Received: 17 November 2020 | Revised: 18 March 2021 | Accepted: 9 April 2021

DOI: 10.1002/mrm.28822

FULL PAPER

Magnetic Resonance in Medicine

Deep learning-based cardiac cine segmentation: Transfer learning application to 7T ultrahigh-field MRI

Markus Johannes Ankenbrand[1] | **David Lohr**[1] | **Wiebke Schlötelburg**[1,2] | **Theresa Reiter**[1,3] | **Tobias Wech**[2] | **Laura Maria Schreiber**[1]

[1] Chair of Cellular and Molecular Imaging, Comprehensive Heart Failure Center (CHFC), University Hospital Wuerzburg, Wuerzburg, Germany

[2] Department of Radiology, University Hospital Wuerzburg, Wuerzburg, Germany

[3] Department of Internal Medicine I, University Hospital Wuerzburg, Wuerzburg, Germany

Correspondence
Markus J. Ankenbrand, Chair of Cellular and Molecular Imaging, Comprehensive Heart Failure Center (CHFC), University Hospital Wuerzburg, Am Schwarzenberg 15, 97078 Wuerzburg, Germany.
Email: markus.ankenbrand@uni-wuerzburg.de

Funding information
This work was supported by the German Ministry of Education and Research (01EO1504). The funding body took no role in the design of the study, collection, analysis, and interpretation of data and in writing the manuscript.

Purpose: Artificial neural networks show promising performance in automatic segmentation of cardiac MRI. However, training requires large amounts of annotated data and generalization to different vendors, field strengths, sequence parameters, and pathologies is limited. Transfer learning addresses this challenge, but specific recommendations regarding type and amount of data required is lacking. In this study, we assess data requirements for transfer learning to experimental cardiac MRI at 7T where the segmentation task can be challenging. In addition, we provide guidelines, tools, and annotated data to enable transfer learning approaches by other researchers and clinicians.

Methods: A publicly available segmentation model was used to annotate a publicly available data set. This labeled data set was subsequently used to train a neural network for segmentation of left ventricle and myocardium in cardiac cine MRI. The network is used as starting point for transfer learning to 7T cine data of healthy volunteers ($n = 22$; 7873 images) by updating the pre-trained weights. Structured and random data subsets of different sizes were used to systematically assess data requirements for successful transfer learning.

Results: Inconsistencies in the publically available data set were corrected, labels created, and a neural network trained. On 7T cardiac cine images the model pre-trained on public imaging data, acquired at 1.5T and 3T, achieved $\text{DICE}_{\text{LV}} = 0.835$ and $\text{DICE}_{\text{MY}} = 0.670$. Transfer learning using 7T cine data and ImageNet weight initialization improved model performance to $\text{DICE}_{\text{LV}} = 0.900$ and $\text{DICE}_{\text{MY}} = 0.791$. Using only end-systolic and end-diastolic images reduced training data by 90%, with

Markus Johannes Ankenbrand and David Lohr Contributed equally to this work.

Parts of this work will be used in the doctoral thesis of David Lohr.

Parts of this work have been accepted at the ISMRM 2020.

no negative impact on segmentation performance (DICE$_{LV}$ = 0.908, DICE$_{MY}$ = 0.805).

Conclusions: This work demonstrates and quantifies the benefits of transfer learning for cardiac cine image segmentation. We provide practical guidelines for researchers planning transfer learning projects in cardiac MRI and make data, models, and code publicly available.

KEYWORDS

7T, cardiac function, cardiac magnetic resonance, deep learning, neural networks, segmentation, transfer learning, ultrahigh-field

1 | INTRODUCTION

Image segmentation is an essential step in the functional analysis of cardiac magnetic resonance imaging. It allows for the extraction of quantitative static measures such as myocardial mass, left ventricular (LV) volume, right ventricular (RV) volume, and wall thickness, as well as dynamic measures such as analysis of wall motion and the ejection fraction (EF). Cardiac cine MRI is the accepted gold standard for this assessment of cardiac function[1] and anatomy and is, therefore, of paramount clinical importance.[2,3] Proper segmentation of such data sets is a tedious and time-consuming process that has increasingly been tackled using various deep learning approaches.[4-7]

Artificial neural networks have been shown to outperform other methods on several high profile image analysis benchmarks and, therefore, so-called deep learning models have become state-of-the-art for a wide variety of computer vision tasks. Multiple factors like the wide application area of deep learning, available compute power, and increasing investments as well as user-friendly open source software have enabled a rapid development of the field of artificial intelligence. This led to ever increasing applications in medical imaging such as MRI[8] where tasks nowadays range from data acquisition and image reconstruction,[9-11] image restoration,[12,13] to image registration,[14,15] and segmentation[16-19] as well as classification[20,21] and outcome prediction.[22,23]

There is consensus in the field that the limited availability of labeled or annotated data because of data access, privacy issues, missing data harmonization, and data protection is one of the main obstacles for future clinical applications of deep neural networks.[17,19,24] Whereas some resources like the UK Biobank[25] already exist to address this issue, the high quality standards and the amount of work required to organize and maintain such a resource makes data access expensive. In addition, such data may already exceed the quality that is available in clinical routine cardiac MRI. This leads to neural networks, which perform very well for a very specific task within a confined data space, where training and testing data share the same distribution. However, these networks usually lack generalization capabilities. Whereas methods such as data augmentation, transfer learning, weakly-supervised, self-supervised, and unsupervised learning have been applied to overcome the issue of small data sets in research, it is unclear how much data are really required to create a well-generalizing network or to apply transfer learning. In transfer learning, the target model is not created from scratch. Instead (part of) an existing pre-trained model is used as starting point. This model was usually trained using a lot of data on a task that is similar to the target task (eg, image classification using ImageNet data as pre-training for semantic segmentation). A related kind of transfer learning is re-training a network on the same task using data with different characteristics (eg, pre-training on 3T MRI images for transfer learning to 7T images). Notable alternative methods for domain transfer include Cycle-GAN[26] and neural style transfer.[27]

In this work, we aim to enable researchers and clinicians in cardiology to apply deep learning-based segmentation models in their respective research by providing guidelines and easily accessible tools as well as annotated data for transfer learning.

2 | METHODS

2.1 | Kaggle data

As mentioned above, cardiac MRI is the gold standard for the assessment of cardiac function, a key indicator of cardiac disease. The 2015 Data Science Bowl challenged participants to create an algorithm for automatic assessment of end-systolic and end-diastolic volumes (ESV and EDV) and therefore, ejection fraction, based on cardiac cine MRI. The data set consists of pre-defined training, validation, and test sets and once the challenge has ended, all sets and their corresponding volume information (end-systolic and end-diastolic)

TABLE 1 Data composition and measurement parameters of the Kaggle data

Metric	Count
Male	670
Female	470
Age: 0-17 (y)	202
Age: 18-30 (y)	173
Age: 31-50 (y)	298
Age: 51+ (y)	467
Max Age (y)	88
Min Age (y)	0.04
1.5 T	1025
3.0 T	115
Metric	**Range**
Echo Time (ms)	1.04-1.54
Repetition Time (ms)	14-54.72
Bandwidth (Hz/Pixel)	915-1235
Slice Thickness (mm)	5-8
Matrix Size	120-608 × 160-736
Resolution (mm)	0.59-1.95
Phases	112-416

were made available for research and academic pursuits, leading to a total of 1140 "annotated" cardiac MRI examinations of normal and abnormal cardiac function. Images are in DICOM format resolving up to 30 phases of the cardiac cycle. Although we will focus on short axis images in this study, the Kaggle data set also contains alternative views. Examinations were done on 1.5T and 3.0T systems (Siemens Magnetom Aera and Skyra, Siemens Healthineers, Erlangen, Germany) with applications of both FLASH and TrueFISP sequences. An overview of the complete data set and its variation in patient data and sequence parameters is given in Table 1.

