2008**, *40*, 284–290. [CrossRef] [PubMed]
5. Leufkens, A.; Van Oijen, M.; Vleggaar, F.; Siersema, P. Factors influencing the miss rate of polyps in a back-to-back colonoscopy study. *Endoscopy* **2012**, *44*, 470–475. [CrossRef]
6. Van Rijn, J.C.; Reitsma, J.B.; Stoker, J.; Bossuyt, P.M.; Van Deventer, S.J.; Dekker, E. Polyp miss rate determined by tandem colonoscopy: A systematic review. *Off. J. Am. Coll. Gastroenterol. ACG* **2006**, *101*, 343–350. [CrossRef]
7. Kim, N.H.; Jung, Y.S.; Jeong, W.S.; Yang, H.J.; Park, S.K.; Choi, K.; Park, D.I. Miss rate of colorectal neoplastic polyps and risk factors for missed polyps in consecutive colonoscopies. *Intest. Res.* **2017**, *15*, 411–418. [CrossRef]
8. Ahn, S.B.; Han, D.S.; Bae, J.H.; Byun, T.J.; Kim, J.P.; Eun, C.S. The miss rate for colorectal adenoma determined by quality-adjusted, back-to-back colonoscopies. *Gut Liver* **2012**, *6*, 64. [CrossRef]
9. Puyal, J.G.B.; Brandao, P.; Ahmad, O.F.; Bhatia, K.K.; Toth, D.; Kader, R.; Lovat, L.; Mountney, P.; Stoyanov, D. Polyp detection on video colonoscopy using a hybrid 2D/3D CNN. *Med. Image Anal.* **2022**, *82*, 102625. [CrossRef]
10. Misawa, M.; Kudo, S.-E.; Mori, Y.; Cho, T.; Kataoka, S.; Yamauchi, A.; Ogawa, Y.; Maeda, Y.; Takeda, K.; Ichimasa, K.; et al. Artificial intelligence-assisted polyp detection for colonoscopy: Initial experience. *Gastroenterology* **2018**, *154*, 2027–2029. [CrossRef]
11. Misawa, M.; Kudo, S.-E.; Mori, Y.; Hotta, K.; Ohtsuka, K.; Matsuda, T.; Saito, S.; Kudo, T.; Baba, T.; Ishida, F.; et al. Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). *Gastrointest. Endosc.* **2021**, *93*, 960–967. [CrossRef]
12. Ishiyama, M.; Kudo, S.-E.; Misawa, M.; Mori, Y.; Maeda, Y.; Ichimasa, K.; Kudo, T.; Hayashi, T.; Wakamura, K.; Miyachi, H.; et al. Impact of the clinical use of artificial intelligence–assisted neoplasia detection for colonoscopy: A large-scale prospective, propensity score–matched study (with video). *Gastrointest. Endosc.* **2022**, *95*, 155–163. [CrossRef]
13. Sabater, A.; Montesano, L.; Murillo, A.C. Robust and efficient post-processing for Video Object Detection. In Proceedings of the International Conference of Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 25–29 October 2020.
14. Krishnan, S.; Yang, X.; Chan, K.; Kumar, S.; Goh, P. Intestinal abnormality detection from endoscopic images. In Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Vol.20 Biomedical Engineering Towards the Year 2000 and Beyond (Cat. No.98CH36286), Hong Kong, China, 29 October–1 November 1998; Volume 2, pp. 895–898. [CrossRef]
15. Karkanis, S.; Iakovidis, D.; Maroulis, D.; Karras, D.; Tzivras, M. Computer-Aided Tumor Detection in Endoscopic Video Using Color Wavelet Features. *Inf. Technol. Biomed. IEEE Trans.* **2003**, *7*, 141–152. [CrossRef]
16. Hwang, S.; Oh, J.; Tavanapong, W.; Wong, J.; de Groen, P.C. Polyp Detection in Colonoscopy Video using Elliptical Shape Feature. In Proceedings of the 2007 IEEE International Conference on Image Processing, San Antonio, TX, USA, 16–19 September 2007; Volume 2, pp. II-465–II-468. [CrossRef]
17. Bernal, J.; Sanchez, J.; Vilariño, F. Towards Automatic Polyp Detection with a Polyp Appearance Model. *Pattern Recognit.* **2012**, *45*, 3166–3182. [CrossRef]
18. Iakovidis, D.K.; Koulaouzidis, A. Automatic lesion detection in capsule endoscopy based on color saliency: Closer to an essential adjunct for reviewing software. *Gastrointest. Endosc.* **2014**, *80*, 877–883. [CrossRef]
19. Ratheesh, A.; Soman, P.; Nair, M.R.; Devika, R.; Aneesh, R. Advanced algorithm for polyp detection using depth segmentation in colon endoscopy. In Proceedings of the 2016 International Conference on Communication Systems and Networks (ComNet), Thiruvananthapuram, India, 21–23 July 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 179–183.
20. Klare, P.; Sander, C.; Prinzen, M.; Haller, B.; Nowack, S.; Abdelhafez, M.; Poszler, A.; Brown, H.; Wilhelm, D.; Schmid, R.M.; et al. Automated polyp detection in the colorectum: A prospective study (with videos). *Gastrointest. Endosc.* **2019**, *89*, 576–582. [CrossRef]
21. Zhu, R.; Zhang, R.; Xue, D. Lesion detection of endoscopy images based on convolutional neural network features. In Proceedings of the 2015 8th International Congress on Image and Signal Processing (CISP), Shenyang, China, 14–15 October 2015; pp. 372–376. [CrossRef]
22. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, Nevada, 3–6 December 2012; pp. 1097–1105.
23. Tajbakhsh, N.; Shin, J.Y.; Gurudu, S.R.; Hurst, R.T.; Kendall, C.B.; Gotway, M.B.; Liang, J. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Trans. Med. Imaging* **2016**, *35*, 1299–1312. [CrossRef]
24. Yuan, Z.; IzadyYazdanabadi, M.; Mokkapati, D.; Panvalkar, R.; Shin, J.Y.; Tajbakhsh, N.; Gurudu, S.; Liang, J. Automatic polyp detection in colonoscopy videos. In Proceedings of the Medical Imaging 2017: Image Processing. International Society for Optics and Photonics, Orlando, FL, USA, 24 February 2017; Volume 10133, p. 101332K.

25. Yuan, Y.; Qin, W.; Ibragimov, B.; Zhang, G.; Han, B.; Meng, M.Q.H.; Xing, L. Densely connected neural network with unbalanced discriminant and category sensitive constraints for polyp recognition. *IEEE Trans. Autom. Sci. Eng.* **2019**, *17*, 574–583. [CrossRef]
26. Liu, Y.; Tian, Y.; Maicas, G.; Pu, L.Z.C.T.; Singh, R.; Verjans, J.W.; Carneiro, G. Photoshopping colonoscopy video frames. In Proceedings of the 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), Iowa City, IA, USA, 3–7 April 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–5.
27. Wang, D.; Zhang, N.; Sun, X.; Zhang, P.; Zhang, C.; Cao, Y.; Liu, B. Afp-net: Realtime anchor-free polyp detection in colonoscopy. In Proceedings of the 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), Portland, OR, USA, 4–6 November 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 636–643.
28. Liu, M.; Jiang, J.; Wang, Z. Colonic polyp detection in endoscopic videos with single shot detection based deep convolutional neural network. *IEEE Access* **2019**, *7*, 75058–75066. [CrossRef]
29. Zhang, P.; Sun, X.; Wang, D.; Wang, X.; Cao, Y.; Liu, B. An efficient spatial-temporal polyp detection framework for colonoscopy video. In Proceedings of the 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), Portland, OR, USA, 4–6 November 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1252–1259.
30. Zheng, Y.; Zhang, R.; Yu, R.; Jiang, Y.; Mak, T.W.; Wong, S.H.; Lau, J.Y.; Poon, C.C. Localisation of colorectal polyps by convolutional neural network features learnt from white light and narrow band endoscopic images of multiple databases. In Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 18–21 July 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 4142–4145.
31. Mo, X.; Tao, K.; Wang, Q.; Wang, G. An efficient approach for polyps detection in endoscopic videos based on faster R-CNN. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 3929–3934.
32. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1–9.
33. Shin, Y.; Qadir, H.A.; Aabakken, L.; Bergsland, J.; Balasingham, I. Automatic Colon Polyp Detection Using Region Based Deep CNN and Post Learning Approaches. *IEEE Access* **2018**, *6*, 40950–40962. [CrossRef]
34. Szegedy, C.; Ioffe, S.; Vanhoucke, V. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *arXiv* **2016**, arXiv:1602.07261.
35. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
36. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
37. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
38. Zhang, X.; Zou, J.; He, K.; Sun, J. Accelerating Very Deep Convolutional Networks for Classification and Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 1943–1955. [CrossRef]
39. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
40. Zhang, X.; Chen, F.; Yu, T.; An, J.; Huang, Z.; Liu, J.; Hu, W.; Wang, L.; Duan, H.; Si, J. Real-time gastric polyp detection using convolutional neural networks. *PLoS ONE* **2019**, *14*, e0214133. [CrossRef]
41. Bagheri, M.; Mohrekesh, M.; Tehrani, M.; Najarian, K.; Karimi, N.; Samavi, S.; Reza Soroushmehr, S.M. Deep Neural Network based Polyp Segmentation in Colonoscopy Images using a Combination of Color Spaces. In Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 23–27 July 2019; pp. 6742–6745. [CrossRef]
42. Sornapudi, S.; Meng, F.; Yi, S. Region-Based Automated Localization of Colonoscopy and Wireless Capsule Endoscopy Polyps. *Appl. Sci.* **2019**, *9*, 2404. [CrossRef]
43. Yuan, Y.; Meng, M.Q.H. Deep learning for polyp recognition in wireless capsule endoscopy images. *Med Phys.* **2017**, *44*, 1379–1389. [CrossRef]
44. Ng, A. Sparse autoencoder. *CS294A Lect. Notes* **2011**, *72*, 1–19.
45. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein GAN. *arXiv* **2017**, arXiv:1701.07875.
46. Huang, G.; Liu, Z.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.
47. Zhang, Y.; Liu, H.; Hu, Q. TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation. *arXiv* **2021**, arXiv:2102.08005.
48. Jha, D.; Ali, S.; Tomar, N.K.; Johansen, H.D.; Johansen, D.; Rittscher, J.; Riegler, M.A.; Halvorsen, P. Real-time polyp detection, localization and segmentation in colonoscopy using deep learning. *IEEE Access* **2021**, *9*, 40496–40510. [CrossRef]
49. Sharma, P.; Balabantaray, B.K.; Bora, K.; Mallik, S.; Kasugai, K.; Zhao, Z. An ensemble-based deep convolutional neural network for computer-aided polyps identification from colonoscopy. *Front. Genet.* **2022**, *13*, 844391. [CrossRef]
50. Livovsky, D.M.; Veikherman, D.; Golany, T.; Aides, A.; Dashinsky, V.; Rabani, N.; Shimol, D.B.; Blau, Y.; Katzir, L.; Shimshoni, I.; et al. Detection of elusive polyps using a large-scale artificial intelligence system (with videos). *Gastrointest. Endosc.* **2021**, *94*, 1099–1109. [CrossRef]

51. Itoh, H.; Roth, H.; Oda, M.; Misawa, M.; Mori, Y.; Kudo, S.E.; Mori, K. Stable polyp-scene classification via subsampling and residual learning from an imbalanced large dataset. *Healthc. Technol. Lett.* **2019**, *6*, 237–242. [CrossRef]
52. Misawa, M.; Kudo, S.; Mori, Y.; Cho, T.; Kataoka, S.; Maeda, Y.; Ogawa, Y.; Takeda, K.; Nakamura, H.; Ichimasa, K.; et al. Tu1990 Artificial intelligence-assisted polyp detection system for colonoscopy, based on the largest available collection of clinical video data for machine learning. *Gastrointest. Endosc.* **2019**, *89*, AB646–AB647. [CrossRef]
53. Nogueira-Rodríguez, A.; Dominguez-Carbajales, R.; Campos-Tato, F.; Herrero, J.; Puga, M.; Remedios, D.; Rivas, L.; Sánchez, E.; Iglesias, A.; Cubiella, J.; et al. Real-time polyp detection model using convolutional neural networks. *Neural Comput. Appl.* **2022**, *34*, 10375–10396. [CrossRef]
54. Xu, J.; Zhao, R.; Yu, Y.; Zhang, Q.; Bian, X.; Wang, J.; Ge, Z.; Qian, D. Real-time automatic polyp detection in colonoscopy using feature enhancement module and spatiotemporal similarity correlation unit. *Biomed. Signal Process. Control* **2021**, *66*, 102503. [CrossRef]
55. Qadir, H.A.; Balasingham, I.; Solhusvik, J.; Bergsland, J.; Aabakken, L.; Shin, Y. Improving Automatic Polyp Detection Using CNN by Exploiting Temporal Dependency in Colonoscopy Video. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 180–193. [CrossRef]
56. Fitting, D.; Krenzer, A.; Troya, J.; Banck, M.; Sudarevic, B.; Brand, M.; Böck, W.; Zoller, W.G.; Rösch, T.; Puppe, F.; et al. A video based benchmark data set (ENDOTEST) to evaluate computer-aided polyp detection systems. *Scand. J. Gastroenterol.* **2022**, *57*, 1397–1403. [CrossRef]
57. Silva, J.; Histace, A.; Romain, O.; Dray, X.; Granado, B. Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. *Int. J. Comput. Assist. Radiol. Surg.* **2014**, *9*, 283–293. [CrossRef]
58. Bernal, J.; Sánchez, F.J.; Fernández-Esparrach, G.; Gil, D.; Rodríguez, C.; Vilariño, F. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Comput. Med. Imaging Graph. Off. J. Comput. Med, Imaging Soc.* **2015**, *43*, 99–111. [CrossRef] [PubMed]
59. Angermann, Q.; Bernal, J.; Sánchez-Montes, C.; Hammami, M.; Fernández-Esparrach, G.; Dray, X.; Romain, O.; Sánchez, F.J.; Histace, A. Towards Real-Time Polyp Detection in Colonoscopy Videos: Adapting Still Frame-Based Methodologies for Video Sequences Analysis. In *Proceedings of the Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures, Quebec City, QC, Canada, 14 September 2017*; Cardoso, M.J., Arbel, T., Luo, X., Wesarg, S., Reichl, T., González Ballester, M.Á., McLeod, J., Drechsler, K., Peters, T., Erdt, M., et al., Eds.; Springer International Publishing: Cham, Switzerland, 2017; pp. 29–41.
60. Vázquez, D.; Bernal, J.; Sánchez, F.J.; Fernández-Esparrach, G.; López, A.M.; Romero, A.; Drozdzal, M.; Courville, A. A Benchmark for Endoluminal Scene Segmentation of Colonoscopy Images. *J. Healthc. Eng.* **2017**, *2017*, 4037190. [CrossRef] [PubMed]
61. Fernández-Esparrach, G.; Bernal, J.; López-Cerón, M.; Córdova, H.; Sánchez-Montes, C.; Rodríguez de Miguel, C.; Sánchez, F.J. Exploring the clinical potential of an automatic colonic polyp detection method based on the creation of energy maps. *Endoscopy* **2016**, *48*, 837–842. [CrossRef] [PubMed]
62. Jha, D.; Smedsrud, P.H.; Riegler, M.A.; Halvorsen, P.; de Lange, T.; Johansen, D.; Johansen, H.D. Kvasir-seg: A segmented polyp dataset. In Proceedings of the International Conference on Multimedia Modeling, Daejeon, Republic of Korea, 5–8 January 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 451–462.
63. Ali, S.; Braden, B.; Lamarque, D.; Realdon, S.; Bailey, A.; Cannizzaro, R.; Ghatwary, N.; Rittscher, J.; Daul, C.; East, J. Endoscopy Disease Detection and Segmentation (EDD2020). *IEEE DataPort* **2020** . [CrossRef]
64. Krenzer, A.; Makowski, K.; Hekalo, A.; Fitting, D.; Troya, J.; Zoller, W.G.; Hann, A.; Puppe, F. Fast machine learning annotation in the medical domain: A semi-automated video annotation tool for gastroenterologists. *BioMed. Eng. OnLine* **2022**, *21*, 33. [CrossRef]
65. Lambert, R.f. Endoscopic classification review group. Update on the Paris classification of superficial neoplastic lesions in the digestive tract. *Endoscopy* **2005**, *37*, 570–578.
66. Kang, J.; Gwak, J. Ensemble of Instance Segmentation Models for Polyp Segmentation in Colonoscopy Images. *IEEE Access* **2019**, *7*, 26440–26447. [CrossRef]
67. Jha, D.; Smedsrud, P.H.; Riegler, M.A.; Johansen, D.; De Lange, T.; Halvorsen, P.; Johansen, H.D. Resunet++: An advanced architecture for medical image segmentation. In Proceedings of the 2019 IEEE International Symposium on Multimedia (ISM), San Diego, CA, USA, 9–11 December 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 225–2255.
68. Guo, Y.B.; Matuszewski, B. Giana polyp segmentation with fully convolutional dilation neural networks. In Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. SCITEPRESS-Science and Technology Publications, Prague, Czech Republic, 25–27 February 2019; pp. 632–641.
69. de Almeida Thomaz, V.; Sierra-Franco, C.A.; Raposo, A.B. Training data enhancements for robust polyp segmentation in colonoscopy images. In Proceedings of the 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), Cordoba, Spain, 5–7 June 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 192–197.
70. Qadir, H.A.; Solhusvik, J.; Bergsland, J.; Aabakken, L.; Balasingham, I. A Framework With a Fully Convolutional Neural Network for Semi-Automatic Colon Polyp Annotation. *IEEE Access* **2019**, *7*, 169537–169547. [CrossRef]
71. Ali, S.; Zhou, F.; Daul, C.; Braden, B.; Bailey, A.; Realdon, S.; East, J.; Wagnières, G.; Loschenov, V.; Grisan, E.; et al. Endoscopy artifact detection (EAD 2019) challenge dataset. *arXiv* **2019**, arXiv:1905.03209.
72. Soberanis-Mukul, R.D.; Kayser, M.; Zvereva, A.A.; Klare, P.; Navab, N.; Albarqouni, S. A learning without forgetting approach to incorporate artifact knowledge in polyp localization tasks. *arXiv* **2020**, arXiv:2002.02883.

73. Xu, R.; Lin, H.; Lu, K.; Cao, L.; Liu, Y. A Forest Fire Detection System Based on Ensemble Learning. *Forests* **2021**, *12*, 217. [CrossRef]
74. Wang, C.Y.; Liao, H.Y.M.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh, I.H. CSPNet: A new backbone that can enhance learning capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 13–19 June 2020; pp. 390–391.
75. Tajbakhsh, N.; Gurudu, S.R.; Liang, J. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Trans. Med. Imaging* **2015**, *35*, 630–644. [CrossRef]
76. Sharma, P.; Bora, K.; Kasugai, K.; Balabantaray, B.K. Two Stage Classification with CNN for Colorectal Cancer Detection. *Oncologie* **2020**, *22*, 129–145. [CrossRef]
77. Wang, K.; Liew, J.H.; Zou, Y.; Zhou, D.; Feng, J. Panet: Few-shot image semantic segmentation with prototype alignment. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 9197–9206.
78. Mirjalili, S. Genetic algorithm. In *Evolutionary Algorithms and Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 43–55.
79. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
80. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
81. Wu, Y.; Kirillov, A.; Massa, F.; Lo, W.Y.; Girshick, R. Detectron2. 2019. Available online: https://github.com/facebookresearch/detectron2 (accessed on 18 December 2022).
82. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
83. Mongan, J.; Moy, L.; Kahn, C.E., Jr. Checklist for artificial intelligence in medical imaging (CLAIM): A guide for authors and reviewers. *Radiol. Artif. Intell.* **2020**, *2*, e200029. [CrossRef]
84. Itoh, H.; Misawa, M.; Mori, Y.; Kudo, S.E.; Oda, M.; Mori, K. Positive-gradient-weighted object activation mapping: Visual explanation of object detector towards precise colorectal-polyp localisation. *Int. J. Comput. Assist. Radiol. Surg.* **2022**, *17*, 2051–2063. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

CEUR-WS.org/Vol-3148/paper2.pdf

Deep Learning using temporal information for automatic polyp detection in videos

Adrian Krenzer[1], Philipp Sodmann[2], Nico Hasler[1] and Frank Puppe[1]

[1]*Department of Artificial Intelligence and Knowledge Systems, University of Würzburg, Germany*
[2]*Gastroenterology department of the University Hospital of Würzburg, University of Würzburg, Germany*

Abstract

Previous research in the field of endoscopic computer vision has mainly focused on the detection of polyps using single images, but not videos or streams of images. The Endoscopic computer vision challenges 2.0 (EndoCV 2.0) is designed specifically to use streams of image sequences for the detection of polyps. In this paper, we describe our approach based on Gong et al. [1] by leveraging deep convolutional neural networks (CNNs) combined with temporal information to improve upon existing solutions for polyp detection. We demonstrate a detection system that combines similar ROI features across multiple frames with temporal attention to predict the final polyp detections for an emerging frame. For evaluation, we compare our approach to two classical image detection algorithms on a validation set based on training data provided by the challenge. The first one is a Single Shot Detector (SSD) called "YOLOv3", and the second one is a two-step region proposal-based CNN called "Faster R-CNN". To minimize the generalization error, we apply data augmentation and add additional open-source data for our training.

Keywords

Machine learning, Deep learning, Endoscopy, Automation, Video object detection, Attention

1. Introduction

The second leading cause of cancer-related deaths worldwide is Colorectal cancer (CRC) [2]. An excellent method to prevent CRC is to detect pre-cancerous lesions (colorectal polyps) of the disease as early as possible, using a colonoscopy. During a colonoscopy, a long flexible tube that is inserted through the rectum into the colon. The end of the tube has a small camera, allowing the physician to examine the colon thoroughly [1]. Computer science researchers are developing new methods to support physicians with this procedure. Polyp detection using computers is called *computer-aided detection (CAD)*. This process of polyp detection has already been subject to numerous publications.

However, these published solutions mostly focus on detection on still images [3]. Therefore, most of the published algorithms do not consider temporal dependencies and do compare themselves on benchmarks which do not consider temporal connections. To predict the final polyp detections for an emerging frame, our approach based on Gong et al. [1] utilizes temporal dependencies by combining similar ROI features across successive frames with temporal attention. Nevertheless, there are already some approaches in the literature addressing temporal dependency in polyp detection: In Itoh et al. [4], temporal information is included through a *3D-ResNet*. The 3D ResNet is thereby combining present and future frames for the detection of a new frame.

Furthermore, Qadir et al. [5] work with a traditional localization model, such as SSD [6] or Faster R-CNN [7], and post-process the output with an *FP Reduction Unit*. This approach considers the area of the generated bounding boxes over the 7 preceding and following frames and identifies and adjusts the outliers. The use of future frames causes a small delay, however, the actual calculation of the *FP Reduction Unit* is fast. A second promising method by Qadir et al. uses a two-step process which aims to decrease the proportion of false predictions. Furthermore, the CNN that flags several regions of interest (ROIs) for classification. The marked ROIs are then compared with subsequent frames and their corresponding ROIs and classified into true positives and false positives. The underlying assumption here is that each frame in a video is similar to its adjacent frames [5].

Xu et al. [8] designed a 2D CNN detector, which takes the spatiotemporal information into account and uses an ISTM network to improve its polyp detection efficiency while maintaining real-time speed. The model was trained on custom data. In addition, there is another approach which includes the temporal dependencies via post-processing. This approach uses fast image detection algorithms like YOLO and, afterwards, combines these predictions with an efficient real-time post-processing technic. This post-processing technique includes the predictions of polyps detected in past frames for future

---

4th International Workshop and Challenge on Computer Vision in Endoscopy (EndoCV2022) in conjunction with the 19th IEEE International Symposium on Biomedical Imaging ISBI2022, March 28th, 2022, IC Royal Bengal, Kolkata, India

adrian.krenzer@uni-wuerzburg.de (A. Krenzer)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

[1] https://www.mayoclinic.org/tests-procedures/colonoscopy/about/pac-20393569

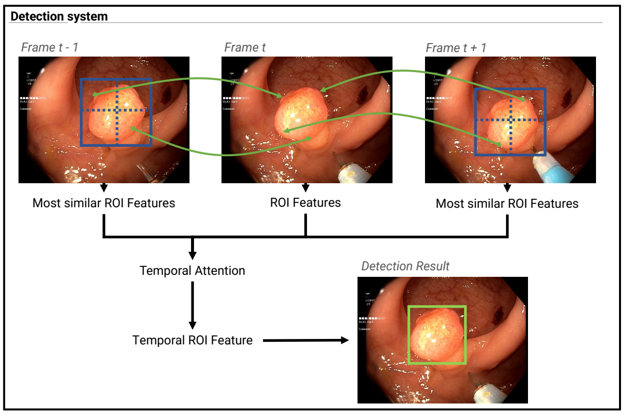


Figure 1: Overview of the polyp detection approach. t denotes the current frame for the detection. t - 1 denotes the frame before frame t and t+1 the frame after frame t. The ROIs are aligned through temporal attention for different frames. This figure is adopted from Gong et al. [1]

.

detections [9]. Taking these ideas forward, we implemented a polyp-detection model using the "ROI-Align Module" of Gong et al. [1] This allows the neural network to attend to information in previous frames and to combine ROI features from different frames for new predictions.

2. Data

To train the model, we used two public available datasets in addition to the challenge dataset:

- Kvasir-SEG [10]: 1000 polyp frames are included in the data collection, along with 1071 masks and bounding boxes. The sizes range from 332×487 pixels to 1920×1072 pixels. Gastroenterologists at Norway's *Vestre Viken Health Trust* confirmed the annotations. The majority of the frames show basic information on the left side, while others have a black box in the lower-left corner that contains data from ScopeGuide's endoscope position marking probe (Olympus). The data is available in the Kvasir-SEG repository[2].
- SUN Colonoscopy Video Database [11]: This dataset comprises 49,136 polyp frames from 100 distinct polyps, all of which are thoroughly documented. These frames were taken at Showa University Northern Yokohama and annotated by Showa University's specialist endoscopists. There are also 109,554 non-polyp frames present. The frames have a resolution of 1240×1080 pixels. The data is available in the SUN Colonoscopy Video repository[3].
- PolypGen2.0 (Polyp Generalization) [12, 13, 14]: This dataset is one of the two sets from the challenge and an extended version of the datasets from the 2020 and 2021 challenges. Both sub-challenges provide multi-center and diverse population datasets with tasks for both detection and segmentation, but the emphasis is on evaluating algorithm generalizability. The goal was to incorporate additional sequence/video data as well as multimodal data from various sites. PolyGen2.0 consists of 46 sequences with a total of 3290 images. All frames have a resolution of 1920×1080 pixels.

We split the PolyGen2.0 dataset into training and validation. For this purpose, 20 random sequences were assigned to validation (1366 images) and the rest to training (1924 images). The resulting validation set was used for all training steps.

3. Methods

In this section, we illustrate our approaches for the EndoCV2022 challenge, depicted in figure 1. All our models are trained on a NVIDIA QUADRO RTX 8000. After exploring the data, we decided to choose an algorithm which includes temporal information for the challenge, since the test data provided includes entire videos rather

---

[2] https://datasets.simula.no/kvasir-seg/

[3] http://sundatabase.org/

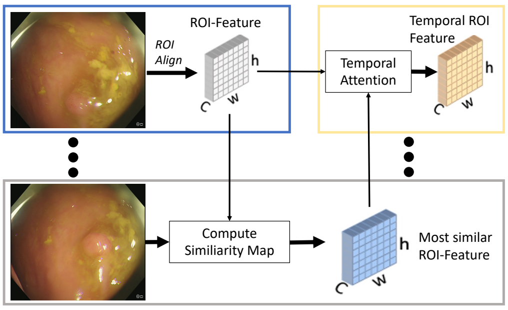


Figure 2: This figure illustrates temporal ROI align design and how its similarity map aggregation and temporal attention are used to compute the temporal ROI feature. This figure is adopted from Gong et al. [1]

than just images. The model is based on Gong et al. [1] and will be explained in the following.

Most state-of-the-art single-frame object detectors use the paradigm of region-based detection. When these detectors are used directly for video object detection (VID), object appearances in videos such as motion blur, video defocus, and object occlusions can degrade detection accuracy. These are frequent problems in endoscopy videos, which make the detection of polyps more difficult. Therefore, the main challenge is to design a method that can utilize the temporal redundancy of the information efficiently for the same object instance in a sequence of images or videos. To extract ROI features, most region-based detectors use ROI Align. However, ROI Align only uses the current frame feature map to extract features for current frame proposals, resulting in ROI features that lack the temporal information of the same object instance in the video. Using feature maps of other frames to perform ROI Align for the current frame proposals is a straightforward and clear technique for using temporal information. However, since the exact placement of the current frame proposals in other frame feature maps is unknown, the basic solution is ineffective.

Temporal ROI Align, on the other hand, defines a target frame as a frame in which the final prediction is made in real-time. In figure 2 the temporal ROI algin process is illustrated. Temporal ROI algin also allows the target frame to have multiple support frames, which are used to refine the features of the target frame. To achieve this refinement, the proposed operator selects the most comparable ROI features from the feature maps of the available support frames. The temporally redundant information of the same object instance in a video is contained in the extracted most comparable ROI characteristics. The main target now is to effectively capture diverse ROI features. Average is inefficient, because a polyp may seem blurry in some frames and clear in others. It is self-evident that the ROI characteristics of clear object instances should take precedence over the features of blurry instances in aggregate. To aggregate the ROI characteristics and the most comparable ROI features, multi-temporal attention blocks are used to perform the temporal feature aggregation. A major advantage of Temporal ROI Align is that it can extract the object features from support frames even when a polyp is partially occluded in the target frame. Therefore, the visible parts are dominant and features at these locations can still get enhanced.

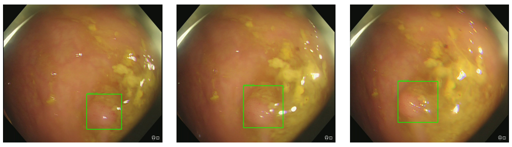


Figure 3: This figure shows a sequence of detections results with our algorithm on the test dataset provided by the challenge. Time is in this sequence running from the left side image to the right side while the polyp is moving to the left.

For our approach, the nerual network is trained for 10 epochs on our full dataset and then finetuned for 3 epochs on the challenge dataset. We choose the stochastic gradient descent (SGD) optimizer with a learning rate of 0.01, momentum of 0.9, and a weight decay of 0.0001. Additionally, we use a linear training warm-up schedule for 1 epoch. To enhance the generalization capabilities of our model, we use the following augmentation-schema: We applied a probability of 0.3 for upward and downward flips and a vertical flipping probability of 0.5. In addition, we rescaled the image with a probability of 0.64. We also use a translation along the horizontal axis with a probability of 0.5.

4. Results

In this section, we describe our results of the EndoCV2022 challenge. We highlight the performance of our approach and compare it to two classic benchmarking algorithms. One is an SSD algorithm called YOLOv3 [15] and the other is the ROI Proposal algorithm called Faster RCNN [16]. We trained both algorithms on the same data. For the validation, we create a validation set. The validation set consists of 20 sequences randomly chosen from the provided data (no additional data is included). We test the detection-created validation set. To enable the comparison of our results with the other participants of the challenge we do also declare our final scores: Score(mAP) 13.12 % and score(mAP50) 27.05 % are our final detection scores on the second round of the challenge evaluation.

Table 1 shows our results on our created validation set for the detection task where YOLOv3 is a benchmark

SSD algorithm, Faster R-CNN is the FASTER R-CNN algorithm with ResNet-101 backbone. For the evaluation, we report the F1-score. The F1-score describes the harmonic mean of precision and recall as shown in the following equations:

$$\text{Precision} = \frac{TP}{TP+FP} \qquad \text{Recall} = \frac{TP}{TP+FN}$$

$$F_1 = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} = \frac{2*TP}{2*TP+FP+FN}$$

We count an annotation as true positive (TP) if the boxes of our prediction and the boxes from the ground truth overlap at least 50%. Additionally, we display the mean average precision (mAP) and the mAP50 with a minimum IoU of 0.5 [17]. The mAP is calculated by the integral of the area under the precision-recall curve. Thereby, all predicted boxen are first ranked by their confidence value given by the polyp detection system. Then we computed precision and recall for different thresholds of these confidence values. When reducing the confidence threshold recall increases and precision decreases. This results in a precision-recall curve. Finally, for this precision-recall curve, the area under the curve is measured. This results in the mAP.

Table 1 shows that our approach is outperforming classical benchmarks on our validation data; this is mostly due to our temporal dependencies included in the algorithm which are not included in the Faster-RCNN approach. Notably, SSD algorithms like YOLOv3 are still 20 FPS faster than our approach in detecting single images. Nevertheless, our approach yield a huge recall increase of 9.5 % compared to the fast YOLOv3. We do especially emphasize this as recall is one of the most important metrics in real clinical use. As it is more important to find a missing polyp than to have additional false positiv detections. Figure 3 shows a sequence of detections results with our algorithm on the test dataset provided by the challenge. Furthermore, figure 4 shows a qualitative comparison of the three detection algorithms. We can see that all algorithms are detecting the polyp. Nevertheless, Yolov3 and Faster-RCNN are distracted by light reflections and therefore also draw wrong detections. Through temporal ROI align, our approach can incorporate the detections from previous frames and therefore does not get distracted by the light reflections.

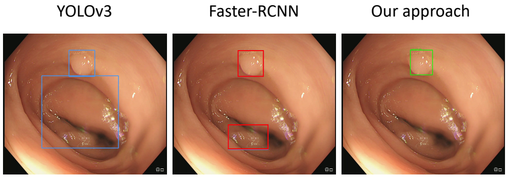


Figure 4: This figure shows a qualitative comparison of the three detection algorithms.

Table 1
Evaluation results of our validation split. We compare our approach based on Gong et al. [1] to two different polyp detection baselines on the same validation split from the challenge. Precision, Recall, F1, and mAP are given in %, and the speed is given in FPS.

	YOLOv3	Faster-RCNN	Our approach
mAP	13.8	14.2	**18.8**
mAP50	27.5	28.9	**32.8**
Precision	32.2	**34.5**	32.4
Recall	30.1	32.4	**39.6**
F1	31.1	33.4	**35.6**
Speed	**44**	15	24

5. Discussion

In this section, we like to discuss two main points: First, the limitations of our approach, and second how to use our approach in clinical useful settings. The first limitation is the current speed of our system. With an inference performance of 24 FPS, the algorithm is not capable of detecting every image with an endoscopy processor processing at 30 FPS. This can be mitigated by pruning and quantization-aware retraining. This on the other hand reduces the accuracy of the algorithm. Additionally, in the literature, a lot of benchmarking scores on still polyp images are already exceeding 80 % F1 score [18, 19]. Nevertheless, those are not directly comparable with our evaluation as they are using different data sets and do not include sequences of images.

The second and most drastic issue is that the system in its current form only works with video data and not a real-time stream of videos due to the dependencies in the algorithm, including preceding and future frames in the prediction. This issue may be solved by changing the algorithm to only use the preceding frames. In its current form, the algorithm can be used to evaluate endoscopies after they are completed or to detect polyps with wireless capsule endoscopy (WCE).

6. Conclusion

Overall, we demonstrate our approach to the Endoscopic computer vision challenges 2.0. We show a detection system that combines similar ROI Features across frames with temporal attention to create the final for polyp detections for a new emerging frame. The system thereby uses present, past, and future features on the temporal axis to create new polyp localizations. We show that the system exceeds classical benchmarks algorithms based on individual frames on our validation data from the challenge.

7. Compliance with ethical standards

This research study was conducted retrospectively using human subject data made available in open access [10, 11, 12, 13, 14]. Ethical approval was not required as confirmed by the license attached with the open access data.

8. Acknowledgments

This research is supported using public funding from Interdisziplinäres Zentrum für Klinische Forschung (IZKF) from the University of Würzburg.

References

[1] T. Gong, K. Chen, X. Wang, Q. Chu, F. Zhu, D. Lin, N. Yu, H. Feng, Temporal roi align for video object recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 1442–1450.

[2] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, A. Jemal, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA: A Cancer Journal for Clinicians 68 (2018) 394–424. URL: https://doi.org/10.3322/caac.21492. doi:`10.3322/caac.21492`.

[3] A. Krenzer, A. Hekalo, F. Puppe, Endoscopic detection and segmentation of gastroenterological diseases with deep convolutional neural networks., in: EndoCV@ ISBI, 2020, pp. 58–63.

[4] H. Itoh, H. Roth, M. Oda, M. Misawa, Y. Mori, S.-E. Kudo, K. Mori, Stable polyp-scene classification via subsampling and residual learning from an imbalanced large dataset, Healthcare Technology Letters 6 (2019) 237–242. URL: https://doi.org/10.1049/htl.2019.0079. doi:`10.1049/htl.2019.0079`.

[5] H. A. Qadir, I. Balasingham, J. Solhusvik, J. Bergsland, L. Aabakken, Y. Shin, Improving automatic polyp detection using CNN by exploiting temporal dependency in colonoscopy video, IEEE Journal of Biomedical and Health Informatics 24 (2020) 180–193. URL: https://doi.org/10.1109/jbhi.2019.2907434. doi:`10.1109/jbhi.2019.2907434`.

[6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C.-Y. Fu, A. C. Berg, Ssd: Single shot multibox detector, ArXiv abs/1512.02325 (2016).

[7] S. Ren, K. He, R. Girshick, J. Sun, Faster r-CNN: Towards real-time object detection with region proposal networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (2017) 1137–1149. URL: https://doi.org/10.1109/tpami.2016.2577031. doi:`10.1109/tpami.2016.2577031`.

[8] X. Liu, X. Guo, Y. Liu, Y. Yuan, Consolidated domain adaptive detection and localization framework for cross-device colonoscopic images, Medical image analysis 71 (2021) 102052.

[9] A. Krenzer, M. Banck, K. Makowski, A. Hekalo, D. Fitting, J. Troya, B. Sudarevic, W. G. Zoller, A. Hann, F. Puppe, A real-time polyp detection system with clinical application in colonoscopy using deep convolutional neural networks (2022).

[10] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, H. D. Johansen, Kvasir-seg: A segmented polyp dataset, in: International Conference on Multimedia Modeling, Springer, 2020, pp. 451–462.

[11] M. Misawa, S.-e. Kudo, Y. Mori, K. Hotta, K. Ohtsuka, T. Matsuda, S. Saito, T. Kudo, T. Baba, F. Ishida, et al., Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video), Gastrointestinal Endoscopy 93 (2021) 960–967.

[12] S. Ali, D. Jha, N. Ghatwary, S. Realdon, R. Cannizzaro, O. E. Salem, D. Lamarque, C. Daul, K. V. Anonsen, M. A. Riegler, et al., Polypgen: A multi-center polyp detection and segmentation dataset for generalisability assessment, arXiv preprint arXiv:2106.04463 (2021). doi:`10.48550/arXiv.2106.04463`.

[13] S. Ali, M. Dmitrieva, N. Ghatwary, S. Bano, G. Polat, A. Temizel, A. Krenzer, A. Hekalo, Y. B. Guo, B. M. et al., Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy, Medical Image Analysis 70 (2021) 102002. URL: https://www.sciencedirect.com/science/article/pii/S1361841521000487. doi:`https://doi.org/10.1016/j.media.2021.102002`.

[14] S. Ali, N. Ghatwary, D. Jha, E. Isik-Polat, G. Polat, C. Yang, W. Li, A. Galdran, M.-Á. G. Ballester, V. Thambawita, et al., Assessing generalisability of deep learning-based polyp detection and segmentation methods through a computer vision challenge, arXiv preprint arXiv:2202.12031 (2022). doi:`10.48550/arXiv.2202.12031`.

[15] J. Redmon, A. Farhadi, Yolov3: An incremental improvement, arXiv preprint arXiv:1804.02767 (2018).

[16] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, Advances in neural information processing systems 28 (2015).

[17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona,

D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: European conference on computer vision, Springer, 2014, pp. 740–755.

[18] D. Wang, N. Zhang, X. Sun, P. Zhang, C. Zhang, Y. Cao, B. Liu, Afp-net: Realtime anchor-free polyp detection in colonoscopy, in: 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), IEEE, 2019, pp. 636–643.

[19] X. Mo, K. Tao, Q. Wang, G. Wang, An efficient approach for polyps detection in endoscopic videos based on faster r-cnn, in: 2018 24th international conference on pattern recognition (ICPR), IEEE, 2018, pp. 3929–3934.

A User Interface for Automatic Polyp Detection Based on Deep Learning with Extended Vision

Adrian Krenzer[1,2](✉), Joel Troya[2], Michael Banck[1,2], Boban Sudarevic[2,3], Krzysztof Flisikowski[4], Alexander Meining[2], and Frank Puppe[1]

[1] Julius-Maximilians University of Würzburg, Sanderring 2,
97070 Würzburg, Germany
`adrian.krenzer@uni-wuerzburg.de`
[2] University Hospital Würzburg, Oberdürrbacher Straße 6,
97080 Würzburg, Germany
[3] Katharinenhospital, Heidelberg, Kriegsbergstrasse 60, 70174 Stuttgart, Germany
[4] Lehrstuhl für Biotechnologie der Nutztiere, School of Life Sciences,
Technische Universität München, Munich, Germany

Abstract. Colorectal cancer (CRC) is a leading cause of cancer-related deaths worldwide. To prevent CRC, the best method is screening colonoscopy. During this procedure, the examiner searches for colon polyps. Colon polyps are mucosal protrusions that protrude from the intestinal mucosa into the intestinal lumen. During the colonoscopy, all those polyps have to be found by the examiner. However, as the colon is folding and winding, polyps may hide behind folds or in uninvestigated areas and be missed by the examiner. Therefore, some publications suggest expanding the view of the examiner with multiple cameras. Nevertheless, expanding the examiner's view with multiple cameras leads to overwhelming and cumbersome interventions. Therefore, we suggest maintaining the examiner's classical endoscope view but extending the endoscope with side cameras. Those side camera views are only shown to an Artificial Intelligence (AI) trained for polyp detection. This AI system detects polyps on the side cameras and alarms the examiner if a polyp is found. Therefore, the examiner can easily move the main endoscope view on the AI detected area without being overwhelmed with too many camera images. In this study, we build a prototype of the endoscope with extended vision and test the automatic polyp detection system on gene-targeted pigs. Results show that our system outperforms current benchmarks and that the AI is able to find additional polyps that were not visualized with the main endoscope camera.

Keywords: Machine learning · Deep learning · Endoscopy · Gastroenterology · Automation · Object detection · Computer vision

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
G. Yang et al. (Eds.): MIUA 2022, LNCS 13413, pp. 851–868, 2022.
https://doi.org/10.1007/978-3-031-12053-4_62

1 Introduction

Colorectal cancer is one of the most common types of cancer globally. In most cases, the cause is unknown. Only three to five percent of all cases can be traced back to known genetic mutations that can be inherited and trigger colon cancer. Nevertheless, colorectal cancer almost always develops from growths that form in the mucosa of the colon, so-called intestinal polyps [8].

One of the most effective methods to prevent CRC is to detect the potential disease as early as possible using a colonoscopy. A colonoscopy inspects the large intestine (colon) with a long flexible tube inserted via the rectum. The tube carries a small camera to allow the physician to look inside the colon. The physician searches for polyps and analyses them carefully. Polyps are protrusions of the mucosal surface of various shapes and sizes that can be benign or malignant. Malignant polyps are at risk to turn into colorectal cancer. Polyps appear on the lining of the colon and rarely cause any symptoms. The two main types of polyps are non-neoplastic and neoplastic polyps. Non-neoplastic polyps are usually harmless, while polyps of type neoplastic can turn cancerous [6].

Therefore, even if many polyps are not cancerous, they always risk turning into colon cancer. In theory, the colonoscopist identifies all polyps of the patient during a colonoscopy and decides if it needs to be removed. However, there is always a potential risk of a polyp being missed during the colonoscopy. Previous studies showed that up to 27% of diminutive polyps are overlooked by physicians, which may be caused by the physician's lack of experience or fatigue and untypical appearance or bad visibility of the polyps [12,24]. Furthermore, a general error rate of 20%-24% during exams leads to a high risk for patients to die from CRC [18,25].

In conclusion, smaller polyps have a higher risk of being missed by the examiner than bigger polyps. Missed polyps are not removed and stay inside the colon, where they can have fatal consequences for the patient. Therefore, the colonoscopist must find and afterward remove all potential cancerous polyps to minimize the risk of colorectal cancer for the patient [1].

Additionally, there are challenges that increase the chance of polyps being missed. One of the fundamental challenges are folds in the colon. These folds can hide polyps from the examiner and increase the risk of developing CRC. In the literature are already different approaches to tackle the issue of hidden polyps by increasing the camera view of the examiner [10,28]. However, these approaches do always incorporate the additional views and monitors and therefore have the potential risk to overwhelm the examiner [11]. Accordingly, these procedures have not yet been implemented in practice. We propose an interface for automatic polyp detection, which includes an extended view. This extended view includes two additional side cameras to the endoscope. We run an artificial intelligence polyp detection system on these side-view cameras, which then alarms the examiner about missing a polyp. Therefore instead of the examiner being overwhelmed with different views, we let the AI handle the additional views, and the examiner can entirely focus on the classic view of the endoscope.

The main contributions of our paper are:

1) *We create an interface for automatic polyp detection with extends the vision of the endoscopists and shows seamless integration for the classic automatic polyp detection task.*
2) *We show that our system outperforms state of the art architectures on our dataset and present that additional polyps are found by the AI through adding extended vision to the system.*
3) *We create a prototype of an endoscope with side cameras and applied and test it during an animal trial with gene-targeted pigs.*

The interface with extended vision is publicly funded and developed by computer scienctists, engineers and endoscopists.

2 Data

One of the biggest problems in implementing deep learning methods is getting adequate qualitative data. Accordingly, getting high-quality colonoscopy video or image data is challenging for automated polyp detection. The challenge of data acquisition is caused by data protection issues and the expensive and time-consuming data annotation by experienced medical experts. We, therefore, used different technics to challenge these issues illustrated below. Further, we split our data into human and animal data to evaluate our system with extended vision. We could only apply our system to animal data as we did not have consent to use the system on humans. Nevertheless, all human data we got is used to pretrain our system for the polyp detection task.

2.1 Animal Data

Formerly, we published a framework that consists of two steps, a small expert annotation part and a large non-expert annotation part [16]. This moves most of the workload from the expert to a non-expert while ensuring high-quality data. Both annotation steps are supplemented with AI assistance to enhance the annotation efficiency further. We used the software Fast Colonoscopy Annotation Tool (FastCat) to process the entire annotation process. This tool supports the annotation process in endoscopic videos by enabling us to label these videos 20 times faster than traditional labeling techniques. The annotation process is split between at least two people. In the first step, an expert analyses the video and annotates a small set of video frames to verify the object's annotations.

In the second step, a non-expert has visual confirmation of the given object and annotates all following and preceding frames with AI assistance. To annotate individual frames, all video frames must be extracted first. Then, relevant frames can be pre-selected by an automated system, and this averts the expert from examining the entire video every single time. After the expert annotation, relevant frames are selected and can be passed on to an AI model. Then, the AI model detects and marks the desired object on all following and preceding frames with an annotation.

Afterward, the non-expert can adjust the AI annotations and further export the results used to train the AI model further. Additionally, the expert annotates the Paris classification [17], the size of the polyp and its location, as well as the start and end frame of the detected polyp, and one box for the non-expert annotators. Overall, as we were filming with extended vision, this data involved three camera angles. First is the classic endoscope view, the standard camera of the endoscope filming in front. It is an entire HD endoscope with a resolution of 1920 × 1080 px. Then we attached two side cameras to the endoscope. These side cameras capture other videos with a quality 320 × 320 px. The endoscope with extended vision is then inserted into four different pigs to create a dataset of 6185 side camera images. Those images are annotated by a medical expert, as illustrated in the previous paragraph. We pretrained our model on the human data illustrated below and then fine-tuned it on our collected animal data.

2.2 Human Data

We use our own data and all publicly available data for the development of our model. We merged the data from online resources and our own data to forge a data set of 506,338 images. The details about the creation of this training data set will are the same as presented in animal data. The data set is made of images and bounding box coordinates of boxes referring to the image. Here we list all the publicly available data we incorporated into the training: CVC-ColonDB [4] 2012, ETIS-Larib [27] 2014:, CVC-VideoClinicDB [3] 2017 CVC-EndoSceneStill [29] 2017, Kvasir-SEG [15] 2020, SUN Colonoscopy Video Database [21] 2020, CVC-Segementation-HD [30] 2017 and Endoscopy Disease Detection Challenge 2020 (EDD2020) [2]. Overall we built a team of advanced gastroenterologists, computer scientists, engineers and medical assistance staff. Together we produced a data set of 506,338 human images, including the open-source images listed above. Our data set includes 361 polyp sequences and 312 non-polyp sequences. This data set is then used for the pretraining of our model.

3 Methods

This section explains the software and hardware used in this work. Figure 1 illustrates the structure of our system. The illustration is split into three phases: video capture of endoscopic images with and without extended vision, AI detection system, and User interface. First, the endoscope coupled with two additional micro cameras captures the frames of the surface of the colon. Those frames are afterward input to the Artificial intelligence. This AI processes the frames in real-time and draws bounding boxes on the detected polyps. The detection results of the main camera are then shown to the endoscopist. The AI inspects only the extended views (side cameras) to avoid disturbing the endoscopist with the classical examination. If the AI detects a polyp in the extended views, the examiner is alarmed via an arrow on the screen on which camera (left camera or right camera) the polyp is detected. Afterward, the examiner can inspect

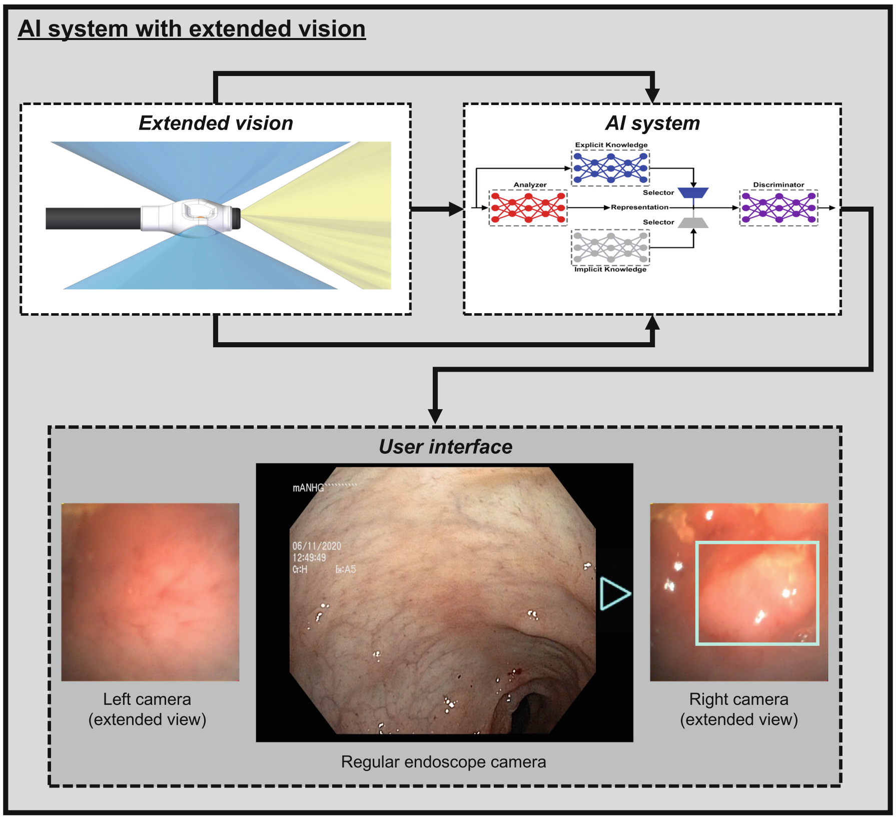


Fig. 1. Overview of the AI system with extended vision. This figure shows all the steps of the whole detection system. The start is an endoscope equipped with two additional micro cameras. The images taken from these cameras are then input into the AI system. If the AI detects a polyp in the extended views, the examiner is alarmed via an arrow on the screen. Afterward, the examiner can inspect the missed polyp with the main camera view. The examiner does never see the side camera views, just the arrows pointing him in the direction of the polyp.

the missed polyp with the main camera view. To further express the novelty of our approach using the YOLOR algorithm [31]. We like to highlight the combination of YOLOR and the side cameras. The architecture of YOLOR enables high detection accuracy while maintaining good detection speed, especially with low-resolution images. Therefore we consider it the best for detection of the side cameras, which operate in real-time with low resolution.

3.1 Video Processing System

As illustrated in Fig. 2, three types of optical camera signals were captured by our system in real-time: the endoscope image and the two lateral micro cameras.

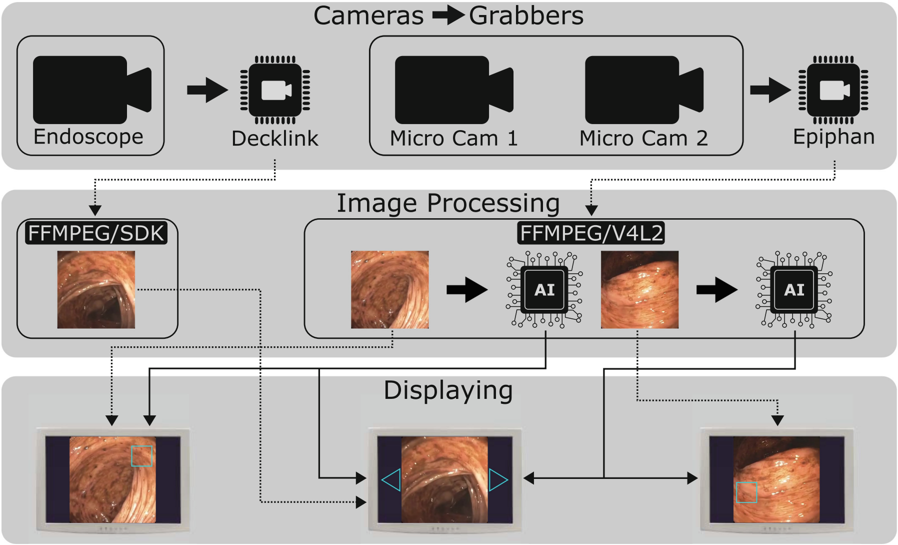


Fig. 2. Overview of the developed video processing system.

An endoscope can be described as a long, thin, illuminated flexible tube with a camera on the tip. They use fiber optics, which allow for the effective transmission of light to illuminate the inner mucosa of the gastrointestinal tract. A charge-coupled device (CCD) image sensor captures the image and transmits it using fine electrical wires to the video processor. The video processor outputs the image to display it on the screen, mainly using a serial digital interface (SDI) with Bayonet Neil-Concelman (BNC) connectors and/or digital visual interface (DVI). Image resolution has increased as technology has advanced. At the moment, most devices in the clinic can provide between 25–60 high-definition (HD) images per second, depending on the manufacturer and model. However, the first models with 4K resolution already appear on the market. In our case, the Olympus 180 CV endoscope processor was connected via SDI to a Blackmagic - DeckLink Mini Recorder 4K [20] on a 1080i signal.

The two micro cameras used for this study were the Osiris M from Optasensor GmbH. These micro cameras provide a resolution of 320 × 320 pixels and have the best focus point at 15mm. Their field of view is 90 degrees. Each microcamera was connected to an image processing system (Osiris M IPS, Optasensor GmbH.), providing a single output image of concatenated micro camera images. The final HDMI image output format is 8 Bit RGB 1080p. Therefore, only one grabber was needed for the micro cameras. To capture this signal, an Epiphan DVI2USB 3.0 grabber [13] was used. Both sources, the endoscope and the micro cameras, were processed by FFmpeg [7]. The micro cameras stream was captured within the V4L2 API. Since Blackmagic is not implementing V4L2, we compiled FFmpeg with the Decklink SDK to enable access to the device through FFmpeg. Processed by FFmpeg to BGR byte arrays, the data was converted to OpenCV [14] matrices for further steps. Since the data of the micro cameras are fused, splitting up and cropping is required to handle the two independent micro cameras individually. This results in a smaller BGR matrix for each camera and frame. The main matrix of the endoscope was forwarded as fast as possible to the main display pipeline. Since this is displayed to the monitor observed by the physician, keeping the delay is a priority for the best user experience. Each frame is forwarded to a second and third display pipeline for the micro cameras and simultaneously to the AI pipelines. The matrices are first cropped, scaled, and padded to the required 320×320 resolution to suit the convolutional neural network (CNN). Afterward, the color channels are swapped from BGR to RGB, before the data is normalized, uploaded to the GPU memory and injected into the CNN. Every microcamera uses its own copy of the CNN for inference. The outcome is a list of scored bounding boxes. If this list is not empty, an arrow is drawn on the main monitor, pointing to the direction of the camera in which the polyp has been detected. The boxes themselves are filtered by the confidence threshold. However, the boxes still contain relative coordinates based on the modified matrix. Therefore, coordinates are transformed to reverse the smaller matrix's crop-, scale, and pad operations. The resulting boxes are added to the secondary displays to highlight the detected polyp. Since the boxes result from the AI pipelines, they are 1–3 frames delayed. This means bounding boxes are always drawn on a more recent frame. However, by a framerate of 50 frames per second, the delay is only 20 ms per frame. Since during the withdrawal, the camera moves slowly, the boxes are still accurate. The arrows and boxes are displayed until the next cycle of the AI pipeline has finished and passed to the display pipeline. In addition, at the same time, all streams are recorded as h264 encoded video files, together with all the polyp detections triggered by the two micro cameras. This asynchronously recording process opens the possibility for a retrospective system evaluation.

3.2 Endoscope Assembly

To assemble the micro cameras on the endoscope, an add-on-cap was 3D printed. This add-on-cap was inserted and fixed 5 mm from the tip of the endoscope. The cap dimensions were 27.3 mm long, the maximum radius was 10.5 mm, and the thickness was 5.5 mm. The material used was nylon PA12, and a selective laser sintering printer (Lisa Pro, Sinterit sp. Z o.o.) was used to produce it. The cap contained two openings to allow the integration of the micro cameras into the normal endoscope's axis. The micro cameras included four mini light-emitting diodes (OSRAM Opto Semiconductors GmbH) arranged around it that allowed the illumination of the mucosa. The total dimensions of the micro camera with the diodes was $3.8 \times 3.8 \times 2\ \mathrm{mm}^3$ ($height \times width \times depth$). The design of the add-on-cap incorporated cut-out areas to allow the micro cameras to have a full field of view. To secure the micro cameras on the add-on-cap, silicone epoxy was used. Additionally, a 2 m length cable has to be connected to use the micro cameras. A tube-like plastic foil was used to protect the cable of the micro cameras. This allowed the flexible endoscope could maintain its normal mobility. Figure 3 shows the 3D printed cap that was assembled to the endoscope. As illustrated, the micro cameras are fixed to each of the sides, thus extending the field of view of mucosa inspected. The position of the add-on-cap does not disturb the mobility of the endoscope because the bending section starts further back. The design does not alter any of the functions that endoscopists normally perform: the instrument channel remains free for therapeutic interventions as well as the illumination, and the suction and water irrigation channels.

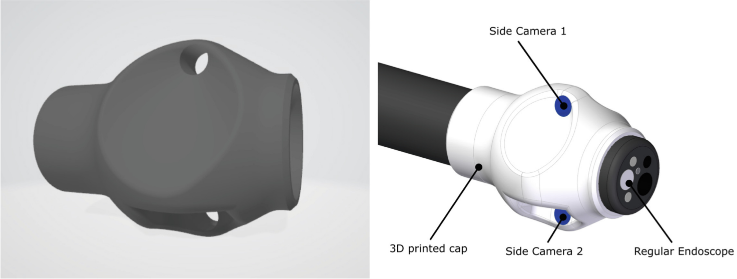


Fig. 3. Left: 3D printed cap used to assemble the micro cameras into the endoscope. Right: Assemble of the cap over the endoscope.

3.3 Polyp Detection System Using AI

Preprocessing and Data Augmentation. To ensure a fast processing speed while keeping up high detection accuracy, we rescale the images to 640×640 pixels. The change in image size allows the detection system to perform with high quality and a speed of 20 ms on an NVIDIA RTX 3080 GPU. In the clinical application subsection, we further define the use of different GPUs and the GPU requirements for a system able to process in real-time. Furthermore, we move the image and model to a half-precision binary floating-point (FP16). However, most machine learning models are in a precision binary floating-point (FP32). With FP16 the model calculates faster but also delivers high quality results. Afterwards, we normalize the image pixels in the following way: The min-max normalization function linearly scales each feature to the interval between 0 and 1. We rescale to the interval 0 and 1 by shifting the values of each feature with the minimum value being 0. Then, a division by the new maximum value is done to see the difference between the original maximum and minimum value.

The values in the column are transformed using the following formula:

$$X_{sc} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

After normalization, the data augmentation follows. In the context of deep learning, augmenting image data means using various processes to modify the original image data. We use the following augmentations: Vertical flip, horizontal flip, rotation, scaling mosaic. The most basic augmentation done is the flip augmentation. This is well suited for polyp images as the endoscope is often rotated during colonoscopy. Here, the image is flipped horizontally, vertically, or both. We use a probability of 0.3 for up and down flips and a vertical flipping probability of 0.5. In addition, we rescale the images with a probability of 0.638. Rescaling creates polyps in different sizes and therefore adds additional data to our data set. The translation moves the image along the horizontal axis. Furthermore, we apply a low probability of 0.1 to rotate the image with a random degree, e.g. 20-degree rotation clockwise. As the last augmentation step, we use mosaic data augmentation. Mosaic data augmentation merges four images into one image. Thereby, the image is rescaled, causing the images to appear in a different context. We use mosaic data augmentation with a probability of 0.944. These data augmentations are only applied to the training data.

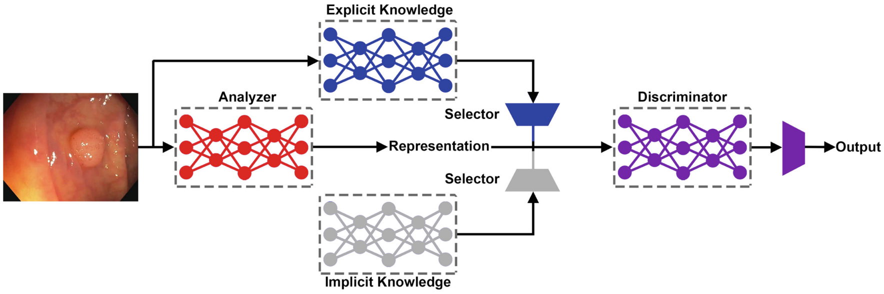


Fig. 4. Overview of the polyp detection system. Adopted from Wang et al. [31].

AI Architecture. For the AI architecture, we used a fast object detector system called YOLOR [31]. An overview of the network's architecture is illustrated in Fig. 4. Humans can look at the same piece of data from different perspectives. A trained CNN model, on the other hand, can generally only accomplish one goal. In general, the characteristics that may be recovered from a trained CNN are not well suited to other problems. The fundamental source of the aforementioned issue is that we extract features from neurons and do not employ implicit information, which is rich in CNN. YOLOR distinguishes between two different types of knowledge, implicit and explicit knowledge. The information directly corresponds to observation is referred to as explicit knowledge in the study. The authors of YOLOR call implicit knowledge that is implicit in the model but has nothing to do with observation. The implicit knowledge is represented as the deeper layers of the network and thereby contains more detail. This implicit and explicit information are especially useful in the case of real-time polyp detection as the structure of the network reduces the overall computational complexity and thereby allows the calculation to remain fast and accurate. This is an ideal scenario for real-time detection systems. To combine implicit and explicit information and enable the learned model to contain a generic representation, from which sub representations suited for particular tasks may be created, they propose a unified network, which is explained below.

What is unique about the unified network is that it can be effectively trained with implicit knowledge. To achieve this, explicit and implicit knowledge was combined to model the error term. The multi-purpose network training process can then be guided by it. This results in the following formula:

$$\begin{gathered} y = f_\theta(x) + \epsilon + g_\phi(\epsilon_{ex}(x), \epsilon_{im}(z)) \\ \text{minimize } \epsilon + g_\phi \phi(\epsilon_{ex}(x), \epsilon_{im}(z)) \end{gathered} \tag{1}$$

where ϵ_{ex} and ϵ_{im} are operations modeling the explicit error and implicit error from observation x and latent code z. g_ϕ here is a task-specific process for combining or selecting information from explicit and implicit knowledge. There are a few approaches for incorporating explicit knowledge into f_θ, now we can rewrite (1) into (2).

$$y = f_\theta(x) \star g_\phi(z) \tag{2}$$

where $\star$ is the approach used to combine f_θ and g_ϕ. In the paper, manifold space reduction and kernel space alignment are used.

Manifold space reduction uses the inner product of the projection vector and implicit representation, which is a constant tensor $Z = \{z_1, z_2, ..., z_k\}$, to reach a reduction of the dimensionality of manifold space.

Kernel space alignment deals with the frequent misalignment problem in multi-task and multi-head networks. We may add and multiply output features and implicit representations to address this issue, allowing Kernel space to be translated, rotated, and scaled to match each neural network's output kernel space.

If we expand the error term derivation procedure to several tasks, we get the following equation:

$$F(x, \theta, Z, \phi, Y, \psi) = 0 \tag{3}$$

where $Z = \{z_1, z_2, ..., z_T\}$ denotes a collection of implicit latent codes for T separate jobs, and *phi* denotes the parameters that may be utilized to build implicit representation from Z. The final output parameters are calculated using ψ from various combinations of explicit and implicit representation.

