Assessing anatomy and function of the heart using 4D cardiac MRI and deep learning

Sander, J.

Publication date
2023

Document Version
Final published version

Assessing anatomy and function of the heart using 4D cardiac MRI and deep learning

Jörg Sander

Assessing anatomy and function of the heart using 4D cardiac MRI and deep learning

Jörg Sander

This research was part of the research project Deep Learning for Medical Image Analysis (DLMedIA), funded by the Dutch Technology Foundation (TTW) with participation of Pie Medical Imaging. The work presented in this thesis was conducted at the Image Sciences Institute, University Medical Center Utrecht, Utrecht, The Netherlands, Informatics Institute, University of Amsterdam, The Netherlands and Department of Biomedical Engineering & Physics, Amsterdam University Medical Center location University of Amsterdam, The Netherlands.

ISBN: 978-94-6419-911-6
Cover design: Foto copyright by Eva Branzini

Printed by Gildeprint.

Assessing anatomy and function of the heart using 4D cardiac MRI and deep learning

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P. P. C. C. Verbeek
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op donderdag 19 october 2023, te 16:00 uur

door

Jörg Sander

geboren te Münster, Duitsland

Promotiecommissie

Promotores:	prof. dr. I. Išgum	AMC-UvA
	prof. dr. T. Leiner	Universiteit Utrecht
Copromotores:	dr. B. D. de Vos	Nostics
	prof. dr. ir. M. A. Viergever	Universiteit Utrecht
Overige leden:	prof. dr. ir. J. J. Sonke	AMC-UvA
	prof. dr. C. G. M. Snoek	Universiteit van Amsterdam
	prof. dr. ir. M. Breeuwer	TU Eindhoven
	prof. dr. H. J. Lamb	Universiteit Leiden
	prof. dr. R. Vliegenthart	Rijksuniversiteit Groningen
	prof. dr. H. A. Marquering	AMC-UvA

Faculteit der Geneeskunde

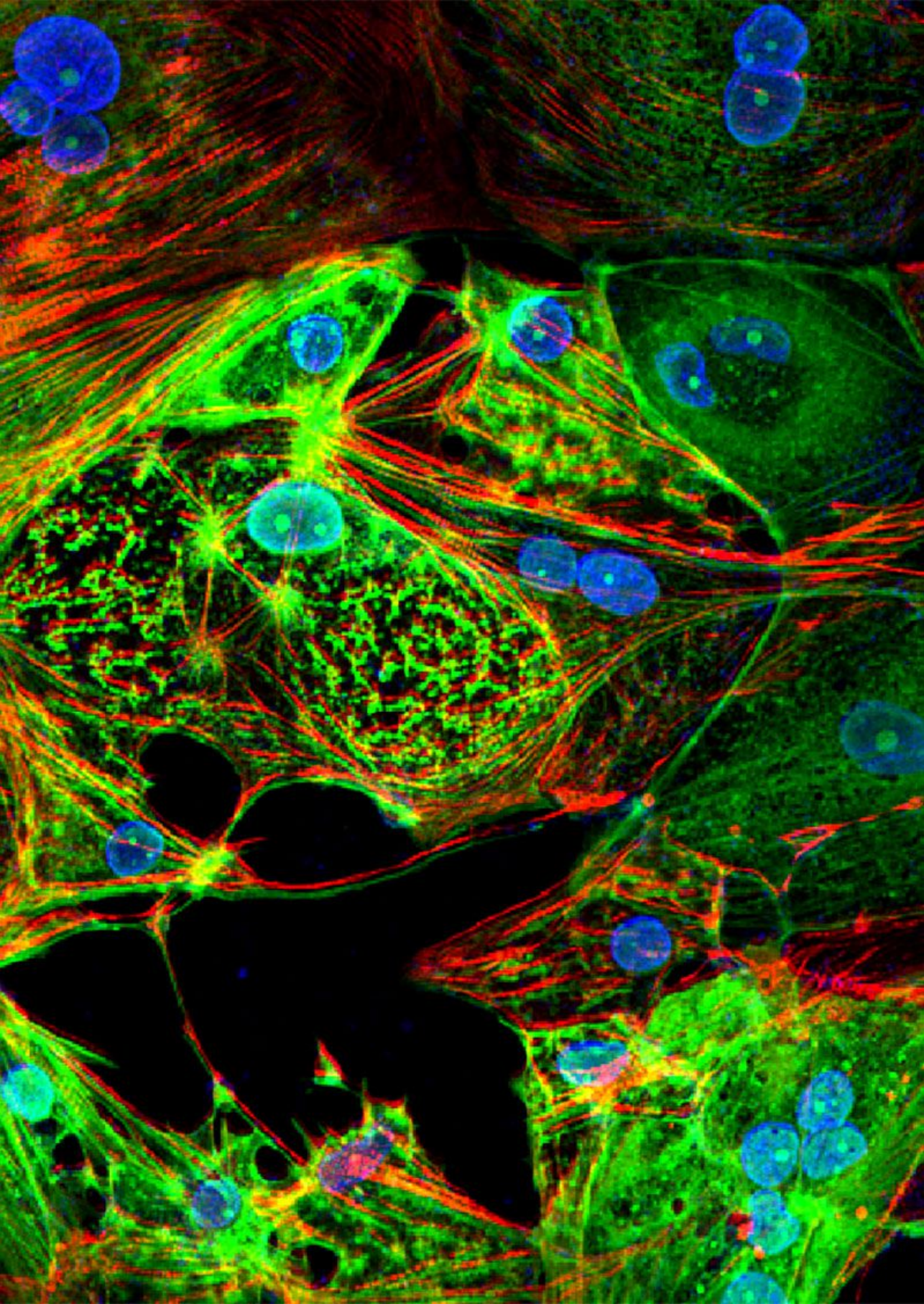
To Elsa († 2018), *my mother*

"So come, my friends, be not afraid
We are so lightly here
It is in love that we are made
In love we disappear."
Leonard Cohen, "Boogie Street"

To Marko Olavi († 2020), *my dear friend*

"And did you get what you wanted from this life, even so?
I did.
And what did you want?
To call myself beloved,
to feel myself beloved on the earth."
Raymond Carver

Contents



CHAPTER 1

Introduction

Illustration (left) copyright by Lina Greenberg

1.1 Introduction

The ability to recognize and respond to health threats is an essential part of the survival and reproductive success of any living organism. Throughout Anthropocene history it seems that homo sapiens have always been concerned about their health. Therefore, we have sought for ways to prevent and treat various health conditions, including herbal remedies, physical activity, and dietary changes. One of the earliest written records of health care is the Egyptian Ebers Papyrus,[1] which dates back to around 1550 BCE. The papyrus contains description of anatomy and function of the human body, the instruments used by doctors, and different diseases and their remedies.[1–3] Furthermore, this document may represent one of the earliest documented observations of the syndrome of heart failure.[3] Meanwhile, according to the 1990-2010 Global Atlas of Cardiovascular Disease,[4] humanity transitioned from a global burden of disease dominated by infectious and maternal conditions to a new world in which cardiovascular disease (CVD) is the leading cause of morbidity and mortality worldwide.

1.2 Cardiovascular magnetic resonance imaging

Medical imaging plays an important role in the diagnosis, prognosis and management of CVDs. Currently, cardiovascular magnetic resonance (CMR) imaging is the reference modality for non-invasive assessment of the morphology and function of the heart[5,6] (see Figure 1.1). CMR imaging is non-invasive and, in contrast to cardiac computed tomography, does not expose the patient to ionizing radiation. Furthermore, the wide variety of soft tissue contrast available on CMR (LGE, T1, T2, lipid-saturation) can be applied to vascular imaging to assess features of vessel wall, inflammation, and atherosclerotic plaque.[5] Cardiac MR images (CMRI) used in this thesis were acquired by using a balanced steady state free precession (balanced SSFP) imaging sequence. CMRIs acquired with a balanced SSFP pulse sequence contain a high contrast-to-noise ratio between the dark myocardium and the bright blood pool. This enables accurate and reproducible assessment of important volumetric (e.g. end-diastolic and end-systolic volumes) and functional parameters (e.g. ejection fraction) of left and right ventricles.[7] The advantages of CMR are particularly valuable for right ventricular functional assessment because its complex shape and position behind the breastbone make it difficult to reliably assess by transthoracic echocardiography.[5] In addition, CMR imaging over time, referred to as cine CMR imaging, enables assessment of cardiac motion. Figure 1.2 shows an example of cardiac CMR short- and long-axis views that are typically acquired in current clinical practice.

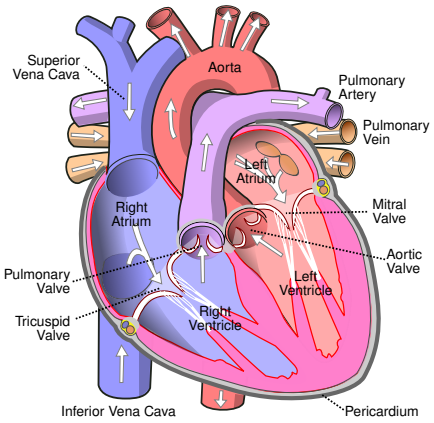
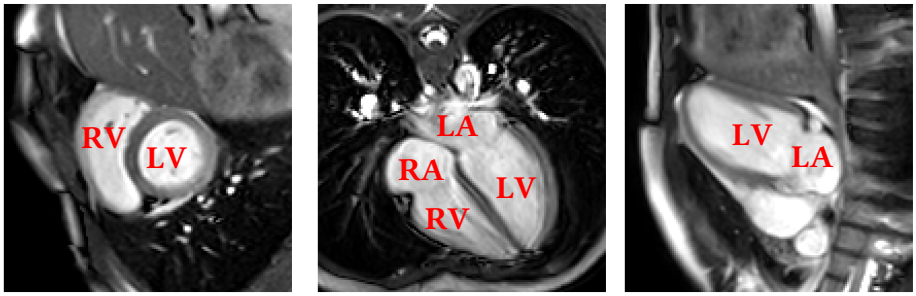


Figure 1.1: Anatomy of the human heart (source Wikipedia)

1.3 Deep learning for automatic cardiac MR image analysis

To compute myocardial mass, volumetric and functional parameters, accurate image segmentation i.e., delineation of myocardial and ventricular structures is essential. Manual segmentation of cine CMRI short-axis volumes is a laborious task, taking about 20 to 30 minutes to segment both ventricles in end-diastole and end-systole.[8] Moreover, manual segmentation across a complete cardiac cycle, comprising 20 to 40 phases per patient, enables computation of parameters quantifying cardiac strain and motion with potential diagnostic implications but due to the required workload, this is practically infeasible. In addition, reproducibility of assessment of cardiac functional indices based on cine CMRI segmentations is hampered by large intra- and inter-observer variability.[9,10] Automatic CMRI segmentation methods may overcome such limitations. Furthermore, to identify new biomarkers for improved stratified diagnosis of cardiomyopathies, automatic CMRI analysis methods, like quality control, segmentation, localization, super-resolution and registration, are required to analyze large-scale multi-center CMRI datasets.

To perform aforementioned complex CMR image analysis tasks automatically, over the past 10 years, deep learning based approaches have become the *de facto* standard.[11] Deep learning in artificial neural networks is a subfield of machine learning which is, in turn, a subfield of artificial intelligence. A deep learning algorithm searches for patterns in data (input e.g., paintings) in order to solve a task e.g., classification of paintings (Vermeer, Rembrandt, Rubens, Hals etc.). Hence, it is not explicitly programmed using handcrafted rules or heuristics. Quintessentially, a deep learning algorithm aims to find a mapping from input to output. Running the deep learning algorithm can be expressed as a function.[12] The function consists of parameters that are adjusted and



(a) (b) (c)

Figure 1.2: Examples of cardiac MRI (CMRI) (a) short-axis view with left (LV) and right (RV) ventricles; (b) 4-chamber long-axis view with LV, RV, left (LA) and right (RA) atrium; (c) 2-chamber long-axis view with LV and LA.

optimized during *training* (learning phase). A deep learning algorithm is *supervised* when trained with labelled i.e., paired input and output data and *unsupervised*, if training is performed on input data only. After training, i.e., at test time, the ability of the algorithm to correctly process new input examples that differ from those used for training is known as *generalization*.[12]

An artificial neural network is a machine learning algorithm that can in principle learn *any* mapping from input to output data.[13] Compared with conventional machine learning approaches that require handcrafted features, deep learning methods automatically extract relevant features from the training data while learning to solve the task at hand. According to Schmidhuber[14] origins of deep learning with convolutional neural networks date back to the early work of Fukushima[15] (1979). The deep learning breakthrough at the beginning of the twenty-first century was mainly caused by digitization and the rapid decline of computation cost in the from of cheap, multi-processor graphics cards[14] (GPU).

Deep learning approaches are well suited to simplify and/or augment every step in the pipeline of cardiac MRI, including optimization of imaging protocols, image acquisition, image reconstruction, image analysis, disease classification, report creation, and derivation of prognostic information.[16] For example, deep learning methods for cardiac MR image analysis have achieved state-of-the-art performance in, e.g., segmentation and disease diagnosis,[17,18] super-resolution,[19,20] landmark detection,[21] motion analysis,[22,23] survival and outcome prediction.[24,25] However, acquisition of cardiac MR images is slow and complicated by cardiac and respiratory motion.[26] To acquire stacks of short-axis 3D cine CMR images simultaneous multi-slice 2D cine CMR imaging is performed under multiple breath-holds. Consequently, in current clinical practice, CMR examination imposes a burden on patients in terms of scan time and breath-holds. To mitigate the risk for motion artifacts and to sustain patient

comfort fast scanning is often required. As a result, short-axis cine CMR scans with high temporal resolution are often highly anisotropic and suffer from respiratory motion induced inter-slice misalignment. Moreover, caused by the low through-plane resolution, short-axis CMR volumes often lack whole-heart coverage predominately at the apex and base of the heart. In addition, recent comparison of a number of state-of-the-art CMRI segmentation methods[17,18] revealed that automatic segmentations are often anatomically implausible. These shortcomings may hamper correct assessment of cardiac anatomy and subsequently hinder accurate analysis of cardiac function.

1.4 Thesis outline

This thesis presents approaches to tackle the aforementioned challenges that can hinder accurate automatic diagnosis and prognosis of cardiovascular diseases using deep learning for automatic cardiac MR image analysis.

CHAPTER 2 presents a method for automatic segmentation of cardiac chambers and left ventricle myocardium in CMRIs. In addition, to investigate the model's trustworthiness, prediction uncertainties were extracted from the model. Moreover, the work examines whether different loss functions effect the reliability of the model. The approach reveals that image areas indicated as highly uncertain, regarding the obtained segmentation, almost entirely cover regions of incorrect segmentations.

Based on these findings, in **CHAPTER 3** an approach is presented that aims to increase reliability of automatic cardiac segmentation methods. The method combines automatic CMRI segmentation with detection of image regions containing local segmentation failures. Highly uncertain predictions were referred to an expert for (simulated) manual correction. Such a human-in-the-loop setting can result in more reliable and increased semi-automatic segmentation performance.

In **CHAPTER 4**, to assess right ventricular function in subjects suspected of arrhythmogenic right ventricular cardiomyopathy, method developed in **CHAPTER 3** is utilized for automatic deep learning CMRI segmentation. The approach is evaluated using the automatic CMRI segmentations for classification of CMR Task Force Criteria.

In current clinical setting, CMR scans with high spatial and temporal resolution are impractical or even impossible to acquire. Hence, accurate assessment of cardiac geometry and function is typically hampered by low through-plane resolution of CMR short-axis images. Therefore, **CHAPTER 5** presents an automatic semantic interpolation approach to increase through-plane resolution of anisotropic CMR short-axis images. The approach can increase spatial resolution by synthesizing new slices in though-plane direction. Furthermore, the model can be trained in an unsupervised fashion i.e., using only low-resolution examples.

Left ventricle segmentations obtained from anisotropic CMRIs lack whole-heart coverage and suffer from motion induced inter-slice misalignment. Consequently, the obtained left ventricle shapes provide limited information about the true 3D cardiac

anatomy. Super-resolution method presented in **CHAPTER 5** cannot alleviate such shortcomings. Hence, **CHAPTER 6** presents a deep learning approach to learn a continuous implicit function representing 3D left ventricle shapes. The proposed method alleviates aforementioned shortcomings by reconstructing and completing of high-resolution 3D cardiac shapes from anisotropic incomplete CMRI segmentations.

The final **CHAPTER 7** provides a general discussion of the presented approaches and most important findings, as well as limitations, and indicates possible future directions.

References

[1] G. Ebers and L. Stern. "Papyros ebers: das hermetische buch über die arzneimittel der alten ägypter in hieratischer schrift," (1875).

[2] J. H. Breasted. "The edwin smith surgical papyrus: published in facsimile and hieroglyphic transliteration with translation and commentary in two volumes," (1930).

[3] M. M. Saba, H. O. Ventura, M. Saleh, and M. R. Mehra. "Ancient egyptian medicine and the concept of heart failure," *Journal of cardiac failure*, vol. 12 (2006), pp. 416–421.

[4] A. E. Moran, G. A. Roth, J. Narula, and G. A. Mensah. "1990-2010 global cardiovascular disease atlas," *Glob Heart*, vol. 9 (2014), pp. 3–16.

[5] T. Leiner, J. Bogaert, M. G. Friedrich, R. Mohiaddin, V. Muthurangu, S. Myerson, A. J. Powell, S. V. Raman, and D. J. Pennell. "SCMR position paper (2020) on clinical indications for cardiovascular magnetic resonance," *Journal of Cardiovascular Magnetic Resonance*, vol. 22 (2020), pp. 1–37.

[6] P. S. Rajiah, C. J. François, and T. Leiner. "Cardiac mri: state of the art," *Radiology* (2023), p. 223008.

[7] F. Grothues, G. C. Smith, J. C. Moon, N. G. Bellenger, P. Collins, H. U. Klein, and D. J. Pennell. "Comparison of interstudy reproducibility of cardiovascular magnetic resonance with two-dimensional echocardiography in normal subjects and in patients with heart failure or left ventricular hypertrophy," *The American journal of cardiology*, vol. 90 (2002), pp. 29–34.

[8] C. P. Corona-Villalobos, I. R. Kamel, N. Rastegar, R. Damico, T. M. Kolb, D. M. Boyce, A.-E. S. Sager, J. Skrok, M. L. Shehata, J. Vogel-Claussen, D. A. Bluemke, R. E. Girgis, S. C. Mathai, P. M. Hassoun, and S. L. Zimmerman. "Bidimensional measurements of right ventricular function for prediction of survival in patients with pulmonary hypertension: comparison of reproducibility and time of analysis with volumetric cardiac magnetic resonance imaging analysis." *Pulmonary circulation*, vol. 5 (3 2015), pp. 527–537.

[9] F. Grothues, J. C. Moon, N. G. Bellenger, G. S. Smith, H. U. Klein, and D. J. Pennell. "Interstudy reproducibility of right ventricular volumes, function, and mass with cardiovascular magnetic resonance." *American heart journal*, vol. 147 (2 2004), pp. 218–223.

[10] T. D. Karamitsos, L. E. Hudsmith, J. B. Selvanayagam, S. Neubauer, and J. M. Francis. "Operator induced variability in left ventricular measurements with cardiovascular magnetic resonance is improved after training," *Journal of Cardiovascular Magnetic Resonance*, vol. 9 (2007), pp. 777–783.

[11] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez. "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42 (2017), pp. 60–88.

[12] C. M. Bishop and N. M. Nasrabadi. "Pattern recognition and machine learning," vol. 4 (Springer, 2006).

[13] K. Hornik, M. Stinchcombe, and H. White. "Multilayer feedforward networks are universal approximators," *Neural networks*, vol. 2 (1989), pp. 359–366.

[14] J. Schmidhuber. "Deep learning in neural networks: an overview," *Neural networks*, vol. 61 (2015), pp. 85–117.

[15] K. Fukushima. "Neural network model for a mechanism of pattern recognition unaffected by shift in position-neocognitron," *IEICE Technical Report, A*, vol. 62 (1979), pp. 658–665.

[16] T. Leiner, D. Rueckert, A. Suinesiaputra, B. Baeßler, R. Nezafat, I. Išgum, and A. A. Young. "Machine learning in cardiovascular magnetic resonance: basic concepts and applications." *Journal of cardiovascular magnetic resonance : official journal of the Society for Cardiovascular Magnetic Resonance*, vol. 21 (1 2019), p. 61.

[17] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester, et al. "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?" *IEEE Transactions on Medical Imaging* (2018).

[18] V. M. Campello, P. Gkontra, C. Izquierdo, C. Martin-Isla, A. Sojoudi, P. M. Full, K. Maier-Hein, Y. Zhang, Z. He, J. Ma, et al. "Multi-centre, multi-vendor and multi-disease cardiac segmentation: the m&ms challenge," *IEEE Transactions on Medical Imaging*, vol. 40 (2021), pp. 3543–3554.

[19] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. De Marvao, T. Dawes, D. P. O'Regan, et al. "Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation," *IEEE transactions on medical imaging*, vol. 37 (2017), pp. 384–395.

[20] Y. Xia, N. Ravikumar, J. P. Greenwood, S. Neubauer, S. E. Petersen, and A. F. Frangi. "Super-resolution of cardiac MR cine imaging using conditional GANs and unsupervised transfer learning," *Medical Image Analysis* (2021), p. 102037.

[21] H. Xue, J. Artico, M. Fontana, J. C. Moon, R. H. Davies, and P. Kellman. "Landmark detection in cardiac MRI by using a convolutional neural network," *Radiology: Artificial Intelligence*, vol. 3 (2021).

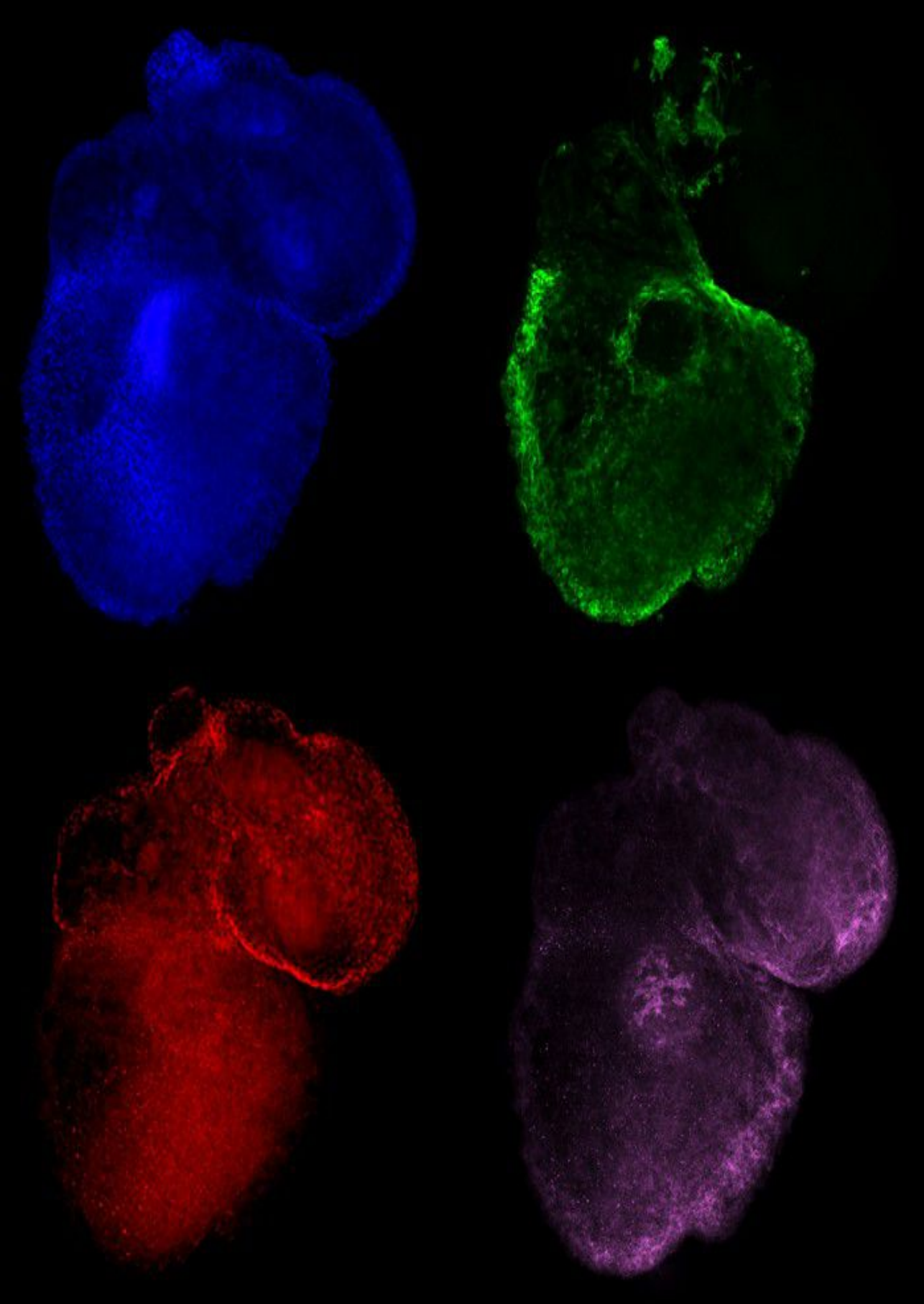
[22] M. A. Morales, M. Van den Boomen, C. Nguyen, J. Kalpathy-Cramer, B. R. Rosen, C. M. Stultz, D. Izquierdo-Garcia, and C. Catana. "Deepstrain: a deep learning workflow for the automated characterization of cardiac mechanics," *Frontiers in Cardiovascular Medicine* (2021), p. 1041.

[23] Q. Meng, C. Qin, W. Bai, T. Liu, A. de Marvao, D. P. O'Regan, and D. Rueckert. "Mulvimotion: shape-aware 3d myocardial motion tracking from multi-view cardiac MRI," *IEEE transactions on medical imaging*, vol. 41 (2022), pp. 1961–1974.

[24] G. A. Bello, T. J. Dawes, J. Duan, C. Biffi, A. de Marvao, L. S. Howard, J. S. R. Gibbs, M. R. Wilkins, S. A. Cook, D. Rueckert, et al. "Deep-learning cardiac motion analysis for human survival prediction," *Nature machine intelligence*, vol. 1 (2019), pp. 95–104.

[25] T. J. Dawes, A. de Marvao, W. Shi, T. Fletcher, G. M. Watson, J. Wharton, C. J. Rhodes, L. S. Howard, J. S. R. Gibbs, D. Rueckert, et al. "Machine learning of three-dimensional right ventricular motion enables outcome prediction in pulmonary hypertension: a cardiac mr imaging study," *Radiology*, vol. 283 (2017), p. 381.

[26] T. Leiner, D. Rueckert, A. Suinesiaputra, B. Baeßler, R. Nezafat, I. Išgum, and A. A. Young. "Machine learning in cardiovascular magnetic resonance: basic concepts and applications," *Journal of Cardiovascular Magnetic Resonance*, vol. 21 (2019), p. 61.



CHAPTER 2

Towards increased trustworthiness of deep learning segmentation methods on cardiac MRI

This chapter is based on: J. Sander, B. D. de Vos, J. M. Wolterink, and I. Išgum. "Towards increased trustworthiness of deep learning segmentation methods on cardiac MRI," *Medical Imaging 2019: Image Processing*, vol. 10949 International Society for Optics and Photonics. (2019), p. 1094919.

Abstract

Current state-of-the-art deep learning segmentation methods have not yet made a broad entrance into the clinical setting in spite of high demand for such automatic methods. One important reason is the lack of reliability caused by models that fail unnoticed and often locally produce anatomically implausible results that medical experts would not make. This paper presents an automatic image segmentation method based on (Bayesian) dilated convolutional networks (DCNN) that generates segmentation masks and spatial uncertainty maps for the input image at hand. The method was trained and evaluated using segmentation of the left ventricle (LV) cavity, right ventricle (RV) endocardium and myocardium (Myo) at end-diastole (ED) and end-systole (ES) in 100 cardiac 2D MR scans from the MICCAI 2017 Challenge (ACDC). Combining segmentations and uncertainty maps and employing a human-in-the-loop setting, we provide evidence that image areas indicated as highly uncertain, regarding the obtained segmentation, almost entirely cover regions of incorrect segmentations. The fused information can be harnessed to increase segmentation performance. Our results reveal that we can obtain valuable spatial uncertainty maps with low computational effort using DCNNs.

2.1 Purpose

Decisions by medical experts are increasingly enriched and augmented by intelligent machines, e.g., through computer aided diagnosis (CAD). The quality of the joint decision process would improve if the automatic systems were able to indicate their uncertainty. This assumes that the provided uncertainty information is reliable i.e., valuable to be considered. A system indicating high uncertainty in image areas of incorrect segmentations could be used to detect and subsequently refer these regions to medical experts. Applying such a human-in-the-loop setting would result in increased segmentation performance. In addition, such a setting could mitigate a severe deficiency of current state-of-the-art deep learning segmentation methods which occasionally generate anatomically implausible segmentations[1] that a medical expert would never make.

Previous research has mainly focused on the assessment of uncertainty in disease prediction,[2] tissue segmentation[3] and pulmonary nodule detection[4] by utilizing Bayesian neural networks (BNN) or test-time data augmentation techniques.[5] Additional methods to estimate uncertainty are Deep Ensembles[6] and Learned Confidence Estimates.[7] In the former multiple models are trained and the variance of their predictions is used as confidence measure, whereas in the latter the model outputs a confidence measure simultaneously with the prediction.

In this work, using multi-structures segmentation in cardiac MR images, we introduce a method that simultaneously generates segmentation masks and uncertainty maps by using a dilated convolutional network (DCNN). We compare two approaches to obtain uncertainty maps. First, we use entropy maps that can be efficiently generated by any probabilistic classifier as entropy is a theoretically grounded quantification of uncertainty in information theory. Second, we employ Bayesian uncertainty maps that are obtained by Bayesian DCNNs (B-DCNN). In addition, we reveal that a valuable uncertainty measure can be obtained if the applied model is *well calibrated*, i.e. if generated probabilities represent the likelihood of being correct. We demonstrate this by simulating a human-in-the-loop setting and provide evidence that image areas indicated as highly uncertain regarding the obtained segmentation almost entirely cover regions of incorrect segmentations. Hence, the fused information can be employed in clinical practice to inform an expert whether and where the generated segmentation should be adjusted.

2.2 Data description

Data from the MICCAI 2017 Challenge on automated cardiac diagnosis (ACDC)[1] was used. The dataset consists of cardiac cine MR images (CMRI) from 150 patients who have been clinically diagnosed in five classes: normal, dilated cardiomyopathy, hypertrophic cardiomyopathy, heart failure with infarction, or right ventricular abnormality. Cases

are uniformly distributed over classes. Manual reference segmentations of the left ventricle (LV) cavity, right ventricle (RV) endocardium and LV myocardium (Myo) at end-diastole (ED) and end-systole (ES) are provided for 100 cases. For each patient, short-axis CMRIs with 28-40 frames are available, in which the ED and ES frame have been indicated. On average images consist of nine slices where each slice has a spatial resolution of 235×263 voxels (on average). The image slices cover the LV from the base to the apex. In-plane voxel spacing varies from 1.37 to 1.68 mm, with slice thickness from 5 to 10 mm and sometimes inter-slice gap of 5 mm. To correct for differences in voxel size, all 2D image slices were resampled to $1.4 \times 1.4\,\text{mm}^2$. Furthermore, to correct for intensity differences among images, each MR volume was normalized between [0.0, 1.0] according to the 5^{th} and 95^{th} percentile of intensities in the image. For detailed specifications about the acquisition protocol we refer the reader to Bernard *et al.*[1]

2.3 Method

To perform segmentation of tissue classes in cardiac 2D MR scans, we used the DCNN developed by Wolterink *et al.*[8] The DCNN architecture comprises a sequence of ten convolutional layers with increasing levels of kernel dilation which results in a receptive field for each voxel of 131×131 voxels, or $18.3 \times 18.3\,\text{cm}^2$. The network has two input channels which take ED and ES slices as its input. We assume that the network leverages cardiac motion differences between ED and ES time points in order to better localize the target structures. To simultaneously segment the LV, RV, LV myocardium and background in ED and ES, the network has eight output channels where each channel provides a probability for one of the classes. Softmax probabilities are calculated over the four tissue classes for images acquired in ED and ES. To enhance generalization performance, the model uses batch normalization and weight decay.