2.2 | 7T data

All assessments regarding transfer learning to 7T data are done using model: r34_CE_p5_s2. As initial point of comparison we used the UKBB model to create labels for 7T data, to assess generalization capability of a model, which was trained on a very homogeneous data set (UKBB).

Following approval of the local ethics committee (7/17-SC), n = 22 (14 female, 8 male) healthy volunteers were examined using a 7T whole body MRI system (Siemens MAGNETOM Terra, Erlangen, Germany) and a 1TX/16RX thorax coil (MRI Tools, Berlin, Germany).[28] Written informed consent was obtained before all measurements. All human volunteers gave their consent for publication using our institutional consent form. Volunteer age was 22-53 years, body weight was 52-95 kg, and height was 151-185 cm. For triggering, both the integrated ECG and an external acoustic triggering system (MRI Tools, Berlin, Germany) were used to synchronize measurements with the heartbeat, choosing whichever method provided a more stable trigger signal during the examination. Images were obtained during initial sequence implementation and optimization for 7T cardiac MRI using a cardiovascular (CV) GRE cine-sequence and protocol parameters, therefore, vary to some degree. The parameters were: TE = 3.57 ms, FOV = 340 mm × 320 mm, interpolated voxel size = 0.66 × 0.66 × 6 mm, and GRAPPA acceleration factors: R = 2 and R = 3. Depending on the heart rate, 6 to 11 segments and 20 to 35 cardiac phases were measured using retrospective gating. Short axis CINE stacks for volumetric evaluation varied in the number of slices (14-17) and multiple breath-holds (~13 s) were necessary to acquire the whole stack. Volunteers were assigned randomly into training, validation and test sets (14, 5, and 3 subjects and 5076, 1842, and 955 images, respectively). All images were manually segmented by an expert radiologist (WS). Three data sets of the test set were additionally segmented by an expert cardiologist (TR), to obtain an estimate of interobserver-variability.

2.3 | Data curation

The complete Kaggle data set is a compilation of real, clinical data from several sites and as such, subject to inconsistencies within individual examinations. Those can be a combination of:

- missing time points,
- inconsistent slice spacing,
- inconsistent image dimension,
- repeat measurements (identical slice location),
- scaled images, and
- image rotations.

Before the application of the published segmentation network of Bai et al,[4] we performed data curation, correcting inconsistencies in all but 8 examinations. More detailed information and curated data are available online (https://github.com/chfc-cmi/cmr-seg-tl, https://doi.org/10.5281/zenodo.3876351).

2.4 | Creating labels

Once the data were corrected for inconsistencies we ran the Python-based segmentation model of Bai et al.[4] for the complete data set, generating RV, LV, and blood pool labels as well as LV ESV and EDV volumes. ESV and EDV values were then compared to the ground truth values provided by Kaggle to determine the accuracy of the network prediction. Based on this comparison, we created confidence sets where the predicted values were in the range of ±5% (p5), ±10% (p10), and ±15% (p15) of the true value. Respectively, these sets contained 175, 520, and 763 examinations and 54,540, 162,480, and 238,350 images. Volume and ejection fraction statistics for each sub-cohort are shown in Supporting Information Table S1 and Supporting Information Figure S1. As additional quality control, 552 images from 9 patients (randomly selected from the p15 confidence set, 6 phases per patient, all slices) were manually annotated by an expert cardiologist (TR) to check for systematic differences in annotation and to validate the automatic labels using DICE scores. Volumes calculated from these masks were compared to ground truth values through Pearson correlation coefficient.

2.5 | Hardware

To deal with the extensive computation demands we used a custom workstation and a high performance cluster, both with graphical processing units.

2.6 | Framework: deep neural network

All implementations were realized using Pytorch[29] and fastai[30] V1. All networks use the U-Net architecture consisting of an encoder and decoder part with skip connections in between.[31] The encoder part is often referred to as the backbone. Training of neural networks (with varying backbones: Restnet34,[32] ResNet50,[33] and VGG16[34]) was performed using fastai's implementation of the *1 cycle* policy[35] with adjusted learning rates (lr) and the confidence sets p5, p10, and p15. The p5, p10, and p15 confidence sets include patients with a maximum of 5%, 10%, and 15% difference between predicted and true volume, respectively.

2.7 | Parameter search

During the parameter search, we evaluated the influence of different training parameters on the efficacy of the trained model. Training with a weight-decay of 0.02 and a batch-size of 32 was done for 30 epochs with frozen weights ($\text{lr} = 1\text{e}^{-4}$) and another 30 epochs with unfrozen weights ($\text{lr} = 1\text{e}^{-5}$). The smallest training set (p5) was used initially, image size was 256 × 256, and moderate data augmentation transforms (s1: flip [none], rotation [20°], lighting [0.4], zoom [1.2], and padding [zeros]) were applied.

To avoid an extensive parameter grid search, we assessed parameter-dependent performance changes in incremental steps. After each step, we determined the best-performing model using EF predictions and introduced subsequent parameter variations on this respective model.

In the first step, we evaluated the influence of the architecture (VGG16, ResNet34, ResNet50) compared to the fully convolutional Network by Bai et al[4] trained on UKBB data (further referred to as UKBB model). Because of memory limitations, we had to reduce the batch size for training of the VGG16 and the ResNet50 models.

In the second step, we assessed variations in the loss function such as cross-entropy (default), generalized DICE,[36] and focal loss. In the third and last step we evaluated the influence of the number of training images using the confidence sets p5, p10, and p15.

2.8 | Data augmentation

Because transfer learning applications assessed in this study are based on 7T data, we expect somewhat different image contrast and artefacts compared to conventional, clinical data sets. In addition, we intended to account for the heterogeneous training data, which led to the following set of augmentations as implemented in fastai[30] for the initial networks (s1: flip [none], rotation [20°], lighting [0.4], zoom [1.2], and padding [zeros]). Furthermore, we aimed to introduce some robustness to forms of data variations, such as 90°-rotations and flips (left-right) using more extensive data augmentation (s2: flip [left-right], rotation [90°], lighting [0.4], zoom [1.2], and padding [zeros]). To test the efficacy of these transforms, we trained a new model (r34_CE_p5_s2) and compared EF predictions on a data set including rotated and flipped images retained during the data curation process.

2.9 | Starting point for model training: 7T human

To assess the efficacy of transfer learning for LV segmentation based on clinical 1.5T and 3T data and experimental (human) 7T data, we compare models with varying degrees of training and transfer learning. Using a U-Net architecture with a ResNet34 backbone (r34_CE_p5_s2), we generated the following 3 models: (1) initialization with random weights (R), (2) initialization with ImageNet-weights: transfer learning 1 (TL), and (3) model 2, pre-trained on Kaggle data: transfer learning 2 (TL2). All models were used to generate predictions for the 7T test set. Model performance was always evaluated using the Soerensen-DICE[37] coefficient between predictions and respective ground truth labels.

2.10 | Data requirements for model training: 7T human

To assess how much and what data are required for convergence of a model we trained all models (R, TL, TL^2) with subsets of the training data. These subsets were created in 2 ways:

1. Complete subject data (all slices and all phases) from either 14, 7, 3, 1 subjects (5076, 2626, 1001, 306 images, respectively); partial subject data (only end-systolic and end-diastolic images) from all subjects (448 images); and
2. Random sets of images with sizes matching the subsets in (1); each smaller set is a real subset of all larger sets.