We may use the following formula to get a prediction for all $z \in Z$ for various tasks.

$$d_\psi(f_\theta(x), g_\phi(z), y) = 0 \tag{4}$$

We begin with a common unified representation $f_\theta(x)$, then go on to task-specific implicit representation $g_\phi(z)$, and eventually accomplish various tasks with task-specific discriminator d_ψ.

We assume that it starts with no past implicit knowledge to train the model. It will not influence explicit representation $f_\theta(x)$. When the combining operator

$\star$ is an addition or concatenation, the first implicit prior is $z \sim N(0, \sigma)$, and when the combining operator $\star$ is multiplication, $z \sim N(1, \sigma)$. σ is an extremely tiny number that is nearly zero. Both z and x are taught using the backpropagation method during the training procedure.

The inference is relatively simple because implicit information has no bearing on observation x. Any implicit model g, no matter how complicated, may be reduced to a collection of constant tensors before the inference phase begins. That means that it has no negligible impact on the algorithm's computational complexity.

The resulting network thus achieves a better or comparable AP than state-of-the-art methods in object detection. In addition, the network can be used in real-time, to be more precise in 39 FPS, which makes it very attractive for polyp detection in endoscopy. Further, implicit representations can be added to the output layer for prediction refinement. This leads in object detection to the fact that although one does not provide any prior knowledge for the implicit representation, the proposed learning mechanism can automatically learn the patterns (x, y), (w, h), (obj), and (classes) of each anchor.

Training Details. For the training of the AI, we first have a pretraining process. This involves the training of the AI on the human data set of over 500.000 images. In this process, we run the AI for 110 epochs on the human dataset with a learning rate of 0.001. We implemented a learning rate schedule that slowly increases the learning rate. We use a batch size of 64 images and train on four NVIDIA Quadro RTX 8000 GPUs with 48 GB RAM each. The model is trained using stochastic gradient descent. Afterward, the AI is finetuned on the animal data. We use the pretrained checkpoint of the human data trained YOLOR as initialization. As the dataset involves smaller amounts of images with lower quality the AI is only trained for 40 epochs with a batch size of 128. In this case, only two of the NVIDIA Quadro RTX 8000 GPUs are used for training. We also implemented a learning rate schedule and slowly increased the learning rate in the initialization. Nevertheless, the learning rate increase was faster than with the human data and the learning rate is set to 0.0001. The model is also trained using stochastic gradient descent.

3.4 Animal Model

An animal model was used to test our concept and obtain all the data. Four gene-targeted pigs (*sus scrofa domesticus*) with the "truncating 1311" mutation in the adenomatous polyposis coli (APC) were endoscopically examined with our system [9]. This mutation is orthologous to the hotspot APC^{1309} mutation which causes human familial adenomatous polyposis with aberrant crypt foci and low- and high-grade dysplastic adenomas in the large intestine. As shown in previous studies, the $APC^{1311/+}$ pigs are a suitable model for experimental endoscopy [26,32]. All animal experiments were approved by the Government of Upper Bavaria (permit number ROB-55.2-2532.Vet_02-18-33) and performed according to the German Animal Welfare Act and European Union Normative for Care and Use of Experimental Animals.

4 Results

This section describes our results on our own created test dataset. For our evaluation, we compare our approach to two classic benchmarking algorithms and a newer approach called YOLOv4 [5]. The benchmarking algorithms are an SSD algorithm called YOLOv3 [22], and the ROI Proposal algorithm called Faster RCNN [23]. We train all algorithms on the same data listed in the data chapter. For the test data, we create a test set. The test set consists of three videos filmed in the colon of three different pigs. As in this example, we like to evaluate the detection of the extended view, our evaluation is only done on the side cameras of the endoscope. The three videos consist of 800 frames having a frame size of 320 × 320 px.

Table 1. Evaluation on the test data set. This table shows our comparison of four different polyp detection approaches on our benchmarking data. The YOLOv3 and Faster-RCNN are baseline models, the third as a model for comparison called YOLOv4 and the last is our polyp detection system. Precision, Recall, F1, and mAP are given in %, and the speed is given in FPS.

	Precision	Recall	mAP	F1	Speed	RT capable
YOLOv3	50.21	54.57	60.52	51.98	44	Yes
YOLOv4	52.02	56.76	62.49	53.99	**47**	Yes
Faster-RCNN	57.22	62.38	67.52	59.33	15	No
Ours	**61.87**	**66.80**	**72.13**	**63.93**	39	Yes

Table 2. Detailed evaluation on the test data set. This table shows our comparison of three different polyp detection approaches on our benchmarking data. The first two models are baseline models, and the third is our polyp detection system. Precision (P), Recall (R), F1, and mAP are given in %.

Video	YOLOv4				F-RCNN				**Ours**			
	P	R	mAP	F1	P	R	mAP	F1	P	R	mAP	F1
1	54.24	52.30	65.35	53.31	61.94	57.88	69.68	59.84	68.60	65.40	74.60	66.96
2	50.75	50.13	51.47	50.42	55.44	56.63	55.11	56.03	59.10	59.40	61.70	59.25
3	51.06	67.85	70.65	58.25	54.27	72.63	77.78	62.12	57.90	75.60	80.10	65.58
Mean	52.02	56.76	62.49	53.99	57.22	62.38	67.52	59.33	61.87	66.80	72.13	63.93

Table 1 presents the results on our test set for the detection task with YOLOv4, a fast detection algorithm, and Faster R-CNN, a FASTER R-CNN algorithm with a ResNet-101 backbone. For the evaluation, we provide the

F1-score. The F1-score consists of the harmonic mean of precision and the recall, as described in the following equations:

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{Recall} = \frac{TP}{TP + FN}$$

$$F_1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 * TP}{2 * TP + FP + FN}$$

We consider an annotation as true positive (TP) when the predicted boxes and the ground truth boxes overlap at least 50%. In addition, we disclose the mean average precision (mAP) and the mAP50 with a minimum IoU of 0.5 [19]. We calculate the mAP using the integral of the area under the precision-recall curve. Thereby, all predicted boxes are first ranked by their confidence value given by the polyp detection system. Afterward, we compute precision and recall with different thresholds of the confidence values. When the confidence threshold is reduced, the recall increases while the precision decreases, resulting in a precision-recall curve. Finally, we measure the area under the curve precision-recall to receive the mAP value.

Table 2 presents a more detailed view of our results, showing the performance for every test video. Table 1 shows that our approach is outperforming classical benchmarks on our test data. Our approach increases the detection results by 4.6 % points compared to the F-RCNN algorithm. This is due to the architecture of the YOLOR algorithm, which allows fast but still accurate detections. Notably, the algorithm YOLOv4 is still 8 FPS faster than our approach to detect single images. Nevertheless, our approach yielded a huge recall increase of 12.23 % points compared to the fast YOLOv3 and 10.04 % compared to YOLOv4. A recall increase is beneficial for clinical practice as examiners care more about finding a missed polyp than getting distracted by a false positive detection. Figure 5 show a sequence of detection results with our algorithm on the test dataset provided.

Found Polyps Through Extended Vision: To test our extended vision user interface, we tried to test if polyps were missed by the classic front view endoscope but found them through a side camera of the extended view. Therefore we compared the detected polyps of our test set with the annotations of our polyps on the main view camera endoscopy. We then checked how many polyps were found through the extended view. We did this by comparing the polyp detections in the classic front view of the endoscope with the detections of the side camera. If there was no detection in the main camera before a true positive detection in the side camera appeared, we counted the polyp as being missed by the classic detection system but detected by our system with extended vision. Overall, the main view detected 84 different polyps in the test data. 13 polyps in the extended view were not seen in the classic endoscope view.

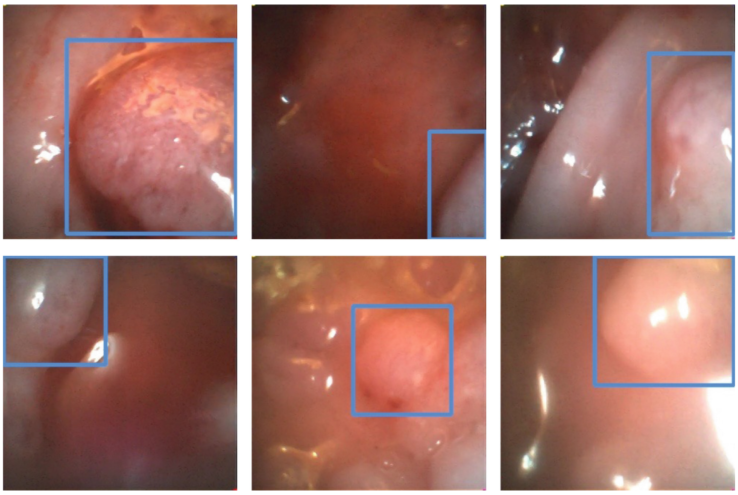


Fig. 5. Detection results. This figure shows some of the true positiv detection results of the side cameras used for extended vision. After the detection the examiner can visualize the polyp with the front camera.

5 Discussion

This chapter shows the limitations of our interface. We mainly show a failure analysis of our system, as well as showing potential risks and challenges that have to be addressed in future work. As our tests and data are based on animal trials, we also discuss the system's clinical application in human interventions.

5.1 Limitations

Fig. 6. Examples of errors in video 17 of the CVC-VideoClinicDB data set. The left image shows a clean detection of a rather flat polyp. The middle image shows a miss of the same polyp due to being blocked by a colon wall, while the right shows a (short) re-detection.

We initiate the discussion of our limitations with failure analysis of our model. First, we refer to Tables 1 and 2, specifically to video 2 which has significantly worse performance compared to the rest. Therefore the examples we choose for the failure analysis are exclusively from video 2. Nevertheless, they differ as some polyps are harder to detect than others. Different lighting in a camera place of the endoscope especially influences the side cameras' detection results (extended view). In addition, bad contrast, diagonal angles, and unusual shapes do enhance the detection difficulty. Hence, multiple reasons can be attributed to the worse performance in some situations.

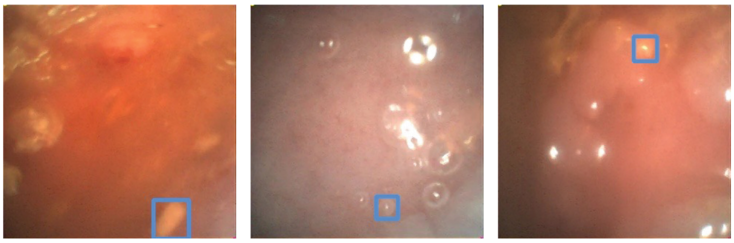


Fig. 7. Examples of errors in video 17 of the CVC-VideoClinicDB data set. The left image shows a clean detection of a rather flat polyp. The middle image shows a miss of the same polyp due to being blocked by a colon wall, while the right shows a (short) re-detection.

E.g., contrast and lighting are one of the main causes of missing or misidentifying a polyp. This is especially true with our extended vision, as the examiner does not see the side cameras. The view of the side cameras is impacted higher by bad lighting conditions. Figure 6 shows some of these bad lighting condition. The right polyps can not be detected because there is no light to make the polyp appear clear on the camera. In the image in the middle, the lighting is reflected very bright. This may be due to the camera being too close too the mucosa. Sometimes those lighting conditions cause FP detection, as seen in the last image on the left side.

Additionally, many FPs created by our system are due to feces and bubbles. Feces are high in contrast, and some polyps are too. Therefore, the neural network is making FP detections, as seen in the left picture of Fig. 7. The FP detection is set on the lower part of the screen; nevertheless, the top of the screen shows a polyp covered by feces and, therefore, is hard to detect. The algorithm is blinded by the lower feces and can not detect the polyp. Another problem is bubbles. Often, the endoscopist has to clean the bowel with water. While doing so, there are constantly emerging bubbles. The detection system sometimes detects these bubbles as their shape may be similar to the shape of polyps.

For clinical use, expanding the examiner's view results in more detected polyps. Therefore, such a system could help the examiner during an actual intervention. Nevertheless, we could only show new detections in animal examples. Our user interface can help the examiner without having to change classical procedures. Nevertheless, first the endoscope with extended vision has to be developed to apply to humans and tested there in future work.

6 Conclusion

We present a prototype that maintains the examiner's classical endoscope view but extends the endoscope with side cameras and AI polyp detection. This AI system detects polyps on the side cameras and alarms the examiner if a polyp is found. The prototype is created by adding two micro cameras to the sides of a classic endoscope. The AI system is trained on human data and fine-tuned with animal data. Then we test the prototype with gene-targeted pigs. The AI outperforms current benchmarks and finds polyps by adding the extended vision to the system. Nevertheless, there are limitations to the system. First, the position and light condition of the side cameras have a high impact on the detection results. If light conditions are bad or cameras are too close to the mucosa, the system cannot detect polyps. Second, the system sometimes detects bubbles, feces, or light reflections as polyps. Third, the system is not ready for clinical interventions in humans. Further development and medical product tests have to be done to allow the system to be applied to the human body.

References

1. Ahn, S.B., Han, D.S., Bae, J.H., Byun, T.J., Kim, J.P., Eun, C.S.: The miss rate for colorectal adenoma determined by quality-adjusted, back-to-back colonoscopies. Gut Liver **6**, 64–70 (2012). https://doi.org/10.5009/gnl.2012.6.1.64
2. Ali, S., et al.: Endoscopy disease detection and segmentation (edd2020) (2020). https://doi.org/10.21227/f8xg-wb80
3. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized Medical Imaging and Graphics 43, 99–111, July 2015. https://doi.org/10.1016/j.compmedimag.2015.02.007
4. Bernal, J., Sánchez, J., Vilarino, F.: Towards automatic polyp detection with a polyp appearance model. Pattern Recogn. **45**(9), 3166–3182 (2012)
5. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
6. Colucci, P.M., Yale, S.H., Rall, C.J.: Colorectal polyps. Clin. Med. Res. **1**(3), 261–262 (2003)
7. Fabrice Bellard, F.t.: Ffmpeg 4.4 (2000). http://www.ffmpeg.org/, [Online; Stand 25.03.2022]
8. Favoriti, P., Carbone, G., Greco, M., Pirozzi, F., Pirozzi, R.E.M., Corcione, F.: Worldwide burden of colorectal cancer: a review. Updat. Surg. **68**(1), 7–11 (2016). https://doi.org/10.1007/s13304-016-0359-y
9. Flisikowska, T., et al.: A porcine model of familial adenomatous polyposis. Gastroenterology **143**(5), 1173–1175 (2012)
10. Gralnek, I.M., et al.: Standard forward-viewing colonoscopy versus full-spectrum endoscopy: an international, multicentre, randomised, tandem colonoscopy trial. Lancet Oncol. **15**(3), 353–360 (2014)
11. Hassan, C., et al.: Full-spectrum (fuse) versus standard forward-viewing colonoscopy in an organised colorectal cancer screening programme. Gut **66**(11), 1949–1955 (2017)

12. Heresbach, D., et al.: Miss rate for colorectal neoplastic polyps: a prospective multicenter study of back-to-back video colonoscopies. Endoscopy **40**(04), 284–290 (2008). https://doi.org/10.1055/s-2007-995618
13. Inc, E.S.: Epiphan dvi2usb 3.0. https://www.epiphan.com/products/dvi2usb-3-0/tech-specs/, [Online; Stand 25.03.2022]
14. Intel Corporation, Willow Garage, I.: Opencv (2000). https://opencv.org/, [Online; Stand 25.03.2022]
15. Jha, D., et al.: Kvasir-SEG: a segmented polyp dataset. In: Ro, Y.M., Cheng, W.-H., Kim, J., Chu, W.-T., Cui, P., Choi, J.-W., Hu, M.-C., De Neve, W. (eds.) MMM 2020. LNCS, vol. 11962, pp. 451–462. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-37734-2_37
16. Krenzer, A., et al.: Fast machine learning annotation in the medical domain: a semi-automated video annotation tool for gastroenterologists (2021)
17. Lambert, R.F.: Endoscopic classification review group. update on the Paris classification of superficial neoplastic lesions in the digestive tract. Endoscopy **37**(6), 570–578 (2005)
18. Leufkens, A., van Oijen, M., Vleggaar, F., Siersema, P.: Factors influencing the miss rate of polyps in a back-to-back colonoscopy study. Endoscopy **44**(05), 470–475 (2012). https://doi.org/10.1055/s-0031-1291666
19. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
20. Ltd, B.D.P.: Blackmagic - decklink mini recorder 4k. https://www.blackmagicdesign.com/pl/products/decklink/techspecs/W-DLK-33, [Online; Stand 25.03.2022]
21. Misawa, M., et al.: Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). Gastrointestinal Endoscopy **93**(4), 960–967 (2021)
22. Redmon, J., Farhadi, A.: Yolov3: an incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
23. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems 28 (2015)
24. Rex, D., et al.: Colonoscopic miss rates of adenomas determined by back-to-back colonoscopies. Gastroenterology **112**(1), 24–28 (1997). https://doi.org/10.1016/s0016-5085(97)70214-2
25. van Rijn, J.C., Reitsma, J.B., Stoker, J., Bossuyt, P.M., van Deventer, S.J., Dekker, E.: Polyp miss rate determined by tandem colonoscopy: a systematic review. Am. J. Gastroenterol. **101**(2), 343–350 (2006). https://doi.org/10.1111/j.1572-0241.2006.00390.x
26. Rogalla, S., et al.: Biodegradable fluorescent nanoparticles for endoscopic detection of colorectal carcinogenesis. Adv. Func. Mater. **29**(51), 1904992 (2019)
27. Silva, J., Histace, A., Romain, O., Dray, X., Granado, B.: Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. Int. J. Comput. Assist. Radiol. Surg. **9**(2), 283–293 (2014)
28. Triadafilopoulos, G., Li, J.: A pilot study to assess the safety and efficacy of the third eye retrograde auxiliary imaging system during colonoscopy. Endoscopy **40**(06), 478–482 (2008)

29. Vázquez, D., et al.: A benchmark for endoluminal scene segmentation of colonoscopy images. J. Healthcare Eng. **2017**, 1–9 (2017). https://doi.org/10.1155/2017/4037190
30. Vázquez, D., et al.: A benchmark for endoluminal scene segmentation of colonoscopy images. J. Healthcare Eng. **2017**, 4037190 (2017). https://doi.org/10.1155/2017/4037190
31. Wang, C.Y., Yeh, I.H., Liao, H.Y.M.: You only learn one representation: Unified network for multiple tasks. arXiv preprint arXiv:2105.04206 (2021)
32. Yim, J.J., et al.: A protease-activated, near-infrared fluorescent probe for early endoscopic detection of premalignant gastrointestinal lesions. Proceedings of the National Academy of Sciences 118(1) (2021)

RESEARCH

Automated classification of polyps using deep learning architectures and few-shot learning

Adrian Krenzer[1,2*], Stefan Heil[1], Daniel Fitting[2], Safa Matti[1], Wolfram G. Zoller[3], Alexander Hann[2] and Frank Puppe[1]

[*]Correspondence:
adrian.krenzer@uni-wuerzburg.de
[1]Department of Artificial Intelligence and Knowledge Systems, Julius-Maximilians University of Würzburg, Sanderring 2, 97070 Würzburg, Germany
Full list of author information is available at the end of the article

Abstract

Background: Colorectal cancer is a leading cause of cancer-related deaths worldwide. The best method to prevent CRC is a colonoscopy. However, not all colon polyps have the risk of becoming cancerous. Therefore, polyps are classified using different classification systems. After the classification, further treatment and procedures are based on the classification of the polyp. Nevertheless, classification is not easy. Therefore, we suggest two novel automated classifications system assisting gastroenterologists in classifying polyps based on the NICE and Paris classification.

Methods: We build two classification systems. One is classifying polyps based on their shape (Paris). The other classifies polyps based on their texture and surface patterns (NICE). A two-step process for the Paris classification is introduced: First, detecting and cropping the polyp on the image, and secondly, classifying the polyp based on the cropped area with a transformer network. For the NICE classification, we design a few-shot learning algorithm based on the Deep Metric Learning approach. The algorithm creates an embedding space for polyps, which allows classification from a few examples to account for the data scarcity of NICE annotated images in our database.

Results: For the Paris classification, we achieve an accuracy of 89.35 %, surpassing all papers in the literature and establishing a new state-of-the-art and baseline accuracy for other publications on a public data set. For the NICE classification, we achieve a competitive accuracy of 81.13 % and demonstrate thereby the viability of the few-shot learning paradigm in polyp classification in data-scarce environments. Additionally, we show different ablations of the algorithms. Finally, we further elaborate on the explainability of the system by showing heat maps of the neural network explaining neural activations.

Conclusion: Overall we introduce two polyp classification systems to assist gastroenterologists. We achieve state-of-the-art performance in the Paris classification and demonstrate the viability of the few-shot learning paradigm in the NICE classification, addressing the prevalent data scarcity issues faced in medical machine learning.

Keywords: Machine learning; Deep learning; Endoscopy; Gastroenterology; Automation; Image Classification; Transformer; Deep metric learning; Few-shot learning

Background

Colorectal cancer (CRC) is the second leading cause of cancer-related deaths worldwide [1]. This cancer develops from lesions inside the colon called polyps. However, not all colon polyps have the risk of becoming cancerous. Therefore, polyps are classified using different classification systems. After the classification, further treatment and procedures are based on the classification of the polyp. Since young physicians often do not have the necessary experience to make the correct decision reliably, computer-assisted procedures are being developed that can assist with the classification.

In the field of automated gastroenterological assistance systems, a significant area of research involves the detection of polyps using deep learning. Polyps are mucosal growths in various body parts, such as the intestine or stomach. In some cases, unusual skin changes can become dangerous and even cancerous. Deep Learning object recognition methods such as CNNs detect and classify polyps automatically during examinations to assist endoscopists [2–4]. This may be beneficial for the future, to detect polyps more accurately by automated methods and to simplify or confirm the prognosis for the proper polyp treatment.

The polyp classification is essential as it helps the endoscopist decide on further treatment methods. For classification, different approaches are used to categorize polyps, such as schemes based on the shape (PARIS) [5] or based on the surface structure (NICE) [6]. The classification of polyps can give first insights into their dangerousness and the appropriate treatment options [5]. Furthermore, van Doorn et al. demonstrated a moderate interobserver agreement among Western international experts for the Paris classification system. Automated classification systems could help increase experts' interobserver agreement on the Paris classification [7].

We consider the Paris and the NICE classification for our automated classification algorithms as they are the most commonly used classification in Europe. Furthermore, the Paris classification is recommended for documentation in the ESGE European Society of Gastrointestinal Endoscopy guidelines and it is also recommended to use advanced endoscopic imaging like NBI [8].

This paper shows therefore two automated classification networks. The first is classifying the polyp based on white light using the Paris classification scheme [5]. A two-step process is introduced: first, detecting and cropping the polyp on the image, and secondly classifying the polyp based on the cropped area with a transformer network. Figure 1 shows some example results of the Paris polyp classification system.

The second is the NICE classification, which is based on Narrow band imaging (NBI). NBI is a variation of endoscopy that uses blue and green light to enhance the visibility of surface patterns and texture of the mucosa. The presented NICE classification system is designed as a Deep Metric Learning based approach of few-shot learning to account for the data scarcity of NICE annotated images in our database.

In the following, the main contributions of the paper are shown:

1) *We introduce a Paris classification system with state-of-art performance on clinical data.*

2) *We created a data set of polyp classification data to train and further enhance the models.*

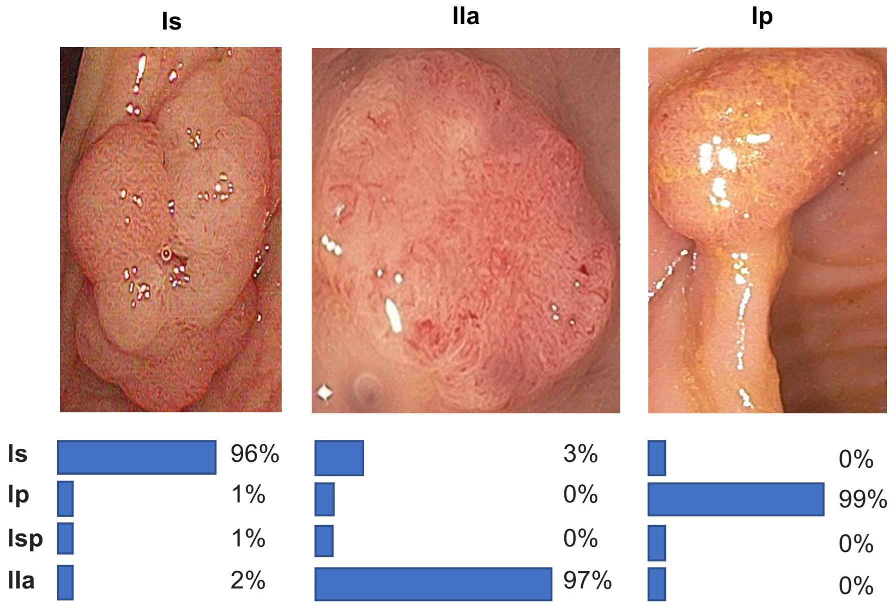


Figure 1 Automated Paris classification examples. This figure illustrated some classification examples of the polyp classification system on our data (EndoData) [9]. The percentage values show the confidence of the classification system.

3) *We present and validate a new approach for the automated NICE classification in data scarce scenarios leveraging few-shot learning.*

Additionally, both polyp classification systems were publicly funded and developed by computer scientists and endoscopists in the same workgroup to ensure the high quality of the polyp classifications. In the next subsection a summary of the medical classification methods of polyps will be given. Furthermore, to overview existing work and properly allocate our paper to the literature, we describe a brief history from general polyp detection to state-of-the-art polyp classification with deep learning techniques.

Medical backgroud

Polyps are small, fungal, or flat mucosal growths in various body regions, such as the intestines, stomach, uterus, or nose. The different-looking skin lesions most commonly occur in the stomach or intestines and affect in particular older people. They often appear after inflammation, leading to higher cell division in the mucosa. Additionally, polyps can become malignant or even cancerous due to unusual cell growth. Polyps can be divided into three types: hyperplastic, neoplastic, and inflammatory. While the hyperplastic and inflammatory types have no or lower risk of degeneration, the neoplastic polyps represent the most dangerous type of polyp. These can increase the risk of cancer, especially as they grow. In order to prevent a severe progression due to polyps, repeated examination by an endoscopist through endoscopy is necessary. In this process, hollow organs such as the intestine are examined with an endoscope, a flexible tube equipped with a camera, and light.

Paris classification In order to categorize polyps and to select appropriate treatment strategies, polyps are classified considering various aspects. One of the most widely used classifications is the Paris classification. Based on a Japanese classification scheme, the Paris classification characterizes the potentially high-risk polyps according to their shape [5]. Figure 2 visualizes the shapes of different polyps:

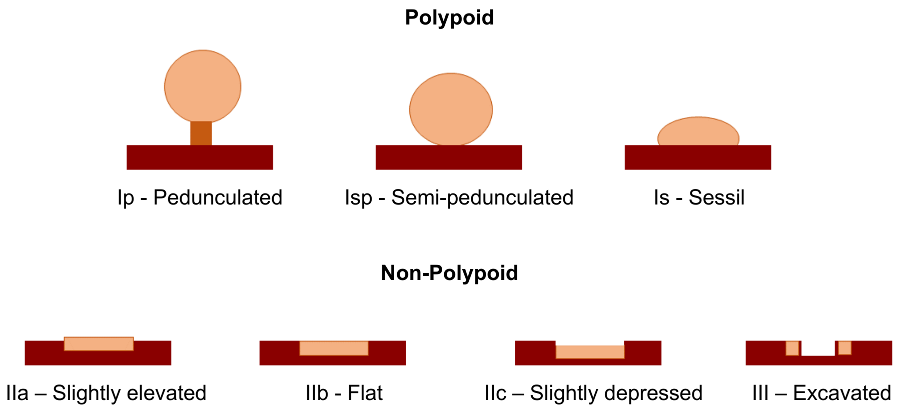


Figure 2 Subdivision of polyps according to Paris classification. Adopted from [5].

Type I polyps are referred to as elevated or polypoid. A distinction is made between the following polyp types:

- Ip Pedunculated
- Isp Semipedunculated
- Is Sessile

Type II polyps are described as flat. In addition, the following distinctions are made:

- IIa Slightly elevated
- IIb Completely flat
- IIc Depressed

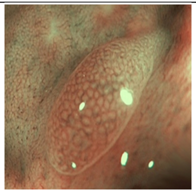
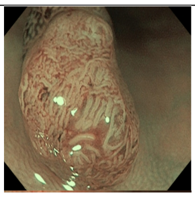
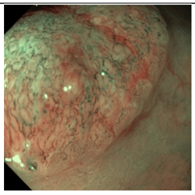
Furthermore, lastly, type III describes the excavated form. Unlike type I, type II and III are not considered polypoid. A prognosis can be obtained through the Paris classification to conclude the type of polyp, and future treatment [5]. The Paris classification is sometimes given in the literature with a preceding 0 before the type. As the preceding is irrelevant to our approach, the leading zero is omitted for clarity.

NICE classification The NICE classification is an established diagnosis scheme classifying polyps into three categories, which specify the most likely pathology ranging from benign hyperplastic to cancerous polyps deeply invading the mucosa underneath the polyp.

The scheme hereby utilizes the Narrow-Band-Imaging technology (NBI) to render the surface texture visible and to characterize the different polyp classes according to features such as the vessel patterns discernible on the polyp surface [10]. An overview of the different NICE classes[1], their characteristics and most likely pathology can be seen in table 1.

The NICE classification has been well established as an informative feature for the classification of polyps ([11], [12]) and the clinical performance of the scheme, as well as the classification performance of human experts using the scheme, have been subject to numerous studies ([10], [12]). In the treatment assessment guideline of the European Society of Gastrointestinal Endoscopy, the degree of submucosal invasion is a decisive criterion for the requirement of surgical removal of neoplastic polyps [13].

Table 1 Overview of the NICE categories. Adopted from Endoscopy-Campus GmbH[1].

	Typ 1	**Typ 2**	**Typ 3**
Color	Same or lighter than background	Browner than background	Brown to dark
Vessels	None or isolated lacy vessels	Brown vessels around white structures	disrupted or missing vessels
Surface	Dark or white spots or homogeneous	Oval, tubular or branched white strucutres	amorphous or absent patterns
Likely pathology	**hyperplastic**	**Adenoma**	**Deep submucosal invasive cancer**
Examples			

A brief history of automated polyp classification

This section gives a brief overview of the current state of the art in automated polyp detection and classification research with respect to deep learning methods. Here, there are mainly two ways deep learning methods can be used to assist gasteroenterlogists with the assessment of polyps. On the one hand, for detecting polyps in polyp images or videos to find them as early as possible. On the other hand, classifying polyps to categorize them and do a proper treatment and analysis. Classifying polyps is based on various superficial features such as shape or structure. In this context, the detection and classification of polyps can be challenging due to numerous aspects.

Since this decade, deep learning has been the leading technology in developing computer-aided polyp detection. Most methods do use Convolutional Neural Networks (CNNs) for the detection of polyps. E.g Zhu et al. show a seven-CNN paired with a support vector machine (SVM) to detect anomalies in endoscopy images Zhu.2015. Another paper is the paper by Zhang et al., which presents a CNN for

---

[1] https://www.endoscopy-campus.com/en/classifications/polyp-classification-nice/

polyp detection and localization. They use a single-shot multibox detector that reused shifted information through max-pooling layers to achieve higher accuracy. They achieved a real-time detection speed of 50 frames per second (FPS) and an average accuracy of 90.4 [14]. Another idea from Bagheri et al. used sophisticated preprocessing involving the colors of the images to correlate the information to locate and segment polyps. In this way, their polyp detection achieved 97.7 % accuracy on the CVC-ColonDB dataset [15]. Another approach from Qadir et al. utilizes a two-step method. In the first step, they used a CNN that generated multiple regions of interest (RoIs) that are then used for classification. These proposed RoIs were compared with subsequent frames and their RoIs. The rationale of this method is that the frame in a video should be similar to the next frame, and this is to reduce the percentage of false predictions.

Sornapudi et al. also utilized region-based CNNs to localize polyps in colonoscopy images but in wireless capsule endoscopy (WCE) images. Therefore, the detection is not done in real-time. During localization, images were segmented and detected based on polyp-like pixels Sornapudi.2019. Currently, also transformer architectures are relevant for polyp detection. For example, a particular sparse autoencoder method called stacked sparse autoencoder with image manifold constraint has been used by Yuan and Meng [16] to detect polyps in WCE images. A sparse autoencoder is an artificial neural network commonly used for unsupervised learning methods [17]. Their approach achieved an accuracy of 98 % in polyp detection [16]. Another approach used transformers in combination with CNNs. Zhang et al. used the ability to view global information of the whole image through the attention layers of transformers and the detailed local detection of CNNs to segment polyps efficiently. They used a new fusion technique called BiFusion to connect the features obtained from the transformers and the CNNs. The method ran in real-time with 98.7 FPS [18].

Not only the localization of polyps represents a goal of computer-specific polyp research, but also the classification according to specific characteristics. For example, Ribeiro et al. used the feature extraction capability of CNNs to classify polyps into "healthy" (average) and "abnormal" (adenoma) classes using Kudo's pit-pattern classification. Pit-pattern classification is a variant of categorizing types of polyps based on their surface structure [19]. The authors achieved an accuracy of 90.96 % by their classification using the CNN [20].

Using pit-pattern classification, a deep learning model was presented in the paper [21] to classify polyps into ,,Benign," ,,Malignant," and ,,Nonmalignant. Here, the model was trained with a private data set and achieved reliability of 84 %. Another popular polyp classification method using a CNN is used in [22]. Here, the authors used the Narrow-Band Imaging International Colorectal Endoscopic Classification (NICE for short) [6], similar to pit-pattern classification using surface features. Here, however, the polyps were additionally categorized by color or vascular structure and classified as polyp type 1 or 2. Thus, a preliminary prognosis can be determined whether the polyp is a hyperplastic or an adenoma tumor. For classification, the authors used a CNN with an SVM. The CNN was pre-trained on a non-medical data set to compensate for the lack of polyp data. They achieved an accuracy of nearly 86 % [22] with their proposed model.

Bryne et al. also used the NICE classification to characterize polyps. They classified them as hyperplastic or adenoma polyps. The authors created a CNN model for real-time application, which was trained and validated using only narrow-band imaging (NBI) video frames. In doing so, they achieved an accurate prediction of 94 % [23] on a sample of 125 testing polyps. Furthermore, Komeda et al. presented a specific CNN model to classify polyps into "adenoma" and "non-adenoma" polyps based on NBI and white-lighted images [24]. In the paper by Lui et al., another automatic classification model is presented to characterize polyps into endoscopically curable lesions and noncurable lesions based on the NBI and white-lighted images. The division into curable and noncurable is based on the types of polyps, such as hyperplastic or tubular. Lui et al. achieved an overall accuracy of 85.5 % with their model, with higher performance on NBI images [4]. In addition, Ozawa et al. used a CNN based on a single-shot multibox detector to detect and classify polyps. They trained and validated the model with a non-public data set and achieved a true-positive rate of 92 % during detection and characterized the detected polyps with an accuracy of 83 % [3]. In 2021, Hsu et al. considered the classification of polyp pathology using gray scale images and a customly designed classification network embedded into a detection and classification pipeline. They achieved an accuracy in the decision between neoplastic or hyperplastic polyps of 82.8% using NBI and 72.2% using white light [25]. An overview over the methods discussed here is presented in table 2.

Regarding the NICE classification, our work can be considered as a polyp classification system categorizing the polyps into the classes hyperplastic and adenoma according the pathological interpretation of the NICE classes I and II. The same methodology has already been applied in the mentioned works in [22] and [23], and we consider therefore the literature outlined in this section as the peer group of our work. But in contrast to most of the previous works, which learn a blackbox pathology classification system, we aim to factorize the pathological assessment by embedding the classifications into the previously introduced well-established classification schemes Paris and NICE, in order to make the pathology assessments more explainable. Instead of the prediction of the pathology directly, we therefore make the prediction of the NICE and Paris class of a polyp to the subject of our study.

To the best of our knowledge, just one similar approach concerning the Paris classification has been published [2]. Bour et al. trained several well-known CNN architectures to classify polyps based on shape. The polyp images were divided into "Not Dangerous", "Dangerous" and "Cancer" concerning the Paris classification. They labeled the Paris classes Is, Ip, Isp, IIa and IIb as "Not Dangerous", class IIc as "Dangerous" and class III as "Cancer". Their algorithms are trained on 785 images. They achieved an accuracy of 87.1 % with ResNet50 as backbone [2].

Table 2 Related methods occupied with the pathological assessment of colorectal polyps.

Author	Year	Method	Data	Classification	Accuracy
Ribeiro et al. [20]	2016	custom CNN	private	healthy abnormal	90.96 %
Zhang et al. [22]	2016	CaffeNet	private and [26]	hyperplastic adenoma	85.9 %
Bryne et al. [23]	2017	InceptionNet	private	hyperplastic adenoma	94 %
Komeda et al. [24]	2017	custom CNN	private	adenoma non-adenoma	75.1 %
Lui et al. [4]	2019	custom CNN	private	curable non-curable	85.5 %
Bour et al. [2]	2019	ResNet-50	private	not dangerous dangerous cancer	87.1 %
Tanwar et al. [21]	2020	VGG-16	private	Benign Malignant Nonmalignant	84 %
Ozawa et al. [3]	2020	SSD	private	hyperplastic adenoma	83 %
Hsu et al. [25]	2021	custom CNN	private	hyperplastic neoplastic	72.2 % (Weight light) 82.8 % (NBI light)

Data and methods

The following chapter describes the methodology of this paper. The section starts with outlining the data sets used for the training process. Furthermore, the chapter involves one section for the methodology of the Paris classification and one section for the NICE classification. For the Paris classifcation, we use a two-step process involving first the detection of the polyp and the cropping of the image to the region of the detected polyp. In a second step, the cropped polyp is provided to a transformer architecture to classify it. For the NICE classification, we deploy a metric learning CNN pre-trained on a texture transfer learning and a self-supervision data set, which is subsequently fine-tuned on the extracted and cropped polyp images.

Data sets

The current chapter will outline the data sets involved in the training of the NICE and Paris classification systems, which were compiled from different sources.
Due to the data sets containing only a subset of the required annotation types (NICE or Paris), the sources for the two classification tasks only partially overlapped.

Paris

For the training and evaluation of the Paris classification system, we used two data sets. The first is an open-source data set called SUN (Showa University and Nagoya University) colonoscopy video data set. The Sun Colonoscopy Video data set consists of approximately 160,000 images, of which approximately 50,000 images contain polyps. Other open source polyp data sets do mostly not attain the Paris classification type. The polyp images contain 100 different polyps annotated by experienced endoscopists from the Showa University. The distribution of the images among the polyp types can be found in the table 3 [27]. Because only polyp images are needed for this work, polypless images were sorted out. Since the images in the data set are single video frames, images that were too small or blurred with unrecognizable content were removed manually to train the networks on recognizable images.

Table 3 Distribution of the images in the SUN Colonoscopy Video data set [27].

Type of polyp	Number of polyps by type	Number of images by polyp type
Is	49 cases	23.154 images
Ip	8 cases	4.162 images
Isp	9 cases	4.684 images
IIa	34 cases	17.136 images

The second data set is EndoData this was created by us at the University clinic of Würzbug [9]. In the next section the proccess of the data creation will be outlined briefly.

Own data creation Previously we created a framework for faster endoscopic annotation. It involves a two-step process. First, a small expert annotation part and then a large non-expert annotation part [28]. Thereby shifting most of the workload away from the expert to the non-expert while retaining high data quality. We combined both tasks using AI to increase the annotation speed further. To speed up is up to 20 times compared to a traditional annotation tool. Thereby the process is divided between at least two people. First, an expert watches the video and labels some video frames to verify the object labeling. In the second step, a non-expert receives a visual confirmation of the given object and can label all following and preceding frames with AI support. In order to label individual frames, all of the frames have to be extracted from the video. Our system is then pre-selecting relevant frames automatically.

Thereby experts can focus on those keyframes. After the expert completes his annotations, the AI model gives the relevant frames. The AI is then detecting the polyps in the image and pre-labeling those. The non-expert can adjust and modify the AI predictions and use them for training the AI model.

In addition, the expert annotates the Paris and, if possible, the NICE classification [5], the size of the polyp and its position, as well as the start and end image of the polyp and a box for the non-expert annotators. Afterward, Endodata [9] is filtered and the relevant Paris and NICE classification parts are extracted to create the final data set used in this paper.

We assembled a team of experienced gastroenterologists and medical assistants to create this data set. The EndoData data set contains 79,625 images with Paris classification involving 364 polyp sequences. The polyp sequences were selected in high quality because we usually annotated only the first 1-3 seconds of polyp appearance, which is critical for polyp detection in a real clinical scenario. We only used the NBI light images and videos from the Olympus processor for the NICE classification.

NICE

As the SUN database does not contain NICE class annotations and little data with a direct NICE annotation is publicly available, only a very limited data set of NICE annotated colorectal polyps was available for this study, comprising the images of not more than 61 different polyps. The data set contained polyp images of two different sources, namely the examples provided for the different NICE classes curated on the Endoscopy Campus [2] and images extracted from the closed source endoscopic data set of the University of Würzburg, which were annotated by an expert gastroenterologist. As the data from the Endoscopy Campus provides only a single image per polyp and the usable frames of a specific polyp in the closed source data were nearly identical, the data set has been constructed to contain only a single image for each polyp.

Due to a lack of data, the third category of the NICE classification scheme has been dropped and the study focuses on the prediction of the first two classes, corresponding in the canonical interpretation to the two classes of hyperplastic and adenomatous polyps. Similar restrictions have already been made in other studies, such as in [2], discussed in the related work of this study. The data set comprises overall 27 images of class NICE II polyps and 34 images of class NICE I.

Due to the data set containing only a single image per polyp, the splits of the data set were disjoint concerning the contained polyp specimens and did not introduce any immediate or latent correlations between training and testing data.

As preprocessing measures, the images were cropped to the polyp region and down- or upsampled to a common shape of 224 × 224. The images have not been made subject to further preprocessing methods.

Paris classification

The first classification method will focus on the Paris classification using white light endoscopy. The following subsection will illustrate the automated NICE classification.

Reason for leaving out classes of the Paris classification As explained earlier polyps are divided into polypoid and non-polypoid in the Paris classification. Type I polyps are polypoid, and type II and III polyps are non-polypoid. Due to the composition of available data, only Is, IIa, Ip, and Isp forms were considered and used to classify polyps. Here, Is denotes the sessile type, IIa the flat raised polyps, Ip a pedunculated form, and Isp the semi-pedunculated polyps [5]. We do not have any data examples for the Paris categories IIb, IIc, and III in our data and the open source SUN data set. This may be due to the acquisition of most of the data from screening coloscopies where Paris types IIb, IIc and III are very rare. Therefore we had to remove those categories in our classification model. By classifying polyps into different types, it is also possible to make statements about the probability of a polyp being cancerous. In one study, it was shown that certain types in the Paris classification can lead to an increase in submucosal invasion. This correlates with a greater risk of developing lymph node metastases from polyp disease in the stomach, which may lead to a poorer prognosis. This revealed that polypoid type I (57 %) and types IIc (37 %) and III (40 %) had a higher risk of submucosal invasion. In comparison, forms IIa and IIb (29% and 20%) showed a lower probability of [5, 29].

Since the images in the data set are single video frames, images that were too small or blurred with unrecognizable content were removed manually to train the networks on recognizable images. Finally, the obtained images were prepared for the models and examined with respect to resolution.

---

[2] https://www.endoscopy-campus.com/en/classifications/polyp-classification-nice/

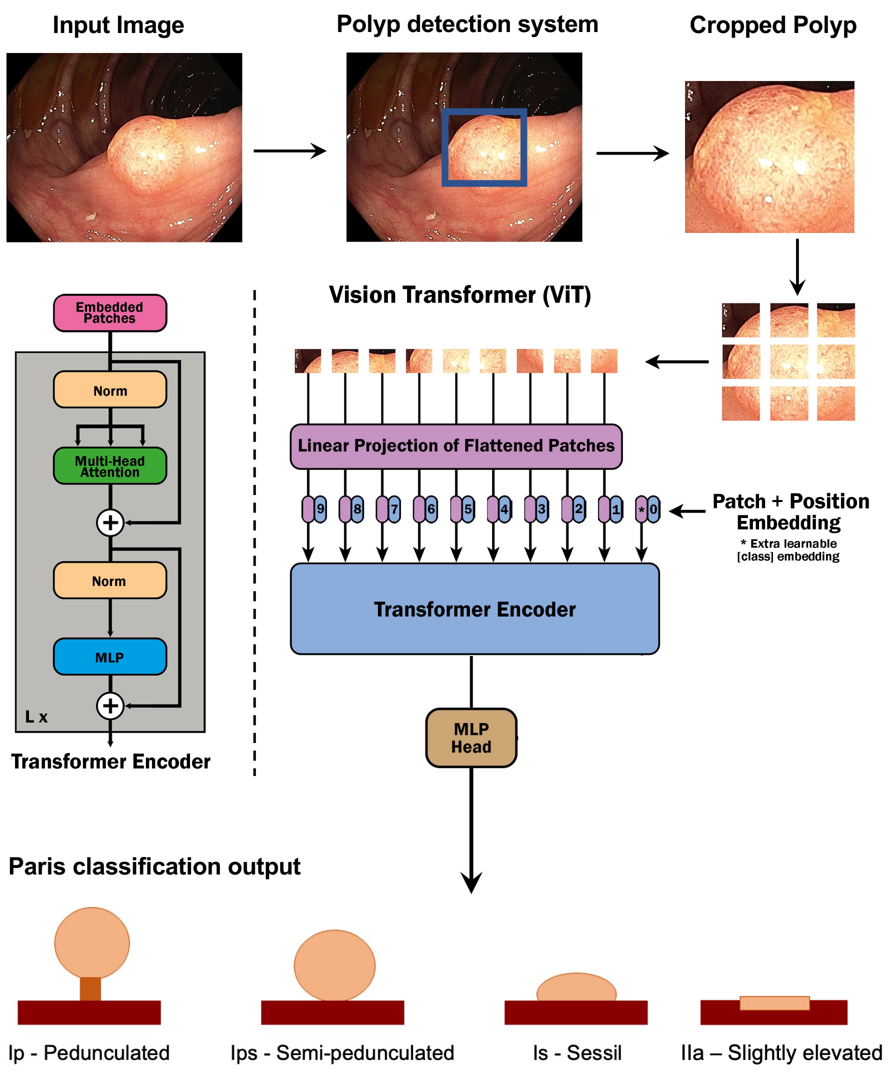


Figure 3 Structure of the polyp classification system. Adopted from [30]. Polyp images are from our data (EndoData) [9].

Figure 3 outlines the structure of our polyp classification system. At the left site, you can see the photo taken from the endoscope processor, which was done after finding the polyp. This image is the input image to our system, and the next step is the polyp detection system. For the polyp detection system, we used ENDOMIND-Adcanced [9], which is a polyp detection system. The system was developed by us using a post-processing technique based on video detection to work in real-time with a stream of images. This allows leveraging the incoming stream context of the endoscope while maintaining real-time performance. The system, therefore, can predict a bounding box surrounding the polyp. In the next step, the image is cropped at the box corners. The background, which is unnecessary for the classification, is cropped so that the polyp is better processed by the following classification step. In the classification step, the resulting polyp image is inserted into the Vision transformer (ViT) [30].

The use of transformers in computer vision is a relatively new field but is a significant competitor to CNNs. The paper Vision Transformer (ViT) introduces the use of transformers in the image processing domain without using a CNN. The Vision Transformer is based on a classical transformer for NLP, which has been adapted for the computer vision task. The input image is brought into fixed-size image sections, also called patches, as visualized in Figure 3. Then, the image patches are passed to the transformer as a sequence, like a sentence sequence. The image sections are converted into computable vectors for the transformer using the patch embedding layer. Furthermore, the positions of the image sections are marked by Positional Embedding, as in a classical Transformer. In addition, a learnable classification token is added. The prepared sequence is then passed to one or more standard Transformer encoders. Unlike the classical transformer, the ViT model does not have a decoder, but a MLP head linked to the previous layers for classification [30, 31]. For pretraining the vision transformer, a large data set is used. For fine-tuning, the pre-trained classification part, the MLP Head, is then removed and replaced by a feed-forward layer specified for the desired task and adapted [30].

The developers of ViT provide three different transformer models for image classification: ViT-Base (12 encoder layers), ViT-Large (24 encoder layers), and ViT-Huge (32 encoder layers), which are available in the following variants: ViT-B/16, Vit-B/32, Vit-L/16, ViT-L/32 and Vit-H/14, the latter not being provided. The trailing number represents the number of image sections during processing. The models were pre-trained with the ImageNet-21k data set [30].

For our classification model, we used the ViT-L-16 model. In the end, the transformer outputs a number between 0 and 3, corresponding the Paris classification.

Benchmark models We used two CNN benchmark models to contest our Paris classification system:

The first is Big Transfer (BiT). It uses the principle of transfer learning, in which a convolutional neural network is pre-trained on a huge data set. The pre-trained network is then selected and re-adapted to the relevant problem, also known as finetuning. The tranfer learning principle is used to compensate for deficiencies in training and testing examples in a data set for training a CNN. Transfer learning can be particularly relevant in the medical classification domain, as many medical data sets contain only a small number of data [32].

The second is Efficient Net. Convolutional Neural Networks have dominated the field of computer vision for years due to their good performance. However, CNNs are dependent on the resources available to build and scale the neural networks. Due to limited resources, scaling a neural network is one of the core problems that Google (Research) is trying to solve with its CNN models called EfficientNet [33]. Scaling a Convolutional Neural Network refers to adjusting certain dimensions that can lead to higher accuracy. Common model scaling is performed on the depth, the width of a CNN, or the resolution of an input image. Here, the depth of a model refers to the number of layers in a Convolutional Neural Network. Width is the number of channels in a layer, while resolution refers to image ratios such as height and width [33].

NICE classification

The data situation faced in the NICE classification outlined in the preceding sections is frequently encountered in artificial intelligence, but is a particularly ubiquitous problem in the medical domain of machine learning: Few data sets are made publicly available, but retained as private resources, the amount of data is limited, especially for rare conditions and cases, and the expertise requiring annotations are costly and time-consuming to acquire. This core issue of artificial intelligence has been subject to inquiry in recent years and the prolific branches of zero-shot and few-shot learning have emerged as potential remedies for the data scarcity issues in many machine learning domains [34]. The former refers to algorithms attempting classifications without having been trained on an example of the target classification task, while the latter refers to strategies in which the availability of a few training examples is leveraged for the fine-tuning of zero-shot classification systems.

few-shot learning (FSL) is an active and promising research branch aiming to cross the chasm between the learning behavior of current machine learning systems and that of humans, who achieve high generalization capabilities from a few examples. Given the data situation faced in the NICE classification of this study, we will explore the performance of FSL approaches in the context of polyp classification. The following section will provide a brief outline of the relevant background of FSL.

Few-shot learning

The FSL literature comprises a large stock of different strategies and philosophies to approach the data scarcity issue. The approaches range from the intensive application of data augmentation methods expanding the data set in order to enforce desired invariances in the classification model, transfer learning strategies and even complex meta-learning algorithms, which are trained to provide parameterizations for a model given a few, or even only single example of the target task [35].

A popular and well-established approach in the transfer learning branch of FSL is embedding learning [36], in which an embedding model $f : R^m \rightarrow R^n$, where $n << m$, is trained, such that task-specific notions of similarity between inputs, manifest as trivially quantifiable similarities between their latent representations generated by the model f. In the desired structure of the latent space, the samples of classes do not form a complex manifold but form clusters, allowing distance metrics, such as the euclidean or the cosine distance, to quantify the similarity and class affiliations of samples. A latent space exhibiting such structural properties might then allow the construction of simple class discrimination hypotheses, which are within reach with little data available for the target task. Frequent choices for hypothesis are as simple as a k-nearest neighbour classification ([37], [34]).

The embedding model f can be learned through transfer learning from a task-unrelated but extensive data set and might subsequently be fine-tuned to the target task data depending on the specific amount of data available.

There are many strategies for training the embedding model f, such as the Matching Networks [37] or the Prototype Networks [38]. In this study, we selected concepts of Deep Metric Learning to enforce the desired structure on the latent embedding space.

Deep Metric Learning

The field of Deep Metric Learning is occupied with the training of encoder models, which enforce the previously discussed properties of the latent space in order to provide a semantic metric in conjunction with a specified distance measure [39].

In the field of metric learning, the approach of Siamese networks is an established training paradigm for the encoder. The concept of Siamese networks has first been considered in the field of signature verification [40], but has since then been ported to CNNs and numerous applications including few-shot scenarios [41].

Conceptually, a siamese network comprises a neural network and a weight-sharing clone, which are subsequently trained on pairs of data points, which might constitute a positive pair, demonstrating semantic similarity or a negative pair demonstrating semantic dissimilarity. The neural network and its clone are then trained to produce embeddings with small in the former, respectively high distance in the latter case w.r.t. a selected distance metric.

Hoffer et al., however, realized that the standard approach of the siamese neural network produces sub-optimal results, if the metric is subsequently to be used for classification tasks, as the minimization and maximization of distances between positive and negative pairs does not necessarily lead to the intra-class distances being smaller than inter-class distances [42]. Hoffer et al. proposed to extend the siamese network to a triplet neural network, which comprises three weight-sharing clones of a neural network and is trained on triplets of data points consisting of an anchor instance x, a positive x^+ and a negative instance x^- exemplifying semantic similarity and dissimilarity to the anchor instance respectively [42].

The training of the network f is then designed to enforce a class-consistent distance metric $\|f(x), f(x^+)\|_{\mathcal{D}} < \|f(x), f(x^-)\|_{\mathcal{D}}$ for a metric $\mathcal{D}$ and for all triplets (x, x^+, x^-).

A variety of losses for the triplet network has been proposed for specific scenarios (such as in [43], [44]), but they are generally based on variations of the contrastive loss for siamese networks. For this study, an adaption of the contrastive triplet loss given in [45] is deployed:

$$\mathcal{L}_{triplet}(x, x^-, x^+) = \|f(x), f(x^+)\|_{\mathcal{D}} + max(0, m - \|f(x), f(x^-)\|_{\mathcal{D}}) \tag{1}$$

where m is a margin parameter, which limits the total decrease in loss value achievable by high distances between the negative pair of the triplet and thus prevents network degeneration tendencies. The concept is illustrated in figure 4.

Considered approaches and methodology

With the background regarding few-shot and deep metric learning outlined, this section will discuss the methods in more detail and provide technical aspects regarding the selected hyperparameters used.

Specifically, we will deploy the triplet neural network concept with the loss given in equation 1, with a margin of $m = 20$ and with the metric being the l2-norm.

For the encoder itself, a member of the ResNet-family, ResNet-18, has been selected as the feature extraction backbone, as no performance gains were achievable using

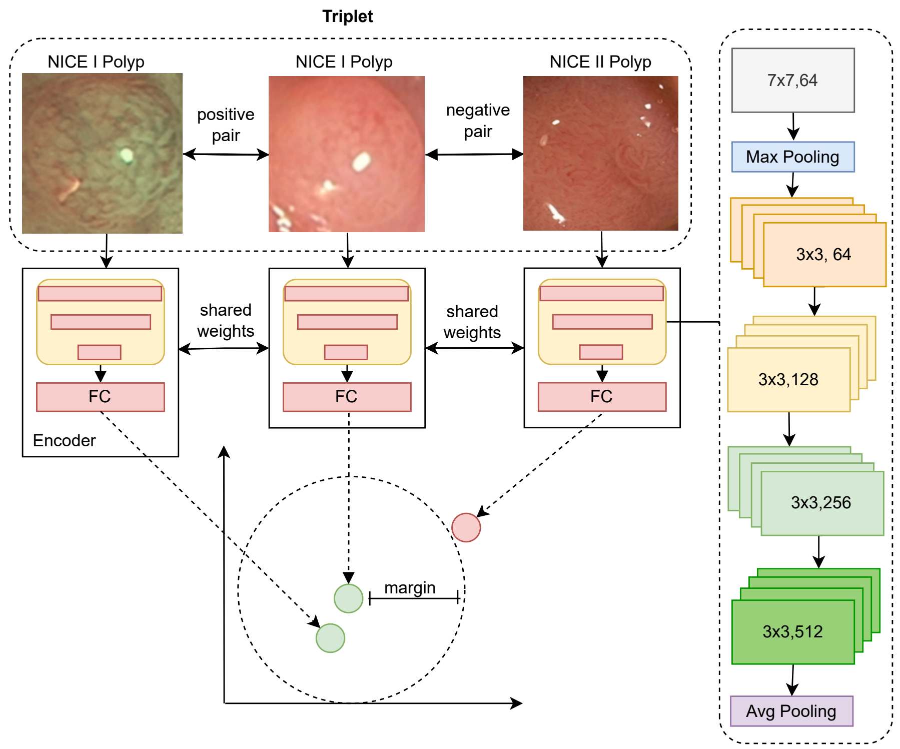


Figure 4 Illustration of the network architecture of a triplet network. According to equation 1, the distance between the vectors of the negative pair is increased to the selected margin

the larger conspecifics such as ResNet-50. The downstream classification layer of the ResNet-18 has been truncated and substituted with a single feedforward encoding layer embedding the average pooled feature map of the backbone into a 64-dimensional latent space.

The encoder has been pre-trained on a transfer learning data set and has been fine-tuned with the available polyp data. Importantly, the fine-tuning did not operate on the classification performance directly, but improved the consistency of the learned metric w.r.t. to the NICE data set using again the triplet loss of equation 1. For the fine-tuning the triplets were formed according to the NICE class affiliation. The fine-tuning scenario is depicted in figure 4.

During the fine-tuning, the training data set, comprising 75% of the available labeled polyp image, has been expanded using a data augmentation process.

As augmentations, random flips along all image axes, as well as random modifications of image hue, contrast, brightness and saturation, have been implemented. The fine-tuning and model selection were subject to an early stopping strategy facilitated by 25% of the train set held back for validation purposes. A single training epoch consisted here of 100 randomly generated triplets.

The embeddings have finally been tested in conjunction with different classification strategies, namely nearest-neighbour (referred to as 1-nn), the smallest average distance (referred to as centroid), or the Support Vector Machine (SVM) [46] equipped with the radial-basis-function kernel. For the 1-nn and centroid approach, the embedded images of the training set served as the latent space population for the test data classification. In the case of the SVM, the embeddings of the training data were used to fit the Support Vector Machine.

In this study, we are particularly interested in the effects of the pretraining and the considered transfer learning data set. We will therefore consider the usage of an out-of-domain, labeled data set and a within-domain, self-supervision-based data set for the pretraining.

Supervised pretraining The challenge of transfer learning is to select a transfer learning data set where the learned notions of semantic similarity are to a large degree aligned with the similarity notions of the target domain, especially if the potential transfer learning data sets exhibit significant domain gaps to the target data regime (such as endoscopic videos).
As the NICE classification scheme is largely based on surface patterns and the textures of polyps [6], we opted in this study for the texture classification data set Describable Texture Data set [3] (DTD for short) [47].
The DTD data set provides a texture database containing 5640 images belonging to 47 different classes of human-distinguishable textures.
As the encoder model is trained with the loss given in equation 1, the construction of triplets is a mandatory preprocess. While the literature has discussed the usefulness of the mining of informative triplets both for the efficiency of training and quality of the discrimination capability (for instance [48]), for the study at hand, the triplets have been randomly mined with positive pairs originating from the same texture classes and negative image pairs from different. Since the DTD data set is a multilabel data set, with some training instances displaying characteristics of different textures simultaneously, the triplet mining selected the negative instances x^- as completely class-disjoint with the anchor instance x.
As a measure to reduce the domain gap between the DTD data set and the polyp images and to provide the encoder with an organic invariance towards highlight corruptions, a preprocessing step has been implemented by grafting random specular highlights extracted from the SUN data set with the detection algorithm of Arnold et al. [49] onto the DTD images. The effect of this preprocessing step will later be discussed in an ablation experiment.

Self-supervised pretraining An alternative approach for pretraining neural networks is the strategy of self-supervised learning. The advantage of self-supervised learning algorithms is their defining independence of labeled ground truth data resulting from their eponymous capability to produce their supervision signal.
A further advantage of the self-supervised approaches is the possibility of tapping into available domain-related data sets. While these data sets lack the relevant ground truth annotation, they might still allow for a pretraining of networks exhibiting smaller domain gaps concerning the target tasks.
Especially in the medical domain, the independence of labeled training data of self-supervised approaches can therefore enable the leveraging of as much of the

---

[3] https://www.robots.ox.ac.uk/~vgg/data/dtd/

available medical data as possible, which is often idiosyncratic (endoscopic images, X-ray scans, etc.).

At a high level, the self-supervised approaches can be divided into generative and discriminative approaches [50], with the former category comprising strategies such as AutoEncoders [51] and the latter comprising again contrastive approaches [50]. The fundamental insight and rationale of using contrastive approaches in self-supervision is that the representations of images and heavily augmented versions of them should be close in the latent space. In contrast, the distance to entirely unrelated images should be more significant. Hence, the self-supervision is again formulated as a triplet metric learning application and the network is enticed to embed the images into representations, which encode features, which are for one invariant towards all applied augmentation methods and for another discriminative towards other images. The concept of the self-supervised training of the encoder is illustrated in figure 5.

This latter discriminative approach has been used as a self-supervised pretraining strategy for the study. The already introduced SUN data set has been used as a source of endoscopic images. For the training, only images containing polyps have been used, which were cropped to the polyp regions and scaled to a common shape of 224×224. Only a fraction of the images in the SUN data set have been deployed for training. The roughly 50000 polyp images have been condensed into a set of approximately 2500 images, which were extracted using an ORB-feature matching based temporal downsampling of the video sequences proposed in [52]. Utilizing the feature matching, the videos were decomposed into a sequence of scenes, out of which the sharpest frames were automatically selected.

As augmentation steps, random flips along all image axes, histogram altering modifications of image hue, contrast, saturation and brightness, and a random gaussian noise have been applied to the images. To further avoid encoding the prevalent specular highlights in the images as a kind of fingerprint, random specular highlights have been grafted onto the images, which have been again extracted from endoscopic images with the specular highlight detection algorithm of Arnold et al. [49].

Results

In this section, we present the results of our two polyp classification systems. We will consider the two subsystems for the Paris and NICE classification separately, starting with the latter classification problem.

Nice classification

The evaluation of the NICE classification system will consider the classification performances of both the full system, comprising the pre-trained encoder network and the subsequent fine-tuning, as well as the stand-alone pre-trained encoder without subsequent fine-tuning.

Beyond that, a range of classification algorithms applied to the embedded polyp images will be considered.

Finally, some design choices will be revisited through ablation experiments.

The experimental design, which has been outlined in the preceding section, will

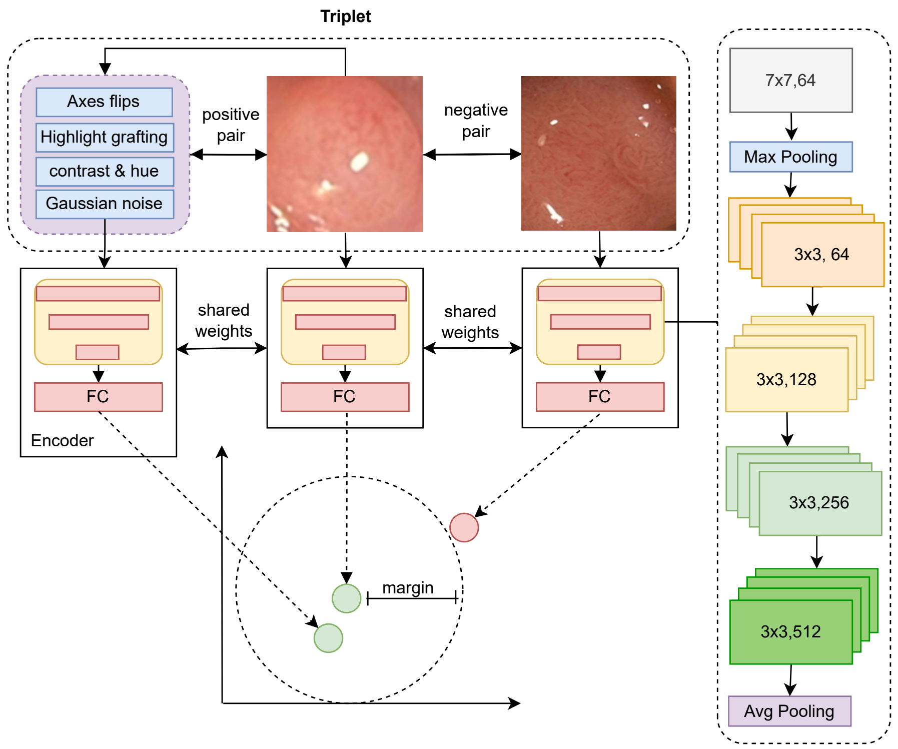


Figure 5 Illustration of the network architecture of a self-supervised triplet network. According to equation 1, the distance between the vectors of the negative pair is increased to the selected margin. The positive pair is built from an image and an augmented view of it.

here be briefly recapitulated concisely: Roughly 75% of the data has been used for the fitting of the classification algorithm and optionally for the fine-tuning of the encoder network. The test data comprised a class-balanced set of roughly 25% of the polyp data. Due to the nature of the data set containing only one image per polyp specimen, the train and test set did not overlap concerning the contained polyp specimens.

As the data split is not negligible in the case of small data sets, we report the average performance of the system across 100 random train/test data splits and the 90% confidence intervals. We expected the confidence intervals to be rather large, as the small data set was unlikely to support a completely split-robust decision boundary. The same train/test splits were used for all experiments. Note at this point, that due to the nonlinearity of the also reported F1-score, the average F1-score is not necessarily equal to the F1-score of the average precision and average recall.