To acquire spatial uncertainty maps of the segmentation during testing, two different approaches were evaluated. First, to obtain entropy maps (e-maps) we computed the multi-class entropy per voxel. Second, to obtain Bayesian uncertainty maps (u-maps), we implemented *Monte Carlo dropout* (MC dropout) introduced by Gal & Ghahramani[9] for approximate Bayesian inference. We added dropout as the last operation in all but the final layer (by randomly switching off 10 percent of a layer's hidden units). By enabling dropout during testing, softmax probabilities are obtained with 10 samples per voxel. As an overall measure of uncertainty we used the maximum softmax variance per voxel over all classes. The variance per voxel per class is obtained from the softmax samples for each class. We chose to use the maximum instead of the mean (as e.g., utilized by Leibig *et al.*[2]) because we found that averaging attenuates the uncertainties.

The quality of e-maps and u-maps depends on the calibration of the acquired probabilities. Previous work[6] revealed that loss functions differ regarding how well the generated probabilities represent the likelihood of being correct. Therefore, we trained the model with three different loss functions: soft-Dice (SD), cross-entropy (CE), and

the Brier score (BS),[10] which is equal to the average gap between softmax probabilities and the references. This provides information about accuracy and uncertainty of the model. Computationally the Brier score loss is equal to the squared error between the one-hot encoding of the correct label and its associated probability.

To use four-fold cross-validation we split the dataset into 75 and 25 training and test patients, respectively. Each model is evaluated on the holdout test images and we report combined results for all 100 patients. During training we used images with 151×151 voxel samples, padded to 281×281 to accommodate the 131×131 voxel receptive field. Training samples were augmented by 90 degree rotations of the images and references. The model was trained for 150,000 iterations using the snapshot ensemble technique described in Huang *et al.*,[11] while after every 10,000th iteration we reset the learning rate to its original value of 0.02 and stored the model. We used mini-batches of size 16 and applied Adam[12] as stochastic gradient descent optimizer. To compare u-maps with e-maps at test time each model was evaluated twice. First, to obtain u-maps we used the last six stored models (iterations 100,000 to 150,000) of each fold to obtain segmentation results. Tissue class per voxel was determined using the mean softmax probabilities over 60 samples (10 samples per voxel per model). In addition, these probabilities served to compute the maximum variance (as described in the beginning of this section). Second, to obtain e-maps we solely employed the last stored model of each fold to acquire segmentation results. We disabled dropout during inference and used one forward pass to compute the softmax probabilities and determine the tissue class per voxel. The corresponding e-maps were computed as the entropy in the four-class probability distribution. Finally, for both evaluations as a post-processing step, the 3D probability volumes were filtered by selection of the largest 3D 6-connected component for each class.

The models were implemented using the PyTorch[13] framework and trained on one Nvidia GTX Titan X GPU with 12 GB memory.

2.4 Results and Discussion

To evaluate whether the obtained per voxel probabilities represent the likelihood of being correct i.e. are well calibrated, we created *Reliability Diagrams*[14] (RD). Figures. 2.1a, 2.1b and 2.1c show the predicted probabilities discretized into ten bins and plotted against the true positive fraction for each bin. If the model is perfectly calibrated, the diagram should match the dashed line. We conclude that a model trained with the soft-Dice loss produces inferior calibrated probabilities compared to the other two loss functions. We conjecture that this could be caused by the relatively low penalty induced by the soft-Dice loss for the model being *overconfident* for true positive tissue labels (see Figure 2.1d).

To compare the quality of the obtained uncertainty maps, we simulate a human-in-the-loop setting. We combine the information of predicted segmentation masks

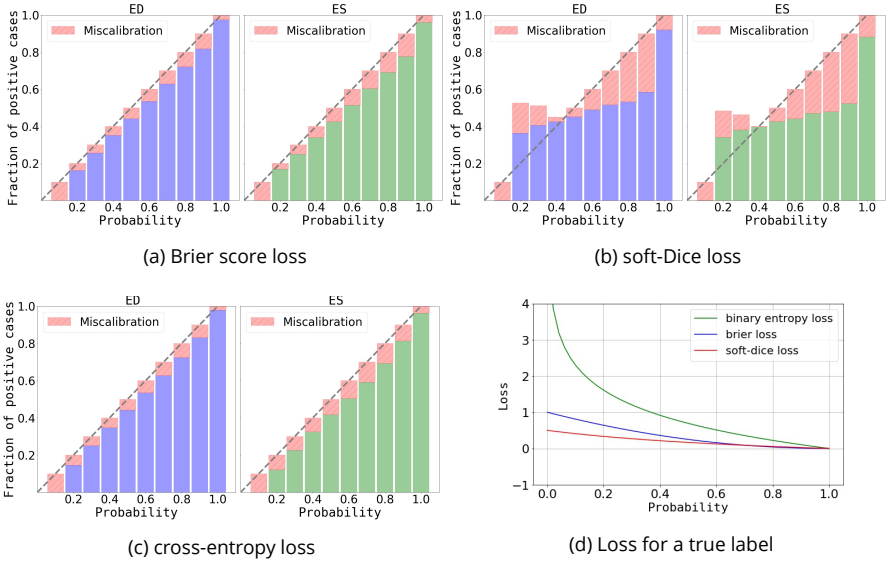
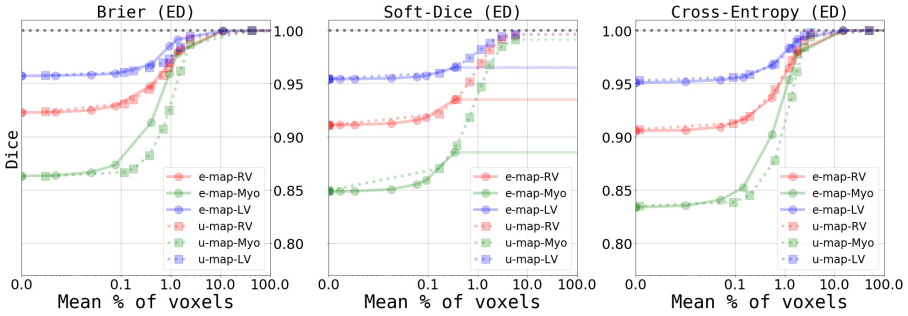
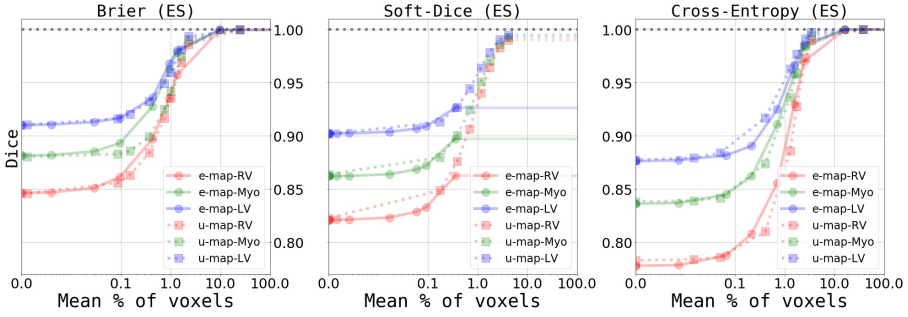


Figure 2.1: Reliability diagrams over all tested ED and ES images and tissue classes for Brier, soft-Dice and cross-entropy loss functions. Blue (end-diastole) and green (end-systole) bars quantify the true positive fraction for each probability bin. Red bars quantify the miscalibration of the model where smaller indicates better. If the model is perfectly calibrated, the diagram should match the dashed line.

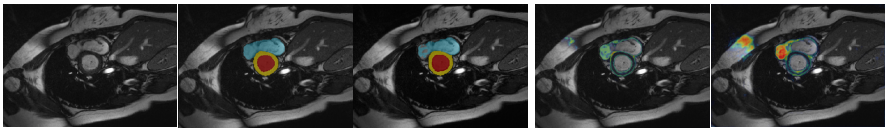


(a) End-diastole



(b) End-systole

Figure 2.2: Comparison between entropy and Bayesian uncertainty maps for different loss functions (RV in red, myocardium in green and LV in blue). Figures visualize Dice score of the corrected segmentation mask when voxels above a tolerated uncertainty or entropy threshold are corrected to their reference label. x-axis shows mean percentage of voxels referred in an image.



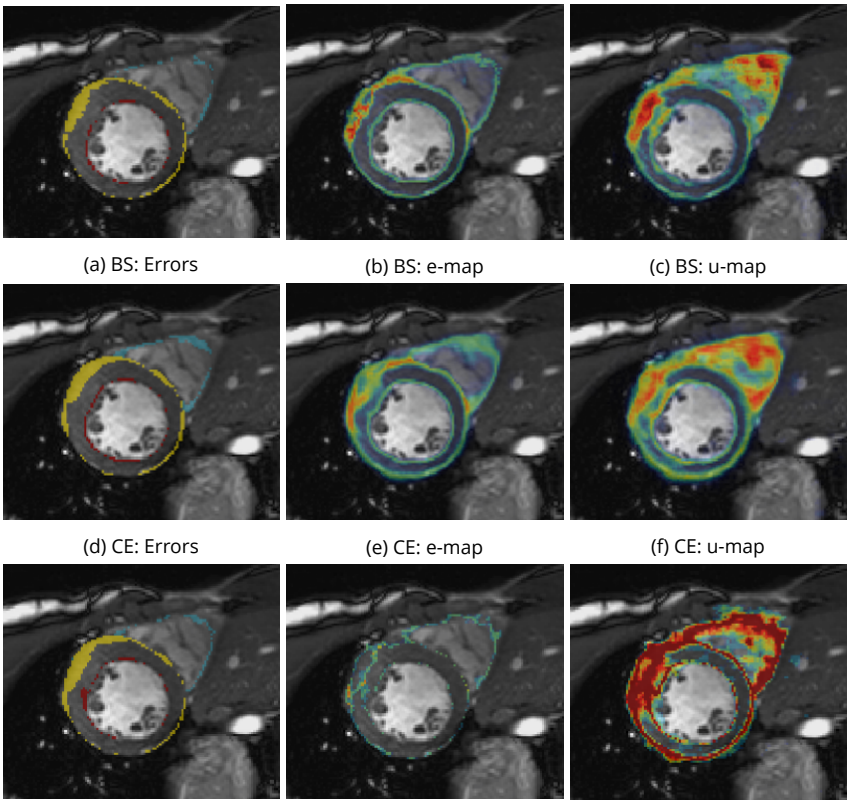
(a) MRI slice (b) Reference (c) Automatic (d) e-map (e) u-map

Figure 2.3: Example of segmentation errors that are covered by high uncertainties. (a) original MRI slice to be segmented; (b) manual reference segmentation; (c) automatic segmentation mask generated by the model when trained with the Brier score loss. Segmentation errors for the right ventricle are covered by (d) high entropy (e-map) and (e) Bayesian uncertainties (u-map).

with the e-maps or u-maps and assume that voxels above a tolerated uncertainty or entropy threshold are corrected to their reference label by an expert. For each threshold we compute the Dice score for the corrected segmentation mask. Figures 2.2a and 2.2b visualize the Dice score as a function of the average percentage of voxels thus referred. We observe a monotonic increase in prediction accuracy when more voxels are referred. E.g., inspecting the referral curves for the Brier score loss in Figure 2.2b we note that referring on average 1% of the voxels in an image, increases performance for 8, 7 and 5% for RV, Myo and LV, respectively. These results are similar for the u-maps and the e-maps. In each experiment, the case in which no voxels are referred for correction is considered the baseline (left most y-axis values). We observe that baseline segmentation performance is highest when the model is trained with the Brier score loss, slightly lower for the soft-Dice, and lowest when cross-entropy is used. Except for the soft-Dice loss we note that u-maps and e-maps follow each other quite closely, which suggests that both carry similar information. Not including the soft-Dice loss, segmentation performance with referral using u-maps or e-maps reaches a Dice score of nearly one when sufficient number of voxels are referred. Hence, we may conclude that areas of uncertainty and entropy almost completely cover the regions of incorrect segmentations[1]. Results obtained after the referral using entropy maps for a model trained with the soft-Dice loss are clearly inferior compared to the performance achieved when using the u-maps. We assume that this is due to the miscalibration of the model (see Figure 2.1b). Compared to e-maps, u-maps tend to exhibit more uncertainty. This is visually expressed for the cross-entropy loss in Figure 2.2a, where the Myo referral-curve obtained with u-maps lags behind the corresponding curve that uses the entropy information.

An example result of the segmentation task performed by a model trained with the Brier score loss is shown in Figure 2.3. The model obviously failed to segment parts of the right ventricle (blue) and we can observe that these errors are covered by entropy and Bayesian uncertainty maps. Figure 2.4 shows a qualitative comparison

[1]Without covering the complete image in which case all voxels would be referred (corresponding to a trivial solution).



(a) BS: Errors (b) BS: e-map (c) BS: u-map

(d) CE: Errors (e) CE: e-map (f) CE: u-map

(g) SD: Errors (h) SD: e-map (i) SD: u-map

Figure 2.4: Comparison of (left column) segmentation errors of left ventricle (red), myocardium (yellow) and right ventricle (blue); (middle column) Entropy maps; and (right column) Bayesian uncertainty maps for the Brier score (BS), cross-entropy (CE) and soft-Dice (SD) loss (per row). High uncertainties correspond to red and low uncertainties to blue colors.

of the uncertainty maps for the three different loss functions (corresponding to rows in the figure) that were used during training. Images in the left column visualize the segmentation errors for the three different tissue types using distinct colors. Although we can observe that the performed errors are roughly the same for the different loss functions, we clearly see significant differences between the uncertainty maps. E.g., when inspecting the e-maps (middle column) we notice that errors with respect to the segmentation of the myocardium are not entirely covered by regions of high uncertainties for a model trained with the soft-Dice loss. In contrast the same regions are almost completely covered by the e-map for a model trained with the Brier score or cross-entropy loss. Furthermore, a model trained with the soft-Dice loss generated u-maps that contain higher uncertainties than u-maps induced by the other two loss functions. We conjecture that this is caused by the miscalibration of the model (see Figure 2.1b) which has a bias towards generating probabilities that are close to zero or one, leading to large softmax variances per voxels (we used 10 samples per voxel). This does not affect the e-maps because we do not sample predictions for these maps during testing. Besides, the provided examples corroborate our earlier finding that the u-maps contain more uncertain, yet often correctly segmented voxels than the e-maps.

2.5 Conclusions

Using a publicly available cardiac cine MRI dataset, we showed that a (Bayesian) dilated CNN trained with the Brier loss produces valuable Bayesian uncertainty and entropy maps. Our results convey that regions of high uncertainty almost completely cover areas of incorrect segmentations. Well calibrated models enable us to obtain useful spatial entropy maps, which can be used to increase the segmentation performance of the model.

References

[1] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester, et al. “Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?” *IEEE Transactions on Medical Imaging* (2018).

[2] C. Leibig, V. Allken, M. S. Ayhan, P. Berens, and S. Wahl. “Leveraging uncertainty information from deep neural networks for disease detection,” *Scientific reports*, vol. 7 (2017), p. 17816.

[3] Y. Kwon, J.-H. Won, B. J. Kim, and M. C. Paik. “Uncertainty quantification using bayesian neural networks in classification: application to ischemic stroke lesion segmentation,” *Medical Imaging with Deep Learning Conference*, 2018.

[4] O. Ozdemir, B. Woodward, and A. A. Berlin. "Propagating uncertainty in multi-stage bayesian convolutional neural networks with application to pulmonary nodule detection," *NIPS Workshop on Bayesian Deep Learning*, 2017.

[5] M. S. Ayhan and P. Berens. "Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks," *Medical Imaging with Deep Learning Conference*, 2018.

[6] B. Lakshminarayanan, A. Pritzel, and C. Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in Neural Information Processing Systems*, 2017, pp. 6402–6413.

[7] T. DeVries and G. W. Taylor. "Learning confidence for out-of-distribution detection in neural networks," *arXiv preprint arXiv:1802.04865* (2018).

[8] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum. "Automatic segmentation and disease classification using cardiac cine MR images," *International Workshop on Statistical Atlases and Computational Models of the Heart*, Springer. 2017, pp. 101–110.

[9] Y. Gal and Z. Ghahramani. "Dropout as a bayesian approximation: representing model uncertainty in deep learning," *International Conference on Machine Learning (ICML)*, 2016, pp. 1050–1059.

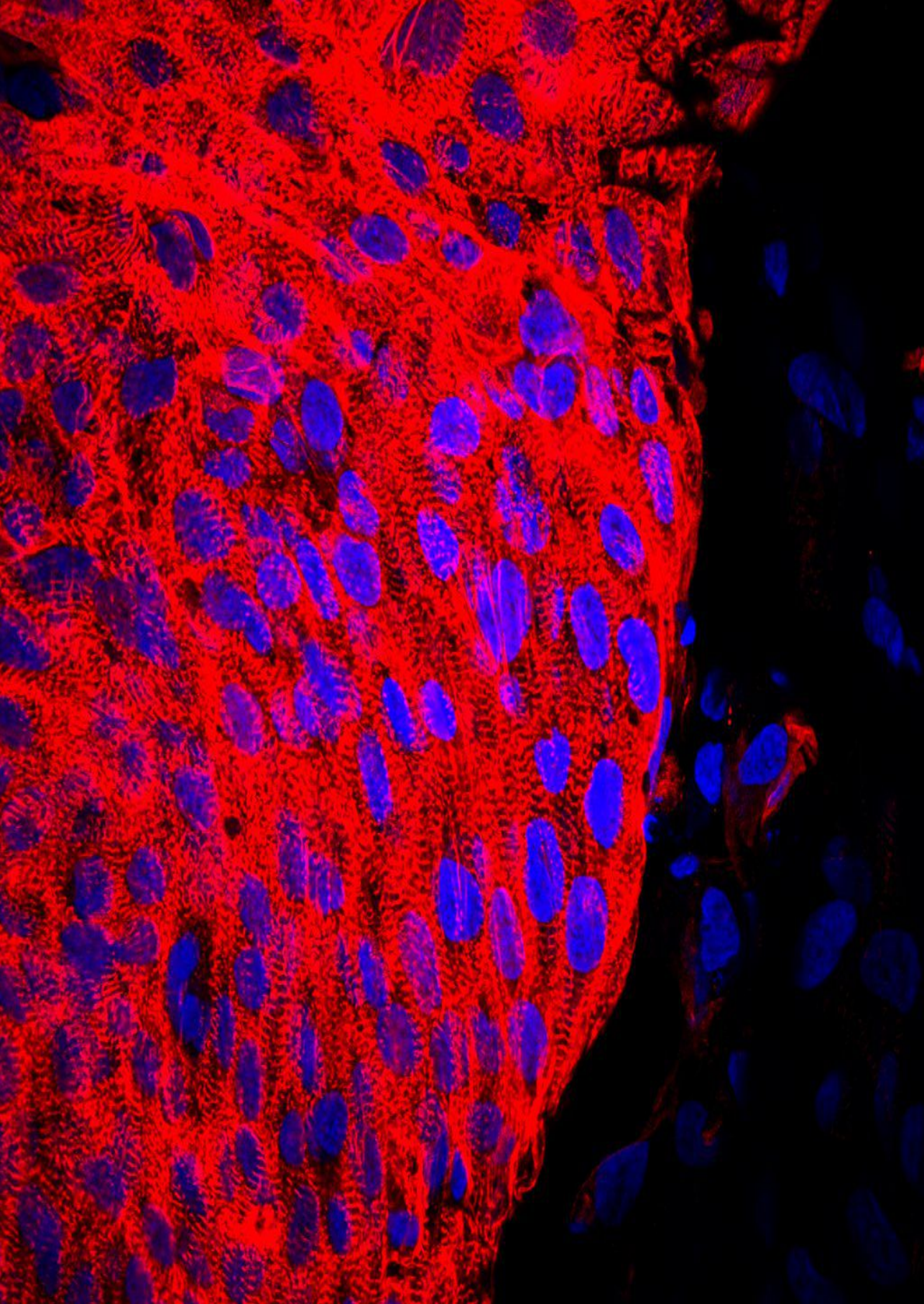
[10] G. W. Brier. "Verification of forecasts expressed in terms of probability," *Monthey Weather Review*, vol. 78 (1950), pp. 1–3.

[11] G. Huang, Y. Li, G. Pleiss, Z. Liu, J. E. Hopcroft, and K. Q. Weinberger. "Snapshot ensembles: train 1, get m for free," *arXiv preprint arXiv:1704.00109* (2017).

[12] D. Kingma and J. Ba. "Adam: a method for stochastic optimization," *ICLR*, vol. 5 (2015).

[13] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. "Automatic differentiation in PyTorch," *NIPS Autodiff Workshop*, 2017.

[14] M. H. DeGroot and S. E. Fienberg. "The comparison and evaluation of forecasters," *The statistician* (1983), pp. 12–22.



CHAPTER 3

Automatic segmentation with detection of local segmentation failures in cardiac MRI

This chapter is based on: J. Sander, B. D. de Vos, and I. Išgum. "Automatic segmentation with detection of local segmentation failures in cardiac MRI," *Scientific Reports*, vol. 10 (2020), pp. 1–19.

Illustration (left) copyright by Richard Mills

Abstract

Segmentation of cardiac anatomical structures in cardiac magnetic resonance images (CMRI) is a prerequisite for automatic diagnosis and prognosis of cardiovascular diseases. To increase robustness and performance of segmentation methods this study combines automatic segmentation and assessment of segmentation uncertainty in CMRI to detect image regions containing local segmentation failures. Three state-of-the-art convolutional neural networks (CNN) were trained to automatically segment cardiac anatomical structures and obtain two measures of predictive uncertainty: entropy and a measure derived by MC-dropout. Thereafter, using the uncertainties another CNN was trained to detect local segmentation failures that potentially need correction by an expert. Finally, manual correction of the detected regions was simulated. Using publicly available CMR scans from the MICCAI 2017 ACDC challenge, the impact of CNN architecture and loss function for segmentation, and the uncertainty measure was investigated. Performance was evaluated using the Dice coefficient and 3D Hausdorff distance between manual and automatic segmentation. The experiments reveal that combining automatic segmentation with simulated manual correction of detected segmentation failures leads to statistically significant performance increase.

3.1 Introduction

To perform diagnosis and prognosis of cardiovascular disease (CVD) medical experts depend on the reliable quantification of cardiac function.[1] Cardiac magnetic resonance imaging (CMRI) is currently considered the reference standard for quantification of ventricular volumes, mass and function.[2] Short-axis CMR imaging, covering the entire left and right ventricle (LV resp. RV) is routinely used to determine quantitative parameters of both ventricle's function. This requires manual or semi-automatic segmentation of corresponding cardiac tissue structures for end-diastole (ED) and end-systole (ES).

Existing semi-automated or automated segmentation methods for CMRIs regularly require (substantial) manual intervention caused by lack of robustness. Manual or semi-automatic segmentation across a complete cardiac cycle, comprising 20 to 40 phases per patient, enables computation of parameters quantifying cardiac motion with potential diagnostic implications but due to the required workload, this is practically infeasible. Consequently, segmentation is often performed at end-diastole and end-systole precluding comprehensive analysis over complete cardiac cycle.

Recently,[3,4] deep learning segmentation methods have shown to outperform traditional approaches such as those exploiting level set, graph-cuts, deformable models, cardiac atlases and statistical models.[5,6] However, recent comparison of a number of automatic methods showed that even the best performing methods generated anatomically implausible segmentations in more than 80% of the CMRIs.[7] Such errors do not occur when experts perform segmentation. To achieve acceptance in clinical practice these shortcomings of the automatic approaches need to be alleviated by further development. This can be achieved by generating more accurate segmentation result or by development of approaches that automatically detect segmentation failures.

In manual and automatic segmentation of short-axis CMRI, largest segmentation inaccuracies are typically located in the most basal and apical slices due to low tissue contrast ratios.[8] To increase segmentation performance, several methods have been proposed.[9–12] Tan *et al.*[9] used a convolutional neural network (CNN) to regress anatomical landmarks from long-axis views (orthogonal to short-axis). They exploited the landmarks to determine most basal and apical slices in short-axis views and thereby constraining the automatic segmentation of CMRIs. This resulted in increased robustness and performance. Other approaches leverage spatial[10] or temporal[11,12] information to increase segmentation consistency and performance in particular in the difficult basal and apical slices.

An alternative approach to preventing implausible segmentation results is by incorporating knowledge about the highly constrained shape of the heart. Oktay *et al.*[13] developed an anatomically constrained neural network (NN) that infers shape constraints using an auto-encoder during segmentation training. Duan *et al.*[14] developed a deep learning segmentation approach for CMRIs that used atlas propagation to

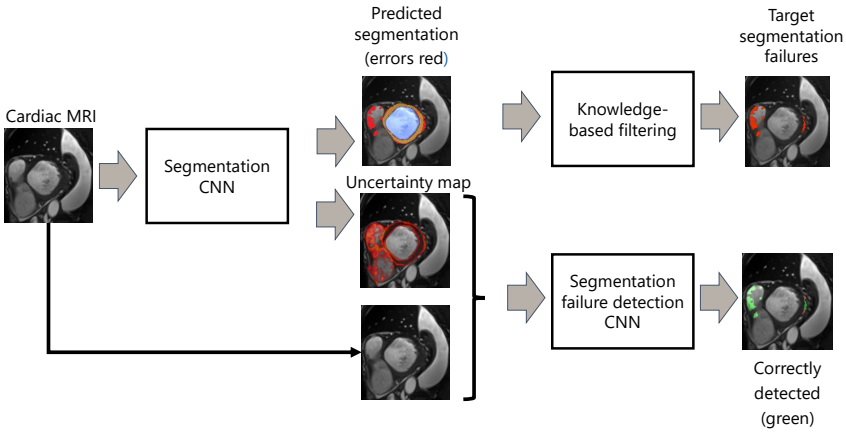


Figure 3.1: Overview of proposed two step approach. Step 1 (left): Automatic CNN segmentation of CMR images combined with assessment of segmentation uncertainties. Step 2 (right): Differentiate tolerated errors from segmentation failures (to be detected) using distance transform maps based on reference segmentations. Detection of image regions containing segmentation failures using CNN which takes CMR images and segmentation uncertainties as input. Manual corrected segmentation failures (green) based on detected image regions.

explicitly impose a shape refinement. This was especially beneficial in the presence of image acquisition artifacts. Recently, Painchaud *et al.*[15] developed a post-processing approach to detect and transform anatomically implausible cardiac segmentations into valid ones by defining cardiac anatomical metrics. Applying their approach to various state-of-the-art segmentation methods the authors showed that the proposed method provides strong anatomical guarantees without hampering segmentation accuracy.

A different research trend focuses on detecting segmentation failures, i.e. on automated quality control for image segmentation. These methods can be divided in those that predict segmentation quality using image at hand or corresponding automatic segmentation result, and those that assess and exploit predictive uncertainties to detect segmentation failure.

Recently, two methods were proposed to detect segmentation failures in large-scale cardiac MR imaging studies to remove these from subsequent analysis.[16,17] Robinson *et al.*[17] using the approach of Reverse Classification Accuracy[18] (RCA) predicted CMRI segmentation metrics to detect failed segmentations. They achieved good agreement between predicted metrics and visual quality control scores. Alba *et al.*[16] used statistical, pattern and fractal descriptors in a random forest classifier to directly detect segmentation contour failures without intermediate regression of segmentation accuracy metrics.

Methods for automatic quality control were also developed for other applications in medical image analysis. Frounchi *et al.*[19] extracted features from the segmentation results of the left ventricle in CT scans. Using the obtained features the authors trained a classifier that is able to discriminate between consistent and inconsistent segmentations. To distinguish between acceptable and non-acceptable segmentations Kohlberger *et al.*[20] proposed to directly predict multi-organ segmentation accuracy in CT scans using a set of features extracted from the image and corresponding segmentation.

A number of methods aggregate voxel-wise uncertainties into an overall score to identify insufficiently accurate segmentations. For example, Nair *et al.*[21] computed an overall score for target segmentation structure from voxel-wise predictive uncertainties. The method was tested for detection of Multiple Sclerosis in brain MRI. The authors showed that rejecting segmentations with high uncertainty scores led to increased detection accuracy indicating that correct segmentations contain lower uncertainties than incorrect ones. Similarly, to assess segmentation quality of brain MRIs Jungo *et al.*[22] aggregated voxel-wise uncertainties into a score per target structure and showed that the computed uncertainty score enabled identification of erroneous segmentations.

Unlike approaches evaluating segmentation directly, several methods use predictive uncertainties to predict segmentation metrics and thereby evaluate segmentation performance.[23,24] For example, Roy *et al.*[23] aggregated voxel-wise uncertainties into four scores per segmented structure in brain MRI. The authors showed that computed scores can be used to predict the Intersection over Union and hence, to determine segmentation accuracy. Similar idea was presented by DeVries *et al.*[24] that predicted segmentation accuracy per patient using an auxiliary neural network that leverages the dermoscopic image, automatic segmentation result and obtained uncertainties. The researchers showed that a predicted segmentation accuracy is useful for quality control.

We build on our preliminary work where automatic segmentation of CMR images using a dilated CNN was combined with assessment of two measures of segmentation uncertainties.[25] For the first measure the multi-class entropy per voxel (entropy maps) was computed using the output distribution. For the second measure Bayesian uncertainty maps were acquired using Monte Carlo dropout[26] (MC-dropout). In our previous work[25] we showed that the obtained uncertainties almost entirely cover the regions of incorrect segmentation i.e. that uncertainties are calibrated. In the current work we extend our preliminary research in two ways. First, we assess impact of CNN architecture on the segmentation performance and calibration of uncertainty maps by evaluating three existing state-of-the-art CNNs. Second, we employ an auxiliary CNN (detection network) that processes a cardiac MRI and corresponding spatial uncertainty map (Entropy or Bayesian) to automatically detect segmentation failures. We differentiate errors that may be within the range of inter-observer variability and hence do not necessarily require correction (tolerated errors) from the errors that an expert would not make and hence require correction (segmentation failures). Given that overlap measures do not capture fine details of the segmentation results and preclude us to

differentiate two types of segmentation errors, in this work, we define segmentation failure using a metric of boundary distance. In our previous work[25] we found that degree of calibration of uncertainty maps is dependent on the loss function used to train the CNN. Nevertheless, in the current work we show that uncalibrated uncertainty maps are useful to detect local segmentation failures. In contrast to previous methods that detect segmentation failure per-patient or per-structure,[23,24] we propose to detect segmentation failures per image region. We expect that inspection and correction of segmentation failures using image regions rather than individual voxels or images would simplify correction process. To show the potential of our approach and demonstrate that combining automatic segmentation with manual correction of the detected segmentation failures per region results in higher segmentation performance we performed two additional experiments. In the first experiment, correction of detected segmentation failures was simulated in the complete data set. In the second experiment, correction was performed by an expert in a subset of images. Using publicly available set of CMR scans from MICCAI 2017 ACDC challenge,[7] the performance was evaluated before and after simulating the correction of detected segmentation failures as well as after manual expert correction.