When training with subsets, the model is exposed to a smaller number of images in every epoch. We, therefore, increased the number of epochs for the subsets to correct for this effect.

3 | RESULTS

3.1 | Validation of Kaggle labels

The calculated volumes from manual annotations correlate strongly with ground truth values from Kaggle (Pearson correlation coefficient r = 0.98, Supporting Information Figure S2). Most calculated values slightly overestimate the LV volume (Supporting Information Figure S2). DICE scores between manual masks and automatically generated labels using the UKBB model show high agreement with median DICE scores of 0.93 and 0.79 for left ventricle and myocardium, respectively (Supporting Information Figure S3).

3.2 | Framework: deep neural network

3.2.1 | Parameter search

Results of the parameter search are illustrated in Figure 1, showing the absolute distance between the EF predictions

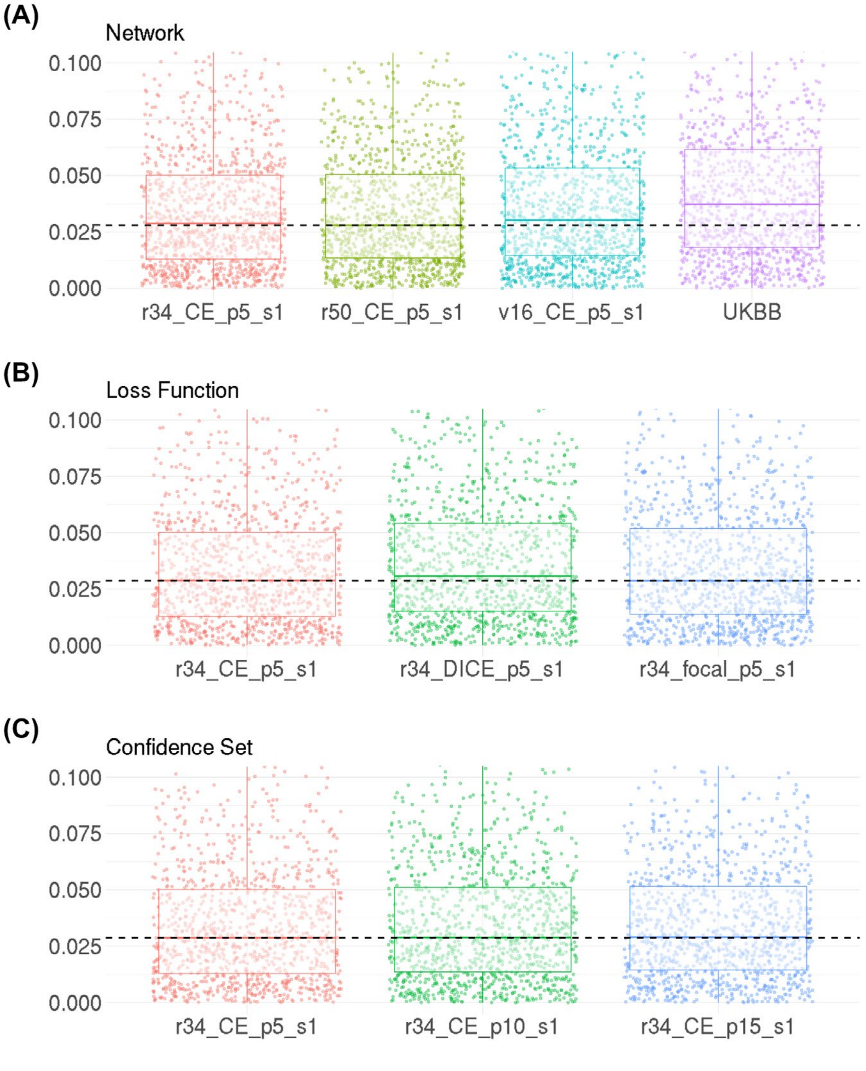

FIGURE 1 Model evaluation during incremental parameter search. Plots show the absolute distance between the EF prediction based on model segmentation and ground truth data provided by Kaggle. The range of the y-axis is restricted for better comparability, dashed lines indicates lowest median. Model performance with A: Architectures (r34: ResNet34, r50: ResNet50, VGG16: v16, UKBB). B: Loss functions (Cross-entropy: CE, DICE, focal loss). C: Confidence sets (p5: 5%, p10: 10%, p15: 15%)

based on model segmentation and ground truth data provided by Kaggle. Overall, the impact of parameter variation on model performance was small (3.64%-4.06% mean distance to ground truth EF).

In a first approach to interpret these results, we compared varying architectures, such as ResNet34, ResNet50, and VGG16, with the UKBB model (Figure 1A). All models led to lower mean and median distance values compared to the UKBB model (Table 2). The lowest median distance values were found using a ResNet50 (2.79%), whereas the lowest mean distance values were found using a ResNet34 (3.64%). Differences in the absolute distance between the models (r34, r50) were rather small (Δ0.08%), however. Considering computational demand, we selected the ResNet34.

TABLE 2 Summary statistics of absolute deviation of predicted and true EF in % for the parameter search

Model	Mean	SD	Median	IQR
r34_CE_p5_s1	3.64	3.38	2.87	3.72
r50_CE_p5_s1	3.71	3.73	2.79	3.70
r34_CE_p15_s1	3.73	3.38	2.91	3.72
r34_focal_p5_s1	3.75	3.90	2.86	3.80
r34_CE_p10_s1	3.77	4.44	2.89	3.76
r34_DICE_p5_s1	3.93	3.43	3.07	3.91
v16_CE_p5_s1	4.06	4.94	3.02	3.87
UKBB	5.42	8.83	3.72	4.34

Sorted from lowest to highest mean value. Models are named by architecture (ResNet34: r34, ResNet50: r50, and VGG16: v16), loss function (cross entropy: CE, focal, DICE), confidence set (p5, p10, p15), and data augmentation (s1, s2).

In the next step of the parameter search, we evaluated model performance using varying loss functions, namely cross-entropy, generalized DICE, and focal loss (Figure 1B). Using the generalized DICE score led to the highest mean (3.93%) and median (3.07%) distance values. Median distance values were similar for cross-entropy and focal loss (2.87% vs 2.86%), whereas the mean distance value was lowest using cross-entropy (3.64%).

We therefore selected cross-entropy for the next step of the parameter search, where we evaluated model performance using varying confidence sets: 5%, 10%, 15% (Figure 1C). Using the various confidence sets only slightly affected median distance values (2.87%, 2.89%, 2.91%). Based on EF predictions, the model: r34_CE_p5_s1 performed best, achieving a mean distance value of 3.64%.

3.2.2 | Data augmentation

Figure 2 shows the performance of our models on the image set containing rotated images, plus the performance of an additional model where data augmentation allowed left-right flips, as well as rotations of up to 90°. Median and mean absolute distance values were lowest (3.06%, 4.08%) using the model with extended data augmentation (r34_CE_p5_s2).