Classification without fine-tuning

This section considers the classification results without a fine-tuning step of the encoder model. The not fine-tuned models were considered to elucidate, how or if at all the differences in the pretraining strategy would manifest in the direct classification performance. The results are given in table 4.

While the Support Vector Machine is the most complex discriminator considered, it displays better performance by a large margin compared to the nearest neighbour and average distance classifier, which indicates, that the two NICE classes are not completely separated in the latent space. Another point of view on this circumstance can be gained in table 5, where the inter- and intra-class variances are reported for the embeddings of the differently pre-trained encoders. Table 5 shows, that the not fine-tuned DTD encoder fails at producing a compact cluster for the NICE II class. The encoder trained on the SUN images using self-supervision produces more consistent embeddings for the polyp images, which is also reflected in its better performance in the classification in table 4. We attribute this difference in performance to the domain gap between the polyp images and the images in the DTD data set.

Table 4 Classification evaluation results not fine-tuned versions of the model pre-trained on the DTD data set and an endoscopic data set using self-supervison. The table shows the average scores for 100 random training/test splits with 90% confidence intervals.

Model	Classification	Acc	Pre	Rec	F1
DTD	1-nn	68.13 (± 14.21)	69.48 (± 12.25)	68.13 (± 14.21)	0.688 (± 0.152)
	centroid	67.69 (± 10.81)	72.55 (± 11.37)	67.69 (± 10.88)	0.700 (± 0.126)
	SVM	**68.64** (± 11.83)	**71.93** (± 10.13)	**68.64** (± 11.83)	**0.701** (± 0.137)
self-sv.	1-nn	65.34 (± 10.50)	65.91 (± 11.43)	65.34 (± 10.50)	0.657 (± 0.129)
	centroid	65.38 (± 15.32)	67.72 (± 14.71)	65.38 (± 15.32)	0.665 (± 0.159)
	SVM	**72.55** (± 13.82)	**73.95** (± 12.61)	**72.55** (± 13.82)	**0.733** (± 0.147)

Table 5 Intra- and interclass variances of the non fine-tuned polyp image embeddings of the models trained on the DTD data set and endoscopic data set with self-supervision. The interclass variance is normalized to 1.

Model	intra NICE I	intra NICE II	inter
DTD	0.62	1.27	1.0
self-supervision	0.88	0.85	1.0

Classification with fine-tuning

This section considers the performance of the two encoder systems with a fine-tuning step. To that end, the train data of the polyp images have been used to produce triplets with negative and positive triplet components selected according to their NICE class affiliation. Besides, a set of augmentations has been applied to the triplet images, encompassing random flipping along all image axes and heavy histogram modifying operations acting upon hue, contrast, brightness and saturation of the images. The training used early stopping facilitated by a held-out validation part of the train set. The results are reported in table 6. Fine-tuning increased the top performance for both pretraining strategies, especially for the model trained on the DTD data set, which exhibits the overall top performance. We attribute this strong increase in performance of the DTD trained model to closing the domain gap between the DTD and polyp images. The results of the DTD trained encoder vis-à-vis the fine-tuned self-supervision system indicate however, that the pretraining on the texture data set bestowed the model with a superior and better generalizing feature extraction capability, which constituted a better initialization for the refinement of the representations.

The SVM classification performed well for both pretraining strategies in relative terms, with the smallest average distance producing even slightly better results on the DTD pre-trained model.

Table 6 Classification evaluation results in fine-tuned model versions pre-trained on the DTD data set and an endoscopic data set using self-supervision. The table shows the average scores for 100 random training/test splits with 90% confidence interval.

Model	Classification	Acc	Pre	Rec	F1
DTD	1-nn	75.31 (± 9.41)	75.94 (± 8.63)	75.31 (± 9.41)	0.757 (± 0.098)
	centroid	**81.39** (± 8.53)	**82.05** (± 8.61)	**81.39** (± 8.53)	**0.817** (± 0.084)
	SVM	81.34 (± 8.74)	81.52 (± 8.39)	81.34 (± 8.74)	0.810 (± 0.086)
self-sv.	1-nn	71.59 (± 8.74)	75.09 (± 8.13)	71.59 (± 8.74)	0.733 (± 0.095)
	centroid	68.88 (± 8.45)	70.30 (± 8.82)	68.88 (± 8.45)	0.696 (± 0.097)
	SVM	**75.04** (± 8.59)	**75.24** (± 8.38)	**75.04** (± 8.59)	**0.751** (± 0.083)

In summary of the results of the preceding two experiments and following the methodology of [23], who base their pathology assessment of polyps on the classes I and II of NICE, we conclude, that the here presented FSL model displays performances comparable to the results reported in the literature reviewed in the related work section of this study, despite the very limited amount of data available and the partially suboptimal acquisition of the images (without the NBI mode activated). Moreover, we conclude that in the case of sufficient fine-tuning data being available, it is advantageous to conduct the pretraining on transfer learning data sets, in which the alignment of the presumed feature extraction capabilities learned from the data set, and the required capabilities for the target task is easier to foresee, as it has been the case with the texture DTD data set. While a smaller domain gap proved advantageous in our experiments (refer back to table 4), when fine-tuning was not conducted, the self-supervision primed the encoder model in a way that allowed only for a minor refinement of the embeddings, which could be converted only into a small gain in performance, before the overfitting to the training data set in. Furthermore, the fine-tuning consolidated the confidence intervals significantly across the considered data splits.

Ablation considerations

This section will discuss the effect and influence of a few design choices made throughout the description of the NICE classification model. The average results of the 100 considered random train/test splits are reported.

First, we consider the influence of the data augmentation applied on the training data during fine-tuning. The results are presented in table 7. While the augmentation yields for both pretraining strategies the best models concerning the F1-score, the performance difference is only small. The main incentive for introducing the training augmentation in the first place was to ensure that the classification was not based on spurious correlations in the small data set. But as the not augmented runs did not produce better results, even slightly worse, it is concluded that this worry was not justified, to begin with.

Table 7 Effect of augmentation during fine-tuning for differently pre-trained embedding models. The classification was performed using a Support Vector Machine. The table shows the average scores for 100 random training/test splits with 90% confidence interval.

Model	Augm.	Acc	Pre	Rec	F1
DTD	N	80.73 (± 8.48)	**82.05** (± 8.74)	80.73 (± 8.48)	0.805 (± 0.082)
	Y	**81.34** (± 8.74)	81.52 (± 8.39)	**81.34** (± 8.74)	**0.810** (± 0.086)
self-sv.	N	74.03 (± 8.97)	**76.13** (± 8.49)	74.03 (± 8.97)	0.750 (± 0.081)
	Y	**75.04** (± 8.59)	**75.24** (± 8.38)	**75.04** (± 8.59)	**0.751** (± 0.083)

Finally, we consider the effect of the augmentation strategy of grafting random specular highlights on the images of the DTD data set during the pretraining of the encoder. In this experiment, we analyze whether accounting for the invariance towards these image corruptions can be fully substituted through fine-tuning and how it affects the not fine-tuned models. To that end, we considered an encoder trained on the DTD data set without the highlight augmentation grafting vis-à-vis the previous encoder in both the fine-tuning and no fine-tuning setting. The results are reported in table 8. As the results indicate, the effects of the highlight-grafting operation depend heavily on the subsequent fine-tuning. While the augmentation increases the performance in all cases, the fine-tuning can catch up with the invariance towards the specular highlights. However, the non-finetuned model without the augmented pretraining suffers to a larger extent from interferences of the image corruptions.

Table 8 Effect on the specular highlight grafting augmentation during pretraining of the encoder with the DTD data set. The average performance on 100 random train/test splits is reported

Model	**finetuning**	**highlight grafting**	**Acc**	**Pre**	**Rec**	**F1**
DTD	Y	N	80.44	80.51	80.44	0.803
	Y	Y	**81.34**	**81.52**	**81.34**	**0.810**
	N	N	65.81	63.86	65.81	0.647
	N	Y	**68.61**	**71.92**	**68.61**	**0.701**

Error analysis

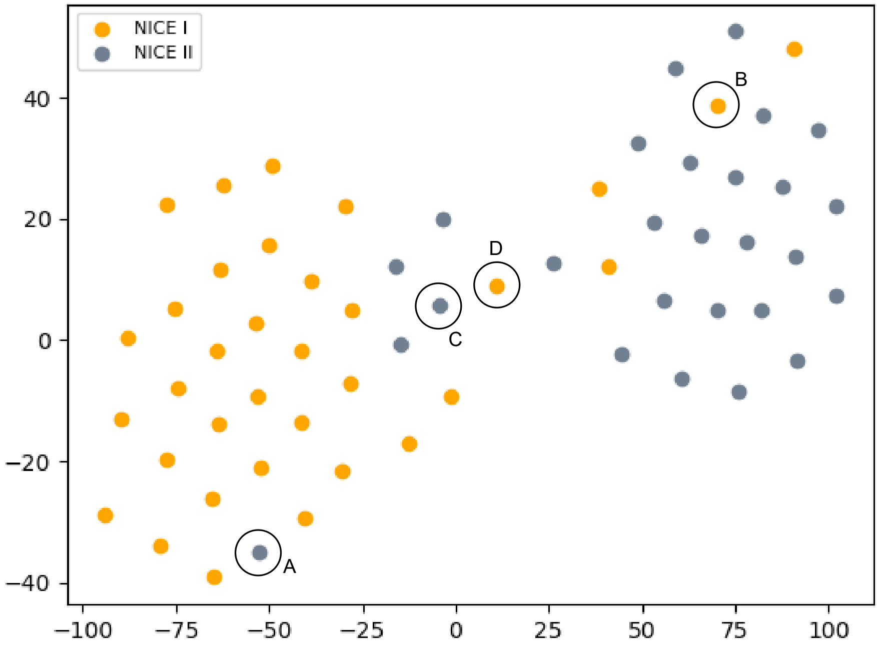


Figure 6 t-SNE embeddings [53] into 2D of the polyp images using the DTD trained encoder. The highlighted data points will be subject of a discussion.

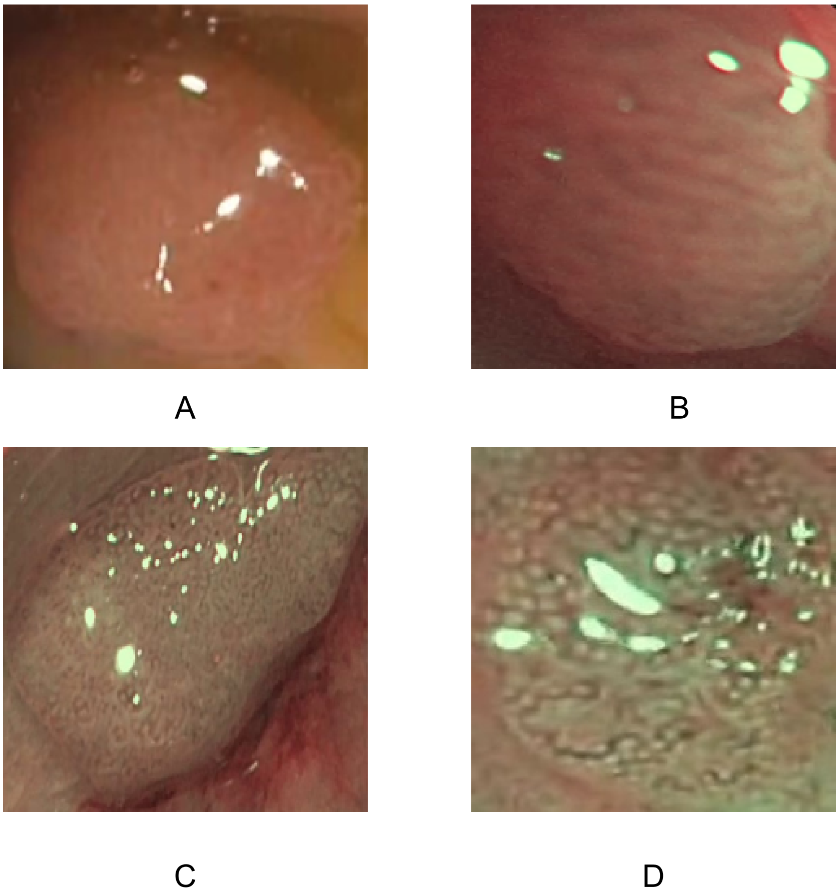


Figure 7 Samples of misclassified polyps of our system. The images A and C belong to the class NICE II. The images B and D appertain to class NICE I.

This section will conclude the NICE classification with a short error analysis of the developed classification system.

The overall quality of the learned embedding can be seen in figure 6, which displays the t-SNE projections [53] into 2D of the embeddings generated by the DTD pre-trained encoder model. The projections reveal that the different NICE classes form two distinct clusters in the latent space, which possess however an overlapping zone, which reflects the classification performances given in table 6.

We will now consider two kinds of problematic embeddings to gain further insights into the performance. Firstly, we consider two data points embedded well into the clusters of the wrong NICE class. The data points are denoted with A and B in figure 6 and in figure 7, where they are depicted in the upper row. As shown in figure 7, the image A is heavily blurred, such that its surface appears feature less. Note, that image A has also not been taken with the NBI-light activated. With the surface patterns not discernible, the homogeneous polyp has been embedded into the NICE I cluster of the latent space. Similarly, polyp B's surface exhibits discernible tubular structures, which have likely been picked up by the encoder and led to an embedding into the NICE II cluster of the latent space.

Secondly, we will consider two polyps that populate the overlapping zone of the two latent clusters. The polyps concerned are denoted C and D in both figure 6 and 7. Both polyps display a pronounced surface texture and rich patterns. While in both cases, the features of their correct NICE class dominate the patterns (tubular in case of polyp C and spotted in D), both polyps display at close inspection also structures of the respective other NICE class.

We conclude from the presented error analysis that the NICE classification system facilitated by the polyp encoding neural network presented in this paper succeeds at generating semantically viable representations of polyps and embedding the polyps into a well-structured latent space apt for downstream usage in classification.

Paris classification

For the Paris classification we compare two additional state-of-the-art algorithms to our approach for a fair comparison. For the comparison, we are using BiT-R152x4, and EfficientNet-B7. BiT-R152x4 and EfficientNet-B7 are both CNN architectures. Our model (ViT-L-16) with different learning rates, data augmentation methods, and dropout rates. This will help decide which hyperparameters and settings are needed for each model to train the best possible polyp classifiers.

Experimental Design

For the evaluation of the Paris classification the images were divided into training, validation, and testing data sets based on the number of different polyps, with approximately 70 % of the polyp images from the SUN Colonoscopy Video data set being used for training, 15 % for validation, and 15 % for testing. The sun data set was thereby split in cases so that there is no polyp training data in which the same case would also be in the test data. The final test data consist of the 15% of polyps in the SUN data set split in cases and 15% of our own data set also split in individual cases.

Transfer learning models were used for training, pre-trained on existing data sets and refined for the polyp classification task. BiT-R152x4 and ViT-L-16 are used with the weights pre-trained on ImageNet-21k. ViT-L-16 was also finetuned on the ILSVRC-2012 data set [30, 32, 33]. In addition, EfficientNet has the special case that training can proceed in two phases. First, all weights in the network are frozen and only the last layers are adjusted. The second phase is optional and offers training in the deeper layers. For this work, both methods were used and the best results were presented.

Finding the correct hyperparameters for the models is essential for the accuracy of the models. Therefore, different parameters and settings were trained and tested for each model. The related results are presented in the ablation study subsection. For this purpose, this paper tested and selected different learning and dropout rates. Furthermore, different data augmentation methods were additionally tested to boost the performance of the models.

In addition to the different dropout rates and data augmentation, the early stopping method was used to avoid overfitting and long training times. For Big Transfer, training was stopped after seven epochs without improvement, while for EfficientNet, training was stopped after 20 epochs without improvement. For our model, the training was stopped after 11 epochs.

Evaluation

The evaluation is done via the F1-score and the accuracy. The F1-score describes the harmonic mean of precision and recall. The F1-score, the accuracy, the recall and precision are shown in following equations:

$$\text{Precision} = \frac{TP}{TP+FP} \qquad \text{Recall} = \frac{TP}{TP+FN}$$

$$F_1 = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} = \frac{2*TP}{2*TP+FP+FN}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

We count an annotation as true positive (TP) if the classification of our prediction and GT do have the same label. If a polyp is predicted in a wrong class but the polyp is another class we count it as a false positive (FP). We calculate the TP, FP, true negatives (TN), false negatives (FN) for every class and calculate the scores according to the equations above.

For the testing BiT-R152x4 from Big Transfer, our model using ViT-L-16 from Vision Transformer, and B7 from EfficientNet were tested. The results are illustrated in the table below:

Table 9 Test results of each model on two different test data sets, the SUN Colonoscopy Video data set and our own data set (EndoData) [9]. All values are given in %.

Model	Data set	Acc	Pre	Rec	F1
BiT-R152x4	SUN	80.45	69.57	77.25	73.21
	EndoData	76.31	76.24	72.28	74.20
EfficientNet-B7	SUN	84.25	72.82	**80.27**	76.36
	EndoData	73.94	72.11	71.01	71.46
Ours	SUN	**89.35**	**84.76**	79.10	**81.28**
	EndoData	**87.42**	**80.09**	**78.83**	**79.45**

Table 9 shows that our approach using a transformer architecture outperforms the two other CNN approaches in nearly all metrics. Especially on the harder-to-classify EndoData [9]. The improvement from BiT-R152x4 to our model shows an accuracy of 76.31% to 87.42 %. A significant approvement considering our approach compared to the CNN approach. Nevertheless, the EfficientNet-B7 algorithm achieves a minimal improvement considering the recall on the SUN data set with an increase from 79.10 % to 80.27 % compared to our approach. As shown in table 2, comparing these algorithms to the published literature in the domain is challenging because the algorithms are evaluated on different data sets and using different classes. Nevertheless, Bour et al., which is the best approach using three classes, achieved an accuracy of 87.1 % [2] on their test data set. With our model, we are surpassing this accuracy by 2.04 %. Nevertheless, in the paper of Bour et al. [2], 785 different polyps are used for training and validation, and the authors did not specify the amount and composition of the test data. Therefore, it is hard to make a fair comparison between the algorithms.

To further elaborate on the results of our model we computed the accuracy, precision, recall and F1-score for every Paris class individually. The results are shown in table 10. For the accuracy the results indicate that classes Is and Ip are best classified by the model.

Table 10 In this figure, the test results of our model on the SUN Colonoscopy Video data set are shown for each Paris class individually. All values are given in %.

Paris class	**Acc**	**Pre**	**Rec**	**F1**
Is	92.97	91.87	93.27	92.56
Ip	94.30	90.66	55.64	68.96
Isp	85.94	68.84	42.41	52.49
IIa	84.43	78.90	76.27	77.56
Mean	89.35	84.76	79.10	81.28

Ablation study

In this section, we present the results of the BiT-R152x4, EfficientNet-B7, and our model with different learning rates, data augmentation methods, and dropout rates. This will help decide which hyperparameters and settings are needed for each model to train the best possible polyp classifiers.

Learning rate To find a suitable learning rate for each model, the models were trained and tested with different learning rates. All models have, if applicable, a dropout rate of 0.5. For the data augmentation, our model and BiT-R152x4 were set to random flipping, while the EfficientNet-B7 results were computed with the combination of random flip, random rotation and random contrast. Table 11 shows the results for each model considering different learning rates. In addition, the time of one training epoch per minute and the required number (#) of epochs until reaching the best accuracy on the validation data set are given.

Table 11 Results on the validation data set considering different learning rates.

Model	**Learning rate**				**Val-acc**	**Training speed**	
	0.01	0.001	0.0001	0.00016		Min/Epoch	#Epochs
BiT-R152x4	✓				0.7890	≈ 30	4
		✓			**0.8213**	≈ 30	8
			✓		0.8140	≈ 30	10
				✓	0.8156	≈ 30	10
EfficientNet-B7	✓				0.7903	≈ 5.7	6
		✓			**0.8212**	≈ 5.7	10
			✓		0.7924	≈ 5.7	30
				✓	0.7969	≈ 5.7	28
Ours	✓				0.4668	≈ 3	19
		✓			0.5938	≈ 3	23
			✓		0.8242	≈ 3	10
				✓	**0.8950**	≈ 3	8

Thereby, the results provide the first indications that for the CNN models BiT-R152x4 and EfficientNet-B7, the best results are obtained with the learning rate of 10^{-3}. Our model achieved better results with a lower learning rate. In addition, this required less time for one training epoch since the computational effort is lower for the Vision Transformer compared to the CNN models [30]. Another interesting aspect of the results in table 11 is that for the CNN methods, the number of epochs increases when decreasing the learning rate, but for our transformer model, considering the first two learning rates of 0.01 and 0.001, the number of epochs is decreasing. This is contradictory and could be attributed to the fact that it is hard to learn for the transformer model with these learning rates and therefore, the training goes longer than it should. For the subsequent analysis to investigate data augmentation and dropout, the learning rate that provided the best validation accuracy in table 11 was used for each model.

Data augmentation In the second step of this analysis, various data augmentation methods were explored to adjust the models to best fit the polyp classification. Data augmentation helps combat overfitting and can create critical diversity in a data set. The increased diversity in the training data set improved the performance. The data augmentation methods used for this training are random flipping (random flip) or rotating the images (random rotation), and changing the contrast (random contrast). Table 12 presents the obtained training results considering different augmentation techniques.

Table 12 Results on the validation data set considering different data augmentation methods.

Model	**Data augmentation**			**Acc**
	random flip	random rotation	random contrast	
BiT-R152x4				0.8155
	✓			**0.8213**
	✓		✓	0.4543
	✓	✓		0.7968
	✓	✓	✓	0.4469
EfficientNet-B7				0.7551
	✓			0.7903
	✓		✓	0.7936
	✓	✓		0.8091
	✓	✓	✓	**0.8212**
Ours				0.7930
	✓			**0.8950**
	✓		✓	0.8210
	✓	✓		0.8242
	✓	✓	✓	0.6016

The table shows that all models benefit from data augmentation. Training runs without data augmentation gave much worse results. This indicates that data augmentation is important for polyp classification. Especially the random horizontal and vertical flipping of the images seems to have a great effect for polyp classification. For the subsequent analysis to investigate dropout, the data augmentation that provided the best validation accuracy in table 12 was used for each model. Random flipping and changing the contrast had different effects on the models. EfficientNet provided improved performance to 82.12 %. The other options in combination with flipping caused deterioration of the results for our model and BiT-R152x4. Nevertheless, their results achieved increased validation accuracy by random flipping alone. 89.50 % for our model and 82.13 % for BiT-R152x4.

Dropout Dropout is a regularization technique to avoid overfitting on the data set. As a further step, this section experiments with different dropout rates to make the models less susceptible to overfitting and thus achieve better values on the validation data set. With one exception for BiT-R152x4, dropout rates of 0.4, 0.5, and 0.6 were tested on the remaining models. The authors of BiT-R152x4 did not use dropout to avoid overfitting, but attempted to train stable models using the learning rate schedule method [32]. In the learning rate schedule method, no fixed learning rate is set for training, but varying learning rates are used. For example, at the beginning of the training, a large learning rate is used to move the gradient faster towards the minimum. Then the learning rate is decreased during training so that at the end the gradient does not skip the minimum. This results in reaching the minimum faster and the model gains higher accuracy.

Table 13 Results on the validation data set considering different dropout rates. BiT-R152x4 did not use dropout and is therefore not included in this table.

Model	**Dropout rate** 0.4	0.5	0.6	**Val-acc**
EfficientNet-B7	✓			0.8094
		✓		**0.8212**
			✓	0.7908
Ours	✓			0.8593
		✓		**0.8950**
			✓	0.8513

The results in the 13 table show that the models produce solid results at all dropout rates, but show the best results at a dropout of 0.5 on the validation data set.

Few-shot learning As a last ablation, we want to briefly revisit the overall selection of the classification model and compare the performances of the Vision Transformer with the model underlying the few-shot learning system presented in the NICE classification section of this paper.

We deployed the outlined self-supervision approach, as the texture dataset DTD is inadequate for pretraining of a shape-centric classification task. As an augmentation engine facilitating the self-supervised pretraining, we deployed the style-transfer algorithm of [54], which provides a model capable of applying the style of arbitrary images to the content of another image. We selected the style-transfer as an augmentation step, as it allows the suppressing of most of the texture and style-related information of the original image and retains the structure and shape information as the main source of discriminative features. For the training, we selected pencil drawing styles, which we found to introduce almost no artificial texture to the images and highlight the structure and shape of the polyps in a very pronounced way. An overview of the deployed triplet generation is given in figure 8. The pretraining was again followed by a fine-tuning phase during which the triplets were constructed according to Paris class affiliation.

The configurations and parameters of the model and training remained identical to the setting described in the NICE classification sections of this paper.

Especially, the ResNet-18 has been retained as a feature extraction backbone and the SVM was used for the subsequent classification of the embeddings generated by the encoder.

The SUN data and the identical split of the 100 cases used in the preceding experiments involving the transformer were used to train and evaluate the model. Similarly to the pretraining of the self-supervised NICE classification system, we used a fully automated key frame selection pipeline to condense the training data down to 1081 images.

The system results are given in table 14. As can be seen in the table, the system achieves high precision in the Paris class IIa and the minority classes Ip and Isp. However, the downside of the high precision is a weak recall, especially in the classes

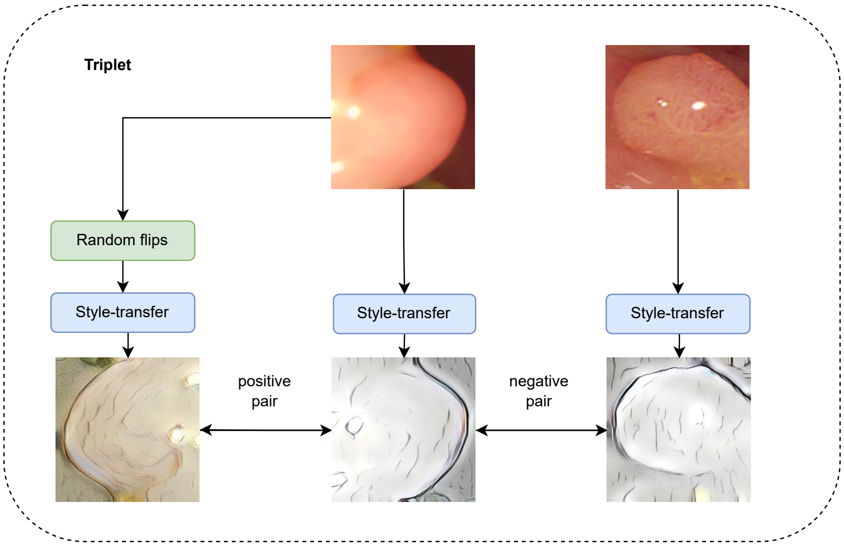


Figure 8 Triplet generation during the self-supervised pretraining for the Paris classification. The same style was used for the images of the negative pair, while different styles were used for the images of the positive pair.

Ip and Isp, where all misclassified images were confused with the class Is or with Is and Ip in case of class Isp. The high precisions in the pedunculated classes allow the model to determine the presence of a peduncluation (Ip or Isp) with a 96.56% precision. The low recall however is also reflected in the precision of the class Is under which many images showing protrusions are subsumed.

The proposed transformer displayed therefore the overall best results in the discussed task, albeit the metric-based system displays performances comparable to those of the other considered models, such as the EfficientNet, despite of the again considered scenario of little available data. Nevertheless, the approach using a state-of-the-art vision model above shows superior results considering the Paris classification.

Table 14 Results of the few-shot model in the SUN Colonoscopy Video data for each Paris class individually. All values are given in %.

Paris class	Acc	Pre	Rec	F1
Is	74.85	66.39	95.42	78.26
Ip	95.43	88.66	39.81	54.92
Isp	90.76	92.08	19.47	32.05
IIa	87.93	92.44	70.68	80.04
Mean	82.97	79.75	75.69	73.19

Discussion

In this chapter, we discuss the limitations and the explainability of the system. We primarily focus on wrong detections of the polyp classification system and discuss those system failures on the data sets. Additionally, we create heat maps showing the networks neural activation to gain deeper insight into the reasons for the classification results of the network. In this paper, two pre-trained CNN models as well as a pre-trained special transformer were used for the Paris classification. Especially the use of different data augmentation methods strongly improved the results of the models. Specifically, random image flipping seems to play an essential role in polyp characterization and should be looked at more closely in future research. This could be due to the reason that the Vision Transformer can understand and learn information about the whole image in the first layers of the model through the Attention layers. This presumably allows the model to better recognize the polyp features. CNNs, in turn, try to classify based on the locally recognized features [30], which profit from different augmentations.

Limitations

First, assessing the test results, the distribution of images on the test data sets was unbalanced. Looking at the two test data sets, it is noticeable that the images with polyp types Is and IIa are particularly strongly represented, while the other classes are less represented. This may weaken the significance of the test results. However, the proportion of classes Ip and Isp in the training and validation data set is also low, and this may cause the models to classify these two classes moderately. This is due to the lack of labeled data sets for the polyp domain, which leads to the following limitation.

The lack of data is a significant problem, specifically in computational medical research, as a large amount of training data is required to build and train stable and accurate deep learning models. However, the number of annotated data sets, specially labeled polyp data sets for Paris classification, are severely limited. In addition, the existing polyp data sets still contain few polyp images for a deep learning task. For, e.g., the SUN Colonoscopy video data set [27], the data set consists of just 100 different polyps, of which nearly 70 are different polyps for training. This number tends to be too small to train a stable classifier. Therefore the diversity of polyps is missing. Moreover, the individual polyp cases of the data set consist of image frames of colonoscopy videos. This leads to the next problem, which may further impact the trained object recognition models. First, a colonoscopy video is many image sequences of one polyp. If we exclude the possible blur and distortion in the frames, the sequences consist of barely or slightly distinguishable images of polyps that are used to train the network. On the other hand, the videos are occasionally based on distant images of polyps, which were cropped and used again in this work based on the annotations. Thus, the data set used contains mostly small images, making them difficult to recognize, as shown by image section (a) in Figure 9.

An additional obstacle in training the classifiers relates to the Paris classification. Since the SUN Colonoscopy Video data set contains polyp images for classes Ip, Isp, Is, and IIa, the object recognition models were examined to classify these four types. Here, it was noticeable that class Isp, the mixed form of Is and Ip, is difficult to identify for the classification models. Here, tests have shown that the mixed form is usually classified as one of the two primary forms due to the high similarity, as shown in an image section (b) in Figure 9. Another reason for the confusion is the angle at which the image is acquired. Because a polyp is imaged from multiple sides during a colonoscopy, images of polyps are produced that cannot lead to a definite conclusion about the shape. For example, an image above of a pedunculated polyp (Ip) does not provide any information about the shape because, most likely, no pedicle can be seen. This problem mainly affects the classes Ip and Isp.

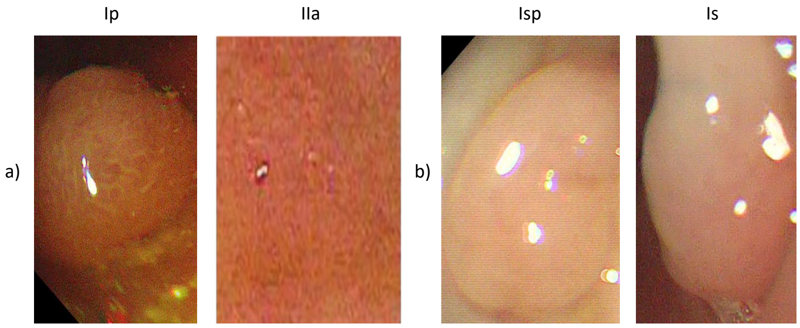


Figure 9 Model detection problems due to (a) difficult to detect polyps due to poor resolution and due to (b) the high similarity of the mixed form Isp class to Is. Images are taken from the SUN data set [27].

Lastly, extending the classification to all Paris classes would be very important. Since classes are missing and there is no "other" class, inherent errors are made when a polyp has a non-modeled class. To create a system with all classes, it would be necessary to construct bigger data sets in which those uncommon classes are highly represented.

Heat maps for the Paris classification

In this section, we demonstrate the use of GradCAM to see what areas are essential for the network to classify a polyp. For this, we used GradCAM with Eigen smooth, a method to remove much noise in the heatmap. We picked three examples for each class to demonstrate the results (see figure 10). This paragraph presents a methodology to generate visual explanations for deriving insight into our polyp classification systems decisions using the Grad-CAM algorithm [55]. We follow the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) [56].

Analyzing figure 10, throughout the examples, the network mostly looks at the polyp's surface and not the background. Furthermore, there are gaps in the heat maps at areas of light reflections, which shows that the network can filter unnecessary information. Especially, for example, Isp with images c1) and c2) shows the AI ignores the background and the light reflections and only considers the structure of the polyp for the classification. In the Ip class, in image a1), we can see a red mark on the polyp. Even that mark is excluded and is not considered by the network, see image a2).

Conclusion

In this paper, we show two novel automated classifications system assisting gastroenterologists in classifying polyps based on the NICE and Paris classification. We introduce a two-step process for the Paris classification: first, detecting and cropping the polyp on the image, and subsequently classifying the polyp with a transformer network. For the NICE classification, we designed a few-shot learning algorithm based on the Deep Metric Learning approach. The algorithm creates an embedding space for polyps, which allows classification from a few examples to account for the data scarcity of NICE annotated images in our database. Overall,

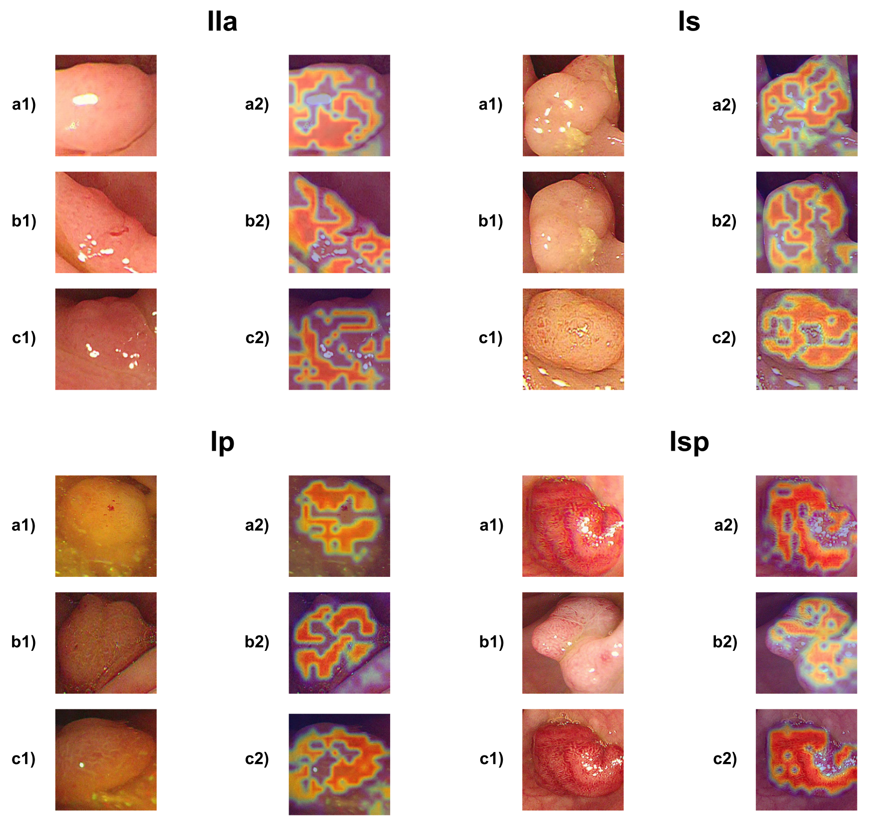


Figure 10 Heat maps for polyp classification. This figure illustrates the classifications of the model using the GRAD-CAM algorithm [55]. Thereby, pixels most relevant for the classification are marked in warm colors like red, and pixels less relevant for the neural network in cold colors like blue. Images are taken from the SUN data set [27].

our Paris classification system shows state-of-the-art results on a publicly available data set with an accuracy of 89.35 %, surpassing all papers in the literature. For the NICE classification, we achieve a competitive accuracy of 81.34 % demonstrating thereby the viability of the FSL approach in data-scarce environments in the endoscopic domain.

Acknowledgements
We kindly thank the University Hospital of Würzburg, the Interdisziplinäres Zentrum für Klinische Forschung (IZKF) and the Forum Gesundheitsstandort Baden-Württemberg for supporting the research. Furthermore, we kindly thank Bayern Innovativ for the support.

Funding
AH and WGZ receive public funding from the state government of Baden-Württemberg, Germany (Funding cluster Forum Gesundheitsstandort Baden-Württemberg) to research and develop artificial intelligence applications for polyp detection in screening colonoscopy. FP receives funding by Interdisziplinäres Zentrum für Klinische Forschung (IZKF) from the University of Würzburg and from Bayern Innovativ (Projectnumber LSM-2007-0009).

Abbreviations
CRC: Colorectal cancer; CNN: Convolutional neural network; CAD: Computer-aided detection; CADx: Computer-aided diagnosis; JSON: JavaScript Object Notation; AI: Artificial intelligence; SUN: Showa University and Nagoya University; WCE: Wireless Capsule Endoscopy; CEM: Context enhancement module; GAN Generative

Adversarial Network; FastCat: Fast Colonoscopy Annotation Tool; FPS: Frames per second; GPU: Graphical processing unit; R-CNN: Region based convolutional neural network; FSL: Few-shot learning; DTD: Describable Texture Data set; CLAIM: Checklist for Artificial Intelligence in Medical Imaging; Acc: Accuracy; Pre: Precision; Rec: Recall; Val: Validation; NBI: Narrow Band Imaging; NICE: NBI International Colorectal Endoscopic;

Availability of data and materials
The first data set used for the analysis of this article is available at the following link (http://sundatabase.org/). The second data set (EndoData) used during the analysis is available from the corresponding author on reasonable request.

Ethics approval and consent to participate
The study including retrospective and prospective collection of examination videos and reports was approved by the responsible institutional review board (Ethical committee Landesärztekammer Baden-Württemberg, 21st of January 2021, F-2020-158). All methods were carried out in accordance with relevant guidelines and regulations. Informed consent was obtained from all subjects and/or their legal guardian(s).

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Authors' contributions
AK implemented and coordinated the study, drafted the manuscript, interpreted the data and implemented the software. SH designed, implemented and evaluated the NICE classification system and its derivative for the Paris classification system and contributed to the appertaining sections of the manuscript. SM contributed to the completion of the manuscript. DF helped with the creation of the data. FP, AH and WZ provided funding and reviewed the manuscript. All authors read and approved the final manuscript.

Author details
[1]Department of Artificial Intelligence and Knowledge Systems, Julius-Maximilians University of Würzburg, Sanderring 2, 97070 Würzburg, Germany. [2]Interventional and Experimental Endoscopy (InExEn), Department of Internal Medicine II, University Hospital Würzburg, Oberdürrbacher Straße 6, 97080 Würzburg, Germany. [3]Department of Internal Medicine and Gastroenterology, Katharinenhospital, Kriegsbergstrasse 60, 70174 Stuttgart, Germany.

References

1. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A.: Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians **68**(6), 394–424 (2018). doi:10.3322/caac.21492
2. Bour, A., Castillo-Olea, C., Garcia-Zapirain, B., Zahia, S.: Automatic colon polyp classification using convolutional neural network: A case study at basque country. In: 2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), pp. 1–5 (2019). doi:10.1109/ISSPIT47144.2019.9001816
3. Ozawa, T., Ishihara, S., Fujishiro, M., Kumagai, Y., Shichijo, S., Tada, T.: Automated endoscopic detection and classification of colorectal polyps using convolutional neural networks. Therapeutic Advances in Gastroenterology **13**, 175628482091065 (2020). doi:10.1177/1756284820910659
4. Lui, T., Wong, K., Mak, L., Ko, M., Tsao, S., Leung, W.: Endoscopic prediction of deeply submucosal invasive carcinoma with use of artificial intelligence. Endoscopy International Open **07**, 514–520 (2019). doi:10.1055/a-0849-9548
5. Lambert, R.f.: Endoscopic classification review group. update on the paris classification of superficial neoplastic lesions in the digestive tract. Endoscopy **37**(6), 570–578 (2005)
6. Hewett, D.G., Kaltenbach, T., Sano, Y., Tanaka, S., Saunders, B.P., Ponchon, T., Soetikno, R., Rex, D.K.: Validation of a simple classification system for endoscopic diagnosis of small colorectal polyps using narrow-band imaging. Gastroenterology **143**(3), 599–607 (2012)
7. Van Doorn, S.C., Hazewinkel, Y., East, J.E., Van Leerdam, M.E., Rastogi, A., Pellisé, M., Sanduleanu-Dascalescu, S., Bastiaansen, B.A., Fockens, P., Dekker, E.: Polyp morphology: an interobserver evaluation for the paris classification among international experts. Official journal of the American College of Gastroenterology— ACG **110**(1), 180–187 (2015)
8. Ferlitsch, M., Moss, A., Hassan, C., Bhandari, P., Dumonceau, J.-M., Paspatis, G., Jover, R., Langner, C., Bronzwaer, M., Nalankilli, K., *et al.*: Colorectal polypectomy and endoscopic mucosal resection (emr): European society of gastrointestinal endoscopy (esge) clinical guideline. Endoscopy **49**(03), 270–297 (2017)
9. Krenzer, A., Banck, M., Makowski, K., Hekalo, A., Fitting, D., Troya, J., Sudarevic, B., Zoller, W.G., Hann, A., Puppe, F.: A real-time polyp detection system with clinical application in colonoscopy using deep convolutional neural networks. `https://assets.researchsquare.com/files/rs-1310139/v1_covered.pdf?c=1644335078`(2022)
10. Sano, Y., Hirate, D., Saito, Y.: Japan nbi expert team classification: Narrow-band imaging magnifying endoscopic classification of colorectal tumors. Digestive Endoscopy **30** (2018)
11. Neilson, L.J., Rutter, M.D., Saunders, B.P., Plumb, A., Rees, C.J.: Assessment and management of the malignant colorectal polpy. Frontline Gastroenterology **6**, 117–126 (2015)
12. Hayashi, N., Tanaka, S., Hewett, D.G., Kaltenbach, T.R., Sano, Y., Ponchon, T., Saunders, B.P., Rex, D.K., Soetikno, R.M.: Endoscopic prediction of deep submucosal invasive carcinoma: validation of the narrow-band imaging internationl colorectal endoscopic (nice) classification. Clinical Endoscopy **78** (2013)

13. Ferlitsch, M., Moss, A., Hassan, C., Bhandari, P., Dumonceau, J., Paspatis, G., Jover, R., Langner, C., Bronzwaer, M., Nalankilli, K., Lockers, P., Hazzan, R., Gralnek, I.M., Gschwantler, M., Waldmann, E., Jeschek, P., Penz, D., Heresbach, D., Moons, L., Lemmers, A., Paraskeva, K., Pohl, J., Ponchon, T., Regula, J., Repici, A., Rutter, M.D., Burgess, N.G., Bourke, M.J.: Colorectal polypectomy and endoscopic mucosal resection (emr): European society of gastrointestinal endoscopy (esge) clinical guideline. Endoscopy **49** (2017)
14. Zhang, X., Chen, F., Yu, T., An, J., Huang, Z., Liu, J., hu, W., Wang, L., Duan, H., Si, J.: Real-time gastric polyp detection using convolutional neural networks. PloS one **14**, 0214133 (2019). doi:10.1371/journal.pone.0214133
15. Bagheri, M., Mohrekesh, M., Tehrani, M., Najarian, K., Karimi, N., Samavi, S., Reza Soroushmehr, S.M.: Deep neural network based polyp segmentation in colonoscopy images using a combination of color spaces. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 6742–6745 (2019). doi:10.1109/EMBC.2019.8856793
16. Yuan, Y., Meng, M.Q.-H.: Deep learning for polyp recognition in wireless capsule endoscopy images. Medical Physics **44**(4), 1379–1389 (2017). doi:10.1002/mp.12147. https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/mp.12147
17. Ng, A., *et al.*: Sparse autoencoder. CS294A Lecture notes **72**(2011), 1–19 (2011)
18. Zhang, Y., Liu, H., Hu, Q.: Transfuse: Fusing transformers and cnns for medical image segmentation. CoRR **abs/2102.08005** (2021). 2102.08005
19. Kudo, S., Hirota, S., Nakajima, T., Hosobe, S., Kusaka, H., Kobayashi, T., Himori, M., Yagyuu, A.: Colorectal tumours and pit pattern. Journal of Clinical Pathology **47**(10), 880–885 (1994). doi:10.1136/jcp.47.10.880. https://jcp.bmj.com/content/47/10/880.full.pdf
20. Ribeiro, E., Uhl, A., Häfner, M.: Colonic polyp classification with convolutional neural networks. In: 2016 IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS), pp. 253–258 (2016). doi:10.1109/CBMS.2016.39
21. Tanwar, S., Goel, P., Johri, P., Diván, M.: Classification of benign and malignant colorectal polyps using pit pattern classification. SSRN Electronic Journal (2020). doi:10.2139/ssrn.3558374
22. Zhang, R., Zheng, Y., Mak, W., Yu, R., Wong, S., Poon, C.: Automatic detection and classification of colorectal polyps by transferring low-level cnn features from nonmedical domain. IEEE Journal of Biomedical and Health Informatics **PP**, 1–1 (2016). doi:10.1109/JBHI.2016.2635662
23. Byrne, M., Chapados, N., Soudan, F., Oertel, C., Pérez, M., Kelly, R., Iqbal, N., Chandelier, F., Rex, D.: Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model. Gut **68**, 2017 (2017). doi:10.1136/gutjnl-2017-314547
24. Komeda, Y., Handa, H., Watanabe, T., Nomura, T., Kitahashi, M., Sakurai, T., Okamoto, A., Minami, T., Kono, M., Arizumi, T., Takenaka, M., Hagiwara, S., Matsui, S., Nishida, N., Kashida, H., Kudo, M.: Computer-aided diagnosis based on convolutional neural network system for colorectal polyp classification: Preliminary experience. Oncology **93**, 30–34 (2017). doi:10.1159/000481227
25. Hsu, C., Hsu, C., Hsu, Z., Shih, F., Chang, M., Chen, T.: Colorectal polyp image detection and classification through grayscale images and deep learning. sensors (2021)
26. Mesejo, P., Pizarro, D., Abergel, A., Rouquette, O., Beorchia, S., Poincloux, L., Bartoli, A.: Computer-aided classification of gastrointestinal lesions in regular colonoscopy. IEEE Transactions on Medical Imaging **35** (2016)
27. Misawa, M., Kudo, S.-e., Mori, Y., Hotta, K., Ohtsuka, K., Matsuda, T., Saito, S., Kudo, T., Baba, T., Ishida, F., *et al.*: Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). Gastrointestinal endoscopy **93**(4), 960–967 (2021)
28. Krenzer, A., Makowski, K., Hekalo, A., Fitting, D., Troya, J., Zoller, W.G., Hann, A., Puppe, F.: Fast machine learning annotation in the medical domain: a semi-automated video annotation tool for gastroenterologists. BioMedical Engineering OnLine **21**(1), 1–23 (2022)
29. Ribeiro, H., Libanio, D., Castro, R., Ferreira, A., Barreiro, P., Carvalho, P., Capela, T., Pimentel-Nunes, P., Santos, C., Dinis-Ribeiro, M.: Reliability of paris classification for superficial neoplastic gastric lesions improves with training and narrow band imaging. Endoscopy International Open **07**, 633–640 (2019). doi:10.1055/a-0828-7541
30. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
31. Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y., Tao, D.: A Survey on Visual Transformer (2021). 2012.12556
32. Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., Houlsby, N.: Large scale learning of general visual representations for transfer. CoRR **abs/1912.11370** (2019). 1912.11370
33. Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. CoRR **abs/1905.11946** (2019). 1905.11946
34. Wang, Y., Quanming, Y., Kwok, J.T., Ni, L.M.: Generalizing from a few examples: A survey on few-shot learning. ACM Comput. Surv. **1** (2020). doi:10. 1145/3386252
35. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaption of deep networks. Proceedings of the 34th International Conference on Machine Learning, 1126–1135 (2017)
36. Edwards, H., Storkey, A.: Towards a neural statistician. International Conference on Learning Representations (2017)
37. Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., Wierstra, D.: Matching networks for one shot learning. Advances in Neural Informtion Processing Systems **29** (2016)
38. Snell, J., Swersky, K., Zemel, R.S.: Prototypical networks for few-shot learning. Advances in Neural Information Processing Systems **30** (2017)
39. Musgrave, K., Belongie, S., Lim, S.: A metric learning reality check (2020)

40. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature verification using a "siamese" time delay neural network. Advances in Neural Information Processing Systems **6**, 737–744 (1994)
41. Koch, G.: Siamese neural networks for one-shot image recognition. Proceedings of the 32nd International Conference on Machine Learning **37** (2015)
42. Hoffer, E., Ailon, N.: Deep metric learning using triplet network. Similarity-Based Pattern Recognition, 84–92 (2015)
43. Liu, W., Wen, Y., Yu, Z., Yang, M.: Large-margin softmax loss for convolutional neural networks. Proceedings of the Internatinoal Conference on Machine Learning, 507–516 (2016)
44. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (2019)
45. Kaya, M., Bilge, H.: Deep metric learning: A survey. Symmetry **11** (2019)
46. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20**, 273–297 (1995)
47. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2014)
48. Harwood, B., Kumar, B., Carneiro, G., Reid, I., Drummond, T.: Smart mining for deep metric learning. Proceedings of the IEEE International Conference on Computer Vision, 2821–2829 (2017)
49. Arnold, M., Ghosh, A., Ameling, S., Lacey, G.: Automatic segmentation and inpainting of specular highlights for endoscopic imaging. EURASIP Journal on Image and Video Processing (2010)
50. Grill, J., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M.: Bootstrap your own latent. a new approach to self-supervised learning. Advances in Neural Information Processing Systems **33** (2020)
51. Kramer, M.A.: Nonlinear principal component analysis using autoassociative neural networks. AIChE **37** (1991)
52. Schoeffmann, K., Szkaliczki, T., Fabro, M.D., Böszörmenyi, L.: Keyframe exraction in endoscopic video. Multimedia Tools and Applications **74** (2014). doi:0.1007/s11042-014-2224-7
53. Roweis, S., Hinton, G.: Stochastic neighor embedding. Neural Information Processing Systems **15** (2002)
54. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. Proceedings of the International Conference on Computer Vision (2017)
55. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
56. Mongan, J., Moy, L., Kahn Jr, C.E.: Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. Radiological Society of North America (2020)

B Co-author publications

Medical Image Analysis 70 (2021) 102002

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Challenge Report

Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy[☆]

Sharib Ali[a,c,*], Mariia Dmitrieva[a], Noha Ghatwary[g], Sophia Bano[h], Gorkem Polat[j], Alptekin Temizel[j], Adrian Krenzer[k], Amar Hekalo[k], Yun Bo Guo[l], Bogdan Matuszewski[l], Mourad Gridach[x], Irina Voiculescu[m], Vishnusai Yoganand[n], Arnav Chavan[o], Aryan Raj[o], Nhan T. Nguyen[p], Dat Q. Tran[p], Le Duy Huynh[q], Nicolas Boutry[q], Shahadate Rezvy[r], Haijian Chen[s], Yoon Ho Choi[t], Anand Subramanian[u], Velmurugan Balasubramanian[v], Xiaohong W. Gao[r], Hongyu Hu[w], Yusheng Liao[w], Danail Stoyanov[h], Christian Daul[i], Stefano Realdon[e], Renato Cannizzaro[f], Dominique Lamarque[d], Terry Tran-Nguyen[b], Adam Bailey[b,c], Barbara Braden[b,c], James E. East[b,c], Jens Rittscher[a]

[a] *Institute of Biomedical Engineering and Big Data Institute, Old Road Campus, University of Oxford, Oxford, UK*
[b] *Translational Gastroenterology Unit, Experimental Medicine Div., John Radcliffe Hospital, University of Oxford, Oxford, UK*
[c] *Oxford NIHR Biomedical Research Centre, Oxford, UK*
[d] *Université de Versailles St-Quentin en Yvelines, Hôpital Ambroise Paré, France*
[e] *Instituto Onclologico Veneto, IOV-IRCCS, Padova, Italy*
[f] *CRO Centro Riferimento Oncologico IRCCS, Aviano, Italy*
[g] *Computer Engineering Department, Arab Academy for Science and Technology, Alexandria, Egypt*
[h] *Wellcome/EPSRC Centre for Interventional and Surgical Sciences(WEISS) and Department of Computer Science, University College London, London, UK*
[i] *CRAN UMR 7039, University of Lorraine, CNRS, Nancy, France*
[j] *Graduate School of Informatics, Middle East Technical University, Ankara, Turkey*
[k] *Department of Artificial Intelligence and Knowledge Systems, University of Würzburg, Germany*
[l] *School of Engineering, University of Central Lancashire, UK*
[m] *Department of Computer Science, University of Oxford, UK*
[n] *Mimyk Medical Simulations Pvt Ltd, Indian Institute of Science, Bengaluru, India*
[o] *Indian Institute of Technology (ISM), Dhanbad, India*
[p] *Medical Imaging Department, Vingroup Big Data Institute (VinBDI), Hanoi, Vietnam*
[q] *EPITA Research and Development Laboratory (LRDE), F-94270 Le Kremlin-Bicêtre, France*
[r] *School of Science and Technology, Middlesex University London, UK*
[s] *Department of Computer Science, School of Informatics, Xiamen University, China*
[t] *Dept. of Health Sciences & Tech., Samsung Advanced Institute for Health Sciences & Tech. (SAIHST), Sungkyunkwan University, Seoul, Republic of Korea*
[u] *Claritrics India Pvt Ltd, Chennai, India*
[v] *School of Medical Science and Technology, Indian Institute of Technology, Kharagpur, West Bengal, India*
[w] *Shanghai Jiaotong University, Shanghai, China*
[x] *Ibn Zohr University, Computer Science HIT, Agadir, Morocco*

ARTICLE INFO

Article history:
Received 21 October 2020
Revised 4 February 2021
Accepted 11 February 2021
Available online 17 February 2021

ABSTRACT

The Endoscopy Computer Vision Challenge (EndoCV) is a crowd-sourcing initiative to address eminent problems in developing reliable computer aided detection and diagnosis endoscopy systems and suggest a pathway for clinical translation of technologies. Whilst endoscopy is a widely used diagnostic and treatment tool for hollow-organs, there are several core challenges often faced by endoscopists, mainly: 1) presence of multi-class artefacts that hinder their visual interpretation, and 2) difficulty in identifying subtle precancerous precursors and cancer abnormalities. Artefacts often affect the robustness of deep learning methods applied to the gastrointestinal tract organs as they can be confused with tissue of interest. EndoCV2020 challenges are designed to address research questions in these remits. In this paper,

---

[☆] Endoscopy Computer Vision Challenge (EndoCV2020).

[*] Corresponding author.

E-mail address: sharib.ali@eng.ox.ac.uk (S. Ali).

https://doi.org/10.1016/j.media.2021.102002

1361-8415/© 2021 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Keywords:
Endoscopy
Challenge
Artefact
Disease
Detection
Segmentation
Gastroenterology
Deep learning

we present a summary of methods developed by the top 17 teams and provide an objective comparison of state-of-the-art methods and methods designed by the participants for two sub-challenges: i) artefact detection and segmentation (EAD2020), and ii) disease detection and segmentation (EDD2020). Multi-center, multi-organ, multi-class, and multi-modal clinical endoscopy datasets were compiled for both EAD2020 and EDD2020 sub-challenges. The out-of-sample generalization ability of detection algorithms was also evaluated. Whilst most teams focused on accuracy improvements, only a few methods hold credibility for clinical usability. The best performing teams provided solutions to tackle class imbalance, and variabilities in size, origin, modality and occurrences by exploring data augmentation, data fusion, and optimal class thresholding techniques.

© 2021 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license
(http://creativecommons.org/licenses/by-nc-nd/4.0/)

1. Introduction

Endoscopy is a widely used imaging technique for both diagnosis and treatment of patients with complications in hollow organs such as esophagus, stomach, colon, bladder, kidney and nasopharynx. During the endoscopic procedure, an endoscope, a long thin tube with a light source and a camera at its tip, is inserted into the organ cavity. The imaging procedure is usually displayed on a monitor on-the-fly and is often recorded for post analysis. Each organ imposes very specific constraints to the use of endoscopes, but the most common obstructions in all endoscopic surveillance consists of artefacts caused by motion, specularities, low contrast, bubbles, debris, bodily fluid and blood. These artefacts hinder the visual interpretation of clinical endoscopists (Ali et al., 2020c). Missed detection rates of precancerous and cancerous lesions are another limitation. Gastrointestinal (GI) cancer (especially colorectal cancer) has high mortality rates and 5-year relative survival rates for stage IIB is around 65% (Rawla et al., 2019). In general, the missed detection rates in endoscopic surveillance is considerably high, at over 15% (Lee et al., 2017). Therefore, the requirement for technology that can be effectively used in clinical settings during endoscopy imaging is necessary.

While a dedicated endoscopic procedure is followed for each specific organ, often these procedures are very similar, in particular for the GI tract organs like the esophagus, stomach, small intestine, colon and rectum. Notably, some precancerous abnormalities such as inflammation or dysplasia and even cancer lesions in these GI organs naturally look very similar. Often automated methods are only trained for a specific abnormality, organ and imaging modality (Zhang et al., 2019), whereas multiple different types of abnormalities can be present in different organs and several imaging protocols are used during endoscopy. Also, methods that are built for colonoscopy cannot be used during a gastroscopy (in the esophagus, stomach and small intestine), despite the nature and occurrence of many abnormalities being similar in these organs. Artefacts are prevalent in all endoscopy surveillance and are usually confused with lesions, which can lead to unreliable outcomes.

A pathway to develop and reliably deploy methods in clinical settings is by benchmarking methods on a curated multi-center, multi-modal, multi-organ and multi-disease dataset and through a thorough evaluation of built methods using standard imaging metrics and metrics that can test their clinical applicability, for example ranking based on accuracy, robustness and computational efficiency (Ali et al., 2020c). Most publicly available datasets are specific to a particular organ, modality or a single abnormality class, e.g., polyp detection and segmentation challenges (Bernal et al., 2017; Jorge and Aymeric, 2017). While dedicated organ specific challenges help to identify one particular disease type, they do not resemble the clinical workflow where the endoscopists are interested in biopsy and treatment of such abnormalities when of potential threat. For polyp class, it is required to identify different stages of polyp such as benign, dysplastic or cancer. Recently, it was shown that polyps and artefacts can be confused mostly due to specularity (Soberanis-Mukul et al., 2020). Artefacts are the fundamental and inevitable issue in endoscopy that often add confusion in detecting tissue abnormalities in these organs. It is therefore vital to accelerate research in identifying these classes and restore frames where possible (Ali et al., 2021) or reduce the false detections by adding uncertainties for such confusions (Soberanis-Mukul et al., 2020). Other ways to address artefact problems in the endoscopy data is by using synthetically generated frames (Mahmood et al., 2018; Formosa et al., 2020; Incetan et al., 2020). Mahmood et al. (2018) used self-regularized transformer network that allowed to transform the real images into synthetic-like images with preserved clinically-relevant features. This allowed the authors to estimate depth in colonoscopy data robustly without being affected by adverse artefact problems. Incetan et al. (2020) demonstrated the use of a virtual active capsule environment that can simulate wide range of normal and abnormal tissue conditions such as inflated, dry and wet; organ types and endoscopy camera designs in capsule endoscopy. This allowed to optimize the analysis software for varied real conditions.

The Endoscopy Computer Vision Challenge (EndoCV2020)[1] is another crowd-sourcing initiative to address fundamental problems in clinical endoscopy and consists of: 1) Endoscopy artefact detection and segmentation (EAD2020), and 2) Endoscopy disease detection and segmentation (EDD2020). EndoCV2020 releases diverse datasets that include multi-center, multi-modal, multi-organ, multi-disease/abnormality, and multi-class artefacts. Among the two sub-challenges, EAD2020 is an extended sub-challenge of EAD2019 (Ali et al., 2019), however, unlike EAD2019 it includes both frame and sequence data with an addition of nearly 500 frames and a total of 41,832 annotations for detection task and 10,739 for segmentation task.

In this paper, we summarise and analyze the results of the top 17 (out of 43) teams participating in the EndoCV2020 challenge. Additionally, we benchmark these methods with the current state-of-the-art detection and segmentation methods. Each method is also evaluated for its efficacy to detect and segment multi-class instances. In addition to the standard computer vision metrics used to evaluate methods during the challenge, we perform a holistic analysis of individual methods to measure their clinical applicability.

2. Related work

With the advancements in deep learning for computer vision, object detection and segmentation algorithms have shown rapid

---

[1] https://endocv.grand-challenge.org.

development in recent years. This is due to the hidden feature representations provided by Convolutional Neutral Networks (CNNs) that show significant improvement over hand-crafted features. CNN-based methods quickly gained the attention of the Medical Imaging community and are now widely used for automating the diagnosis and treatment for a range of imaging modalities, e.g. radiographs, CT, MRI, and endoscopy imaging. Below we present an overview of the recent deep learning-based object detection and segmentation techniques and discuss the related work in the context to medical image analysis with a particular focus on endoscopy imaging applications.

2.1. Detection and localization

Object detection and localization refers to determining the instances of an object (from a list of predefined object categories) that exist in an image. Object detection approaches can be broadly divided into three categories: single-stage, multi-stage and anchor-free detectors. A brief survey of these is presented below. *Single-stage detectors* Single-stage networks perform a single pass on the data and incorporate anchor boxes to tackle multiple object detection on the same image grid such as in YOLO-v2 (Redmon et al., 2016). Similarly, Liu et al. (2016) proposed the Single Shot MultiBox Detector (SSD) with additional layers to allow detection of multiple scales and aspect ratios. RetinaNet was introduced by Lin et al. (2017b) where the authors introduced focal loss that puts the focus on the sparse hard examples enabling a boost in performance and speed.

The domain of Gastroenterology has started to benefit from the success of single-stage object detectors. Wang et al. (Wang et al., 2018) proposed a model that is based on SegNet (Badrinarayanan et al., 2017) architecture to detect polyps during colonoscopy. Urban et al. (2018) used YOLO to detect polyps from colonoscopy images in real-time. Horie et al. (2019) used SSD to detect superficial and advanced esophagal cancer. RetinaNet was the most popular detector in the first EAD challenge held in 2019. RetinaNet detector with focal loss was used by some top performing teams (Kayser et al., 2019; Oksuz et al., 2019) Multi-stage detectors use a region proposal network to find regions of interest for objects and then a classifier to refine the search to get the final predictions. A two-stage architecture R-CNN using the classical region proposal method was proposed by Girshick et al. (2014) whose speed was improved later by integrating an end-to-end trainable region proposal network (RPN), widely known as Faster R-CNN (Ren et al., 2015). Due to the high precision of the Faster R-CNN, its architecture has become the base for many successful models in the object detection and segmentation domains, such as Cascade R-CNN (Cai and Vasconcelos, 2018) and Mask R-CNN (He et al., 2017). Although these two-stage networks have shown successful results on public datasets such as Pascal VOC (Everingham et al., 2012) and COCO (Lin et al., 2014), they are slow compared to the single-stage object detectors due to their region proposal mechanism.

In the field of Gastroenterology, Yamada et al. (2019) used Faster R-CNN with VGG16 as the backbone to detect challenging lesions which are generally missed by colonoscopy procedures. Their reported prediction speed was not suitable for real-time examination. Shin et al. (2018) detected Polyps using the Fast R-CNN architecture with a region proposal network and an inception ResNet backbone. The two-stage detectors tend to yield better results than their single-stage contemporaries and have performed better at medical image analysis challenges. In the EAD2019 challenge, the top performing team (Suhui Yang, 2019) used a Cascade R-CNN with a feature pyramid network (FPN) module and a ResNet backbone. Similarly, Pengyi and Xiaoqiong (2019) who used Mask aided R-CNN with an ensemble of different ResNet backbones finished second.

Anchor-free detectors A newly emerging detector type are the anchor-free detectors. Single and multi-stage detectors rely on the presence of anchors. Anchor free architectures claim to detect objects while skipping the process of anchor definition. They rely on different geometrical characteristics like the center or corner points of objects (Law and Deng, 2018; Duan et al., 2019). Duan et al. (2019) utilized the upper left and lower right corner to mark an object. The authors used classical backbones to generate a heatmap from the feature map showing potential spots of the object corners. A corner pooling technique was then used to create the classic bounding box of object detection. Zhou et al. (2019) used a similar approach but instead they used a single point as the center of the bounding box.

Because of real-time dependencies in medical applications like the detection of polyps which have to be removed directly (Wang et al., 2019), anchor-free detectors are receiving more attention. Wang et al. (2019) designed an anchor-free automatic polyp detector which achieved the state-of-the-art results while maintaining real-time applicability. Liu et al. (2020) showed an anchor-free detector with state-of-the-art performance while maintaining real-time performance.

2.2. Semantic segmentation

Semantic segmentation involves pixel-level partitioning of an image into multiple segments where each segment represents a pre-defined object or scene category. Based on the success of deep learning approaches on medical imaging data for segmentation, we can divide these approaches broadly into the following groups: *Models based on fully convolutional networks* Fully Convolutional Network (FCN) architectures include only convolutional layers that enable them to take any arbitrary size input image to output a segmentation mask of the same size. These models are mostly based on the architecture developed by Long et al. (2015) for semantic image segmentation.