3.2 Data

In this study data from the MICCAI 2017 Automated Cardiac Diagnosis Challenge (ACDC)[7] was used. The dataset consists of cardiac cine MR images (CMRIs) from 100 patients uniformly distributed over normal cardiac function and four disease groups: dilated cardiomyopathy, hypertrophic cardiomyopathy, heart failure with infarction, and right ventricular abnormality. Detailed acquisition protocol is described by Bernard *et al.*[7] Briefly, short-axis CMRIs were acquired with two MRI scanners of different magnetic strengths (1.5 and 3.0 T). Images were made during breath hold using a conventional steady-state free precession (SSFP) sequence. CMRIs have an in-plane resolution ranging from 1.37 to 1.68 mm (average reconstruction matrix 243 × 217 voxels) with slice spacing varying from 5 to 10 mm. Per patient 28 to 40 volumes are provided covering partially or completely one cardiac cycle. Each volume consists of on average ten slices covering the heart. Expert manual reference segmentations are provided for the LV cavity, RV endocardium and LV myocardium (LVM) for all CMRI slices at ED and ES time frames. To correct for intensity differences among scans, voxel intensities of each volume were scaled to the [0.0, 1.0] range using the minimum and maximum of the volume. Furthermore, to correct for differences in-plane voxel sizes, image slices were resampled to $1.4 \times 1.4\,\text{mm}^2$.

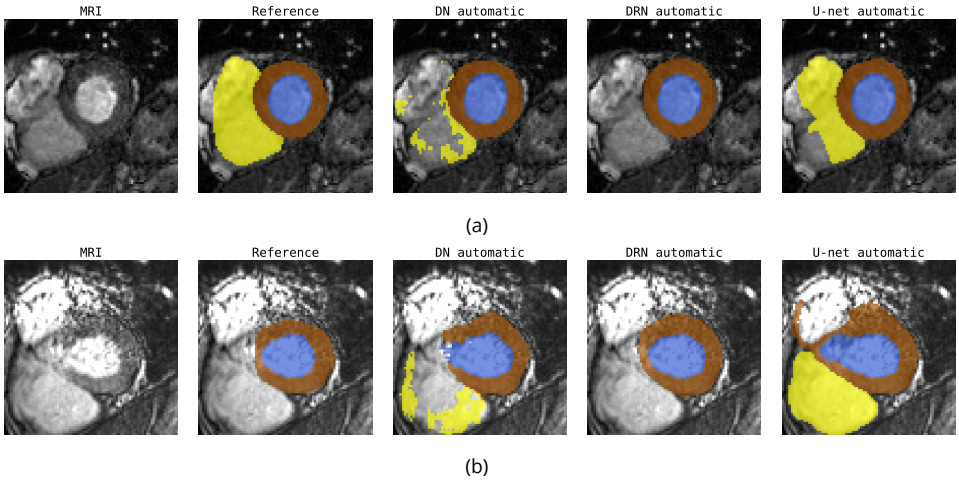


Figure 3.2: Examples of automatic segmentations generated by different segmentation models for two cardiac MRI scans (rows) at ES at the base of the heart.

3.3 Methods

To investigate uncertainty of the segmentation, anatomical structures in CMR images are segmented using a CNN. To investigate whether the approach generalizes to different segmentation networks, three state-of-the-art CNNs were evaluated. For each segmentation model two measures of predictive uncertainty were obtained per voxel. Thereafter, to detect and correct local segmentation failures an auxiliary CNN (detection network) that analyzes a cardiac MRI was used. Finally, this leads to the uncertainty map allowing detection of image regions that contain segmentation failures. Figure 3.1 visualizes this approach.

3.3.1 Automatic segmentation of cardiac MRI

To perform segmentation of LV, RV, and LVM in cardiac MR images i.e. 2D CMR scans, three state-of-the-art CNNs are trained. Each of the three networks takes a CMR image as input and has four output channels providing probabilities for the three cardiac structures (LV, RV, LVM) and background. Softmax probabilities are calculated over the four tissue classes. Patient volumes at ED and ES are processed separately. During inference the 2D automatic segmentation masks are stacked into a 3D volume per patient and cardiac phase. After segmentation, the largest 3D connected component for each class is retained and volumes are resampled to their original voxel resolution. Segmentation networks differ substantially regarding architecture, number of parameters and receptive field size. To assess predictive uncertainties from the segmentation models *Monte Carlo dropout* (MC-dropout) introduced by Gal &

Ghahramani[26] is implemented in every network. The following three segmentation networks were evaluated: Bayesian Dilated CNN, Bayesian Dilated Residual Network, Bayesian U-net.

Bayesian Dilated CNN (DN): The Bayesian DN architecture comprises a sequence of ten convolutional layers. Layers 1 to 8 serve as feature extraction layers with small convolution kernels of size 3×3 voxels. No padding is applied after convolutions. The number of kernels increases from 32 in the first eight layers, to 128 in the final two fully connected classification layers, implemented as 1×1 convolutions. The dilation level is successively increased between layers 2 and 7 from 2 to 32 which results in a receptive field for each voxel of 131×131 voxels, or 18.3×18.3 cm^2. All trainable layers except the final layer use rectified linear activation functions (ReLU). To enhance generalization performance, the model uses batch normalization in layers 2 to 9. In order to convert the original DN[27] into a Bayesian DN, dropout is added as the last operation in all but the final layer and 10 percent of a layer's hidden units are randomly switched off.

Bayesian Dilated Residual Network (DRN): The Bayesian DRN is based on the original DRN from Yu *et al.*[28] for image segmentation. More specifically, the DRN-D-22[28] is used which consists of a feature extraction module with output stride eight followed by a classifier implemented as fully convolutional layer with 1×1 convolutions. Output of the classifier is upsampled to full resolution using bilinear interpolation. The convolutional feature extraction module comprises eight levels where the number of kernels increases from 16 in the first level, to 512 in the two final levels. The first convolutional layer in level 1 uses 16 kernels of size 7×7 voxels and zero-padding of size 3. The remaining trainable layers use small 3×3 voxel kernels and zero-padding of size 1. Level 2 to 4 use a strided convolution of size 2. To further increase the receptive field convolutional layers in level 5, 6 and 7 use a dilation factor of 2, 4 and 2, respectively. Furthermore, levels 3 to 6 consist of two residual blocks. All convolutional layers of the feature extraction module are followed by batch normalization, ReLU function and dropout. Adding dropout and switching off 10 percent of a layer's hidden units converts the original DRN[28] into a Bayesian DRN.

Bayesian U-net (U-net): The standard architecture of the U-net[29] is used. The network is fully convolutional and consists of a contracting, bottleneck and expanding path. The contracting and expanding path each consist of four blocks i.e. resolution levels which are connected by skip connections. The first block of the contracting path contains two convolutional layers using a kernel size of 3×3 voxels and zero-padding of size 1. Downsampling of the input is accomplished by employing a max pooling operation in block 2 to 4 of the contracting path and the bottleneck using a convolutional kernel of size 2×2 voxels and stride 2. Upsampling is performed by a transposed convolutional layer in block 1 to 4 of the expanding path using the same kernel size and stride as the max pooling layers. Each downsampling and upsampling layer is followed by two convolutional layers using 3×3 voxel kernels with zero-padding

size 1. The final convolutional layer of the network acts as a classifier and uses 1×1 convolutions to reduce the number of output channels to the number of segmentation classes. The number of kernels increases from 64 in the first block of the contracting path to 1024 in the bottleneck. In contrast, the number of kernels in the expanding path successively decreases from 1024 to 64. In deviation to the standard U-net instance normalization is added to all convolutional layers in the contracting path and ReLU non-linearities are replaced by LeakyReLU functions because this was found to slightly improve segmentation performance. In addition, to convert the deterministic model into a Bayesian neural network dropout is added as the last operation in each block of the contracting and expanding path and 10 percent of a layer's hidden units are randomly switched off.

3.3.2 Assessment of predictive uncertainties

To detect failures in segmentation masks generated by CNNs in testing, spatial uncertainty maps of the obtained segmentations are generated. For each voxel in the image two measures of uncertainty are calculated. First, a computationally cheap and straightforward measure of uncertainty is the entropy of softmax probabilities over the four tissue classes which are generated by the segmentation networks. Using these, normalized entropy maps $\mathbf{E} \in [\mathbf{0}, \mathbf{1}]^{\mathbf{H}\times\mathbf{W}}$ (e-map) are computed where H and W denote the height and width of the original CMRI, respectively.

Second, by applying MC-dropout in testing, softmax probabilities with a number of samples T per voxel are obtained. As an overall measure of uncertainty the mean standard deviation of softmax probabilities per voxel over all tissue classes C is computed

$$\mathbf{B}(\mathrm{I})^{(\mathrm{x,y})} = \frac{1}{\mathrm{C}} \sum_{\mathrm{c}=1}^{\mathrm{C}} \sqrt{\frac{1}{\mathrm{T}-1} \sum_{\mathrm{t}=1}^{\mathrm{T}} \left(\mathrm{p_t}(\mathrm{I})^{(\mathrm{x,y,c})} - \hat{\mu}^{(\mathrm{x,y,c})}\right)^2}\,, \tag{3.1}$$

where $\mathbf{B}(\mathrm{I})^{(\mathrm{x,y})} \in [0, 1]$ denotes the normalized value of the Bayesian uncertainty map (b-map) at position (x, y) in 2D slice I, C is equal to the number of classes, T is the number of samples and $\mathrm{p_t}(\mathrm{I})^{(\mathrm{x,y,c})}$ denotes the softmax probability at position (x, y) in image I for class c. The predictive mean per class $\hat{\mu}^{(\mathrm{x,y,c})}$ of the samples is computed as follows:

$$\hat{\mu}^{(\mathrm{x,y,c})} = \frac{1}{\mathrm{T}} \sum_{\mathrm{t}=1}^{\mathrm{T}} \mathrm{p_t}(\mathrm{I})^{(\mathrm{x,y,c})}\,. \tag{3.2}$$

In addition, the predictive mean per class is used to determine the tissue class per voxel.

3.3.3 Calibration of uncertainty maps

Ideally, incorrectly segmented voxels as defined by the reference labels should be covered by higher uncertainties than correctly segmented voxels. In such a case the spatial uncertainty maps are perfectly calibrated. *Risk-coverage curves* introduced by Geifman *et al.*[30] visualize whether incorrectly segmented voxels are covered by higher uncertainties than those that are correctly segmented. Risk-coverage curves convey the effect of avoiding segmentation of voxels above a specific uncertainty value on the reduction of segmentation errors (i.e. risk reduction) while at the same time quantifying the voxels that were omitted from the classification task (i.e. coverage).

To generate risk-coverage curves first, each patient volume is cropped based on a minimal enclosing parallelepiped bounding box that is placed around the reference segmentations to reduce the number of background voxels. Note that this is only performed to simplify the analysis of the risk-coverage curves. Second, voxels of the cropped patient volume are ranked based on their uncertainty value in descending order. Third, to obtain uncertainty threshold values per patient volume the ranked voxels are partitioned into 100 percentiles based on their uncertainty value. Finally, per patient volume each uncertainty threshold is evaluated by computing a coverage and a risk measure. Coverage is the percentage of voxels in a patient volume at ED or ES that is automatically segmented. Voxels in a patient volume above the threshold are discarded from automatic segmentation and would be referred to an expert. The number of incorrectly segmented voxels per patient volume is used as a measure of risk. Using bilinear interpolation risk measures are computed per patient volume between $[0, 100]$ percent.

3.3.4 Detection of segmentation failures

To detect segmentation failures uncertainty maps are used but direct application of uncertainties is infeasible because many correctly segmented voxels, such as those close to anatomical structure boundaries, have high uncertainty. Hence, an additional patch-based CNN (detection network) is used that takes a cardiac MR image together with the corresponding spatial uncertainty map as input. For each patch of 8×8 voxels the network generates a probability indicating whether it contains segmentation failure. In the following, the terms patch and region are used interchangeably.

The detection network is a shallow Residual Network (S-ResNet)[31] consisting of a feature extraction module with output stride eight followed by a classifier indicating the presence of segmentation failure. The first level of the feature extraction module consists of two convolutional layers. The first layer uses 16 kernels of 7×7 voxels and zero-padding of size 3 and second layer 32 kernels of 3×3 voxels and zero-padding of 1 voxel. Level 2 to 4 each consist of one residual block that contains two convolutional layers with 3×3 voxels kernels with zero-padding of size 1. The first convolutional layer of each residual block uses a strided convolution of 2 voxels to downsample the

input. All convolutional layers of the feature extraction module are followed by batch normalization and ReLU function. The number of kernels in the feature extraction module increases from 16 in level 1 to 128 in level 4. The network is a 2D patch-level classifier and requires that the size of the two input slices is a multiple of the patch-size. The final classifier consists of three fully convolutional layers, implemented as 1×1 convolutions, with 128 feature maps in the first two layers. The final layer has two channels followed by a softmax function which indicates whether the patch contains segmentation failure. Furthermore, to regularize the model dropout layers ($p = 0.5$) were added between the residual blocks and the fully convolutional layers of the classifier.

3.4 Evaluation

Automatic segmentation performance, as well as performance after simulating the correction of detected segmentation failures and after manual expert correction was evaluated. For this, the 3D Dice-coefficient (DC) and 3D Hausdorff distance (HD) between manual and (corrected) automatic segmentation were computed. Furthermore, the following clinical metrics were computed for manual and (corrected) automatic segmentation: left ventricle end-diastolic volume (EDV); left ventricle ejection fraction (EF); right ventricle EDV; right ventricle ejection fraction; and left ventricle myocardial mass. Following Bernard *et al.*[7] for each of the clinical metrics three performance indices were computed using the measurements based on manual and (corrected) automatic segmentation: Pearson correlation coefficient; mean difference (bias and standard deviation); and mean absolute error (MAE).

To evaluate detection performance of the automatic method precision-recall curves of identification of slices that require correction were computed. A slice is considered positive in case it consists of at least one image region with a segmentation failure. To achieve accurate segmentation in clinic, identification of slices that contain segmentation failures might ease manual correction of automatic segmentations in daily practice. To further evaluate detection performance detection rate of segmentation failures was assessed on a voxel level. More specific, sensitivity against the number of false positive regions was evaluated because manual correction is presumed to be performed at this level.

Finally, after simulation and manual correction of the automatically detected segmentation failures, segmentation was re-evaluated and significance of the difference between the DCs, HDs and clinical metrics was tested with a Mann–Whitney U test.

3.5 Experiments

To use stratified four-fold cross-validation the dataset was split into training (75%) and test (25%) set. The splitting was done on a patient level, so there was no overlap

Table 3.1: Segmentation performance of different combination of model architectures, loss functions and evaluation modes (without or with MC dropout enabled during testing) in terms of Dice coefficient (top) and Hausdorff distance (bottom) (mean ± standard deviation). Each combination comprises a block of two rows. A row in which column *Uncertainty map for detection* indicates e- or b-map shows results for the combined segmentation and detection approach. Numbers accentuated in black/bold are ranked first in the segmentation only task whereas numbers accentuated in red/bold are ranked first in the combined segmentation & detection task. The last row states the performance of the winning model[32] in the ACDC challenge (on 100 patient images). Number with asterisk indicates statistical significant at 5% level w.r.t. the segmentation-only approach. Best viewed in color.

(a) Dice coefficient

Model	Uncertainty map for detection	End-diastole LV	End-diastole RV	End-diastole LVM	End-systole LV	End-systole RV	End-systole LVM
DN-Brier		0.962±0.02	0.928±0.04)	0.875±0.03)	0.901±0.11	0.832±0.10)	0.884±0.04
	e-map	*0.965±0.01	*0.949±0.02	*0.885±0.03	*0.937±0.06	*0.905±0.05	*0.909±0.03
DN-Brier+MC		0.961±0.02	0.922±0.04	0.875±0.04	0.912±0.08	0.839±0.11	0.882±0.04
	b-map	*0.966±0.01	*0.950±0.01	*0.886±0.03	***0.942**±0.03	*0.916±0.04	*0.912±0.03
DN-soft-Dice		0.960±0.02	0.921±0.04	0.870±0.04	0.909±0.08	0.812±0.12	0.879±0.04
	e-map	*0.965±0.01	*0.945±0.02	*0.879±0.04	*0.938±0.03	*0.891±0.06	*0.905±0.03
DN-soft-Dice+MC		0.958±0.02	0.913±0.05	0.868±0.04	0.907±0.07	0.818±0.12	0.875±0.04
	b-map	*0.964±0.01	*0.944±0.02	*0.877±0.04	*0.939±0.03	*0.900±0.05	*0.904±0.03
DRN-CE		0.961±0.02	0.929±0.03	0.878±0.03	0.912±0.06	0.850±0.09	0.891±0.03
	e-map	0.964±0.01	*0.943±0.02	*0.886±0.03	*0.937±0.03	*0.899±0.04	*0.908±0.03
DRN-CE+MC		0.961±0.02	0.926±0.03	0.877±0.03	0.913±0.06	0.847±0.10	0.890±0.03
	b-map	*0.965±0.01	*0.948±0.01	*0.887±0.03	*0.939±0.03	*0.911±0.04	*0.909±0.03
DRN-soft-Dice		0.964±0.01	**0.937**±0.02	0.888±0.03	**0.919**±0.06	0.856±0.09	**0.900**±0.03
	e-map	0.967±0.01	*0.945±0.02	0.893±0.03	0.934±0.04	*0.892±0.06	*0.911±0.03
DRN-soft-Dice+MC		0.963±0.02	0.935±0.03	0.886±0.03	0.921±0.06	**0.857**±0.09	0.899±0.03
	b-map	0.967±0.01	*0.947±0.02	0.893±0.03	*0.938±0.03	*0.907±0.04	*0.912±0.03
U-net-CE		0.962±0.02	0.923±0.05	0.878±0.03	0.907±0.07	0.840±0.08	0.885±0.03
	e-map	0.966±0.01	*0.946±0.02	*0.890±0.03	*0.935±0.04	*0.901±0.06	*0.909±0.03
U-net-CE+MC		0.962±0.02	0.926±0.04	0.879±0.03	0.909±0.07	0.849±0.07	0.887±0.03
	b-map	0.967±0.01	***0.954**±0.02	*0.893±0.03	*0.940±0.04	***0.920**±0.04	***0.914**±0.03
U-net-soft-Dice		**0.965**±0.02	0.928±0.04	0.888±0.03	0.914±0.08	0.844±0.09	0.896±0.03
	e-map	0.968±0.01	*0.943±0.03	*0.898±0.03	0.930±0.05	*0.886±0.07	*0.911±0.03
U-net-soft-Dice+MC		**0.965**±0.02	0.929±0.04	**0.889**±0.03	0.911±0.10	0.845±0.09	0.897±0.03
	b-map	***0.968**±0.01	*0.948±0.03	***0.900**±0.03	0.928±0.09	*0.895±0.06	***0.914**±0.03
Isensee et al.		0.966	0.941	0.899	0.924	0.875	0.908

(b) Hausdorff Distance

Model	Uncertainty map for detection	End-diastole LV	End-diastole RV	End-diastole LVM	End-systole LV	End-systole RV	End-systole LVM
DN-Brier		6.7±3.1	13.5±5.9	10.2±6.9	10.7±7.7	16.7±6.8	12.3±5.8
	e-map	*5.7±2.7	*11.7±5.2	* 8.3±5.9	* 8.0±6.5	*14.2±5.6	* 9.7±5.0
DN-Brier+MC		6.9±3.3	13.1±5.2	9.9±5.9	9.9±5.7	15.0±6.1	12.0±5.2
	b-map	*5.5±2.6	*10.6±5.1	* 7.4±4.2	* 7.5±6.0	*12.6±5.6	* 8.8±4.0
DN-soft-Dice		7.1±3.5	14.8±6.8	11.0±6.6	10.2±5.6	17.7±7.8	12.9±6.2
	e-map	*5.6±2.8	*12.6±5.5	* 8.6±4.6	* 8.0±5.0	*14.6±5.9	* 9.6±4.5
DN-soft-Dice+MC		7.7±3.9	14.4±6.0	10.5±4.9	10.1±5.3	17.2±8.0	12.5±5.3
	b-map	*6.3±3.4	*11.5±4.0	* 8.6±4.8	* 7.8±4.6	*13.6±4.9	* 9.6±4.7
DRN-CE		5.5±2.6	11.7±5.4	8.2±6.2	9.1±6.4	13.7±5.6	8.9±5.3
	e-map	*4.5±1.9	* 9.0±4.5	* 6.3±4.1	* 6.2±4.4	*11.1±5.3	* **6.7**±4.2
DRN-CE+MC		5.6±2.6	11.9±5.5	8.0±5.9	8.7±5.5	13.5±5.9	**8.5**±4.5
	b-map	***4.2**±1.6	* **8.1**±3.7	* **6.1**±4.2	* **5.4**±3.6	***10.1**±5.5	* 6.8±3.8
DRN-soft-Dice		**5.5**±2.8	11.9±6.1	**7.7**±5.9	8.5±5.0	13.5±5.5	8.9±5.1
	e-map	*4.6±2.2	* 9.4±4.5	6.7±4.7	* 6.7±4.4	*11.6±5.4	* 7.0±3.3
DRN-soft-Dice+MC		5.7±3.2	**11.5**±5.1	8.0±5.5	**8.3**±4.5	**13.3**±5.1	8.9±5.1
	b-map	*4.5±2.2	* 9.3±4.5	* 6.3±4.0	* 6.2±4.1	*10.4±5.0	* 7.0±3.4
U-net-CE		6.4±4.3	15.7±8.6	9.0±6.0	9.7±5.3	17.0±7.7	12.7±8.2
	e-map	*4.9±3.9	*12.2±8.1	* 7.1±5.6	* 6.1±3.2	*12.6±6.5	* 8.4±6.3
U-net-CE+MC		6.2±4.2	15.3±8.4	8.8±5.8	9.2±5.0	16.5±7.6	12.0±8.0
	b-map	*4.3±1.6	* 9.9±6.6	* 6.7±4.8	* 5.4±2.8	*10.3±4.7	* 7.6±6.2
U-net-soft-Dice		6.1±3.9	14.1±7.6	10.6±8.4	9.2±7.1	16.3±7.5	12.6±9.6
	e-map	*4.6±2.3	*11.3±7.2	* 7.5±5.5	* 7.3±6.5	*13.7±7.6	* 9.8±8.0
U-net-soft-Dice+MC		6.2±3.9	14.1±7.7	10.5±8.7	9.0±7.0	15.8±7.5	12.1±9.2
	b-map	*4.5±2.1	*10.4±7.2	* 7.6±7.0	* 7.3±6.9	*12.9±6.6	* 9.8±8.4
Isensee et al.		7.1	14.3	8.9	9.8	16.3	10.4

in patient data between training and test sets. Furthermore, patients were randomly chosen from each of the five patient groups w.r.t. disease. Each patient has one volume for ED and ES time points, respectively.

3.5.1 Training segmentation networks

DRN and U-net were trained with a patch size of 128×128 voxels which is a multiple of their output stride of the contracting path. In the training of the dilated CNN (DN) images with 151×151 voxel samples were used. Zero-padding to 281×281 was performed to accommodate the 131×131 voxel receptive field that is induced by the dilation factors. Training samples were randomly chosen from training set and augmented by 90 degree rotations of the images. All models were initially trained with three loss functions: soft-Dice[33] (SD); cross-entropy (CE); and Brier loss.[34] However, for the evaluation of the combined segmentation and detection approach for each model architecture the two best performing loss functions were chosen: soft-Dice for all models; cross-entropy for DRN and U-net and Brier loss for DN. For completeness, we provide the equations for all three used loss functions.

$$\text{soft-Dice}_c = \frac{\sum_{i=1}^{N} R_c(i)\, A_c(i)}{\sum_{i=1}^{N} R_c(i) + \sum_{i=1}^{N} A_c(i)}\,, \tag{3.3}$$

where N denotes the number of voxels in an image, R_c is the binary reference image for class c and A_c is the probability map for class c.

$$\text{Cross-Entropy}_c = -\sum_{i=1}^{N} t_{ic}\ \log\ p(y_i = c|x_i)\,, \quad \text{where } t_{ic} = 1 \text{ if } y_i = c, \text{ and 0 otherwise.} \tag{3.4}$$

$$\text{Brier}_c = \sum_{i=1}^{N} \left(t_{ic} - p(y_i = c|x_i)\right)^2\,, \quad \text{where } t_{ic} = 1 \text{ if } y_i = c, \text{ and 0 otherwise.} \tag{3.5}$$

where N denotes the number of voxels in an image and p denotes the probability for a specific voxel x_i with corresponding reference label y_i for class c. Choosing Brier loss to train the DN model instead of CE was motivated by our preliminary work which showed that segmentation performance of DN model was best when trained with Brier loss.[25] All models were trained for 100,000 iterations. DRN and U-net were trained with a learning rate of 0.001 which decayed with a factor of 0.1 after every 25,000 steps. Training DN used the snapshot ensemble technique,[35] where after every 10,000 iterations the learning rate was reset to its original value of 0.02.

All three segmentation networks were trained using mini-batch stochastic gradient descent using a batch size of 16. Network parameters were optimized using the Adam

optimizer.[36] Furthermore, models were regularized with weight decay to increase generalization performance.

3.5.2 Training detection network

To train the detection model a subset of the errors performed by the segmentation model is used. Segmentation errors that presumably are within the range of inter-observer variability and therefore do not inevitably require correction (tolerated errors) are excluded from the set of errors that need to be detected and corrected (segmentation failures). To distinguish between tolerated errors and the set of segmentation failures $\mathcal{S}_I$ the Euclidean distance of an incorrectly segmented voxel to the boundary of the reference target structure is used. For each anatomical structure a 2D distance transform map is computed that provides for each voxel the distance to the anatomical structure boundary. To differentiate between tolerated errors and the set of segmentation failures $\mathcal{S}_I$ an acceptable tolerance threshold is applied. A more rigorous threshold is used for errors located inside compared to outside of the anatomical structure because automatic segmentation methods have a tendency to undersegment cardiac structures in CMRI. Hence, in all experiments the acceptable tolerance threshold was set to three voxels (equivalent to on average 4.65 mm) and two voxels (equivalent to on average 3.1 mm) for segmentation errors located outside and inside the target structure. Furthermore, a segmentation error only belongs to $\mathcal{S}_I$ if it is part of a 2D 4-connected cluster of minimum size 10 voxels. This value was found in preliminary experiments by evaluating values $\{1, 5, 10, 15, 20\}$. However, for apical slices all segmentation errors are included in $\mathcal{S}_I$ regardless of fulfilling the minimum size requirement because in these slices anatomical structures are relatively small and manual segmentation is prone to large inter-observer variability.[7] Finally, segmentation errors located in slices above the base or below the apex are always included in the set of segmentation failures.

Using the set $\mathcal{S}_I$ a binary label t_j is assigned to each patch $P_j^{(I)}$ indicating whether $P_j^{(I)}$ contains at least one voxel belonging to set $\mathcal{S}_I$ where $j \in \{1 \dots M\}$ and M denotes the number of patches in a slice I. The detection network is trained by minimizing a weighted binary cross-entropy loss:

$$\mathcal{L}_{DT} = -\sum_{j \in P^{(I)}} w_{pos}\, t_j \log p_j + (1 - t_j) \log(1 - p_j)\,, \tag{3.6}$$

where w_{pos} represents a scalar weight, t_j denotes the binary reference label and p_j is the softmax probability indicating whether a particular image region $P_j^{(I)}$ contains at least one segmentation failure. The average percentage of regions in a patient volume containing segmentation failures ranges from 1.5 to 3 percent depending on the segmentation architecture and loss function used to train the segmentation model. To train a detection network w_{pos} was set to the ratio between the average percentage of negative samples divided by the average percentage of positive samples.

Each fold was trained using spatial uncertainty maps and automatic segmentation masks generated while training the segmentation networks. Hence, there was no overlap in patient data between training and test set across segmentation and detection tasks. In total 12 detection models were trained and evaluated resulting from the different combination of 3 model architectures (DRN, DN and U-net), 2 loss functions (DRN and U-net with CE and soft-Dice, DN with Brier and soft-Dice) and 2 uncertainty maps (e-maps, b-maps).

Table 3.2: Segmentation performance of different combination of model architectures, loss functions and evaluation modes (without or with MC dropout (MC) enabled during testing) in terms of clinical metrics: left ventricle (LV) end-diastolic volume (EDV); LV ejection fraction (EF); right ventricle (RV) EDV; RV ejection fraction; and LV myocardial mass. Quantitative results compare clinical metrics based on reference segmentations with 1) automatic segmentations and 2) simulated manual correction of automatic segmentations using spatial uncertainty maps. ρ denotes the Pearson correlation coefficient, *bias* denotes the mean difference between the two measurements (mean ± standard deviation) and *MAE* denotes the mean absolute error between the two measurements. Each combination comprises a block of two rows. A row in which column *Uncertainty map for detection* indicates e- or b-map shows results for the combined segmentation and detection approach. Numbers accentuated in black/bold are ranked first in the segmentation only task. Numbers in red indicate statistical significant at 5% level w.r.t. the segmentation-only approach for the specific clinical metric. Best viewed in color.