3.3 | Transfer learning

Representative cine images from the Kaggle and the 7T cine data set as well as respective data augmentation are shown in

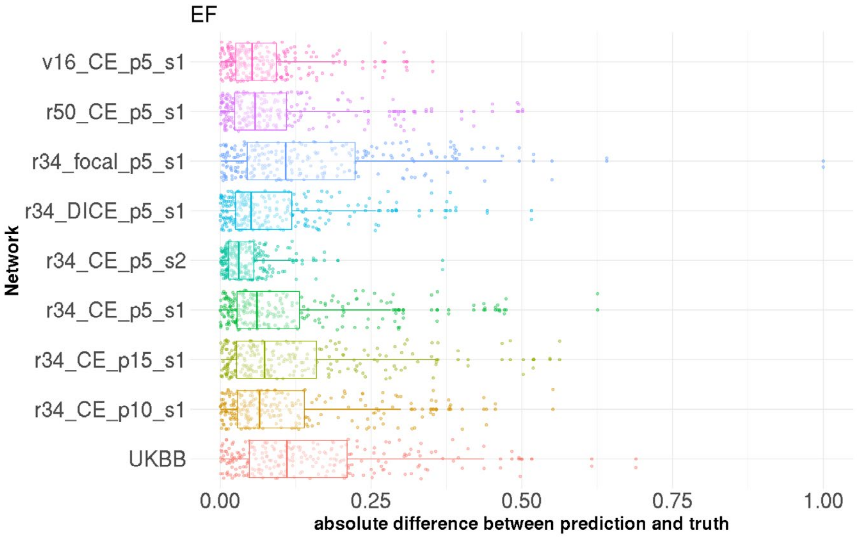

FIGURE 2 Model evaluation based on data including rotated images. Plots show the absolute distance between the EF predictions based on model segmentation and ground truth data provided by Kaggle for all models of the parameter search, plus 1 model trained with extended data augmentation (s2). Models are named by architecture (ResNet34: r34, ResNet50: r50, VGG16: v16), loss function (cross entropy: CE, focal, DICE), confidence set (p5, p10, p15), and data augmentation (s1: standard data augmentation, s2: extended data augmentation, enabling LR-flips and rotations up to 90°)

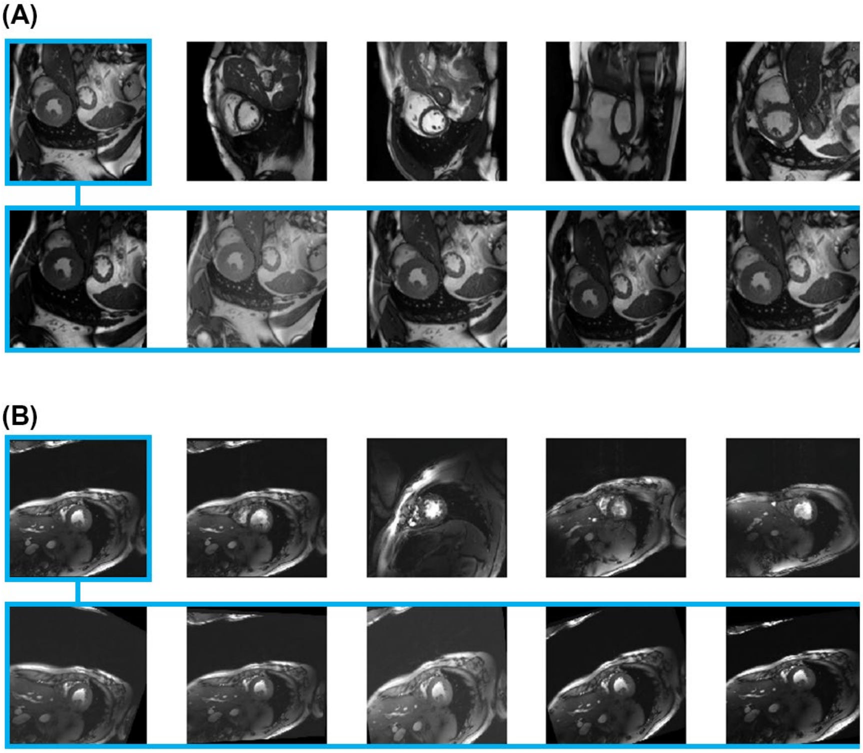

FIGURE 3 Representative cine images and respective data augmentation. Random selection of 5 images (top) with 5 data augmentation examples (bottom) for the first image of the random selection. A: Kaggle data. B: 7T human cine data

Figure 3A,B. Although the Kaggle data set includes images with varying field of views and resolution, the 7T data are consistent.

Figure 4 presents the inter-observer variability as difference of LV volume within each image in ml for all slices and phases of the 7T human cine test set (n = 3). The slice count starts with 0 at the most apical slice and moves toward the most basal slice with increasing slice number. Overall expert 2 achieved $DICE_{LV} = 0.94$ and $DICE_{MY} = 0.81$ and deviations in LV volume of individual images were lower than ±5 mL in all but 1 image (set 3, slice 12, phase 4). Compared to expert one (who labeled our training data) and expert 2, the AI model achieved $DICE_{LV} = 0.90$, $DICE_{MY} = 0.79$ as well as $DICE_{LV} = 0.91$, $DICE_{MY} = 0.81$. Deviations in LV volume of individual images were smaller than 5 mL in >95% of the cases. Representative predictions of the AI and deviations to expert 1 are shown in Figure 5. All apical slices labeled by the AI were in excellent agreement with that of our experts. The largest deviations between AI and both experts was found for the very basal slice where myocardial tissue moves in and out of plane throughout the cardiac cycle.

3.3.1 | Starting point for model training

Results of model training using varying degrees of transfer learning are displayed in Figure 6. The DICE scores for the left ventricle and the myocardium in dependence of the number of images seen during training are plotted, showing performance and overall convergence for the 3 models analyzed. All curves have been smoothed to increase interpretability.

Starting with the full data set, there are clear differences in starting points (DICE after first epoch) and peak performance (highest performance reached) for the 3 models.

R: Random weight initialization followed by training using 7T data led to the:

- lowest starting points with $DICE_{LV}$ ~ 0.57 and $DICE_{MY}$ ~ 0.25, and
- lowest peak performance with $DICE_{LV}$ ~ 0.89 and $DICE_{MY}$ ~ 0.77.

TL: ImageNet weight initialization followed by training using 7T data led to the:

- starting points of $DICE_{LV}$ ~ 0.77 and $DICE_{MY}$ ~ 0.51, and
- higher peak performance with $DICE_{LV:}$ 0.91 and $DICE_{MY}$: 0.79.

TL^2: ImageNet weight initialization, pre-trained (Kaggle data), re-trained 7T data led to the:

- highest starting points with $DICE_{LV}$ ~ 0.90 and $DICE_{MY}$ ~ 0.78, and
- higher peak performance with $DICE_{LV:}$ 0.92 and $DICE_{MY}$: 0.81.

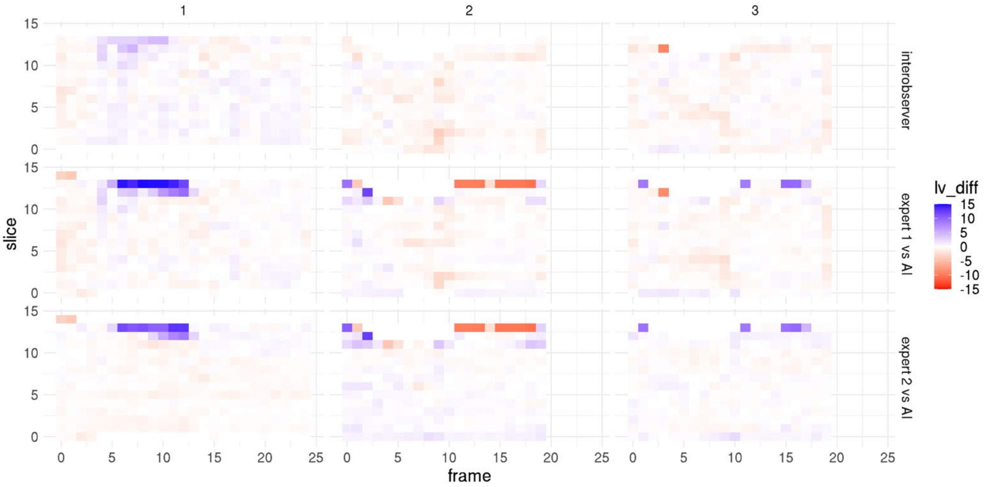

FIGURE 4 Inter-observer variability. Difference in LV volume in [ml] for all slices and phases of the 7T cine images of the test set. The slice count starts with 0 at the most apical slice and moves toward the most basal slice with increasing slice number. Top: inter-observer variability of the 2 experts. Middle: inter-observer variability expert 1 (labeled training data as well) versus AI. Bottom: inter-observer variability expert 2 versus AI