Sun et al. (2017) proposed a multi-channel FCN (MC-FCN) to segment liver tumors from multi-phase contrast-enhanced CT images. Kaul et al. (2019) proposed FocusNet for skin cancer and lung lesion segmentation. A benchmark study for polyp segmentation using FCNs was conducted by Gao et al. (2017). Similarly, Brandao et al. (2017) used FCN architecture with VGG backbone for a polyp segmentation task. The same group explored integration of depth information to improve segmentation accuracy in their FCN-based model (Brandao et al., 2018).

Models based on encoder-decoder architecture U-Net (Ronneberger et al., 2015), an encoder-decoder architecture, has become widely popular in medical image analysis community. U-Net based models have shown tremendous success, from cell segmentation (Falk et al., 2019) to liver tumor segmentation (Chlebus et al., 2017) and beyond (Sevastopolsky, 2017; Norman et al., 2018).

In endoscopy imaging, U-Net-based models were used for instrument segmentation on GI endoscopy data (Jha et al., 2020). Khan and Choo (2019) developed a model based on U-Net architecture for endoscopy artefact segmentation. Bano et al. (2020) directly used U-Net architecture for segmenting placental vessels from Fetoscopy imaging. Motion induced segmentation exploiting U-Net in the framework was used to segment kidney stones in the Uteroscopy data (Gupta et al., 2020). *Models based on pyramid-based architecture* In both detection and segmentation tasks, a crucial part is being able to identify objects and features of varying scales and sizes. One approach to this problem is to incorporate convolutional feature maps of varying resolutions during classification, which yields information about different scales of the image, making it easier to detect both small and big objects. Such architectures are referred to as *pyramid networks*. PSPNet (Zhao et al., 2017) is one of such design that incorporates global context information for the task of scene parsing using a pyramid pooling module. A similar pyramid-based approach can be found in the task of object detection with Feature Pyramid Network (FPN) (Lin et al., 2017a). FPN extracts feature maps on a per-resolution-basis from the two bottom-up and top-down pathways of a pretrained architecture. The output maps can then be upsampled and concatenated to output a segmentation map (Seferbekov et al., 2018).

Guo et al. (2019) used PSPNet as part of an ensemble model including a U-Net and SegNet architecture for the task of automated polyp segmentation in colonoscopy images. Jia et al. (2020) trained a two-stage polyp detector named PLPNet which utilizes FPN for multiscale feature representation using both CVC-ColonDB (Bernal et al., 2012) and CVC-ClinicDB (Bernal et al., 2015). Their experimental results show that PLPNet outperforms other architectures in most regions on CVC-612 dataset (Bernal et al., 2015) and performs similarly on the ETIS dataset (Silva et al., 2014). Zhang and Xie (2019) utilized an FPN combined with a Cascade R-CNN for artefact detection in endoscopic images. *Models based on dilated convolution architecture* One of the challenges in the construction of semantic segmentation networks is to effectively control the size of the receptive field, providing adequate contextual information for pixel-level decisions while, at the same time, maintaining high spatial resolution and computational efficiency. The *dilated* or *atrous* convolution was proposed to address these challenges (Yu and Koltun, 2015). Chen at al. (2018) proposed a family of very effective semantic segmentation architectures, collectively named DeepLab (also an *encoder-decoder* network), all using the dilated convolution. DeepLabv3+ uses atrous kernels within the spatial pyramid pooling (ASPP) module and depth-wise separable convolution to improve the computational efficiency.

Guo et al. (2020a) proposed a fully convolutional network based on atrous kernels to segment polyps in endoscopy images, with their network winning the GIANA 2017 challenge (Jorge and Aymeric, 2017). Nguyen et al. (2020a) augmented DeepLabv3+ architecture, showing its favourable performance when compared with other state-of-the-art methods on the CVC-ClinicDB Bernal et al. (2015) and ETIS-Larib (Silva et al., 2014) datasets. Ali et at. (2020a) used DeepLabv3+ with ResNet50 backbone to segment Barrett's area from esophageal endoscopy data. Yang and Cheng (2019) developed a model based on DeepLabv3+ for multi-class artefact segmentation used with different backbone architectures.

2.3. Endoscopy computer vision challenges

Biomedical challenges allow to set-up a benchmark for different computer vision methods. Several sub-challenge categories for the development of automated methods for wide-range of problems in endoscopy including surgical instrument segmentation (Ross et al., 2020), robotic scene segmentation (Allan et al., 2020), and computer aided detection and segmentation for polyps (Bernal et al., 2017; 2018) and Barrett's cancer detection[2] have been initiated under MICCAI EndoVis challenge[3]. Endoscopy artefact detection (EAD2019) is another challenge which was first initiated in 2019 and launched in conjunction with IEEE International Symposium on Biomedical Imaging (ISBI) 2019 (Ali et al., 2020c).

3. The endocv challenge: Dataset, evaluation and submission

In this section, we present the dataset compiled for the EndoCV2020 challenge, the protocol used to obtain the ground truth for this data, evaluation metrics that were defined to assess participants methods and a brief summary on the challenge setup and ranking procedure.

3.1. Dataset and challenge tasks

The EndoCV2020 challenge consists of two sub-challenges critical in clinical endoscopy. The EAD2020[4] sub-challenge comprises of diverse endoscopy video frames collected from seven institutions worldwide, including three different modalities and five different human organs (see Fig. 2). Endoscopy video frames were annotated for detection and localization of eight different artefact class occurrences identified by clinical experts in the challenge team. These include specularity, saturation, misc. artefacts, blur, contrast, bubbles, instrument and blood. A total of 280 patient videos from multiple organs and institutions have been used for curating this dataset. Over 45,478 annotations were performed for this challenge on both single frame and sequence video data. Example annotations are shown in Fig. 1. Training data for the detection task consisted of total 2531 frames with 31,069 bounding boxes while 643 frames with 7511 binary masks were released for the segmentation task (except for blur, blood and contrast). The sequence data were sampled by manually observing the amount of changes in artefact categories in the selected sequence. Sequences were required to change from large areas of artefacts to small or no artefact frames and vice versa mimicking natural occurrence in endoscopic procedures. Sequence data for training included 5 sequences (232 frames) for detection and 2 sequences (70 frames) for semantic segmentation tasks sampled from 3 videos of 3 different patients. For the test set, two sequence (80 frames) for detection task were used from 2 independent patient videos. As observed in Fig. 2, due to the nature of occurrence of various artefact classes, the proportion of annotations for each class is different (Fig. 3). However, the proportion of training and test samples per-class were matched in the test data (also see Table 1).

Separately, EDD2020[5] is a new disease detection and segmentation sub-challenge that consists of five disease categories (Ali et al., 2020b). The provided training set consisted of total 385 video frames comprising of 137 different patients used in this study with a total of 817 individual annotations. The annotations included non-dysplastic Barrett's esophagus (NDBE), suspicious, high-grade dysplasia (HGD), cancer, and polyp categories (also see Fig. 1). These disease classes were from three different endoscopic modalities (white light, narrow-band imaging, and chromoendoscopy) acquired from four different clinical centers, investigating four different GI organs. By including varied range of endoscopy data acquired from multiple organs like GI tract and liver in EAD sub-challenge and both upper and lower GI tract data for EDD sub-challenge, EndoCV2020 challenge aimed at developing more general methods that can potentially be applied in different endoscopy routine procedures independent to organ type. To our knowledge, this is the first comprehensive dataset for the multi-class detection and segmentation tasks. More details on the dataset are provided in Fig. 2. The detailed breakdown of training set and test set for each specific task is provided in Table 1.

EndoCV2020 posed three specific challenge tasks (see Fig. 4) that included: 1) detection and localization task, 2) semantic segmentation task and 3) out-of-sample generalization task. For detection and generalization tasks, participants were provided with

---

[2] https://endovissub-barrett.grand-challenge.org.

[3] https://endovis.grand-challenge.org.

[4] https://ead2020.grand-challenge.org.

[5] https://edd2020.grand-challenge.org.

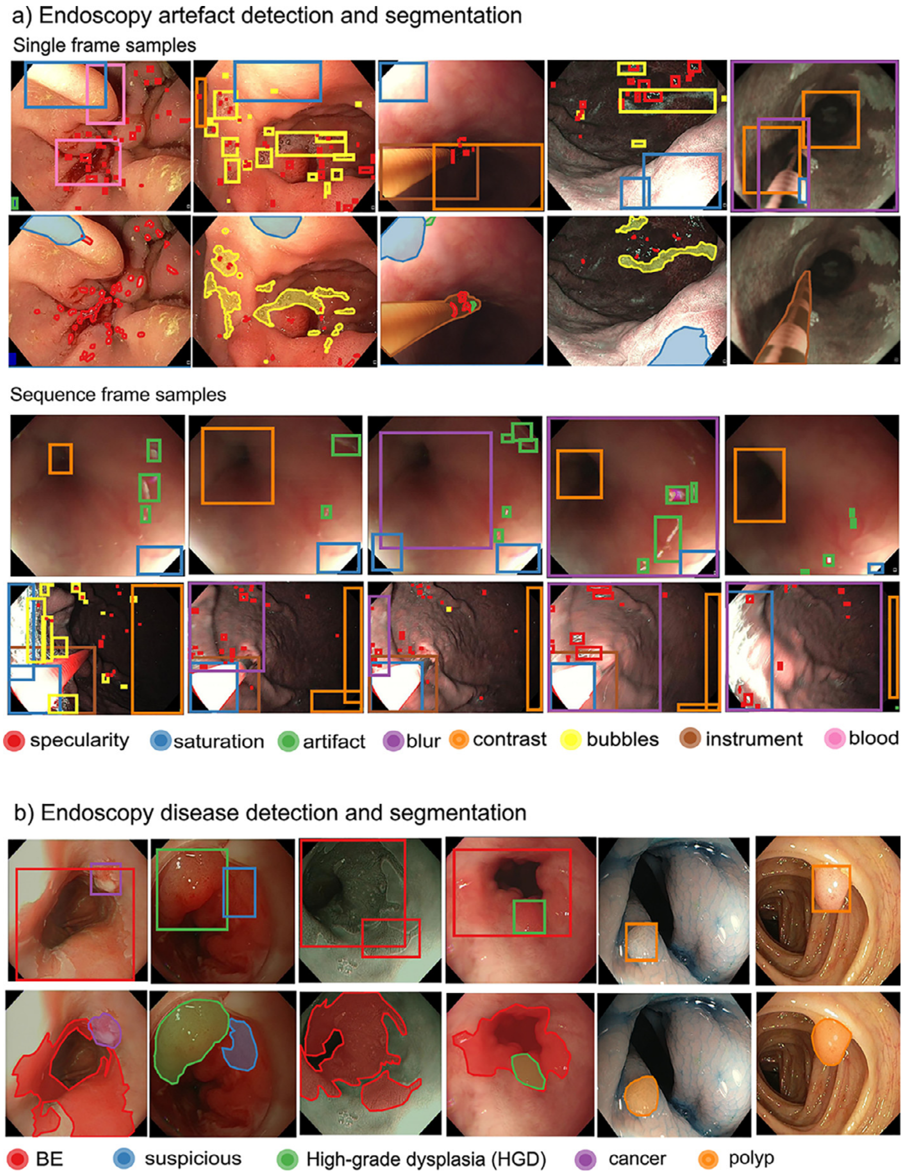
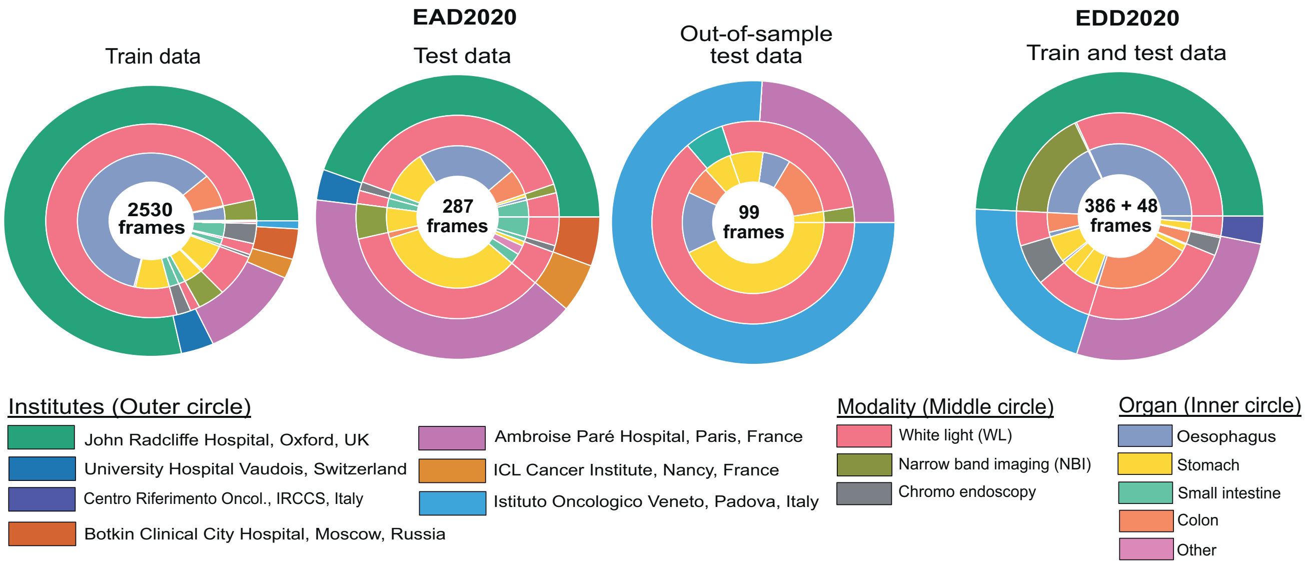


Fig. 1. EndoCV2020 train data samples. (a) Endoscopy artefact detection and segmentation sub-challenge (EAD2020) samples. Both single frame samples (top) and sequence frames (bottom) were released. While detection annotations involve 8 classes, segmentation classes were limited to 5 distinct class instances, mostly large indefinable shapes that include specularity, saturation, imaging artefact, bubbles and instrument. It can be observed that for sequence data most artefact instances follow upto few sequential frames so it is desirable to achieve such training datasets. 4th sample in the single frame data for segmentation shows that even though bounding boxes for detection are provided for all specular regions, some segmentation labels were missing. This shows the presence of annotator variability in the data. (b) Endoscopy disease detection and segmentation training samples for sub-challenge EDD2020. First four samples belong to esophageal endoscopy while the last two frames were acquired during colonoscopy. It can be observed that disease classes in esophagus confuse often, mostly the patient choice here is Barrett's where clearly suspected and high-grade dysplasia appear jointly. Similarly, for colonoscopy data protruded polyps can easily be confused with the surrounding ridge-like openings and specular areas.

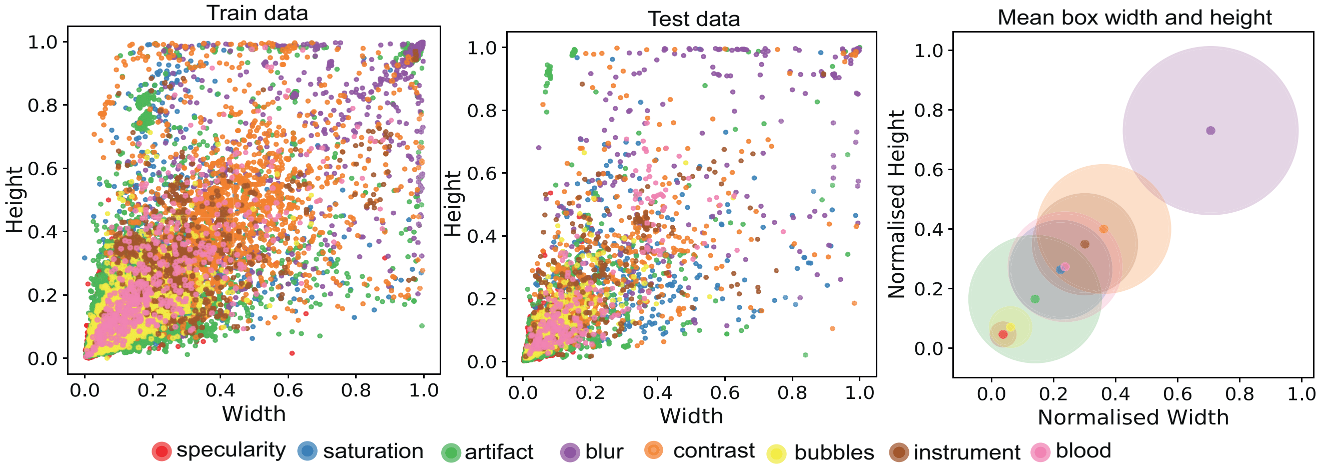
Table 1
Breakdown of data: Number of samples and annotations released for EndoCV2020 challenge.

EndoCV	Tasks	# of classes	# of frames		# of annotations	
			Train	Test	Train	Test
EAD2020	Detection task	8	single: 2299 seq.: 232	single: 237 seq.: 80	31,069	7750
	Segmentation task	5	643	162	7511	3228
	Generalization task	8	na	99	na	3013
EDD2020	Detection task	5	386	43	749	68
	Segmentation task	5	386	43	749	68

a. EndoCV2020 multi-center data cohort: Train and test data for each sub-challenge



b. EAD2020 train and test sample with per class width and height for detection dataset



c. EDD2020 train and test sample with per class width and height for detection dataset

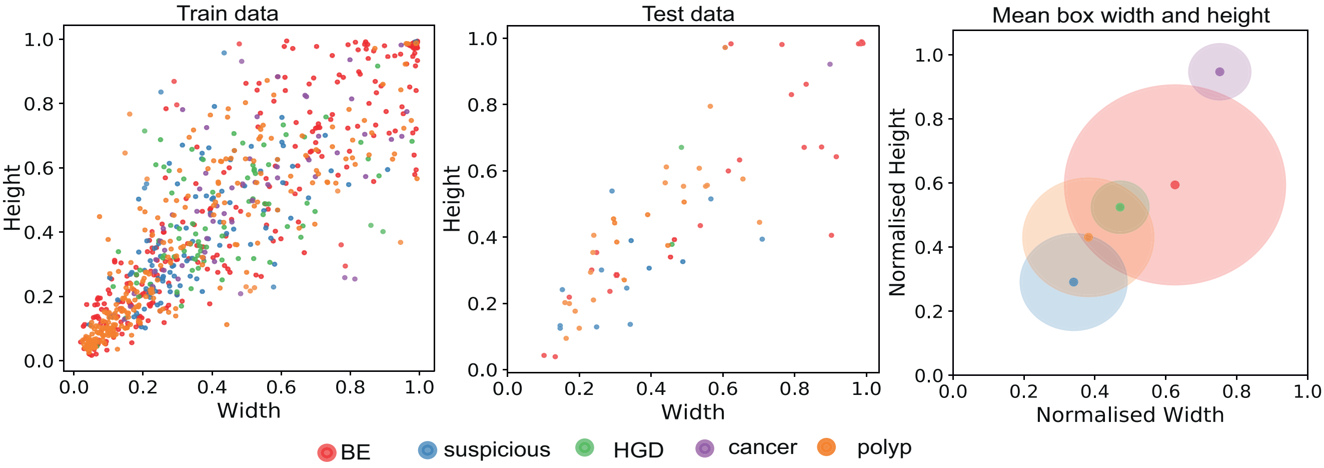


Fig. 2. Endoscopy computer vision EndoCV2020 challenge dataset details. (a) Multi-center, multi-modality and multi-organ dataset for EAD and EDD sub-challenges. For EAD2020, 2532 frames with 8 class bounding boxes for the detection task out-of which 573 included ground truth masks for segmentation task were provided. Participants were assessed on 317 frames for detection and 162 frames for segmentation tasks. An additional 99 frames were used to test out-of-sample generalization task for EAD sub-challenge. While EDD2020 consisted of 384 train samples and 43 test samples for 5 disease classes. (b-c) The distribution of 8 artefact classes of EAD and 5 disease classes of EDD w.r.t. their size compared to their height and width of image is provided. Each class size variability is also shown on right as blobs with mean at center and radius as standard deviation.

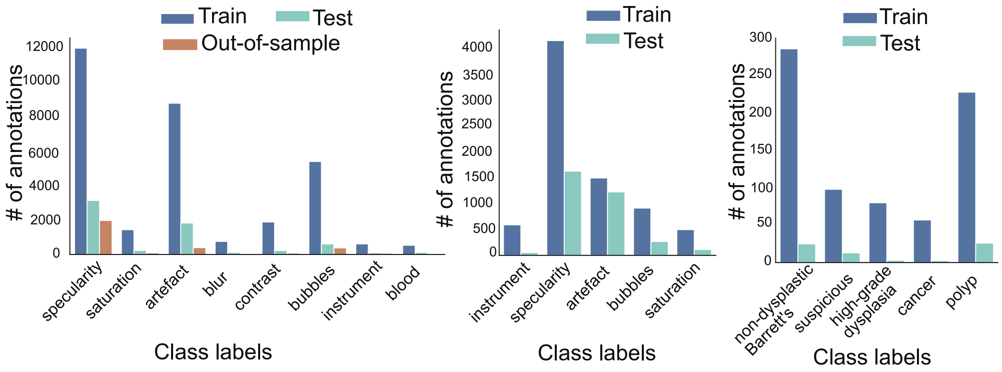


Fig. 3. EndoCV2020 train and test per-class sample proportion: Train and test annotations for sub-challenge on artefact (A,B) and disease (C) detection and segmentation for each class label.

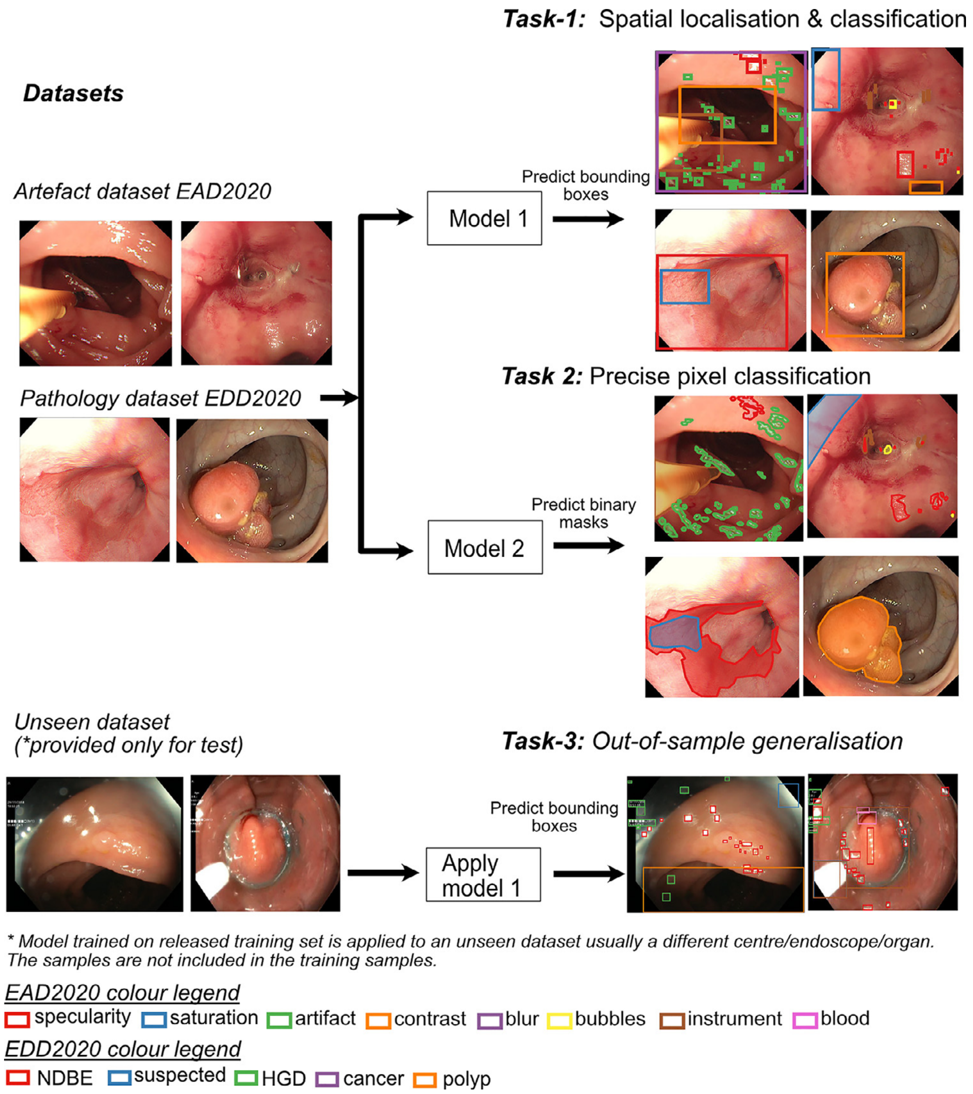

Task-1: Spatial localisation & classification
Datasets
Artefact dataset EAD2020
Predict bounding boxes
Model 1
Pathology dataset EDD2020
Task 2: Precise pixel classification
Predict binary masks
Model 2
Unseen dataset
(*provided only for test)
Task-3: Out-of-sample generalisation
Predict bounding boxes
Apply model 1
* Model trained on released training set is applied to an unseen dataset usually a different centre/endoscope/organ. The samples are not included in the training samples.
EAD2020 colour legend
specularity
saturation
artifact
contrast
blur
bubbles
instrument
blood
EDD2020 colour legend
NDBE
suspected
HGD
cancer
polyp

Fig. 4. EndoCV2020 challenge task descriptions for each sub-challenge. The three tasks of the EndoCV2020 challenge includes: (a) The "detection" task aimed at the coarse localization and classification. Given an input image (left) a detection model (middle) outputs the artefact/disease class and coordinates of the containing bounding box. (b) The "segmentation" task is aimed at precise delineation of artefact/disease object boundaries. The model predicts binary output images denoting the presence ('1') or absence ('0') of each class. (c) The "out-of-sample generalization" task is aimed at assessing the ability of a model trained on different dataset to generalize on an unseen dataset usually coming from a different center.

Table 2
Data collection information for each center: Data acquisition system and patient consenting information.

Centers	System info.	Ethical approval	Patient consenting type
John Radcliffe Hospital, Oxford, UK	Olympus GIF-H260Z, EVIS Lucera CV260	REC Ref: 16/YH/0247	Universal consent
Ambroise Paré Hospital, Paris, France	Olympus Exera 195	N° IDRCB: 2019-A01602-55	Endospectral study
Istituto Oncologico Veneto, Padova, Italy	Olympus endoscope H190	NA	Generic patients consent
Centro Riferimento Oncologico, IRCCS, Italy	Olympus VG-165, CV180, H185	NA	Generic patients consent
ICL, Cancer Institute, Nancy, France	Karl Storz 27005BA	NA	Generic patients consent
University Hospital Vaudois, Switzerland	NA (flexible cystoscopy)	NA	Generic patients consent
Botkin Clinical City Hospital, Moscow, Russia	BioSpec	NA	Generic patients consent

both frame label annotations for single and sequence images for the EAD2020 challenge while only single frames were released for EDD2020. The generalization task was only evaluated for the EAD2020 and only consisted of test data from an unseen institution that was not present in any training set. It is to be noted that test samples for all other tasks were taken from different patients as well even though they were collected from the same centers as that in the training set. EAD2020 attracted nearly 700 participants with 29 teams on the leaderboard and EDD2020 recorded nearly 550 participants with 14 teams on the leaderboard. Participation was permitted in either one or both sub-challenges. Both challenge datasets are publicly available for research and education. EAD2020 challenge data is available at Mendeley Data (https://doi.org/10.17632/c7fjbxcgj9.3) and EDD2020 dataset is available at IEEE dataPort (https://doi.org/10.21227/f8xg-wb80).

3.1.1. Ethical and privacy aspects of the data

Data for EAD2020 were collected from 7 different centers while for EDD2020 were from 4 centers. Each center was responsible for handling the ethical, legal and privacy of the relevant data sent to the challenge organizers. The data collection from each center included either two or all essential steps described below:

1. Patient consenting procedure at the home institution (required)
2. Review of the data collection plan by a local medical ethics committee or an institutional review board
3. Anonymization of the video or image frames (including demographic information) prior to sending to the organizers (required)

Table 2 illustrates the ethical and legal processes fulfilled by each center along with the endoscopy equipment used for the data collected for this challenge.

3.1.2. Annotation protocol

A team of two clinical experts and one post-doctoral researcher determined the class labels for the artefact detection challenge while for the disease detection challenge we consulted with four senior Gastroenterologists (over 20 years experience) regarding the class labels in the GI tract endoscopy. For each sub-challenge senior Gastroenterologists sampled the video frames from a small sub-set of video data collected from various institutions and multi-patient data cohort (see Fig. 2). These frames were then taken as reference to produce bounding box annotations for the remaining train-test dataset by four experienced postdoctoral fellows. Finally, further validation by three clinical endoscopists independently was carried out to assure the reference standard. The ground-truth labels were randomly sampled (1 per 20 frames) during this process. However, after the completion of this phase the entire annotation was discussed and reviewed together with the team of senior Gastroenterologists. Priority was given to indecisive frame annotations to have a collective opinion from experts. Following general annotation strategies were used by clinical experts and researchers:

- For the same region, multiple boxes (for detection/ generalization) or pixel-wise delineation (for semantic segmentation) were performed if the region belonged to more than 1 class
- The minimal box sizes were used to describe the class region and similarly possible small annotation areas for semantic segmentation were merged instead of having multiple small boxes/regions
- Each class type was determined to be distinctive and general across all datasets

For EAD dataset, defined class categories used included below descriptions (Ali et al., 2021). Related samples are presented in Fig. 1(a).

1. blur → fast camera motion
2. bubbles → a thin film of liquid with air that distorts tissue appearance
3. specularity → mirror-like reflection
4. saturation → overexposed bright pixel areas
5. contrast → low contrast areas from underexposure
6. misc. artefact → chromatic aberration, debris etc.
7. instrument → biopsy or any other instrument
8. blood → flow of red colored liquid due to biopsy or surgery

For EDD dataset, both upper-GI (gastroscopy) and lower-GI (colonoscopy) data were used with below defined class categories (please refer to the samples in Fig. 1(b)):

1. NDBE or BE → non-dysplastic Barrett's esophagus determined by a squamo-columnar junction above the gastric fold in the esophagus (Eluri and Shaheen, 2017)
2. HDG → high-grade dysplasia or early adenocarcinoma determined by irregular mucosal appearance (Wang et al., 2012)
3. suspected → aka low-grade dysplasia, an early sign of pathology (Eluri and Shaheen, 2017)
4. cancer → abnormal growth (Boland et al., 2005)
5. polyp → abnormal protrusion of the mucosa (Williams et al., 2013)

For the annotations of disease classes, pathology reports were also used to validate the class category for non-dysplastic Barrett's esophagus (BE), high-grade dysplasia (HGD), suspected (dysplasia or low-grade dysplasia), and cancer categories. That is, expert annotations (three senior gastroenterologists) were taken and supported with the pathology report for most disease categories including some indecisive cases. However, for the polyp class, both the protruded and flat polyps were marked by two experienced post-doctoral researchers and checked by a senior lower-GI specialist (no further categorization based on pathology report was done except for cancer cases).

3.2. Evaluation metrics

The challenge problems fall into three distinct categories. For each there already exist well-defined evaluation metrics used by the wider imaging community which we use for evaluation here.

Codes related to all evaluation metrics used in this challenge are also available online[6].

3.2.1. Spatial localization and classification task

Metrics used for multi-class disease detection:

- IoU - intersection over union: This metric measures the overlap between two bounding boxes A and B, where A is segmented region and B is annotated GT. It is evaluated as the ratio between the overlapped area $A \cap B$ over the total area $A \cup B$ occupied by the two boxes:

$$\mathrm{IoU} = \frac{A \cap B}{A \cup B} \tag{1}$$

where $\cap$, $\cup$ denote the intersection and union respectively. In terms of numbers of true positives (TP), false positives (FP) and false negatives (FN), IoU (aka Jaccard JC) can be defined as:

$$IoU/JC = \frac{TP}{TP + FP + FN} \tag{2}$$

- mAP - mean average precision: mAP of detected class instances is evaluated based on precision (p) defined as $p = \frac{TP}{TP+FP}$ and recall (r) as $r = \frac{TP}{TP+FN}$. This metric measures the ability of an object detector to accurately retrieve all instances of the ground truth bounding boxes. Average precision (AP) is computed as the Area Under Curve (AUC) of the precision-recall curve of detection sampled at all unique recall values $(r_1, r_2, \ldots)$ whenever the maximum precision value drops:

$$\mathrm{AP} = \sum_n \left\{ (r_{n+1} - r_n) p_{\mathrm{interp}}(r_{n+1}) \right\}, \tag{3}$$

with $p_{\mathrm{interp}}(r_{n+1}) = \max_{\tilde{r} \geq r_{n+1}} p(\tilde{r})$. Here, $p(r_n)$ denotes the precision value at a given recall value. This definition ensures monotonically decreasing precision. The mAP is the mean of AP over all N classes given as

$$\mathrm{mAP} = \frac{1}{N} \sum_{i=0}^{N} \mathrm{AP}_i \tag{4}$$

This definition was popularised in the PASCAL VOC challenge (Everingham et al., 2012). The final mAP (mAP_d) was computed as an average mAPs for IoU from 0.25 to 0.75 with a step-size of 0.05 which means an average over 11 IoU levels is used for 5 categories in the competition (mAP @[.25 : .05 : .75]).

Participants were finally ranked on a final mean score ($\mathrm{score_d}$), a weighted score of mAP and IoU represented as:

$$\mathrm{score_d} = 0.6 \times \mathrm{mAP_d} + 0.4 \times \mathrm{IoU_d} \tag{5}$$

Standard deviation between the computed mAPs ($\pm\sigma_{score_d}$) are taken into account when the participants have the same $\mathrm{score_d}$. Scores on both single frame data and sequence data were first separately computed and then averaged to get the final $score_d$ of the detection task.

3.2.2. Segmentation task

Metrics widely used for multi-class semantic segmentation of disease classes have been used for scoring semantic segmentation. The final semantic score $scores_s$ comprises of an average score of F_1-score (Dice Coefficient, DSC), F_2-score, precision (PPV), recall (Rec) and accuracy (Acc).

Precision, recall, F_β-scores:

These measures evaluate the fraction of correctly predicted instances. Given a number of true instances #GT (ground-truth bounding boxes or pixels in image segmentation) and number of predicted instances #Pred by a method, precision is the fraction of predicted instances that were correctly found, $PPV = \frac{\#\mathrm{TP}}{\#\mathrm{Pred.}}$ where #TP denotes number of true positives and recall is the fraction of ground-truth instances that were correctly predicted, $Rec = \frac{\#\mathrm{TP}}{\#\mathrm{GT}}$. Ideally, the best methods should have jointly high precision and recall. F_β-scores gives a single score to capture this desirability through a weighted (β) harmonic means of precision and recall, $\mathrm{F}_\beta = (1 + \beta^2) \cdot \frac{PPV \cdot Rec}{(\beta^2 \cdot PPV) + Rec}$.

Participants are ranked based on the value of their semantic performance score given by:

$$\mathrm{score_s} = 0.25 \times (p + r + \mathrm{F}_1 + \mathrm{F}_2) \tag{6}$$

Standard deviation between each of the subscores are computed and averaged to obtain the final $\pm\sigma_{scores_s}$ which is used during evaluation for participants with same final semantics score. We have also used provided accuracy of each semantic method in this paper for scientific completeness. Accuracy (Acc) can be defined as $Acc = \frac{TP+TN}{TP+TN+FP+FN}$.

3.2.3. Out-of-sample generalization task

Out-of-sample generalization of disease detection is defined as the ability of an algorithm to achieve similar performance when applied to a completely different institution data. To assess this, participants were challenged to apply their trained models on video frames that were neither included in the training nor in the test data of the other tasks. Assuming that participants applied the same trained weights, the out-of-sample generalization ability was estimated as the mean deviation between the mAP score of the detection and out-of-sample generalization test datasets of each class i for deviation greater than a tolerance of $\{0.1 \times \mathrm{mAP_d^i}\}$.

$$\mathrm{dev_g} = \frac{1}{N} \sum_i \mathrm{dev_g}^i \tag{7}$$

$$\mathrm{dev_g}^i = \begin{cases} 0, & \text{for } |\mathrm{mAP_d}^i - \mathrm{mAP_g}^i| / \mathrm{mAP_d}^i \leq 0.1 \\ |\mathrm{mAP_d}^i - \mathrm{mAP_g}^i|, & \text{for } |\mathrm{mAP_d}^i - \mathrm{mAP_g}^i| / \mathrm{mAP_d}^i > 0.1 \end{cases} \tag{8}$$

The best algorithm should have high $\mathrm{mAP_g}$ and low $\mathrm{dev_g}(\rightarrow 0)$. Participants were finally ranked using a weighted ranking *score for out-of-sample generalization* as $\mathrm{R_{gen}} = 1/3 \cdot \mathrm{Rank}(\mathrm{dev_g}) + 2/3 \cdot \mathrm{Rank}(\mathrm{mAP_g})$ where $\mathrm{Rank}(\mathrm{mAP_g})$ is the rank of a participant when sorted by $\mathrm{mAP_g}$ in ascending order.

3.3. Challenge setup, and ranking procedure

The challenge proposal was submitted to the IEEE ISBI challenge organisers and was peer-reviewed by two reviewers. Upon the acceptance, the challenge website[7] was launched on 1st November 2019. Training datasets for each sub-challenge (EAD and EDD) were first provided (via AWS amazon S3 for EAD data and IEEE data portal for EDD data[8]). The test data was released nearly 20 days before the leaderboard closing through a docker container set-up. A docker based online leaderboard was established separately for EAD2020[9] and EDD2020[10] where each participating team was allowed to submit a maximum of 2 submissions per day on the final test data. A wiki-page[11] was set-up for the submission guidelines

---

[6] https://github.com/sharibox/EndoCV2020.

[7] https://endocv.grand-challenge.org.

[8] https://ieee-dataport.org/competitions/endoscopy-disease-detection-and-segmentation-edd2020.

[9] https://ead2020.grand-challenge.org/evaluation/leaderboard/.

[10] https://edd2020.grand-challenge.org/evaluation/leaderboard/.

[11] https://github.com/sharibox/EndoCV2020/wiki.

and a code repository with evaluation metrics used in the challenge was also provided[12].

For the ranking of different task categories, we used the metrics described in Section 3.2. The participants were able to see only the final score in the leaderboard and all other sub-scores were hidden for the final test data. This was done to avoid any class specific refinement on the released test set. Notably, the detection task was bounded by two IoU thresholds (mAP @ IoU thresholds $[.25 : .05 : .75]$) and the overall IoU scores itself. For the detection task, participants were ranked on a final weighted score of mAP and IoU (see Eq. (5)), while for the segmentation task, participants were ranked based on a final weighted average of DSC or F1-score, F2-score, precision and recall (see Eq. (6)). For the generalization task, both the mAP score gap dev_g and mAP on generalization data mAP_g were taken into account.

4. Method summary of the participants

In this Section, we present summary of top participating teams for both EAD2020 and EDD2020 sub-challenges. Each of these teams has participated in either detection task or segmentation task or both.

4.1. EAD2020 Participating teams

- **Team *polatgorkem*** (Polat et al., 2020) The team used an ensemble of three object detectors: Faster R-CNN (ResNet50 with FPN), Cascade R-CNN (ResNet50 with FPN), RetinaNet (ResNet101 with FPN). Class-agnostic NMS operation, where the model predictions were passed through the NMS procedure together for all classes, was applied to the output of each individual model. During ensemble, only the bounding boxes for which majority of the models agree were kept. False-positive elimination was applied as a post-processing step to eliminate same-type predicted boxes located close to each other. For each class, an IoU threshold was determined.
- **Team *CVML*** (Guo et al., 2020b) CVML team's model was inspired by DeepLabV3+. The team experimented with several changes including the backbone, the global pooling, the dilated kernels and the convolution kernels with dilation rates. Moreover, the squeeze-and-excitation module is added behind the balanced ASPP module to introduce attention gating at the output of the original encoder to better utilize the information available in the computed feature maps. In addition, the original multi-class classifier is replaced with 5 binary classifiers to enable segmentation of the overlapping objects. At test time, they used some post-processing techniques such as rotation, holes filling and removal of objects from the image boundary.
- **Team *mouradai_ox*** (Gridach and Voiculescu, 2020) The team proposed a novel neural network called OxEndoNet to tackle the segmentation challenge. The network uses the pyramid dilated module (PDM) consisting of multiple dilated convolutions stacked in parallel. For each input image, pre-trained ResNet50 (on ImageNet) was used as the backbone to extract the feature map followed by multiple PDM layers to form an end-to-end trainable network. In the final architecture, they used four PDM layers; each layer used four parallel dilated convolutions with a filter size of 3×3 and dilation rates of 1, 2, 3, and 4. They fed the final PDM layer to a convolution layer followed by a bilinear interpolation to up-scale the feature map to the original image size.
- **Team *mimykgcp*** (Y et al., 2020) The team re-trained the ResNeXt101 backbone with the cardinality parameter set to 64. To enable detection of artefacts at different scales, an FPN was integrated into the object detectors. Data-Augmentation techniques based on RandAugment (Cubuk et al., 2019) were incorporated to improve the generalization capability. For the segmentation task, a U-Net with an ImageNet pre-trained ResNext50 backbone was used.
- **Team *DuyHUYNH*** (Huynh and Boutry, 2020) For segmentation, the team exploited a model based on U-Net++ using pre-trained EfficientNet on ImageNet as the backbone. The model was trained to minimize F2-loss using the Adam optimizer. At the test-time the team used five transformations: horizontal, vertical flipping, and three rotations. For detection, the team used the bounding boxes deduced from the results of their segmentation model on the EDD dataset, while for EAD, they used YOLOv3 pre-trained on COCO.
- **Team *mathew666*** (Hu and Guo, 2020) The team used Cascade RCNN architecture with the ResNeXt backbone in a FPN based feature extraction paradigm. Data augmentation with probability of 0.5 for horizontal flip was applied. The team also utilised multi-scale detection to tackle with variable sized object detection.
- **Team *arnavchavan04*** (Jadhav et al., 2020) For the object detection task, the team used an ensemble of three models: Faster R-CNN (ResNext101 + FPN), RetinaNet (ResNet101 + FPN) and Faster R-CNN (ResNext101 + DC5). For the segmentation task, an ensemble of multiple depth EfficientNet models with FPN trained on multiple optimization plateaus (DSC, BCE, IoU) was designed. Data augmentation techniques like horizontal and vertical flip, cutout (random holes), random contrast, gamma, brightness, rotation along with CutMix (Yun et al., 2019) strategy for the segmentation task were incorporated to improve generalization capability.
- **Team *anand_subu*** (Subramanian and Srivatsan, 2020) The team used RetinaNet with ResNet101 backbone. For the segmentation task, the team used an ensemble network with U-Net with a ResNet50 backbone and DeepLabV3. However, the team reported U-Net with ResNet101 as their best architecture of choice. All the backbones were pre-trained on the ImageNet. Real-time augmentation techniques like rotation, shear, random-image-flip, image contrast, brightness, saturation, and hue variations were incorporated while training to improve the generalization capability of the network.
- **Team *higersky*** (Chen et al., 2020) The team implemented Hyper Task Cascade and Cascade R-CNN with ResNeXt101 backbone as a feature extractor and FPN module for multi-scale feature representation for the object detection task. They applied Soft-NMS (Bodla et al., 2017) to avoid mistakenly discarded bounding-boxes. For the semantic segmentation task, the team incorporated DeepLabV3+ with ResNet101 backbone and trained with BCE and DICE losses. The backbones for both tasks were pre-trained on ImageNet.
- **Team *MXY*** (Yu and Guo, 2020) The team used a Cascade R-CNN with an ImageNet pre-trained ResNet101 backbone and a FPN module. Post-detection, soft-NMS was added to remove false predictions. The dataset was augmented by random resizing technique to improve the final output scores. The team used more weight for the losses of specularity, artefact, and bubbles classes to overcome classification difficulties between those classes.
- **Team *StarStarG*** The team used Cascade-RCNN as network architecture and adopted COCO2017 pre-trained ResNeXt as backbone with FPN and multi-stage RCNN framework. The authors also integrated Deformable Convolutional Networks in backbone to improve the model performance.
- **Tesam *xiaohong1*** (Gao and Braden, 2020) The team built their detection and segmentation method upon Yolact-based instance

---

[12] https://github.com/sharibox/EndoCV2020.

Table 3
Endoscopy artefact detection and segmentation (EAD2020) method summary for top 13 teams (out-of 33 valid submissions).

Team EAD2020	Algorithm	Preprocessing	Nature	Basis-of-choice	Backbone	Data aug.	Pretrained	Computation		code
Detection								GPU	Test time	
polatgorkem (METU_DLCV)	Faster RCNN + CascadeRCNN + Retinanet	Resize Normalise	Ensemble	Accuracy+	ResNet50, ResNet101	Yes (R, F)[a]	COCO	RTX 2080	0.76	GorkemP/EAD
qzheng5 (CVML)	Faster RCNN	Resize Normalise	Context	Accuracy+	ResNet101	Yes (R, T, LD)[a]	COCO	GTX1060	0.20	CVML/EAD2020
xiaohong1	YOLACT + NMS-within-class	None	Context	Accuracy+, speed+	ResNet101	None	ImageNet	Tesla K80	0.14	yolact
mathew666	Faster RCNN + NMS	None	Context	Accuracy+	ResNet101	Yes	NA	RTX 2080	NA	NA
VinBDI	EfficientDet D0	Resize (512x512)	Multiscale scalable	Speed+	EfficientNet B0	Yes (S, Sc, R, N, MU)[a]	COCO	RTX 2080TI	NA	endocv2020-seg
higersky	Cascade R-CNN	None	Cascading	Accuracy+	ResNeXt101	Yes	NA	GTX1080 Ti	NA	NA
StarStarG	Cascade R-CNN	Resize Normalise	Cascading	Accuracy+	ResNeXt101	Yes (F, S)[a]	NA	RTX 2080	NA	NA
anand_subu	RetinaNet	Resize Normalise	Context	Accuracy+, speed+	ResNet101	Yes (R, Sh, F, C, B, St, H)[a]	ImageNet	GTX1050Ti	0.36	anand-subu/EAD2020
arnavchavan04	RetinaNet + FasterRCNN (FPN + DC5)	Resize (512x512)	Ensemble	Accuracy+	ResNet50; ResNeXt101	Yes (F, C, R)[a]	ImageNet	Tesla T4	NA	ubamba98/EAD2020
MXY	Cascase RCNN + FPN	Resize Normalise	Cascading	Accuracy+	ResNet101	Yes (F)[a]	ImageNet	RTX 2080 Ti	0.80	Carboxy/EAD2020
mimykgcp	Faster RCNN + + RetinaNet	Resize Normalise	Ensemble	Accuracy+, speed+	ResNeXt101	Yes (RA)[a]	COCO	GTX 1080Ti	0.58	NA
DuyHUYNH (LRDE)	YOLOv3	Normalise	Multiscale	Accuracy+, speed++	Darknet53	Yes (RA)[a]	COCO	GTX1080 Ti	0.07	dhuynh/endocv2020
Segmentation										
qzheng5 (CVML)	DeepLabv3+	Resize (513x513) Normalise	Encoder-decoder, mutiscale	Accuracy+	SE-ResNeXt50	(R, T, LD + TTA)[a]	ImageNet	GTX1080Ti	0.50; 5 (+TTA)	CVML/EAD2020
mouradai_ox	Pyramid dilated module	Resize (512x512) Normalise	Multiscale	Accuracy+, speed+	ResNet50	Yes (T, R, LD)[a]	ImageNet	Colab	0.37	NA
arnavchavan04	FPN + EfficientNet	Resize (512x512)	Ensemble	Accuracy+	EfficientNet	Yes (F, C, R)[a]	ImageNet	Tesla T4	NA	ubamba98/EAD2020
VinBDI	U-Net + BiFPN	Resize (512x512)	Ensemble, Endcoder-decoder	Accuracy+, speed+	EfficientNet B4; ResNet50	Yes (S, Sc, R, F)[a]	COCO ImageNet	RTX 2080TI	NA	endocv2020-seg
higersky	DeepLabv3+	None	Encoder-decoder, mutiscale	Accuracy+	ResNet101	Yes (F;S;Sc;Bl)[a]	ImageNet	GTX1080 Ti	NA	NA
anand_subu	U-Net	Resize (512x512)	Encoder-decoder	Accuracy+	ResNet50	Yes (S, F, R, N, Cr, Bl, H, St, C, Sp)[a]	ImageNet	GTX1050Ti	0.17	anand-subu/EAD2020
DuyHUYNH (LRDE)	U-Net+	Normalise	Encoder-decoder	Accuracy+, speed+	EfficientNet B1	Yes (R, S, F, Sc, LD, TTA)[a]	ImageNet	GTX1080 Ti	0.97	dhuynh/endocv2020
mimykgcp	U-Net	Resize Normalise	Encoder-decoder	Accuracy+, speed+	ResNeXt50	Yes (RA)[a]	ImageNet	RTX 2070	0.25	NA

[a] B: brightness, C: contrast, F: Flip, H: hue, LD: Local deformation, N: noise, R: Rotation, RA: RandAugment, S: Shift, Sc: scaling Sh: shear, St: saturation, Mu: mixup, T: Translation, TTA: test-time augmentation

Table 4
Endoscopy disease detection and segmentation (EDD2020) method summary for top 7 teams (out-of 14 submission).

Team EDD2020	Algorithm	Preprocessing	Nature	Basis-of-choice	Backbone	Data aug.	Pretrained	Computation		code
Detection								GPU	Test time	
Adrian	YOLOv3+ Faster R-CNN	Resize	Ensemble	Accuracy+, speed+	Darnet53 ResNet101	Yes (F, D)[a]	COCO public polyp dataset	Tesla P100	0.41	Adrian398/EDD
shahadate	Mask R-CNN	Resize Normalise	Multiscale	Accuracy, speed+	ResNet101	Yes (Sc, R, F, Cr, S, N)[a]	COCO	RTX2060	NA	EDD-Mask-rcnn
VinBDI	EfficientDet D0	Resize (512x512)	Ensemble	Speed+	EfficientNet B0	Yes (S, Sc, R, N, MU)[a]	COCO	RTX 2080TI	NA	endocv2020-seg
YH_Choi	CenterNet	NA	Context	Accuracy+	ResNet50	Yes(Du, R, F, C, B)[a]	PASCAL VOC2012	RTX 2080	2	NA
DuyHUYNH (LRDE)	U-Net+	Normalise	Encoder-decoder	Speed	EfficientNet B1	Yes (R, S, F, Sc, LD, TTA)[a]	ImageNet	GTX1080 Ti	1.53	dhuynh/endocv2020
mimykgcp (vishnusai)	Faster RCNN + RetinaNet	Resize (256x256) normalise	Ensemble	Accuracy+, speed+	ResNeXt101	Yes (RA)[a]	COCO	GTX1080Ti	0.58	NA
Segmentation										
Adrian	YOLOv3 + Faster R-CNN + Cascade RCNN	Resize	Ensemble	Accuracy+	Darnet53 ResNet101	Yes (F, D)[a]	COCO public polyp dataset	Tesla P100		Adrian398/EDD2020
shahadate	MaskRCNN	Resize Normalise	Multiscale	Accuracy, speed+	ResNet101	Yes (Sc, R, F, Cr, S, N)[a]	COCO	RTX2060		EDD-Mask-rcnn
VinBDI	U-Net + BiFPN	Resized (512x512)	Ensemble Endcoder-decoder	Accuracy+, speed+	EfficientNet B4 ResNet50	Yes (S, Sc, R, F)[a]	COCO ImageNet	RTX 2080 Ti	NA	endocv2020-seg
YH_Choi	U-Net	NA	Encoder-decoder	Accuracy+	ResNet50	Yes(Du, R, F, C, B)[a]	PASCAL VOC2012	RTX 2080	7	NA
DuyHUYNH (LRDE)	U-Net+	Normalise	Encoder-decoder	Accuracy+, speed+	EfficientNet B1	Yes (R, S, F, Sc, LD, TTA)[a]	ImageNet	GTX1080 Ti	1.53	endocv2020
drvelmuruganb	SUMNet	NA	Encoder-decoder	Accuracy+, speed++	VGG11	Yes(R, A, Sc, P, and Cr)[a]	ImageNet	GTX1080 Ti	0.16	drvelmuruganb/EDD2020
mimykgcp	U-Net	Resize Normalise	Encoder-decoder	Accuracy+	ResNeXt50	Yes (RA)[a]	ImageNet	RTX2070	1.25	NA

[a] A: affine, B: brightness, C: contrast, Cr: cropping, D: distortion, Du: duplication, F: flip, H: hue, LD: local deformation, Mu: mixup, N: noise, P: perspective transformation, R: rotation, RA: RandAugment library, S: shift, Sc: scaling, Sh: shear, St: saturation, T: translation, TTA: test-time augmentation

segmentation system. Yolact (Bolya et al., 2019) adds a segmentation component to the RetinaNet to ensure the tasks of detection, classification and delineation which are performed simultaneously. The network uses ResNet101 as an imageNet pretrained backbone.

4.2. EDD2020 Participating teams

- **Team *Adrian*** (Krenzer et al., 2020) The team compared two different models: YOLOv3 with darknet-53 backbone and Faster R-CNN with ResNet-101 backbone. For post-processing, both algorithms in the final architecture were combined. For the second task, the team leveraged the state-of-the-art Cascade Mask R-CNN with ResNeXt-151 as a backbone. The team trained YOLOv3 using categorical cross-entropy for classification and default localization loss, while for Cascade Mask-RCNN, they used binary cross entropy for classification and mask, and L1 smooth for boundary box regression.
- **Team *Shahadate*** (Rezvy et al., 2020) The team implemented a modified benchmark Mask R-CNN infrastructure model on the EDD2020 dataset. They used COCO trained weights and biases with the ResNet101 backbone as an initial feature extractor. The network head of the backbone model was replaced with new untrained layers that consisted of a fully-connected classifier with five classes and an additional background class. Non-maximum suppression was used to reduce overlapped detection. Finally, the team merged multiple bounding boxes for the same class label as one bounding box to match with the mask annotation.
- **Team *VinBDI*** (Nguyen et al., 2020b) For the object detection task, the team designed an ensemble of six EfficientDet models (with BiFPN modules) trained on six different EfficientNet backbones. A total of eleven augmentation techniques were incorporated to increase the output prediction scores of the model. For the segmentation task, an ensemble of U-Net and EfficientNet-B4 and BiFPN with the ResNet50 backbone was devised. The same team also participated in the EAD2020 sub-challenge.
- **Team *YH_Choi*** (Choi et al., 2020) The team implemented a CenterNet-based model with the PASCAL VOC pretrained ResNet50 backbone for the object detection task. A similar backbone with U-Net was devised for the segmentation task. The dataset was randomly duplicated to tackle class-imbalance. To improve generalization performance, each image was augmented 86 times by randomly choosing augmentation techniques from the pool of rotation, flipping, contrast enhancement and brightness adjustment.
- **Team *drvelmuruganb*** (Balasubramanian et al., 2020) For the segmentation of disease classes the team used an encoder-decoder based SUMNet architecture with the ImageNet pretrained VGG11 backbone. The authors also applied several augmentation strategies including variable brightness and HSV values, multiple crops and geometric transformations such as rotation, affine, scaling and projective were also applied to improve the accuracy.

5. Results

For the EAD2020 sub-challenge, we present the results of 12 participating teams for multi-class artefact detection task and 8 teams for segmentation task. Similarly, for EDD2020 sub-challenge, we have included top 6 teams for detection and 7 teams for segmentation of multi-class diseases. In this section we present the quantitative and qualitative results for each team based on the evaluation metrics discussed in Section 3.2. For the EAD2020 sub-challenge, 3 different test dataset were released: 1) single-frame data for detection and segmentation, 2) sequence dataset for detection only and 3) out-of-sample data for generalization task only. For the detection task, the average of the aggregated sum of the detection scores for the single frame data and the sequence data were considered for final scoring. While, for the EDD2020 challenge only single frame detection and segmentation data were released. Below we present the result for each sub-challenges separately.

5.1. Quantitative results

5.1.1. EAD2020 Sub-challenge

In this section, the results of the participant teams in the EAD2020 challenge to detect and segment artefacts are presented.

Detection task for EAD2020

Table 5 and Table 6 present the mAP values computed at different IoU thresholds (i.e., 25%, 50%, and 75%), overall mAP, overall IoU, and the final score for the detection of the artefacts from single frame and sequence data, respectively. Additionally, we also provide results of baseline methods that include YOLOv3 and RetinaNet with darknet53 and ResNet101 backbones, respectively. In Table 5 (i.e., single frame detection), it can be observed that the team *polatgorkem* that implemented ensemble technique with Cascaded RCNN, Faster-RCNN and RetinaNet surpassed the other teams by achieving the highest final score on the leaderboard (score_d, Eq. 5) of 25.123 ± 7.124 with the best overall mIoU of 36.579 providing a high overlap ratio between the generated bounding box with ground truth per frame. The method proposed by the team *arnavchavan04* comes in the second place with score_d of 24.079 ± 9.342 with 9% more mAP than the winning team but large sacrifice in the mean IoU. Similarly, for sequence data in Table 6, team *polatgorkem* maintained the first position with a final score of 25.529 ± 10.326. While the second scorer team *VinBDI* suggested a method that obtained a better balanced between mAP and mIoU scores.

Furthermore, Table 7 shows the overall ranking for the teams in terms of Score (R_{score_d}), mAP (R_{mAP}), and generalizability performance (R_g) in addition to, mAP_d, mAP_{seq}, score_d, mAP_g and dev_g. The baseline RetinaNet recorded the least deviation but also the least mAPs. On considering the mAP_g and dev_g together for the final ranking of the generalization task, teams VinBDI and StarStarG secured the first place. On observing at the class-wise performance in Fig. 5 (a) (i.e., single frame), it can be seen that there was a high detection score (score_d) and AP for larger artefact instances such as saturation and contrast. Similarly, most of the teams had a high IoU with the ground truth when detecting the instrument class. On the other hand, the detection and localization of smaller artefact instances such as bubble and saturation showed the degraded performances by all the participating teams and by the baseline methods.

Segmentation task for EAD2020

Table 8 presents the JC, DSC, F2, PPV, recall, and accuracy obtained by each team and baseline methods. As shown, the method proposed by team *arnavchavan04* and team *VinBDI* had the best performance in terms of JC ($>$ 62%), DSC ($>$ 67%), F2 ($>$ 67%) and PPV ($>$ 80%) proving the ability to segment less false positive regions. However, the method suggested by team *qzheng5* and team *DuyHUYNH* segmented more true positive regions compared to other teams obtaining top recall values of 0.8352 and 0.828. The baseline methods showed a low performance in terms of final score compared to the methods proposed by the participants. Furthermore, Fig. 6 (a) shows class-wise scores for DSC, PPV and Recall. Similar to detection, segmenting larger instances like the saturation and the instrument obtained the high scores. Specularity, bubble and the artefact classes were among least performing classes for many teams and baseline methods.

Table 5
EAD2020 results for the detection task on the single frame dataset. mAP at IoU thresholds 25%, 50% and 75% are provided along with overall mAP and overall IoU computations. Overall scores are computed at 11 IoU thresholds and averaged. Weighted detection score $score_d$ is computed between overall mAP and IoU scores only. Three best scores for each metric criteria are in bold.

Team names	mAP$_{25}$	mAP$_{50}$	mAP$_{75}$	overall mAP$_d$	overall mIoU$_d$	mAP$_\delta$	score$_d$ ± δ
polatgorkem	26.886	17.883	5.608	17.486	**36.579**	7.124	**25.123 ± 7.124**
qzheng5	33.134	20.084	5.570	19.720	27.185	8.820	22.706 ± 8.820
xiahong1	30.627	19.384	4.935	18.512	26.388	8.428	21.663 ± 8.428
mathew666	20.360	19.440	7.783	18.091	**32.692**	5.617	**23.931 ±5.617**
VinBDI	38.429	25.426	7.053	24.069	12.644	**10.291**	19.499 ± 10.291
higersky	36.920	25.770	**9.452**	24.771	17.298	8.707	21.781 ± 8.707
StarStarG	**41.800**	**29.984**	**10.733**	**28.380**	16.250	**10.042**	23.528 ± 10.042
anand_subu	29.755	19.893	5.271	18.886	24.029	7.619	20.943 ± 7.619
arnavchavan04	**38.752**	**27.247**	**9.858**	**26.021**	21.165	9.342	**24.079 ±9.342**
MXY	25.373	18.967	7.171	17.82	**28.056**	5.754	21.914 ± 5.754
mimykgcp	**39.897**	**26.296**	6.839	**25.082**	10.209	**10.765**	19.133 ± 10.765
DuyHUYNH	20.512	12.234	2.978	11.894	27.063	5.671	17.962 ± 5.671
baselines							
YOLOv3	22.798	13.736	2.804	13.249	24.883	6.525	17.903 ± 6.525
RetinaNet$ResNet101)	15.270	8.927	2.061	8.754	23.202	4.275	14.533 ± 4.275

Table 6
EAD2020 results for the sequence dataset. mAP at IoU thresholds 25%, 50% and 75% are provided along with overall mAP and overall IoU computations. Overall scores are averaged with 11 IoU thresholds. Weighted detection score $score_d$ is computed between overall mAP and IoU scores only. Three best scores for each metric criteria are in bold.

Team names	mAP$_{25}$	mAP$_{50}$	mAP$_{75}$	overall mAP$_{seq}$	overall mIoU$_{seq}$	mAP$_\delta$	score$_d$ ± δ
polatgorkem	38.464	24.803	4.138	23.137	**29.117**	10.326	**25.529 ± 10.326**
qzheng5	**48.210**	25.717	3.997	25.665	20.949	**14.222**	**23.779 ± 14.222**
xiahong1	46.087	25.813	2.684	25.136	18.398	**15.128**	22.441 ± 15.128
mathew666	31.599	21.878	3.053	19.623	20.858	9.718	20.117 ± 9.718
VinBDI	45.295	**26.723**	4.396	25.285	**23.426**	**13.972**	**24.542 ± 13.972**
higersky	**47.716**	**29.841**	**4.473**	**28.334**	12.865	14.579	22.147 ± 14.579
StarStarG	**46.965**	**30.202**	**5.432**	**28.107**	8.371	13.367	20.213 ± 13.367
anand_subu	38.352	25.535	3.843	23.014	20.703	10.859	22.089 ± 10.859
arnavchavan04	34.511	21.524	**4.886**	20.700	11.827	9.839	17.151 ± 9.839
MXY	31.391	19.838	3.620	18.601	**21.504**	8.688	19.762 ± 8.688
mimykgcp	44.972	26.780	4.400	**25.937**	6.892	13.697	18.319 ± 13.697
DuyHUYNH	28.632	15.524	0.815	15.468	16.968	9.381	16.068 ± 9.381
baselines							
YOLOv3	32.199	18.473	1.137	17.176	16.351	10.596	16.846 ± 10.596
RetinaNet$ResNet101)	17.646	6.447	0.767	8.079	10.000	5.151	9.252 ± 5.151

Table 7
EAD2020 team ranking based on different metric criteria for detection and generalization task. Overall mAPs (mAP$_d$ and mAP$_{seq}$) computed on single frame and sequence data are averaged. Final score$_d$ is then computed as the weighted value between the final IoU$_d$ and the averaged mAP. Rankings for each metric are also provided based on ascending order of the scores except for deviation score for out-of-sample data. Three best scores for each metric criteria are in bold.

Team Names	mAP$_d$	mAP$_{seq}$	final IoU	final score$_d$	mAP$_g$	dev$_g$	R$_{score_d}$	R$_{mAP}$	R$_{gen}$
polatgorkem	17.486	23.137	**32.848**	**25.326**	21.008	9.359	**1**	9	6
qzheng5	19.720	24.174	23.751	**22.668**	23.749	8.522	**2**	6	5
xiahong1	18.512	25.136	22.393	**22.051**	24.579	8.169	**3**	7	**3**
mathew666	18.091	19.651	**26.783**	22.035	16.714	5.674	4	10	4
VinBDI	24.069	25.282	18.033	22.018	**24.140**	5.607	5	4	**1**
higersky	24.771	**28.252**	15.061	21.931	**24.850**	7.686	6	**2**	**2**
StarStarG	**28.380**	**28.107**	12.311	21.870	**25.340**	7.537	7	**1**	**1**
anand_subu	18.886	23.004	22.359	21.510	20.203	7.896	8	8	5
arnavchavan04	**26.021**	20.700	16.496	20.614	21.138	6.968	10	5	**3**
MXY	17.820	18.597	**24.779**	20.836	17.294	6.077	9	11	4
mimykgcp	**25.082**	**25.843**	8.536	18.691	23.929	7.999	11	**3**	4
DuyHUYNH	11.894	15.468	22.016	17.015	11.304	**4.807**	13	13	4
baselines									
YOLOv3	13.249	17.176	20.617	17.374	15.456	**4.397**	12	12	**3**
RetinaNet (ResNet101)	8.754	8.079	16.601	11.690	7.763	**1.985**	14	14	**3**

5.1.2. EDD2020 Sub-challenge

In this section, we report the performance of the participating teams in the EDD2020 challenge for the detection and segmentation.