Method	Uncertainty map for detection	LV_{EDV} ρ	LV_{EDV} bias±σ	LV_{EDV} MAE	LV_{EF} ρ	LV_{EF} bias±σ	LV_{EF} MAE	RV_{EDV} ρ	RV_{EDV} bias±σ	RV_{EDV} MAE	RV_{EF} ρ	RV_{EF} bias±σ	RV_{EF} MAE	LVM_{Mass} ρ	LVM_{Mass} bias±σ	LVM_{Mass} MAE
DN-Brier		0.997	**0.0±6.1**	4.5	0.892	2.2± 9.2	4.2	0.977	**-0.2±11.8**	8.5	0.834	5.3±10.3	8.5	0.984	-2.7± 9.0	**7.0**
	e-map	0.997	0.0±5.5	4.0	0.982	0.1± 3.8	2.2	0.992	0.0± 6.9	5.2	0.955	1.9± 5.5	4.1	0.986	-2.1± 8.4	6.6
DN-Brier+MC		0.997	1.6±6.0	4.4	0.921	1.1± 7.9	3.9	0.975	6.7±12.4	9.6	0.854	3.5± 9.9	7.7	0.984	0.7± 9.2	7.1
	b-map	0.998	1.0±5.3	3.9	0.991	0.0± 2.7	1.9	0.993	3.2± 6.7	5.7	0.975	0.8± 4.0	3.0	0.987	0.1± 8.3	6.5
DN-soft-Dice		0.996	1.2±6.5	4.9	0.918	1.5± 8.0	3.9	0.972	**0.2±13.0**	9.6	0.802	7.2±11.3	10.2	0.982	-4.5± 9.6	8.5
	e-map	0.997	1.0±5.5	4.2	0.989	0.2± 3.0	2.2	0.990	0.2± 7.6	5.9	0.940	3.3± 6.2	5.2	0.983	-4.3± 9.3	8.2
DN-soft-Dice+MC		0.996	3.2±7.1	5.6	0.958	0.4± 5.7	3.6	0.964	8.1±14.9	12.3	0.827	4.8±11.0	8.9	0.978	-0.7±10.7	8.3
	b-map	0.997	2.2±5.6	4.4	0.988	-0.2± 3.1	2.2	0.990	4.0± 7.7	7.0	0.959	1.8± 5.1	4.1	0.982	-1.0± 9.5	7.6
DRN-CE		0.997	-0.2±5.5	4.1	0.968	1.2± 5.0	3.5	0.976	1.5±12.1	8.5	0.870	1.3± 9.2	6.9	0.980	**0.6±10.2**	7.8
	e-map	0.998	0.2±4.5	3.5	0.992	0.2± 2.5	1.9	0.988	1.4± 8.5	6.2	0.952	0.8± 5.6	4.2	0.985	0.4± 8.7	6.8
DRN-CE+MC		**0.998**	1.0±4.9	**3.9**	0.972	0.8± 4.6	3.1	0.973	4.8±12.8	9.4	0.876	**0.4± 9.1**	**6.6**	0.981	1.9± 9.9	7.6
	b-map	0.998	0.7±4.6	3.6	0.992	-0.1± 2.5	1.8	0.992	2.9± 6.9	5.7	0.967	0.6± 4.6	3.4	0.987	1.2± 8.3	6.6
DRN-soft-Dice		**0.998**	0.8±5.1	4.0	0.976	0.2± 4.4	3.0	0.980	**0.2±11.0**	**7.5**	**0.882**	3.1± 8.7	6.8	0.984	-3.5± 9.1	7.5
	e-map	0.998	0.7±4.4	3.5	0.987	- 0.1± 3.1	2.2	0.987	0.1± 9.1	6.4	0.938	1.9± 6.3	4.9	0.986	-3.5± 8.7	7.1
DRN-soft-Dice+MC		**0.998**	1.8±5.1	**3.9**	**0.979**	-0.3± 4.1	2.9	0.977	3.5±11.7	8.1	0.868	1.7± 9.5	6.8	0.983	-1.4± 9.5	7.4
	b-map	0.998	1.7±4.7	3.7	0.990	-0.2± 2.9	2.1	0.989	2.3± 8.1	5.8	0.959	0.8± 5.2	3.8	0.986	-1.3± 8.5	6.8
U-net-CE		0.995	-4.7±7.2	6.1	0.954	4.1± 6.0	5.1	0.963	-7.6±15.2	12.1	0.870	5.6± 9.0	8.1	0.971	-8.5±12.2	11.5
	e-map	0.998	-3.2±4.8	4.4	0.992	1.7± 2.6	2.4	0.987	-4.1± 9.1	6.7	0.957	2.6± 5.2	4.1	0.983	-5.7± 9.3	8.2
U-net-CE+MC		0.995	-4.3±7.2	5.9	0.958	3.8± 5.8	4.9	0.968	-4.8±14.1	10.7	0.867	5.0± 9.1	7.9	0.972	-8.1±12.0	11.1
	b-map	0.997	-3.5±5.5	4.9	0.990	1.6± 2.9	2.6	0.992	-1.8± 7.0	4.9	0.974	1.6± 4.1	3.3	0.981	-6.8±10.0	9.4
U-net-soft-Dice		0.997	-2.0±6.0	4.5	0.853	3.6±10.9	5.0	0.968	-1.0±14.1	10.0	0.782	4.8±11.6	9.0	**0.985**	-7.7± 8.8	9.2
	e-map	0.997	-1.7±5.3	4.1	0.969	1.9± 4.9	3.3	0.981	-0.1±10.9	7.5	0.919	3.3± 7.0	5.9	0.984	-6.6± 9.0	8.7
U-net-soft-Dice+MC		0.997	-1.8±5.9	4.4	0.941	3.0± 6.7	4.4	0.969	0.6±13.9	9.8	0.792	4.4±11.3	8.7	**0.985**	-7.2± 8.9	8.9
	b-map	0.997	-1.5±5.3	4.1	0.979	1.1± 4.1	2.9	0.985	1.2± 9.4	6.5	0.939	2.9± 6.2	4.9	0.984	-5.9± 9.0	8.5

Table 3.3: Segmentation performance of different combination of model architectures, loss functions and evaluation modes (without or with MC dropout (MC) enabled during testing) in terms of clinical metrics: left ventricle (LV) end-diastolic volume (EDV); LV ejection fraction (EF); right ventricle (RV) EDV; RV ejection fraction; and LV myocardial mass. Quantitative results compare clinical metrics based on reference segmentations with 1) automatic segmentations and 2) simulated manual correction of automatic segmentations using spatial uncertainty maps. ρ denotes the Pearson correlation coefficient, *bias* denotes the mean difference between the two measurements (mean ± standard deviation) and *MAE* denotes the mean absolute error between the two measurements. Each combination comprises a block of two rows. A row in which column *Uncertainty map for detection* indicates e- or b-map shows results for the combined segmentation and detection approach. Numbers accentuated in black/bold are ranked first in the segmentation only task. Numbers in red indicate statistical significant at 5% level w.r.t. the segmentation-only approach for the specific clinical metric. Best viewed in color.

	Uncertainty map for detection	LV_{EDV}			LV_{EF}			RV_{EDV}			RV_{EF}			LVM_{Mass}		
Method		ρ	**bias**±σ	**MAE**	ρ	**bias**±σ	**MAE**	ρ	**bias**±σ	**MAE**	ρ	**bias**±σ	**MAE**	ρ	**bias**±σ	**MAE**
DN-Brier		0.997	**0.0±6.1**	4.5	0.892	2.2± 9.2	4.2	0.977	**-0.2±11.8**	8.5	0.834	5.3±10.3	8.5	0.984	-2.7± 9.0	**7.0**
	e-map	0.997	0.0±5.5	4.0	0.982	0.1± 3.8	2.2	0.992	0.0± 6.9	5.2	0.955	1.9± 5.5	4.1	0.986	-2.1± 8.4	6.6
DN-Brier+MC		0.997	1.6±6.0	4.4	0.921	1.1± 7.9	3.9	0.975	6.7±12.4	9.6	0.854	3.5± 9.9	7.7	0.984	0.7± 9.2	7.1
	b-map	0.998	1.0±5.3	3.9	0.991	0.0± 2.7	1.9	0.993	3.2± 6.7	5.7	0.975	0.8± 4.0	3.0	0.987	0.1± 8.3	6.5
DN-soft-Dice		0.996	1.2±6.5	4.9	0.918	1.5± 8.0	3.9	0.972	**0.2±13.0**	9.6	0.802	7.2±11.3	10.2	0.982	-4.5± 9.6	8.5
	e-map	0.997	1.0±5.5	4.2	0.989	0.2± 3.0	2.2	0.990	0.2± 7.6	5.9	0.940	3.3± 6.2	5.2	0.983	-4.3± 9.3	8.2
DN-soft-Dice+MC		0.996	3.2±7.1	5.6	0.958	0.4± 5.7	3.6	0.964	8.1±14.9	12.3	0.827	4.8±11.0	8.9	0.978	-0.7±10.7	8.3
	b-map	0.997	2.2±5.6	4.4	0.988	-0.2± 3.1	2.2	0.990	4.0± 7.7	7.0	0.959	1.8± 5.1	4.1	0.982	-1.4± 9.5	7.6
DRN-CE		0.997	-0.2±5.5	4.1	0.968	1.2± 5.0	3.5	0.976	1.5±12.1	8.5	0.870	1.3± 9.2	6.9	0.980	**0.6±10.2**	7.8
	e-map	0.998	0.2±4.5	3.5	0.992	0.2± 2.5	1.9	0.988	1.4± 8.5	6.2	0.952	0.8± 5.6	4.2	0.985	0.4± 8.7	6.8
DRN-CE+MC		**0.998**	1.0±4.9	**3.9**	0.972	0.8± 4.6	3.1	0.973	4.8±12.8	9.4	0.876	**0.4± 9.1**	**6.6**	0.981	1.9± 9.9	7.6
	b-map	0.998	0.7±4.6	3.6	0.992	-0.1± 2.5	1.8	0.992	2.9± 6.9	5.7	0.967	0.6± 4.6	3.4	0.987	1.2± 8.3	6.6
DRN-soft-Dice		**0.998**	0.8±5.1	4.0	0.976	0.2± 4.4	3.0	0.980	**0.2±11.0**	**7.5**	**0.882**	3.1± 8.7	6.8	0.984	-3.5± 9.1	7.5
	e-map	0.998	0.7±4.4	3.5	0.987	-0.1± 3.1	2.2	0.987	0.1± 9.1	6.4	0.938	1.9± 6.3	4.9	0.986	-3.5± 8.7	7.1
DRN-soft-Dice+MC		**0.998**	1.8±5.1	**3.9**	**0.979**	-0.3± 4.1	2.9	0.977	3.5±11.7	8.1	0.868	1.7± 9.5	6.8	0.983	-1.4± 9.5	7.4
	b-map	0.998	1.7±4.7	3.7	0.990	-0.2± 2.9	2.1	0.989	2.3± 8.1	5.8	0.959	0.8± 5.2	3.8	0.986	-1.3± 8.5	6.8
U-net-CE		0.995	-4.7±7.2	6.1	0.954	4.1± 6.0	5.1	0.963	-7.6±15.2	12.1	0.870	5.6± 9.0	8.1	0.971	-8.5±12.2	11.5
	e-map	0.998	-3.2±4.8	4.4	0.992	1.7± 2.6	2.4	0.987	-4.1± 9.1	6.7	0.957	2.6± 5.2	4.1	0.983	-5.7± 9.3	8.2
U-net-CE+MC		0.995	-4.3±7.2	5.9	0.958	3.8± 5.8	4.9	0.968	-4.8±14.1	10.7	0.867	5.0± 9.1	7.9	0.972	-8.1±12.0	11.1
	b-map	0.997	-3.5±5.5	4.9	0.990	1.6± 2.9	2.6	0.992	-1.8± 7.0	4.9	0.974	1.6± 4.1	3.3	0.981	-6.8±10.0	9.4
U-net-soft-Dice		0.997	-2.0±6.0	4.5	0.853	3.6±10.9	5.0	0.968	-1.0±14.1	10.0	0.782	4.8±11.6	9.0	**0.985**	-7.7± 8.8	9.2
	e-map	0.997	-1.7±5.3	4.1	0.969	1.9± 4.9	3.3	0.981	-0.1±10.9	7.5	0.919	3.3± 7.0	5.9	0.984	-6.6± 9.0	8.7
U-net-soft-Dice+MC		0.997	-1.8±5.9	4.4	0.941	3.0± 6.7	4.4	0.969	0.6±13.9	9.8	0.792	4.4±11.3	8.7	**0.985**	-7.2± 8.9	8.9
	b-map	0.997	-1.5±5.3	4.1	0.979	1.1± 4.1	2.9	0.985	1.2± 9.4	6.5	0.939	2.9± 6.2	4.9	0.984	-5.9± 9.0	8.5

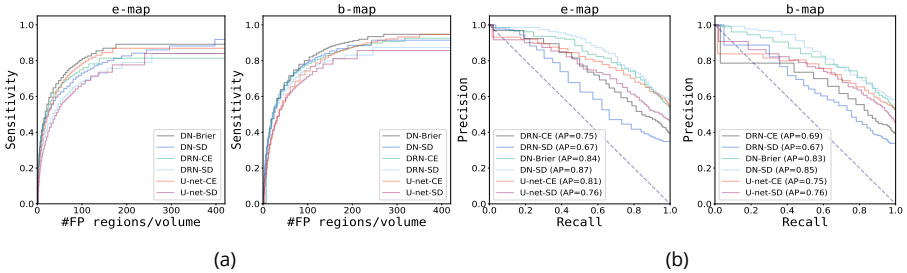


Figure 3.3: Detection performance of segmentation failures generated by different combination of segmentation architectures and loss functions. (a) Sensitivity for detection of segmentation failures on voxel level (y-axis) as a function of number of false positive image regions (x-axis). (b) Precision-recall curve for detection of slices containing segmentation failures (where AP denotes average precision). Results are split between entropy and Bayesian uncertainty maps. Each figure contains a curve for the six possible combination of models (three) and loss functions (two). SD denotes soft-Dice and CE cross-entropy, respectively.

The patches used to train the network were selected randomly (2/3), or were forced (1/3) to contain at least one segmentation failure by randomly selecting a scan containing segmentation failure, followed by random sampling of a patch containing at least one segmentation failure. During training the patch size was fixed to 80×80 voxels. To reduce the number of background voxels during testing, inputs were cropped based on a minimal enclosing, rectangular bounding box that was placed around the automatic segmentation mask. Inputs always had a minimum size of 80×80 voxels or were forced to a multiple of the output grid spacing of eight voxels in both direction required by the patch-based detection network. The patches of size 8×8 voxels did not overlap. In cases where the automatic segmentation mask only contains background voxels (scans above the base or below apex of the heart) input scans were center-cropped to a size of 80×80 voxels.

Models were trained for 20,000 iterations using mini-batch stochastic gradient descent with batch-size 32 and Adam as optimizer.[36] Learning rate was set to 0.0001 and decayed with a factor of 0.1 after 10.000 steps. Furthermore, dropout percentage was set to 0.5 and weight decay was applied to increase generalization performance.

3.5.3 Segmentation using correction of the detected segmentation failures

To investigate whether correction of detected segmentation failures increases segmentation performance two scenarios were performed. In the first scenario manual correction of the detected failures by an expert was simulated for all images at ED and ES time points of the ACDC dataset. For this purpose, in image regions that were detected to contain segmentation failure predicted labels were replaced with reference labels. In the second scenario manual correction of the detected failures was performed

by an expert in a random subset of 50 patients of the ACDC dataset. The expert was shown CMRI slices for ED and ES time points together with corresponding automatic segmentation masks for the RV, LV and LV myocardium. Image regions detected to contain segmentation failures were indicated in slices and the expert was only allowed to change the automatic segmentations in these indicated regions. Annotation was performed following the protocol described in Bernard *et al.*[7] Furthermore, expert was able to navigate through all CMRI slices of the corresponding ED and ES volumes.

3.6 Results

In this section we first present results for the segmentation-only task followed by description of the combined segmentation and detection results.

3.6.1 Segmentation-only approach

Table 3.1 lists quantitative results for segmentation-only and combined segmentation and detection approach in terms of Dice coefficient and Hausdorff distance. These results show that DRN and U-net achieve similar Dice coefficients and outperformed the DN network for all anatomical structures at end-systole. Differences in the achieved Hausdorff distances among the methods are present for all anatomical structures and for both time points. The DRN model achieved the highest and the DN network the lowest Hausdorff distance.

Table 3.3 lists results of the evaluation in terms of clinical metrics. These results reveal noticeable differences between models for ejection fraction (EF) of left and right ventricle, respectively. We can observe that U-net trained with the soft-Dice and the Dilated Network (DN) trained with Brier or soft-Dice loss achieved considerable lower accuracy for LV and RV ejection fraction compared to DRN. Overall, the DRN model achieved highest performance for all clinical metrics.

Effect of model architecture on segmentation: Although quantitative differences between models are small, qualitative evaluation discloses that automatic segmentations differ substantially between the models. Figure 3.2 shows that especially in regions where the models perform poorly (apical and basal slices) the DN model more often produced anatomically implausible segmentations compared to the DRN and U-net. This seems to be correlated with the performance differences in Hausdorff distance.

Effect of loss function on segmentation: The results indicate that the choice of loss function only slightly affects the segmentation performance. DRN and U-net perform marginally better when trained with soft-Dice compared to cross-entropy whereas DN performs better when trained with Brier loss than with soft-Dice. For DN this is most pronounced for the RV at ES.

A considerable effect of the loss function on the accuracy of the LV and RV ejection fraction can be observed for the U-net model. On both metrics U-net achieved the

Table 3.4: Average precision and percentage of slices with segmentation failures generated by Dilated Network (DN), Dilated Residual Network (DRN) and U-net when trained with soft-Dice (SD), CE or Brier loss. Per patient, average precision of detected slices with failure using e- or b-maps (2nd and 3rd columns). Per patient, average percentage of slices containing segmentation failures (reference for detection task) (4th and 5th columns).

Model	Average precision		% of slices with segmentation failures	
	e-map	b-map	e-map	b-map
DN-Brier	84.0	83.0	53.7	52.4
DN-SD	87.0	85.0	58.3	58.1
DRN-CE	75.0	69.0	39.5	39.4
DRN-SD	67.0	67.0	34.9	33.7
U-net-CE	81.0	75.0	54.8	52.5
U-net-SD	76.0	76.0	46.7	45.5

lowest accuracy of all models when trained with the soft-Dice loss.

Effect of MC dropout on segmentation: The results show that enabling MC-dropout during testing seems to result in slightly improved HD while it does not affect DC.

3.6.2 Detection of segmentation failures

Detection of segmentation failures on voxel level: To evaluate detection performance of segmentation failures on voxel level Figure 3.3a shows average voxel detection rate as a function of false positively detected regions. This was done for each combination of model architecture and loss function exploiting e- (Figure 3.3a, left) or b-maps (Figure 3.3a, right). These results show that detection performance of segmentation failures depends on segmentation model architecture, loss function and uncertainty map.

The influence of (segmentation) model architecture and loss function on detection performance is slightly stronger when e-maps were used as input for the detection task compared to b-maps. Detection rates are consistently lower when segmentation failures originate from segmentation models trained with soft-Dice loss compared to models trained with CE or Brier loss. Overall, detection rates are higher when b-maps were exploited for the detection task compared to e-maps.

Detection of slices with segmentation failures: To evaluate detection performance w.r.t. slices containing segmentation failures precision-recall curves for each combination of model architecture and loss function using e-maps (Figure 3.3b, left) or b-maps (Figure 3.3b, right) are shown. The results show that detection performance of slices containing segmentation failures is slightly better for all models when using e-maps. Furthermore, the detection network achieves highest performance using uncertainty maps obtained from the DN model and the lowest when exploiting e- or b-maps ob-

Table 3.5: Comparing performance of segmentation-only approach (auto-only) with combined segmentation and detection approach for two scenarios: simulated correction of detected segmentation failures (auto+simulation); and manual correction of detected segmentation failures by an expert (auto+expert). Automatic segmentations were obtained from a U-net trained with cross-entropy. Evaluation was performed on a subset of 50 patients from the ACDC dataset. Scenarios are compared against segmentation-only approach (auto-only) in terms of (a) Dice Coefficient (b) Hausdorff Distance and (c) Clinical metrics. Results obtained from simulated manual correction represent an upper bound on the maximum achievable performance. Detection network was trained with e-maps. Number with asterisk indicates statistical significant at 5% level w.r.t. the segmentation-only approach.

(a) **Dice coefficient:** Mean ± standard deviation for left ventricle (LV), right ventricle (RV) and left ventricle myocardium (LVM).

	End-diastole			End-systole		
Scenario	**LV**	**RV**	**LVM**	**LV**	**RV**	**LVM**
auto-only	0.964±0.02	0.927±0.04	0.883±0.03	0.916±0.05	0.854±0.08	0.886±0.04
auto+simulation	0.967±0.01	*0.948±0.03	*0.894±0.03	*0.939±0.03	*0.915±0.04	*0.910±0.03
auto+expert	0.965±0.02	0.940±0.03	0.885±0.03	0.927±0.04	0.868±0.07	0.894±0.03

(b) **Hausdorff Distance:** Mean ± standard deviation for left ventricle (LV), right ventricle (RV) and left ventricle myocardium (LVM).

	End-diastole			End-systole		
Scenario	**LV**	**RV**	**LVM**	**LV**	**RV**	**LVM**
auto-only	5.6±3.3	15.7±9.7	8.5±6.4	9.2±5.8	16.5±8.8	13.4±10.5
auto+simulation	4.5±2.1	*9.0±4.6	*5.9±3.4	*5.2±2.5	*10.3±3.7	*6.6±2.9
auto+expert	4.9±2.8	*9.8±4.3	7.3±4.3	7.2±3.3	*12.5±4.7	*8.3±3.5

(c) **Clinical metrics:** a) Left ventricle (LV) end-diastolic volume (EDV) b) LV ejection fraction (EF) c) Right ventricle (RV) EDV d) RV ejection fraction e) LV myocardial mass. Quantitative results compare clinical metrics based on reference segmentations with 1) automatic segmentations; 2) simulated manual correction and 3) manual expert correction of automatic segmentations using spatial uncertainty maps. ρ denotes the Pearson correlation coefficient, *bias* denotes the mean difference between the two measurements (mean ± standard deviation) and *MAE* denotes the mean absolute error between the two measurements.

	LV_{EDV}			LV_{EF}			RV_{EDV}			RV_{EF}			LVM_{Mass}		
Scenario	ρ	**bias** ±σ	**MAE**	ρ	**bias** ±σ	**MAE**	ρ	**bias** ±σ	**MAE**	ρ	**bias** ±σ	**MAE**	ρ	**bias** ±σ	**MAE**
auto-only	0.995	-4.4 ±7.0	5.7	0.927	5.0 ±7.1	5.8	0.962	-6.4 ±16.2	11.9	0.878	5.8 ±8.7	8.0	0.979	-6.4 ±10.6	9.5
auto+simulation	0.998	-3.9 ±5.2	4.8	0.989	2.3 ±2.9	2.9	0.984	-3.7 ±10.4	6.8	0.954	2.7 ±5.5	4.5	0.983	-5.5 ±9.6	8.1
auto+expert	0.996	-4.3 ±6.5	5.5	0.968	2.7 ±4.8	4.3	0.976	-3.2 ±12.9	8.3	0.883	5.1 ±8.6	7.7	0.980	-6.2 ±10.2	9.1

Table 3.6: Effect of number of Monte Carlo (MC) samples on segmentation performance in terms of (a) Dice coefficient (DC) and (b) Hausdorff Distance (HD) (mean ± standard deviation). Higher DC and lower HD is better. Abbreviations: Cross-Entropy (CE), Dilated Residual Network (DRN) and Dilated Network (DN).

(a) Dice coefficient

Number of MC samples	DRN-CE	U-net-CE	DN-soft-Dice
1	0.894±0.07	0.896±0.07	0.871±0.09
3	0.900±0.07	0.901±0.07	0.883±0.08
5	0.902±0.07	0.901±0.07	0.887±0.08
7	0.903±0.07	0.901±0.07	0.888±0.08
10	0.904±0.06	0.902±0.07	0.890±0.08
20	0.904±0.07	0.902±0.07	0.890±0.08
30	0.904±0.07	0.902±0.07	0.891±0.08
60	0.904±0.07	0.902±0.07	0.891±0.08

(b) Hausdorff Distance

Number of MC samples	DRN-CE	U-net-CE	DN-soft-Dice
1	9.88±5.76	11.79±8.23	13.54±7.14
3	9.70±6.13	11.40±7.78	12.71±6.79
5	9.54±6.07	11.37±7.81	12.06±6.29
7	9.38±5.86	11.29±7.86	12.08±6.38
10	9.38±5.91	11.24±7.71	11.85±6.34
20	9.37±5.83	11.27±7.79	11.90±6.52
30	9.39±5.91	11.32±7.93	11.90±6.48
60	9.39±5.93	11.22±7.83	11.89±6.56

tained from the DRN model. Table 3.4 shows the average precision of detected slices with segmentation failures per patient, as well as the average percentage of slices that do contain segmentation failures (reference for detection task). The results illustrate that these measures are positively correlated i.e. that precision of detected slices in a patient volume is higher if the volume contains more slices that need correction. On average the DN model generates cardiac segmentations that contain more slices with at least one segmentation failure compared to U-net (ranks second) and DRN (ranks third). A higher number of detected slices containing segmentation failures implies an increased workload for manual correction.

3.6.3 Calibration of uncertainty maps

Figure 3.4 shows risk-coverage curves for each combination of model architectures, uncertainty maps and loss functions (Figure 3.4 left: CE or Brier loss, Figure 3.4 right: soft-Dice). The results show an effect of the loss function on slope and convergence of the curves. Segmentation errors of models trained with the soft-Dice loss are less frequently covered by higher uncertainties than models trained with CE or Brier loss (steeper slope and lower minimum are better). This difference is more pronounced for e-maps. Models trained with the CE or Brier loss only slightly differ concerning convergence and their slopes are approximately identical. In contrast, the curves of the models trained with the soft-Dice differ regarding their slope and achieved minimum. Comparing e- and b-map of the DN-SD and U-net-SD models the results reveal that the curve for b-map has a steeper slope and achieves a lower minimum compared to the e-map. For the DRN-SD model these differences are less striking. In general for a specific combination of model and loss function the risk-coverage curves using b-maps achieve a lower minimum compared to e-maps.

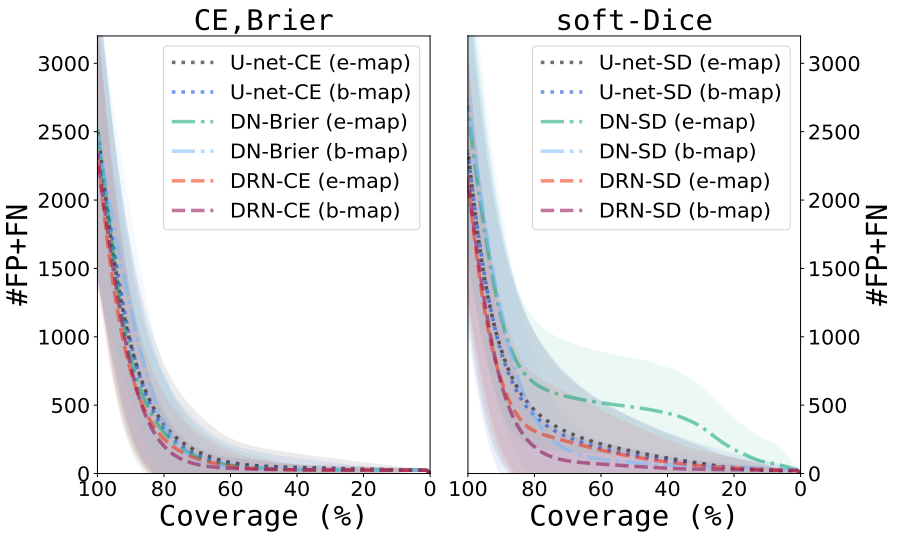


Figure 3.4: Comparison of risk-coverage curves for different combination of model architectures, loss functions and uncertainty maps. Results are separated for loss functions (left cross-entropy and Brier, right soft-Dice loss). 100% coverage means that none of the voxels is discarded based on its uncertainty whereas a coverage of 0% denotes the scenario in which all predictions are replaced by their reference labels. Note, all models were trained with two different loss functions (1) soft-Dice (SD) for all models (2) cross-entropy (CE) for DRN and U-net and Brier loss for DN.

3.6.4 Correction of automatically identified segmentation failures

Simulated correction: The results listed in Table 3.1 and 3.3 show that the proposed method consisting of segmentation followed by simulated manual correction of detected segmentation failures delivers accurate segmentation for all tissues over ED and ES points. Correction of detected segmentation failures improved the performance in terms of DC, HD and clinical metrics for all combinations of model architectures, loss functions and uncertainty measures. Focusing on the DC after correction of detected segmentation failures the results reveal that performance differences between evaluated models decreased compared to the segmentation-only task. This effect is less pronounced for HD where the DRN network clearly achieved superior results in the segmentation-only and combined approach. The DN performs the least of all models but achieves the highest absolute DC performance improvements in the combined approach for RV at ES. Overall, the results in Table 3.1 disclose that improvements attained by the combined approach are almost all statistically significant ($p \leq 0.05$) at ES and frequently at ED (96% resp. 83% of the cases). Moreover, improvements are in 99% of the cases statistically significant for HD compared to 81% of the cases for DC.

Results in terms of clinical metrics shown in Table 3.3 are inline with these findings. We observe that segmentation followed by simulated manual correction of detected

segmentation failures resulted in considerably higher accuracy for LV and RV ejection fraction. Achieved improvements for clinical metrics are only statistically significant ($p \leq 0.05$) in one case for RV ejection fraction.

In general, the effect of correction of detected segmentation failures is more pronounced in cases where the segmentation-only approach achieved relatively low accuracy (e.g. DN-SD for RV at ES). Furthermore, performance gains are largest for RV and LV at ES and for ejection fraction of both ventricles.

The best overall performance is achieved by the DRN model trained with cross-entropy loss while exploiting entropy maps in the detection task. Moreover, the proposed two step approach attained slightly better results using Bayesian maps compared to entropy maps.

Manual correction: Table 3.5 lists results for the combined automatic segmentation and detection approach followed by *manual* correction of detected segmentation failures by an expert. The results show that this correction led to improved segmentation performance in terms of DC, HD and clinical metrics. Improvements in terms of HD are in 50 percent of the cases statistically significant ($p \leq 0.05$) and most pronounced for RV and LV at end-systole.

Qualitative examples of the proposed approach are visualized in Figures 3.5 and 3.6 for simulated correction and manual correction of the automatically detected segmentation failures, respectively. For the illustrated cases (simulated) manual correction of detected segmentation failures leads to increased segmentation performance. On average manual correction of automatic segmentations took less than 2 minutes for ED and ES volumes of one patient compared to 20 minutes that is typically needed by an expert for the same task.

3.7 Ablation Study

To demonstrate the effect of different hyper-parameters in the method, a number of experiments were performed. In the following text these are detailed.

3.7.1 Impact of number of Monte Carlo samples on segmentation performance

To investigate the impact of the number of Monte Carlo (MC) samples (T) on the segmentation performance validation experiments were performed for all three segmentation architectures (Dilated Network, Dilated Residual Network and U-net) using $T \in \{1, 3, 5, 7, 10, 20, 30, 60\}$ samples. Results of these experiments are listed in Table 3.6. We observe that segmentation performance started to converge using 7 samples only. Performance improvements using an increased number of MC samples were largest for the Dilated Network. Overall, using more than 10 samples did not increase segmentation performance. Hence, in the presented work T was set to 10.

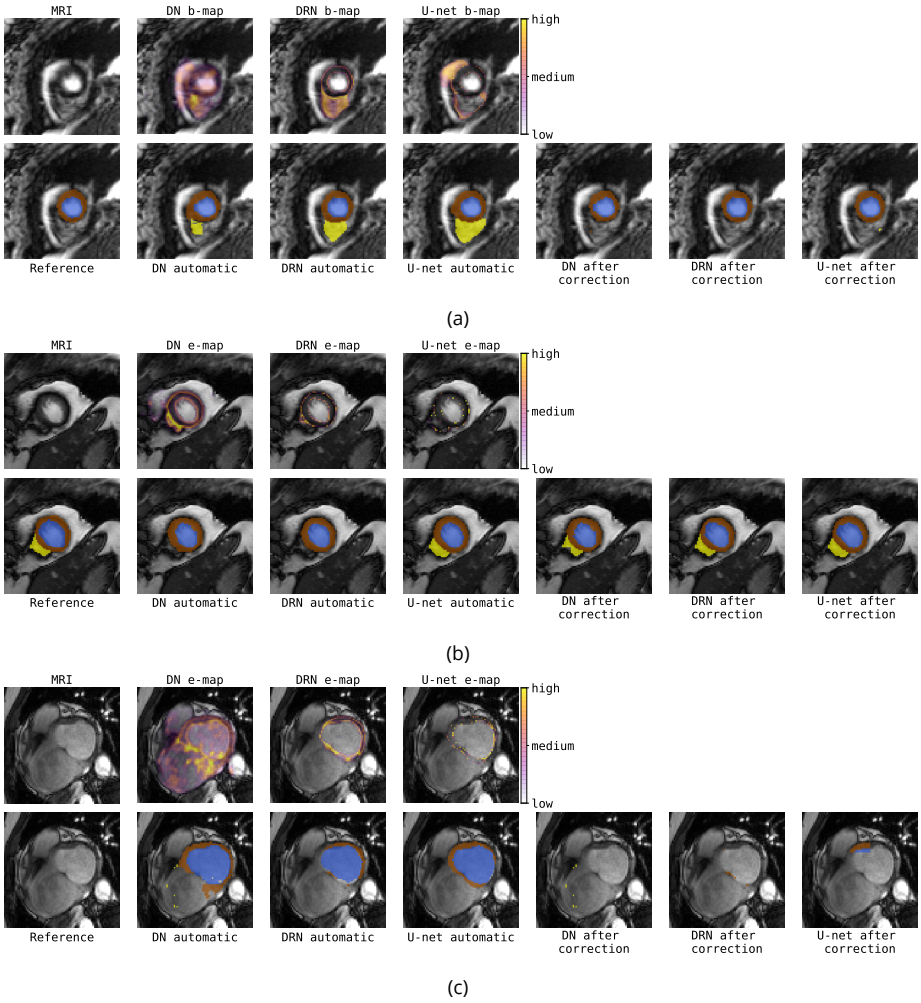


Figure 3.5: Three patients showing results of combined segmentation and detection approach consisting of segmentation followed by simulated manual correction of detected segmentation failures. First column shows MRI (top) and reference segmentation (bottom). Results for automatic segmentation and simulated manual correction respectively achieved by: Dilated Network (DN-Brier, 2nd and 5th columns); Dilated Residual Network (DRN-soft-Dice, 3rd and 6th columns); and U-net (soft-Dice, 4th and 7th columns).