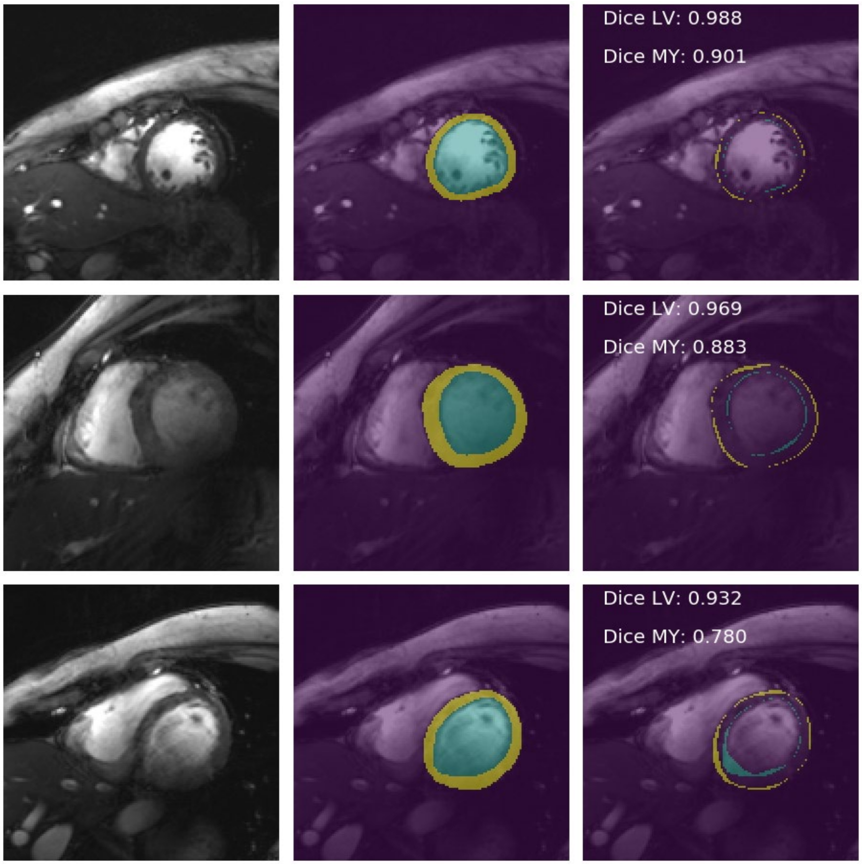

FIGURE 5 Predictions of TL^2 on the 7T human test set. Examples of mid-cavity segmentation results with high (top), intermediate (middle) and low (bottom) DICE scores. Images (left), with predicted classes (middle, background: purple, LV: blue, MY: yellow) and differences to the ground truth (right, LV-error: blue, MY-error: yellow)

3.3.2 | Data requirements for model training

Results of model training using varying degrees of transfer learning and a smaller amount of training data are displayed in Figure 6 as well. The full training data set consists of 14 volunteers, whereas the subsets consist of 7, 3, and 1 volunteer. For the most part, curves follow the trend described for the full data set, although each reduction in volunteers led to lower starting points. Peak performances remain similar with a reduction to 7 volunteers, but drop using subset n3, in particular for models R and TL. Only for a very small number of training images (n1) peak performances are higher for model R ($DICE_{LV}$: 0.86; $DICE_{MY}$: 0.72) compared to TL ($DICE_{LV}$: 0.83; $DICE_{MY}$: 0.70).

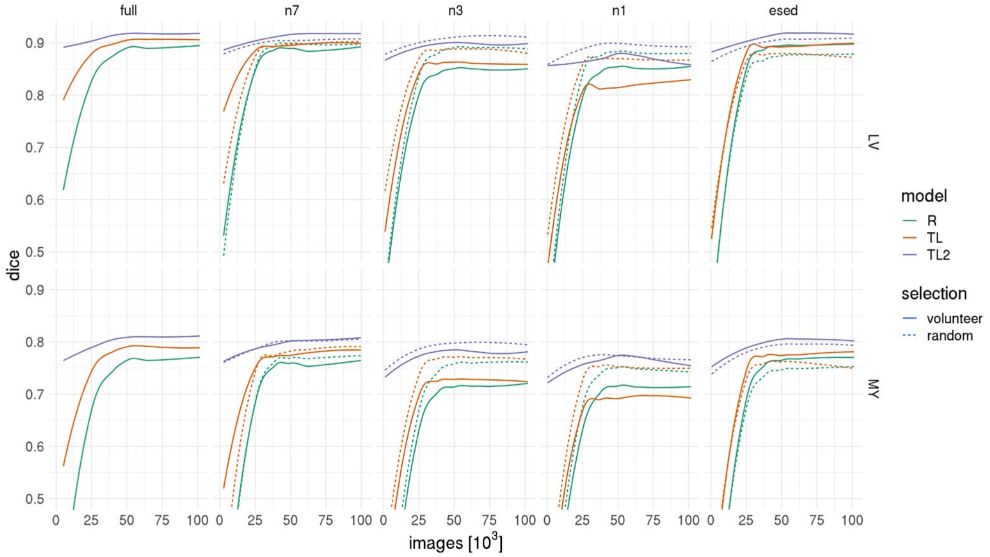

FIGURE 6 Training evaluation based on the validation set. DICE scores of the left ventricle and the myocardium in 7T human cardiac cine images as function of number of images seen during training. Varying degrees of transfer learning (R: ResNet34 initialized with random weights and trained using 7T cine images, TL: ResNet34 initialized with ImageNet weights and trained using 7T cine images, TL^2: ResNet34 initialized with ImageNet weights, pre-trained on the 1.5T and 3T Kaggle cine images and re-trained on 7T cine images) are shown for the 2 subsets 1 (line): subset of whole volunteers (full = 14, 7, 3, 1), 2 (dotted line): subset of random images with image numbers corresponding to first subset. In addition, there is 1 model ("esed") trained using only end-systolic and end-diastolic images from all volunteers and a corresponding model trained with a number of random images equivalent to the "esed"-set

For small subsets, such as n3 and n1, starting points as well as peak performances of all models is higher using the random selection of training images instead of all images from a set (3/1) of volunteers. The same trend is shown for the set n7 using models R and TL.

Using only end-systolic and end-diastolic images led to similar peak performance regarding DICE scores compared to the full data set ($R_{LV,MY}$: 0.90, 0.77; $TL_{LV,MY}$: 0.90, 0.78; $TL^2_{LV,MY}$: 0.92, 0.81 versus $R_{LV,MY}$: 0.89, 0.77; $TL_{LV,MY}$: 0.91, 0.79; $TL^2_{LV,MY}$: 0.92, 0.81). In addition, the selection of end-systolic and end-diastolic images led to increased DICE-scores as starting points and higher peak performance for all models, when compared to the same number of randomly selected images.