Detection task for EDD2020

In Table 9, the team *adrian* achieved the highest score among other participants and the baseline methods with a final $score_d$ of 33.602 ± 8.523 with the highest overall mAP (37.594) and the second highest overall mIoU (27.614). The best localization score was obtained by the team *sahadate* but with nearly 5% lower mAP than the top scorer team. Furthermore, the baseline method *RetinaNet* with the ResNet101 backbone performed better than most of the participating teams. From Table 10, it is evident that most teams and baselines failed to detect suspicious class instance while

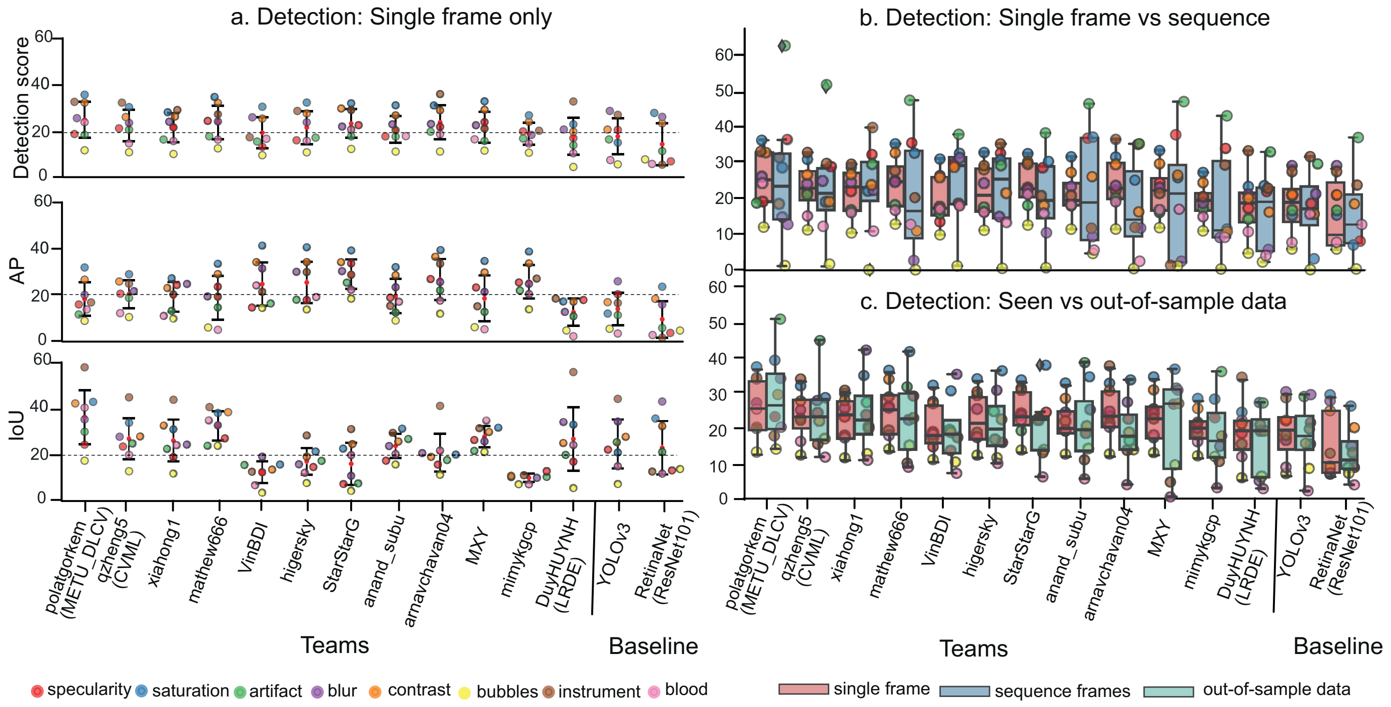


Fig. 5. Detection and out-of-sample generalization tasks for EAD2020 sub-challenge. a) Error bars and swarm plots for the intersection over union (IoU, top), average precision (AP, middle) and challenge detection score (mAP$_d$, bottom) for each team is presented on 237 single frame test data. b-c) Comparison of mAP$_d$ w.r.t. mAP$_{seq}$ (mAP on sequence test data with 80 frames) and mAP$_g$ (mAP on out-of-sample data 99 frames) are provided. a-c) On the right, results from baseline detection methods: YOLOv3 and RetinaNet (with ResNet101 backbone) are also presented. Teams are arranged by decreasing overall detection ranking $\mathbf{R}_{score_d}$ (see Table 7).

Table 8
Evaluation of the artefact segmentation task. Top three best scores for each metric criteria are in bold.

Team Names	JC	DSC	F2	PPV	Rec	Acc	Score$_s$	R$_{score_s}$
qzheng5	0.477	0.532	0.561	0.556	**0.835**	0.973	0.621	8
VinBDI	**0.628**	**0.673**	**0.670**	**0.837**	0.738	**0.978**	**0.730**	**2**
higersky	0.529	0.579	0.587	0.675	0.758	0.975	0.650	5
anand_subu	0.304	0.354	0.361	0.430	0.747	0.975	0.473	14
arnavchavan04	**0.622**	**0.673**	**0.683**	**0.800**	0.767	**0.977**	**0.731**	**1**
DuyHUYNH	0.502	0.557	0.583	0.593	**0.829**	0.974	0.640	6
mimykgcp	0.531	0.576	0.579	**0.723**	0.726	**0.977**	0.651	4
mouradai_ox	**0.581**	**0.632**	**0.647**	0.711	**0.800**	0.974	**0.697**	**3**
baselines								
FCN8	0.500	0.548	0.550	0.670	0.708	0.976	0.619	9
UNet-ResNet34	0.310	0.364	0.373	0.419	0.766	0.974	0.481	13
PSPNet	0.497	0.541	0.534	0.698	0.680	0.975	0.613	10
DeepLabv3 (ResNet50)	0.448	0.495	0.492	0.599	0.704	0.974	0.572	12
DeepLabv3+ (ResNet50)	0.485	0.533	0.535	0.646	0.726	0.976	0.610	11
DeepLabv3+ (ResNet101)	0.501	0.547	0.546	0.683	0.718	0.973	0.624	7

Table 9
EDD2020 results for the detection task on the single frame dataset. mAP at IoU thresholds 25%, 50% and 75% are provided along with overall mAP and overall IoU computations. Overall scores are computed at 11 IoU thresholds and averaged. Weighted detection score *score$_d$* is computed between overall mAP and IoU scores only. Three best scores for each metric criteria are in bold.

Team names	mAP$_{25}$	mAP$_{50}$	mAP$_{75}$	overall mAP$_d$	overall mIoU$_d$	mAP$_\delta$	score$_d$ ± δ
adrian	**48.402**	**33.562**	**27.098**	**37.594**	**27.614**	8.523	**33.602 ± 8.523**
sahadate	**37.612**	**23.284**	**15.837**	**26.834**	**32.420**	8.325	**29.068 ± 8.325**
VinBDI	**43.202**	**26.981**	**17.001**	**30.219**	17.773	**9.478**	25.241 ± 9.478
YHChoi	23.183	11.082	8.800	15.783	24.623	6.216	19.319 ± 6.216
DuyHUYNH	23.959	9.587	5.659	12.479	13.829	6.284	13.019 ± 6.284
mimykgcp	34.884	20.982	4.463	20.742	2.270	**9.359**	13.353 ± 9.359
drvelmuruganb	31.018	18.421	11.768	21.790	7.322	7.424	16.002 ± 7.424
baselines							
YOLOv3	34.305	21.227	14.650	22.980	24.351	6.456	23.528 ± 6.456
RetinaNet (ResNet50)	26.833	14.441	9.907	17.552	25.580	6.464	20.763 ± 6.464
RetinaNet (ResNet101)	42.579	27.000	11.194	27.974	**26.434**	**11.949**	**27.358 ±11.949**

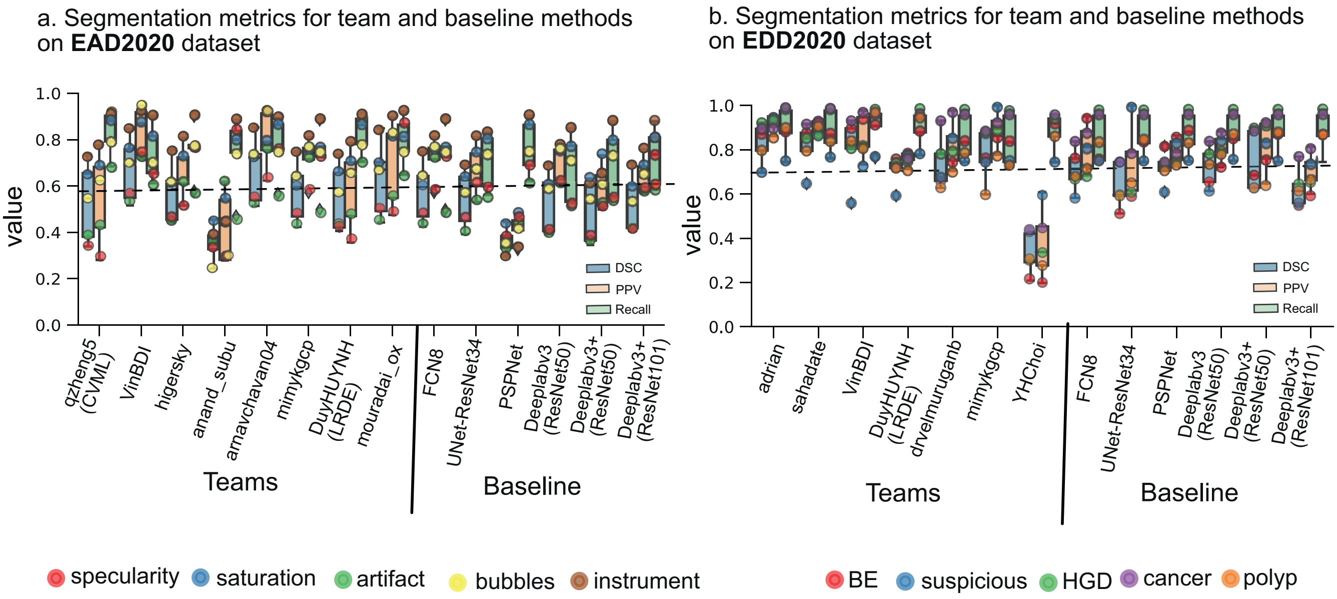


Fig. 6. Semantic segmentation for EAD and EDD sub-challenges: Error bars with overlayed swarm plots for dice similarity coefficient (DSC), positive predictive value (PPV) or precision and recall are presented for each team and baseline methods for the EAD2020 (a) and EDD2020 (b) challenges. 6 different baseline methods are also provided for comparison.

Table 10
Per class evaluation results for the detection task of the EDD2020 sub-challenge.

Teams EDD2020	NDBE	suspicious	HGD	cancer	polyp	δ
adrian	28.911	1.776	32.727	64.286	60.269	22.841
sahadate	**46.193**	1.099	22.727	10.000	54.152	**20.414**
VinBDI	**48.489**	**3.497**	**25.852**	10.000	**63.260**	**22.660**
YHChoi	26.900	0.000	22.727	0.000	29.289	13.057
DuyHUYNH	20.281	1.499	11.364	0.000	29.254	11.134
mimykgcp	**50.089**	**4.592**	**23.064**	5.852	20.112	16.429
drvelmuruganb	34.775	0.000	22.727	0.000	51.446	19.993
baselines						
YOLOv3 (darknet53)	38.839	0.000	6.970	**16.667**	52.426	19.712
RetinaNet (ResNet50)	23.636	0.000	18.182	0.000	45.943	17.086
RetinaNet (ResNet101)	29.483	0.000	22.727	**31.818**	**55.840**	17.909

Table 11
Evaluation of the disease segmentation methods proposed by the participating teams and the baseline methods. Top three evaluation criteria are highlighted in bold.

Team Names	JC	DSC	F2	PPV	Rec	Acc	Score_s	R_{score_s}
adrian	**0.820**	**0.836**	**0.842**	**0.921**	0.894	0.955	**0.873**	**1**
sahadate	**0.797**	**0.816**	**0.819**	**0.906**	0.883	0.955	**0.856**	**2**
VinBDI	**0.788**	**0.805**	**0.812**	**0.859**	**0.912**	0.952	**0.847**	**3**
DuyHUYNH	0.6843	0.7058	0.718	0.762	**0.905**	0.931	0.773	9
drvelmuruganb	0.7166	0.7349	0.734	0.819	0.857	0.959	0.786	6
mimykgcp	0.7561	0.7721	0.770	0.893	0.845	0.957	0.820	4
YHChoi	0.314	0.340	0.356	0.385	**0.896**	0.892	0.494	13
baselines								
FCN8	0.687	0.705	0.709	0.811	0.850	0.953	0.769	10
UNet-ResNet34	0.617	0.637	0.638	0.732	0.868	0.958	0.719	11
pspnet	0.698	0.721	0.723	0.797	0.876	0.959	0.779	8
DeepLabv3 (RetinaNet50)	0.704	0.724	0.724	0.810	0.878	**0.962**	0.784	7
DeepLabv3+ (RetinaNet50)	0.725	0.744	0.749	0.818	0.882	**0.960**	0.798	5
DeepLabv3+ (RetinaNet1010	0.608	0.627	0.629	0.698	0.880	**0.962**	0.709	12

most teams performed comparatively better on polyp and NDBE classes. Only the winning team *adrian* and RetinaNet (ResNet101) provided a descent score for cancer class with most teams recording mAP below 10. For HGD class category, top performing teams were *adrian* and*VinBDI* with mAP over 25.

Segmentation task for EDD2020

From Table 11, it can be observed that the three teams *(Adrian, sahadate* and *nhanthanhnguyen94*) achieved a DSC over 0.80. Moreover, they maintained the high performance for other metrics as well that include JC (>0.78), F2 (>0.81), and PPV (>0.85) securing first, second and third ranks, respectively. Teams *VinBDI* and *DuyHUYNH* were able to segment more true positive regions reaching the top recall values. Fig. 6(b) represents per-class metric values. It can be observed that unlike detection task, most teams reported high performance for cancer class. Also, most teams showed higher DSC, PPV and recall for BE class instance as well (> 0.8 for top

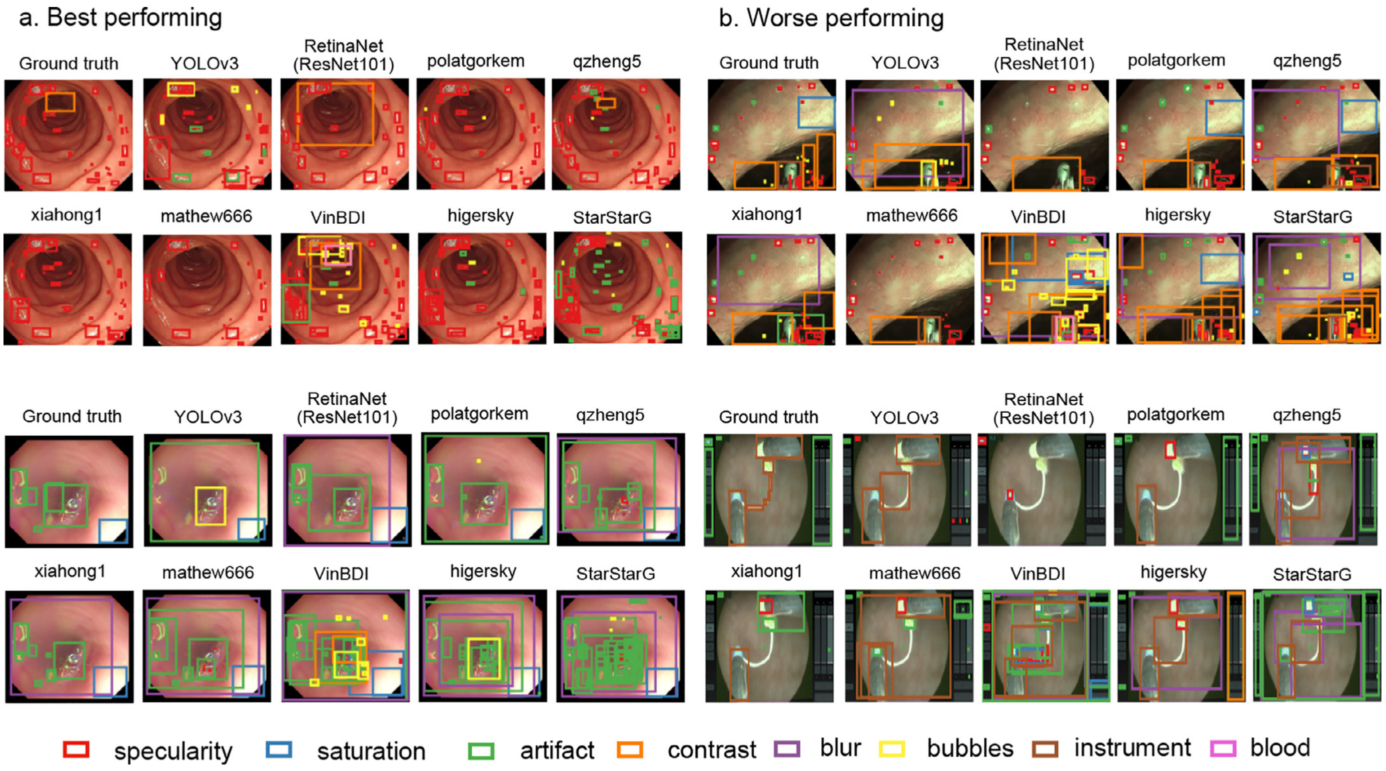


Fig. 7. EAD2020 best and worse performing samples for the detection task. a) Best performing samples for 6 top ranked team results. b) Worse performing samples for the same teams in (a). Results with baseline methods are also included together with ground truth sample.

three teams). However, similar to the detection task, most team and baseline methods reported least values for the suspicious class.

5.2. Qualitative results

Detection task

Fig. 7 shows the best (panel a) and the worse (panel b) performing frames from single frame dataset for EAD2020. It can be observed that specularity and artefacts are detected and well localized by top teams (see Fig. 7 a). Similarly, in the bottom example, saturation is also detected by all the participants. Even though, blur is not present for this sample, most methods also detected it. While for the worse performing frame (see Fig. 7 b), instrument class is confused with contrast or artefact on the top sample, while in the bottom sample instrument is detected by some teams but often either detected only partially or overlapped by different classes such as saturation or artefact.

For out-of-sample generalization task, it can be seen in Fig. 8 (a) that besides YOLOv3 baseline method, all the baselines and teams detected saturation class. While some teams (*mathew666, VinBDI, higersky*) detected multiple bounding boxes for the same class, they also detected blur class for this frame. While for worse performing frame (see Fig. 8(b)), instrument class (at the center of the image) is well localized only by the team *xiahong1* while most teams either partially detected the instrument (e.g., team *qzheng5*) or could not detect the instrument class at all (e.g., team *polatgorkem*). In both cases, the three teams *VinBDI, higersky* and *StarStarG* produced multiple overlapping and different size bounding boxes.

Qualitative results for the EDD2020 challenge is shown in Fig. 9. The best performing samples in Fig. 9(a) shows polyp class (at the top); non-dysplastic Barrett's esophagus (NDBE) and suspicious classes on the bottom. It can be observed that polyp class is detected and well localized by all the teams and baseline methods. However, for bottom row NDBE is detected by most of the methods while confusion is observed across the suspicious class with high-grade dysplasia (HGD) class. Team *mimykgcp* produced numerous bounding boxes failing to optimally localize adherent disease classes. For the worse performing frames (Fig. 9(b)), cancer class (top) in the ground truth is confused with the polyp class instance for most of the teams and the baseline methods. While, for the NDBE class in the bottom of Fig. 9(b), teams were either not able to detect the NDBE class (except team *adrian*, team *YH-Choi* and YOLOv3) at all or partially detected the NDBE areas (e.g., teams *VinBDI* and *drvvelmuruganb*). Again, for the presented case, team *mimykgep* detected numerous bounding boxes.

Segmentation task

Endoscopic artefact segmentation samples representing best and worse performing teams is provided in Fig. 10. For the sample with only the instrument class (see Fig. 10 a, top panel) it can be observed that almost all the baseline and teams were able to predict precise delineation of the instrument class. Similarly, in the bottom panel of Fig. 10(a), specularity, saturation and artefact classes were segmented well by most of the teams and baseline methods. Even though, a single instrument class is present in the sample image in Fig. 10(b), none of the methods were able to segment the instrument. Also, for the bottom panel in the Fig. 10(b), specularity areas were segmented well by the teams *mouradaiox* and *mimykgcp*. However, saturation area was under segmented by most of the teams and baseline methods. Fig. 11 (a) represents the polyp class (at the top); NDBE and suspicious classes (at the bottom). It can be observed that polyp is segmented well by all the baselines and most teams (except team *drvelmuruganb* who misclassified the pixels to suspicious class). While, most teams and baselines were able to precisely delineate NDBE class for the frame in the bottom panel but missed suspicious area. In the worse performing sample (see Fig. 11(b)), most teams were able to segment NDBE area but large HGD area was missed by all the teams. Also, some teams confused HGD area with suspicious class. For the bot-

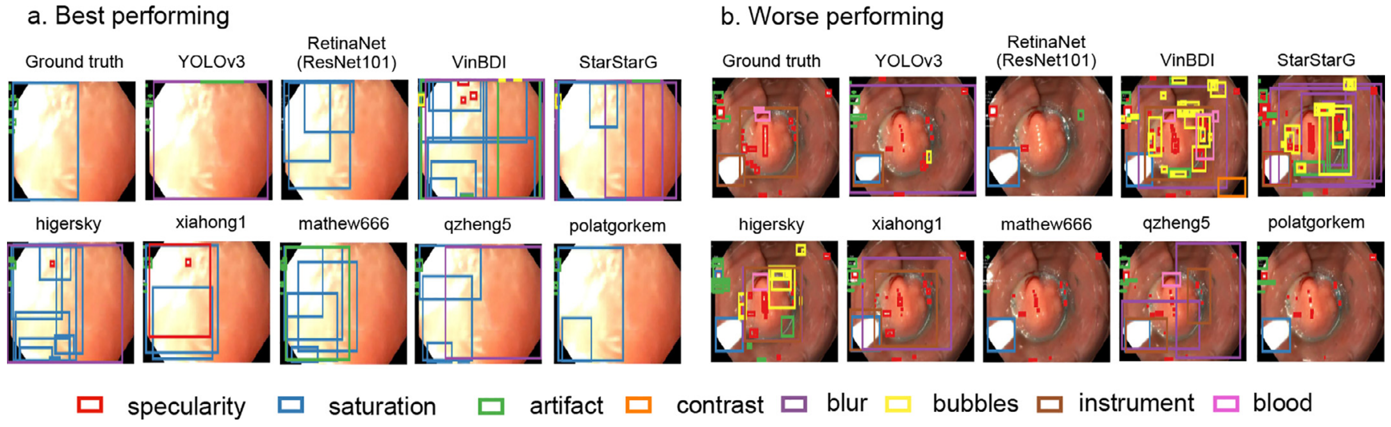


Fig. 8. EAD2020 best and worse performing samples for the generalization task. a) Best performing samples for 7 top ranked team results. b) Worse performing samples for the same teams in (a). Results with baseline methods are also included together with ground truth sample.

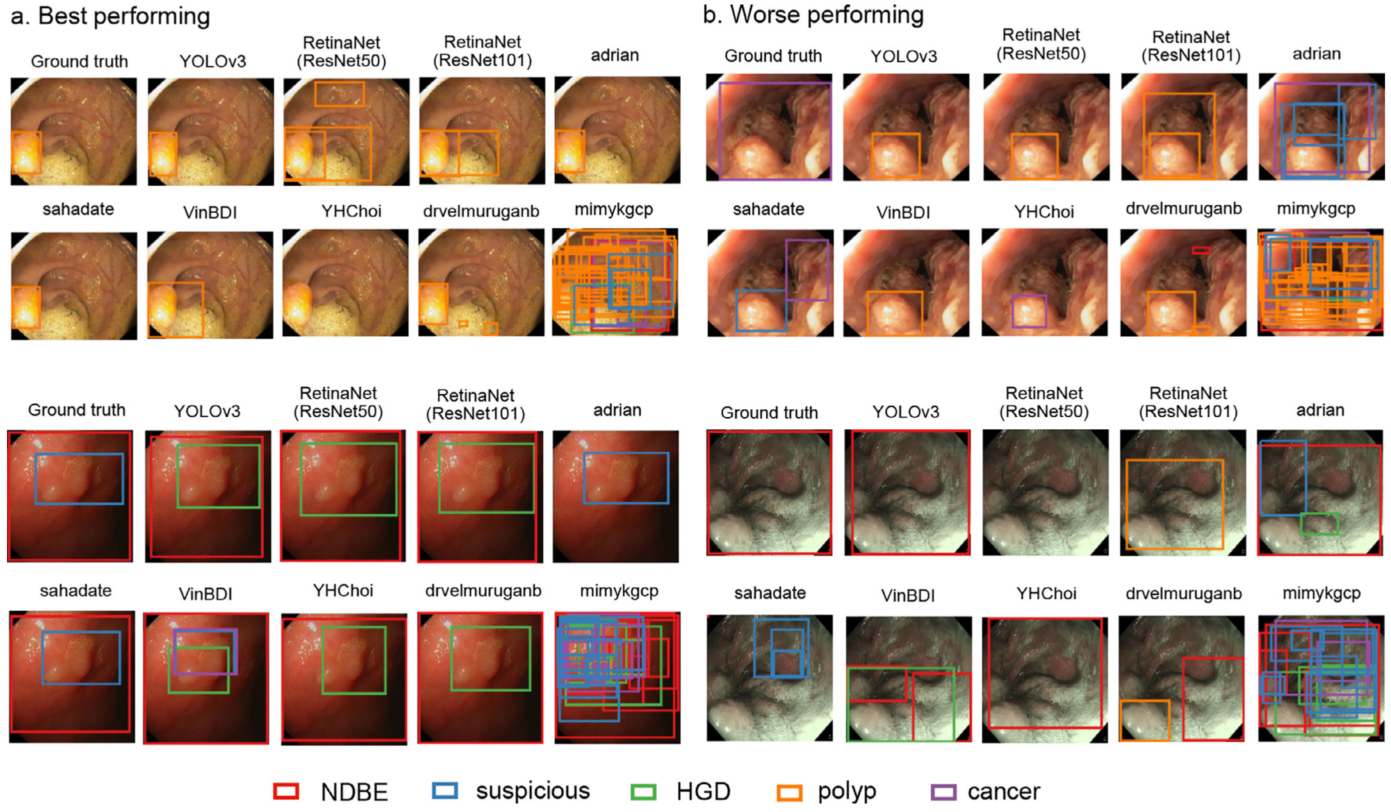


Fig. 9. EDD2020 best and worse performing samples for the detection task. a) Best performing samples for 6 top ranked team results. b) Worse performing samples for the same teams in (a). Results with baseline methods are also included together with ground truth sample.

tom panel in Fig. 11(b), instead of suspicious class present in the ground truth, almost all the teams detected this as polyp or cancer. However, the region delineation was close to the ground truth for most teams.

6. Discussion

Deep learning methods are rapidly being translated for the use of computer aided detection (CADe) and diagnosis (CADx) of diseases in complex clinical settings including endoscopy. However, the amount of data variability particularly in endoscopy is significantly higher than in natural scenes which possess a significant challenge in the process. It is therefore vital to determine an effective translational pathway in endoscopy. Majority of challenges in endoscopy are due to its complex surveillance that lead to severe artefacts that may confuse with disease. Similarly, a system designed for a particular organ may not generalize to be used in the other.

Most deep learning methods that were used in the EndoCV2020 challenge can be categorised into multiscale, symbiotic, ensemble, encoder-decoder and cascading nature, or a combination of these (see Table 3 and Table 4). Fig. 12 presents the overview of the used methods for the detection (a) and segmentation (b) challenge tasks based on the architecture usage. It can be observed that the majority of detection methods used two-stage Faster-RCNN with 4/7 teams combining it with one-stage RetinaNet or YOLOv3 or a combination of all. Cascade R-CNN which is built upon Faster R-CNN cascaded architecture was exploited by 4 teams. Similarly, U-Net-based architectures were utilised by most teams for semantic

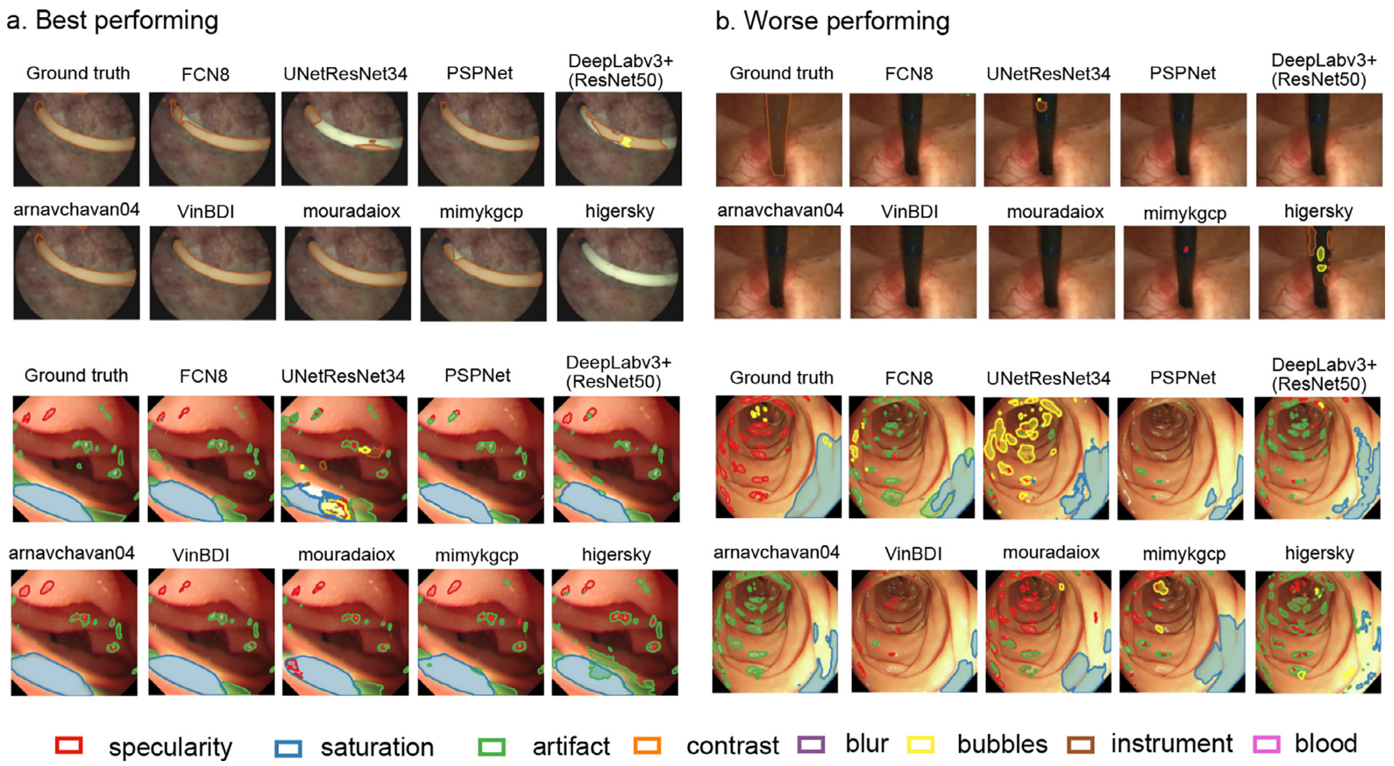


Fig. 10. EAD2020 best and worse performing samples. a) Best performing samples for 5 top ranked team results. b) Worse performing samples for the same teams in (a). Results with baseline methods are also included together with ground truth sample (top). Single class samples are chosen at the top and multi-class samples are at the bottom in each category.

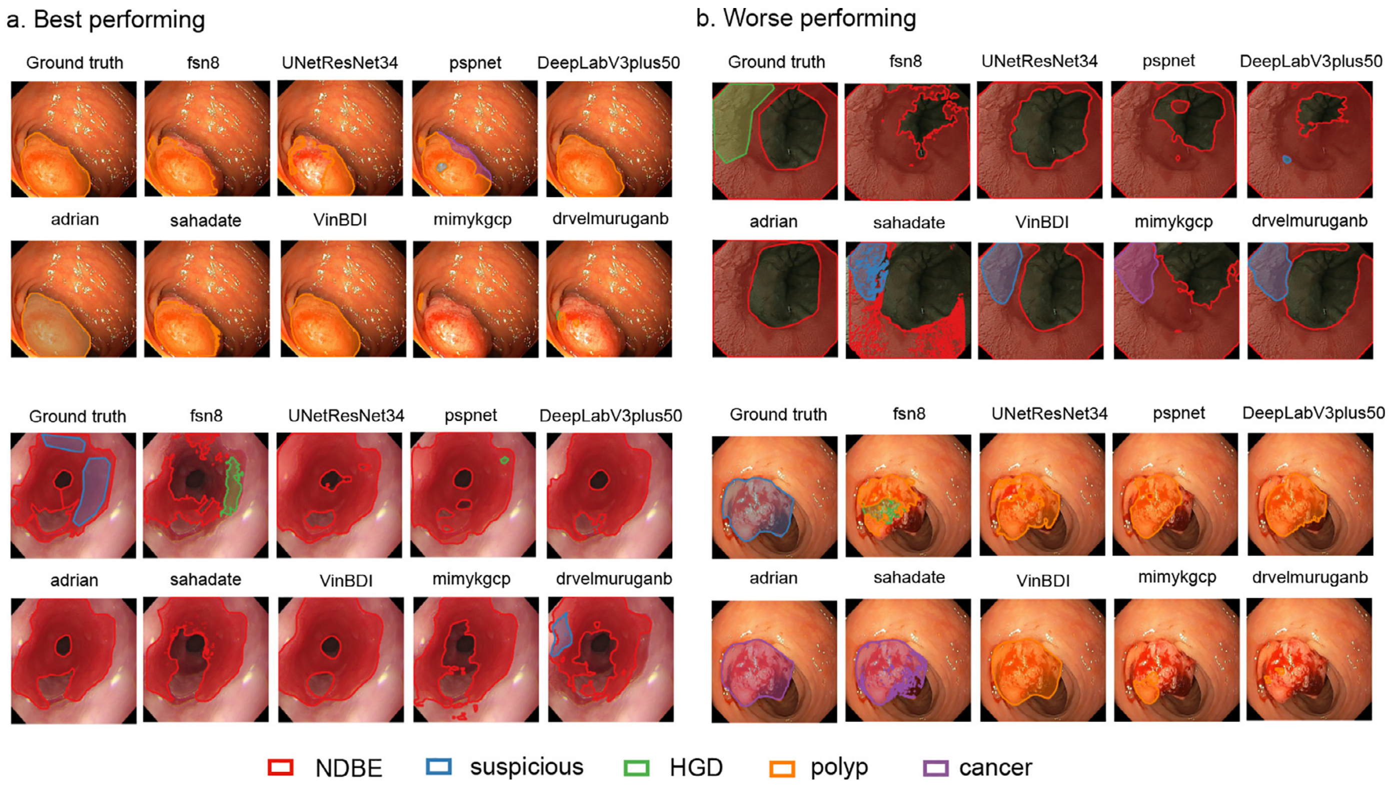


Fig. 11. EDD2020 best and worse performing samples. a) Best performing samples for 5 top team results. b) Worse performing samples for the same teams in (a). Results with baseline methods are also included together with ground truth sample (top).

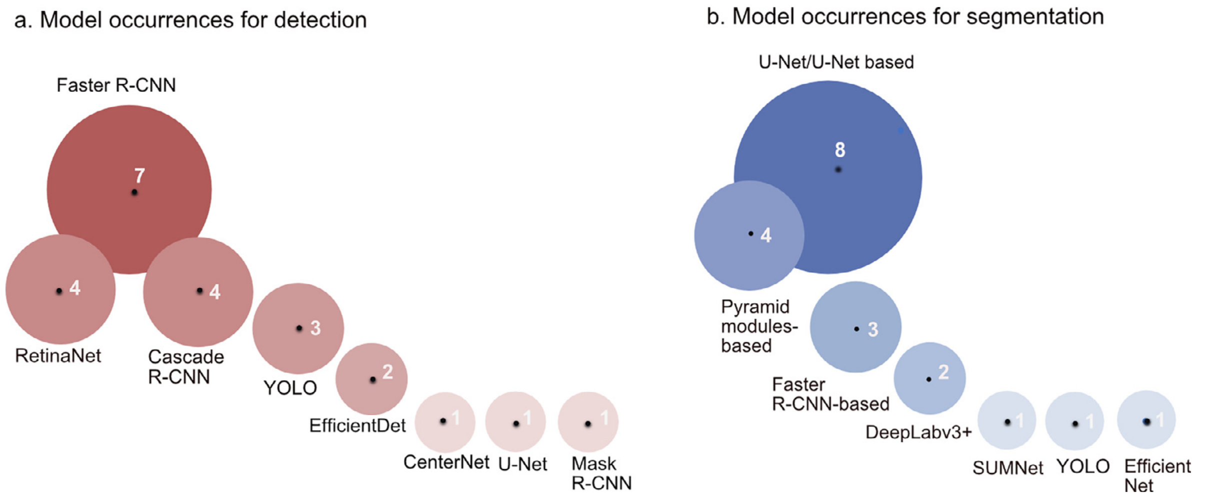


Fig. 12. EndoCV2020 method categories in blob-representation. Model occurrences are presented for detection (a) and segmentation (b) tasks for both EAD2020 and EDD2020 sub-challenges. The number of occurrences is provided inside each blob.

segmentation task with 4 teams exploring pyramid module-based architectures and 2 teams used Deeplabv3+ architecture. Faster RCNN-based model was also explored with additional thresholding (e.g., team *adrian*) or per pixel prediction heads (e.g., team *sahadate*). Even though similar techniques were used in EAD2019 challenge (Ali et al., 2020c), a direct comparison is not possible. This is due to the inclusion of more data for EAD2020 in both train and test sets. Also, EAD2020 includes sequence data which was not provided in EAD2019 challenge.

For the detection task, the top performing teams on the challenge metric in both EAD (team *polatgorkem*) and EDD (team *adrian*) were those using ensemble networks, i.e., maneuvering outputs from multiple architectures. However, these networks sacrifice the speed of detection which can be observed from the computational time which were significantly higher than teams that used a single architecture (see Table 7 and Table 9). Other teams that used such an approach included team *arnavchavan04* and *mimykgcp* who combined Faster R-CNN with RetinaNet but both teams were respectively on 10th and 11th ranking. Just using Faster R-CNN alone with ResNet101 backbone, teams *qzhang5* and *mathew666* were able to detect both small and large size bounding boxes with sub-optimal accuracy that put them at 2nd and 4th positions, respectively. Similarly, team *sahadate* claimed 2nd position on EDD detection task using Mask R-CNN which is based on the Faster R-CNN architecture. For EAD2019 challenge (Ali et al., 2020c), team yangsuhui also used an ensemble network with Cascade RCNN and FPN approach for the detection task similar to the EAD2020 top scorer team *polatgorkem*.

An intelligent choice for improved speed and accuracy using a scalable network was presented by the teams *xiahong1* (used YOLACT) and *VinBDI* (used EfficientDet D0) which were placed 3rd and 5th, respectively, on the final detection score of the EAD2020. On the sequence data, team *VinBDI* was the 2nd best method demonstrating the reliability of the used EfficientNet and FPN architectures. However, for almost all team methods the standard deviation was higher than for single frame data. No team exploited the sequence data provided for training. Team *VinBDI* was also ranked 3rd on the EDD detection task. Teams *higerssky, StarStarG* and *MXY* that used cascaded R-CNN were ranked respectively on 6th, 7th and 9th positions. Additionally, the team StarStarG was ranked 1st and team higersky was ranked 2nd on the overall mAP. However, it is to be noted that taking only mAP scores into account for detection could lead to over detection of the bounding boxes that increases the chance of finding a particular class but at the same time weakens the localization capability of the algorithm (see Fig. 7). Similar observations were found for the EDD dataset where the team *mimykgcp* obtained an overall mAP of 20.742 but only 2.270 for the overall IoU (see Table 9). As a result, over detection of the bounding boxes can be seen in Fig. 9. In order to deal with the over detection of the bounding boxes, YOLACT architecture used by *xiahong1* suppressed the duplicate detections using already-removed detections in parallel (*fast NMS*). Similarly, teams such as *polatgorkem* from the EAD and *adrian* from the EDD were able to eliminate the duplicate detections using ensemble network and a class agnostic NMS.

Hypothesis I: In the presence of multiple class objects, object detection methods may fail to precisely regress the bounding boxes. Methods need better penalization on the bounding box regression or a technique to perform effective non-maximal suppression.

The choice of networks from each team depended on their ambition of either obtaining very high accuracy without focusing on speed or a trade-off between the speed and the accuracy or focusing on both and thinking out-of the box to use more recent developed methods which beats faster networks (such as YOLOv3) that included EfficientDet D0 architecture used by the team VinBDI (see Table 3). Due to the efficiency of the EfficientDet D0 network that used biFPN and efficientNet backbone, team *VinBDI* achieved second least deviation in mAP (i.e., $\text{dev}_g = 5.607$) with competitive mAP_g $(= 24.140)$ and won the generalization task together with the team *StarStarG* who had slightly higher mAP_g $(= 25.340)$ but larger mAP deviation between detection and generalization datasets. Most methods for the detection task on both the EAD and EDD dataset performed better than the baseline one-stage methods (YOLOv3 and RetinaNet). However, it was found that even though team *polatgorkem* won the detection task, the method failed on generalization data where the team was ranked only last. The main reason behind this could be because the generalization gap mAP_g was estimated between two mAP's (mAP_d and mAP_g) and not IoU. Also, the final ranking was done taking into account the rank of dev_g and mAP_g only. It can be observed in Fig. 8 that the bounding box localization of team *polatgorkem* is precise in (a) while it misses instrument area at the center in (b). However, the winning teams *VinBDI* and *StarStarG* both over detect the boxes. The generalization ability of the methods were not explored for EDD dataset.

Hypothesis II: Metrics are critical but using a single metric does not always gives the right answer. Weighted metrics are desired in object detection task to establish a good trade-off between detection and precise localization.

A major problem in the detection of EDD dataset was class confusion mostly for suspicious, HGD and cancer classes. This could be because of smaller number of samples for each of these classes compared to NDBE and polyp (see Fig. 3). While most methods were able to detect and localize NDBE and polyp class in general (3/7 teams with an overall mAP > 45 and 4/7 teams with > 50), all teams failed in suspicious class (overall mAP < 5.0) and most teams for cancer class (overall mAP < 15.0) (see Table 10). Fig. 9 shows that polyp is detected and localized very well by most teams (a, top). Similarly, NDBE is localized by most methods, however, in this case suspicious class is confused mostly with the HGD. Also, in Fig. 9(b, top), it can be observed that the cancer class instance is confused with mostly polyp class.

Hypothesis III: Detection bounding boxes confuse with classes that have similar morphology and smaller number of samples failing to learn the contextual features. To improve detection, such samples need to be identified and more data demonstrating such attributes need to be injected (both positive and negative samples).

Similar to the detection task, teams that used ensemble techniques were among the best performing teams for the segmentation task. Teams *arnavchavan04* and *VinBDI* secured first ($score_s = 0.731$) and second ($score_s = 0.730$) positions, respectively, on the EAD2020 segmentation task (see Table 8) and the team *adrian* won the EDD2020 segmentation task challenge with $score_s$ of 0.873 (see Table 11). The team *arnavchavan04* used multiple augmentation techniques including cutmix and a feature pyramid network with a combination of EfficientNet backbones from B3 to B5. Similarly, team *VinBDI* ensembled a U-Net architecture with EfficientNet B4 and BiFPN network with ResNet50 backbone. Compared to EAD2019 where the winning team yangsuhui used DeepLabV3+ model with two different backbones, both of the top scorer teams of 2020 revealed the strength of recent EfficientNet and FPN-based segmentation approaches.

In the EDD2020 segmentation task, the team *adrian* combined predictions from three object detection architectures where the YOLOv3 and Faster R-CNN class predictions were used to correct the instance segmentation masks from Cascade R-CNN. A direct instance segmentation approach used by the team *sahadate* secured second position ($score_s = 0.856$) on the same while ensemble network of the team *VinBDI* secured the third position ($score_s = 0.847$). Direct usage of a single existing state-of-the-art methods utilising different augmentation techniques (e.g., *DuyHUYNH*) or different backbones (e.g., *mimykgcp, qzheng5*) resulted in improved results compared to the original baseline methods, however, much lower than the top performing methods (see Table 8 and Table 11).

Hypothesis IV: The choice of combinatorial networks that well synthesises width, depth and resolution to capture optimal receptive field, and a domain agnostic knowledge transfer mechanism are critical to tackle heterogeneous (multi-center and variable size) multi-class object segmentation task.

From Fig. 6 it can be observed that the top three performing teams of the EAD2020 segmentation task (*arnavchavan04, VinBDI, mouradai_ox*) has high DSC value (0.538, 0.548 and 0.492 respectively) compared to most methods for the specularity class instance. It is to be noted that the specularities are often confused with either artefact or bubbles which makes them hard to differentiate. For the instrument, saturation and bubbles class instances (see Fig. 10 a.), most methods obtained high performance compared to other classes (e.g., the top three teams obtained 0.853, 0.844, 0.848 for the instrument; 0.722, 0.758, 0.703 for the saturation; and 0.738, 0.693, 0.693 for the bubbles class instance, respectively), artefact (DSC < 0.520) was among the worst class for most teams and for the baseline methods. This is mostly due to the variable size of artefacts; and the bubbles class instance is predominantly confused with either artefact or the specularity class (see Fig. 10 b.). Additionally, due to small sized and sparsely scattered specularity or bubble regions in some cases (for e.g., 4th image from left in Fig. 3(a)), the annotator variability for these samples can have affected method performances for these classes. While checking for such biases is beyond the conducted study, we refer to the work by Rolnick et al. (2017). The authors suggested that in general deep learning models are capable of generalizing from training data where the correct labels are outnumbered by the incorrect ones. However, the authors also acknowledged that a decrease in performance is inevitable and necessary steps such as using larger batch size and downscaling learning rate can help mitigate these issues.

Unlike the EAD2020, the EDD2020 segmentation task comprised of larger shaped regions and only a few classes confused (see 1 b.). Most methods scored comparably high DSC values with over 75% for most of the disease classes except for suspicious class by most of the team. However, Fig. 11(b) (top) shows that while majority of teams were able to segment NDBE class area, the teams either missed the HGD area or miss classified HGD as suspicious class instance. It is to be noted that there is a very subtle difference between the HGD and the suspicious region even for the expert endoscopists. Similar observation can be found for the segmentation of protruded structures (Fig. 11(b), bottom) where most methods confused the class with the polyp class and the top two teams (*adrian, sahadate*) classified it as cancer class. Looking up into our expert consensus notes we found that these samples had hard to reach agreement cases (i.e., suspicious and HGD classes; and cancer and polyp region).

Hypothesis V: Instead of hard scoring of predicted mask classes that penalizes the method performance heavily in presence of marginal visual difference between classes and variability due to existing expert consensus in the dataset, probability maps can be used to mitigate such problem. Additionally, teams should be encouraged to report results for different batch size and learning rates for obtaining better insight regarding performance especially when datasets are prone to have some incorrect labels.

7. Conclusion

We provided a comprehensive analysis of the deep learning methods built to tackle two distinct challenges in the gastrointestinal endoscopy: a) artefact detection and segmentation and b) disease detection and segmentation. It has been possible by the crowd-sourcing initiative of the EndoCV2020 challenges. We have laid out the summary of the methods developed by the top 17 participating teams and compared their methods with the state-of-the-art detection and segmentation methods. Additionally, we dissected-different paradigms used by the teams and present a detailed analysis and discussion of the outcomes. We also suggested pathways to improve the methods for building reliable and clinically transferable methods. In future, we aim towards more holistic comparison of the built methods for clinical deployability by testing for hardware and software reliability in clinical setting.

Author contributions

S. Ali conceptualized the work, led the challenge and workshop, prepared the dataset, software and performed all analyses. M. Dmitrieva, N. Ghatwary, and S. Bano served as organising committee and participated in annotations. A. Bailey, B. Braden, J.E. East, R. Cannizzaro, D. Lamarque, S. Realdon were involved in the validation and quality checks of the annotations used in this challenge. G. Plolat, A. Temizel, A. Krenzer, A. Hekalo, YB. Guo, B. Matuszewski, M. Gridach, V. Yoganand assisted in compiling the related work and method section of the manuscript. S. Ali wrote most of the manuscript with inputs from M. Dmitrieva, N. Ghat-

wary, S. Bano and all co-authors. All authors participated in the revision of this manuscript and provided substantial input.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowlgedgments

The research was supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health. S. Ali, B. Braden, A. Bailey and J. E. East is supported by NIHR BRC and J. Rittscher by Ludwig Institute for Cancer Research and EPSRC Seebibyte Programme Grant (EP/M013774/1). We want to acknowledge Karl Storz for co-sponsoring the challenge workshop. We would also like to acknowledge the annotators and our EndoCV2020 challenge workshop proceedings reviewers and IEEE International Symposium on Biomedical Imaging 2020 challenge committee Michal Kozubek and Hans Johnson.

References

Ali, S., Bailey, A., East, J.E., Leedham, S.J., Haghighat, M., Investigators, T., Lu, X., Rittscher, J., Braden, B., 2020. Artificial intelligence-driven real-time 3D surface quantification of barrett's oesophagus for risk stratification and therapeutic response monitoring. medRxiv doi:10.1101/2020.10.04.20206482.

Ali, S., Ghatwary, N.M., Braden, B., Lamarque, D., Bailey, A., Realdon, S., Cannizzaro, R., Rittscher, J., Daul, C., East, J., 2020. Endoscopy disease detection challenge 2020. CoRR abs/2003.03376.

Ali, S., Zhou, F., Bailey, A., Braden, B., East, J.E., Lu, X., Rittscher, J., 2021. A deep learning framework for quality assessment and restoration in video endoscopy. Med. Image Anal. 68, 101900. doi:10.1016/j.media.2020.101900.

Ali, S., Zhou, F., Braden, B., Bailey, A., Yang, S., Cheng, G., Zhang, P., Li, X., Kayser, M., Soberanis-Mukul, R.D., et al., 2020. An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy. Sci. Rep. 10 (1), 1–15.

Ali, S., Zhou, F., Daul, C., Braden, B., Bailey, A., Realdon, S., East, J., Wagnières, G., Loschenov, V., Grisan, E., et al., 2019. Endoscopy artifact detection (EAD 2019) challenge dataset. arXiv preprint arXiv:1905.03209.

Allan, M., Kondo, S., Bodenstedt, S., Leger, S., Kadkhodamohammadi, R., Luengo, I., Fuentes, F., Flouty, E., Mohammed, A., Pedersen, M., Kori, A., Alex, V., Krishnamurthi, G., Rauber, D., Mendel, R., Palm, C., Bano, S., Saibro, G., Shih, C.-S., Chiang, H.-A., Zhuang, J., Yang, J., Iglovikov, V., Dobrenkii, A., Reddiboina, M., Reddy, A., Liu, X., Gao, C., Unberath, M., Kim, M., Kim, C., Kim, C., Kim, H., Lee, G., Ullah, I., Luna, M., Park, S. H., Azizian, M., Stoyanov, D., Maier-Hein, L., Speidel, S., 2020. 2018 robotic scene segmentation challenge. 2001.11190.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal. Mach. Intell. 39 (12), 2481–2495.

Balasubramanian, V., Kumar, R., Kamireddi, S.J., Sathish, R., Sheet, D., 2020. Semantic segmentation, detection AND localisation of mucosal lesions from gastrointestinal endoscopic images using SUMNET. In: Proceedings of the 2nd International Workshop and Challenge on Computer Vision in Endoscopy, EndoCV@ISBI 2020, Iowa City, Iowa, USA, 3rd April 2020. In: CEUR Workshop Proceedings, 2595, pp. 82–83.

Bano, S., Vasconcelos, F., Shepherd, L.M., Poorten, E.V., Vercauteren, T., Ourselin, S., David, A.L., Deprest, J., Stoyanov, D., 2020. Deep placental vessel segmentation for fetoscopic mosaicking. arXiv preprint arXiv:2007.04349.

Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F., 2015. Wm-dova maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians. Computeri. Med. Imag. and Graph. 43, 99–111.

Bernal, J., Sánchez, J., Vilarino, F., 2012. Towards automatic polyp detection with a polyp appearance model. Patt. Recognit. 45 (9), 3166–3182.

Bernal, J., et al., 2017. Comparative validation of polyp detection methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge. IEEE Trans. Med. Imag 36 (6), 1231–1249.

Bernal, J., et al., 2018. Polyp detection benchmark in colonoscopy videos using gtcreator: A novel fully configurable tool for easy and fast annotation of image databases. In: Proc. Comput. Assist. Radiol. Surg. (CARS).

Bodla, N., Singh, B., Chellappa, R., Davis, L.S., 2017. Soft-nms–improving object detection with one line of code. In: Proceedings of the IEEE international conference on computer vision, pp. 5561–5569.

Boland, C., Luciani, M., Gasche, C., A., G., 2005. Infection, inflammation, and gastrointestinal cancer. Gut 54 (9), 1321–1331. doi:10.1136/gut.2004.060079.

Bolya, D., Zhou, C., Xiao, F., Lee, Y.J., 2019. YOLACT: Real-time instance segmentation. IEEE international conference on computer vision.

Brandao, P., Mazomenos, E., Ciuti, G., Cali, R., Bianchi, F., Menciassi, A., Dario, P., Koulaouzidis, A., Arezzo, A., Stoyanov, D., 2017. Fully convolutional neural networks for polyp segmentation in colonoscopy. In: Medical Imaging 2017: Computer-Aided Diagnosis. International Society for Optics and Photonics. SPIE, pp. 101–107. doi:10.1117/12.2254361.

Brandao, P., Zisimopoulos, O., Mazomenos, E., Ciuti, G., Bernal, J., Visentini-Scarzanella, M., Menciassi, A., Dario, P., Koulaouzidis, A., Arezzo, A., Hawkes, D.J., Stoyanov, D., 2018. Towards a computed-aided diagnosis system in colonoscopy: automatic polyp segmentation using convolution neural networks. Journal of Medical Robotics Research 03 (02), 1840002. doi:10.1142/S2424905X18400020.

Cai, Z., Vasconcelos, N., 2018. Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6154–6162.

Chen, H., Lian, C., Wang, L., 2020. Endoscopy artefact detection and segmentation using deep convolutional neural network. In: Proceedings of the 2nd International Workshop and Challenge on Computer Vision in Endoscopy, EndoCV@ISBI 2020, Iowa City, Iowa, USA, 3rd April 2020. CEUR-WS.org, pp. 37–41.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV), pp. 801–818.

Chlebus, G., Meine, H., Moltz, J.H., Schenk, A., 2017. Neural network-based automatic liver tumor segmentation with random forest-based candidate filtering. arXiv preprint arXiv:1706.00842.

Choi, Y.H., Lee, Y.C., Hong, S., Kim, J., Won, H., Kim, T., 2020. Centernet-based detection model and u-net-based multi-class segmentation model for gastrointestinal diseases. In: Proceedings of the 2nd International Workshop and Challenge on Computer Vision in Endoscopy, EndoCV@ISBI 2020, Iowa City, Iowa, USA, 3rd April 2020. CEUR-WS.org, pp. 73–75.

Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V., 2019. Randaugment: practical data augmentation with no separate search. arXiv preprint arXiv:1909.13719 2 (4), 7.

Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., Tian, Q., 2019. Centernet: Keypoint triplets for object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 6569–6578.

Eluri, S., Shaheen, N., 2017. Barrett'S esophagus: diagnosis and management. Gastrointest. Endosc. 85 (45), 889–903. doi:10.1016/j.gie.2017.01.007.

Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A., 2012. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

Falk, T., Mai, D., Bensch, R., Çiçek, Ö., Abdulkadir, A., Marrakchi, Y., Böhm, A., Deubner, J., Jäckel, Z., Seiwald, K., Dovzhenko, A., Tietz, O., Dal Bosco, C., Walsh, S., Saltukoglu, D., Tay, T.L., Prinz, M., Palme, K., Simons, M., Diester, I., Brox, T., Ronneberger, O., 2019. U-Net: Deep learning for cell counting, detection, and morphometry. Nat. Methods 16 (1), 67–70. doi:10.1038/s41592-018-0261-2.

Formosa, G.A., Micah Prendergast, J., Sean Humbert, J., Rentschler, M.E., 2020. Nonlinear dynamic modeling of a robotic endoscopy platform on synthetic tissue substrates. J Dyn Syst Meas Control 143 (1). doi:10.1115/1.4048190.

Gao, J., Vázquez, D., Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., López, A.M., Romero, A., Drozdzal, M., Courville, A., 2017. A benchmark for endoluminal scene segmentation of colonoscopy images. J. Healthc. Eng. 4037190.

Gao, X.W., Braden, B., 2020. Artefact detection and segmentation based on a deep learning system. In: Proceedings of the 2nd International Workshop and Challenge on Computer Vision in Endoscopy, EndoCV@ISBI 2020, Iowa City, Iowa, USA, 3rd April 2020. In: CEUR Workshop Proceedings, 2595, pp. 80–81.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580–587.

Gridach, M., Voiculescu, I., 2020. OXENDONET: A dilated convolutional neural networks for endoscopic artefact segmentation. In: Proceedings of the 2nd International Workshop and Challenge on Computer Vision in Endoscopy, EndoCV@ISBI 2020, Iowa City, Iowa, USA, 3rd April 2020. CEUR-WS.org, pp. 26–29.

Guo, X., Zhang, N., Guo, J., Zhang, H., Hao, Y., Hang, J., 2019. Automated polyp segmentation for colonoscopy images: a method based on convolutional neural networks and ensemble learning. Med. Phys. 46 (12), 5666–5676. doi:10.1002/mp.13865.

Guo, Y., Bernal, J., J Matuszewski, B., 2020. Polyp segmentation with fully convolutional deep neural networks-extended evaluation study. Journal of Imaging 6 (7), 69.

Guo, Y.B., Zheng, Q., Matuszewski, B.J., 2020. Deep encoder-decoder networks for artefacts segmentation in endoscopy images. In: Proceedings of the 2nd International Workshop and Challenge on Computer Vision in Endoscopy, EndoCV@ISBI 2020, Iowa City, Iowa, USA, 3rd April 2020. CEUR-WS.org, pp. 18–21.

Gupta, S., Ali, S., Goldsmith, L., Turney, B., Rittscher, J., 2020. Mi-unet: Improved segmentation in ureteroscopy. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 212–216.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp. 2961–2969.

Horie, Y., Yoshio, T., Aoyama, K., Yoshimizu, S., Horiuchi, Y., Ishiyama, A., Hirasawa, T., Tsuchida, T., Ozawa, T., Ishihara, S., et al., 2019. Diagnostic outcomes of esophageal cancer by artificial intelligence using convolutional neural networks. Gastrointest. Endosc. 89 (1), 25–32.

Hu, H., Guo, Y., 2020. Endoscopic artefact detection in mmdetection. In: Proceedings of the 2nd International Workshop and Challenge on Computer Vision in Endoscopy, EndoCV@ISBI 2020, Iowa City, Iowa, USA, 3rd April, pp. 78–79. Vol. 2595, CEUR Workshop Proceedings

Huynh, L.D., Boutry, N., 2020. A u-net++ with pre-trained efficientnet backbone for segmentation of diseases and artifacts in endoscopy images and videos. In: Proceedings of the 2nd International Workshop and Challenge on Computer Vision in Endoscopy, EndoCV@ISBI 2020, Iowa City, Iowa, USA, 3rd April 2020. CEUR-WS.org, pp. 13–17.

Incetan, K., Celik, I.O., Obeid, A., Gokceler, G.I., Ozyoruk, K.B., Almalioglu, Y., Chen, R.J., Mahmood, F., Gilbert, H., Durr, N.J., Turan, M., 2020. Vr-caps: A virtual environment for capsule endoscopy. 2008.12949.

Jadhav, S., Bamba, U., Chavan, A., Tiwari, R., Raj, A., 2020. Multi-plateau ensemble for endoscopic artefact segmentation and detection. In: Proceedings of the 2nd International Workshop and Challenge on Computer Vision in Endoscopy, EndoCV@ISBI 2020, Iowa City, Iowa, USA, 3rd April 2020. CEUR-WS.org, pp. 22–25.

Jha, D., Ali, S., Emanuelsen, K., Hicks, S., Thambawita, V., Garcia-Ceja, E., Riegler, M., de Lange, T., Schmidt, P.T., Johansen, H., Johansen, D., Halvorsen, P., 2020. Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy. OSF Preprints.

Jia, X., Mai, X., Cui, Y., Yuan, Y., Xing, X., Seo, H., Xing, L., Meng, M.Q., 2020. Automatic polyp recognition in colonoscopy images using deep learning and two-stage pyramidal feature prediction. IEEE Trans. Autom. Sci. Eng. 17 (3), 1570–1584.

Jorge, B., Aymeric, H., 2017. GastrointestinalG imageI analysisANA challenge. https://endovissub2017-giana.grand-challenge.org/.

Kaul, C., Manandhar, S., Pears, N., 2019. Focusnet: an attention-based fully convolutional network for medical image segmentation. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). IEEE, pp. 455–458.

Kayser, M., Soberanis-Mukul, R.D., Albarqouni, S., Navab, N., 2019. Focal loss for artefact detection in medical endoscopy. In: Proceedings of the 1st International Workshop and Challenge on Computer Vision in Endoscopy.

Khan, M.A., Choo, J., 2019. Multi-class artefact detection in video endoscopy via convolution neural networks. In: Proceedings of the 1st International Workshop and Challenge on Computer Vision in Endoscopy.

Krenzer, A., Hekalo, A., Puppe, F., 2020. Endoscopic detection and segmentation of gastroenterological diseases with deep convolutional neural networks. In: Proceedings of the 2nd International Workshop and Challenge on Computer Vision in Endoscopy, EndoCV@ISBI 2020, Iowa City, Iowa, USA, 3rd April 2020. CEUR-WS.org, pp. 58–63.

Law, H., Deng, J., 2018. Cornernet: Detecting objects as paired keypoints. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 734–750.

Lee, J., Park, S.W., Kim, Y.S., Lee, K.J., Sung, H., Song, P.H., Yoon, W.J., Moon, J.S., 2017. Risk factors of missed colorectal lesions after colonoscopy. Medicine 96 (27). e7468–e7468

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2117–2125.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp. 2980–2988.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context. In: European conference on computer vision. Springer, pp. 740–755.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C., 2016. Ssd: Single shot multibox detector. In: European conference on computer vision. Springer, pp. 21–37.

Liu, Y., Zhao, Z., Chang, F., Hu, S., 2020. An anchor-free convolutional neural network for real-time surgical tool detection in robot-assisted surgery. IEEE Access 8, 78193–78201.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431–3440.

Mahmood, F., Chen, R., Durr, N.J., 2018. Unsupervised reverse domain adaptation for synthetic medical images via adversarial training. IEEE Trans. Med. Imaging 37 (12), 2572–2581. doi:10.1109/TMI.2018.2842767.

Nguyen, N.-Q., Vo, D.M., Lee, S.-W., 2020. Contour-aware polyp segmentation in colonoscopy images using detailed upsamling encoder-decoder networks. IEEE Access 8, 99495–99508.

Nguyen, N.T., Tran, D.Q., Nguyen, D.B., 2020. Detection and segmentation of endoscopic artefacts and diseases using deep architectures. In: Proceedings of the 2nd International Workshop and Challenge on Computer Vision in Endoscopy, EndoCV@ISBI 2020, Iowa City, Iowa, USA, 3rd April 2020. CEUR-WS.org, pp. 64–67.

Norman, B., Pedoia, V., Majumdar, S., 2018. Use of 2d u-net convolutional neural networks for automated cartilage and meniscus segmentation of knee mr imaging data to determine relaxometry and morphometry. Radiology 288 (1), 177–185.

Oksuz, I., Clough, J.R., King, A.P., Schnabel, J.A., 2019. Artefact detection in video endoscopy using retinanet and focal loss function. In: Proceedings of the 1st International Workshop and Challenge on Computer Vision in Endoscopy, EndoCV@ISBI 2019.

Pengyi, Z., Xiaoqiong, L.Y.Z., 2019. Ensemble mask-aided r-cnn. In: Ali, S., Zhou, F. (Eds.), Proceedings of the 1st International Workshop and Challenge on Computer Vision in Endoscopy, EndoCV@ISBI 2019.

Polat, G., Sen, D., Inci, A., Temizel, A., 2020. Endoscopic artefact detection with ensemble of deep neural networks and false positive elimination. In: Proceedings of the 2nd International Workshop and Challenge on Computer Vision in Endoscopy, EndoCV@ISBI 2020, Iowa City, Iowa, USA, 3rd April 2020. CEUR-WS.org, pp. 8–12.

Rawla, P., Sunkara, T., Barsouk, A., 2019. Epidemiology of colorectal cancer: incidence, mortality, survival, and risk factors. Prz. Gastroenterol. 14 (2), 89–103.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779–788.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, pp. 91–99.

Rezvy, S., Zebin, T., Braden, B., Pang, W., Taylor, S., Gao, X.W., 2020. Transfer learning for endoscopy disease detection & segmentation with mask-rcnn benchmark architecture. In: Proceedings of the 2nd International Workshop and Challenge on Computer Vision in Endoscopy, EndoCV@ISBI 2020, Iowa City, Iowa, USA, 3rd April 2020. CEUR-WS.org, pp. 68–72.

Rolnick, D., Veit, A., Belongie, S.J., Shavit, N., 2017. Deep learning is robust to massive label noise. CoRR abs/1705.10694.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer, pp. 234–241.