3.7.2 Effect of patch-size on detection performance

The combined segmentation and detection approach detects segmentation failures on region level. To investigate the effect of patch-size on detection performance three different patch-sizes were evaluated: 4×4, 8×8, and 16×16 voxels. The results are shown in Figure 3.7. We can observe in Figure 3.7a that larger patch-sizes result in a lower

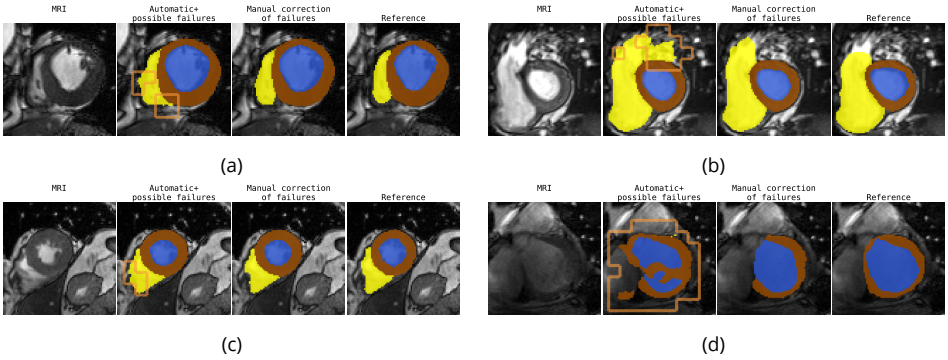


Figure 3.6: Four patients showing results of combined segmentation and detection approach consisting of segmentation followed by manual expert correction of detected segmentation failures. Expert was only allowed to adjust the automatic segmentations in regions where the detection network predicted segmentation failures (orange contour shown in 2nd column). Automatic segmentations were generated by a U-net trained with the cross-entropy loss. Segmentation failure detection was performed using entropy maps.

number of false positive regions. The result is potentially caused by the decreasing number of regions in an image when using larger patch-sizes compared to smaller patch-sizes. Furthermore, Figure 3.7b reveals that slice detection performance is only slightly influenced by patch-size. To ease manual inspection and correction by an expert, it is desirable to keep region-size i.e. patch-size small. Therefore, in the experiments a patch-size of 8×8 voxels was used.

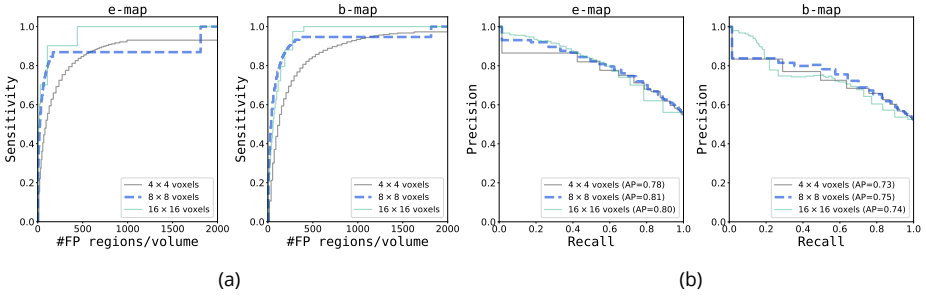


Figure 3.7: Detection performance for three different patch-sizes specified in voxels. (a) Sensitivity for detection of segmentation failures on voxel level (y-axis) versus number of false positive image regions (x-axis). (b) Precision-recall curve for detection of slices containing segmentation failures (where AP denotes average precision). Results are split between entropy and Bayesian uncertainty maps. In the experiments patch-size was set to 8×8 voxels.

3.7.3 Impact of tolerance threshold on number of segmentation failures

To investigate the impact of the tolerance threshold separating segmentation failures and tolerable segmentation errors, we calculated the ratio of the number of segmentation failures and all errors i.e. the sum of tolerable errors and segmentation failures. Figure 3.8 shows the results. We observe that at least half of the segmentation failures are located within a tolerance threshold i.e. distance of two to three voxels of the target structure boundary as defined by the reference annotation. Furthermore, the mean percentage of failures per volume is considerably lower for the Dilated Residual Network (DRN) and highest for the Dilated Network. This result is inline with our earlier finding (see Table 3.4) that average percentage of slices that do contain segmentation failures is lowest for the DRN model.

3.8 Discussion

We have described a method that combines automatic segmentation and assessment of uncertainty in cardiac MRI with detection of image regions containing segmentation failures. The results show that combining automatic segmentation with manual correction of detected segmentation failures results in higher segmentation performance. In contrast to previous methods that detected segmentation failures per patient or per structure, we showed that it is feasible to detect segmentation failures per image region. In most of the experimental settings, simulated manual correction of detected segmentation failures for LV, RV and LVM at ED and ES led to statistically significant improvements. These results represent the upper bound on the maximum achievable performance for the manual expert correction task. Furthermore, results show that manual expert correction of detected segmentation failures led to consistently improved segmentations. However, these results are not on par with the simulated expert correction scenario. This is not surprising because inter-observer variability is high for the presented task and annotation protocols may differ between clinical environments. Moreover, qualitative results of the manual expert correction reveal that manual correction of the detected segmentation failures can prevent anatomically implausible segmentations (see Figure 3.6). Therefore, the presented approach can potentially simplify and accelerate correction process and has the capacity to increase the trustworthiness of existing automatic segmentation methods in daily clinical practice.

The proposed combined segmentation and detection approach was evaluated using three state-of-the-art deep learning segmentation architectures. The results suggest that our approach is generic and applicable to different model architectures. Nevertheless, we observe noticeable differences between the different combination of model architectures, loss functions and uncertainty measures. In the segmentation-only task the DRN clearly outperforms the other two models in the evaluation of the boundary of the segmented structure. Moreover, qualitative analysis of the automatic segmentation

masks suggests that DRN generates less often anatomically implausible and fragmented segmentations than the other models. We assume that clinical experts would prefer such segmentations although they are not always perfect. Furthermore, even though DRN and U-net achieve similar performance in regard to DC we assume that less fragmented segmentation masks would increase trustworthiness of the methods.

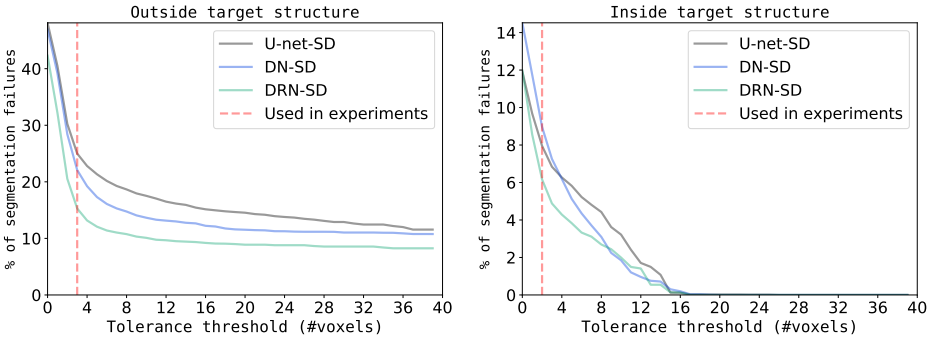


Figure 3.8: Mean percentage of the segmentation failures per volume (y-axis) in the set of all segmentation errors (tolerable errors+segmentation failures) depending on the tolerance threshold (x-axis). Red, dashed vertical line indicates threshold value that was used throughout the experiments. Results are split between segmentation errors located (a) outside and (b) inside the target structure. Each figure contains a curve for U-net, Dilated Network (DN) and Dilated Residual Network (DRN) trained with the soft-Dice (SD) loss. Segmentation errors located in slices above the base or below the apex are always included in the set of segmentation failures and therefore, they are independent of the applied tolerance threshold.

In agreement with our preliminary work we found that uncertainty maps obtained from a segmentation model trained with soft-Dice loss have a lower degree of uncertainty calibration compared to when trained with one of the other two loss functions (cross-entropy and Brier).[25] Nevertheless, the results of the combined segmentation and detection approach showed that a lower degree of uncertainty calibration only slightly deteriorated the detection performance of segmentation failures for the larger segmentation models (DRN and U-net) when exploiting uncertainty information from e-maps. Hendrycks and Gimpel[37] showed that softmax probabilities generated by deep learning networks have poor direct correspondence to confidence. However, in agreement with Geifman *et al.*[30] we presume that probabilities and hence corresponding entropies obtained from softmax function are ranked consistently i.e. entropy can potentially be used as a relative uncertainty measure in deep learning. In addition, we detect segmentation failures per image region and therefore, our approach does not require perfectly calibrated uncertainty maps. Furthermore, results of the combined segmentation and detection approach revealed that detection performance of segmentation failures using b-maps is almost independent of the loss function used to

train the segmentation model. In line with Jungo *et al.*[38] we assume that by enabling MC-dropout in testing and computing the mean softmax probabilities per class leads to better calibrated probabilities and b-maps. This assumption is in agreement with Srivastava *et al.*[39] where a CNN with dropout used at testing is interpreted as an ensemble of models.

Quantitative evaluation in terms of Dice coefficient and Hausdorff distance reveals that proposed combined segmentation and detection approach leads to significant performance increase. However, the results also demonstrate that the correction of the detected failures allowed by the combined approach does not lead to statistically significant improvement in clinical metrics. This is not surprising because state-of-the-art automatic segmentation methods are not expected to lead to large volumetric errors[7] and standard clinical measures are not sensitive to small segmentation errors. Nevertheless, errors of the current state-of-the-art automatic segmentation methods may lead to anatomically implausible segmentations[7] that may cause distrust in clinical application. Besides increase in trustworthiness of current state-of-the-art segmentation methods for cardiac MRIs, improved segmentations are a prerequisite for advanced functional analysis of the heart e.g. motion analysis[40] and very detailed morphology analysis such as myocardial trabeculae in adults.[41]

For the ACDC dataset used in this manuscript, Bernard *et al.*[7] reported inter-observer variability ranging from 4 to 14.1 mm (equivalent to on average 2.6 to 9 voxels). To define the set of segmentation failures, we employed a strict tolerance threshold on distance metric to distinguish between tolerated segmentation errors and segmentation failures (see Ablation study). Stricter tolerance threshold was used because the thresholding is performed in 2D, while evaluation of segmentation is done in 3D. Large slice thickness in cardiac MR could lead to a discrepancy between the two. As a consequence of this strict threshold results listed in Table 3.4 show that almost all patient volumes contain at least one slice with a segmentation failure. This might render the approach less feasible in clinical practice. Increasing the threshold decreases the number of segmentation failures and slices containing segmentation failures (see Figure 3.8) but also lowers the upper bound on the maximum achievable performance. Therefore, to show the potential of our proposed approach we chose to apply a strict tolerance threshold. Nevertheless, we realize that although manual correction of detected segmentation failures leads to increased segmentation accuracy the performance of precision-recall is limited (see Figure 3.3) and hence, should be a focus of future work.

The presented patch-based detection approach combined with (simulated) manual correction can in principle lead to stitching artefacts in the resulting segmentation masks. A voxel-based detection approach could potentially solve this. However, voxel-based detection methods are more challenging to train due to the very small number of voxels in an image belonging to the set of segmentation failures.

Evaluation of the proposed approach for 12 possible combinations of segmentation

models (three), loss functions (two) and uncertainty maps (two) resulted in an extensive number of experiments. Nevertheless, future work could extend evaluation to other segmentation models, loss functions or combination of losses. Furthermore, our approach could be evaluated using additional uncertainty estimation techniques e.g. by means of ensembling of networks[42] or variational dropout.[43] In addition, previous work by Kendall and Gal,[44] Tanno *et al.*[45] has shown that the quality of uncertainty estimates can be improved if model (epistemic) and data (aleatoric) uncertainty are assessed simultaneously with separate measures. The current study focused on the assessment of model uncertainty by means of MC-dropout and entropy which is a combination of epistemic and aleatoric uncertainty. Hence, future work could investigate whether additional estimation of aleatoric uncertainty improves the detection of segmentation failures.

Furthermore, to develop an end-to-end approach future work could incorporate the detection of segmentation failures into the segmentation network. Besides, adding the automatic segmentations to the input of the detection network could increase the detection performance.

Finally, the proposed approach is not specific to cardiac MRI segmentation. Although data and task specific training would be needed the approach could potentially be applied to other image modalities and segmentation tasks.

3.9 Conclusion

A method combining automatic segmentation and assessment of segmentation uncertainty in cardiac MR with detection of image regions containing local segmentation failures has been presented. The combined approach, together with simulated and manual correction of detected segmentation failures, increases performance compared to segmentation-only. The proposed method has the potential to increase trustworthiness of current state-of-the-art segmentation methods for cardiac MRIs.

References

[1] H. D. White, R. M. Norris, M. A. Brown, P. W. Brandt, R. Whitlock, and C. J. Wild. "Left ventricular end-systolic volume as the major determinant of survival after recovery from myocardial infarction." *Circulation*, vol. 76 (1987), pp. 44–51.

[2] F. Grothues, G. C. Smith, J. C. Moon, N. G. Bellenger, P. Collins, H. U. Klein, and D. J. Pennell. "Comparison of interstudy reproducibility of cardiovascular magnetic resonance with two-dimensional echocardiography in normal subjects and in patients with heart failure or left ventricular hypertrophy," *The American journal of cardiology*, vol. 90 (2002), pp. 29–34.

[3] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez. "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42 (2017), pp. 60–88.

[4] T. Leiner, D. Rueckert, A. Suinesiaputra, B. Baeßler, R. Nezafat, I. Išgum, and A. A. Young. "Machine learning in cardiovascular magnetic resonance: basic concepts and applications," *Journal of Cardiovascular Magnetic Resonance*, vol. 21 (2019), p. 61.

[5] C. Petitjean and J.-N. Dacher. "A review of segmentation methods in short axis cardiac mr images," *Medical image analysis*, vol. 15 (2011), pp. 169–184.

[6] P. Peng, K. Lekadir, A. Gooya, L. Shao, S. E. Petersen, and A. F. Frangi. "A review of heart chamber segmentation for structural and functional analysis using cardiac magnetic resonance imaging," *Magnetic Resonance Materials in Physics, Biology and Medicine*, vol. 29 (2016), pp. 155–195.

[7] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester, et al. "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?" *IEEE Transactions on Medical Imaging* (2018).

[8] A. Suinesiaputra, D. A. Bluemke, B. R. Cowan, M. G. Friedrich, C. M. Kramer, R. Kwong, S. Plein, J. Schulz-Menger, J. J. Westenberg, A. A. Young, et al. "Quantification of lv function and mass by cardiovascular magnetic resonance: multi-center variability and consensus contours," *Journal of Cardiovascular Magnetic Resonance*, vol. 17 (2015), p. 63.

[9] L. K. Tan, R. A. McLaughlin, E. Lim, Y. F. Abdul Aziz, and Y. M. Liew. "Fully automated segmentation of the left ventricle in cine cardiac MRI using neural network regression," *Journal of Magnetic Resonance Imaging*, vol. 48 (2018), pp. 140–152.

[10] Q. Zheng, H. Delingette, N. Duchateau, and N. Ayache. "3-d consistent and robust segmentation of cardiac images by deep learning with spatial propagation," *IEEE transactions on medical imaging*, vol. 37 (2018), pp. 2137–2148.

[11] N. Savioli, M. S. Vieira, P. Lamata, and G. Montana. "Automated segmentation on the entire cardiac cycle using a deep learning work-flow," *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, IEEE. 2018, pp. 153–158.

[12] W. Bai, M. Sinclair, G. Tarroni, O. Oktay, M. Rajchl, G. Vaillant, A. M. Lee, N. Aung, E. Lukaschuk, M. M. Sanghvi, et al. "Automated cardiovascular magnetic resonance image analysis with fully convolutional networks," *Journal of Cardiovascular Magnetic Resonance*, vol. 20 (2018), p. 65.

[13] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. De Marvao, T. Dawes, D. P. O'Regan, et al. "Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation," *IEEE transactions on medical imaging*, vol. 37 (2017), pp. 384–395.

[14] J. Duan, G. Bello, J. Schlemper, W. Bai, T. J. Dawes, C. Biffi, A. de Marvao, G. Doumou, D. P. O'Regan, and D. Rueckert. "Automatic 3d bi-ventricular segmentation of cardiac images by a shape-refined multi-task deep learning approach," *IEEE transactions on medical imaging* (2019).

[15] N. Painchaud, Y. Skandarani, T. Judge, O. Bernard, A. Lalande, and P.-M. Jodoin. "Cardiac MRI segmentation with strong anatomical guarantees," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. 2019, pp. 632–640.

[16] X. Albà, K. Lekadir, M. Pereañez, P. Medrano-Gracia, A. A. Young, and A. F. Frangi. "Automatic initialization and quality control of large-scale cardiac MRI segmentations," *Medical image analysis*, vol. 43 (2018), pp. 129–141.

[17] R. Robinson, V. V. Valindria, W. Bai, O. Oktay, B. Kainz, H. Suzuki, M. M. Sanghvi, N. Aung, J. M. Paiva, F. Zemrak, et al. "Automated quality control in image segmentation: application to the uk biobank cardiovascular magnetic resonance imaging study," *Journal of Cardiovascular Magnetic Resonance*, vol. 21 (2019), pp. 1–14.

[18] V. V. Valindria, I. Lavdas, W. Bai, K. Kamnitsas, E. O. Aboagye, A. G. Rockall, D. Rueckert, and B. Glocker. "Reverse classification accuracy: predicting segmentation performance in the absence of ground truth," *IEEE transactions on medical imaging*, vol. 36 (2017), pp. 1597–1606.

[19] K. Frounchi, L. C. Briand, L. Grady, Y. Labiche, and R. Subramanyan. "Automating image segmentation verification and validation by learning test oracles," *Information and Software Technology*, vol. 53 (2011), pp. 1337–1348.

[20] T. Kohlberger, V. Singh, C. Alvino, C. Bahlmann, and L. Grady. "Evaluating segmentation error without ground truth," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. 2012, pp. 528–536.

[21] T. Nair, D. Precup, D. L. Arnold, and T. Arbel. "Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. 2018, pp. 655–663.

[22] A. Jungo, R. Meier, E. Ermis, E. Herrmann, and M. Reyes. "Uncertainty-driven sanity check: application to postoperative brain tumor cavity segmentation," *Medical Imaging with Deep Learning Conference*, 2018.

[23] A. G. Roy, S. Conjeti, N. Navab, C. Wachinger, A. D. N. Initiative, et al. "Bayesian quicknat: model uncertainty in deep whole-brain segmentation for structure-wise quality control," *NeuroImage*, vol. 195 (2019), pp. 11–22.

[24] T. DeVries and G. W. Taylor. "Leveraging uncertainty estimates for predicting segmentation quality," *arXiv preprint arXiv:1807.00502* (2018).

[25] J. Sander, B. D. de Vos, J. M. Wolterink, and I. Išgum. "Towards increased trustworthiness of deep learning segmentation methods on cardiac MRI," *Medical Imaging 2019: Image Processing*, vol. 10949 International Society for Optics and Photonics. (2019), p. 1094919.

[26] Y. Gal and Z. Ghahramani. "Dropout as a bayesian approximation: representing model uncertainty in deep learning," *International Conference on Machine Learning (ICML)*, 2016, pp. 1050–1059.

[27] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum. "Automatic segmentation and disease classification using cardiac cine MR images," *International Workshop on Statistical Atlases and Computational Models of the Heart*, Springer. 2017, pp. 101–110.

[28] F. Yu, V. Koltun, and T. Funkhouser. "Dilated residual networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*, Code available at `https://github.com/fyu/drn`. 2017, pp. 472–480.

[29] O. Ronneberger, P. Fischer, and T. Brox. "U-net: convolutional networks for biomedical image segmentation," *International Conference on Medical image computing and computer-assisted intervention*, Springer. 2015, pp. 234–241.

[30] Y. Geifman and R. El-Yaniv. "Selective classification for deep neural networks," *Advances in neural information processing systems*, 2017, pp. 4878–4887.

[31] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[32] F. Isensee, P. F. Jaeger, P. M. Full, I. Wolf, S. Engelhardt, and K. H. Maier-Hein. "Automatic cardiac disease assessment on cine-mri via time-series segmentation and domain specific features," *International workshop on statistical atlases and computational models of the heart*, Springer. 2017, pp. 120–129.

[33] F. Milletari, N. Navab, and S.-A. Ahmadi. "V-net: fully convolutional neural networks for volumetric medical image segmentation," *2016 fourth international conference on 3D vision (3DV)*, IEEE. 2016, pp. 565–571.

[34] G. W. Brier. "Verification of forecasts expressed in terms of probability," *Monthey Weather Review*, vol. 78 (1950), pp. 1–3.

[35] G. Huang, Y. Li, G. Pleiss, Z. Liu, J. E. Hopcroft, and K. Q. Weinberger. "Snapshot ensembles: train 1, get m for free," *arXiv preprint arXiv:1704.00109* (2017).

[36] D. Kingma and J. Ba. "Adam: a method for stochastic optimization," *ICLR*, vol. 5 (2015).

[37] D. Hendrycks and K. Gimpel. "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv preprint arXiv:1610.02136* (2016).

[38] A. Jungo and M. Reyes. "Assessing reliability and challenges of uncertainty estimations for medical image segmentation," *arXiv preprint arXiv:1907.03338* (2019).

[39] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15 (2014), pp. 1929–1958.

[40] G. A. Bello, T. J. Dawes, J. Duan, C. Biffi, A. de Marvao, L. S. Howard, J. S. R. Gibbs, M. R. Wilkins, S. A. Cook, D. Rueckert, et al. "Deep-learning cardiac motion analysis for human survival prediction," *Nature machine intelligence*, vol. 1 (2019), pp. 95–104.

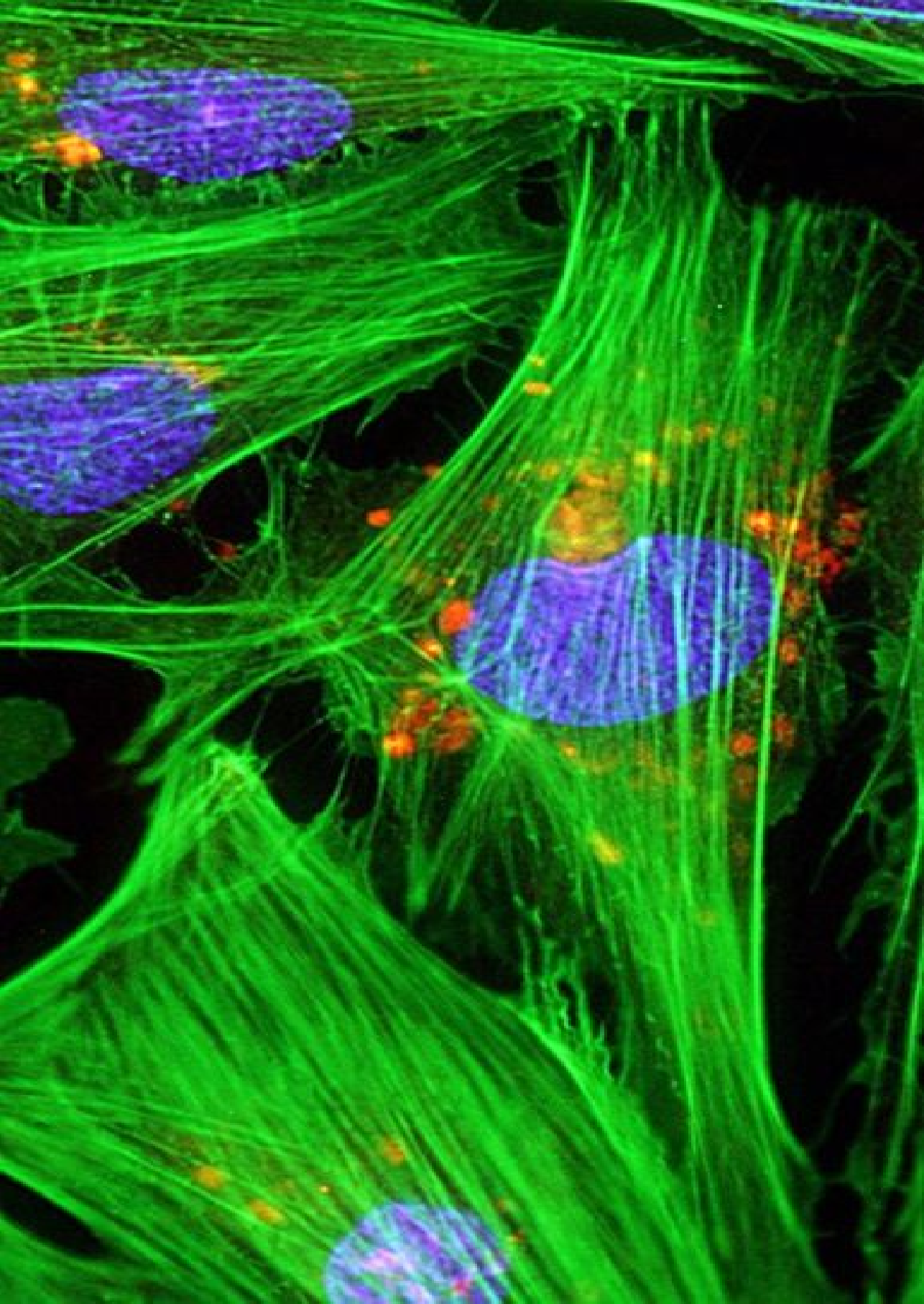
[41] H. V. Meyer, T. J. Dawes, M. Serrani, W. Bai, P. Tokarczuk, J. Cai, A. de Marvao, A. Henry, R. T. Lumbers, J. Gierten, et al. "Genetic and functional insights into the fractal structure of the heart," *Nature* (2020), pp. 1–6.

[42] B. Lakshminarayanan, A. Pritzel, and C. Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in Neural Information Processing Systems*, 2017, pp. 6402–6413.

[43] D. P. Kingma, T. Salimans, and M. Welling. "Variational dropout and the local reparameterization trick," *Advances in neural information processing systems*, 2015, pp. 2575–2583.

[44] A. Kendall and Y. Gal. "What uncertainties do we need in bayesian deep learning for computer vision?" *Advances in neural information processing systems*, 2017, pp. 5574–5584.

[45] R. Tanno, D. Worrall, E. Kaden, A. Ghosh, F. Grussu, A. Bizzi, S. N. Sotiropoulos, A. Criminisi, and D. C. Alexander. "Uncertainty quantification in deep learning for safer neuroimage enhancement," *arXiv preprint arXiv:1907.13418* (2019).



CHAPTER 4

Towards automatic classification of cardiovascular magnetic resonance task force criteria for diagnosis of arrhythmogenic right ventricular cardiomyopathy

---

This chapter is based on: M. Bourfiss*, J. Sander*, B. D. de Vos, A. S. Te Riele, F. W. Asselbergs, I. Išgum, and B. K. Velthuis. "Towards automatic classification of cardiovascular magnetic resonance task force criteria for diagnosis of arrhythmogenic right ventricular cardiomyopathy," *Clinical Research in Cardiology* (2022), pp. 1–16.[1]

[1] ***Both authors contributed equally**.

Abstract

BACKGROUND Arrhythmogenic right ventricular cardiomyopathy (ARVC) is diagnosed according to the Task Force Criteria (TFC) in which cardiovascular magnetic resonance (CMR) imaging plays an important role. Our study aims to apply an automatic deep learning-based segmentation for right and left ventricular CMR assessment and evaluate this approach for classification of the CMR TFC.

METHODS We included 227 subjects suspected of ARVC who underwent CMR. Subjects were classified into 1) ARVC patients fulfilling TFC; 2) at-risk family members; and 3) controls. To perform automatic segmentation, a Bayesian Dilated Residual Neural Network was trained and tested. Performance of automatic versus manual segmentation was assessed using Dice-coefficient and Hausdorff distance. Since automatic segmentation is most challenging in basal slices, manual correction of the automatic segmentation in the most basal slice was simulated (automatic$^{\text{-basal}}$). CMR TFC calculated using manual and automatic$^{\text{-basal}}$ segmentation were compared using Cohen's Kappa (κ).

RESULTS Automatic segmentation was trained on CMRs of 70 subjects (39.6±18.1 years, 47% female) and tested on 157 subjects (36.9±17.6 years, 59% female). Dice-coefficient and Hausdorff distance showed good agreement between manual and automatic segmentations (≥0.89 and ≤10.6mm, respectively) which further improved after simulated correction of the most basal slice (≥0.92 and ≤9.2mm, $p<0.001$). Pearson correlation of volumetric and functional CMR measurements was good to excellent (automatic ($r=0.78$-0.99, $p<0.001$) and automatic$^{\text{-basal}}$ ($r=0.88$-0.99, $p<0.001$) measurements). CMR TFC classification using automatic$^{\text{-basal}}$ segmentations was comparable to manual segmentations (κ 0.98±0.02) with comparable diagnostic performance.

CONCLUSIONS Combining automatic segmentation of CMRs with correction of the most basal slice results in accurate CMR TFC classification of subjects suspected of ARVC.

4.1 Background

Arrhythmogenic right ventricular cardiomyopathy (ARVC) is an inherited heart disease that is characterized by ventricular dysfunction, predominantly affecting the right ventricle (RV), and potentially life-threatening ventricular arrhythmias. Accurate recognition of this disease is vital since the implantation of an implantable cardioverter defibrillator can be life-saving. ARVC is diagnosed according to the revised 2010 Task Force Criteria (TFC).[1] Apart from electrical and family history criteria, an important role is given to the assessment of ventricular dysfunction and structural alterations. Cardiac magnetic resonance (CMR) imaging is the modality of choice for the assessment of cardiac function and dimensions in ARVC[2] since the asymmetric geometry and the position of the RV in the chest hampers visualization of the entire RV by 2-D echocardiography.[3] The CMR TFC are based on RV regional wall motion abnormalities combined with cut-off values for RV ejection fraction (EF) or sex-specific cut-off values for RV indexed end-diastolic volume (EDVI).[1] CMR can deliver one minor or two major points of the necessary four TFC points for an ARVC diagnosis. Therefore, accurate RV assessment is essential. Segmenting CMRs to measure functional and structural parameters is a laborious task, taking about 25 minutes to segment both ventricles in end-diastole (ED) and end-systole (ES).[4,5] Notably, RV segmentation takes two-thirds of this segmentation time and is prone to intra- and inter-observer variability.[6] RV segmentation difficulties can arise from the trabeculated and complex RV geometry.[7,8] In ARVC, RV and left ventricular (LV) anatomy can be further complicated by pathological wall thinning and aneurysms due to fibrofatty replacement of the myocardial wall.[2] As a consequence, CMR misinterpretations are a key cause of over-diagnosis in ARVC.[2] The use of automatic methods for the segmentation of the ventricles may overcome these challenges. Over the last few years many state-of-the-art deep learning segmentation approaches for short-axis CMR have been developed.[4,9–11] For automatic LV segmentation such methods can achieve performance level of human experts.[12,13] However, previous studies also demonstrated that in manual and automatic segmentation of short-axis CMR, the largest disagreements and errors occur in the most basal and apical slices.[8,12–15] Moreover, previous methods have often been evaluated on CMR datasets with limited pathology especially related to the RV. In contrast, this study included a large hospital population being evaluated for ARVC, including subjects with structurally normal hearts and those with complex structural abnormalities. In this work we apply a previously validated state-of-the-art segmentation approach[16] on a large heterogeneous hospital population of patients suspected of ARVC. The purpose of this study was to (i) evaluate our previously developed deep learning segmentation approach for RV and LV CMR assessment in patients suspected of ARVC, and (ii) evaluate the clinical implication of this approach for classification of the CMR TFC in subjects suspected of ARVC.