4 | DISCUSSION

In this study, we successfully used a specialized, publicly available model[4] to produce labels for a public data set of clinical 1.5 and 3T cardiac cine MRI, enabling access to more annotated data. Based on these labels we created a basic AI model, other researchers can use for their individual segmentation tasks. In addition, we applied transfer learning to segmentation of 7T human cine data, demonstrating that models based on these labels and a moderate amount of new domain data enable state-of-the-art segmentation results.

One of the obstacles to get started in deep learning-based segmentation is the large amount of annotated data required to train an initial model. In this study, we circumvent this problem using the public Kaggle data set to which we provide labels. The quality of these labels was evaluated using the volume information (end-systolic and end-diastolic volumes) included in the original Kaggle data set. Therefore, careful data curation had to be applied to avoid data inconsistencies (slice spacing, changes in image dimensions and image resolution, as well as missing slices) within individual patients. In addition, we found that label quality was connected to image orientation and image resolution. Scores (mean distance between labels and Kaggle "ground truth"), data curation scripts, as well as labels are provided in the online repository, enabling future use in other studies. Limited additional validation has been performed through manual annotation of 816 images from 9 patients. Calculated volumes from these masks correlate strongly with ground truth values but indicate a slight systematic overestimation for the manual volumes (Supporting Information Figure S2). This can be because of different annotation strategies, for example, regarding basal slices and treatment of papillary muscles.[38] Because labeling is consistent within each task, we do not expect this difference to have a negative influence. In the manually validated subset, DICE scores between automatically and manually labeled images show high agreement (Supporting Information Figure S3), confirming that the predicted volumes are because of correct labeling. It also shows that images from phases other than systole and diastole are labeled correctly. The thresholds of 5%, 10%, and 15% (deviation to the "ground truth") for the subsets used in this study were chosen arbitrarily. With 54,540, 162,480, and 239,350 images, respectively, we assumed these 3 sets to provide the reasonable compromise between label accuracy and label quantity needed to assess data requirements in this specific transfer learning application.

Based on the now annotated data, we trained initial segmentation models with varying architectures (ResNet34, ResNet50, and VGG16), varying loss functions (cross-entropy, generalized DICE, focal loss), and varying training sets (p5, p10, and p15). The final model we selected was a ResNet34, using cross-entropy as a loss function and the p5 set for training with an image resolution of 256 × 256. We selected this model based on mean distance to ground truth EF as an indicator of performance. Although it is possible for errors in both systolic and diastolic volumes to cancel out an produce good EF estimates, this case rarely happened in our data. We used EF as it is the most clinically relevant value. Overall impacts of parameter variations (3.64-4.06% mean distance to ground truth EF) were rather small. Similar to the use in this study, researchers or clinicians can use this model as a starting point for their respective transfer learning applications.

Considering the performance of this model on 7T human cine data, generalization capability appears limited. This is also true for the UKBB model. As the authors[4] point out, the UKBB model was "trained on a single data set, the UK Biobank data set, which is a relatively homogenous data set" and might therefore "not generalize well to other vendor or sequence data sets." More extensive data augmentation might lead to better generalization, potentially at the expense of performance on the specific data set of interest. The state of generalization and model suitability for a specific data set can be monitored through sensitivity analysis.[39] In addition, it emphasizes why improvements in generalization[40-42] are needed and why we applied an additional step of transfer learning to 7T data.

Because of differences in training data, our initial models based on UKBB labels outperformed the UKBB model on the Kaggle data. Although the UKBB model was trained on the homogeneous UKBB data, our models were trained on the heterogeneous Kaggle data itself. In addition, we applied data augmentation with respect to rotations and contrast and used only Kaggle data with the most accurate (top 15%) labels.

Although multiple studies[4,5,43] have demonstrated great image segmentation results for 1 specific data set, these models have not been tested on other data sets or initially lack generalization capability. In this study, we show that transfer learning leads to improved model performance. DICE scores achieved on 7T human cine data before and after transfer learning were comparable to human inter-observer variability and is within the range of state-of-the-art results, despite the relatively small set of training data.[19] In addition, inter-observer-variability in EDV (3.5%) and ESV (10.5%) between our model and the expert radiologist are in good agreement with literature reports (EDV: 2.5%-5.3%, ESV: 6.8%-13.9%)[44] based on SSFP CMR imaging.

Typically, segmentation of the left ventricle is done to evaluate ejection fraction, a clinically used parameter. In this study, we show that the model-based volume prediction on the test set is very accurate for apical, mid-cavity, and basal slices, with the exception of the most basal slice, where myocardial tissue moves in and out of plane throughout the cardiac cycle. Because we do not have a "ground-truth" segmentation for the Kaggle data, and no information on labeling protocols, we do not know if there is any consistency in the definition of basal slices or the inclusion or exclusion of papillary muscle.

Although transfer learning allows models to adapt to similar tasks and new data sets containing new characteristics and patterns, this step also requires new labels. This aspect is often a limitation, because labeled medical data are difficult to acquire, particularly in areas that require domain-specific knowledge. In addition, the manual labeling process for high quality segmentations itself is often tedious and labor intensive. In this study, we show that transfer learning applications (ImageNet weights to Kaggle data to 7T data) for cardiac cine segmentation of human 7T data can provide state-of-the-art results when training with labeled data from 7 to 14 volunteers. Having labels for 3 volunteers leads to decent results. We consider labels for only 1 volunteer to be insufficient.

For small training data sets ($n \leq 1001$), we show that a random selection of images from multiple volunteers leads to better performance compared to the selection of all images from a smaller number of volunteers. Generalization capabilities of a model increase with the amount of variation provided in the training data, and therefore using data from a multitude of patients or volunteers, where morphology and therefore image content and contrast differ, may be more beneficial than providing the same number of more coherent images from a small number of volunteers. Furthermore, we demonstrate that the number of required images can drastically be reduced (from 5076 to 448 images), using labeled data from specific heart phases, end-diastolic, and end-systolic instead of all images. This may be possible because knowing the 2 extreme states of contraction the model can deal more easily with intermediate states. Considering that $n = 448$ images (roughly equivalent to 2 full cardiac EF examinations) enable close to state-of-the-art results for cardiac cine segmentation, data requirements for transfer learning applications in closely related tasks are low. In addition, labels for end-diastolic and end-systolic images are created in routine clinical cardiac examinations and therefore easily accessible. It is important to note, however, that data requirements depend on the heterogeneity of the target data set. More data are required when training on more heterogeneous data sets (e.g., when training segmentation models on a data set covering multiple diseases or broad age ranges). Although using only ES and ED frames worked really well in our case, where all 7T data were measured in healthy (circular shape of the LV) volunteers, this might not be true in cases of cardiac disease. In such cases, it may be beneficial to introduce more data variation with respect to the shape of the LV.

In summary, how much and which kind of data should be included in the transfer learning process should be carefully considered before labeling new data. In particular, the notion to provide data sequentially by individual may result in higher data requirements than necessary. There are various other routine cardiac MR examinations such as T_2, T_1, LGE, and even T_2^* that require segmentation.[41,42,45] Transfer learning applications to image segmentation of such varying contrasts may benefit from the amount of annotated data and the framework provided in this study.

With respect to future use of this annotated data, we recommend researchers take the following steps:

1. use the pre-trained model we provide (r34_CE_p5_s2),
2. re-train with training data from the new domain and tune hyper parameters using validation data from the new domain, and
3. evaluate model performance on a test set from the new domain.

In this study, we used only the 5%-15% of the most accurate kaggle labels to create our base models. Therefore, researchers attempting to train their own base network using the labeled Kaggle data should always assess label quality.

4.1 | Limitations

The experimental 7T data used in this study is not comparable to clinical cardiac MRI in patients. Future performances on clinical data should be evaluated against the Kaggle data set.