Ross, T., Reinke, A., Full, P.M., Wagner, M., Kenngott, H., Apitz, M., Hempe, H., Mindroc Filimon, D., Scholz, P., Tran, T.N., Bruno, P., Arbelez, P., Bian, G.-B., Bodenstedt, S., Bolmgren, J.L., Bravo-Snchez, L., Chen, H.-B., Gonzlez, C., Guo, D., Halvorsen, P., Heng, P.-A., Hosgor, E., Hou, Z.-G., Isensee, F., Jha, D., Jiang, T., Jin, Y., Kirtac, K., Kletz, S., Leger, S., Li, Z., Maier-Hein, K.H., Ni, Z.-L., Riegler, M.A., Schoeffmann, K., Shi, R., Speidel, S., Stenzel, M., Twick, I., Wang, G., Wang, J., Wang, L., Wang, L., Zhang, Y., Zhou, Y.-J., Zhu, L., Wiesenfarth, M., Kopp-Schneider, A., Müller-Stich, B.P., Maier-Hein, L., 2020. Comparative validation of multi-instance instrument segmentation in endoscopy: results of the robust-mis 2019 challenge. Med. Image Anal. 101920. doi:10.1016/j.media.2020.101920.

Seferbekov, S.S., Iglovikov, V.I., Buslaev, A.V., Shvets, A.A., 2018. Feature pyramid network for multi-class land segmentation. CoRR abs/1806.03510.

Sevastopolsky, A., 2017. Optic disc and cup segmentation methods for glaucoma detection with modification of u-net convolutional neural network. Pattern Recognit. Image Anal. 27 (3), 618–624.

Shin, Y., Qadir, H.A., Aabakken, L., Bergsland, J., Balasingham, I., 2018. Automatic colon polyp detection using region based deep cnn and post learning approaches. IEEE Access 6, 40950–40962.

Silva, J., Histace, A., Romain, O., Dray, X., Granado, B., 2014. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. Int. Jour. of Comput. Assis. Radiol. and Surg. 9 (2), 283–293.

Soberanis-Mukul, R.D., Kayser, M., Zvereva, A., Klare, P., Navab, N., Albarqouni, S., 2020. A learning without forgetting approach to incorporate artifact knowledge in polyp localization tasks. CoRR abs/2002.02883.

Subramanian, A., Srivatsan, K., 2020. Exploring deep learning based approaches for endoscopic artefact detection and segmentation. In: Proceedings of the 2nd International Workshop and Challenge on Computer Vision in Endoscopy, EndoCV@ISBI 2020, Iowa City, Iowa, USA, 3rd April 2020. CEUR-WS.org, pp. 51–56.

Suhui Yang, G.C., 2019. Endoscopic artefact detection and segmentation with deep convolutional neural network. In: Ali, S., Zhou, F. (Eds.), Proceedings of the 1st International Workshop and Challenge on Computer Vision in Endoscopy, EndoCV@ISBI 2019.

Sun, C., Guo, S., Zhang, H., Li, J., Chen, M., Ma, S., Jin, L., Liu, X., Li, X., Qian, X., 2017. Automatic segmentation of liver tumors from multiphase contrast-enhanced ct images based on fcns. Artif. Intell. Med. 83, 58–66.

Urban, G., Tripathi, P., Alkayali, T., Mittal, M., Jalali, F., Karnes, W., Baldi, P., 2018. Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. Gastroenterology 155 (4), 1069–1078.

Vishnusai, V., Prakash, P., Shivashankar, N., 2020. A submission note on EAD 2020: Deep learning based approach for detecting artefacts in endoscopy. In: Proceedings of the 2nd International Workshop and Challenge on Computer Vision in Endoscopy, EndoCV@ISBI 2020, Iowa City, Iowa, USA, 3rd April 2020. CEUR-WS.org, pp. 30–36.

Wang, D., Zhang, N., Sun, X., Zhang, P., Zhang, C., Cao, Y., Liu, B., 2019. Afp-net: Realtime anchor-free polyp detection in colonoscopy. In: 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, pp. 636–643.

Wang, K.K., Tian, J.M., Gorospe, E., Penfield, J., Prasad, G., Goddard, T., Wongkeesong, M., Buttar, N., Lutzke, L., Krishnadath, S., 2012. Diseases of the esophagus : official journal of the international society for diseases of the esophagus. Dis Esophagus 25 (4), 349–355. doi:10.1111/j.1442-2050.2012.01342.x.

Wang, P., Xiao, X., Brown, J.R.G., Berzin, T.M., Tu, M., Xiong, F., Hu, X., Liu, P., Song, Y., Zhang, D., et al., 2018. Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. Nat. Biomed. Eng. 2 (10), 741–748.

Williams, J.G., Pullan, R.D., Hill, J., Horgan, P.G., Salmo, E., Buchanan, G.N., Rasheed, S., McGee, S.G., Haboubi, N., 2013. Management of the malignant colorectal polyp: acpgbi position statement. Colorectal Disease 15 (s2), 1–38. doi:10.1111/codi.12262.

Yamada, M., Saito, Y., Imaoka, H., Saiko, M., Yamada, S., Kondo, H., Takamaru, H., Sakamoto, T., Sese, J., Kuchiba, A., et al., 2019. Development of a real-time endoscopic image diagnosis support system using deep learning technology in colonoscopy. Sci. Rep. 9 (1), 1–9.

Yang, S., Cheng, G., 2019. Endoscopic artefact detection and segmentation with deep convolutional neural network. In: Proceedings of the 1st International Workshop and Challenge on Computer Vision in Endoscopy.

Yu, F., Koltun, V., 2015. Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122.

Yu, Z., Guo, Y., 2020. Endoscopic artefact detection using cascade R-CNN based model. In: Proceedings of the 2nd International Workshop and Challenge on Computer Vision in Endoscopy, EndoCV@ISBI 2020, Iowa City, Iowa, USA, 3rd April 2020. CEUR-WS.org, pp. 42–46.

Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y., 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 6023–6032.

Zhang, X., Chen, F., Yu, T., An, J., Huang, Z., Liu, J., Hu, W., Wang, L., Duan, H., Si, J., 2019. Real-time gastric polyp detection using convolutional neural networks. PLoS ONE 14 (3), 1–16. doi:10.1371/journal.pone.0214133.

Zhang, Y.-y., Xie, D., 2019. Detection and segmentation of multi-class artifacts in endoscopy. Journal of Zhejiang University-SCIENCE B 20 (12), 1014–1020.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2881–2890.

Zhou, X., Zhuo, J., Krahenbuhl, P., 2019. Bottom-up object detection by grouping extreme and center points. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 850–859.

International Journal of Colorectal Disease (2022) 37:1349–1354
https://doi.org/10.1007/s00384-022-04178-8

ORIGINAL ARTICLE

Pilot study of a new freely available computer-aided polyp detection system in clinical practice

Thomas J. Lux[1] · **Michael Banck**[1,2] · **Zita Saßmannshausen**[1] · **Joel Troya**[1] · **Adrian Krenzer**[1,2] · **Daniel Fitting**[1] · **Boban Sudarevic**[1,3] · **Wolfram G. Zoller**[3] · **Frank Puppe**[2] · **Alexander Meining**[1] · **Alexander Hann**[1]

Accepted: 1 May 2022 / Published online: 11 May 2022
© The Author(s) 2022

Abstract

Purpose Computer-aided polyp detection (CADe) systems for colonoscopy are already presented to increase adenoma detection rate (ADR) in randomized clinical trials. Those commercially available closed systems often do not allow for data collection and algorithm optimization, for example regarding the usage of different endoscopy processors. Here, we present the first clinical experiences of a, for research purposes publicly available, CADe system.
Methods We developed an end-to-end data acquisition and polyp detection system named EndoMind. Examiners of four centers utilizing four different endoscopy processors used EndoMind during their clinical routine. Detected polyps, ADR, time to first detection of a polyp (TFD), and system usability were evaluated (NCT05006092).
Results During 41 colonoscopies, EndoMind detected 29 of 29 adenomas in 66 of 66 polyps resulting in an ADR of 41.5%. Median TFD was 130 ms (95%-CI, 80–200 ms) while maintaining a median false positive rate of 2.2% (95%-CI, 1.7–2.8%). The four participating centers rated the system using the System Usability Scale with a median of 96.3 (95%-CI, 70–100).
Conclusion EndoMind's ability to acquire data, detect polyps in real-time, and high usability score indicate substantial practical value for research and clinical practice. Still, clinical benefit, measured by ADR, has to be determined in a prospective randomized controlled trial.

Keywords Colonoscopy · Polyp · Artificial intelligence · Deep learning · CADe

Introduction

Screening colonoscopies are highly effective at reducing the incidence of colorectal cancer (CRC). Previous studies revealed a decrease of 68% regarding CRC-related mortality by performing screening colonoscopies as most of these carcinomas develop over years following the adenoma-carcinoma sequence [1]. Adenoma detection rate (ADR) evolved to one of the most important colonoscopy quality parameters correlated to interval carcinoma rate [1]. As the research of artificial intelligence (AI) progressed, clinical applications were tested for viability [2]. A meta-analysis by Hassan et al. analyzed the current randomized studies regarding deep learning–based polyp detection in colonoscopy (CADe) [3]. They concluded that AI-assisted polyp detection increases the ADR, especially for small (< 5 mm), flat adenomas. Anyhow, only one of the five analyzed studies was performed in Europe [4] while the others are limited to an Asian study population [5–8]. Furthermore, three of the studies included mostly symptomatic patients [5–7]. Regarding generalizability, only one of the CADe systems [4] was evaluated with multiple processor types and only one study was multicentric [4]. Therefore, the authors concluded that more data for non-Asian populations is necessary. Furthermore, examiners focus on the center of the endoscopic image and CADe systems improve detection in the image's periphery [9]. Lastly, to our knowledge there is

---

Thomas J. Lux and Michael Banck contribute equally to this work.

✉ Alexander Hann
hann_a@ukw.de

[1] Interventional and Experimental Endoscopy (InExEn), Internal Medicine II, University Hospital Würzburg, Würzburg, Germany

[2] Artificial Intelligence and Knowledge Systems, Institute for Computer Science, Julius-Maximilians-Universität Würzburg, Würzburg, Germany

[3] Department of Internal Medicine and Gastroenterology, Katharinenhospital, Stuttgart, Germany

no data regarding usability and acceptance of CADe systems in clinical practice.

In this study, we present the pilot phase results of our real-time CADe polyp detection system EndoMind and its framework applied in clinical practice. The proposed framework is an end-to-end solution capable of data acquisition for the training of neural networks as well as clinical application of the AI. The AI was developed utilizing multicentric data acquired by the EndoMind framework itself using different endoscopy processor types. Therefore, it is capable of fast development, evaluation, and real-time application of AI-based video analysis. Lastly, we analyzed the physicians' feedback to evaluate the potential hardships of migrating this powerful tool for polyp detection to clinical application.

Methods

Development of EndoMind hardware and software

EndoMind hardware utilizes regular off-the-shelf components including a high-performance computer and a video grabber card that provide compatibility with a multitude of available endoscopy processors. The components were determined based on optimal requirements for a real-time AI application system while maintaining affordable pricing to make this freely available system easy to implement for clinicians in the future. Supplementary Table 1 lists the hardware composition resulting in a total price of about 2,880 €.

The CADe system, including software and hardware, was developed to perform data acquisition of the video signal and the exact location of the AI predictions as well as real-time polyp detection simultaneously. The software is able to handle a wide range of endoscopy processor video signals, including analog to ultra-high definition standards. The video signal is processed to single images called frames independently of the input source. Those are then forwarded to three processing pipelines (Display, AI, and Recording) in parallel to fit the requirements for real-time application (Supplementary Fig. 1). This parallelization minimizes video delay as only the predictions are visualized on a later frame. Furthermore, the AI predicts only every second to third frame and extrapolates the results to the remaining frames. The AI is based on a convolutional neural network that was trained with 506,338 manually annotated images from endoscopic examinations with and without visible polyps. The software's detailed structure is explained in Supplementary Material. EndoMind software including a detailed installation handbook is freely available for research purposes (https://www.ukw.de/research/inexen/ai-applied-in-real-time/).

Participants

We retrospectively reviewed colonoscopy reports and corresponding videos of our randomized controlled trial's pilot phase data. Here, examiners with at least 10 years of experience in performing colonoscopies were asked to evaluate EndoMind before starting the randomized study phase (NCT05006092). Only complete video recordings were included. The evaluated video recordings originate from four different endoscopy processors (Olympus CV-170 and CV-190 (Olympus Europa SE & Co. KG, Hamburg, Germany), Pentax i7000 (Pentax Europe GmbH, Hamburg, Germany), and Storz TC301 (Karl Storz SE & Co. KG, Tuttlingen, Germany)). Centers included three outpatient gastroenterological practices and one community-based hospital.

Data annotation

A physician (TJL) annotated each video from start to end and a board-certified gastroenterologist (AH) verified annotations. Sequences including polyps were labeled as such. Polyp size, morphology, pathological report if available, location and Boston bowel preparation scale (BBPS) were retrospectively identified. Polyps were categorized as proximal if located between caecum and the left flexure, otherwise as distal. Withdrawal time was determined as the time difference of the last anatomic landmark inspection (ileocecal valve, appendix, or ileum) and last image inside of the body [10]. Time spent on endoscopic interventions was manually annotated and subtracted from withdrawal time as well as all other evaluations. Each CADe prediction was labeled as true or false positive.

Survey

Examiners of the four centers were asked to participate in an online survey about the EndoMind usage (Supplementary Table 2). The survey consisted of the System Usability Scale (SUS) resulting in a total score of 0 to 100 points. Additional questions about the EndoMind performance were rated using a Likert scale from 1 (strongly disagree) to 5 (strongly agree) or percentage estimates.

Statistical analysis

Statistical analysis was performed using Python 3.10. Sensitivity was defined as the number of polyps detected in at least one frame divided by the number of all visible polyps. Time to first detection (TFD) was determined for each polyp as the visible time between polyp appearance and the first frame with correct CADe detection. For histology-based analyses, polyps without available histology due to not performed resection were excluded. Data was tested for normal distribution using SciPy's normal test. For data with normal distribution, mean and standard deviation were calculated. For non-normal distributed data, median and its two-sided 95% confidence intervals (CI) were calculated using bootstrapping ($n = 1000$).

Ethical considerations

The study was approved by the local ethical committee responsible for each study center (Ethik-Kommission Landesärztekammer Baden-Württemberg (F-2021–047), Ethik-Kommission Landesärztekammer Hessen (2021–2531), and Ethik-Kommission der Landesärztekammer Rheinland-Pfalz (2021–15,955)). All procedures were in accordance with the Helsinki Declaration of 1964 and later versions. Signed informed consent from each patient was obtained prior to participation.

Results

Patient characteristics

Using EndoMind (Fig. 1), 41 examinations were recorded during the pilot phase of the study in four centers. Patient characteristics are presented in Table 1. Most examinations were performed for colorectal cancer screening or surveillance (63.4%). BBPS was rated as six or higher in 95.1% of the examinations. Characteristics of the participating examiners are presented in Supplementary Table 3.

CADe performance

In total, 66 polyps were identified in 41 colonoscopies. Figure 2 depicts representative images of EndoMind detections. Polyp characteristics and detection metrics are summarized in Table 2. Of the 37 histologically evaluated polyps, 29 were diagnosed as adenomatous resulting in an ADR of 41.5%. EndoMind detected 29 of 29 adenomas and 66 of 66 polyps. Overall, median TFD was as fast as 130 ms (95%-CI, 80–200 ms).

Manual annotation of all 1,544,063 individual images of which 74,422 (4.82%) contained a visible polyp, revealed an overall CADe accuracy of 95.3%. Median false positive detection rate per examination was 2.2% (95%-CI, 1.7–2.8%).

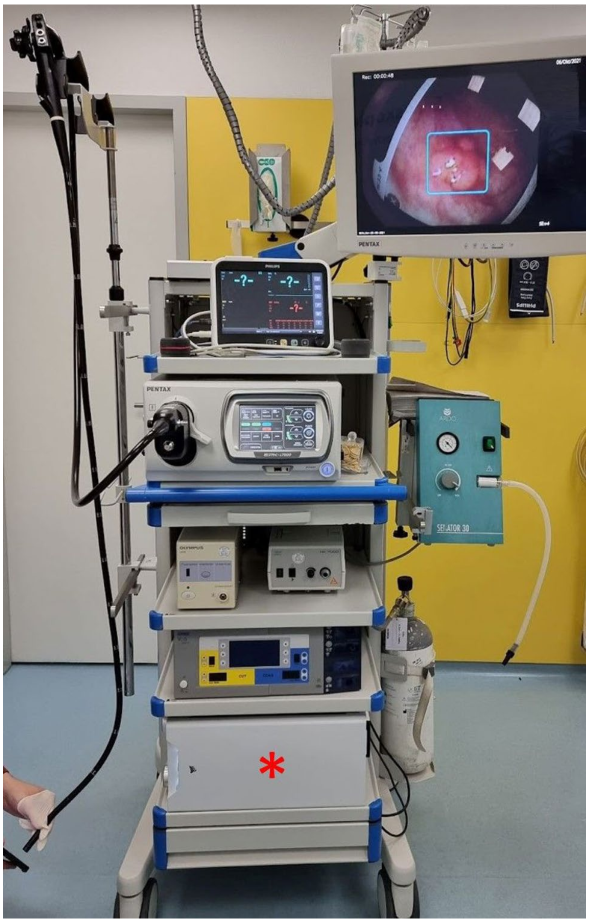


Fig. 1 EndoMind mounted on an endoscopic tower in one of the participating centers. Presentation of a polyp image on a small screen (lower left corner) and proper detection with a bounding box (upper right corner) by EndoMind (asterisk)

Table 1 Patient characteristics

Characteristic	Value
Age in years, median (95% CI)	62.0 (57.0–67.0)
Gender	
Male, *n* (%)	17 (41.5)
Female, *n* (%)	24 (58.5)
Indication	
Screening or surveillance, *n* (%)	26 (63.4)
Symptomatic, *n* (%)	15 (36.6)
BBPS, median (95% CI)	7.0 (7.0–8.0)
BBPS ≥ 6, *n* (%)	39 (95.1)
BBPS < 6, *n* (%)	2 (4.9)

CI confidence interval, *BBPS* Boston bowel preparation scale

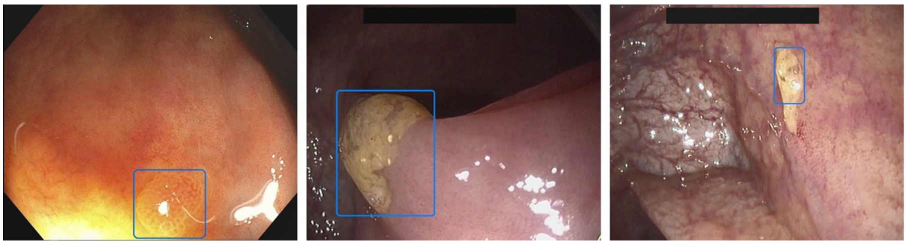


Fig. 2 Representative selection of EndoMind detections. EndoMind correctly marks a well visible (left) and a stool covered (middle) polyp with a blue bounding box. A common cause for false positive detections represented by stool on the bowel wall is displayed in the right image

Usability survey

Examiners participating in the pilot phase rated the usability of EndoMind with a median SUS score of 96.3 (95%-CI, 70–100). The physicians subjectively stated that 89% (95%-CI, 79–94%) of the polyps were detected by our system. Of those polyps 46% (95%-CI, 21–61%) were subjectively detected by EndoMind before the examiner. Anyhow, users partially criticized false detections as distracting (median 3, 95%-CI, 2–3) and as a possible reason for a prolonged withdrawal time (median 2.5, 95%-CI, 2.0–5.0). Lastly, interventionists agreed that the EndoMind system would benefit patient care (median 4.5, 95%-CI, 3.0–5.0) and therefore would like to use it in their clinical routine (median 4.5, 95%-CI, 4.0–5.0).

Table 2 Polyp characteristics and CADe performance

Category	*n* (%)	TFD in ms, median (95%-CI)
All polyps	66 (100)	130 (80–200)
Size		
< 5 mm	41 (62.1)	160 (80–260)
5–10 mm	19 (28.8)	120 (60–340)
> 10 mm	6 (9.1)	80 (40–4,380)
Histology (*n* = 37)		
Non-adenomatous	7	200 (60–2,280)
Tubular adenoma	24	160 (80–520)
Tubulovillous adenoma	3	180 (60–200)
Sessile serrated lesion	2	160 (100–220)
Carcinoma	1	40 (n.a.)
Location		
Proximal	30 (45.5)	160 (80–350)
Distal	36 (54.6)	120 (60–210)

TFD time to first polyp detection, *CI* confidence interval, *n.a.* not available

Discussion

In this work, we present the freely available CADe system EndoMind. It incorporates recording of endoscopy videos with AI predictions. Additionally, it is capable of real-time polyp detection on a variety of endoscopy processors. We could demonstrate successful installation and use of our system in four non-research-focused centers. While previous studies included mostly symptomatic patients of Asiatic origin in a hospital setting [5–7, 11], 63.4% of the colonoscopies included in our pilot phase study were performed as screening or surveillance examinations. Furthermore, we could preliminarily validate high sensitivity (100% of polyps detected) and fast detection (median TFD 130 ms). While this preliminary data may not be directly compared to other studies, the ADR in our pilot phase study was 41.5%. A total of 29 out of 37 (78.4%) histologically evaluated polyps were diagnosed as adenoma which indicates high quality of the performed colonoscopies. Assessing the characteristics of the detected adenomas, we found a similar size distribution compared to previously published studies [4–7]. Other CAD systems report a false positive (FP) rate of 0.9 to 8% [12–14]. Assessment of false detections by EndoMind is located in the lower range with 2.2%. Qualitative screening of coherent false positive detections revealed mainly stool-covered areas, air bubbles, or pseudo-polyps generated by artifacts due to suction of the mucosa as the most common sources. As especially right-sided polyps are initially often covered by mucus, some of those FP detections may not be eliminated without severely affecting detection of these polyps in the early phase when they appear. Nevertheless, as a recent in depth analysis by Spadaccini et al. demonstrated, examiners can quickly disregard these FPs [15].

Our usability-focused survey involved only highly experienced examiners, mostly from outpatient treatment centers. We designed EndoMind to assist in screening colonoscopies; therefore, this group resembles the future target group. The participating physicians found EndoMind to be easy to use and maintain with a median SUS of 96.3 which exceeds the average of 69 [16]. Furthermore, they agreed that their clinical routine would benefit from the regular usage of EndoMind. However, the examiners also stated that false positive detections might increase their withdrawal time. Additionally, even correctly detected polyps might disturb the workflow if the physician has already identified it. Therefore, features to easily and even automatically deactivate the system should be implemented in future. While manual deactivation may be achieved by a foot switch or voice command, automatic deactivation based on the examination state seems also promising. For this, the most practical approaches include activation of the CADe system only after identification of the caecum and deactivation if an instrument is detected in the field of view. This would restrict the CADe detections to the withdrawal time and prevent disturbing activations during resections and biopsies.

Additionally, we evaluated the physician's impressions of how many polyps were missed (11%), as well as how many polyps were detected by the system before the examiner (46%). The discrepancy between our determined sensitivity and the survey result may result from a different definition of detection: while frequently used metrics accept a polyp as detected if it is recognized at all, examiners might define a polyp, which is only detected after it is centered and focused on the image, as missed. As a more realistic measure, we therefore evaluated the TFD. Here, 89.4% of the polyps were detected in less than a second, which closely correlates with the examiners' impression of the percentage of CADe-identified polyps.

While our results imply high clinical value of our freely available CADe system, absence of a control group in this early stage as well as the small sample size demands verification by a larger, randomized, controlled study. The aim of this study was therefore not to present how our system improves the ADR, but instead to demonstrate the application of this new CADe system in a clinical scenario involving multiple processor types and an evaluation of its performance on a frame-by-frame basis.

As our system is easy to use, and preliminary results indicate high practical value, we are confident that patient care would profit if systems like EndoMind are utilized in the daily routine. Furthermore, the implemented recording capabilities reduce the effort for continuously improving the system. By usage of rapid training iterations, our system enables for user- or patient group–specific AI fine-tuning as it is known from other applications like text to speech applications which improve their performance with increasing time of use. We hope that the EndoMind platform might contribute to improving endoscopy by continuously incorporating new AI features.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1007/s00384-022-04178-8.

Acknowledgements The authors acknowledge the support by Prof. J.F. Riemann, "Stiftung Lebensblicke" the Foundation for early detection of colon cancer.

Author contribution AH and TJL: study concept and design. AH, TJL, and ZS: statistical analysis. AK and MB: development of the software and neural network. AH, TJL, and MB: interpretation of results, and drafting of the manuscript. AH, TJL, ZS, JT, AK, DF, BS, WGZ, FP, and AM: acquisition of data and providing study material. All authors: critical revision of the article for important intellectual content and final approval of the article.

Funding Open Access funding enabled and organized by Projekt DEAL. The authors AH and WGZ receive public funding from the state government of Baden-Württemberg, Germany (Funding cluster "Forum Gesundheitsstandort Baden-Württemberg") to research and develop artificial intelligence applications for polyp detection in screening colonoscopy (funding number 5409.0–001.01/15). AM receives funding from the IZKF Würzburg (funding number F-406) and the Bavarian Center for Cancer Research (BZKF) for further implementation and development of artificial intelligence for detection of (pre-) neoplastic lesions.

Declarations

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Corley DA, Levin TR, Doubeni CA (2014) Adenoma detection rate and risk of colorectal cancer and death. N Engl J Med 370:2541. https://doi.org/10.1056/NEJMc1405329
2. Liu P, Wang P, Glissen Brown JR et al (2020) The single-monitor trial: an embedded CADe system increased adenoma detection during colonoscopy: a prospective randomized study. Ther Adv Gastroenterol 13:1756284820979165. https://doi.org/10.1177/1756284820979165
3. Hassan C, Spadaccini M, Iannone A et al (2021) Performance of artificial intelligence in colonoscopy for adenoma and polyp detection: a systematic review and meta-analysis. Gastrointest Endosc 93:77-85.e6. https://doi.org/10.1016/j.gie.2020.06.059
4. Repici A, Badalamenti M, Maselli R et al (2020) Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial. Gastroenterology 159:512-520.e7. https://doi.org/10.1053/j.gastro.2020.04.062
5. Wang P, Liu X, Berzin TM et al (2020) Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): a double-blind randomised study.

Lancet Gastroenterol Hepatol 5:343–351. https://doi.org/10.1016/S2468-1253(19)30411-X

6. Wang P, Berzin TM, Glissen Brown JR et al (2019) Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. Gut 68:1813–1819. https://doi.org/10.1136/gutjnl-2018-317500
7. Liu W-N, Zhang Y-Y, Bian X-Q et al (2020) Study on detection rate of polyps and adenomas in artificial-intelligence-aided colonoscopy. Saudi J Gastroenterol Off J Saudi Gastroenterol Assoc 26:13–19. https://doi.org/10.4103/sjg.SJG_377_19
8. Su J-R, Li Z, Shao X-J et al (2020) Impact of a real-time automatic quality control system on colorectal polyp and adenoma detection: a prospective randomized controlled study (with videos). Gastrointest Endosc 91:415-424.e4. https://doi.org/10.1016/j.gie.2019.08.026
9. Troya J, Fitting D, Brand M et al (2022) The influence of computer-aided polyp detection systems on reaction time for polyp detection and eye gaze. Endoscopy. https://doi.org/10.1055/a-1770-7353
10. Kaminski MF, Thomas-Gibson S, Bugajski M et al (2017) Performance measures for lower gastrointestinal endoscopy: a European Society of Gastrointestinal Endoscopy (ESGE) Quality Improvement Initiative. Endoscopy 49:378–397. https://doi.org/10.1055/s-0043-103411
11. Gong D, Wu L, Zhang J et al (2020) Detection of colorectal adenomas with a real-time computer-aided system (ENDOANGEL): a randomised controlled study. Lancet Gastroenterol Hepatol 5:352–361. https://doi.org/10.1016/S2468-1253(19)30413-3
12. Urban G, Tripathi P, Alkayali T et al (2018) Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. Gastroenterology 155:1069-1078.e8. https://doi.org/10.1053/j.gastro.2018.06.037
13. Hassan C, Wallace MB, Sharma P et al (2020) New artificial intelligence system: first validation study versus experienced endoscopists for colorectal polyp detection. Gut 69:799–800. https://doi.org/10.1136/gutjnl-2019-319914
14. Pfeifer L, Neufert C, Leppkes M et al (2021) Computer-aided detection of colorectal polyps using a newly generated deep convolutional neural network: from development to first clinical experience. Eur J Gastroenterol Hepatol 33:e662–e669. https://doi.org/10.1097/MEG.0000000000002209
15. Spadaccini M, Hassan C, Alfarone L et al (2022) Comparing number and relevance of false activations between two artificial intelligence CADe SystEms: the NOISE study. Gastrointest Endosc S0016–5107(21):01945–01953. https://doi.org/10.1016/j.gie.2021.12.031
16. Bangor A, Kortum P, Miller J (2009) Determining what individual SUS scores mean: Adding an adjective rating scale. J Usability Stud 4:114–123

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

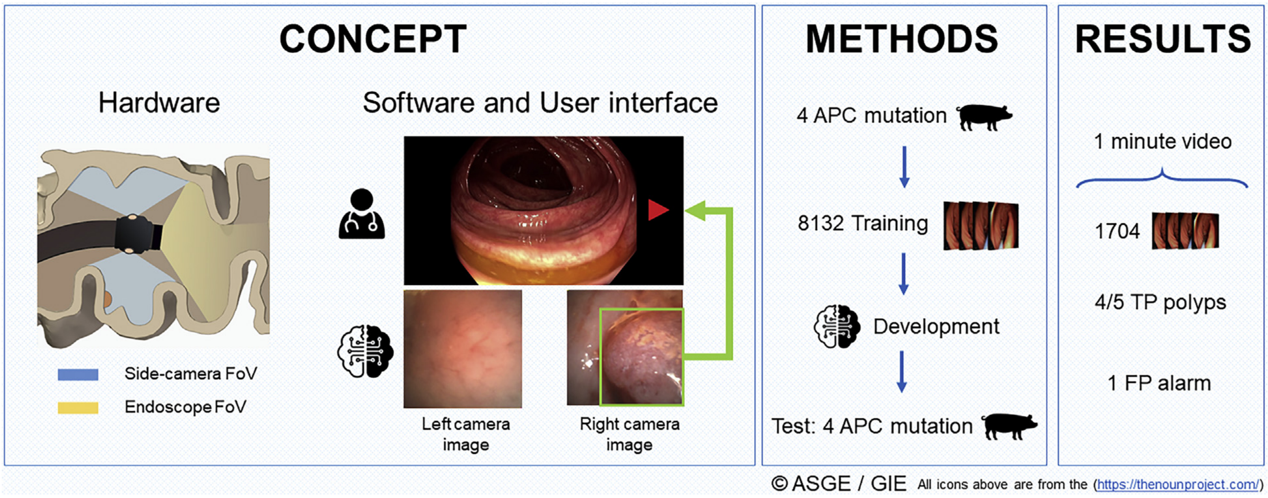
NEW METHODS

New concept for colonoscopy including side optics and artificial intelligence

Joel Troya, MSc,[1] Adrian Krenzer, MSc,[1,2] Krzysztof Flisikowski, Dr habil,[3] Boban Sudarevic, Dipl Ing,[1] Michael Banck, MSc,[1,2] Alexander Hann, MD,[1] Frank Puppe, Prof Dr,[2] Alexander Meining, Prof MD[1]

Würzburg, München, Germany

GRAPHICAL ABSTRACT



© ASGE / GIE All icons above are from the (https://thenounproject.com/)

Background and Aims: Adenoma detection rate is the crucial parameter for colorectal cancer screening. Increasing the field of view with additional side optics has been reported to detect flat adenomas hidden behind folds. Furthermore, artificial intelligence (AI) has also recently been introduced to detect more adenomas. We therefore aimed to combine both technologies in a new prototypic colonoscopy concept.

Methods: A 3-dimensional–printed cap including 2 microcameras was attached to a conventional endoscope. The prototype was applied in 8 gene-targeted pigs with mutations in the adenomatous polyposis coli gene. The first 4 animals were used to train an AI system based on the images generated by microcameras. Thereafter, the conceptual prototype for detecting adenomas was tested in a further series of 4 pigs.

Results: Using our prototype, we detected, with side optics, adenomas that might have been missed conventionally. Furthermore, the newly developed AI could detect, mark, and present adenomas visualized with side optics outside of the conventional field of view.

Conclusions: Combining AI with side optics might help detect adenomas that otherwise might have been missed.

(footnotes appear on last page of article)

Most colonoscopies are performed for the detection and removal of early neoplasms or adenomas. Several studies have demonstrated that the adenoma detection rate is the critical parameter for reducing the incidence of colorectal cancer.[1,2]

So far, several approaches have been mentioned to increase the adenoma detection rate by increasing the endoscopic viewing angle, thereby covering a larger area of the colon's surface.[3,4] However, these approaches are problematic in that the examiner needs to deal with more image information.[5] In addition, images may be distorted.[6] Accordingly, these procedures have not yet been implemented in practice.

Another completely new approach to polyp detection is digital image analysis using artificial intelligence (AI). However, the image presented to any AI corresponds precisely to the image of the examiner and therefore is more of a confirmation of what is visible. Lesions out of the standard field of view have not yet been recognized. Both approaches separately (full-spectrum endoscopy and computer-aided detection devices) have been proven to increase the adenoma detection rate.[4,7] Therefore, it is interesting to try to merge both concepts in a single innovation.

In contrast to endoscopy, solutions to these shortcomings of image information and current AI systems have been offered in the automotive sector. Here, a yellow or red symbol is projected into the side mirror for the driver while driving to draw attention to vehicles passing ("blind spot assistant" or "lane change assistant"). Theoretically, this assistance function can be transferred entirely to colonoscopy in that the examiner focuses on the conventional image "forward" as always, whereas an AI analyzes other image sources "looking back." If the AI detects an adenoma outside of the standard forward field of view, a corresponding warning signal may be presented to the endoscopist.

We reported on the potential benefit of a 3-dimensional–printed, side optic–enhanced cap including 2 microcameras as a feasible add-on to improve adenoma detection rates. More flat lesions were detected using an ex vivo colonoscopy simulator, especially in problematic areas of the colon, such as the flexures.[8] Here we present the further development and early preclinical evaluation of this device, including AI as an assistant system for detecting adenomatous polyps outside the conventional endoscopic image.

METHODS

Pig model

Gene-targeted pigs with the truncating 1311 mutation in the adenomatous polyposis coli (*APC*) gene were endoscopically examined with our system.[9] The *APC*^{1311} mutation is orthologous to the hotspot *APC*^{1309} mutation responsible for human familial adenomatous polyposis and causing aberrant crypt foci and low- and high-grade dysplastic adenomas in the large intestine, similar to the precancerous lesions that patients with familial adenomatous polyposis develop. The *APC*$^{1311/+}$ pigs are a suitable model for experimental endoscopy, as shown in previous studies.[10,11] All animal experiments were approved by the Government of Upper Bavaria (permit number ROB-55.2-2532.Vet_02-18-33) and performed according to the German Animal Welfare Act and European Union Normative for Care and Use of Experimental Animals.

Cap and micro-optics

Two red-green-blue microcamera modules (OsirisM; Optasensor GmbH, Nürnberg, Germany) were integrated into a 3-dimensional–printed cap. Each microcamera had 4 mini light-emitting diodes (OSRAM Opto Semiconductors GmbH, München, Germany) arranged around it. The total dimensions were 3.8 × 3.8 × 2 mm^3 (height × width × depth). The resolution of the obtained image signal was 320 × 320 pixels, the best focus point was 15 mm, and the field of view was 90 degrees.

The 3-dimensional–printed cap was designed to be fixed 5 mm from the tip of a gastroscope (GIF-H180; Olympus, Tokyo, Japan). The material used was nylon PA12, and a selective laser sintering printer (Lisa Pro, Sinterit sp. Z o.o., Kraków, Poland) was used to produce it. The cap contained 2 openings to allow the microcameras to be integrated into the normal endoscope's axis. Therefore, their orientation allowed a new view not achieved with the regular endoscope camera (Fig. 1A). The cap also incorporated cut-out areas to allow a complete field of view. Microcameras and mini light-emitting diodes were secured on the cap with silicone epoxy. A tube-like plastic foil was used to protect the 2-m length cable of the cameras (Fig. 1B).

Figure 2B shows the setup used to perform colonoscopies. The pigs' colons were flushed by rectal water irrigation before the endoscope and side optics were introduced.

User interface

A custom-designed computer grabbed 3 different image signals: the image source from the endoscope and the 2 image sources from the microcameras. The developed deep learning network analyzed these 2 last image sources, and each output was displayed on a second screen. Whenever the algorithm detected a presumptuous adenoma, a green square was drawn over the correspondent image, highlighting the region of interest. At the same time, an arrow was drawn in the conventional endoscope image and displayed on the main screen. The location of the arrow could be right or left, depending on which camera triggered the detection (Fig. 3). The endoscopist remained focused on the conventional endoscopic image until an arrow caught his attention. Then, the cause of the detection was visualized on the second monitor containing the microcameras images.

To avoid a significant number of false detections, an averaging filter was implemented. It analyzed several images together to predict an outcome, making the system more stable.

The 3 image signals were recorded and stored. Moreover, a record file containing the detections of each microcamera was created. The record file included the timestamp and location on the screen of the detection

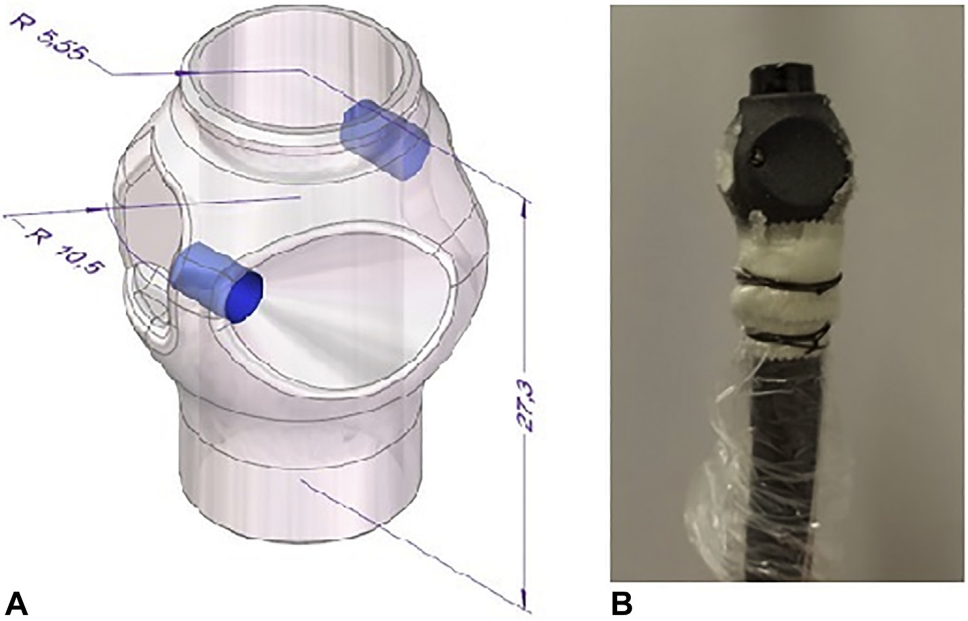


Figure 1. Three-dimensional–printed cap to place the microcameras. **A,** Three-dimensional design with main dimensions. **B,** Image of the final assembly in a gastroscope.

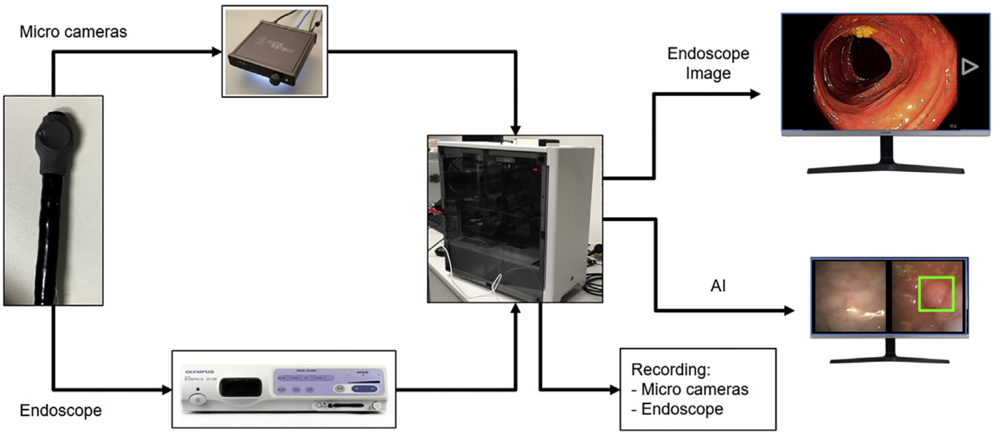


Figure 2. Diagram of the setup. *AI*, Artificial intelligence.

so they could always be correlated with the primary endoscope image source for postprocessing or analysis.

Artificial intelligence

For the detection of the polyps, deep learning was used. Deep learning attempts to imitate the workings of the human brain in data processing and pattern detection. Therefore, deep learning refers to a machine learning method that uses artificial neural networks with numerous intermediate layers between the input and output layers, thereby forming an extensive internal structure. For the detection of polyps, pattern recognition of the neural network is a critical factor. Therefore, techniques of general object detection are used. In particular, convolutional neural networks are the best in detecting patterns and objects in images.[12,13] To implement AI in our case, the output from each of the microcameras was preprocessed (cropping, color transformation, image resize, and normalization) and then analyzed by the AI.

The architecture of the AI used was based on YOLOv5,[14] a neural network design that is state of the art and has its primary focus on being fast and accurate. Therefore, the YOLOv5 architecture allows real-time processing of more than 30 frames per second while keeping a high detection rate. Because we did not have many images of animal colon polyps to train the AI, several techniques had to be used to prevent overfitting. First, we trained with different image augmentations to enhance and generalize the training data. In deep learning, augmenting image data means using various processes to modify the original

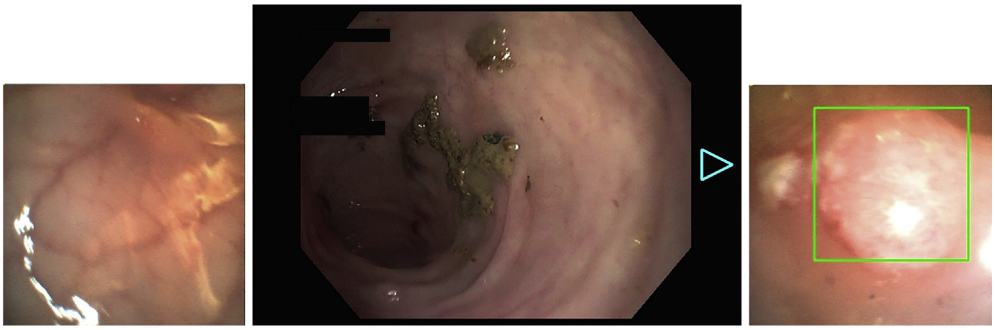


Figure 3. Image of a correct detection. **Left,** Left-sided camera showing normal mucosa. **Center,** Standard endoscopic view image. The *arrowhead* alerts the endoscopist that the right camera has detected an adenoma. **Right,** Right-sided microcamera highlighting the region where the adenoma appears.

image data. For data augmentation, we used techniques including flipping, rotating, translation, scaling, and mosaic augmentation. Second, we used early stopping during training. Early stopping is a regularization technique in iterative machine learning methods where the training is stopped as soon as a significant deterioration of the generalization performance is detected. The training algorithm then returns the model parameters with the best generalization performance up to that point. Third, we used a dropout value of 30%. When the network is trained, 30% of neurons in each layer of the network are turned off ("dropout") and not considered for the upcoming computational step. During training, units and their input and output connections are randomly removed from the network. Therefore, the units should be individually different from each other so their characteristics are considered for the prediction.

RESULTS

In a first animal trial including 4 pigs, 60,329 frames were obtained with the microcameras during colonoscopy. Among these frames, 8132 were annotated with a bounding box whenever a polyp appeared in the image. Because the number of pigs to train the AI was low, the risk of overfitting was high. Therefore, these annotations were used to fine-tune a pretrained model. The basic architecture of the pretrained AI model consisted of 223,566 images from human colonoscopy: 122,323 images contained polyps and 101,243 normal mucosa.

During the second animal trial, another 4 gene-targeted pigs with mutations in the *APC* gene were endoscopically examined. Polyps were aimed to be detected during retrieval of the endoscope. The shaft of the endoscope was torqued with small movements in a clockwise and counterclockwise direction to fully visualize the mucosa with side optics. This generated 32,831 frames from each image source.

Eighteen minutes of examinations were recorded. A significant number of frames were not useful because of direct contact of the microcamera with the mucosa (27.06%) or stool remnants (34.40%). In a minute of a video containing 1704 useful frames, the system recognized 4 of 5 polyps and raised 1 false-positive alarm.

Usually, stool remnants trigger a noteworthy number of false-positive detections. However, in our case, they accounted for less than 1% of the examination.

DISCUSSION

We present early preclinical data to further test a new prototypic concept for polyp detection in the colon of gene-targeted pigs. To the best of our knowledge, for the first time we were able to combine the concept of side-viewing or wide-angle endoscopy with AI. Thereby, a new system was developed with assistant functions outside the conventional field of view. Hence, the endoscopist may focus and concentrate on the standard endoscopic image but receive an optical signal as soon as the AI detects a polyp outside the standard field of view. Thereby, AI helps to omit previous shortcomings of wide-angle endoscopes[5] (ie, to focus on 3 monitors simultaneously).

In addition, our approach is simple, and although we do not have preliminary economic data, we assume that costs could be reduced compared with existing wide-angle colonoscopes because standard endoscopes can be used. Although the microcameras used have a limited field of view and lower resolution compared with a standard endoscopic image, if the AI is adequately trained with those images, detection of polyps is possible as has been shown.

Potential shortcomings of our study could be the pig model. *APC* pigs did not receive a standard lavage for bowel preparation. Although manual cleansing was performed, stool remnants embedded in the microcameras was a major factor in the failure to detect more polyps. In addition, rotating movements were necessary to visualize a greater surface of the colon mucosa because of the limited field of view of the microcameras. Finally, only a few animals were examined.

Nevertheless, the prototypic concepts have been proven valuable with potential perspectives to be integrated into clinical practice once it has been approved as a medical device. The next steps will be to further integrate the microcameras, better train the AI for the detection of polyps using side-viewing optics, and, ultimately, to potentially integrate other imaging modalities, such as optical-coherence tomography or near-infrared imaging,[10,15] for AI-guided scanning of the colon mucosa outside the conventional endoscopic field of view. Future studies must be performed to extensively evaluate the performance of the device and testing if it can lead to the detection of additional polyps. Overall, we believe our new concept for colonoscopy combining AI with side optics might help detect adenomas that otherwise could be missed without significantly disturbing the conventional colonoscopic workflow.

REFERENCES

1. Su JR, Li Z, Shao XJ, et al. Impact of a real-time automatic quality control system on colorectal polyp and adenoma detection: a prospective randomized controlled study (with videos). Gastrointest Endosc 2020;91:415-24.
2. Corley DA, Jensen CD, Marks AR, et al. Adenoma detection rate and risk of colorectal cancer and death. N Engl J Med 2014;370:1298-306.
3. Triadafilopoulos G, Li J. A pilot study to assess the safety and efficacy of the Third Eye retrograde auxiliary imaging system during colonoscopy. Endoscopy 2008;6:478-82.
4. Gralnek IM, Siersema PD, Halpern Z, et al. Standard forward-viewing colonoscopy versus full-spectrum endoscopy: an international, multicentre, randomised, tandem colonoscopy trial. Lancet Oncol 2014;3:353-60.
5. Hassan C, Senore C, Radaelli F, et al. Full-spectrum (FUSE) versus standard forward viewing. Gut 2017;66:1949-55.
6. Uraoka T, Tanaka S, Oka S, et al. Feasibility of a novel colonoscope with extra-wide angle of view: a clinical study. Endoscopy 2015;47:444-8.
7. Hassan C, Spadaccini M, Iannone A, et al. Performance of artificial intelligence in colonoscopy for adenoma and polyp detection: a systematic review and meta-analysis. Gastrointest Endosc 2021;93:77-85.
8. Walter BM, Hann A, Frank R, et al. A 3D-printed cap with side optics for colonoscopy: a randomized ex vivo study. Endoscopy 2017;49:808-12.
9. Flisikowska T, Merkl C, Martina L, et al. A porcine model of familial adenomatous polyposis. Gastroenterology 2012;5:1173-5.
10. Yim JJY, Harmsen S, Flisikowski K, et al. A protease-activated, near-infrared fluorescent probe for early endoscopic detection of premalignant gastrointestinal lesions. Proc Natl Acad Sci USA 2021;1:118.
11. Rogalla S, Flisikowski K, Gorpas D, et al. Biodegradable fluorescent nanoparticles for endoscopic detection of colorectal carcinogenesis. Adv Funct Mater 2019;29:1904992.
12. Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 2017;6:1137-49.
13. Zhao Z-Q, Zheng P, Xu S-T, et al. Object detection with deep learning: a review. IEEE Trans Neural Netw Learn Syst 2019;11:3212-32.
14. Redmon J, Farhadi A. YOLOv3: an incremental improvement. Available at: https://arxiv.org/abs/1804.02767. Accessed June 7, 2021
15. Zulina N, Caravaca O, Liao G, et al. Colon phantoms with cancer lesions for endoscopic characterization with optical coherence tomography. Biomed Opt Expr 2021;2:955-68.

Abbreviations: AI, artificial intelligence; APC, adenomatous polyposis coli.

DISCLOSURE: *The following authors disclosed financial relationships: A. Meining: Consultant for Ovesco GmbH. All other authors disclosed no financial relationships. Research support for this study was provided by IZKF Würzburg, the foundation “Forschung hilft,” and the Bavarian Center for Cancer Research.*

Copyright © 2021 by the American Society for Gastrointestinal Endoscopy
0016-5107/$36.00
https://doi.org/10.1016/j.gie.2021.12.003

Received July 27, 2021. Accepted December 11, 2021.

Current affiliations: Interventional and Experimental Endoscopy (InExEn), Internal Medicine II, University Hospital Würzburg, Würzburg, Germany (1), Artificial Intelligence and Knowledge Systems, Institute for Computer Science, Julius-Maximilians-Universität Würzburg, Würzburg, Germany (2), Lehrstuhl für Biotechnologie der Nutztiere, School of Life Sciences, Technische Universität München, München, Germany (3).

Reprint requests: Joel Troya, MSc, Universitätsklinikum Würzburg, Medizinische Klinik und Poliklinik II, Grombühlstraße 12, 97080, Würzburg, Germany.

If you would like to chat with an author of this article, you may contact Mr Troya at TroyaSebas_J@ukw.de.

SCANDINAVIAN JOURNAL OF GASTROENTEROLOGY
https://doi.org/10.1080/00365521.2022.2085059

ORIGINAL ARTICLE

OPEN ACCESS

A video based benchmark data set (ENDOTEST) to evaluate computer-aided polyp detection systems

Daniel Fitting[a], Adrian Krenzer[a,b], Joel Troya[a], Michael Banck[a,b], Boban Sudarevic[a,c], Markus Brand[a], Wolfgang Böck[d], Wolfram G. Zoller[c], Thomas Rösch[e], Frank Puppe[b], Alexander Meining[a] and Alexander Hann[a]

[a]Interventional and Experimental Endoscopy (InExEn), Internal Medicine II, University Hospital Wuerzburg, Würzburg, Germany; [b]Artificial Intelligence and Knowledge Systems, Institute for Computer Science, Julius-Maximilians-Universität, Würzburg, Germany; [c]Department of Internal Medicine and Gastroenterology, Katharinenhospital, Stuttgart, Germany; [d]Practice for gastroenterology, Ulm, Germany; [e]Department of Interdisciplinary Endoscopy, University Hospital Hamburg-Eppendorf, Hamburg, Germany

ABSTRACT

Background and aims: Computer-aided polyp detection (CADe) may become a standard for polyp detection during colonoscopy. Several systems are already commercially available. We report on a video-based benchmark technique for the first preclinical assessment of such systems before comparative randomized trials are to be undertaken. Additionally, we compare a commercially available CADe system with our newly developed one.

Methods: ENDOTEST consisted in the combination of two datasets. The validation dataset contained 48 video-snippets with 22,856 manually annotated images of which 53.2% contained polyps. The performance dataset contained 10 full-length screening colonoscopies with 230,898 manually annotated images of which 15.8% contained a polyp. Assessment parameters were accuracy for polyp detection and time delay to first polyp detection after polyp appearance (FDT). Two CADe systems were assessed: a commercial CADe system (GI-Genius, Medtronic), and a self-developed new system (ENDOMIND). The latter being a convolutional neuronal network trained on 194,983 manually labeled images extracted from colonoscopy videos recorded in mainly six different gastroenterologic practices.

Results: On the ENDOTEST, both CADe systems detected all polyps in at least one image. The per-frame sensitivity and specificity in full colonoscopies was 48.1% and 93.7%, respectively for GI-Genius; and 54% and 92.7%, respectively for ENDOMIND. Median FDT of ENDOMIND with 217 ms (Inter-Quartile Range(IQR)8–1533) was significantly faster than GI-Genius with 1050 ms (IQR 358–2767, $p = 0.003$).

Conclusions: Our benchmark ENDOTEST may be helpful for preclinical testing of new CADe devices. There seems to be a correlation between a shorter FDT with a higher sensitivity and a lower specificity for polyp detection.

ARTICLE HISTORY
Received 5 April 2022
Revised 22 May 2022
Accepted 27 May 2022

KEYWORDS
Colonoscopy; polyp; artificial intelligence; deep learning; CADe

Introduction

Improvement of surveillance colonoscopy for the prevention of colorectal cancer (CRC) has always been a field of intensive research in gastrointestinal endoscopy. By detection and subsequent resection of adenomatous polyps, patients are preserved from cancer development. Thus, the adenoma detection rate (ADR) was established as a validated marker of colonoscopy quality [1]. An increase in ADR results in a decrease in interval carcinoma [2]. Still, the miss rate of neoplastic lesions is unacceptably high with great variability among individual endoscopists [3, 4]. The implementation of artificial intelligence systems for CADe resulted in increased ADRs in multiple prospective mostly single-center randomized trials [5]. Although surveillance colonoscopy is usually an outpatient procedure, training data of those systems mainly rely on colonoscopy videos recorded in in-hospital settings [6–8]. Several commercial CADe systems have already entered the market [9–11]. GI Genius (Medtronic plc., Dublin, Ireland) was one of the first commercial CADe systems in Europe. The multicenter randomized study by Repici *et. al* reported an increase in ADR of 14.4 percentage points regarding colonoscopies performed without the CADe system [10]. However, direct comparison of different CADe systems using the same benchmark data has to our knowledge never been done. Therefore, in this study we have generated ENDOTEST, a dataset that allows the comparison of different polyp detection systems. ENDOTEST includes polyp and non-polyp video sequences, in the same way as has previously

CONTACT Alexander Hann hann_a@ukw.de Universitätsklinikum Würzburg, Medizinische Klinik und Poliklinik II, Oberdürrbacher Str. 6, Würzburg, 97080, Deutschland.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/00365521.2022.2085059

© 2022 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

been done in other studies [12–14], but it also includes full-length frame-by-frame polyp annotated colonoscopies.

There is a wide variety of databases for colonoscopy fully annotated regarding the presence of polyps. CVC-VideoClinicDB was provided in the context of the GIANA sub-challenge that was part of the MICCAI 2017 Endoscopic Vision Challenge. This data set contains 18,733 frames from 18 videos without ground truth and 11,954 frames with ground truth [15]. SUN Colonoscopy Video Database was developed by Mori Laboratory and it contains 49,135 fully annotated polyp frames from 100 different polyps. It also contains 109,554 non-polyp frames [13]. The biggest and more diverse one is LDPolypVideo dataset which contains 160 colonoscopy video sequences and 40,266 frames with polyp annotations [14]. However, to our knowledge, there is no publicly available dataset containing full-length colonoscopies frame-by-frame annotated regarding the presence of a polyp.

In addition, in this work we also introduce ENDOMIND, a publicly funded investigator-initiated project of artificial intelligence applications for polyp detection in screening colonoscopy. ENDOMIND was developed by computer engineers as well as endoscopists in the same work group.

The aim of this study is to describe the validation, and performance comparison between the newly developed CADe system ENDOMIND trained with multicentric outpatient colonoscopy videos and the commercially available CADe system GI Genius using a defined benchmark data set ENDOTEST that contains screening colposcopies.

Methods

Training data set of ENDOMIND

Videos from routine colonoscopies were recorded in six gastroenterologic practices in Germany. The endoscopic processors were Olympus CV-170 and CV-190 (Olympus Europa SE & Co. KG, Hamburg, Germany), Pentax i7000 (Pentax Europe GmbH, Hamburg, Germany), and Storz TC301 (Karl Storz SE & Co. KG, Tuttlingen, Germany). The recording included retrospectively and prospectively collected endoscopic videos ranging from January 2018 to May 2021. Over 500 colonoscopy videos were screened for polyps. Subsequently, 219 video sequences comprising of 500 to 2000 images displaying one polyp were extracted. Additional sequences of the same video with the same length showing normal mucosa, residual stool, bubbles, or water irrigation images were labeled accordingly and added to the training data.

Polyps were manually classified according to the estimated size (<5 mm, 5–10 mm, >10 mm) and Paris classification [16]. A representative subset of polyps was chosen for box annotation, which included a frame-by-frame annotation of the visible polyp. This was performed by an experienced endoscopist using a custom-made annotation tool as previously described [17]. Annotation of the 219 sequences with and without a visible polyp resulted in 194,983 labeled images out of which 52.4% contained a polyp. Those images were used as training set for the development of ENDOMIND (Figure 1).

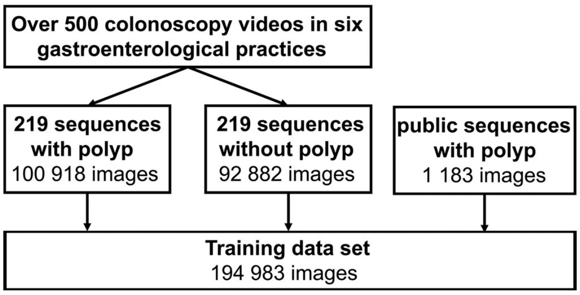

Over 500 colonoscopy videos in six gastroenterological practices
219 sequences with polyp
100 918 images
219 sequences without polyp
92 882 images
public sequences with polyp
1 183 images
Training data set
194 983 images

Figure 1. Composition of the training data set for the development of ENDOMIND.

ENDOMIND training

For polyp detection an artificial intelligence (AI) was trained using an on 1,183 publicly available polyp images [18–20] pre-trained deep learning algorithm (Supplementary material). It refers to a machine learning method that uses artificial neural networks with numerous intermediate layers, forming an extensive internal structure between the input and output layers. We used the YOLOv5 architecture [21]. Different techniques were used to further enhance YOLOv5. Firstly, we used early stopping during training. Early stopping is a form of regularization to prevent overfitting in iterative machine learning methods. Overfitting increases specificity on the training dataset but does not allow generalization into the real-world scenario. Secondly, we use a dropout value of 30%. When the network is trained, 30% of neurons in each layer of the network are turned off ("dropout") and not considered for the upcoming computational step. This technique also prevents overfitting of the model. Thirdly, we trained with different image augmentations. In deep learning, augmenting image data means using various processes to modify the original image data. For data augmentation, we used techniques including flipping, rotating, translation, scaling, and mosaic augmentation [22]. ENDOMIND can be downloaded for research purposes using this link: https://www.ukw.de/research/inexen/ai-for-polyp-detection/

Video based benchmark data set (ENDOTEST)

ENDOTEST was composed of two subsets of video based images (or frames): the validation and the performance dataset. Both were developed from retrospective recordings of colonoscopies from two centers (University Hospital Ulm and Würzburg) that differ from the data used for ENDOMIND training. Both centers used the Olympus CV-190 endoscopy processor. In one center, the recordings were performed using a video frame grabber (SDI2USB 3.0, Epiphan Systems Inc.) and stored with the same input resolution and minimal compression. Later, the video recordings were processed by GI Genius using an image converter (HA5, AJA Video Systems Inc.). In the second center, both signals (raw and GI Genius processed signal) were simultaneously recorded in real-time.

The validation dataset consists of a balanced dataset of 24 polyp and corresponding non-polyp sequences

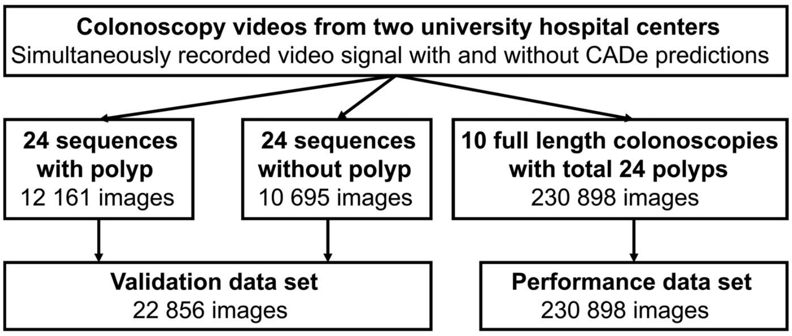

Colonoscopy videos from two university hospital centers
Simultaneously recorded video signal with and without CADe predictions
24 sequences with polyp
12 161 images
24 sequences without polyp
10 695 images
10 full length colonoscopies with total 24 polyps
230 898 images
Validation data set
22 856 images
Performance data set
230 898 images

Figure 2. ENDOTEST components including the validation and performance data set for the head-to-head comparison of the GI Genius and ENDOMIND.

Table 1. Characteristics of the 24 polyps included in the 10 videos of the performance data set.

Characteristic	n (%)
Paris classification	
0-Ip	1 (4.2)
0-Is	4 (16.7)
0-IIa	19 (79.2)
Size	
0–4 mm	11 (45.8)
5–10 mm	8 (33.3)
10–20 mm	5 (20.8)
Location	
Right Colon	16 (66.7)
Left Colon	5 (20.8)
Rectum	3 (12.5)
Histology	
Tubular adenoma	9 (37.5)
Sessile serrated lesion	12 (50)
Non-adenomatous lesion	3 (12.5)

comprising a total of 22,856 images that were manually annotated by marking of a bounding box over polyps (Figure 2). In the validation dataset, we considered the balance of 1:1 to be appropriate as many publicly available datasets do compare themselves on e.g. only polyp images and therefore have a ratio of 1:0 [12,13]. However, in general, equally balanced polyp and non-polyp video sequences do not resemble reality where visible polyps make up a minority of the total examination length. Therefore, additional 10 full colonoscopy videos were manually annotated by an experienced board certified gastroenterologist defining if the frame contained a polyp or not. Criteria for the selection of the videos included screening as the indication for the colonoscopy, the existence of minimum one adenomatous polyp, and a good bowel preparation with a BBPS score of 6 or higher. Annotation resulted in 230,898 images for the performance dataset. 15.8% of these images contained polyps. Those 24 polyps are characterized in Table 1. The performance dataset provides a more realistic scenario.

Validation of ENDOMIND and GI Genius

The manually annotated boxes using the raw colonoscopy video signal were defined as ground truth. For evaluation of the GI Genius system, the 24 polyp and corresponding non-polyp sequences were processed by the GI Genius device in real-time. The video output of GI Genius was recorded to be annotated in a second step. The resulting bounding box containing frames were reviewed by an experienced endoscopist. A frame was considered as true positive (TP) in case of overlap by the CADe bounding box and the annotated box. The absence of a CADe bounding box in a polyp-containing frame counted as a false negative (FN). A false positive detection (FP) was defined as a CADe bounding box that was not in contact with the manually annotated box.

In contrast, ENDOMIND algorithm was directly applied to every single frame of the raw colonoscopy videos for polyp detection. Both CADe systems were analyzed for accuracy, per-frame specificity, per-frame sensitivity, precision and F1-score.

Performance of ENDOMIND in comparison with GI Genius

For performance evaluation, the full-length colonoscopies were processed by the GI Genius device creating videos with bounding boxes. All frames with visible bounding boxes (CADe detections) were automatically identified by a custom-made application.

ENDOMIND for polyp detection was applied on the raw colonoscopy video signal using the performance data set. For performance analysis, a video frame was defined as TP if a CADe bounding box appeared in a frame that had been manually annotated to contain a polyp. Polyp-containing frames without a CADe bounding box were considered as FN. FP frames were considered if the CADe system drew a bounding box on video frame without a visible polyp.

A direct comparison of both CADe systems was performed regarding the standard metrics accuracy $= (TP + TN)/(TP + TN + FP + FN)$, per-frame specificity $= TN/(TN + FP)$, per-frame sensitivity (recall)$=TP/(TP + FN)$, precision $= TP/(TP + FP)$ and F1-score $= 2*precision*recall/(precision + recall)$ calculated from the TP, FP, FN und TN values. Furthermore, videos were analyzed for the median first detection time. FDT was defined for each polyp as the time in between the first appearance of the polyp in a video and the first marking with a bounding box by the CADe system.

Table 2. Comparison of ENDOMIND with GI Genius regarding accuracy, precision, specificity, sensitivity and F1-score using the validation data set with balanced polyp and non-polyp images and the performance data set with more non-polyp images in the video sequences.

	Validation Data Set		Performance Data Set	
	EndoMind	GI Genius	EndoMind	GI Genius
Accuracy	**85.7%**	79%	85%	**89.1%**
Precision	86.2%	**98%**	48.6%	**65.3%**
Per-frame specificity	84.1%	**98%**	92.7%	**93.7%**
Per-frame sensitivity = Recall	**87.1%**	62%	**54%**	48.1%
F1-score	**86.6%**	76%	**45.8%**	**45.8%**

Statistical analysis

Statistical analysis was performed using IBM SPSS Statistics 28. Wilcoxon Signed Ranks test was performed to test for significant differences between the paired groups. A p-value of $<.05$ indicated statistical significance.