Table 4.1: Baseline characteristics of study population. Abbreviations: ARVC= arrhythmogenic right ventricular cardiomyopathy; CMRI= cardiac magnetic resonance imaging; DSP= desmoplakin; PKP2= plakophilin-2; PLN= phospholamban; TFC=Task Force Criteria. *: Significant difference between control and ARVC patients; †: Significant difference between control and at-risk subjects; ‡: Significant difference between ARVC patients and at-risk subjects.

	Study population			
	ARVC patients (n=37)	At-risk ARVC group (n=66)	Control group (n=54)	p-value
Demographics				
Age at CMRI (years)	39.1±19.0	30.7±16.2†‡	42.9±15.9	<0.001
Female (%)	20 (54)	43 (65)	29 (54)	0.37
Proband (%)	10 (27)	0 (0) ‡	na	<0.001
Genetic status				
Pathogenic variant	36 (97)	56 (85)	na	0.06
PKP2 (%)	24 (71)	33 (59)		
PLN (%)	4 (12)	22 (39)		
DSP (%)	4 (12)	1 (2)		
Other (%)	4 (12)	0		
Clinical phenotype				
Total TFC score	5 [4-6]*	2 [1-3]†‡	0	<0.001
Repolarization criteria				
Minor (%)	10 (27)	0 (0)		
Major (%)	8 (22)	3 (5)		
Depolarization criteria				
Minor (%)	23 (62)	9 (14)		
Major (%)	0 (0)	0 (0)		
Arrhythmia criteria				
Minor (%)	25 (68)	6 (9)		
Major (%)	2 (5)	0 (0)		
Structural criteria				
Minor (%)	6 (16)	1 (3)		
Major (%)	25 (68)	0 (0)		

4.2 Methods

4.2.1 Study population

We included a consecutive cohort of subjects suspected of ARVC who underwent CMR as part of their clinical evaluation between 2014 and 2019 at the University Medical Center (UMC) Utrecht. This yielded 241 subjects, of whom 14 were excluded because of an equivocal diagnosis (ARVC neither confirmed nor rejected) (n=12), prior chemotherapy (n=1) and imaging artefacts due to irregular heart rhythm (n=1). This

led to a study population of 227 subjects who were classified into three groups: 1) ARVC patients diagnosed according to the 2010 TFC (n=53), 2) family members at-risk of developing ARVC (n=96), and 3) control subjects initially suspected of ARVC but in whom ARVC was excluded after full clinical assessment (n=78). Diagnosis in the control patients included RV outflow tract tachycardia (n=45), premature ventricular contractions/non-sustained ventricular tachycardia (n=19), mutation-negative family members of mutation-positive ARVC patients (n=3), healthy athletes (n=3), syncope without a cardiac cause (n=3) repolarization abnormalities with a structurally normal heart (n=3) and pectus excavatum (n=2). This study was reviewed by the UMC Utrecht Institutional Review Board and was granted a waiver of informed consent.

4.2.2 ARVC diagnosis

ARVC diagnosis was based on the revised 2010 diagnostic TFC.[1] In short, these consensus-based criteria rely on major and minor criteria for six different categories: 1) global and regional dysfunction and structural alterations, 2) tissue characterization, 3) repolarization abnormalities, 4) depolarization/conduction abnormalities, 5) arrhythmias, and 6) family history/genetics. In each of these six categories subjects can score a minor criterium (one point), a major criterium (two points) or no criteria (0 points). A definite ARVC diagnosis was made if a subject has at least four points. The first category can be assessed by CMR, with minor criteria for regional RV wall motion abnormalities plus RVEF >40 to ≤45% or RVEDVI ≥100 to <110 mL/m^2 (males) or ≥90 to <100 mL/m^2 (females) and major criteria for RV regional wall motion abnormalities plus RVEF ≤40% or RVEDVI ≥110mL/m^2 (males) or ≥100 mL/m^2 (females).[1]

4.2.3 CMR dataset

All subjects underwent CMR using either 1.5 or 3 Tesla Ingenia or Achieva Philips scanners (Best, the Netherlands). The CMR dataset consisted of conventional steady-state free precession sequence short-axis and longitudinal-axis (4-chamber, 2-chamber and 3-chamber of both ventricles) cine CMR images acquired during breath holds. For this work, we only included the short-axis CMR volumes consisting of 12-18 contiguous slices covering both ventricles. The short-axis imaging parameters were as follows: each slice containing 25 to 40 phases covering one cardiac cycle with repetition time 2.6-3.4 ms and echo time 1.3-1.7 ms, flip angle 45-60 degrees. The CMR images have an in-plane resolution ranging from 1.11 to 1.45 with a slice thickness varying from 7 to 10 mm. Furthermore, reconstruction matrix of images ranges from 240x240 to 288x288 voxels. Expert radiology technicians made manual reference segmentations of the RV and LV endocardium for all CMR slices at ED and ES time frames. Both time points were manually chosen by the same experts. The CMR segmentation protocol was published previously[17] and adheres to the guidelines of the Society of Cardiovascular Magnetic Resonance (SCMR).[18] Furthermore, the presence of RV and/or LV wall motion

abnormalities was visually evaluated by an experienced cardiovascular radiologist on all available cine images and used for the calculation of the CMR TFC.

4.2.4 Automatic segmentation of CMR

Prior to segmentation, voxel intensities in CMR scans were normalized by rescaling the values between [0,1] based on their 1st and 99th percentiles per scan. Furthermore, voxels intensities below or above the 1st and 99th percentiles were clamped to 0 and 1, respectively. To perform automatic segmentation of RV and LV in the 2D short-axis CMR images, we trained a Bayesian Dilated Residual Neural Network (DRN)[19] that was previously developed and evaluated by Sander *et al.*[16] The Bayesian DRN was based on the original DRN from Yu *et al.*[19] for image segmentation. To convert the original DRN[19] into a Bayesian DRN, we implemented Monte Carlo dropout (MC dropout) introduced by Gal & Ghahramani.[20] Using a Bayesian i.e. MC dropout approach is advantageous because multiple predictions for the same voxel can be averaged to obtain an improved final prediction per voxel.[16] Furthermore, architecture and parameters of the Bayesian DRN were identical to the model described in Sander *et al.*[16] The network used a 2D CMR image as input and had three output channels, each providing probability for the LV, RV or background. Softmax probabilities were calculated over the three tissue classes. To train the model a combination of soft-Dice[21] and cross-entropy was used as loss function. For completeness, we provide the equations for both loss functions:

$$\text{soft-Dice}_c = 1 - \frac{\sum_{i=1}^{N} R_c(i)\, A_c(i)}{\sum_{i=1}^{N} R_c(i) + \sum_{i=1}^{N} A_c(i)}\,, \tag{4.1}$$

where N denotes the number of voxels in an image, R_c is the binary reference image for class c and A_c is the probability map for class c.

$$\text{Cross-Entropy}_c = 1 - \sum_{i=1}^{N} t_{ic}\ \log\ p(y_i = c|x_i)\,, \tag{4.2}$$

where p denotes the probability for a specific voxel x_i with corresponding reference label y_i for class c, and t_{ic} =1 if y_i = c, and 0 otherwise. Hyper-parameters of the network were determined in our previous work[16] using CMR images from the MICCAI 2017 Automated Cardiac Diagnosis Challenge (ACDC).[12] Therefore, no validation set was required in the current work. To train the model, patches of 160x160 voxels were randomly chosen from the training set. Training data was augmented by 90 degree rotations, elastic deformations and gamma transformations of the images. The model was trained for 160,000 iterations using mini-batch stochastic gradient descent with

batch-size 16 and Adam as optimizer.[22] Learning rate was set to 0.001 and decayed with a factor of 0.1 after every 40,000 steps. To increase generalization performance weight decay was used and set to 0.0005. Furthermore, dropout percentage was set to 0.1. Enabling MC dropout during testing, tissue class per voxel was determined using the mean softmax probabilities over 15 samples. Voxel wise segmentation may result in isolated (small clusters of) voxels. To address this, only the largest 3D connected component for each class was retained in the automatic segmentations.

Simulation of the correction of automatic segmentation Previous research demonstrated that most segmentation inaccuracies occur in the most basal slice on the CMR.[8,12–15] To evaluate whether these inaccuracies of our method impact TFC classification, correction of the automatic segmentation in the most basal slice of each CMR volume was simulated. This was achieved by replacing the automatic segmentation of the most basal slice with the corresponding manual reference defined by specially trained radiology technicians as a part of a regular clinical workup. We refer to the this scenario as automatic$^{\text{-basal}}$ hereafter.

4.2.5 Automatic ED/ES phase selection

Accurate identification of ED and ES phase in the cardiac cycle is a prerequisite to automatically compute RVEDV and RVESV. To show the potential of the method to automatically determine the ED and ES phase we automatically segmented all CMR volumes of the patients in the test set, and derived the RV and LV volumes for all time points of the cardiac cycle. For each patient ED was identified as the phase in which the fully automatically segmented volume was maximal and ES as the phase in which the volume was minimal. Automatically identified phases were compared with the manually selected phases using Bland-Altman analysis. In these plots (e.g. Figure 4.3) the distance between automatically and manually selected phases is expressed as percentage of a complete cardiac cycle. Evaluation was performed for RV and LV separately, and for automatic and automatic$^{\text{-basal}}$ segmentations separately.

4.2.6 Evaluation of automatic segmentation

To evaluate performance of the automatic segmentation method 3D Dice-coefficient and 3D Hausdorff distance between manual and automatic segmentations were computed. For this, the 2D automatic segmentation masks were stacked into a 3D volume per patient and cardiac phase. The Dice-coefficient quantifies overlap between manual and automatic segmentation and its value ranges between 0 and 1. A higher Dice-coefficient indicates better agreement between manual and automatic segmentations. The Hausdorff distance evaluates segmentation along the boundary of the target structure by measuring the maximum distance between manual and automatic segmentation. Qualitative performance of the automatic segmentation method was visually assessed. To investigate whether segmentation errors accumulate at specific slice locations in the

CMR volume the distribution of segmentation errors over slice location was computed. For this, four slice locations in a volume were distinguished: (i) most apical slice, (ii) most basal slice, (iii) mid-ventricular slices, and (vi) slices located below the apex or above the base of the heart. Furthermore, to evaluate the clinical implications of our automatic CMR segmentation approach for the classification of the CMR TFC in subjects suspected of ARVC, the following CMR measurements were computed for manual, automatic and automatic$^{\text{-basal}}$ segmentations: LV end-diastolic volume (EDV), LV end-systolic volume (ESV), LV stroke volume (SV), LVEF, RVEDV, RVESV, RVSV, and RVEF.

4.2.7 Statistical analysis

Statistical analysis was performed using RStudio Version 1.3.1093 (Boston, MA, USA) and IBM SPSS Statistics (version 25, USA). Continuous values were presented as mean ± standard deviation or median [interquartile range]. Categorical data was displayed as absolute frequency (n) and percentages (%). For continuous comparisons of two groups, two-tailed Student's t-test was used. For continuous comparisons of three or more groups, one-way ANOVA was used. Categorical data were compared using the chi-square χ^2 test. A p-value of <0.05 was considered significant. Comparison of automatic and manual absolute CMR measurements were assessed using Bland-Altman analysis and the Pearson correlation coefficient (r). CMR TFC was first classified using visual assessment of wall motion abnormalities and manually derived RVEDVI and RVEF, and next using visual assessment of wall motion abnormalities and automatically derived RVEDVI and RVEF. CMR TFC classification agreement between manually vs. automatically derived CMR measurements was assessed using Cohen's kappa (κ). Furthermore, sensitivity and specificity of CMR TFC by manual and automatic approach was determined and compared using the McNemar test.

4.3 Results

4.3.1 Study population

We included 70 subjects in the training set (mean age 39.6±18.1 years, 47% female) and 157 subjects in the test set (mean age 36.9±17.6 years, 59% female). Patient characteristics are shown in Table 4.1. The test set included 37 ARVC patients, 66 at-risk family members and 54 controls subjects. The distribution of subjects across the three patient categories was the same for training and test sets (34% controls, 42% at risk, 24% ARVC patients). No statistically significant difference in sex existed between the three subgroups ($p=0.37$), but at-risk family members were younger than ARVC patients ($p=0.021$) and controls ($p<0.001$). ARVC patients had a median of 5 [4-6] diagnostic TFC points, while at-risk family members had a median of 2 [1-3] points ($p<0.001$). In total, 84% of ARVC patients and 3% of at-risk family members had minor or major CMR

Table 4.2: Segmentation performance of deep learning segmentation model in terms of Dice-coefficient (higher is better) and Hausdorff distance (in millimeter, lower is better). Automatic^{-basal} refers to the scenario in which the most basal slice of each automatic segmentation volume was replaced with the corresponding manual reference. Depicted values specify mean ± standard deviation. Abbreviations: LV= left ventricle; RV= right ventricle.

	End-diastole		**End-systole**	
	LV	RV	LV	RV
Dice-coefficient				
Automatic	0.96±0.01	0.93±0.03	0.93±0.04	0.89±0.04
Automatic^{-basal}	0.97±0.01	0.95±0.02	0.95±0.02	0.92±0.03
Hausdorff-distance				
Automatic	6.42±2.26	10.42±2.99	6.58±2.73	10.60±3.50
Automatic^{-basal}	5.07±2.27	9.19±3.19	5.52±2.47	9.09±3.05

TFC (RV wall motion abnormalities combined with abnormal RVEF or RVEDVI cut-off values). Among 103 ARVC patients and at-risk family members, 90 (87%) carried a pathogenic variant, mostly in plakophilin-2 (n=57, 63%) followed by phospholamban (n=26, 29%) and desmoplakin (n=5, 6%).

4.3.2 Assessment of segmentation performance

Table 4.2 lists quantitative results of the automatic segmentation. The automatic method achieved mean Dice-coefficient for ED and ES 0.96±0.01 and 0.93±0.03, respectively for the LV and 0.93±0.04 and 0.89±0.04, respectively for the RV. Visual assessment of automatic segmentation results depicted in Figure 4.1 reveal that performance was higher for mid-ventricular slices (second and third rows Figure 4.1) compared with apical and basal slices (first and fourth row Figure 4.1), while an under-segmentation of trabeculated areas occurred in the apical slices (first row Figure 4.1). Furthermore, as depicted in Figure 4.2, visual assessment of the manual reference segmentation revealed a high variability of the RV shape in the basal slices in both ED and ES time points. Furthermore, as listed in Table 4.3, comparison of automatic with manual reference segmentations disclosed that on average 24.5% of the segmentation errors i.e. misclassified voxels were located in the most basal slice (30.7 and 18.3% for RV and LV, respectively). In contrast, on average only 6.5% of the errors were located in an apical slice (5.4 and 7.6% for RV and LV, respectively). Table 4.2 lists segmentation results after the simulated correction of the automatic RV and LV segmentation in the most basal slice. The results show an increased segmentation performance: mean Dice-coefficient for the ED and ES are 0.97±0.01 and 0.95±0.03 (vs. 0.96±0.01 and 0.93±0.03 uncorrected) respectively for the LV and 0.95±0.02 and 0.92±0.03 (vs. 0.93±0.04 and 0.89±0.04 uncorrected) respectively for the RV ($p<0.001$ [one side Wilcoxon signed-rank test]).

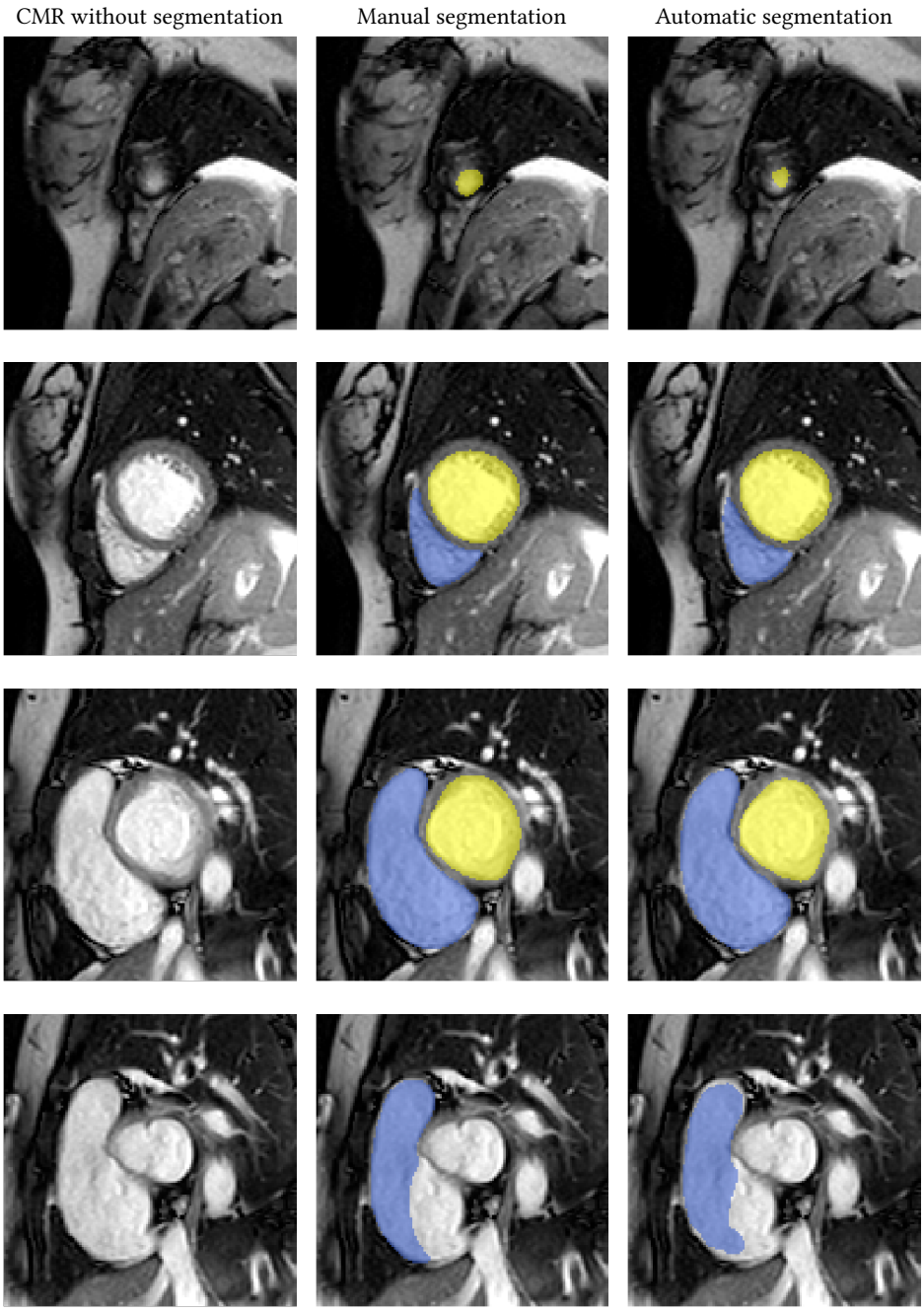


Figure 4.1: Qualitative segmentation results for left (yellow) and right (blue) ventricles at end-systole for a patient included in the test set. Columns depict raw CMR (first column), CMR with manual reference segmentation (second column) and CMR with automatic segmentation (third column). Rows show apical, mid-ventricular and most basal slices for LV (third row) and RV (fourth row), respectively.

Table 4.3: Percentage of segmentation errors per target structure, i.e., left and right ventricle (LV, RV), located in basal, apical, all mid-ventricular or slices above base and below apex.

	Slice location			
	Basal	Mid-ventricular	Apical	Above base & below apex
LV	18.3%	70%	7.5%	4.2%
RV	30.7%	61%	5.4%	2.9%

Table 4.4: Correlation between manual and automatic (second and third columns) and automatic^{-basal} (fourth and fifth columns) measurements. Abbreviations: EF=ejection fraction; SV=stroke volume; EDV=end-diastolic volume; ESV=end-systolic volume. *p-value of correlation <0.001.

	Automatic		**Automatic^{-basal}**	
	Mean absolute difference (vs. manual)	Correlation r (with manual)	Mean absolute difference (vs. manual)	Correlation r (with manual)
Right ventricle				
EF (%)	1.4±4.7	0.82 (0.77-0.87)*	0.9±3.9	0.88 (0.84-0.91)*
SV (mL)	-2.0±10.8	0.89 (0.84-0.91)*	0.7±9.2	0.92 (0.90-0.94)*
EDV (mL)	-9.9±13.9	0.95 (0.94-0.97)*	-5.5±9.6	0.98 (0.97-0.98)*
ESV (mL)	-7.9±11.0	0.95 (0.93-0.96)*	-4.8±8.1	0.97 (0.96-0.98)*
Left ventricle				
EF (%)	2.4±3.6	0.78 (0.71-0.84)*	1.4±2.1	0.92 (0.89-0.94)*
SV (mL)	1.4±7.3	0.93 (0.91-0.95)*	0.04±4.2	0.98 (0.97-0.98)*
EDV (mL)	-4.6±6.1	0.99 (0.98-0.99)*	-4.4±4.1	0.99 (0.99-1.00)*
ESV (mL)	-6.0±6.4	0.95 (0.93-0.96)*	-4.4±4.6	0.97 (0.96-0.98)*

4.3.3 Automatic ED and ES phase selection

The Bland-Altman plots shown in Figure 4.3a demonstrate the comparison between automatically identified cardiac phases using the automatic^{-basal} segmentations to automatically determine the ED and ES phases. For this scenario the bias [limits of agreement] were -0.72 [-5.29, 3.85]% for the ED-LV phase and -3.03 [-10.08, 4.03]% for the ES-LV phase, respectively, and -0.34 [-9.58, 8.89]% for the ED-RV and 0.48 [-7.20, 8.17]% for the ES-RV. Figure 4.3b depicts the same comparison using the automatic segmentations with the manually selected ED and ES phases. The bias [limits of agreement] were -0.87 [-6.26, 4.52]% for the ED-LV phase and -1.64 [-10.28, 6.99]% for the ES-LV phase, respectively, and -0.96 [-11.69, 9.76]% for the ED-RV phase and −0.05 [-7.62, 7.53]% for the ES-RV phase, respectively.

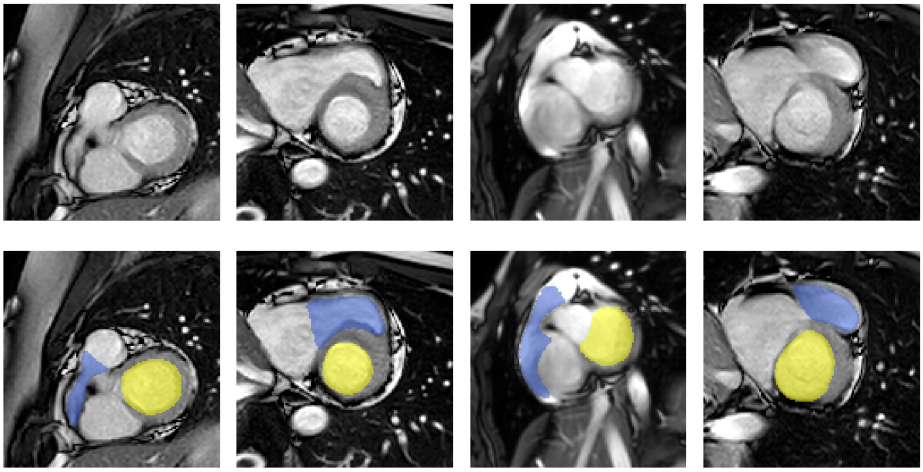


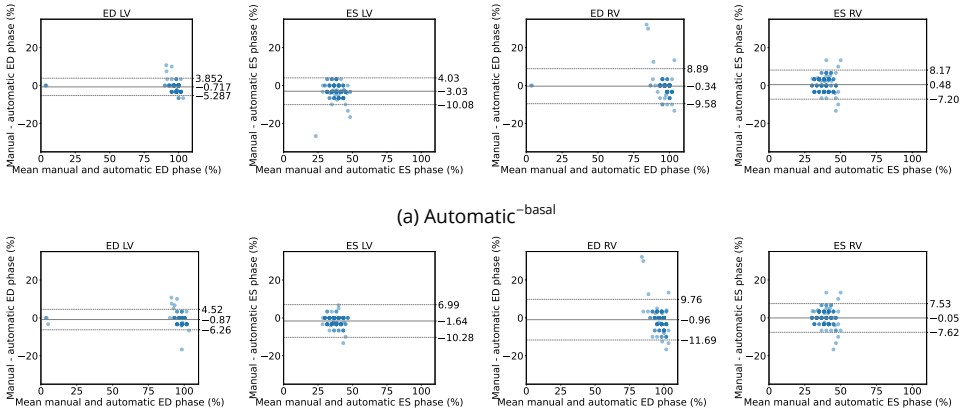
Figure 4.2: Examples illustrating right ventricle (RV) shape variability in manual reference segmentations for basal slices. Shown are original CMR without (top row) and with (bottom row) manual reference segmentations of the left (yellow) and right (blue) ventricle.

4.3.4 Assessment of absolute CMR measurements

Automatically measured volumes (RV and LV EDV and ESV) are slightly underestimated compared to manually measured volumes (see Figures 4.4). However, as shown in Table 4.4, the correlations of both RV and LV volumes were excellent (0.95-0.99, $p<0.001$). For RV and LV EF and SV, automatic measurements seem to be slightly overestimated compared to manual measurements; nonetheless, correlations were excellent 0.82-0.89 for RV and good to excellent (0.78-0.93) for LV. After simulated manual correction of the basal slice, agreement between manual and automated measurements increased, as depicted in the Bland Altman plots (see Figures 4.4). This was also reflected in the Pearson correlation coefficient for both the volumetric (EDV, ESV) ($r=0.97$-0.99, $p<0.001$) as well as the functional (SV, EF) ($r=0.88$-0.98, $p<0.001$) CMR measurements.

4.3.5 Classification of ARVC TFC

Since agreement between manual and automatic measurements was higher in the automatic$^{-\text{basal}}$, we used these results for the further analysis. Table 4.5 depicts the mean and standard deviation of the CMR measurements stratified per subgroup. The trends between the three subgroups (ARVC, at-risk family members and controls) were comparable between manual and automated measurements: ARVC patients had significantly reduced RVEF ($p<0.001$) and LVEF ($p=0.002$), as well as increased RVEDVI ($p<0.001$), RVESVI ($p<0.001$) and LVESVI ($p<0.013$) compared to at-risk family members



(a) Automatic^{-basal}

(b) Automatic

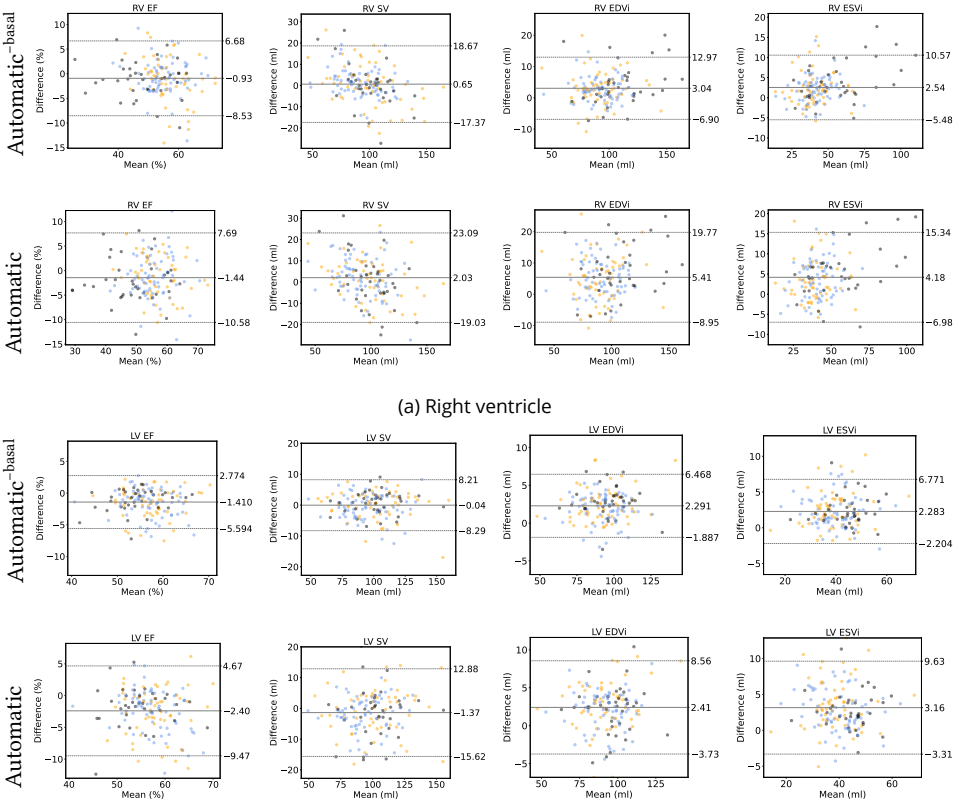
Figure 4.3: Bland-Altman plots with the agreement between the manually and automatically selected end-diastolic (ED) and end-systolic (ES) phases for right ventricle (RV) and left ventricle (LV), respectively, using (a) automatic^{-basal} segmentations and (b) automatic^{-basal}. Distance between automatically and manually selected phases is expressed as percentage of a complete cardiac cycle. Evaluation was performed for LV (first and second columns) and RV (third and fourth columns) separately. Higher opacity of colors correlates to higher density of data points.

and controls. These trends between the subgroups were also observed in the boxplots of Figure 4.5.

We next compared CMR TFC classification using manual vs. automatic^{-basal} CMR measurements. All but one subject (156/157, 99%) were correctly classified, showing an agreement of κ 0.98±0.02. As depicted in Figure 4.6a, subjects who classified as no (n=130) or minor (n=6) CMR TFC were correctly classified using the CMR measurements computed using the automatic^{-basal} segmentations obtained from the deep learning segmentation model. For major TFC, all but one subject were correctly classified; with one female subject being misclassified as minor CMR TFC. This classification discrepancy was based on a 5 mL/m^2 difference in RVEDVI (102 mL/m^2 using manual measurements and 97 mL/m^2 using automatic measurements), whereby the cutoff for major CMR TFC is set at >100mL/m^2 in women. The total TFC in this patient went from 5 to 4, which did not change the ARVC diagnosis. Sensitivity and specificity of minor and major CMR TFC for diagnosis of ARVC were comparable for manual (minor TFC 31% | 99% and major TFC 66% | 100%) and for automatic^{-basal} (minor TFC 35% | 100% and major TFC 65% | 100%, p=0.32). CMR TFC classification using the uncorrected automatic measurements are depicted in Figure 4.6b. This resulted in correct classification of 149/157 (95%) subjects.