There are some limitations connected to the use of the Kaggle data set. Although there are variations in measurement parameters, such as resolution, FOV, matrix size, TE, TR, bandwidth, and slice thickness, most examinations (~90%) were done at 1.5T. In addition, all data were acquired using Siemens whole body MRI systems. Models trained using this data set might therefore not generalize well to other vendor data sets, requiring transfer learning as demonstrated in this study.

Because no disease-related information is provided in the Kaggle data set, we have no knowledge which and how many pathological patterns are currently represented in the data set. In this study, we demonstrate that transfer learning to 7T data of healthy human volunteers enables DICE scores of DICE_{LV}: 0.92 and DICE_{MY}: 0.81. A clinical application would require a performance assessment or transfer learning for specific cardiac pathologies, both beyond the scope of this cardiology-related methodological work.

Furthermore, the accuracy of the labels we created was assessed based on comparison to provided volume information only, and confirmation through manual annotations may be biased because we do not know if the provided volume information is based on consistent definitions of basal slices or the inclusion or exclusion of papillary muscle. This should be considered when creating models based on this data set. Additional automatic quality control could be applied by classifying volume curves as normal or abnormal.[18] In general, there is a need for a standard benchmark data set, where labels are based on standardized protocols and images are representations of diverse clinical phenotypes (diseases, vendors, field strengths, sequences, and protocols).

Our transfer learning task involved only data from a single MRI vendor (Siemens Healthcare GmbH). Therefore, it is uncertain whether transfer learning from the base network trained with Siemens data to data from another vendor requires a comparable amount of training data. However, we assume that differences in contrast because of sequences (TRUFI versus FLASH) and field strength (1.5T, 3T, and 7T) are more pronounced than differences between vendors at the same field strength.

5 | CONCLUSIONS

In this study, we provide access to annotated cardiac cine MRI data, and AI models, which can be used as a starting point for transfer learning applications. Using such a base model, we demonstrate that transfer learning from clinical 1.5T and 3T cine data to 7T cine data are feasible with moderate data requirements, potentially enabling future applications to other cardiac MRI examinations such as T_2, T_1, LGE, and even T_2^*. Furthermore, we show that not all data has the same value with respect to transfer learning approaches, and careful selection of the training data may drastically reduce data requirements.

ACKNOWLEDGMENTS
The authors thank Andreas Hotho for insightful discussions. Open access funding enabled and organized by ProjektDEAL. DL, MJA, LMS & TW designed the study. MJA, DL & TW developed the computational methods. DL, WS and TR collected the data. MJA and DL analyzed and interpreted the data. DL wrote the initial draft of the manuscript. All authors read and approved the final manuscript.

CONFLICT OF INTEREST

The Department of TW (Department of Diagnostic and Interventional Radiology, University Hospital, Wuerzburg, Germany) receives a research grant from Siemens Healthcare GmbH.

DATA AVAILABILITY STATEMENT

The data sets supporting the conclusions of this article are available in the zenodo repository, https://doi.org/10.5281/zenodo.3876350 and source code is available in GitHub and zenodo https://github.com/chfc-cmi/cmr-seg-tl, https://doi.org/10.5281/zenodo.3894647. The raw 7T images are not publicly available because of data privacy regulations.

ORCID

Markus Johannes Ankenbrand https://orcid.org/0000-0002-6620-807X
David Lohr https://orcid.org/0000-0002-6509-3776
Theresa Reiter https://orcid.org/0000-0001-8324-1560
Tobias Wech https://orcid.org/0000-0002-2813-7100
Laura Maria Schreiber https://orcid.org/0000-0002-8827-1838

TWITTER

Markus Johannes Ankenbrand @IIMOG

REFERENCES

1. Moon JC, Lorenz CH, Francis JM, Smith GC, Pennell DJ. Breath-hold FLASH and FISP cardiovascular MR imaging: left ventricular volume differences and reproducibility. *Radiology*. 2002;223:789-797.
2. Curtis JP, Sokol SI, Wang Y, et al. The association of left ventricular ejection fraction, mortality, and cause of death in stable outpatients with heart failure. *J Am Coll Cardiol*. 2003;42:736-742.
3. Karamitsos TD, Francis JM, Myerson S, Selvanayagam JB, Neubauer S. The role of cardiovascular magnetic resonance imaging in heart failure. *J Am Coll Cardiol*. 2009;54:1407-1424.
4. Bai W, Sinclair M, Tarroni G, et al. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *J Cardiovas Magn Reson*. 2018;20:65.
5. Baumgartner CF, Koch LM, Pollefeys M, Konukoglu E. An exploration of 2D and 3D deep learning techniques for cardiac MR image segmentation. *arXiv e-prints*. 2017. https://ui.adsabs.harvard.edu/abs/2017arXiv170904496B. Accessed September 01, 2017.
6. Jang Y, Hong Y, Ha S, Kim S, Chang H-J. Automatic segmentation of LV and RV in cardiac MRI. In: Pop M, Sermesant M, Jodoin PM, Zhuang AL, Yang G, Young A, Bernard O, eds. *Statistical Atlases and Computational Models of the Heart*. ACDC and MMWHS Challenges. STACOM 2017. Lecture Notes in Computer Science. Vol 10663. Cham: Springer. 2018:161-169. https://doi.org/10.1007/978-3-319-75541-0_17
7. Tran PV. A fully convolutional neural network for cardiac segmentation in short-axis MRI. *arXiv e-prints*. 2016. https://ui.adsabs.harvard.edu/abs/2016arXiv160400494T. Accessed April 01, 2016.
8. Liu J, Pan Y, Li M, et al. Applications of deep learning to MRI images: a survey. *Big Data Mining and Analytics*. 2018;1:1-18.
9. Chen F, Taviani V, Malkiel I, et al. Variable-density single-shot fast spin-echo MRI with deep learning reconstruction by using variational networks. *Radiology*. 2018;289:366-373.
10. Schlemper J, Caballero J, Hajnal JV, Price AN, Rueckert D. A deep cascade of convolutional neural networks for dynamic MR image reconstruction. *IEEE Trans Med Imaging*. 2018;37:491-503.
11. Zhu B, Liu JZ, Cauley SF, Rosen BR, Rosen MS. Image reconstruction by domain-transform manifold learning. *Nature*. 2018;555:487-492.
12. Benou A, Veksler R, Friedman A, Riklin RT. Ensemble of expert deep neural networks for spatio-temporal denoising of contrast-enhanced MRI sequences. *Med Image Anal*. 2017;42:145-159.
13. Bermudez C, Plassard AJ, Davis TL, Newton AT, Resnick SM, Landman BA. Learning implicit brain MRI manifolds with deep learning. Proceedings of SPIE--the International Society for Optical Engineering. 2018, p. 10574.
14. de Vos BD, Berendsen FF, Viergever MA, Sokooti H, Staring M, Isgum I. A deep learning framework for unsupervised affine and deformable image registration. *arXiv e-prints*. 2018. https://ui.adsabs.harvard.edu/abs/2018arXiv180906130D. Accessed September 01, 2018.
15. Wu G, Kim M, Wang Q, Munsell BC, Shen D. Scalable high-performance image registration framework by unsupervised deep feature representations learning. *IEEE Trans Biomed Eng*. 2016;63:1505-1516.
16. Akkus Z, Galimzianova A, Hoogi A, Rubin DL, Erickson BJ. Deep learning for brain MRI segmentation: state of the art and future directions. *J Digit Imaging*. 2017;30:449-459.
17. Hesamian MH, Jia W, He X, Kennedy P. Deep learning techniques for medical image segmentation: achievements and challenges. *J Digit Imaging*. 2019;32:582-596.
18. Ruijsink B, Puyol-Antón E, Oksuz I, et al. Fully automated, quality-controlled cardiac analysis from CMR: validation and large-scale application to characterize cardiac function. *JACC Cardiovasc Imaging*. 2019;13:684-695.
19. Chen C, Qin C, Qiu H, et al. Deep learning for cardiac image segmentation: a review. *arXiv e-prints*. 2019. https://ui.adsabs.harvard.edu/abs/2019arXiv191103723C. Accessed November 01, 2019.
20. Liu F, Shen C. Learning deep convolutional features for MRI based Alzheimer's disease classification. *arXiv e-prints*. 2014. https://ui.adsabs.harvard.edu/abs/2014arXiv1404.3366L. Accessed April 01, 2014.
21. Pinaya WHL, Gadelha A, Doyle OM, et al. Using deep belief network modelling to characterize differences in brain morphometry in schizophrenia. *Sci Rep*. 2016;6:38897.
22. Bello GA, Dawes TJW, Duan J, et al. Deep-learning cardiac motion analysis for human survival prediction. *Nat Mach Intell*. 2019;1:95-104.
23. Dawes TJW, de Marvao A, Shi W, et al. Machine learning of three-dimensional right ventricular motion enables outcome prediction in pulmonary hypertension: a cardiac MR imaging study. *Radiology*. 2017;283:381-390.
24. Lundervold AS, Lundervold A. An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift für Medizinische Physik*. 2019;29:102-127.
25. Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. 2015;12:e1001779.