Ethical considerations

The study including retrospective and prospective collection of examination videos and reports was approved by the responsible institutional review board (Ethical committee Stuttgart, 21 January 2021, F-2020-158). The study was registered at the German Clinical Trial Register (26 March 2021, DRKS-ID: DRKS00024150). Signed informed consent from each patient for data recording was obtained for the prospective data collection.

Results

Both systems detected all polyps in both data sets in at least one image. A summary of the comparison of both systems for both data sets included in ENDOTEST with five standard metrics is shown in Table 2. ENDOMIND had a significantly higher per-frame sensitivity (recall) but lower per-frame specificity and precision in both data sets compared to GI Genius. This is represented by a mostly continuous detection of polyps by ENDOMIND in every frame whereas GI Genius often stops detecting the same polyp intermittently. This results in a flickering bounding box. See Supplementary video as an example. In aggregated metrics combining both false positives and false negatives ENDOMIND had a lead of 85.7% accuracy and 86.6% F1-score compared to GI Genius (79% accuracy and 76% F1-score) in the balanced validation data set, while GI Genius was partially better in the unbalanced performance data set with 89.1% accuracy and 45.8% F1-Score compared to ENDOMIND (85% accuracy and 45.8% F1-Score). The rate of false positive detections in the full length colonoscopies was 6% in case of ENDOMIND and 2% in case of GI Genius.

An important factor beside the recognition of a neoplastic lesion is how fast an algorithm detects a polyp and alerts or orientates the examiner through the bounding box to examine a region more closely. We evaluated this period as FDT in performance data set. ENDOMIND had a significantly faster median FDT of 217 milliseconds (ms; with an Inter-Quartile Range (IQR) between 8 and 1533 ms) compared with 1050 ms (IQR 358, 2767 ms) of GI Genius ($p = .003$). ENDOMIND detected 79.2% of the polyps faster than GI Genius (Figure 3).

Discussion

In this work we introduce ENDOTEST, a novel video-based benchmark dataset that includes polyp sequences and full-length screening colonoscopies with minimum one adenomatous polyp. In addition, we have compared our recently developed CADe system, ENDOMIND, with a commercially available CADe system. ENDOMIND has been developed as part of a publicly funded investigator-initiated project. Besides the usage of data obtained from routine colonoscopies in an outpatient setting in gastroenterological practices, we aimed to evaluate the CADe system using a head-to-head comparison with the commercially available CADe system GI Genius with the software version present on March 2020. To our knowledge, there is no previous study comparing different CADe.

As a basis for our deep learning algorithm, we chose YOLOv5 [21]. YOLOv5 offers two significant advantages: First, it is an end-to-end neural network and thereby fast and easy to use in a real-time system. Thus, it enables the system to run during the clinical routine. Second, YOLOv5 is one of the AI models with a very high detection rate while not overfitting detection box accuracy. Finding the polyp is considered more important than drawing the bounding box 100% precisely around the edges of the polyp. YOLOv5 is an algorithm maximizing those aspects.

Beside the new algorithm used for ENDOMIND, we focused on optimal training data. We decided not to include single images of polyps generated for the endoscopic report since those images often present a cleaned polyp viewed from an optimal angle. This does not reflect the look of a polyp at its first appearance. Thus, AI might not learn how to recognize polyps at this early stage of visualization. Our training data consists exclusively of videos. Accordingly, the performed annotation of polyps included the earliest time point of polyp appearance. Additionally, the endoscopies were performed in gastroenterological practices. This is in contrast to many other CADe systems where training data mainly relies on colonoscopy videos recorded in an in-hospital setting [6,8,23]. The vast majority of the screening colonoscopies are performed in gastroenterological practices. Therefore, usage of data coming from this source as training data for the development of polyp detection systems might be better suited for the intended purpose.

Validation data sets used for different CADe systems vary largely [6,23,24]. The proportion of polyp and non-polyp images is mainly composed in a balanced manner. However, differences regarding polyp morphology and size in between those validation data sets prevent direct benchmarking. Still, the performance of ENDOMIND on validation dataset of polyp sequences is comparable to other published CADe systems with a per-frame sensitivity of 82%–86% and specificity of 86%–89%: on publicly available datasets. Indeed, using just still images of polyps results in a marked higher

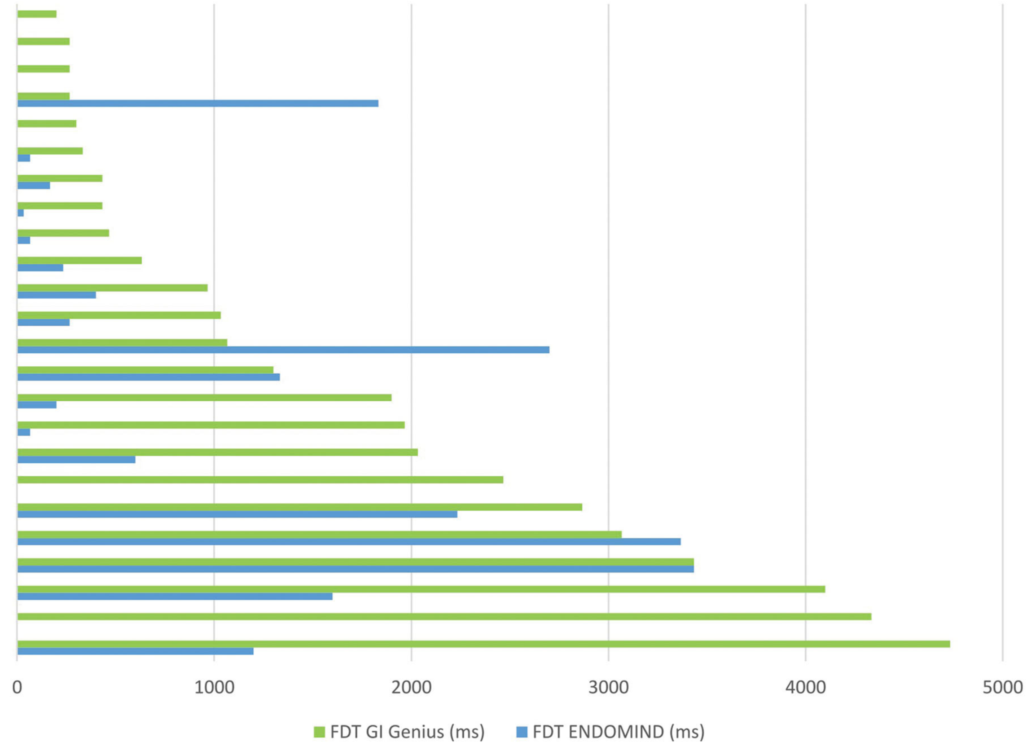


Figure 3. Time delay from the first appearance of a polyp to the detection by ENDOMIND and GI Genius is shown for each of the 24 polyps in the performance data set ordered by the detection time of GI Genius. If there is no bar, the polyp was detected in the earliest possible image within the video.

sensitivity and specificity [25]. Due to the fact that validation data sets of commercially available CADe systems are not publicly available, we applied the already CE-approved CADe system GI Genius to our data set. The direct comparison showed a marked higher per-frame sensitivity of ENDOMIND with inferiority in regard to per-frame specificity and precision and a higher accuracy score. As for many screening tools CADe developers prioritize sensitivity [23,26]. Therefore, polyp images are overrepresented in publicly available validation data sets, and results derived from those datasets are difficult to be extrapolated to performance in real-life use. To overcome this issue, ENDOTEST additionally included full-length screening colonoscopies that directly resemble situations in daily work of an endoscopist. In both evaluation data sets, ENDOMIND has a much higher per-frame sensitivity (recall), while GI Genius showed superior per-frame specificity. If the data set contains many polyps frames like in the validation data set, this results in better aggregate values like accuracy and F1-score for ENDOMIND, and if the data set contains more images without polyps like in the performance data set, GI Genius has better aggregate values. Since ENDOMIND has been trained with a data set where roughly half of the images contained polyp, this behavior could have been expected. Additionally, our data set ENDOTEST provides high-quality compared to other open-source databases like SUN Colonoscopy Video Database, LDPolypVideo-Benchmark, CVC-ClinicVideoDB [12–14]. Those datasets include sequences of polyps and non full-length colonoscopies. These polyp sequences are usually recorded after polyp identification with subsequent cleaning of the polyp and therefore do not provide a realistic scenario of the intervention. As ENDOTEST includes full-length colonoscopies we provide a more realistic approach and therefore consider our data to be of higher quality. Furthermore, the first crucial frames of polyp appearance are included, which allows to measure the FDT. ENDOTEST is therefore divided in two subsets: the classic evaluation using video sequences, and the full-length colonoscopies that represent a more realistic scenario.

The main reason behind optimizing sensitivity (recall) over specificity and precision was, that the system should alert physicians to polyps they might miss during a colonoscopy. Such polyps are visible for a short time period only – much shorter than in the videos used in the performance evaluation, where the polyps were detected by the physicians and therefore were in view for an extended period of time. If it takes longer to detect a polyp, the polyp might go out of view and remain undetected. Therefore, the FDT is a key indicator for the benefit of a CADe. Here, ENDOMIND detected 79.2% of the polyps faster than GI Genius. This is a direct consequence of the higher sensitivity (recall) of ENDOMIND and comes at the cost of more false alarms, as the lower precision of ENDOMIND shows. However, we regard a type II error of missing a polyp as more severe than a type I error of a false alarm. Nasohisa *et al* described that twofold withdrawal velocity dramatically decreases sensitivity per lesion of the CADe system CAD EYE (Fujifilm Europe GmbH, Düsseldorf, Germany) [27]. Indeed withdrawal time shows variability among endoscopists [28] with impact on

ADR. Thus, a faster detection of a polyp visible only for a fraction of a second might help to urge the examiner to revisit the polyp location. The lower sensitivity of GI Genius results in non-consistent recognition of polyps with short interruptions that are visible as a flickering signal (Supplementary video).

A field for future improvement of ENDOMIND is the comparatively low precision corresponding to the lower specificity and the higher FP rate compared to GI Genius. Pfeifer *et al* showed a similar rate of FP in the commercially available CADe system Discovery AI (Pentax Europe GmbH, Hamburg, Germany) [9]. A high rate of FP detections might have the potential to induce distrust of the endoscopist to the CADe system. The examiner might thereby inspect bounding boxes less thoroughly. Still, compared to the GI Genius validation study published by Hassan C *et al.* the current GI Genius FP rate in our dataset is higher [24]. This once more illustrates the effect of different evaluation data sets. A uniform evaluation dataset used to measure performances of different CADe systems is therefore needed.

Our future focus is on adding video sequences of normal mucosa to the training data in order to overcome the high rate of FP detections. Additionally, applying a threshold that suppresses short detections flagged by the CADe with only low precision might help. Alternatively, a different deep learning network architecture can be chosen like in a recent published study with only minimal false positive detections [29]. Another future interesting point would be the comparison of the results here obtained with other publicly available datasets in order to assess the quality of ENDOTEST.

A limitation of our study might be the differing CADe analysis of the data sets. The FDT of the CADe GI Genius might have been influenced by the delay due to the input and output of the video signal into the device whereas ENDOMIND predictions were done on a high-performance computer on every video frame without the delay described above. We estimate the delay of ENDOMIND during real-time endoscopy to be further increased by 25 ms. Liu P *et al* reported a delay of 20 ms of their CADe system [26] which is at the same level as the 50 ms a CAD system reported by Byrne *et al* [30]. Additional evaluation with a complete computer set up of the CADe system directly attached to the endoscopy processor is therefore needed in the next step to overcome the mentioned limitations. Additionally, although not observed, the minimal compression of the videos during the recording of the colonoscopies could affect the CADe system since the signal does not originate directly from the endoscopy processor.

Conclusion

In this work, we have used the generated ENDOTEST database to compare two computer-aided polyp detection systems. ENDOTEST contains full-length screening colonoscopies, frame-by-frame annotated regarding the presence of a polyp. Therefore, it resembles a more realistic scenario than previous existing databases, and allows the calculation of the crucial parameter FDT. In addition, the developed CADe system prototype ENDOMIND, has shown promising performance and sensitivity in this preliminary evaluation when compared with a commercially available CADe system. However, further training data is needed to increase the precision to avoid false alarms. In general, there is a lack of common definitions of quality measures of annotation and validation. Benchmark data sets for validation of CAD systems could overcome these limitations. Even more important, we currently lack realistic data sets for evaluation containing undetected polyps being presented only for a short period of time, since their detections is the main purpose of the systems.

Acknowledgements

The authors acknowledge the support by Prof. J.F. Riemann, "Stiftung Lebensblicke" the Foundation for early detection of colon cancer.

Author contributions

DF, AK, JT, FP and AH: study concept and design. DF, AK and JT: performed the experiments. DF, AK, JT, TR, FP and AH: interpretation of results, and drafting of the manuscript. DF, FP and AH: statistical analysis. DF, JT, MB, BS, WB, and WGZ: acquisition of data. All authors: critical revision of the article for important intellectual content and final approval of the article.

Disclosure statement

The authors report there are no competing interests to declare.

Funding

This work was supported by the State Government of Baden-Württemberg, Germany (Funding cluster "Forum Gesundheitsstandort Baden-Württemberg") under Grant number 5409.0-001.01/15).

ORCID

Alexander Hann http://orcid.org/0000-0001-8035-3559

References

[1] Kaminski MF, Robertson DJ, Senore C, et al. Optimizing the quality of colorectal cancer screening worldwide. Gastroenterology. 2020;158(2):404–417.
[2] Corley DA, Jensen CD, Marks AR, et al. Adenoma detection rate and risk of colorectal cancer and death. N Engl J Med. 2014; 370(14):1298–1306.
[3] Zhao S, Wang S, Pan P, et al. Magnitude, risk factors, and factors associated with adenoma miss rate of tandem colonoscopy: a systematic review and meta-analysis. Gastroenterology. 2019; 156(6):1661–1674.e11.
[4] Brenner H, Altenhofen L, Kretschmann J, et al. Trends in adenoma detection rates during the first 10 years of the German screening colonoscopy program. Gastroenterology. 2015;149(2):356–366.e1.
[5] Hassan C, Spadaccini M, Iannone A, et al. Performance of artificial intelligence for colonoscopy regarding adenoma and polyp detection: a meta-analysis. Gastrointest Endosc. 2020; 93(1):77–85.

[6] Liu W-N, Zhang Y-Y, Bian X-Q, et al. Study on detection rate of polyps and adenomas in artificial-intelligence-aided colonoscopy. Saudi J Gastroenterol. 2020;26(1):13–19.

[7] Wang P, Liu X, Berzin TM, et al. Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): a double-blind randomised study. Lancet Gastroenterol Hepatol. 2020;5(4):343–351.

[8] Su J-R, Li Z, Shao X-J, et al. Impact of a real-time automatic quality control system on colorectal polyp and adenoma detection: a prospective randomized controlled study (with videos). Gastrointest Endosc. 2020;91(2):415–424.e4.

[9] Pfeifer L, Neufert C, Leppkes M, et al. Computer-aided detection of colorectal polyps using a newly generated deep convolutional neural network: from development to first clinical experience. Eur J Gastroenterol Hepatol. 2021;33(1S Suppl 1):e662–e669.

[10] Repici A, Badalamenti M, Maselli R, et al. Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial. Gastroenterology. 2020;159(2):512–520.e7.

[11] Weigt J, Repici A, Antonelli G, et al. Performance of a new integrated computer-assisted system (CADe/CADx) for detection and characterization of colorectal neoplasia. Endoscopy. 2022;54(2): 180–184.

[12] Bernal J, Sánchez FJ, Fernández-Esparrach G, et al. WM-DOVA maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians. Comput Med Imaging Graph. 2015;43:99–111.

[13] Misawa M, Kudo S-E, Mori Y, et al. Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). Gastrointest Endosc. 2021;93(4):960–967.e3.

[14] Ma Y, Chen X, Cheng K, et al. 2021. LDPolypVideo benchmark: a large-scale colonoscopy video dataset of diverse polyps. In de Bruijne M, Cattin PC, Cotin S, Padoy N, Speidel S, Zheng Y, Essert C (eds) Medical image computing and computer assisted intervention – MICCAI 2021. Springer International Publishing, Cham, pp 387–396.

[15] Angermann Q, Bernal J, Sánchez-Montes C, et al. 2017. Towards real-time polyp detection in colonoscopy videos: adapting still frame-based methodologies for video sequences analysis. In: Cardoso Arbel T, Luo X, Wesarg S, Reichl T, González Ballester MÁ, McLeod J, Drechsler K, Peters T, Erdt M, Mori K, Linguraru MG, Uhl A, Oyarzun Laura C, Shekhar R (eds) Computer assisted and robotic endoscopy and clinical Image-Based procedures. Springer International Publishing, Cham, pp 29–41.

[16] Endoscopic Classification Review Group. Update on the Paris classification of superficial neoplastic lesions in the digestive tract. Endoscopy. 2005;37:570–578.

[17] Krenzer A, Makowski K, Hekalo A, et al. Semi-automated machine learning video annotation for gastroenterologists. Stud Health Technol Inform. 2021;281:484–485.

[18] Ali S, Braden B, Lamarque D, et al. 2020. Endoscopy Disease Detection and Segmentation (EDD2020)

[19] Jha D, Smedsrud PH, Riegler MA, et al. 2020. Kvasir-SEG: a segmented polyp dataset. In: Ro YM, Cheng W-H, Kim J, Chu W-T, Cui P, Choi J-W, Hu M-C, De Neve W (eds) MultiMedia modeling. Springer International Publishing, Cham, pp 451–462.

[20] Vázquez D, Bernal J, Sánchez FJ, et al. A benchmark for endoluminal scene segmentation of colonoscopy images. J Healthc Eng. 2017;2017:4037190.

[21] Redmon J, Farhadi A. 2018. YOLOv3: An Incremental Improvement. arXiv:180402767 [cs]

[22] Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data. 2019;6(1):60.

[23] Wang P, Xiao X, Glissen Brown JR, et al. Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. Nat Biomed Eng. 2018;2(10):741–748.

[24] Hassan C, Wallace MB, Sharma P, et al. New artificial intelligence system: first validation study versus experienced endoscopists for colorectal polyp detection. Gut. 2020;69(5):799–800.

[25] Wang D, Zhang N, Sun X, et al. 2019. AFP-Net: realtime anchor-free polyp detection in colonoscopy. In 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, Portland, OR, USA, pp 636–643

[26] Liu P, Wang P, Glissen Brown JR, et al. The single-monitor trial: an embedded CADe system increased adenoma detection during colonoscopy: a prospective randomized study. Therap Adv Gastroenterol. 2020;13:1756284820979165.

[27] Yoshida N, Inoue K, Tomita Y, et al. An analysis about the function of a new artificial intelligence, CAD EYE with the lesion recognition and diagnosis for colorectal polyps in clinical practice. Int J Colorectal Dis. 2021;36(10):2237–2245.

[28] Benson ME, Reichelderfer M, Said A, et al. Variation in colonoscopic technique and adenoma detection rates at an academic gastroenterology unit. Dig Dis Sci. 2010;55(1):166–171.

[29] Livovsky DM, Veikherman D, Golany T, et al. Detection of elusive polyps using a large-scale artificial intelligence system (with videos). Gastrointest Endosc. 2021;94(6):1099–1109.e10.

[30] Byrne MF, Chapados N, Soudan F, et al. Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model. Gut. 2019;68(1):94–100.

Received: 8 February 2022 | Accepted: 31 March 2022

DOI: 10.1002/ueg2.12235

ORIGINAL ARTICLE

UNITED EUROPEAN GASTROENTEROLOGY ueg journal WILEY

Development and evaluation of a deep learning model to improve the usability of polyp detection systems during interventions

Markus Brand[1] | **Joel Troya**[1] | **Adrian Krenzer**[1,2] | **Zita Saßmannshausen**[1] | **Wolfram G. Zoller**[3] | **Alexander Meining**[1] | **Thomas J. Lux**[1] | **Alexander Hann**[1]

[1]Interventional and Experimental Endoscopy (InExEn), Internal Medicine II, Gastroenterology, University Hospital Würzburg, Würzburg, Germany

[2]Artificial Intelligence and Knowledge Systems, Institute for Computer Science, Julius-Maximilians-Universität Würzburg, Würzburg, Germany

[3]Department of Internal Medicine and Gastroenterology, Katharinenhospital, Stuttgart, Germany

Correspondence
Joel Troya, Universitätsklinikum Würzburg, Medizinische Klinik und Poliklinik II, Oberdürrbacher Str. 6, 97080 Würzburg, Germany.
Email: TroyaSebas_J@ukw.de

Funding information
Dieter von Holtzbrinck; Eva Mayr-Stihl Foundation; Fischerwerke GmbH; Forum Gesundheitsstandort Baden-Württemberg
Open access funding enabled and organized by Projekt DEAL.

Abstract

Background: The efficiency of artificial intelligence as computer-aided detection (CADe) systems for colorectal polyps has been demonstrated in several randomized trials. However, CADe systems generate many distracting detections, especially during interventions such as polypectomies. Those distracting CADe detections are often induced by the introduction of snares or biopsy forceps as the systems have not been trained for such situations. In addition, there are a significant number of non-false but not relevant detections, since the polyp has already been previously detected. All these detections have the potential to disturb the examiner's work.

Objectives: Development and evaluation of a convolutional neuronal network that recognizes instruments in the endoscopic image, suppresses distracting CADe detections, and reliably detects endoscopic interventions.

Methods: A total of 580 different examination videos from 9 different centers using 4 different processor types were screened for instruments and represented the training dataset (519,856 images in total, 144,217 contained a visible instrument). The test dataset included 10 full-colonoscopy videos that were analyzed for the recognition of visible instruments and detections by a commercially available CADe system (GI Genius, Medtronic).

Results: The test dataset contained 153,623 images, 8.84% of those presented visible instruments (12 interventions, 19 instruments used). The convolutional neuronal network reached an overall accuracy in the detection of visible instruments of 98.59%. Sensitivity and specificity were 98.55% and 98.92%, respectively. A mean of 462.8 frames containing distracting CADe detections per colonoscopy were avoided using the convolutional neuronal network. This accounted for 95.6% of all distracting CADe detections.

Conclusions: Detection of endoscopic instruments in colonoscopy using artificial intelligence technology is reliable and achieves high sensitivity and specificity.

Markus Brand, Joel Troya, Thomas J. Lux and Alexander Hann contributed equally to the work.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. United European Gastroenterology Journal published by Wiley Periodicals LLC on behalf of United European Gastroenterology.

Accordingly, the new convolutional neuronal network could be used to reduce distracting CADe detections during endoscopic procedures. Thus, our study demonstrates the great potential of artificial intelligence technology beyond mucosal assessment.

KEYWORDS
CADe, colonoscopy, deep learning, instrument, intervention

INTRODUCTION

Artificial intelligence (AI) for colonic polyp detection is the most important application of this new technology in gastrointestinal endoscopy to date. Efficiency and functionality of these computer-aided detection (CADe) systems have been demonstrated in several randomized trials.[1–6] However, CADe systems also show many false positive (FP) detections.[7,8] These false markings can affect the examiner's concentration. If a false detection occurs in addition to a relevant finding, the examiner's attention may be distracted, leading to missed findings in the worst case.[9]

Daily use of CADe systems shows that endoscopic interventions (especially biopsies and polypectomies) lead to many false activations of CADe systems. In this case, false positive activations occur due to the inserted instruments (forceps, needle, snare), but also due to intervention on the mucosa itself (injection, resection, clipping). In addition, there are a significant number of non-false but not relevant detections, since the polyp has already been previously detected. To enable the investigators to put their full concentration on the intervention, no distracting AI signals regarding polyp detection should be visible during the procedure.

Therefore, the aim of the current study was to develop and evaluate an AI system that reliably detects introduced instruments in order to disable the CADe system during an intervention and avoid distracting detections.

Key summary

Summarize the established knowledge on this subject

- Multiple computer-aided diagnosis (CADe) systems for polyp detection are currently introduced into clinical practice.
- CADe results in multiple distracting detections, especially during therapeutic interventions when instruments are visible.

What are the significant and/or new findings of this study?

- Development and evaluation of a deep learning model to recognize visible instruments that are used for therapeutic intervention in gastrointestinal endoscopy.
- Our model automatically prevents distracting CADe detections during therapeutic interventions.

METHODS

Training dataset

Data from nine different center in Germany, two university hospitals, one community-based hospital and six gastroenterology practices, were collected retrospectively from March 2019 to August 2021. A total of 519,856 images were selected from 580 randomly selected different colonoscopy videos for building the training dataset. Of all images in the training dataset, 144,217 (27.7%) contained a visible instrument (Figure 1). The types of instruments used for training the model included graspers, hot and cold snares, injection needles and clips. No minimum or maximum number of images per instrument in a colonoscopy was predefined for training the model. Images of good and poor quality (e.g., blurry images) were chosen for model training in order to represent a real-life scenario. The colonoscopies were performed using different processors including CV-190 and CV-170 (Olympus Europa SE & Co. KG, Hamburg, Germany), Image1 S (Karl Storz SE & Co. KG, Tuttlingen, Germany) and EPK-i7000 (Pentax Europe GmbH, Hamburg, Germany) and were recorded using a standard computer with a video grabber (DeckLink Mini Recorder, Blackmagic Design Pty Ltd., Melbourne, Australia) and a custom recording software. Representative images of the four different processor types are displayed in Supplementary Figure 1.

To reduce almost identical images in the dataset, images from the same colonoscopy were filtered to exclude neighboring images. For training the convolutional neuronal network (CNN), the dataset was split into a train (90%) and a validation (10%) dataset. To prevent bias, all images of one colonoscopy were either included in the train or the validation dataset.

Preprocessing and CNN training

Initially, a region of interest for each used processor type was defined and images were cropped accordingly. Afterward, images were zero padded and resized to a dimension of 512 × 512 pixels to yield uniform images. For train data, the image augmentation pipeline (Supplementary Code Section 1) was applied. All images underwent the standard procedure of image normalization (Supplementary Code Section 2), so that each color and brightness value are standardized. Resulting

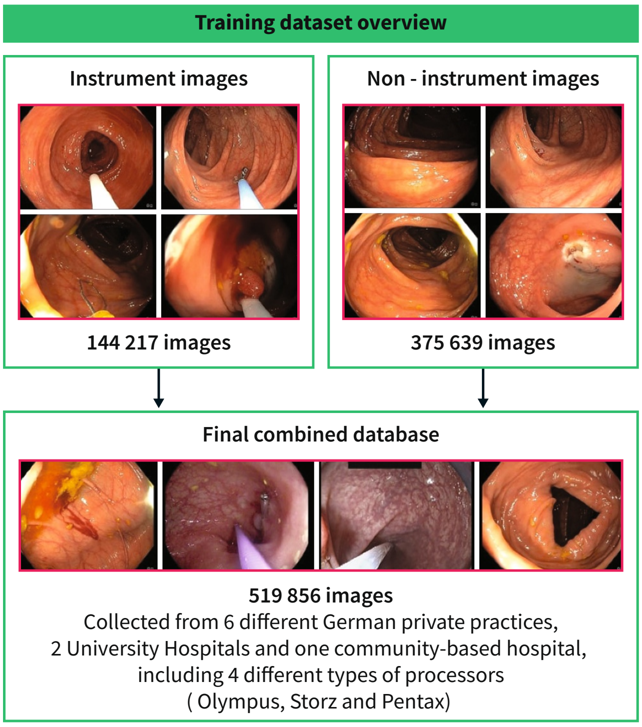

Training dataset overview
Instrument images
144 217 images
Non - instrument images
375 639 images
Final combined database
519 856 images
Collected from 6 different German private practices,
2 University Hospitals and one community-based hospital,
including 4 different types of processors
(Olympus, Storz and Pentax)

FIGURE 1 Characteristics of the training dataset containing images with and without visible instruments captured using the four different endoscopy processor types

constant values for this were calculated on all train data images and were always used to normalize each input image. CNN training is described in detail in Supplementary Material. The current instrument detection software is freely available to download for research purposes (https://github.com/Maddonix/instrument_detection).

Model testing

The CNN for detecting visible instruments was tested in the withdrawal phase of a set of 10 full-length colonoscopy videos by analyzing its performance in each single image of the videos. To stabilize the prediction results, a running mean function was applied. This was performed to avoid erroneous suppressions of AI detections caused by our CNN. Here, we assigned the current video image the majority label of itself as well as the previous 14 video images, representing a threshold of 467 ms. This same dataset was used to test the change in performance of a CADe system (GI Genius, Medtronic Inc., Ireland, Version March 2020) in the reduction of distracting activations with the developed instrument detection system.

Ethics approval

Patients provided written informed consent prior to video recording. The ethics committee of the University hospital Würzburg approved retrospective analysis of the data used in this study.

Statistical analysis

Two evaluations were statistically analyzed: the capabilities of the instrument detection system and the reduction of CADe distracting activations. Per-frame sensitivity and specificity, accuracy, and Receiver Operating Characteristic (ROC) were calculated for both evaluations. Sensitivity has been defined as the ratio between the number of frames with a visible instrument that were correctly detected (TP) and the total number of frames with a visible instrument (TP+FN). Specificity was defined as the ratio between the number of frames without a visible instrument that were correctly assessed (TN) and the sum of the total number of frames with a false detection and the TN frames (FP+TN). Accuracy was defined as the ratio between the number of correct system assessments (TP +TN) and the total number of frames. Metrics where weighted average to compensate for the imbalance of images with/out a visible instrument. For the calculation of the weighted average metrics the parameter "average" in every used function of the sklearn.metrics module from scikit-learn 1.0.2 package was set to "weighted". All calculations were performed using Python Software (version 3.6).

RESULTS

Characteristics of the patient cohort

The test dataset comprised 10 full-length colonoscopy videos from 10 different patients. Men and women were equally represented, the mean age was 57.1 (interquartile range; 46–65) and the mean Boston Bowel Preparation Scale was 6.9 (range; 6–9) (Supplementary Table 1). The total duration of the withdrawal phase, with the duration of interventions included, was 1 h and 25 min, corresponding to 153,623 single video frames. During this time, instruments were visible for a total of 7 min and 12 s on the screen These 10 videos included a total of 12 different interventions, where 19 different endoscopic through the scope instruments were used: 4 cold snares, 11 graspers, 1 hot snare, 2 needles and 1 clip (Table 1).

Performance of the instrument detection system

The CNN overall accuracy achieved in the detection of visible instruments in the test dataset was 98.59%. Sensitivity and specificity were 98.55% and 98.92%, respectively. The grasper was the instrument that was best detected by the system, with a sensitivity of 99.08% and a specificity of 99.36%, whereas the snare, with a sensitivity of 98.21% and a specificity of 98.51%, was the most difficult instrument to detect, probably because often only the wire was visible. Representative images of a grasper, a snare and a false positive detection of the CNN with the corresponding heat map that depicts the image areas that are recognized as an instrument are

TABLE 1 Characteristics and performance of the instrument detection system in the test dataset

	Intervention	Type of instrument	Number of visible instrument frames	Sensitivity (%)	Specificity (%)	Disturbing CADe activations (frames)	Disturbing CADe activations avoided (frames)	False-avoided CADe activations (frames)	Total number of CADe activations
Video 1	Polypectomy	Snare	728	98.60	99.51	377	330	13	3352
	Polypectomy	Snare	1262						
Video 2	Polypectomy	Grasper	142	99.77	99.76	6	6	0	396
Video 3	Polypectomy	Grasper	269	99.21	99.68	137	102	5	1252
Video 4	Polypectomy	Grasper	174	98.87	98.43	8	8	2	302
Video 5	Polypectomy	Needle	407	99.22	99.64	1232	1184	54	7834
		Snare	931						
Video 6	Polypectomy	Grasper	1136	99.31	99.67	161	150	6	531
Video 7	Polypectomy	Snare	2760	98.01	98.35	741	736	50	1577
Video 8	Polypectomy	Needle	2493	96.90	96.58	1923	1906	204	7737
		Hot snare	1048						
		Clip	292						
Video 9	Polypectomy	Snare	255	99.33	99.62	101	84	21	1361
Video 10	Random biopsies	5x Grasper	751	98.14	98.98	153	122	2	1099
	Polypectomy	Grasper	871						

Abbreviation: CADe, Computer-aided detection system.

presented in Figure 2. The ROC curve illustrating the diagnostic ability of our instrument detection system is depicted in Figure 3. No marked differences in performance were observed relating to BBPS value, that ranged from 6 to 9 (Supplementary Table 1).

Reduction of CADe distracting activations

A total of 25,441 activations were triggered by the CADe system in the test dataset. These activations included detected polyps and false positive detections. 4839 activations (19.02%) occurred when an instrument was visible and were regarded as distracting CADe activations. Especially significant was the amount of distracting activation caused by snares or needles. Our system was able to avoid 4628 of these activations, representing a sensitivity value of 95.64%. Regarding the number of CADe activations that were falsely avoided, the value in frames amounts to 357. Out of those, 292 contained polyps that were previously detected. This implies that the system has a specificity of 98.62% in terms of performance in preventing distracting CADe activations. The overall accuracy was of 99.17%.

The metrics of the developed instrument detection system per intervention and its performance in preventing distracting CADe activations are presented in Table 1 and exemplarily illustrated in Figure 4. In addition, Figure 5 and Video S1 present a graphical example of how the AI system works.

DISCUSSION

Since the introduction of commercially available AI-systems for colorectal polyp detection, the use of these promising systems in daily practice is increasing. The great potential of AI-systems is currently in the field of diagnostics, as CADe systems support the examiner in real time and with high sensitivity.[6,10,11] Since CADe systems have been trained with diagnostic polyp images, they achieve high sensitivity for native polyps in the colon.[12] However, changes to the mucosa in the course of an intervention (e.g., injection) lead to false positive detections, as the systems have not been trained for such situations. The instruments used during the intervention also lead to many false positive detections that may disturb the investigator's concentration. In addition, there are many non-false but irrelevant detections because the polyp causing the intervention has been previously detected. To enable the investigators to put their full concentration on the intervention, no distracting AI signals should be visible during the intervention. This could be achieved by suppressing the CADe signal during the intervention, since polyp detection is not necessary during the intervention.

Currently, the endoscopist can only manually turn off the CADe system and turn it back on after the procedure. Some systems require a button to be pressed on the processor, as not all systems can be controlled via a button on the endoscope. However, it is possible that the endoscopist forgets to turn the system back on after the procedures. Therefore, automatically stopping and starting the

CADe system would increase the comfort for the endoscopist and prevent the CADe system from being accidentally switched off permanently.

Our novel AI system detects inserted instruments with high sensitivity and specificity. Therefore, the system can capture the time frame of an endoscopic intervention with high accuracy. This would enable the suppression of the CADe signal for the duration of the intervention to focus the investigator's concentration on the intervention. The suppression relates not only to false positive detections but instead to all CADe detections during an intervention that do not add value to the endoscopic image.

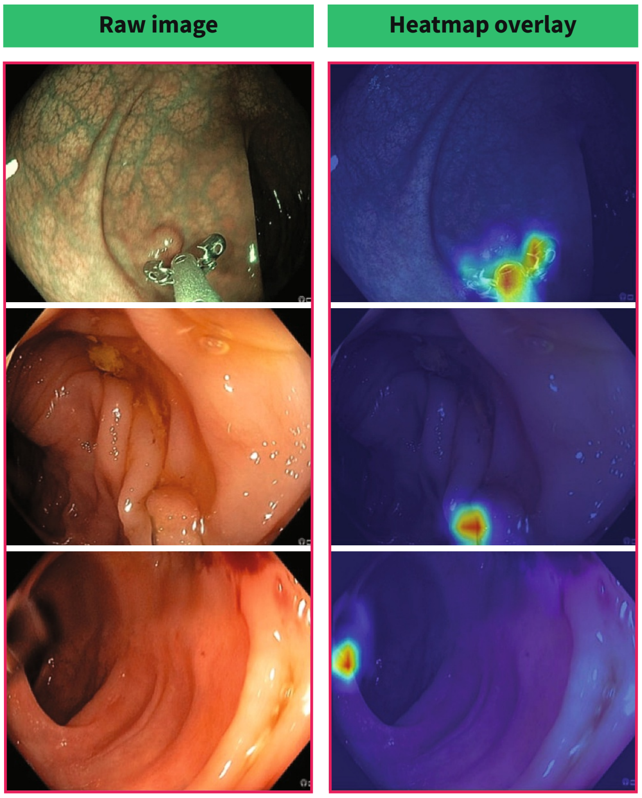


FIGURE 2 Grasper (upper row), snare (middle row), and a false positive detection (lower row) of the instrument detecting CNN with the corresponding gradient-weighted class activation mapping (Grad-CAM). Grad-CAM images on the right side visualize areas responsible for the CNN prediction as an instrument

The requirements for such a tool detection system are very high, as suppression of the CADe signal outside of an intervention (false positive instrument detection) may increase the risk of missing other visible polyps. Our study shows that our new AI system achieves a very high specificity, which is sufficient for this purpose. To obtain this high specificity, our system was trained with a large number of images from multiple centers using different endoscopy processors. The number of training images we used is comparable to the number used in development of other CADe systems.[13,14] In addition, the optimized algorithm presents only a short delay of 467 ms, that allows for the real-time use in combination with a CADe system.

Since the sensitivity of our AI system is in a high range, the instruments introduced were missed in only a few frames during an intervention. This applies in particular to the insertion and removal of an instrument where only a small portion of it is visible at the edge of the endoscopic view. Once the instrument is in the normal working position, it is quickly and reliably detected by the AI system. Thus, the crucial part of the intervention is captured by our instrument detection system. However, a problem with instrument detection arises when an instrument is pressed so firmly into the mucosa that it is barely visible. In this situation, the instrument recognition works accordingly worse. Nevertheless, our video analysis showed that the new AI system significantly reduced the number of false-positive CADe detections during an endoscopic intervention. While many publications on AI systems only use short, specially selected video sequences in the evaluation phase, our system was tested on full-length colonoscopy, which brings the results much closer to the real examination situation.[15]

Interestingly, the commercially available CADe system seems to generate more detections when a snare is used in comparison to a grasper. There might be different explanations for this phenomenon.

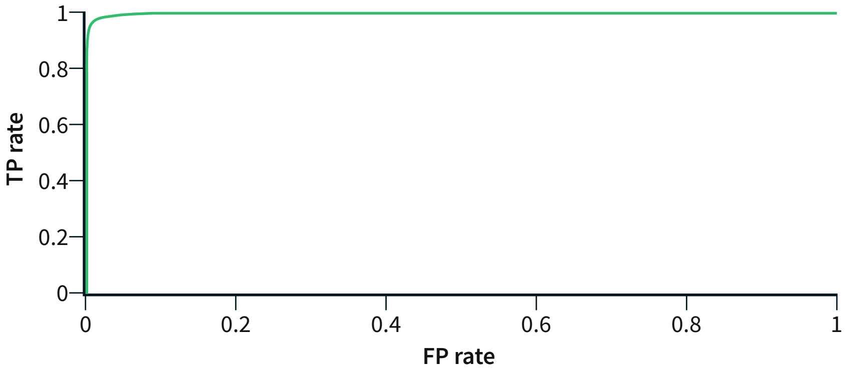


FIGURE 3 Receiver Operating Characteristic Curve of the instrument detection CNN visualizes specificity. While adjusting classification thresholds, the TP rate reaches 96.58% while maintaining a FP rate of 1% resulting in an area under the curve of 0.9971. CNN, convolutional neuronal network; FP, false positive; TP, true positive

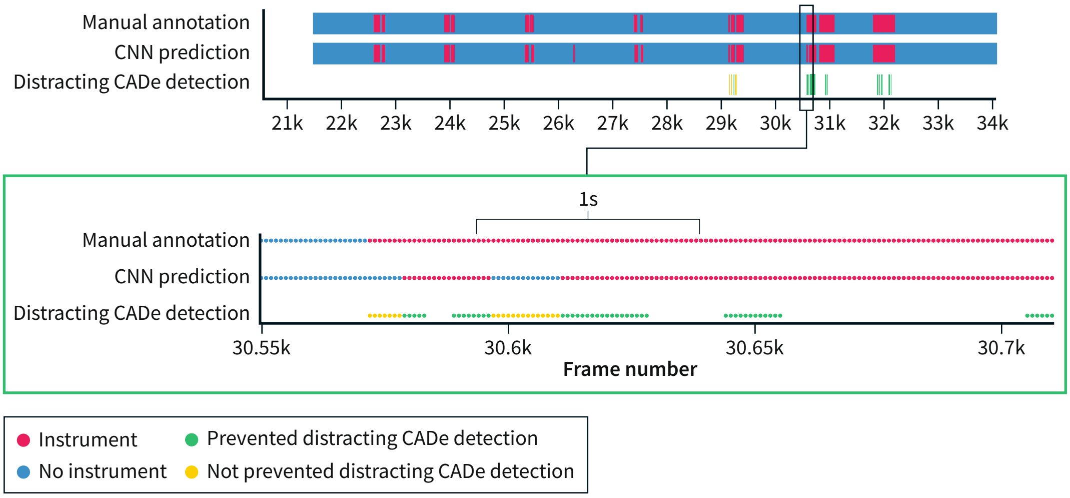


FIGURE 4 Schematic overview of the images with (red) and without (blue) visible instruments in a coloscopy video. The first row represents the manual annotations of whether the corresponding image contains a visible instrument. The second row represents the predictions output by our CNN. The third row represents the distracting CADe activations successfully prevented (green) or unsuccessfully prevented (yellow) by using the developed instrument detection CNN. The inset shows 160 frames (one dot per frame) which correlate to 5.33 s in the video. CADe, Computer-aided detection system; CNN, convolutional neuronal network

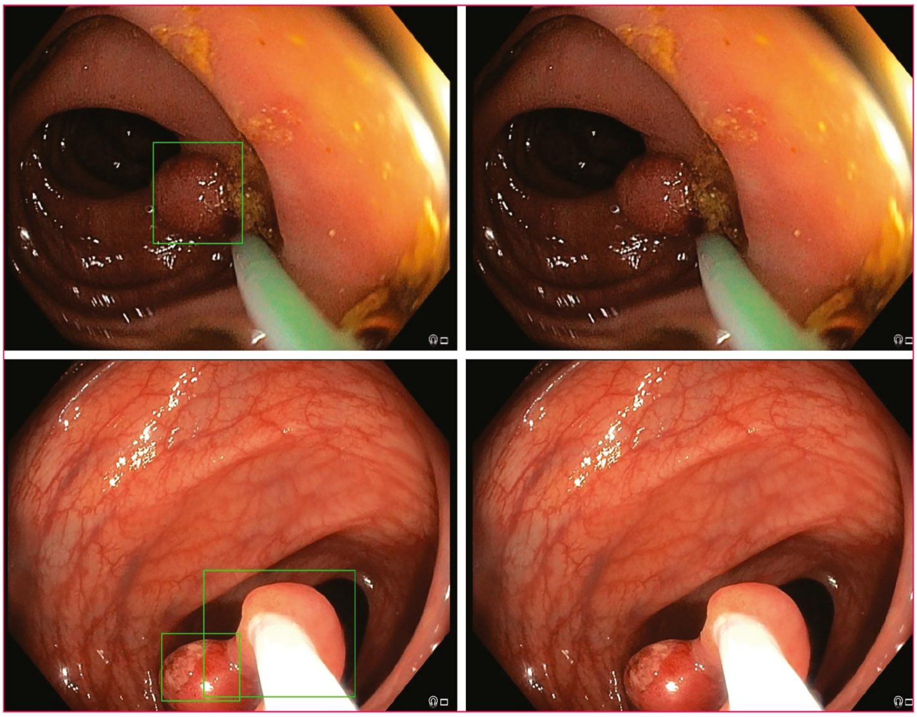


FIGURE 5 Single images of a polypectomy involving a needle for submucosal injection (upper row) and a snare (lower row) using the computer-aided polyp detection system (CADe) (left) and the additional CADe preventing instrument detection system (right). Video S1: Head-to-head comparison of a colonoscopy video sequence with (right) and without (left) the use of the instrument detection convolutional neuronal network

The first one is that the snare produces folds around the polyp by pressing the wire against the mucosa. Those folds are then falsely interpreted as polyps by the commercially available CADe system. Another explanation is that the total time that snares are visible in our dataset is longer than that of graspers. The longer visibility results in more CADe detections. Lastly, the not openly available training dataset of the commercially available CADe system might contain images of graspers and snares in an unbalanced manner.

The implementation of our freely available instrument detection AI system in an existing CADe system could be done by controlling the input signal of the examination monitor. Here, the instrument detection AI would analyze the raw endoscopy processor output signal in parallel with the CADe system. During withdrawal with no visible instrument, the CADe signal would be displayed to the examiner. Thus, allowing the examiner to fully benefit from the high polyp detection rate of a commercially available CADe system. When an instrument is detected by the new AI, an automatic switch would display the raw processor signal instead. Alternatively, the AI system for instrument recognition can also be integrated as a filter directly into an existing or newly developed CADe system. By integrating our system in a single CADe system, the process would be more comfortable.

To the best of our knowledge, an AI system for instrument detection using deep learning methods has not been developed in gastrointestinal endoscopy. Therefore, we present the first AI system in the field that enables recognition of endoscopic interventions by instrument detection. In addition to the mentioned application, the system could also be useful for automated recording of intervention times or withdrawal time. This could potentially help in obtaining objective data to assess the quality of colonoscopies.[16]

However, our study has several limitations. The new AI system was evaluated using previously recorded videos. Therefore, the mentioned implementation of the AI system in daily practice must be tested in future prospective studies to evaluate clinical benefit. Another possibility would be to investigate (e.g. by eye-tracking) whether the examiner's attention could be better focused on the intervention by reducing distracting CADe signals.[17] Other limitations include that, to facilitate the generation of a large-annotated training dataset in a short period of time, no predefined protocol was used for video selection. A quantitative identification of the causes of false positive detections of our CNN need to be evaluated in future studies.

CADe systems already achieved a remarkable benefit in randomized controlled trials. Future developments of those systems include improving usability by adding customizable features. The commercially available system that was used in our study for example, presents a well-studied CADe function.[5] Other systems incorporate computer-aided diagnosis (CADx) that is only turned on when virtual chromoendoscopy is activated by the examiner.[18] In the case of the ENDO-AID CADe system by Olympus the examiner has the possibility to choose how many CADe detection boxes should be maximally displayed on the screen. The examiner can even choose from two different CADe modes that presumably present different sensitivities.[19] In other words, in general, these devices are incorporating customizable modules that increase the usability and, therefore, the value of the device. Our work aligns in this direction by contributing to the prevention of distracting CADe activations during interventions.

In conclusion, our study shows that instrument detection using AI technology is reliable and achieves high sensitivity and specificity. Therefore, the new AI system could be helpful to reduce distracting CADe detections during endoscopic procedures. Although the clinical benefit of the new AI system needs further evaluation, our study demonstrates the great potential of AI technology beyond mucosal assessment.

ACKNOWLEDGMENTS

Alexander Hann and Wolfram G. Zoller receive public funding for this work from the state government of Baden-Württemberg, Germany (Funding cluster Forum Gesundheitsstandort Baden-Württemberg) to research and develop artificial intelligence applications for polyp detection in screening colonoscopy. Wolfram G. Zoller receives additional funding to support this work by the Eva Mayr-Stihl Foundation, Waiblingen, Germany, the Fischerwerke GmbH & Co. KG, Waldachtal, Germany and the Dieter von Holtzbrinck Stiftung GmbH, Stuttgart, Germany.

Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

The authors have no conflicts of interest to declare.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Markus Brand https://orcid.org/0000-0002-3495-5206
Joel Troya https://orcid.org/0000-0002-7992-0146
Adrian Krenzer https://orcid.org/0000-0002-1593-3300
Alexander Hann https://orcid.org/0000-0001-8035-3559

REFERENCES

1. Wang P, Liu X, Berzin TM, Glissen Brown JR, Liu P, Zhou C, et al. Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): a double-blind randomised study. Lancet Gastroenterol & Hepatol. 2020;5(4):343–51. https://doi.org/10.1016/s2468-1253(19)30411-x
2. Wang P, Liu P, Glissen Brown JR, Berzin TM, Zhou G, Lei S, et al. Lower adenoma miss rate of computer-aided detection-assisted colonoscopy vs routine white-light colonoscopy in a prospective tandem study. Gastroenterology. 2020;159(4):1252–61:e1255. https://doi.org/10.1053/j.gastro.2020.06.023
3. Wang P, Berzin TM, Glissen Brown JR, Bharadwaj S, Becq A, Xiao X, et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. Gut. 2019;68(10):1813–19. https://doi.org/10.1136/gutjnl-2018-317500
4. Su JR, Li Z, Shao XJ, Ji CR, Ji R, Zhou RC, et al. Impact of a real-time automatic quality control system on colorectal polyp and adenoma detection: a prospective randomized controlled study (with videos). Gastrointest Endosc. 2020;91(2):415–24:e414. https://doi.org/10.1016/j.gie.2019.08.026
5. Repici A, Badalamenti M, Maselli R, Correale L, Radaelli F, Rondonotti E, et al. Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial. Gastroenterology. 2020;159(2):512–20:e517. https://doi.org/10.1053/j.gastro.2020.04.062
6. Hassan C, Spadaccini M, Iannone A, Maselli R, Jovani M, Chandrasekar VT, et al. Performance of artificial intelligence in colonoscopy for adenoma and polyp detection: a systematic review and meta-analysis. Gastrointest Endosc. 2021;93(1):77–85:e76. https://doi.org/10.1016/j.gie.2020.06.059
7. Hassan C, Badalamenti M, Maselli R, Correale L, Iannone A, Radaelli F, et al. Computer-aided detection-assisted colonoscopy: classification and relevance of false positives. Gastrointest Endosc. 2020;92(4):900–4:e904. https://doi.org/10.1016/j.gie.2020.06.021

8. Holzwanger EA, Bilal M, Glissen Brown JR, Singh S, Becq A, Ernest-Suarez K, et al. Benchmarking definitions of false-positive alerts during computer-aided polyp detection in colonoscopy. Endoscopy. 2021;53(09):937–40. https://doi.org/10.1055/a-1302-2942
9. Hann A, Troya J, Fitting D. Current status and limitations of artificial intelligence in colonoscopy. United European Gastroenterol J. 2021;9(5):527–33. https://doi.org/10.1002/ueg2.12108
10. Ebigbo A, Mendel R, Probst A, Manzeneder J, Prinz F, de Souza LA, Jr., et al. Real-time use of artificial intelligence in the evaluation of cancer in Barrett's oesophagus. Gut. 2020;69(4):615–6. https://doi.org/10.1136/gutjnl-2019-319460
11. Ueyama H, Kato Y, Akazawa Y, Yatagai N, Komori H, Takeda T, et al. Application of artificial intelligence using a convolutional neural network for diagnosis of early gastric cancer based on magnifying endoscopy with narrow-band imaging. J Gastroenterol Hepatol. 2021;36(2):482–9. https://doi.org/10.1111/jgh.15190
12. Misawa M, Kudo SE, Mori Y, Hotta K, Ohtsuka K, Matsuda T, et al. Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). Gastrointest Endosc. 2021;93(4):960–7:e963. https://doi.org/10.1016/j.gie.2020.07.060
13. Wang P, Xiao X, Glissen Brown JR, Berzin TM, Tu M, Xiong F, et al. Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. Nat Biomed Eng. 2018;2(10):741–8. https://doi.org/10.1038/s41551-018-0301-3
14. Urban G, Tripathi P, Alkayali T, Mittal M, Jalali F, Karnes W, et al. Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. Gastroenterology. 2018;155(4):1069–78:e1068. https://doi.org/10.1053/j.gastro.2018.06.037
15. Lu Z, Xu Y, Yao L, Zhou W, Gong W, Yang G, et al. Real-time automated diagnosis of colorectal cancer invasion depth using a deep learning model with multimodal data (with video). Gastrointest Endosc. 2021. https://doi.org/10.1016/j.gie.2021.11.049
16. Kaminski MF, Thomas-Gibson S, Bugajski M, Bretthauer M, Rees CJ, Dekker E, et al. Performance measures for lower gastrointestinal endoscopy: a European Society of Gastrointestinal Endoscopy (ESGE) quality improvement initiative. United European Gastroenterol J. 2017;5(3):309–34. https://doi.org/10.1177/2050640617700014
17. Troya J, Fitting D, Brand M, Sudarevic B, Kather JN, Meining A, et al. The influence of computer-aided polyp detection systems on reaction time for polyp detection and eye gaze. Endoscopy. 2022. https://doi.org/10.1055/a-1770-7353
18. Weigt J, Repici A, Antonelli G, Afifi A, Kliegis L, Correale L, et al. Performance of a new integrated computer-assisted system (CADe/CADx) for detection and characterization of colorectal neoplasia. Endoscopy. 2022;54(02):180–4. https://doi.org/10.1055/a-1372-0419
19. Koo CS, Dolgunov D, Koh CJ. Key tips for using computer-aided diagnosis in colonoscopy - observations from two different platforms. Endoscopy. 2021. https://doi.org/10.1055/a-1701-6201

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Brand M, Troya J, Krenzer A, Saßmannshausen Z, Zoller WG, Meining A, et al. Development and evaluation of a deep learning model to improve the usability of polyp detection systems during interventions. United European Gastroenterol J. 2022;10(5):477–84. https://doi.org/10.1002/ueg2.12235

Digestion

Research Article

Digestion
DOI: 10.1159/000525345

Received: March 8, 2022
Accepted: June 1, 2022
Published online: June 29, 2022

Frame-by-Frame Analysis of a Commercially Available Artificial Intelligence Polyp Detection System in Full-Length Colonoscopies

Markus Brand[a] Joel Troya[a] Adrian Krenzer[a, b] Costanza De Maria[c, d]
Niklas Mehlhase[e] Sebastian Götze[e] Benjamin Walter[e] Alexander Meining[a]
Alexander Hann[a]

[a]Interventional and Experimental Endoscopy (InExEn), Department of Internal Medicine II, University Hospital Würzburg, Würzburg, Germany; [b]Artificial Intelligence and Knowledge Systems, Institute for Computer Science, Julius-Maximilians-Universität, Würzburg, Germany; [c]Department of Gastroenterology and Hepatology, Ente Ospedaliero Cantonale (EOC), Bellinzona, Switzerland; [d]Department of Biomedical Science, University of Italian Switzerland (USI), Lugano, Switzerland; [e]Department of Internal Medicine I, University Hospital Ulm, Ulm, Germany

Keywords
Colonoscopy · Deep learning · Computer-aided detection · Artificial intelligence

Abstract

Introduction: Computer-aided detection (CADe) helps increase colonoscopic polyp detection. However, little is known about other performance metrics like the number and duration of false-positive (FP) activations or how stable the detection of a polyp is. ***Methods:*** 111 colonoscopy videos with total 1,793,371 frames were analyzed on a frame-by-frame basis using a commercially available CADe system (GI-Genius, Medtronic Inc.). Primary endpoint was the number and duration of FP activations per colonoscopy. Additionally, we analyzed other CADe performance parameters, including per-polyp sensitivity, per-frame sensitivity, and first detection time of a polyp. We additionally investigated whether a threshold for withholding CADe activations can be set to suppress short FP activations and how this threshold alters the CADe performance parameters. ***Results:*** A mean of 101 ± 88 FPs per colonoscopy were found. Most of the FPs consisted of less than three frames with a maximal 66-ms duration. The CADe system detected all 118 polyps and achieved a mean per-frame sensitivity of 46.6 ± 26.6%, with the lowest value for flat polyps (37.6 ± 24.8%). Withholding CADe detections up to 6 frames length would reduce the number of FPs by 87.97% ($p < 0.001$) without a significant impact on CADe performance metrics. ***Conclusions:*** The CADe system works reliable but generates many FPs as a side effect. Since most FPs are very short, withholding short-term CADe activations could substantially reduce the number of FPs without impact on other performance metrics. Clinical practice would benefit from the implementation of customizable CADe thresholds.

© 2022 The Author(s).
Published by S. Karger AG, Basel

Introduction

Artificial intelligence (AI) is presumably a powerful tool in colorectal cancer prevention using colonoscopy, as several randomized controlled trials (RCTs) have shown that computer-aided detection (CADe) increases

Markus Brand and Joel Troya contributed equally to this work.

Karger@karger.com
www.karger.com/dig

© 2022 The Author(s).
Published by S. Karger AG, Basel

This article is licensed under the Creative Commons Attribution 4.0 International License (CC BY) (http://www.karger.com/Services/OpenAccessLicense). Usage, derivative works and distribution are permitted provided that proper credit is given to the author and the original publisher.

OPEN ACCESS

Correspondence to:
Alexander Hann, hann_a@ukw.de

adenoma detection rate (ADR) and decreases the miss rate of neoplastic lesions [1–6].

However, there are still many unanswered questions regarding CADe systems. For example, many false-positive (FP) activations of up to 8% of all frames occur during examination with CADe systems [7]. The number and duration of FP activations play an important role regarding the examiners comfort in using those systems, as these activations can affect the examiners attention leading to misinterpretation of normal mucosa [8]. Therefore, an international consensus conference has identified the analysis of FP activations as an important research focus [9]. Current studies on this topic include only small numbers of cases with about 40 colonoscopy examinations and mainly investigate the cause and the clinical impact of FP activations [10, 11]. However, specific data on the duration and pattern of FP activations are not available, although such information is necessary to better understand the operation of CADe systems in order to improve them. An example for improvement might be the reduction of FPs through customizable activation thresholds. In addition, previous RCTs only provide data on per-polyp sensitivity (PPS), i.e., whether a polyp was detected resulting in a yes or no answer. How stable the detection signal is over time, termed per-frame sensitivity, was not assessed as no frame-by-frame analysis of real full-length videos has been performed so far.

Therefore, the objective of this study was to analyze the FP pattern of a commercial CADe system. This was done using a frame-by-frame analysis of full-length real-life videos to determine the effects of different CADe activation thresholds on FPs. Additionally, in a patient-based analysis, we examined performance parameters such as PPS or the mean number of polyps per colonoscopy (PPC).

Materials and Methods

Study Design

Videos from 244 routine colonoscopies performed in two tertiary centers (University Hospital Ulm and Würzburg) were retrospectively analyzed. Recording took place between March 2019 and April 2020. Those colonoscopies (raw signals) were recorded using the high-definition video signal of the endoscopy processor (Olympus CV-190). For the performance analysis of a commercially available CADe system (GI Genius, Medtronic Inc., Ireland, software version of March 2020), this raw video signal was introduced into the AI system, and the output signal (with visible CADe detections) was recorded. Accordingly, a video pair consisting of raw signal and CADe signal was assembled for video analysis of each colonoscopy.

Colonoscopies

Colonoscopies were performed using the colonoscopes CF-HQ190AL and CF H180AI/AL (Olympus Co., Tokyo, Japan). All patients were prepared for the colonoscopy using a standard split-dose regimen with 2L polyethylene glycol with ascorbic acid (Moviprep, Norgine Pharma; Harefield, England). Endoscopies were performed using nurse-assisted propofol sedation [12]. Polyps were removed upon detection by cold or hot snare technique if no contraindication for resection was present. The examiners were classified due to their experience in colonoscopy between junior and senior with 2,000 performed colonoscopies as a threshold.

Video Analysis

All videos were screened by a board-certified gastroenterologist and experienced endoscopist (MB) with over 4,000 performed colonoscopies. Examinations performed for screening reasons or post polypectomy surveillance were included in the analysis. For further analysis, the following exclusion criteria were defined: inflammatory bowel disease, active gastrointestinal bleeding, poor bowel preparation defined by a Boston Bowel Preparation Scale (BBPS) lower than 5, incomplete colonoscopies, advanced neoplasia, altered gut anatomy, endoscopy only performed for an extended resection and polyposis syndrome. Included colonoscopies were analyzed in a deep frame-by-frame manner using a custom-made annotation tool as previously described [13].

Analysis of Non-CADe Signal (Raw Videos)

The start and the end of withdrawal and polypectomies were annotated. Each polyp was counted for the analysis. Additionally, polyps were characterized using the Paris classification and size (<5 mm, 5–10 mm, 11–20 mm, >20 mm). In a frame-by-frame analysis, each frame with a partially or completely visible polyp was annotated as a polyp frame. Frames with even small parts of a polyp visible were regarded as a polyp-containing frame. Polyp annotation stopped at the beginning of the resection (first frame with a visible instrument in the image).

Analysis of CADe Signal (AI Videos)

All frames with visible bounding boxes resembling CADe detections were automatically identified by a custom-made application. Subsequently, each bounding box was classified by an experienced endoscopist (MB) as a true-positive (TP) or FP detection. It was considered TP if the bounding box had contact with the visible polyp, irrespective of how much area of the lesion was covered. Small hyperplastic polyps of the rectosigmoid were excluded from the analysis. The absence of a bounding box in a frame with a visible polyp was regarded a false negative. The absence of a box in a frame without a polyp was considered a true negative. A FP detection was defined as a detected area that was not in contact with a polyp. In case of a FP detection in a frame with a visible polyp, the term distraction was used.

Endpoints

The primary endpoint of the study was the number of FP activations per colonoscopy and the duration of FP activations. For the secondary endpoints, we analyzed further CADe performance parameters, including mean number of PPC of the CADe System, PPS, per-frame sensitivity, and first detection time (FDT) of a polyp.

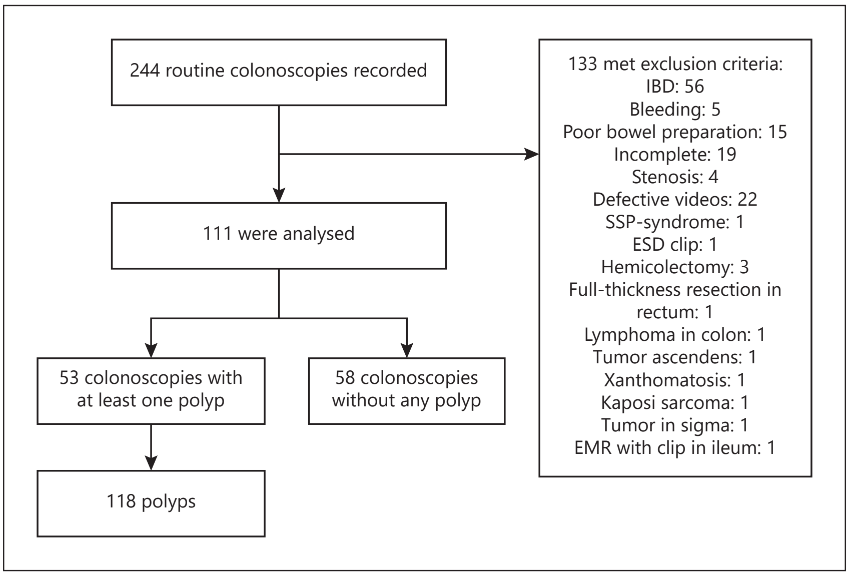

244 routine colonoscopies recorded
133 met exclusion criteria:
IBD: 56
Bleeding: 5
Poor bowel preparation: 15
Incomplete: 19
Stenosis: 4
Defective videos: 22
SSP-syndrome: 1
ESD clip: 1
Hemicolectomy: 3
Full-thickness resection in rectum: 1
Lymphoma in colon: 1
Tumor ascendens: 1
Xanthomatosis: 1
Kaposi sarcoma: 1
Tumor in sigma: 1
EMR with clip in ileum: 1
111 were analysed
53 colonoscopies with at least one polyp
58 colonoscopies without any polyp
118 polyps

Fig. 1. Flowchart of study design. EMR, endoscopic mucosal resection; ESD, endoscopic submucosal dissection; IBD, inflammatory bowel disease; SSP sessile serrated polyposis.

In addition, we investigated whether a threshold for withholding short CADe activations can be set to suppress FP activations and how this threshold alters CADe performance parameters such as PPC, PPS, or per-frame sensitivity.

Data Analysis and Statistics

FP activations were counted in their number, with each contiguous sequence of FP frames counted as one activation. In addition, the duration of FP activations was measured in frames. Each frame had a duration of 33 ms. The mean number of PPC was calculated by dividing the number of detected polyps by the number of performed colonoscopies. PPS was defined as the number of polyps detected by the CADe system in at least one frame divided by the number of polyps annotated in the raw video data. The per-frame sensitivity, previously published as temporal coherence, was calculated by dividing the number of TP frames by the total number of frames where the polyp was visible in the raw signal (TP + false negative), as previously described by Zhou et al. [14]. Additionally, the per-lesion sensitivity, defined as the number of polyps in which more than half of each polyp's frame were detected by the CADe, divided by the total number of polyps, was analyzed as previously described by Misawa et al. [15]. FDT of a polyp was defined as the time interval between the first appearance of a polyp in the raw video and the first frame containing a TP-CADe activation. If the polyp was not permanently visible during this time span, frames without a visible polyp were excluded. By this method, FDT included only frames with a visible polyp. The mean withdrawal time was determined using the recorded videos and defined as the time frame between the coecum and anal canal, excluding time spent for performing biopsies or snare resection [16].

Statistical analysis was performed using Python version 3.8. The χ^2 and Fisher's exact tests were used to test for significant differences between categorical variables. Student's *t* test and Mann-Whitney U test were applied for continuous variables depending on their distribution pattern. A *p* value of <0.05 indicated statistical significance.

Table 1. Patient and polyps characteristics

Characteristic	Value
Sex	
Male, *n* (%)	50 (45.05)
Female, *n* (%)	61 (54.95)
Age, mean (range)	60.46 (19–89)
BBPS, mean (range)	7.50 (6–9)
Withdrawal time, mean, minutes (IQR)	8.98 (5.33–22.04)
Polyps, *n*	118
Polyps per patient, mean (range)	1.06 (0–6)
Paris classification, *n* (%)	
0-Ip	6 (5.08)
0-Is	42 (35.59)
0-IIa	70 (59.32)
Size, *n* (%)	
1–5 mm	65 (55.08)
6–10 mm	30 (25.42)
11–20 mm	23 (19.49)
Location, *n* (%)	
Right colon	66 (55.93)
Left colon	35 (29.66)
Rectum	17 (14.41)

BBPS, Boston Bowel Preparation Scale; IQR, interquartile range.