Table 4.5: Table lists right ventricular (RV) and left ventricular (LV) function and dimension. CMR measurements are given for controls, at-risk family members and ARVC patients, stratified per method: (i) manual, (ii) automatic^{-basal} [light blue], and (iii) automatic [blue]). Significant difference 0.01≤p≤0.05 (*) or p<0.01 (**) between control and ARVC patients; Significant difference 0.01≤p≤0.05 (†) or p<0.01 (‡) between at-risk and ARVC patients.

	Study population			
	ARVC patients (n=37)	At-risk ARVC group (n=66)	Control group (n=54)	p-value
Right ventricle				
EF	47.1±9.0**‡	55.5±5.9	56.2±6.1	<0.001
	48.3±9.6**‡	56.1±6.0	57.2±7.4	<0.001
	49.2±9.0**‡	56.9±5.8	57.4±7.7	<0.001
SV	99.0±15.9‡	92.7±18.3	99.5±23.4	0.116
	98.7±20.7	91.6±21.0	99.0±27.3	0.159
	98.5±21.8	90.2±20.5	97.±26.6	0.139
EDV	218.1±53.2**‡	168.6±34.9	178.1±41.3	<0.001
	210.1±50.5**‡	164.1±35.5	173.2±42.3	<0.001
	204.7±51.0**‡	159.9±35.3	169.2±41.9	<0.001
EDVI	111.6±25.4**‡	93.7±14.8	92.9±18.5	<0.001
	107.4±23.8**‡	91.1±14.9	90.1±18.5	<0.001
	104.6±24.0**‡	88.7±15.1	88.0±18.1	<0.001
ESV	119.1±45.8**‡	76.0±20.8	78.6±22.6	<0.001
	111.4±41.8**‡	72.5±19.3	74.1±22.2	<0.001
	106.2±39.5**‡	69.7±19.4	72.3±22.7	<0.001
ESVI	60.8±22.7**‡	42.1±9.9	41.1±11.2	<0.001
	56.9±20.7**‡	40.2±9.0	38.8±11.1	<0.001
	54.3±19.5**‡	38.6±8.9	37.7±11.1	<0.001
Left ventricle				
EF	53.3±5.7**‡	56.4±4.4	56.7±5.3	0.003
	54.6±5.2**‡	57.8±4.9	58.1±5.2	0.003
	55.6±5.7**‡	58.7±5.6	59.3±5.3	0.005
SV	101.0± 18.1	94.6±18.3	101.3±23.1	0.133
	100.7±18.0	94.9±18.1	101.3±23.7	0.173
	102.8±17.9	95.8±17.0	102.6±22.6	0.091
EDV	190.5±33.4‡	168.6±33.6	179.0±38.3	0.011
	185.4±33.4‡	165.1±33.0	174.2±37.5	0.018
	186.2±32.8‡	164.6±32.3	173.5±36.9	0.010
EDVI	97.5±14.3	93.6±12.8	93.2±16.0	0.322
	94.8±14.4	91.7±12.9	90.7±15.5	0.376
	95.3±14.3	91.4±12.7	90.4±15.4	0.246
ESV	89.6±20.7**‡	74.0±18.3	77.7±19.0	0.001
	84.7±19.5**‡	70.2±18.2	72.9±17.6	0.001
	83.3±20.0**‡	68.8±19.2	70.9±18.3	0.001
ESVI	45.6±8.7**‡	41.0±7.9	40.5±8.9	0.010
	43.2±8.5**†	38.9±8.1	38.0±8.4	0.011
	42.5±9.0**†	38.1±8.7	36.9±8.5	0.009



(a) Right ventricle

(b) Left ventricle

Figure 4.4: Bland-Altman plots of (a) right ventricular (RV) CMR measurements. Absolute agreement between: (first row) manuals vs. automatic$^{-\text{basal}}$ CMR measurements; and (second row) manuals vs. automatic CMR measurements. Data points are stratified by disease classification 1) ARVC patients (in black), 2) at-risk family members (in blue), and 3) control subjects (in orange). (b) depicts the same information for the left ventricular (LV). Abbreviations: EF=ejection fraction; SV=stroke volume; EDVI=end-diastolic volume index; ESVI=end-systolic volume index.

4.4 Discussion

In this study, we evaluated our previously developed deep learning segmentation approach for RV and LV ventricular CMR assessment in patients suspected of ARVC. Moreover, we evaluated the clinical implication of this approach for classification of the CMR TFC in subjects suspected of ARVC. We demonstrated that CMR TFC classification using our automatic segmentation with limited manual correction in the most basal slice was comparable to classification using manual segmentation performed during clinical work-up. Therefore, CMR TFC classification could potentially be performed

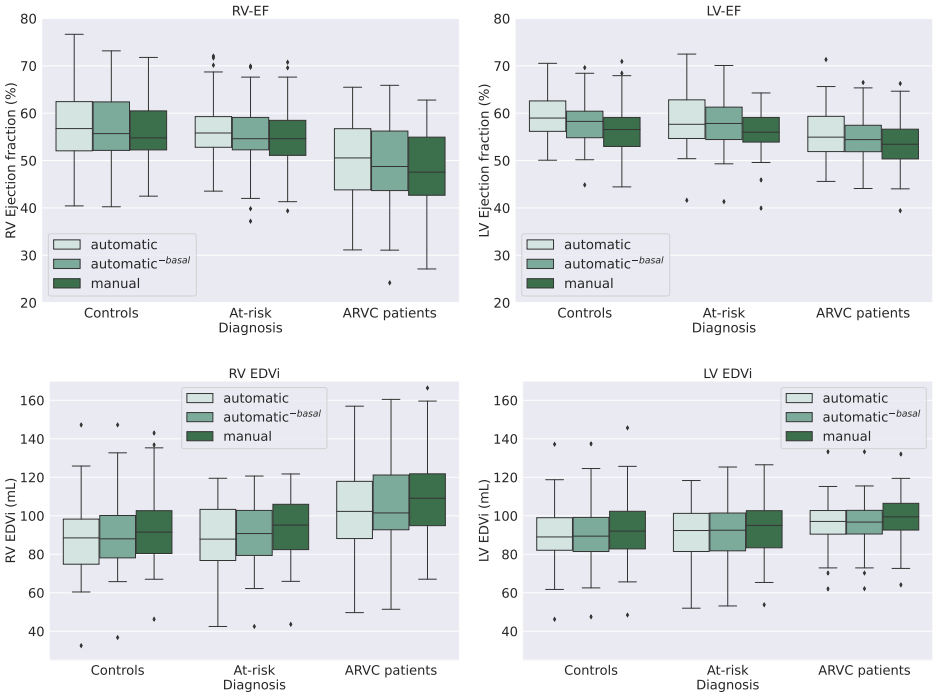


Figure 4.5: Boxplots depicting right ventricular (RV, first column) and left ventricular (LV, second column) function (first row) and dimension (second row). CMR measurements are given for controls, at-risk family members and ARVC patients, stratified per method (automatic, automatic^{-basal} and manual).

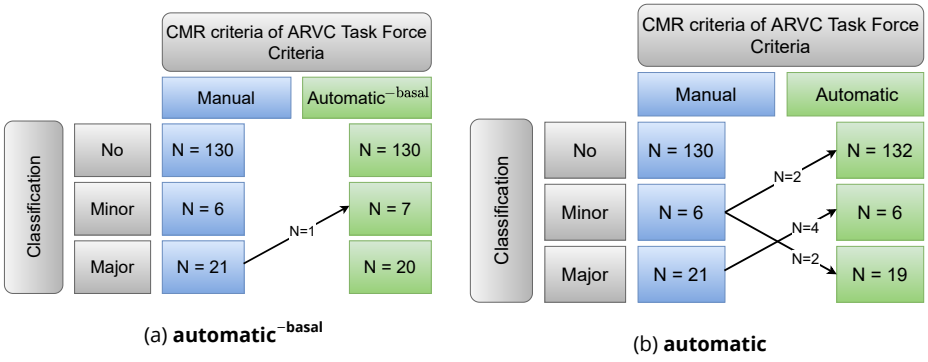


Figure 4.6: Classification of CMR criteria of TFC (no, minor and major) using manual segmentations and (a) automatic^{-basal}; (b) automatic segmentations to compute CMR measurements. The arrows indicate the number of patients that change CMR classification category when using (a) automatic^{-basal}, and (b) automatic measurements. Abbreviations: CMR=cardiovascular magnetic resonance; N=number of subjects.

using automatically measured CMR parameters with limited expert interaction.

4.4.1 Previous studies

Recently studies[15,23,24] have shown that deep learning segmentation methods outperform traditional approaches such as those exploiting level set, graph-cuts, deformable models, cardiac atlases and statistical models.[25,26] Many current state-of-the-art deep learning biventricular segmentation algorithms have been evaluated on publicly available cine CMR data from the MICCAI 2017 ACDC.[12] The dataset contains CMR volumes from 150 patients distributed uniformly over normal cardiac function and four disease groups: dilated cardiomyopathy, hypertrophic cardiomyopathy, ischemic cardiomyopathy, and RV abnormality (RVEDVI greater than 110 mL/m^2 for men, and greater than 100 mL/m^2 for women, and/or a RVEF below 40%). The ACDC challenge showed that the largest segmentation inaccuracies were located in the most basal and apical slices of the short-axis,[12] which is in line with our results presented in Table 4.3. Comparable results were obtained in the recently held Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation (M&Ms) challenge.[13] Importantly, in contrast to the ACDC and M&Ms datasets, the clinical annotation protocol used in our study adheres to the guidelines of the SCMR.[18] Segmentation of the RV, especially in basal slices, is more challenging when following SCMR guidelines compared with the protocol used for the ACDC and M&Ms datasets. For example, in the SCMR guideline the outflow tract is included as part of the RV blood volume which challenges segmentation of the basal slices due to the unclear ventricular-atrial transition.

Researchers[27,28] have also trained and evaluated deep learning CMR segmentation algorithms on the large-scale annotated dataset from the UK Biobank,[29] reaching a performance comparable with human experts. The dataset contains short-axis cine CMR volumes of 5,008 subjects. As the majority of the subjects are healthy, the dataset is considered relatively homogenous.[29] In the present work, we trained and evaluated a previously developed deep learning segmentation algorithm[16] on a real-life dataset with subjects suspected of ARVC who underwent CMR as part of their clinical evaluation. Compared to the previously mentioned datasets,[12,13,29] our dataset contains substantially more subjects with RV complexity caused by ARVC due to possible aneurysms and wall thinning and contained CMR images acquired on different field strengths (1.5 and 3 Tesla), pulse sequences and imaging parameters. Hence, the current work demonstrates that by only correcting a single slice per volume, an existing state-of-the-art segmentation method[16] is sufficiently reliable to be applied to a relevant clinical problem. Furthermore, we are the first to compare classification of the CMR TFC of subjects suspected of ARVC using manually and automatically derived CMR measurements and showing that the deep learning segmentation algorithm we use performs well in this diverse clinical environment.

4.4.2 Comparison to manual segmentation

We showed a good to excellent agreement of manual and automated CMR measurements, which significantly increased after simulated correction of the most basal slice of the RV and LV (automatic (r=0.78-0.99, p<0.001) and automatic^{-basal} (r=0.88-0.99, p<0.001) measurements). This was also reflected in the significant increase of the Dice coefficients and Hausdorff distance after basal correction (p<0.001). This is in agreement with a recent study, showing an improvement of the agreement between automatic and manual segmentation when manually adjusting the most basal slice.[30]

Large intra- and inter-observer variability is currently the greatest source of error when manually segmenting CMRs[8,31] with more variability seen for the RV compared to LV due to the RV geometrical complexity.[18] Previously published inter-observer variability ranges from 2.6 -10.5%[32,33] for the LV and 6.2-14.1%[33,34] for the RV. The largest variability between manual readers also appears in the apical and basal slices[14] presumably due to low tissue contrast ratios, hypertrabeculation and unclear ventricular-atrial transition of especially the RV. The variability in contouring of the basal slice is illustrated in Figure 4.2. The corresponding manual segmentations convey the difficulty to determine the anatomical boundaries of cardiac structures in these slices. We presumed that such variability also hampers performance of the automatic segmentation method. This limitation can be alleviated by increasing the training set size. To further improve performance of deep learning segmentation approaches, especially of basal and apical short-axis slices, future work could exploit anatomical information extracted from long-axis views (2, 3, 4-chamber views) e.g. valve landmarks and apical point.[35,36] Furthermore, deep learning based CMR segmentation methods would benefit from short-axis volumes with higher through-plane resolution e.g., using super-resolution.[10,37,38] This would make application of 3D segmentation approaches more feasible and hence, those models could potentially harness any inter-slice dependencies. Finally, using explicit topological prior information[39] for model optimization is a promising training approach to prevent automatic models from generating anatomically implausible segmentation.

4.4.3 Clinical implementation of deep learning methods

Depending on the stage of disease, ARVC patients show a wide variety of ventricular changes that can be observed on CMR: ventricular wall motion abnormalities (e.g. aneurysms, akinesia, dyskinesia), wall thinning (due to fibrofatty replacement of the myocardium), increased trabeculations, dilatation and decreased functional measurements, that are especially present in the RV.[2] These challenges make ARVC eminently suitable to study the performance of machine learning algorithms on the RV. Previously published algorithms showed better agreement for LV than RV volumes.[40] Although limits of agreement were smaller for the LV compared to the RV, we showed comparable segmentation performance for RV and LV CMR measurements in this heterogenous

study population. Furthermore, segmentation performance was comparable between structurally normal hearts and hearts affected by ARVC.

Importantly, we showed that calculation of ARVC TFC from automatically computed CMR parameters is feasible when combining automatic segmentation with correction of the most basal slice only. The diagnostic performance of the CMR TFC calculated using automatic segmentations (sensitivity 32-58%, specificity 99-100%) were comparable to manual measurements in this and previously published studies (sensitivity 13-69%, specificity 88-100%).[41,42] Although the correlation of manual and automatic measurements is high, the differences in CMR TFC classification without basal correction demonstrates that a fully automatic segmentation approach without human intervention is not yet reliable. However, the conducted experiments reveal that current state-of-the-art deep learning segmentation models can substantially reduce manual effort to semi-automatically segment cardiac structures in a heterogeneous dataset: manual segmentation time would be approximately 2 minutes instead of 25 minutes. Recently, Huellebrand *et al.*[43] proposed a human-in-the-loop approach that combines deep learning-based CMR segmentation and cardiac disease classification. The authors show that manual correction of automatic CMR segmentations by a clinical expert results in increased classification performance compared to a fully automatic segmentation approach. To identify volumes that contain segmentation failures the user can explore parallel coordinates plots that visualize CMR measurements along with cardiac shape and texture features. A similar approach was previously presented by Sander *et al.*[16] that combines automatic segmentation and assessment of segmentation uncertainty in CMR to automatically detect image regions containing local segmentation failures. Subsequently, detected regions are manually corrected by a clinical expert. Such a semi-automatic approach could lead to a large reduction in inter-observer variability. This is not only interesting for specialized tertiary ARVC centers, but even more for less experienced centers, since CMR misinterpretations are an important cause of over-diagnosis in ARVC and only 27% of people referred to a tertiary center with a suspected ARVC diagnosis finally meet diagnostic criteria for ARVC.[44] Our work shows that our previously developed deep learning segmentation method is able to fulfill a diagnostic purpose by simplifying accurate calculation of functional and volumetric measurements for the CMR TFC, showing opportunities to facilitate and improve individual patients health.

4.4.4 Limitations

Although we automated the calculation of the dimensional and functional parameters, wall motion abnormalities are also part of the CMR TFC. This was evaluated visually by experienced cardiovascular radiologists in this work, but it is subject to inter-observer variation in less experienced readers. Due to anatomical challenges of the RV a fully automatic RV strain algorithm is not yet available. Future work should focus on

automatic computation of RV strain and better automatic segmentation of the basal slice, which could contribute to full automatization and standardization of the CMR TFC. Combining automatic segmentation with manual correction of the most basal slice, 99% of the CMR TFC were correctly classified, with misclassification of only one patient from major to minor CMR TFC. Moreover, one could argue that this latter classification falls within measurement error, and it did not change the diagnosis (total TFC score went from 5 to 4). Although the absolute differences in volumetric and functional parameters were small, due to the absolute cutoff values used for the CMR TFC, differences in classification can theoretically exist when the difference is as small as 1 mL/m^2, and clinical interpretation of automatic measurements remains important. Notably, CMR is no gold standard for the diagnosis of ARVC, but rather part of the diagnostic process.

4.5 Conclusions

Automatic deep learning-based CMR segmentation has the ability to provide a fast, standardized and reproducible method to measure RV and LV volumetric parameters on CMR. We demonstrate that the applied automated segmentations have a good agreement with manual segmentations. Furthermore, combining automatic segmentation with manual correction of the segmentation in the most basal slice results in accurate CMR TFC classification of subjects suspected of ARVC.

References

[1] F. I. Marcus, W. J. McKenna, D. Sherrill, C. Basso, B. Bauce, D. A. Bluemke, H. Calkins, D. Corrado, M. G. P. J. Cox, J. P. Daubert, G. Fontaine, K. Gear, R. Hauer, A. Nava, M. H. Picard, N. Protonotarios, J. E. Saffitz, D. M. Y. Sanborn, J. S. Steinberg, H. Tandri, G. Thiene, J. A. Towbin, A. Tsatsopoulou, T. Wichter, and W. Zareba. "Diagnosis of arrhythmogenic right ventricular cardiomyopathy/dysplasia: proposed modification of the task force criteria." *Circulation*, vol. 121 (13 2010), pp. 1533–41.

[2] A. S. J. M. te Riele, H. Tandri, and D. A. Bluemke. "Arrhythmogenic right ventricular cardiomyopathy (arvc): cardiovascular magnetic resonance update." *Journal of cardiovascular magnetic resonance : official journal of the Society for Cardiovascular Magnetic Resonance*, vol. 16 (2014), p. 50.

[3] F. von Knobelsdorff-Brenkenhoff, G. Pilz, and J. Schulz-Menger. "Representation of cardiovascular magnetic resonance in the aha / acc guidelines." *Journal of Cardiovascular Magnetic Resonance*, vol. 19 (1 2017), p. 70.

[4] W. Bai, M. Sinclair, G. Tarroni, O. Oktay, M. Rajchl, G. Vaillant, A. M. Lee, N. Aung, E. Lukaschuk, M. M. Sanghvi, F. Zemrak, K. Fung, J. M. Paiva, V. Carapella, Y. J. Kim, H. Suzuki, B. Kainz, P. M. Matthews, S. E. Petersen, S. K. Piechnik, S. Neubauer, B. Glocker, and D. Rueckert. "Automated cardiovascular magnetic resonance image analysis with fully convolutional networks." *Journal of cardiovascular magnetic resonance : official journal of the Society for Cardiovascular Magnetic Resonance*, vol. 20 (1 2018), p. 65.

[5] C. P. Corona-Villalobos, I. R. Kamel, N. Rastegar, R. Damico, T. M. Kolb, D. M. Boyce, A.-E. S. Sager, J. Skrok, M. L. Shehata, J. Vogel-Claussen, D. A. Bluemke, R. E. Girgis, S. C. Mathai, P. M. Hassoun, and S. L. Zimmerman. "Bidimensional measurements of right ventricular function for prediction of survival in patients with pulmonary hypertension: comparison of reproducibility and time of analysis with volumetric cardiac magnetic resonance imaging analysis." *Pulmonary circulation*, vol. 5 (3 2015), pp. 527–537.

[6] F. Grothues, J. C. Moon, N. G. Bellenger, G. S. Smith, H. U. Klein, and D. J. Pennell. "Interstudy reproducibility of right ventricular volumes, function, and mass with cardiovascular magnetic resonance." *American heart journal*, vol. 147 (2 2004), pp. 218–223.

[7] F. Haddad, R. Doyle, D. J. Murphy, and S. A. Hunt. "Right ventricular function in cardiovascular disease, part ii: pathophysiology, clinical importance, and management of right ventricular failure." *Circulation*, vol. 117 (13 2008), pp. 1717–1731.

[8] L. Bonnemains, D. Mandry, P.-Y. Marie, E. Micard, B. Chen, and P.-A. Vuissoz. "Assessment of right ventricle volumes and function by cardiac mri: quantification of the regional and global interobserver variability." *Magnetic resonance in medicine*, vol. 67 (6 2012), pp. 1740–1746.

[9] F. Isensee, P. Jaeger, P. M. Full, I. Wolf, S. Engelhardt, and K. H. Maier-Hein. "Automatic cardiac disease assessment on cine-mri via time-series segmentation and domain specific features," *arXiv* (2017).

[10] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. De Marvao, T. Dawes, D. P. O'Regan, et al. "Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation," *IEEE transactions on medical imaging*, vol. 37 (2017), pp. 384–395.

[11] Q. Tao, W. Yan, Y. Wang, E. H. M. Paiman, D. P. Shamonin, P. Garg, S. Plein, L. Huang, L. Xia, M. Sramko, J. Tintera, A. de Roos, H. J. Lamb, and R. J. van der Geest. "Deep learning-based method for fully automatic quantification of left ventricle function from cine mr images: a multivendor, multicenter study." *Radiology*, vol. 290 (1 2019), pp. 81–88.

[12] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester, G. Sanroma, S. Napel, S. Petersen, G. Tziritas, E. Grinias, M. Khened, V. A. Kollerathu, G. Krishnamurthi, M.-M. Rohe, X. Pennec, M. Sermesant, F. Isensee, P. Jager, K. H. Maier-Hein, P. M. Full, I. Wolf, S. Engelhardt, C. F. Baumgartner, L. M. Koch, J. M. Wolterink, et al. "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?" *IEEE Transactions on Medical Imaging*, vol. 37 (11 2018), pp. 2514–2525.

[13] V. M. Campello, P. Gkontra, C. Izquierdo, C. Martin-Isla, A. Sojoudi, P. M. Full, K. Maier-Hein, Y. Zhang, Z. He, J. Ma, M. Parreno, A. Albiol, F. Kong, S. C. Shadden, J. C. Acero, V. Sundaresan, M. Saber, M. Elattar, H. Li, B. Menze, F. Khader, C. Haarburger, C. M. Scannell, M. Veta, A. Carscadden, K. Punithakumar, X. Liu, S. A. Tsaftaris, X. Huang, X. Yang, et al. "Multi-centre, multi-vendor and multi-disease cardiac segmentation: the m&ms challenge." *IEEE transactions on medical imaging*, vol. 40 (12 2021), pp. 3543–3554.

[14] A. Suinesiaputra, D. A. Bluemke, B. R. Cowan, M. G. Friedrich, C. M. Kramer, R. Kwong, S. Plein, J. Schulz-Menger, J. J. M. Westenberg, A. A. Young, and E. Nagel. "Quantification of lv function and mass by cardiovascular magnetic resonance: multi-center variability and consensus contours." *Journal of cardiovascular magnetic resonance : official journal of the Society for Cardiovascular Magnetic Resonance*, vol. 17 (1 2015), p. 63.

[15] T. Leiner, D. Rueckert, A. Suinesiaputra, B. Baeßler, R. Nezafat, I. Išgum, and A. A. Young. "Machine learning in cardiovascular magnetic resonance: basic concepts and applications." *Journal of cardiovascular magnetic resonance : official journal of the Society for Cardiovascular Magnetic Resonance*, vol. 21 (1 2019), p. 61.

[16] J. Sander, B. D. de Vos, and I. Išgum. "Automatic segmentation with detection of local segmentation failures in cardiac MRI," *Scientific Reports*, vol. 10 (2020), pp. 1–19.

[17] N. Prakken, B. Velthuis, E. Vonken, W. P. Mali, and M. Cramer. "Cardiac mri: standardized right and left ventricular quantification by briefly coaching inexperienced personnel," *The Open Magnetic Resonance Journal*, vol. 1 (2008), pp. 104–111.

[18] J. Schulz-Menger, D. A. Bluemke, J. Bremerich, S. D. Flamm, M. A. Fogel, M. G. Friedrich, R. J. Kim, F. von Knobelsdorff-Brenkenhoff, C. M. Kramer, D. J. Pennell, S. Plein, and E. Nagel. "Standardized image interpretation and post-processing in cardiovascular magnetic resonance - 2020 update : society for cardiovascular magnetic resonance (scmr): board of trustees task force on standardized post-

processing." *Journal of cardiovascular magnetic resonance : official journal of the Society for Cardiovascular Magnetic Resonance*, vol. 22 (1 2020), p. 19.

[19] F. Yu, V. Koltun, and T. Funkhouser. "Dilated residual networks," 2017, pp. 636–644.

[20] Y. Gal and Z. Ghahramani. "Dropout as a bayesian approximation: representing model uncertainty in deep learning," edited by M. F. Balcan and K. Q. Weinberger. Vol. 48 (PMLR, 2016), pp. 1050–1059.

[21] F. Milletari, N. Navab, and S.-A. Ahmadi. "V-net: fully convolutional neural networks for volumetric medical image segmentation," (2016).

[22] D. P. Kingma and J. Ba. "Adam: a method for stochastic optimization," *arXiv* (1412.6980 2014).

[23] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez. "A survey on deep learning in medical image analysis." *Medical image analysis*, vol. 42 (2017), pp. 60–88.

[24] C. Chen, C. Qin, H. Qiu, G. Tarroni, J. Duan, W. Bai, and D. Rueckert. "Deep learning for cardiac image segmentation: a review." *Frontiers in cardiovascular medicine*, vol. 7 (2020), p. 25.

[25] C. Petitjean and J.-N. Dacher. "A review of segmentation methods in short axis cardiac mr images." *Medical image analysis*, vol. 15 (2 2011), pp. 169–184.

[26] P. Peng, K. Lekadir, A. Gooya, L. Shao, S. E. Petersen, and A. F. Frangi. "A review of heart chamber segmentation for structural and functional analysis using cardiac magnetic resonance imaging." *Magma (New York, N.Y.)*, vol. 29 (2 2016), pp. 155–195.

[27] R. Robinson, V. V. Valindria, W. Bai, O. Oktay, B. Kainz, H. Suzuki, M. M. Sanghvi, N. Aung, J. M. Paiva, F. Zemrak, et al. "Automated quality control in image segmentation: application to the uk biobank cardiovascular magnetic resonance imaging study," *Journal of Cardiovascular Magnetic Resonance*, vol. 21 (2019), pp. 1–14.

[28] R. Attar, M. Pereañez, A. Gooya, X. Albà, L. Zhang, M. H. de Vila, A. M. Lee, N. Aung, E. Lukaschuk, M. M. Sanghvi, K. Fung, J. M. Paiva, S. K. Piechnik, S. Neubauer, S. E. Petersen, and A. F. Frangi. "Quantitative cmr population imaging on 20,000 subjects of the uk biobank imaging study: lv/rv quantification pipeline and its evaluation." *Medical image analysis*, vol. 56 (2019), pp. 26–42.

[29] S. E. Petersen, P. M. Matthews, F. Bamberg, D. A. Bluemke, J. M. Francis, M. G. Friedrich, P. Leeson, E. Nagel, S. Plein, F. E. Rademakers, et al. "Imaging in population science: cardiovascular magnetic resonance in 100,000 participants of uk biobank-rationale, challenges and approaches," *Journal of Cardiovascular Magnetic Resonance*, vol. 15 (2013), pp. 1–10.

[30] S. G.J.H., S. Poort, B. Velthuis, V. van Deursen, C. Nguyen, D. Sosnovic, R. Dierckx, R. Slart, R. Borra, and N. Prakken. "Balancing speed and accuracy in cardiac magnetic resonance function post-processing: comparing 2 levels of automation in 3 vendors to manual assessment," *Diagnostics (Basel, Switzerland)*, vol. 11 (2021), p. 1758.

[31] A. Bhuva, W. Bai, C. Lau, R. Davies, Y. Ye, H. Bulluck, E. McAlindon, V. Culotta, P. Swoboda, G. Captur, T. Treibel, J. Augusto, K. Knott, A. Seraphim, G. Cole, S. Petersen, N. Edwards, J. Greenwood, C. Bucciarelli-Ducci, A. Hughes, D. Rueckert, J. Moon, and C. Manisty. "A multicenter, scan-rescan, human and machine learning cmr study to test generalizability and precision in imaging biomarker analysis." *Circulation. Cardiovascular imaging*, vol. 12 (10 2019), p. e009214.

[32] J. C. C. Moon, C. H. Lorenz, J. M. Francis, G. C. Smith, and D. J. Pennell. "Breath-hold flash and fisp cardiovascular mr imaging: left ventricular volume differences and reproducibility." *Radiology*, vol. 223 (3 2002), pp. 789–797.

[33] C. F. Mooij, C. J. de Wit, D. A. Graham, A. J. Powell, and T. Geva. "Reproducibility of MRI measurements of right ventricular size and function in patients with normal and dilated ventricles." *Journal of magnetic resonance imaging : JMRI*, vol. 28 (1 2008), pp. 67–73.

[34] F. Grothues, G. C. Smith, J. C. C. Moon, N. G. Bellenger, P. Collins, H. U. Klein, and D. J. Pennell. "Comparison of interstudy reproducibility of cardiovascular magnetic resonance with two-dimensional echocardiography in normal subjects and in patients with heart failure or left ventricular hypertrophy." *The American journal of cardiology*, vol. 90 (1 2002), pp. 29–34.

[35] H. Xue, J. Artico, M. Fontana, J. C. Moon, R. H. Davies, and P. Kellman. "Landmark detection in cardiac MRI using a convolutional neural network." *Radiology: Artificial Intelligence* (2021). doi: 10.1148/ryai.2021200197, p. e200197.

[36] N. Painchaud, Y. Skandarani, T. Judge, O. Bernard, A. Lalande, and P.-M. Jodoin. "Cardiac segmentation with strong anatomical guarantees." *IEEE transactions on medical imaging*, vol. 39 (11 2020), pp. 3703–3713.

[37] K. K. Bhatia, A. N. Price, W. Shi, J. V. Hajnal, and D. Rueckert. "Super-resolution reconstruction of cardiac mri using coupled dictionary learning," *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, IEEE. 2014, pp. 947–950.

[38] J. Sander, B. D. de Vos, and I. Išgum. "Autoencoding low-resolution MRI for semantically smooth interpolation of anisotropic MRI," *Medical Image Analysis*, vol. 78 (2022), p. 102393.

[39] J. Clough, N. Byrne, I. Oksuz, V. A. Zimmer, J. A. Schnabel, and A. King. "A topological loss function for deep-learning based image segmentation using persistent homology." *IEEE transactions on pattern analysis and machine intelligence*, vol. PP (2020).