26. Zhu J-Y, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. 2017;2242-2251.
27. Gatys LA, Ecker AS, Bethge MJA. A neural algorithm of artistic style. 2015;abs/1508.06576.
28. Lohr D, Terekhov M, Kosmala A, Stefanescu MR, Hock M, Schreiber LM. Cardiac MRI with the Siemens Terra 7T system: initial experience and optimization of default protocols. Paper presented at: Proc. of the 26th Annual Meeting of ISMRM; April, 2018; Paris, France.
29. Paszke A, Gross S, Massa F, et al. PyTorch: an imperative style, high-performance deep learning library. *arXiv e-prints*. 2019. arXiv:1912.01703. https://ui.adsabs.harvard.edu/abs/2019arXiv191201703P. Accessed December 01, 2019.
30. Howard J, Gugger S. Fastai: a layered API for deep learning. *Information*. 2020;11:108.
31. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A, eds. *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015*. MICCAI 2015. Lecture Notes in Computer Science, Vol 9351. Cham: Springer; 2015. https://doi.org/10.1007/978-3-319-24574-4_28
32. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *arXiv e-prints*. 2015:arXiv:1512.03385. https://ui.adsabs.harvard.edu/abs/2015arXiv151203385H. Accessed December 01, 2015.
33. Abbasi-Sureshjani S, Amirrajab S, Lorenz C, Weese J, Pluim J, Breeuwer M. 4D semantic cardiac magnetic resonance image synthesis on XCAT anatomical model. *arXiv e-prints*. 2020. https://ui.adsabs.harvard.edu/abs/2020arXiv200207089A. Accessed February 01, 2020.
34. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv e-prints*. 2014:arXiv:1409.1556. https://ui.adsabs.harvard.edu/abs/2014arXiv1409.1556S. Accessed September 01, 2014.
35. Smith LN. A disciplined approach to neural network hyper-parameters: Part 1 -- learning rate, batch size, momentum, and weight decay. *arXiv e-prints*. 2018:arXiv:1803.09820. https://ui.adsabs.harvard.edu/abs/2018arXiv180309820S. Accessed March 01, 2018.
36. Sudre CH, Li W, Vercauteren T, Ourselin S, Cardoso MJ. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Cardoso J, Arbel T, Carneiro G, et al. eds. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. DLMIA 2017, ML-CDS 2017. Lecture Notes in Computer Science, Vol 10553. Cham: Springer; 2017. https://doi.org/10.1007/978-3-319-67558-9_28
37. Zijdenbos AP, Dawant BM, Margolin RA, Palmer AC. Morphometric analysis of white matter lesions in MR images: method and validation. *IEEE Trans Med Imaging*. 1994;13:716-724.
38. Suinesiaputra A, Bluemke DA, Cowan BR, et al. Quantification of LV function and mass by cardiovascular magnetic resonance: multi-center variability and consensus contours. *J Cardiovasc Magn Reson*. 2015;17:63.
39. Ankenbrand MJ, Shainberg L, Hock M, Lohr D, Schreiber LM. Sensitivity analysis for interpretation of machine learning based segmentation models in cardiac MRI. *BMC Med Imaging*. 2021;21:27.
40. Feng X, Yang J, Laine AF, Angelini ED. Discriminative localization in CNNs for weakly-supervised segmentation of pulmonary nodules. *arXiv e-prints*. 2017:arXiv:1707.01086. https://ui.adsabs.harvard.edu/abs/2017arXiv170701086F. Accessed July 01, 2017.
41. Chen J, Li H, Zhang J, Menze B. Adversarial convolutional networks with weak domain-transfer for multi-sequence cardiac MR images segmentation. *arXiv e-prints*. 2019. https://ui.adsabs.harvard.edu/abs/2019arXiv190809298C. Accessed August 01, 2019.
42. Wang J, Huang H, Chen C, Ma W, Huang Y, Ding X. Multi-sequence cardiac MR segmentation with adversarial domain adaptation network. *arXiv e-prints*. 2019. https://ui.adsabs.harvard.edu/abs/2019arXiv191012514W. Accessed October 01, 2019.
43. Tran PV. A fully convolutional neural network for cardiac segmentation in short-axis MRI. *arXiv e-prints*. 2016:arXiv:1604.00494. https://ui.adsabs.harvard.edu/abs/2016arXiv160400494T. Accessed April 01, 2016.
44. Luijnenburg SE, Robbers-Visser D, Moelker A, Vliegen HW, Mulder BJM, Helbing WA. Intra-observer and interobserver variability of biventricular function, volumes and mass in patients with congenital heart disease measured by CMR imaging. *Int J Cardiovasc Imaging*. 2010;26:57-64.
45. Vesal S, Ravikumar N, Maier A. Automated multi-sequence cardiac MRI segmentation using supervised domain adaptation. *arXiv e-prints*. 2019. https://ui.adsabs.harvard.edu/abs/2019arXiv190807726V. Accessed August 01, 2019.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the Supporting Information section.

FIGURE S1 Volume and EF distribution in each confidence set

FIGURE S2 Systolic and diastolic volumes of 9 random patients from the Kaggle data set as provided (ground truth) and calculated from manual annotations (manual annotation). The gray dashed line is the diagonal and indicates identical values whereas the solid blue line is a linear regression of the data points ($R^2 = 0.94$ [diastole] and $R^2 = 0.99$ [systole])

FIGURE S3 DICE scores of images from 9 random Kaggle data sets labeled manually by an expert cardiologist compared to automatically generated labels from the ukbb_cardiac network for left ventricular cavity (LV) and myocardium (MY)

TABLE S1 Volume and EF statistics of confidence sets. The mean and standard deviation of end systolic volume (ESF), end diastolic volume (EDV), and ejection fraction (EF) are shown for each of the confidence sets (p05, p10, and p15) and the rest (not part of any confidence set)

How to cite this article: Ankenbrand MJ, Lohr D, Schlötelburg W, Reiter T, Wech T, Schreiber LM. Deep learning-based cardiac cine segmentation: Transfer learning application to 7T ultrahigh-field MRI. *Magn Reson Med*. 2021;86:2179–2191. https://doi.org/10.1002/mrm.28822