Results

Baseline Characteristics

From 244 routine colonoscopies, 133 colonoscopies met the exclusion criteria. Thus, a total of 111 pairs of colonoscopy videos including the raw video signal and the CADe signal were analyzed (Fig. 1) in a deep frame-

Table 2. Per-frame sensitivity as a measure of time percentage in which polyps were correctly detected by the CADe system

Characteristic	Value
Per-frame sensitivity, mean% ± SD (*n*)	47.73±26.50 (118)
Polyps with less than 50% per-frame sensitivity, *n* (%)	56 (47.46)
Paris classification	
0-Ip, mean% ± SD (*n*)	74.86±21.78 (6)
0-Is, mean% ± SD (*n*)	60.10±22.31 (42)
0-IIa, mean% ± SD (*n*)	37.99±24.65 (70)
Size	
1–5 mm, mean% ± SD (*n*)	49.10±25.96 (65)
6–10 mm, mean% ± SD (*n*)	50.91±27.74 (30)
11–20 mm, mean% ± SD (*n*)	39.72±25.98 (23)
Location	
Right colon, mean% ± SD (*n*)	43.56±25.04 (66)
Left colon, mean% ± SD (*n*)	53.43±26.00 (35)
Rectum, mean% ± SD (*n*)	52.21±31.44 (17)

SD, standard deviation.

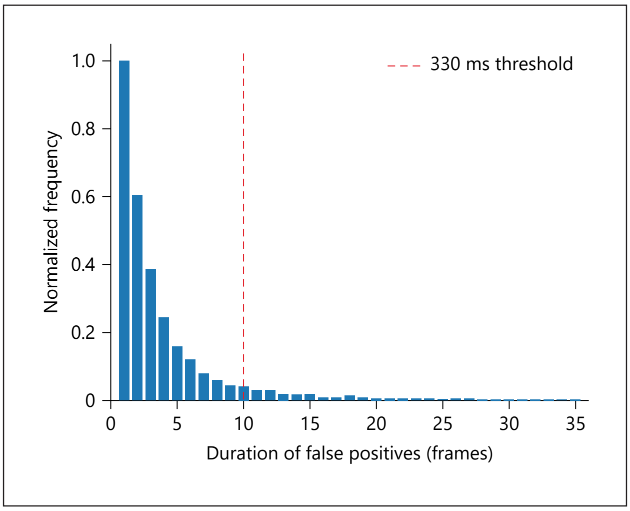


Fig. 2. Histogram displaying the different length of FP CADe activation durations measured in consecutive frames. The bars to left of the dotted red line represent more than 90% of all activations. CADe, computer-aided detection.

by-frame manner. Most of the examinations (65.8%) were done by experienced investigators with over 2,000 performed colonoscopies. The mean BBPS score was 7.5, with the lowest value being 6. The mean withdrawal time was 8:58 min. A total of 118 polyps were identified and annotated in the 111 videos. Most of the polyps were diminutive with 1–5 mm in size (55.08%) with flat or sessile shape (Paris 0-Is/IIa; 35.59%/59.32%). Baseline characteristics of the colonoscopies and detailed characterization of the polyps are shown in Table 1. In total, the 111 examinations analyzed contained 1,793,371 frames, including 173,959 frames (9.7%) with polyps and 1,619,412 frames (90.3%) without polyps. Three polyps were detected by the CADe but not perceived by the endoscopist.

Primary Endpoint

Rate of FP Detections and Distracting Detections

A total of 11,188 FP activations were detected in the 111 coloscopies (101 ± 88 FPs per colonoscopy). The mean duration of a FP activation was 135 ms. In relation to the withdrawal time, the FPs account for a mean of 2.48% resembling 13.61 s. Most of the FP detections consisted of one to two frames, corresponding to a period of max. 66 ms (Fig. 2). Only a minority of detections accounted for continuous detections consisting of 10 frames or more, resembling more than 330 ms. In the subgroup of colonoscopies with at least one polyp, we examined the frames with a FP CADe detection in an image with a visible polyp, termed distracting detection. Here we found that 1.6 ± 2.1% of the frames with polyps contain this distraction.

Secondary Endpoints

PPC and PPS

The CADe system detected all 118 polyps that were visible in the videos, resulting in a PPS of 100%. The mean number of PPC was 1.06.

Per-Frame Sensitivity and Per-Lesion Sensitivity

The mean per-frame sensitivity of the CADe system for all 118 polyps was 47.73 ± 26.5% (Table 2). The mean per-lesion sensitivity of the CADe system was 47.46%. In

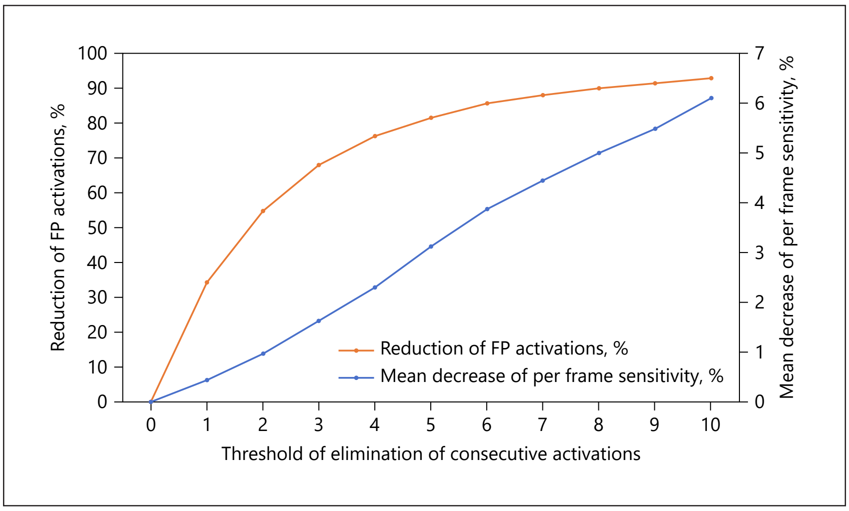


Fig. 3. Effect of elimination of short-lasting CADe activations on per-frame sensitivity (per frame sensitivity, blue line) and FP activations (red line). As shown, the progressive elimination of activations with increasing duration has a higher impact on reducing FP activations than on per-frame sensitivity reduction. CADe, computer-aided detection; FP, false positive.

a subgroup analysis, we found that the per-frame sensitivity was significantly lower in flat polyps (Paris 0-IIa) compared to 0-Ip or 0-Is configuration (37.99 ± 24.65% vs. 74.86 ± 21.78%, $p < 0.001$ or 37.99 ± 24.65% vs. 60.10 ± 22.31%, $p < 0.001$). While polyp size did not influence the per-frame sensitivity, polyp localization in the right-sided colon segments was associated with lower mean per-frame sensitivity, compared to the left-sided colon segments (43.56 ± 25.04% vs. 53.43 ± 26.00, $p = 0.017$).

FDT of a Polyp

FDT was available for each of the 118 polyps. The mean FDT was 1,692 ± 2,052 ms with a wide range from 33.3 to 12,033 ms. In a subgroup analysis, we found the highest FDT in the polyp size group 11–20 mm with mean 2,179 ± 3,174 ms (Table 3). However, this was not significant when compared to size groups 1–5 mm and 6–10 mm. In contrast, we found a significantly higher FDT in Paris 0-IIa polyps in comparison to 0-Ip or 0-Is polyps (2,068 ± 2,413 ms vs. 522 ± 216 ms or 1,233 ± 1,247 ms, $p = 0.023$ and $p = 0.046$).

Impact of Different CADe Activation Thresholds on FPs and CADe Performance Parameters

To estimate the effect of withholding CADe activations of a defined frame length on the FPs and the CADe performance, a subgroup analysis was performed using only activations of a defined frame length or longer. Figure 3 shows graphically how withholding short activations of 1–10 frames significantly reduces the rate of FPs while having little effect on the per-frame sensitivity of the CADe system. For example, withholding activations up to a length of 10 frames representing 330 ms reduced FP activations by 92.79% ($p < 0.001$), while the per-frame sensitivity decreased by only 6.07% ($p = 0.07$). In addition, we examined whether withholding short activations influenced PPC or PPS. Up to a threshold of 3 frames (100 ms), no polyps were missed. In contrast, a threshold of 10 frames representing 330 ms resulted in 7 missed polyps. In this case, all missed polyps were of flat shape (Paris 0-IIa) and had previously low per-frame sensitivity values of <28%. PPC was not significantly affected by withholding CADe activations up to a threshold of 10 frames ($p = 0.71$), whereas initial significant changes in PPS occurred at a threshold of 7 frames ($p = 0.02$). Detailed information

Table 3. Time to first detection of a visible polyp by the CADe system

Characteristic	Mean, ms ± SD (*n*)
Time to first detection	1,692.66±2,052.73 (118)
Paris classification	
0-Ip	522.22±216.71 (6)
0-Is	1,233.33±1,246.96 (42)
0-IIa	2,068.57±2,413.86 (70)
Size	
1–5 mm	1,621.03±1,792.53 (65)
6–10 mm	1,474.44±1,420.34 (30)
11–20 mm	2,179.71±3,174.02 (23)

SD, standard deviation; ms, milliseconds; CADe, computer-aided detection.

Table 4. Information on the impact of different CADe detection thresholds on the mean number of PPC, PPS, and per-frame sensitivity

Threshold, frames	FP reduction % (*p* value)	PPC, value (*p* value)	PPS, value (*p* value)	Per-frame sensitivity, mean ± SD (*p* value)
None	0%	1.06	1	47.73±26.50
1	34.25 (<0.001)	1.06 (1.00)	1 (1.00)	47.29±26.54 (0.82)
2	54.78 (<0.001)	1.05 (0.99)	0.99 (1.00)	46.76±26.63 (0.70)
3	67.96 (<0.001)	1.04 (0.91)	0.97 (0.25)	46.10±26.72 (0.59)
4	76.16 (<0.001)	1.03 (0.83)	0.96 (0.13)	45.43±26.76 (0.45)
5	81.46 (<0.001)	1.02 (0.83)	0.96 (0.13)	44.61±26.97 (0.34)
6	85.54 (<0.001)	1.02 (0.83)	0.96 (0.07)	43.86±27.06 (0.24)
7	87.92 (<0.001)	1 (0.75)	0.94 (0.02)	43.29±27.21 (0.19)
8	89.97 (<0.001)	1 (0.75)	0.94 (0.02)	42.73±27.23 (0.14)
9	91.42 (<0.001)	1 (0.75)	0.94 (0.02)	42.25±27.2 (0.10)
10	92.79 (<0.001)	0.99 (0.71)	0.93 (0.01)	41.63±27.4 (0.07)

The threshold value indicates that activations of the value or shorter have been eliminated. SD, standard deviation; PPS, per-polyp sensitivity; PPC, polyps per colonoscopy. *p* values represent comparisons to the CADe signal without a threshold.

on the impact of different thresholds on PPC, PPS, and per-frame sensitivity are shown in Table 4. In addition, online supplementary Video 1 (for all online suppl. material, see www.karger.com/doi/10.1159/000525345) shows an example of how a threshold of 6 frames (no significant changes in PPC, PPS, or per-frame sensitivity) affects FP activations in the endoscopic view.

Discussion

The development of an AI system for polyp detection using deep learning techniques applied on a larger dataset was first described by Wang et al. [17]. Subsequently, several commercially available CADe systems have been developed for colonoscopy. In prospective RCTs, CADe systems showed a significantly higher ADR compared to expert colonoscopists [1–4, 7, 18–21]. Moreover, a recently published meta-analysis found a significant increase of ADR [6]. While prospective studies have extensively evaluated the ADR of various CADe systems, little is known about the detailed performance of CADe systems, e.g., FP rate, FP duration, or per-frame sensitivity, especially in a real-life scenario. Only a few studies about CADe systems include a single-frame analysis. However, these studies used single polyp frames, short video sequences, or videos consisting of less than 160,000 frames [14, 15, 17]. Thus, we present the largest frame-by-frame dataset, to our knowledge, with 111 full-length videos consisting of over 170,000 polyp frames and a total of over 1,700,000 frames. Additionally, to the best of our knowledge, our study is the first evaluating CADe performance in a frame-by-frame analysis in real-life videos.

The PPS of 100% highlights the effectiveness of CADe systems in clinical practice; however, the number of FP activations is not negligible and is higher than the previously published values [10, 11]. While previous studies analyzed the cause and clinical relevance of FPs, the use of frame-by-frame analysis allowed us to determine the exact duration and distribution pattern of FPs. As shown, most FPs were shorter than 330 ms, hence they are perceived by the endoscopist only as a brief flashing of the bounding box. However, it is not yet clear whether the short activations do or do not affect the normal mucosa visualization pattern of endoscopists. Some retrospective studies suggest that FP activations result in the negligible increase of the total withdrawal time, as most of them are immediately discarded by the endoscopists [10, 11]. Other studies using for example eye-tracking glasses suggest that CADe and FPs activations might have an impact on the visualization pattern of the endoscopists [8, 22]. Therefore, further studies using eye tracking technology during endoscopic examinations in a prospective manner should be performed in order to analyze the influence of short FP activations on the examiner and the withdrawal time. Nevertheless, many short FPs may impair the endoscopist's concentration in the long run; certainly, they reduce the comfort of the CADe application.

An option to reduce the FP rate, especially for short FP, could be withholding of short CADe activations. As shown, withholding short detections up to 10 frames length reduced the number of FP by up to 92.79% without having a significant effect on per-frame sensitivity. However, above a threshold of 3 frames representing 100 ms, this is at the expense of a few missed polyps, especially those with a flat shape. Another effect to consider should be the impact that the withholding of short CADe activations could have on FDT. Unfortunately, there are no studies that demonstrate the effect of different FDTs on the detection of polyps. However, considering the big effect in the reduction of FP activations and since there was no significant change in PPC or PPS up to a threshold of 6 frames (200 ms), an appropriate threshold for optimization of the CADe system could be in this range.

Besides PPC, PPS, and FP rate, per-frame sensitivity is another important performance parameter of CADe systems, particularly since the temporal stability of polyp detection indicates how well CADe detection works for different polyp types. The per-frame sensitivity determined in our study is lower than in previous publications [14, 15, 17]. However, in previous studies, only several single images of polyps or selected video sequences were used to evaluate the self-developed systems. For example, the study by Misawa et al. [15] analyzed video clips with a total of 152,560 frames. In our study, full-length real-life videos containing 1,793,371 frames were used, so the conditions for CADe detection may have been more challenging, yet more realistic. Another important reason is that small hyperplastic polyps in the rectosigmoid were excluded in our study due to clinical irrelevance, whereas these polyps, which can often be reliably identified, were included in the evaluation in previous studies.

Since flat polyps (Paris 0-IIa) and sessile serrated adenomas have higher miss rates, the effect of CADe systems on polyp detection could be substantial if these lesions were reliably detected [23]. However, our data show that in clinical practice, per-frame sensitivity and FDT tend to be worse in these polyps. These findings are consistent with previous data, reporting lower per-frame sensitivity for laterally spreading tumors and sessile serrated adenoma, showing that there is an urgent need for improvement in this point [14].

There are several limitations to our study. Since this is a retrospective analysis of previously stored videos, histologic differentiation of colonic polyps was not possible. In order to increase the relevance of the detected polyps, we excluded hyperplastic polyps in the rectosigmoid. Due to the exclusion of examinations with a BBPS score of <6 points, the mean BBPS score is 7.5 points, which is relatively high [24]. However, recently published papers on CADe performance metrics describe similarly high BBPS values [10, 11]. To shorten the time-consuming deep frame analysis, we have dispensed with a detailed analysis of the FPs with respect to their cause. However, Hassan et al. [10] performed such an analysis using the same CADe system – they found bubbles, stool, and colonic folds to be the main reasons for FP activation. We also did not manually annotate each polyp-containing frame with bounding boxes. Thus, subsequent analysis of, for example, intersection over the union of the CADe boxes with ground truth was not performed.

Conclusion

This commercially available CADe system is a powerful tool to facilitate polyp detection even under daily clinical conditions, but at the expense of many FP activations. Through a frame-by-frame video analysis, we were able to show that many of these FPs are of very short duration. Withholding short-term CADe detections could substantially reduce the number of FP activations, but at higher thresholds at the expense of a few missed polyps. This applies in particular to flat polyps, which generally have poorer per-frame sensitivity values. Since we could not detect any significant change in the mean number of PPC and PPS up to a threshold of 6 frames, an appropriate threshold for optimization of the CADe system could be in this range. Nevertheless, further detailed analysis of CADe systems is needed to better understand the strengths and weaknesses of this promising technology and to further optimize the systems. A customizable CADe detection threshold that can be adjusted to the needs of the examiner would be useful in clinical practice.

Statement of Ethics

This study protocol involving retrospective analysis of data was reviewed and approved by the Ethics Committee of the University Hospital Würzburg, approval number 2021032901. According to the Ethics Committee of the University Hospital Würzburg, patients were not required to give informed consent for this retrospective analysis.

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Funding Sources

Alexander Hann receives public funding from the state government of Baden-Württemberg, Germany (funding cluster “Forum Gesundheitsstandort Baden-Württemberg”), to research and develop artificial intelligence applications for polyp detection in screening colonoscopy. Alexander Meining receives funding from the IZKF Würzburg and the Bavarian Center for Cancer Research (BZKF) for further implementation and development of artificial intelligence for detection of (pre)neoplastic lesions.

Author Contributions

Markus Brand, Joel Troya, Alexander Meining, and Alexander Hann: study concept and design, interpretation of results, and drafting of the manuscript. Joel Troya: statistical analysis. Markus Brand: annotation of videos. Joel Troya, Adrian Krenzer, Costanza De Maria, Niklas Mehlhase, Sebastian Götze, and Benjamin Walter: acquisition of data. Markus Brand, Joel Troya, Adrian Krenzer, Costanza De Maria, Niklas Mehlhase, Sebastian Götze, Benjamin Walter, Alexander Meining, and Alexander Hann: critical revision of the article for important intellectual content and final approval of the article.

Data Availability Statement

The data underlying this article will be shared on reasonable request to the corresponding author.

References

1 Wang P, Berzin TM, Glissen Brown JR, Bharadwaj S, Becq A, Xiao X, et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. Gut. 2019 Oct;68(10):1813–9.
2 Liu WN, Zhang YY, Bian XQ, Wang LJ, Yang Q, Zhang XD, et al. Study on detection rate of polyps and adenomas in artificial-intelligence-aided colonoscopy. Saudi J Gastroenterol. 2020 Jan–Feb;26(1):13–9.
3 Repici A, Badalamenti M, Maselli R, Correale L, Radaelli F, Rondonotti E, et al. Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial. Gastroenterology. 2020 Aug;159(2):512–20 e7.
4 Su JR, Li Z, Shao XJ, Ji CR, Ji R, Zhou RC, et al. Impact of a real-time automatic quality control system on colorectal polyp and adenoma detection: a prospective randomized controlled study (with videos). Gastrointest Endosc. 2020 Feb;91(2):415–24 e4.
5 Deliwala SS, Hamid K, Barbarawi M, Lakshman H, Zayed Y, Kandel P, et al. Artificial intelligence (AI) real-time detection vs. routine colonoscopy for colorectal neoplasia: a meta-analysis and trial sequential analysis. Int J Colorectal Dis. 2021 Nov;36(11):2291–303.
6 Hassan C, Spadaccini M, Iannone A, Maselli R, Jovani M, Chandrasekar VT, et al. Performance of artificial intelligence in colonoscopy for adenoma and polyp detection: a systematic review and meta-analysis. Gastrointest Endosc. 2021 Jan;93(1):77–85.e6.
7 Pfeifer L, Neufert C, Leppkes M, Waldner MJ, Hafner M, Beyer A, et al. Computer-aided detection of colorectal polyps using a newly generated deep convolutional neural network: from development to first clinical experience. Eur J Gastroenterol Hepatol. 2021 Dec 1;33:e662–9.
8 Troya J, Fitting D, Brand M, Sudarevic B, Kather JN, Meining A, et al. The influence of computer-aided polyp detection systems on reaction time for polyp detection and eye gaze. Endoscopy. 2022 Feb 14. Epub ahead of print.
9 Ahmad OF, Mori Y, Misawa M, Kudo SE, Anderson JT, Bernal J, et al. Establishing key research questions for the implementation of artificial intelligence in colonoscopy: a modified Delphi method. Endoscopy. 2021 Sep;53(9):893–901.
10 Hassan C, Badalamenti M, Maselli R, Correale L, Iannone A, Radaelli F, et al. Computer-aided detection-assisted colonoscopy: classification and relevance of false positives. Gastrointest Endosc. 2020 Oct;92(4):900–4.e4.
11 Spadaccini M, Hassan C, Alfarone L, Da Rio L, Maselli R, Carrara S, et al. Comparing the number and relevance of false activations between 2 artificial intelligence computer-aided detection systems: the NOISE study. Gastrointest Endosc. 2022 May;95(5):975–81.e1.
12 Riphaus A, Wehrmann T, Hausmann J, Weber B, von Delius S, Jung M, et al. S3-guidelines “sedation in gastrointestinal endoscopy” 2014 (AWMF register no. 021/014). Z Gastroenterol. 2015 Aug;53(8):802–42.
13 Krenzer A, Makowski K, Hekalo A, Puppe F. Semi-automated machine learning video annotation for gastroenterologists. Stud Health Technol Inform. 2021 May 27;281:484–5.
14 Zhou G, Xiao X, Tu M, Liu P, Yang D, Liu X, et al. Computer aided detection for laterally spreading tumors and sessile serrated adenomas during colonoscopy. PLoS One. 2020;15(4):e0231880.
15 Misawa M, Kudo SE, Mori Y, Hotta K, Ohtsuka K, Matsuda T, et al. Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). Gastrointest Endosc. 2021 Apr;93(4):960–7.e3.
16 Kaminski MF, Thomas-Gibson S, Bugajski M, Bretthauer M, Rees CJ, Dekker E, et al. Performance measures for lower gastrointestinal endoscopy: a European Society of Gastrointestinal Endoscopy (ESGE) quality improvement initiative. United European Gastroenterol J. 2017 Apr;5(3):309–34.
17 Wang P, Xiao X, Glissen Brown JR, Berzin TM, Tu M, Xiong F, et al. Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. Nat Biomed Eng. 2018 Oct;2(10):741–8.
18 Klare P, Sander C, Prinzen M, Haller B, Nowack S, Abdelhafez M, et al. Automated polyp detection in the colorectum: a prospective study (with videos). Gastrointest Endosc. 2019 Mar;89(3):576–82.e1.
19 Luo Y, Zhang Y, Liu M, Lai Y, Liu P, Wang Z, et al. Artificial intelligence-assisted colonoscopy for detection of colon polyps: a prospective, randomized cohort study. J Gastrointest Surg. 2021 Aug;25(8):2011–8.
20 Wang P, Liu P, Glissen Brown JR, Berzin TM, Zhou G, Lei S, et al. Lower adenoma miss rate of computer-aided detection-assisted colonoscopy vs routine white-light colonoscopy in a prospective tandem study. Gastroenterology. 2020 Oct;159(4):1252–61.e5.
21 Wang P, Liu X, Berzin TM, Glissen Brown JR, Liu P, Zhou C, et al. Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): a double-blind randomised study. Lancet Gastroenterol Hepatol. 2020;5(4):343–51.
22 Zhao SB, Yang W, Wang SL, Pan P, Wang RD, Chang X, et al. Establishment and validation of a computer-assisted colonic polyp localization system based on deep learning. World J Gastroenterol. 2021 Aug 21;27(31):5232–46.
23 Heresbach D, Barrioz T, Lapalus MG, Coumaros D, Bauret P, Potier P, et al. Miss rate for colorectal neoplastic polyps: a prospective multicenter study of back-to-back video colonoscopies. Endoscopy. 2008 Apr;40(4):284–90.
24 Hagege H, Laugier R, Nahon S, Coulom P, Isnard-Bagnis C, Albert-Marty A. Real-life conditions of use of sodium phosphate tablets for colon cleansing before colonoscopy. Endosc Int Open. 2015 Aug;3(4):E346–53.

C Declaration of own Contributions

1. Adrian Krenzer, Amar Hekalo and Frank Puppe. Endoscopic Detection and Segmentation of Gastroenterological Diseases with Deep Convolutional Neural Networks. *EndoCV@ISBI*, 58-63, 2020

   AK implemented and coordinated the study, drafted the manuscript, interpreted the data and implemented the software. AH contributed to the completion of the manuscript. FP provided funding and contributed to the manuscript.

2. Adrian Krenzer and Frank Puppe. Bigger Networks are not Always Better: Deep Convolutional Neural Networks for Automated Polyp Segmentation. `https://web.archive.org/web/20220127160735id_/http://ceur-ws.org/Vol-2882/paper12.pdf`, 2020

   AK drafted the manuscript, interpreted the data and implemented the software. FP provided funding and contributed to the manuscript.

3. Sharib Ali, Mariia Dmitrieva, Noha Ghatwary, Sophia Bano, Gorkem Polat, Alptekin Temizel, Adrian Krenzer, Amar Hekalo, et al. Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy. *Medical image analysis*, 70:102002, 2021

   SA drafted the manuscript and gathered all the results of the challenge. AK and AH created the winning algorithm of the EndoCV 2020 challenge and drafted the related work of the manuscript. The other authors contributed to the algorithms created in the challenge or helped with the manuscript's creation.

4. Adrian Krenzer, Kevin Makowski, Amar Hekalo, Daniel Fitting, Joel Troya, Wolfram G. Zoller, Alexander Hann and Frank Puppe. Fast machine learning annotation in the medical domain: a semi-automated video annotation tool for gastroenterologists. *BioMedical Engineering OnLine*, 21:1-23, 2022

   AK implemented and coordinated the study, drafted the manuscript, and interpreted the data. KM and AK designed and implemented the software. AH1 and KM contributed to completing the manuscript. DF helped with the creation of the data. JT helped with the data preprocessing. FP, AH2 and WZ provided funding and reviewed the manuscript.

5. Markus Brand, Joel Troya, Adrian Krenzer, Zita Saßmannshausen, Wolfram G. Zoller, Alexander Meining, Thomas J. Lux and Alexander Hann. Development and evaluation of

a deep learning model to improve the usability of polyp detection systems during interventions. *United European Gastroenterology Journal*, Wiley Online Library, 2022

MB, JT, TJL, AH implemented and coordinated the study, drafted the manuscript, and interpreted the data. AK developed the AI for polyp detection used in the study, helped with the evaluation and contributed to the AI sections of the manuscript. ZS helped with the creation of the data. AH, AM, WZ provided funding and reviewed the manuscript.

6. Joel Troya, Adrian Krenzer, Krzysztof Flisikowski, Boban Sudarevic, Michael Banck, Alexander Hann, Frank Puppe, Alexander Meining. New concept for colonoscopy including side optics and artificial intelligence. *Gastrointestinal Endoscopy*, 95:794-798, 2022

   JT, AM and AK implemented and coordinated the study. FK helped with the execution of the study. JT drafted the manuscript. AK developed the AI for polyp detection used in the study, processed the data, evaluated the data with JT and contributed to the AI sections of the manuscript. MB helped with the implementation of the software. SB helped with the endoscope's creation and the study's execution. AH, AM and FP provided funding and reviewed the manuscript.

7. Thomas J. Lux, Michael Banck, Zita Saßmannshausen, Joel Troya, Adrian Krenzer, Daniel Fitting, Boban Sudarevic, Wolfram G. Zoller, Frank Puppe, Alexander Meining and Alexander Hann. Pilot study of a new freely available computer-aided polyp detection system in clinical practice. *International Journal of Colorectal Disease*, 1-6, 2022

   AH and TJL: study concept and design. AH, TJL, and ZS: statistical analysis. AK and MB: development of the software and neural network. AH, TJL, and MB: interpretation of results, and drafting of the manuscript. AH, TJL, ZS, JT, AK, DF, BS, WGZ, FP, and AM: acquisition of data and providing study material. All authors: critical revision of the article for important intellectual content and final approval of the article.

8. Daniel Fitting, Adrian Krenzer, Joel Troya, Michael Banck, Boban Sudarevic, Markus Brand, Wolfgang Böck, Wolfram G. Zoller, Thomas Rösch, Frank Puppe, Alexander Meining and Alexander Hann. A video based benchmark data set (ENDOTEST) to evaluate computer-aided polyp detection systems. *Scandinavian Journal of Gastroenterology*, 1-7, 2022.

   DF, AK, JT, FP and AH: study concept and design. DF, AK and JT: performed the experiments. DF, AK, JT, TR, FP and AH: interpretation of results, and drafting of the manuscript. DF, FP and AH: statistical analysis. DF, JT, MB, BS, WB, and WGZ: acquisition of data. All authors: critical revision of the article for important intellectual content and final approval of the article.

9. Adrian Krenzer, Kevin Makowski, Amar Hekalo, Frank Puppe. Semi-Automated Machine Learning Video Annotation for Gastroenterologists. *Public Health and Informatics*, 484–485, 2021.

AK implemented and coordinated the study, drafted the manuscript, interpreted the data and implemented the software. KM helped with the implementation of the software. AH contributed to the completion of the manuscript. FP provided funding and contributed to the manuscript.

10. Adrian Krenzer, Michael Banck, Kevin Makowski, Amar Hekalo, Daniel Fitting, Joel Troya, Boban Sudarevic, Wolfram G. Zoller, Alexander Hann and Frank Puppe. A Real-Time Polyp Detection System with Clinical Application in Colonoscopy Using Deep Convolutional Neural Networks. Accepted by *MDPI Journal of Imaging*, 2023.

    AK implemented and coordinated the study, drafted the manuscript, interpreted the data and implemented the software. MB contributed to the creation of the prototype. KM helped with the implementation of the software. AH1, KM, MB contributed to the completion of the manuscript. DF helped with the creation of the data. JT helped with the data preprocessing. BS contributed to the installation of the software. FP, AH2 and WZ provided funding and reviewed the manuscript. All authors read and approved the final manuscript.

11. Adrian Krenzer, Joel Troya, Michael Banck, Boban Sudarevic, Krzysztof Flisikowski, Alexander Meining and Frank Puppe. A User Interface for Automatic Polyp Detection Based on Deep Learning with Extended Vision. *Annual Conference on Medical Image Understanding and Analysis*, 851-868, Cambridge, 2022

    AK implemented and coordinated the study, drafted the manuscript, interpreted the data and implemented the software. JT, MB, BS, AK created the prototype. FK helped with the execution of the study. AM and FP provided funding and contributed to the manuscript.

12. Adrian Krenzer, Philipp Sodmann, Nico Hasler and Frank Puppe. Deep Learning using temporal information for automatic polyp detection in videos. *EndoCV@ISBI*, 14-19, 2022

    AK implemented and coordinated the study, drafted the manuscript, interpreted the data and implemented the software. NH contributed to the software. PS contributed to the manuscript. FP provided funding and contributed to the manuscript.

13. Markus Brand, Joel Troya, Adrian Krenzer, Costanza De Maria, Niklas Mehlhase, Sebastian Götze, Benjamin Walter, Alexander Meining, Alexander Hann. Frame-by-Frame Analysis of a Commercially Available Artificial Intelligence Polyp Detection System in Full-Length Colonoscopies. *Digestion*, 103:378-385, 2022

    MB, JT, AM, and AH: study concept and design, interpretation of results, and drafting of the manuscript. JT: statistical analysis. AK: Development of the AI for polyp detection. MB: annotation of videos. JT, AK, CDM, NM, SG, and BW: acquisition of data. MB, JT, AK, CDM, NM, SG, BW, AM, and AH: critical revision of the article for important intellectual content and final approval of the article.

14. Adrian Krenzer, Stefan Heil, Daniel Fitting, Safa Matti, Wolfram G. Zoller, Alexander Hann and Frank Puppe. Automated classification of polyps using deep learning architectures and few-shot learning. Under minor revision by *BMC Medical Imaging*, 2023.

    AK implemented and coordinated the study, drafted the manuscript, interpreted the data and implemented the software. SH designed, implemented and evaluated the NICE classification system and its derivative for the Paris classification system and contributed to the appertaining sections of the manuscript. SM contributed to the completion of the manuscript. DF helped with the creation of the data. FP, AH and WZ provided funding and reviewed the manuscript. All authors read and approved the final manuscript.

Bibliography

[1] L. V. Ahn and Laura A. Dabbish. Esp: Labeling images with a computer game. In *AAAI Spring Symposium: Knowledge Collection from Volunteer Contributors*, volume 2, 2005.

[2] Sharib Ali, Barbara Braden, Dominique Lamarque, Stefano Realdon, Adam Bailey, Renato Cannizzaro, Noha Ghatwary, Jens Rittscher, Christian Daul, and James East. Endoscopy disease detection and segmentation, 2020. URL `http://dx.doi.org/10.21227/f8xg-wb80`.

[3] Sharib Ali, Mariia Dmitrieva, Noha Ghatwary, Sophia Bano, Gorkem Polat, Alptekin Temizel, Adrian Krenzer, Amar Hekalo, Yun Bo Guo, Bogdan Matuszewski, et al. Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy. *Medical image analysis*, 70:102002, 2021.

[4] Quentin Angermann, Jorge Bernal, Cristina Sánchez-Montes, Maroua Hammami, Gloria Fernández-Esparrach, Xavier Dray, Olivier Romain, F. Javier Sánchez, and Aymeric Histace. Towards real-time polyp detection in colonoscopy videos: Adapting still frame-based methodologies for video sequences analysis. In *Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures*, pages 29–41, Cham, 2017. Springer International Publishing. ISBN 978-3-319-67543-5.

[5] Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *ArXiv*, abs/1701.07875, 2017.

[6] Mahnoosh Bagheri, Majid Mohrekesh, Milad Tehrani, Kayvan Najarian, Nader Karimi, Shadrokh Samavi, and S. M. Reza Soroushmehr. Deep neural network based polyp segmentation in colonoscopy images using a combination of color spaces. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6742–6745, 2019. doi: 10.1109/EMBC.2019.8856793.

[7] Ishita Barua, Daniela Guerrero Vinsard, Henriette C Jodal, Magnus Løberg, Mette Kalager, Øyvind Holme, Masashi Misawa, Michael Bretthauer, and Yuichi Mori. Artificial intelligence for polyp detection during colonoscopy: a systematic review and meta-analysis. *Endoscopy*, 53(03):277–284, 2021.

[8] Jorge Bernal, Javier Sánchez, and Fernando Vilarino. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition*, 45(9):3166–3182, 2012.

[9] Jorge Bernal, Javier Sanchez, and Fernando Vilariño. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition*, 45:3166–3182, 09 2012. doi: 10.1016/j.patcog.2012.03.002,.

[10] Jorge Bernal, Javier Sanchez, and Fernando Vilariño. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition*, 45:3166–3182, 09 2012. doi: 10.1016/j.patcog.2012.03.002,.

[11] Jorge Bernal, F. Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 43:99–111, July 2015. doi: 10.1016/j.compmedimag.2015.02.007. URL `https://doi.org/10.1016/j.compmedimag.2015.02.007`.

[12] Jorge Bernal, F. Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society*, 43: 99–111, July 2015. ISSN 1879-0771. doi: 10.1016/j.compmedimag.2015.02.007.

[13] P K Bhagat and Prakash Choudhary. Image annotation: Then and now. *Image and Vision Computing*, 80, 10 2018. doi: 10.1016/j.imavis.2018.09.017.

[14] Abhishek Bhandari, Melissa Woodhouse, and Samir Gupta. Colorectal cancer is a leading cause of cancer incidence and mortality among adults younger than 50 years in the usa: a seer-based analysis with comparison to other young-onset cancers. *Journal of Investigative Medicine*, 65(2):311–315, 2017.

[15] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[16] Aurélien Bour, Cristián Castillo-Olea, Begonya Garcia-Zapirain, and Sofia Zahia. Automatic colon polyp classification using convolutional neural network: A case study at basque country. In *2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pages 1–5, 2019. doi: 10.1109/ISSPIT47144.2019.9001816.

[17] Michael Byrne, Nicolas Chapados, Florian Soudan, Clemens Oertel, Milagros Pérez, Raymond Kelly, Nadeem Iqbal, Florent Chandelier, and Douglas Rex. Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model. *Gut*, 68:gutjnl–2017, 10 2017. doi: 10.1136/gutjnl-2017-314547.

[18] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[19] Ching-Yao Chan. Advancements, prospects, and impacts of automated driving systems. *International journal of transportation science and technology*, 6(3):208–216, 2017.

[20] Victor de Almeida Thomaz, Cesar A Sierra-Franco, and Alberto B Raposo. Training data enhancements for robust polyp segmentation in colonoscopy images. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 192–197. IEEE, 2019.

[21] Daniel C DeMarco, Elizabeth Odstrcil, Luis F Lara, David Bass, Chase Herdman, Timothy Kinney, Kapil Gupta, Leon Wolf, Thomas Dewar, Thomas M Deas, et al. Impact of experience with a retrograde-viewing device on adenoma detection rates and withdrawal times during colonoscopy: the third eye retroscope study group. *Gastrointestinal endoscopy*, 71 (3):542–550, 2010.

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[23] Abhishek Dutta and Andrew Zisserman. The VIA annotation software for images, audio and video. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, New York, NY, USA, 2019. ACM. ISBN 9781450368896. doi: 10.1145/3343031.3350535. URL `https://doi.org/10.1145/3343031.3350535`.

[24] Andre Esteva, Katherine Chou, Serena Yeung, Nikhil Naik, Ali Madani, Ali Mottaghi, Yun Liu, Eric Topol, Jeff Dean, and Richard Socher. Deep learning-enabled medical computer vision. *NPJ digital medicine*, 4(1):1–9, 2021.

[25] Andriy Fedorov, Reinhard Beichel, Jayashree Kalpathy-Cramer, Julien Finet, Jean-Christophe Fillion-Robin, Sonia Pujol, Christian Bauer, Dominique Jennings, Fiona Fennessy, Milan Sonka, et al. 3d slicer as an image computing platform for the quantitative imaging network. *Magnetic resonance imaging*, 30(9):1323–1341, 2012.

[26] Glòria Fernández-Esparrach, Jorge Bernal, Maria López-Cerón, Henry Córdova, Cristina Sánchez-Montes, Cristina Rodríguez de Miguel, and Francisco Javier Sánchez. Exploring the clinical potential of an automatic colonic polyp detection method based on the creation of energy maps. *Endoscopy*, 48:837–842, September 2016. ISSN 1438-8812. doi: 10.1055/s-0042-108434.

[27] Faming Gong, Hanbing Yue, Xiangbing Yuan, Wenjuan Gong, and Tao Song. Discriminative Correlation Filter for Long-Time Tracking. *The Computer Journal*, 63(3):460–468, 05 2019. ISSN 0010-4620. doi: 10.1093/comjnl/bxz049. URL `https://doi.org/10.1093/comjnl/bxz049`.

[28] Ian M Gralnek, Peter D Siersema, Zamir Halpern, Ori Segol, Alaa Melhem, Alain Suissa, Erwin Santo, Alan Sloyer, Jay Fenster, Leon MG Moons, et al. Standard forward-viewing colonoscopy versus full-spectrum endoscopy: an international, multicentre, randomised, tandem colonoscopy trial. *The lancet oncology*, 15(3):353–360, 2014.

[29] Yun Bo Guo and Bogdan Matuszewski. Giana polyp segmentation with fully convolutional dilation neural networks. In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 632–641. SCITEPRESS-Science and Technology Publications, 2019.

[30] Gaurav Gupta. Trainingdata.io: Ai assisted image & video training data labeling scale, 2019. URL `https://github.com/trainingdata/AIAssistedImageVideoLabelling/`. [Online; 11/13/2022].

[31] Gaurav Gupta and Anu Gupta. Trainingdata.io, 2019. URL `https://docs.trainingdata.io/`. [Online; 11/13/2022].

[32] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 3154–3160, 2017.

[33] Cesare Hassan, Carlo Senore, Franco Radaelli, Giovanni De Pretis, Romano Sassatelli, Arrigo Arrigoni, Gianpiero Manes, Arnaldo Amato, Andrea Anderloni, Franco Armelao, et al. Full-spectrum (fuse) versus standard forward-viewing colonoscopy in an organised colorectal cancer screening programme. *Gut*, 66(11):1949–1955, 2017.

[34] Cesare Hassan, Marco Spadaccini, Andrea Iannone, Roberta Maselli, Manol Jovani, Viveksandeep Thoguluva Chandrasekar, Giulio Antonelli, Honggang Yu, Miguel Areia, Mario Dinis-Ribeiro, et al. Performance of artificial intelligence in colonoscopy for adenoma and polyp detection: a systematic review and meta-analysis. *Gastrointestinal endoscopy*, 93(1): 77–85, 2021.

[35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[36] Nicholas Hoerter, Seth A Gross, and Peter S Liang. Artificial intelligence and polyp detection. *Current treatment options in gastroenterology*, 18(1):120–136, 2020.

[37] CM. Hsu, CC. Hsu, Z. Hsu, F. Shih, M. Chang, and T. Chen. Colorectal polyp image detection and classification through grayscale images and deep learning. *sensors*, 2021.

[38] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017.

[39] Sae Hwang, JungHwan Oh, Wallapak Tavanapong, Johnny Wong, and Piet C. de Groen. Polyp detection in colonoscopy video using elliptical shape feature. In *2007 IEEE International Conference on Image Processing*, volume 2, pages II – 465–II – 468, 2007. doi: 10.1109/ICIP.2007.4379193.

[40] Dimitris K Iakovidis and Anastasios Koulaouzidis. Automatic lesion detection in capsule endoscopy based on color saliency: closer to an essential adjunct for reviewing software. *Gastrointestinal endoscopy*, 80(5):877–883, 2014.

[41] Hussein Ibrahim, Xiaoxuan Liu, and Alastair K Denniston. Reporting guidelines for artificial intelligence in healthcare research. *Clinical & experimental ophthalmology*, 49(5):470–476, 2021.

[42] Hayato Itoh, Holger Roth, Masahiro Oda, Masashi Misawa, Yuichi Mori, Shin-Ei Kudo, and Kensaku Mori. Stable polyp-scene classification via subsampling and residual learning from an imbalanced large dataset. *Healthcare Technology Letters*, 6(6):237–242, December 2019. doi: 10.1049/htl.2019.0079. URL `https://doi.org/10.1049/htl.2019.0079`.

[43] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Dag Johansen, Thomas De Lange, Pål Halvorsen, and Håvard D Johansen. Resunet++: An advanced architecture for medical image segmentation. In *2019 IEEE International Symposium on Multimedia (ISM)*, pages 225–2255. IEEE, 2019.

[44] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International Conference on Multimedia Modeling*, pages 451–462. Springer, 2020.

[45] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1): 221–231, 2012.

[46] Jaeyong Kang and Jeonghwan Gwak. Ensemble of instance segmentation models for polyp segmentation in colonoscopy images. *IEEE Access*, 7:26440–26447, 2019. doi: 10.1109/access.2019.2900672. URL `https://doi.org/10.1109/access.2019.2900672`.

[47] Stavros Karkanis, Dimitris Iakovidis, Dimitris Maroulis, Dimitrios Karras, and M. Tzivras. Computer-aided tumor detection in endoscopic video using color wavelet features. *Information Technology in Biomedicine, IEEE Transactions on*, 7:141 – 152, 10 2003. doi: 10.1109/TITB.2003.813794.

[48] Vivek Kaul, Sarah Enslin, and Seth A Gross. History of artificial intelligence in medicine. *Gastrointestinal endoscopy*, 92(4):807–812, 2020.

[49] Kadircan H Keskinbora. Medical ethics considerations on artificial intelligence. *Journal of clinical neuroscience*, 64:277–282, 2019.

[50] Peter Klare, Christoph Sander, Martin Prinzen, Bernhard Haller, Sebastian Nowack, Mohamed Abdelhafez, Alexander Poszler, Hayley Brown, Dirk Wilhelm, Roland M Schmid, et al. Automated polyp detection in the colorectum: a prospective study (with videos). *Gastrointestinal endoscopy*, 89(3):576–582, 2019.

[51] Yoriaki Komeda, Hisashi Handa, Tomohiro Watanabe, Takanobu Nomura, Misaki Kitahashi, Toshiharu Sakurai, Ayana Okamoto, Tomohiro Minami, Masashi Kono, Tadaaki Arizumi, Mamoru Takenaka, Satoru Hagiwara, Shigenaga Matsui, Naoshi Nishida, Hiroshi Kashida, and Masatoshi Kudo. Computer-aided diagnosis based on convolutional neural network system for colorectal polyp classification: Preliminary experience. *Oncology*, 93: 30–34, 12 2017. doi: 10.1159/000481227.

[52] Adrian Krenzer and Frank Puppe. Bigger networks are not always better: Deep convolutional neural networks for automated polyp segmentation. In *MediaEval*, volume 2882, 2020.

[53] Adrian Krenzer, Amar Hekalo, and Frank Puppe. Endoscopic detection and segmentation of gastroenterological diseases with deep convolutional neural networks. In *EndoCV@ISBI*, pages 58–63, 2020.

[54] Adrian Krenzer, Kevin Makowski, Amar Hekalo, and Frank Puppe. Semi-automated machine learning video annotation for gastroenterologists. In *Public Health and Informatics*, pages 484–485. IOS Press, 2021.

[55] Adrian Krenzer, Stefan Heil, Daniel Fitting, Safa Matti, Wolfram G Zoller, Alexander Hann, and Frank Puppe. Automated classification of polyps using deep learning architectures and few-shot learning. *under minor revision of Medical Imaging*, 2022.

[56] Adrian Krenzer, Kevin Makowski, Amar Hekalo, Daniel Fitting, Joel Troya, Wolfram G Zoller, Alexander Hann, and Frank Puppe. Fast machine learning annotation in the medical domain: a semi-automated video annotation tool for gastroenterologists. *BioMedical Engineering OnLine*, 21(1):1–23, 2022.

[57] Adrian Krenzer, Philipp Sodmann, Nico Hasler, and Frank Puppe. Deep learning using temporal information for automatic polyp detection in videos. In *EndoCV@ ISBI*, pages 14–19, 2022.

[58] Adrian Krenzer, Joel Troya, Michael Banck, Boban Sudarevic, Krzysztof Flisikowski, Alexander Meining, and Frank Puppe. A user interface for automatic polyp detection based on deep learning with extended vision. In *Annual Conference on Medical Image Understanding and Analysis*, pages 851–868. Springer, 2022.

[59] Adrian Krenzer, Michael Banck, Kevin Makowski, Amar Hekalo, Daniel Fitting, Joel Troya, Boban Sudarevic, Wolfgang G Zoller, Alexander Hann, and Frank Puppe. A real-time polyp-detection system with clinical application in colonoscopy using deep convolutional neural networks. *Journal of Imaging*, 9(2):26, 2023.

[60] S.M. Krishnan, X. Yang, K.L. Chan, S. Kumar, and P.M.Y. Goh. Intestinal abnormality detection from endoscopic images. In *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Vol.20 Biomedical Engineering Towards the Year 2000 and Beyond (Cat. No.98CH36286)*, volume 2, pages 895–898 vol.2, 1998. doi: 10.1109/IEMBS.1998.745583.

[61] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[62] S Kudo, S Hirota, T Nakajima, S Hosobe, H Kusaka, T Kobayashi, M Himori, and A Yagyuu. Colorectal tumours and pit pattern. *Journal of Clinical Pathology*, 47(10): 880–885, 1994. ISSN 0021-9746. doi: 10.1136/jcp.47.10.880. URL `https://jcp.bmj.com/content/47/10/880`.

[63] A. Leibetseder, B. Münzer, K. Schoeffmann, and J. Keckstein. Endometriosis annotation in endoscopic videos. In *2017 IEEE International Symposium on Multimedia (ISM)*, pages 364–365, 2017. doi: 10.1109/ISM.2017.69.

[64] Anke M Leufkens, Daniel C DeMarco, Amit Rastogi, Paul A Akerman, Kassem Azzouzi, Richard I Rothstein, Frank P Vleggaar, Alessandro Repici, Giacomo Rando, Patrick I Okolo, et al. Effect of a retrograde-viewing device on adenoma detection rate during colonoscopy: the terrace study. *Gastrointestinal endoscopy*, 73(3):480–489, 2011.

[65] Ming Liu, Jue Jiang, and Zenan Wang. Colonic polyp detection in endoscopic videos with single shot detection based deep convolutional neural network. *IEEE Access*, 7:75058–75066, 2019.

[66] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. *ArXiv*, abs/1512.02325, 2016.

[67] Yuyuan Liu, Yu Tian, Gabriel Maicas, Leonardo Zorron Cheng Tao Pu, Rajvinder Singh, Johan W. Verjans, and Gustavo Carneiro. Photoshopping colonoscopy video frames. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, April 2020. doi: 10.1109/isbi45749.2020.9098406. URL `https://doi.org/10.1109/isbi45749.2020.9098406`.

[68] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12009–12019, 2022.

[69] Thomas Lui, Kenneth Wong, Loey Mak, Michael Ko, Stephen Tsao, and Wai Leung. Endoscopic prediction of deeply submucosal invasive carcinoma with use of artificial intelligence. *Endoscopy International Open*, 07:E514–E520, 04 2019. doi: 10.1055/a-0849-9548.

[70] Thomas J Lux, Michael Banck, Zita Saßmannshausen, Joel Troya, Adrian Krenzer, Daniel Fitting, Boban Sudarevic, Wolfram G Zoller, Frank Puppe, Alexander Meining, et al. Pilot study of a new freely available computer-aided polyp detection system in clinical practice. *International Journal of Colorectal Disease*, pages 1–6, 2022.

[71] Niall O' Mahony, Sean Campbell, Anderson Carvalho, Suman Harapanahalli, Gustavo Velasco-Hernandez, Lenka Krpalkova, Daniel Riordan, and Joseph Walsh. Deep learning vs. traditional computer vision. doi: 10.1007/978-3-030-17795-9.

[72] Petar Mamula, William M Tierney, Subhas Banerjee, David Desilets, David L Diehl, Francis A Farraye, Vivek Kaul, Sripathi R Kethu, Richard S Kwon, Marcos C Pedrosa, et al. Devices to improve colon polyp detection. *Gastrointestinal endoscopy*, 73(6):1092–1097, 2011.

[73] Pablo Martinez, Mohamed Al-Hussein, and Rafiq Ahmad. A scientometric analysis and critical review of computer vision applications for construction. *Automation in Construction*, 107:102947, 2019.

[74] Camilla Mattiuzzi, Fabian Sanchis-Gomar, and Giuseppe Lippi. Concise update on colorectal cancer epidemiology. *Annals of translational medicine*, 7(21), 2019.

[75] Ateev Mehrotra, Michele Morris, Rebecca A Gourevitch, David S Carrell, Daniel A Leffler, Sherri Rose, Julia B Greer, Seth D Crockett, Andrew Baer, and Robert E Schoen. Physician characteristics associated with higher adenoma detection rate. *Gastrointestinal endoscopy*, 87(3):778–786, 2018.

[76] P. Mesejo, D. Pizarro, A. Abergel, O. Rouquette, S. Beorchia, L. Poincloux, and A. Bartoli. Computer-aided classification of gastrointestinal lesions in regular colonoscopy. *IEEE Transactions on Medical Imaging*, 35, 2016.

[77] Microsoft. Visual object tagging tool, 2019. URL `https://github.com/microsoft/VoTT`. [15.03.2022].

[78] Masashi Misawa, Shinei Kudo, Yuichi Mori, Tomonari Cho, Shinichi Kataoka, Yasuharu Maeda, Yushi Ogawa, Kenichi Takeda, Hiroki Nakamura, Katsuro Ichimasa, et al. Tu1990 artificial intelligence-assisted polyp detection system for colonoscopy, based on the largest available collection of clinical video data for machine learning. *Gastrointestinal Endoscopy*, 89(6):AB646–AB647, 2019.

[79] Masashi Misawa, Shin-ei Kudo, Yuichi Mori, Kinichi Hotta, Kazuo Ohtsuka, Takahisa Matsuda, Shoichi Saito, Toyoki Kudo, Toshiyuki Baba, Fumio Ishida, et al. Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). *Gastrointestinal Endoscopy*, 93(4):960–967, 2021.

[80] Xi Mo, Ke Tao, Quan Wang, and Guanghui Wang. An efficient approach for polyps detection in endoscopic videos based on faster r-cnn. In *2018 24th international conference on pattern recognition (ICPR)*, pages 3929–3934. IEEE, 2018.

[81] Andrew Ng et al. Sparse autoencoder. *CS294A Lecture notes*, 72(2011):1–19, 2011.

[82] Fangwei Ning, Yan Shi, Maolin Cai, Weiqing Xu, and Xianzhi Zhang. Manufacturing cost estimation based on a deep-learning method. *Journal of Manufacturing Systems*, 54:186–195, 2020.

[83] Alba Nogueira-Rodríguez, Ruben Dominguez-Carbajales, Fernando Campos-Tato, Jesús Herrero, Manuel Puga, David Remedios, Laura Rivas, Eloy Sánchez, Agueda Iglesias, Joaquín Cubiella, et al. Real-time polyp detection model using convolutional neural networks. *Neural Computing and Applications*, 34(13):10375–10396, 2022.

[84] Benjamin Nulsen, Ryan C Ungaro, Natalie Davis, Elliot Turvall, Lisa Deutsch, and Blair Lewis. Changes in adenoma detection rate with implementation of full-spectrum endoscopy. *Journal of Clinical Gastroenterology*, 52(10):885–890, 2018.

[85] Tsuyoshi Ozawa, Soichiro Ishihara, Mitsuhiro Fujishiro, Youichi Kumagai, Satoki Shichijo, and Tomohiro Tada. Automated endoscopic detection and classification of colorectal polyps using convolutional neural networks. *Therapeutic Advances in Gastroenterology*, 13: 175628482091065, 03 2020. doi: 10.1177/1756284820910659.

[86] Kenneth Philbrick, Alexander Weston, Zeynettin Akkus, Timothy Kline, Panagiotis Korfiatis, Tomas Sakinis, Petro Kostandy, Arunnit Boonrod, Atefeh Zeinoddini, Naoki Takahashi, and Bradley Erickson. Ril-contour: a medical imaging dataset annotation tool for and with deep learning. *Journal of Digital Imaging*, 32, 05 2019. doi: 10.1007/s10278-019-00232-0.

[87] Hemin Ali Qadir, Ilangko Balasingham, Johannes Solhusvik, Jacob Bergsland, Lars Aabakken, and Younghak Shin. Improving automatic polyp detection using CNN by exploiting temporal dependency in colonoscopy video. *IEEE Journal of Biomedical and Health Informatics*, 24(1):180–193, January 2020. doi: 10.1109/jbhi.2019.2907434. URL `https://doi.org/10.1109/jbhi.2019.2907434`.

[88] Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358, 2019.

[89] Aparna Ratheesh, Pooja Soman, M Revathy Nair, RG Devika, and RP Aneesh. Advanced algorithm for polyp detection using depth segmentation in colon endoscopy. In *2016 International Conference on Communication Systems and Networks (ComNet)*, pages 179–183. IEEE, 2016.

[90] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[91] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[92] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, June 2017. doi: 10.1109/tpami.2016.2577031. URL `https://doi.org/10.1109/tpami.2016.2577031`.

[93] Eduardo Ribeiro, Andreas Uhl, and Michael Häfner. Colonic polyp classification with convolutional neural networks. In *2016 IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 253–258, 2016. doi: 10.1109/CBMS.2016.39.

[94] Samantha Cruz Rivera, Xiaoxuan Liu, An-Wen Chan, Alastair K Denniston, and Melanie J Calvert. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the spirit-ai extension. *bmj*, 370, 2020.

[95] Rubin-Lab. epad: web-based platform for quantitative imaging in the clinical workflow, 2014. URL `https://epad.stanford.edu/`. [Online; 8/13/2022].

[96] Bryan C. Russell, Antonio Torralba, Kevin P Murphy, and William T. Freeman. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77, 05 2008. doi: 10.1007/s11263-007-0090-8. URL `https://doi.org/10.1007/s11263-007-0090-8`.

[97] Nabile M Safdar, John D Banja, and Carolyn C Meltzer. Ethical considerations in artificial intelligence. *European journal of radiology*, 122:108768, 2020.

[98] Robert R Schaller. Moore's law: past, present and future. *IEEE spectrum*, 34(6):52–59, 1997.

[99] Boris Sekachev, Nikita Manovich, and Andrey Zhavoronkov. Computer vision annotation tool: A universal approach to data annotation, 2019. URL `https://software.intel.com/content/www/us/en/develop/articles/computer-vision-annotation-tool-a-universal-approach-to-data-annotation.html`.

[100] Hoo-Chang Shin, Neil A Tenenholtz, Jameson K Rogers, Christopher G Schwarz, Matthew L Senjem, Jeffrey L Gunter, Katherine P Andriole, and Mark Michalski. Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In *International workshop on simulation and synthesis in medical imaging*, pages 1–11. Springer, 2018.

[101] Younghak Shin, Hemin Ali Qadir, Lars Aabakken, Jacob Bergsland, and Ilangko Balasingham. Automatic colon polyp detection using region based deep CNN and post learning approaches. *IEEE Access*, 6:40950–40962, 2018. doi: 10.1109/access.2018.2856402. URL `https://doi.org/10.1109/access.2018.2856402`.

[102] Jenni AM Sidey-Gibbons and Chris J Sidey-Gibbons. Machine learning in medicine: a practical introduction. *BMC medical research methodology*, 19(1):1–18, 2019.

[103] Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International Journal of Computer Assisted Radiology and Surgery*, 9(2):283–293, 2014.

[104] Roger D. Soberanis-Mukul, Maxime Kayser, A. A. Zvereva, Peter Klare, Nassir Navab, and Shadi Albarqouni. A learning without forgetting approach to incorporate artifact knowledge in polyp localization tasks. *ArXiv*, abs/2002.02883, 2020.

[105] Jeong-Yeop Song, Youn Hee Cho, Mi A Kim, Jeong-Ae Kim, Chun Tek Lee, and Moon Sung Lee. Feasibility of full-spectrum endoscopy: Korea's first full-spectrum endoscopy colonoscopic trial. *World Journal of Gastroenterology*, 22(8):2621, 2016.

[106] S. Sornapudi, Frank Meng, and Steven Yi. Region-based automated localization of colonoscopy and wireless capsule endoscopy polyps. *Applied Sciences*, 9:2404, 2019.

[107] D.G. Stork. Character and document research in the open mind initiative. In *Proceedings of the Fifth International Conference on Document Analysis and Recognition. ICDAR '99 (Cat. No.PR00318)*, pages 1–12, 1999. doi: 10.1109/ICDAR.1999.791712.

[108] Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. *CoRR*, abs/1602.07261, 2016. URL `http://arxiv.org/abs/1602.07261`.

[109] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.

[110] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*, 35(2):630–644, 2015.

[111] Nima Tajbakhsh, Jae Y. Shin, Suryakanth R. Gurudu, R. Todd Hurst, Christopher B. Kendall, Michael B. Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging*, 35(5):1299–1312, May 2016. doi: 10.1109/tmi.2016.2535302. URL `https://doi.org/10.1109/tmi.2016.2535302`.

[112] Sushama Tanwar, Pallavi Goel, Prashant Johri, and Mario Diván. Classification of benign and malignant colorectal polyps using pit pattern classification. *SSRN Electronic Journal*, 01 2020. doi: 10.2139/ssrn.3558374.

[113] G Triadafilopoulos and J Li. A pilot study to assess the safety and efficacy of the third eye retrograde auxiliary imaging system during colonoscopy. *Endoscopy*, 40(06):478–482, 2008.

[114] Oleg O Varlamov, Dmitry A Chuvikov, Larisa E Adamova, Maxim A Petrov, Irina K Zabolotskaya, and Tatyana N Zhilina. Logical, philosophical and ethical aspects of ai in medicine. *International Journal of Machine Learning and Computing*, 9(6):868, 2019.

[115] David Vázquez, Jorge Bernal, F. Javier Sánchez, Gloria Fernández-Esparrach, Antonio M. López, Adriana Romero, Michal Drozdzal, and Aaron Courville. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of Healthcare Engineering*, 2017:1–9, 2017. doi: 10.1155/2017/4037190. URL `https://doi.org/10.1155/2017/4037190`.

[116] Ioannis Vourgidis, Shadreck Joseph Mafuma, Paul Wilson, Jenny Carter, and Georgina Cosma. Medical expert systems–a study of trust and acceptance by healthcare stakeholders. In *UK Workshop on Computational Intelligence*, pages 108–119. Springer, 2018.

[117] David Vázquez, Jorge Bernal, F. Javier Sánchez, Gloria Fernández-Esparrach, Antonio M. López, Adriana Romero, Michal Drozdzal, and Aaron Courville. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of healthcare engineering*, 2017: 4037190, 2017. ISSN 2040-2295. doi: 10.1155/2017/4037190.

[118] Dechun Wang, Ning Zhang, Xinzi Sun, Pengfei Zhang, Chenxi Zhang, Yu Cao, and Benyuan Liu. Afp-net: Realtime anchor-free polyp detection in colonoscopy. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 636–643. IEEE, 2019.

[119] Fei Wang, Lawrence Peter Casalino, and Dhruv Khullar. Deep learning in medicine—promise, progress, and challenges. *JAMA internal medicine*, 179(3):293–294, 2019.

[120] Jerome D Waye, Russell I Heigh, David E Fleischer, Jonathan A Leighton, Suryakanth Gurudu, Leslie B Aldrich, Jiayi Li, Sanjay Ramrakhiani, Steven A Edmundowicz, Dayna S Early, et al. A retrograde-viewing device improves detection of adenomas in the colon: a prospective efficacy evaluation (with videos). *Gastrointestinal endoscopy*, 71(3):551–556, 2010.

[121] Rüdiger Wirth and Jochen Hipp. Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, volume 1, pages 29–39. Manchester, 2000.

[122] Jianwei Xu, Ran Zhao, Yizhou Yu, Qingwei Zhang, Xianzhang Bian, Jun Wang, Zhizheng Ge, and Dahong Qian. Real-time automatic polyp detection in colonoscopy using feature enhancement module and spatiotemporal similarity correlation unit. *Biomedical Signal Processing and Control*, 66:102503, 2021.

[123] Yixuan Yuan and Max Q.-H. Meng. Deep learning for polyp recognition in wireless capsule endoscopy images. *Medical Physics*, 44(4):1379–1389, 2017. doi: https://doi.org/10.1002/mp.12147. URL `https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.12147`.

[124] Yixuan Yuan, Wenjian Qin, Bulat Ibragimov, Guanglei Zhang, Bin Han, Max Q.-H. Meng, and Lei Xing. Densely connected neural network with unbalanced discriminant and category sensitive constraints for polyp recognition. *IEEE Transactions on Automation Science and*

Engineering, 17(2):574–583, April 2020. doi: 10.1109/tase.2019.2936645. URL `https://doi.org/10.1109/tase.2019.2936645`.

[125] Zijie Yuan, Mohammadhassan IzadyYazdanabadi, Divya Mokkapati, Rujuta Panvalkar, Jae Y Shin, Nima Tajbakhsh, Suryakanth Gurudu, and Jianming Liang. Automatic polyp detection in colonoscopy videos. In *Medical Imaging 2017: Image Processing*, volume 10133, page 101332K. International Society for Optics and Photonics, 2017.

[126] P. A. Yushkevich, Y. Gao, and G. Gerig. Itk-snap: An interactive tool for semi-automatic segmentation of multi-modality biomedical images. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3342–3345, 2016. doi: 10.1109/EMBC.2016.7591443.

[127] Pengfei Zhang, Xinzi Sun, Dechun Wang, Xizhe Wang, Yu Cao, and Benyuan Liu. An efficient spatial-temporal polyp detection framework for colonoscopy video. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1252–1259. IEEE, 2019.

[128] Ruikai Zhang, Yali Zheng, Wing Mak, Ruoxi Yu, Sunny Wong, and Carmen Poon. Automatic detection and classification of colorectal polyps by transferring low-level cnn features from nonmedical domain. *IEEE Journal of Biomedical and Health Informatics*, PP:1–1, 12 2016. doi: 10.1109/JBHI.2016.2635662.

[129] Xiangyu Zhang, Jianhua Zou, Kaiming He, and Jian Sun. Accelerating very deep convolutional networks for classification and detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:1943–1955, 2016.

[130] Xu Zhang, Fei Chen, Tao Yu, Jiye An, Zhengxing Huang, Jiquan Liu, Weiling Hu, Liangjing Wang, Huilong Duan, and Jianmin Si. Real-time gastric polyp detection using convolutional neural networks. *PLOS ONE*, 14(3):e0214133, March 2019. doi: 10.1371/journal.pone.0214133. URL `https://doi.org/10.1371/journal.pone.0214133`.

[131] Xu Zhang, Fei Chen, Tao Yu, Jiye An, Zhengxing Huang, Jiquan Liu, Weiling Hu, Liangjing Wang, Huilong Duan, and Jianmin Si. Real-time gastric polyp detection using convolutional neural networks. *PLOS ONE*, 14(3):1–16, 03 2019. doi: 10.1371/journal.pone.0214133. URL `https://doi.org/10.1371/journal.pone.0214133`.

[132] Yali Zheng, Ruikai Zhang, Ruoxi Yu, Yuqi Jiang, Tony WC Mak, Sunny H Wong, James YW Lau, and Carmen CY Poon. Localisation of colorectal polyps by convolutional neural network features learnt from white light and narrow band endoscopic images of multiple databases. In *2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 4142–4145. IEEE, 2018.

[133] Rongsheng Zhu, Rong Zhang, and Dixiu Xue. Lesion detection of endoscopy images based on convolutional neural network features. In *2015 8th International Congress on Image and Signal Processing (CISP)*, pages 372–376, 2015. doi: 10.1109/CISP.2015.7407907.