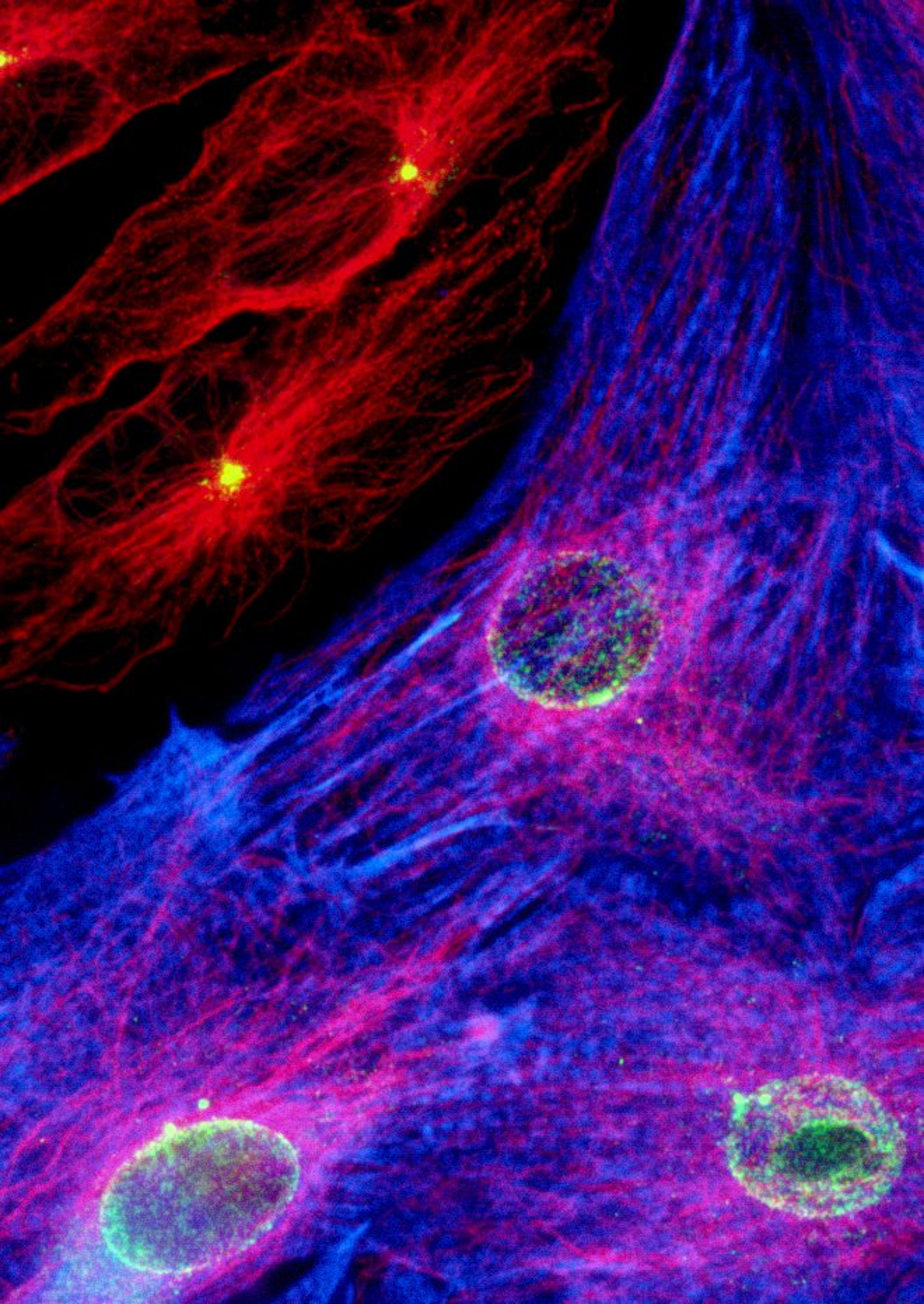
[40] S. J. Backhaus, G. Metschies, M. Billing, J. T. Kowallick, R. J. Gertz, T. Lapinskas, B. Pieske, J. Lotz, B. Bigalke, S. Kutty, G. Hasenfuss, P. Beerbaum, S. Kelle, and A. Schuster. "Cardiovascular magnetic resonance imaging feature tracking: impact of training on observer performance and reproducibility." *PloS one*, vol. 14 (1 2019), p. e0210127.

[41] L. P. Bosman, J. Cadrin-Tourigny, M. Bourfiss, M. A. Ghasabeh, A. Sharma, C. Tichnell, R. W. Roudijk, B. Murray, H. Tandri, P. Khairy, I. R. Kamel, S. L. Zimmerman, J. B. Reitsma, F. W. Asselbergs, J. P. van Tintelen, J. F. van der Heijden, R. N. W. Hauer, H. Calkins, C. A. James, and A. S. J. M. T. Riele. "Diagnosing arrhythmogenic right ventricular cardiomyopathy by 2010 task force criteria: clinical performance and simplified practical implementation." *Europace : European pacing, arrhythmias, and cardiac electrophysiology : journal of the working groups on cardiac pacing, arrhythmias, and cardiac cellular electrophysiology of the European Society of Cardiology*, vol. 22 (5 2020), pp. 787–796.

[42] G. D. Aquaro, A. Barison, G. Todiere, C. Grigoratos, L. A. Ali, G. D. Bella, M. Emdin, and P. Festa. "Usefulness of combined functional assessment by cardiac magnetic resonance and tissue characterization versus task force criteria for diagnosis of arrhythmogenic right ventricular cardiomyopathy." *The American journal of cardiology*, vol. 118 (11 2016), pp. 1730–1736.

[43] M. Huellebrand, M. Ivantsits, L. Tautz, S. Kelle, and A. Hennemuth. "A collaborative approach for the development and application of machine learning solutions for CMR-based cardiac disease classification," *Frontiers in Cardiovascular Medicine*, vol. 9 (2022).

[44] C. Bomma, J. Rutberg, H. Tandri, K. Nasir, A. Roguin, C. Tichnell, R. Rodriguez, C. James, E. Kasper, P. Spevak, D. A. Bluemke, and H. Calkins. "Misdiagnosis of arrhythmogenic right ventricular dysplasia/cardiomyopathy." *Journal of cardiovascular electrophysiology*, vol. 15 (3 2004), pp. 300–306.



CHAPTER 5

Autoencoding Low-Resolution MRI for Semantically Smooth Interpolation of Anisotropic MRI

This chapter is based on: J. Sander, B. D. de Vos, and I. Išgum. "Autoencoding low-resolution MRI for semantically smooth interpolation of anisotropic MRI," *Medical Image Analysis*, vol. 78 (2022), p. 102393.

Which is an extension of: J. Sander, B. D. de Vos, and I. Išgum. "Unsupervised super-resolution: creating high-resolution medical images from low-resolution anisotropic examples," *Medical Imaging 2021: Image Processing*, vol. 11596 International Society for Optics and Photonics. (2021), 115960E.

Abstract

High-resolution medical images are beneficial for analysis but their acquisition may not always be feasible. Alternatively, high-resolution images can be created from low-resolution acquisitions using conventional upsampling methods, but such methods cannot exploit high-level contextual information contained in the images. Recently, better performing deep-learning based super-resolution methods have been introduced. However, these methods are limited by their supervised character, i.e. they require high-resolution examples for training. Instead, we propose an unsupervised deep learning semantic interpolation approach that synthesizes new intermediate slices from encoded low-resolution examples. To achieve semantically smooth interpolation in through-plane direction, the method exploits the latent space generated by autoencoders. To generate new intermediate slices, latent space encodings of two spatially adjacent slices are combined using their convex combination. Subsequently, the combined encoding is decoded to an intermediate slice. To constrain the model, a notion of semantic similarity is defined for a given dataset. For this, a new loss is introduced that exploits the spatial relationship between slices of the same volume. During training, an existing *in-between* slice is generated using a convex combination of its neighboring slice encodings. The method was trained and evaluated using publicly available cardiac cine, neonatal brain and adult brain MRI scans. In all evaluations, the new method produces significantly better results in terms of Structural Similarity Index Measure and Peak Signal-to-Noise Ratio ($p < 0.001$ using one-sided Wilcoxon signed-rank test) than a cubic B-spline interpolation approach. Given the unsupervised nature of the method, high-resolution training data is not required and hence, the method can be readily applied in clinical settings.

5.1 Introduction

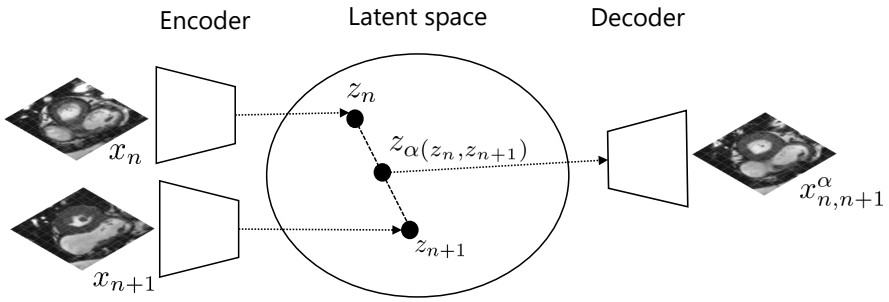
High spatial resolution of medical images is considered a key quality component for accurate disease diagnosis and prognosis. However, in case of magnetic resonance imaging (MRI) acquiring high-resolution images comes at the cost of reduced signal-to-noise ratio (SNR) or decreased temporal resolution. Higher image resolution can be achieved by increasing acquisition time. However, in clinical practice, fast scanning is often required to mitigate the risk for motion artifacts and to sustain patient comfort.

As a result, MRI scans with high temporal resolution are often highly anisotropic, which may hamper accurate analysis. There is ample research on methods that enable faster acquisition while maintaining high SNR and high spatial resolution such as compressed sensing[1] and parallel MRI.[2] However, these methods are mainly available in a research setting and are not available for the majority of MRIs obtained in clinical practice.

Conventional interpolation methods like Linear, B-spline, and Lanczos resampling[3] are often used to increase through-plane resolution of anisotropic MRIs at post-processing. The possibility to apply such methods retrospectively, i.e. after image acquisition, is often advantageous because they do not require raw image data. Conventional interpolation methods are easily and broadly applicable. Nevertheless, they cannot exploit high-level contextual information contained in the images. Furthermore, to upsample low-resolution images these methods quintessentially compute weighted intensity averages using existing image voxel values.

More sophisticated super-resolution methods have been developed that either perform denoising, deblurring, anti-aliasing, upsampling, or a combination thereof, aiming to recover a high quality of medical images from their degraded versions. In the context of this work, super-resolution for medical images refers to the process of recovering information that was lost or degraded during the low-resolution sampling process. Hence, such methods can recover anatomical structures that are finer than the sampling grid. As a result of this process anatomical structures also appear smooth and plausible in through-plane direction.

Super-resolution methods can be broadly divided into two categories. First, approaches that combine several low-resolution images to estimate, or reconstruct, the high-resolution image.[4,5] Typically, such methods require registering the low-resolution scans with each other. Therefore, performance of these approaches depends on the quality of image alignment. Given that alignment of non-moving organs can be achieved easier than for moving organs like the heart, early super-resolution approaches in the medical imaging domain were first developed for brain MRI (e.g., Peled *et al.*[6,7]). Second, to extend the applicability of super-resolution methods to moving organs approaches were developed that learn a non-linear mapping between (paired) low-resolution (LR) and high-resolution (HR) image patches.[8–12] Recently, these methods were superseded by deep-learning based super-resolution approaches using convolutional neural



$$z_{\alpha(z_n,z_{n+1})} = (1-\alpha)z_n + \alpha z_{n+1} \text{ for } \alpha \in [0,1]$$

Figure 5.1: Visualization of proposed method. To perform upsampling in anisotropic medical images we exploit the ability of autoencoders to interpolate in latent space. The trained autoencoder is used to project two spatially adjacent slices onto a latent space. Thereafter, latent space encodings z_n and z_{n+1} are combined using their convex combination. Increasing α from 0 to 1 results in a sequence of new slices where each subsequent slice is progressively less semantically similar to x_n and more semantically similar to x_{n+1}. Anatomy in the obtained stack of slices appears semantically smooth in the direction that is perpendicular to the imaging plane.

networks (CNN).[13–27]

To learn the non-linear mapping between low and high-resolution images the aforementioned methods require high-resolution training examples. However, in clinical practice such images are impractical or even impossible to acquire. Hence, to circumvent this short-coming, several unsupervised methods[28–34] have been proposed to increase resolution of images without using high-resolution training data. These approaches take advantage of anisotropic images and exploit the high in-plane resolution to increase the low through-plane resolution. Jog *et al.*[28] proposed a Fourier-based method[35] that combines multiple augmentations of a single 3D low-resolution input image to estimate Fourier coefficients of higher frequency ranges absent in the anisotropic images. To create required image augmentations the method learns a regression function between blurred low-resolution slices and their corresponding high-resolution slices extracted from the high-resolution in-plane direction. Later Zhao *et al.*[29] proposed to alter the method of Jog *et al.*[28] by replacing the regression approach with the deep-learning super-resolution approach by Lim *et al.*[17] Both methods were evaluated on brain MRI with relatively mild anisotropy. Subsequently, Zhao *et al.*[30,31] extended their approach[29] to suppress aliasing artifacts. In parallel, Dalca *et al.*[32] proposed a Gaussian Mixture Model that learns to encode anatomical similarities extracted from a large collection of anisotropic brain MRIs. Using the learned latent structure, low-resolution scans are upsampled by imputing missing slices.

We propose a deep learning semantic interpolation approach that synthesizes

new intermediate slices from encoded low-resolution examples. Specifically, in our preliminary study[33] we trained an autoencoder to compress and reconstruct high-resolution slices taken from highly anisotropic volumes. Using the latent space of the trained autoencoder new intermediate slices are synthesized by mixing the encodings of two adjacent slices. The process is depicted in Figure 5.1. Our method can synthesize an arbitrary number of intermediate slices exploiting the mixing coefficient of the neighboring slice's latent vectors. Note that the method effectively imputes image slices that are of similar appearance, i.e. slice thickness is retained, but slice spacing will be decreased, resulting in anatomically and semantically smooth transitions in through-plane direction. The method was evaluated on cardiac cine MRI. In parallel to our study, Xia *et al.*[34] presented a super-resolution approach for cardiac cine MRI that employs a conditional generative adversarial network (GAN) that takes two spatially adjacent cardiac MRI (CMRI) slices as input to synthesize a slice in-between the input slices. To guide adversarial training the approach generates an auxiliary image using previously developed optical[36] and depth-aware[37] flow-based interpolation approaches. The approach increases anisotropic resolution with upsampling-factor of two, synthesizing one slice in through-plane direction. By recursively applying the method, higher upsampling-factors of two can be achieved.

Building further on our preliminary method using convolutional autoencoders, we introduce an additional training loss function that exploits the spatial relationship between neighboring slices in 3D images. This enables us to define a notion of semantic similarity for a given dataset. As a result, the autoencoder is encouraged to generate new slices that provide a semantically smooth morphing between two input images. Furthermore, we provide evidence that our extended approach leads to improved upsampling performance when compared to Sander *et al.*[33] Unlike Xia *et al.*,[34] our approach can be applied with any desired upsampling factor, i.e. it can synthesize an arbitrary number of slices between two given slices in a straightforward fashion. Moreover, our approach uses only a single encoder-decoder structure and does not rely on auxiliary networks. Compared to Dalca *et al.*[32] our approach does not require a common atlas space to operate in. Moreover, our method is easy to implement and requires little GPU memory.

Like in our preliminary work,[33] we evaluate performance on cardiac cine MRI. Moreover, to show that our approach generalizes to other anatomies the evaluation has been substantially extended. In the experiments, three publicly available MRI datasets were used: cardiac cine MRI from the MICCAI 2017 Automated Cardiac Diagnosis Challenge[38] (ACDC); neonatal brain MRI of the developing Human Connectome Project[39] (dHCP) and adult brain MRI from the OASIS project.[40] Evaluation on neonatal and adult brain MRI enabled performance comparison with related unsupervised[28,29] and supervised[41] super-resolution methods. Finally, to demonstrate that our model is invariant to MRI scanners and voxel intensity distributions, we apply a model trained on cardiac MRI scans from the ACDC dataset to cardiac MRIs of the Sunnybrook dataset.[42]

5.2 Data

5.2.1 Cardiac Cine MRI

AUTOMATED CARDIAC DIAGNOSIS CHALLENGE Cardiac cine MRIs from the MICCAI 2017 Automated Cardiac Diagnosis Challenge (ACDC)[38] were used. The dataset consists of short-axis cardiac cine MRIs from 100 patients uniformly distributed over normal cardiac function and four disease groups: dilated cardiomyopathy, hypertrophic cardiomyopathy, heart failure with infarction, and right ventricular abnormality. Detailed acquisition protocol is described by.[38]

Briefly, MRIs have an in-plane resolution ranging from 1.37 to 1.68 mm (average reconstruction matrix 243×217 voxels) with slice thickness and spacing varying from 5 to 10 mm. The ACDC dataset specifies slice spacing for each image volume while slice thickness is only specified as a range for the complete dataset. Per patient 28 to 40 time points cover the cardiac cycle. Each volume consists of on average ten slices including the heart. To correct for intensity differences among scans, in the current work, image intensities of each volume were rescaled and clamped between [0, 1] based on their 1^{st} and 99^{th} percentiles. Furthermore, to correct for differences in-plane voxel sizes, image slices were resampled to $1.4 \times 1.4\,mm^2$.

SUNNYBROOK CARDIAC DATA To demonstrate the ability of our proposed method to generalize to other datasets with the same modality and anatomy, cardiac cine MRI from the publicly available Sunnybrook Cardiac dataset[42] was used for additional model evaluation. The dataset contains 45 short-axis cine MRI images distributed over four pathology categories: healthy subjects, patients with hypertrophy, patients with heart failure and infarction, and patients with heart failure without infarction.

Each scan contains 20 time points (i.e. volumes) encompassing the entire cardiac cycle, which results in 45×20 volumes in total. All scans have a slice thickness and spacing of 8 mm and an in-plane resolution of $1.25 \times 1.25\,mm^2$. Scans are made with a 256×256 reconstruction matrix and consist of about 10 slices. In this work, image intensities of each volume were rescaled and clamped between [0, 1] based on their 1^{st} and 99^{th} percentiles.

5.2.2 Neonatal Brain MRI

In this study neonatal brain MRIs of the developing Human Connectome Project (dHCP)[39] were used (second release). The dataset constists of 508 infants with gestational age at birth ranging from 24 to 42 weeks. All infants were scanned without sedation in a scanner environment optimized for safe and comfortable neonatal imaging. A comprehensive description of the acquisition protocol can be found in Hughes *et al.*[39]

The T_2-weighted (T2w) images are provided with an isotropic resolution of

$0.5 \times 0.5 \times 0.5\,\text{mm}^3$. To reduce image size, in this work, volumes were cropped to cortical brain structures resulting in an axial reconstruction matrix of 256×256 voxels for all images. Finally, to correct for intensity differences among scans, voxel intensities of each volume were scaled to the [0, 1] range.

5.2.3 Adult Brain MRI

Brain MRIs of 416 subjects aged 18 to 96 from the OASIS project[40] were used. Detailed information about the acquisition can be found in the paper by Marcus *et al.*[40] and on the OASIS website[1].

Briefly, T_1-weighted brain MRIs were provided with an isotropic resolution of $1.0 \times 1.0 \times 1.0\,\text{mm}^3$. To correct for intensity differences among scans, in this work, voxel intensities of each volume were scaled to the [0, 1] range.

5.3 Method

We propose a method to synthesize new slices in anisotropic 3D medical images by using the ability of a trained autoencoder to interpolate in latent space. The autoencoder is trained to compress and reconstruct high-resolution 2D slices taken from volumes with low through-plane resolution. We postulate that the autoencoder learns to encode anatomy from a collection of training images. While an individual image may only capture a partial aspect of a complete anatomical structure, such as the heart, the autoencoder may infer the missing information from similarly appearing images that captured different aspect of the anatomical structure.

After training, input slices can be *reconstructed* with minimal information loss. More important, new slices are *synthesized* through convex combinations of latent space encodings of the two adjacent slices, which is followed by decoding of the convex combinations to the new intermediate slices. Note that increasing the mixing coefficient from 0 to 1 results in a sequence of new slices where each subsequent slice is progressively less semantically similar to the first input slice and more semantically similar to the second input slice. Although thickness of synthesized slices will be similar to the input slices, slice spacing will decrease. As a result, anatomical structures appear smooth and plausible in through-plane direction.

5.3.1 Autoencoder

An autoencoder[43] is an unsupervised learning algorithm that aims to learn a lower-dimensional representation of the input. It consists of an encoder and decoder implemented as neural network. The encoder f_θ parametrized by θ compresses the input

[1] https://www.oasis-brains.org/

$x \in \mathbb{R}^{d_x}$ into a lower-dimensional space $z = f_\theta(x), z \in \mathbb{R}^{d_z}$, i.e. the latent space representation, which captures the most salient features of the input. Typically, this layer has the least amount of neurons and is also referred to as the bottleneck of an autoencoder.

The decoder g_ϕ parametrized by ϕ uses the latent space representation to generate an approximate reconstruction of the input, $\hat{x} = g_\phi(z)$. The network layers in encoder and decoder are fully connected. In general, training an autoencoder aims to minimize a loss function $\mathscr{L}(x, \hat{x})$ that quantifies dissimilarity between the input and the corresponding reconstruction.

This work uses a convolutional autoencoder[44] (CAE)[2] that has the same overall structure as a standard autoencoder but replaces the fully connected layers with convolutions. Latent space encodings generated by standard autoencoders are vectors with dimensionality equal to the size of the lower-dimensional space. In comparison, latent space representation of an input tensor generated by a convolutional autoencoder is a tensor with rank equal to the rank of the input tensor. The rank of an image is three (width, height, number of input channels). Throughout this work the number of input channels is one for gray scale images.

5.3.2 Autoencoder Architecture

The architecture of the convolutional autoencoder used in this work is the same for all datasets and experiments. The architecture of the encoder consists of two blocks, each with two consecutive convolutional layers using a kernel size of 3×3 voxels and zero-padding of size 1, followed by batch normalization and 2×2 voxels average pooling. The first and second block use 32 and 64 kernels, respectively. The last block is followed by two additional convolutional layers of 128, and 128 kernels for the final output layer. The output of the final convolutional layer is used as latent space representation of the input. All convolutional layers except for the final use a leaky ReLU nonlinearity. The combination of two average pooling layers of size 2×2 voxels and 128 kernels for the latent space representation results in an *over-complete* autoencoder. In other words, the information that can potentially be stored in the latent space is larger than the information contained in the grayscale input image.

The architecture of the decoder is reverse of the encoder. It consists of two blocks of two consecutive convolutional layers with leaky ReLU nonlinearities followed by batch normalization and 2×2 voxels nearest neighbor upsampling. The number of kernels is halved after each upsampling layer. The last block is followed by two additional convolutional layers of 32 kernels, and 1 kernel for the last layer. All convolutional layers of the autoencoder use a kernel size of 3×3 voxels and zero-padding of size 1. To ensure that output values are in the range of $[0, 1]$ the final layer uses the sigmoid function. Moreover, using two average pooling layers of size 2×2 voxels in the encoder requires the width and height of the input image each to be divisible by four.

[2]The terms autoencoder and convolutional autoencoder will be used interchangeably hereafter.

Nevertheless, test images do not have to match the size of the training patches.

5.3.3 Autoencoding for Semantic Interpolation

Our interpolation approach projects two spatially adjacent slices (x_n, x_{n+1}) onto a latent space. It requires high in-plane resolution. Thereafter, latent representations $z_n = f_\theta(x_n)$ and $z_{n+1} = f_\theta(x_{n+1})$ are combined using a convex combination:

$$z_{\alpha(z_n, z_{n+1})} = (1 - \alpha)z_n + \alpha z_{n+1} \text{ for } \alpha \in [0, 1] \,. \tag{5.1}$$

where α denotes the mixing coefficient. Finally, the decoder generates a new slice $x^{\alpha}_{n,n+1} = g_\phi(z_{\alpha(z_n, z_{n+1})})$ by decoding the mixture of latent codes. We refer to this process as *synthesizing* slices for values of $\alpha \in (0, 1)$, as opposed to *reconstructing* encoded input slices when $\alpha \in \{0, 1\}$ for slices x_n and x_{n+1}, respectively. Increasing α from 0 to 1 results in a sequence of new slices where each subsequent slice is progressively less semantically similar to x_n and more semantically similar to x_{n+1}. As a result, anatomy in the obtained stack of slices also appears semantically smooth in the direction from which the slices were extracted. We refer to this approach as Autoencoding for Semantic Interpolation (ASI).

To attain upsampling of anisotropic images by factor K in through-plane direction, K – 1 slices need to be synthesized where the set of alpha values $\mathscr{A}$ is defined as follows:

$$\begin{aligned} \mathscr{A} &= \left\{\alpha_i | \alpha_i = \frac{i}{K}\right\}_{i=1}^{K-1}, \text{ where } K = \{m | m \in \mathbb{N}, m > 1\} \\ &\text{and } |\mathscr{A}| = K - 1 \,, \end{aligned} \tag{5.2}$$

and $|.|$ denotes the cardinality of a set.

5.3.4 Loss Function

The autoencoder is trained to compress and reconstruct high-resolution slices taken from an anisotropic 3D medical imaging dataset. The model aims to minimize the reconstruction loss between the original x and the reconstructed $\hat{x}$ slice.

To further constrain the model a notion of semantic similarity is defined for a given dataset. For this, the spatial relationship between slices of the same volume is exploited. During training an existing *in-between* slice x_n (where $n \in \mathbb{N}^+$) that has two neighboring slices, x_{n-1} and x_{n+1}, is synthesized using a convex combination of the neighboring slice encodings where α (the mixing coefficient) of Equation 5.1 is set to 0.5. The new slice encoding is mapped through the decoder to an approximation $\hat{x}_n^{\alpha=0.5}$ of the original in-between slice x_n.

Finally, a distance loss is computed between the original in-between x_n slice and its approximation i.e. synthesized slice $\hat{x}_n^{\alpha=0.5}$ resulting in the following combined loss:

$$\mathscr{L} = \underbrace{d(x_n, \hat{x}_n)}_{\mathscr{L}_{\text{reconstruction}}} + \lambda \underbrace{d(x_n, \hat{x}_n^{\alpha=0.5})}_{\mathscr{L}_{\text{synthesis}}} \text{ where } \lambda \geq 0, \tag{5.3}$$

d denotes a distance function between two images and λ is a hyperparameter weighting the contribution of the synthesis loss. Setting λ to zero disables the synthesis loss during training. Minimizing the synthesis loss during training should encourage the autoencoder to linearize the latent space of images. Therefore, a convex combination of slice encodings should result in smooth nonlinear interpolations in image space.

To compute the reconstruction loss, this work used the pixel-wise mean squared error between original x_n and reconstructed $\hat{x}_n$ image. In addition, to compute the synthesis loss between reference x_n and synthesized image $\hat{x}_n^{\alpha=0.5}$ the Learned Perceptual Image Patch Similarity (LPIPS) metric[45] was used. The LPIPS metric is a perceptually-based pairwise image distance that is calculated as a weighted difference between the VGG-16[46] embedding of the reference and synthesized image. LPIPS uses the embeddings of VGG-16 layers `conv_1` to `conv_5`. The VGG-16 CNN is pretrained on ImageNet and the weights to compute the weighted difference were fit so that the metric agrees with human perceptual similarity judgments.

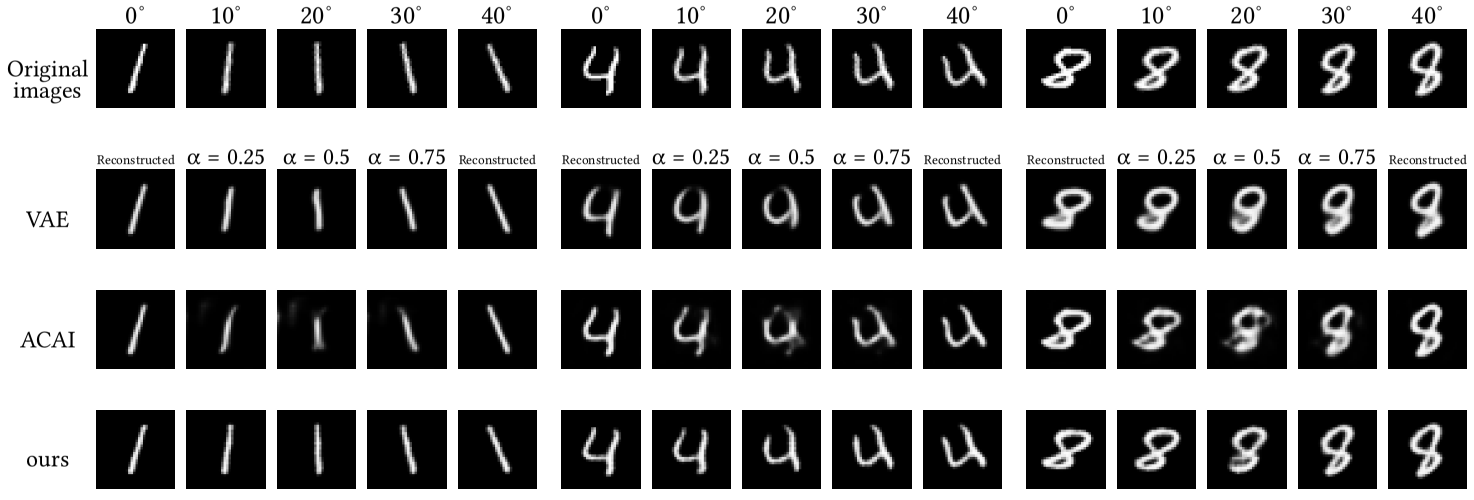


Figure 5.2: Examples of rotation performed by interpolating the latent space representation of MNIST digits. The top row shows the original input images (rotated 0° and 40°) and the intermediate target rotations (10°, 20°, 30°). The results in the rows below show reconstructions (most left and right) and synthesis (intermediate rotation angles) of intermediate rotations using latent space interpolation. Note that our approach follows the original image rotation more closely than VAE[47] and ACAI approach.[48] α denotes the mixing coefficient as specified in Equation 5.1.

.

5.4 Implementation details

The method was implemented using the PyTorch framework[49] and trained on one Nvidia GTX Titan X GPU with 12 GB memory. Model weights were initialized as zero-mean Gaussian random variables with a standard deviation of $1/\sqrt{n_l\ (1 + 0.2^2)}$ set in accordance with the leaky ReLU slope of 0.2. n_l denotes the number of incoming network connections to layer l.

A model was trained using mini-batch stochastic gradient descent with a learning rate of 1×10^{-5}. Image slices were provided once per epoch to the autoencoder in random order. In each experiment the training set was augmented by 90 degree rotations of the images and random intensity changes. Network parameters were optimized using the Adam optimizer[50] minimizing the reconstruction and synthesis loss.

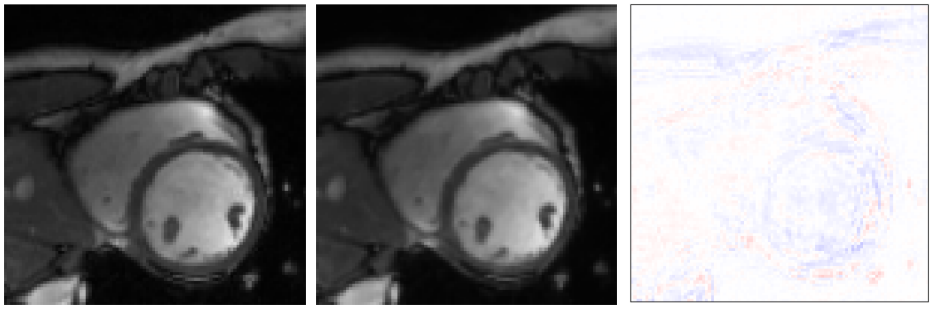
To compute the synthesis loss as described in Section 5.3.4, this study used the LPIPS metric implementation[3] of Zhang *et al.*[45] Furthermore, in order to compute the synthesis loss mini-batches of T slice pairs were randomly selected from the training set. A slice pair consisted of two slices (x_{n-1} and x_{n+1}) originating from the same volume that are spatially separated by one in-between slice x_n. To determine the optimal value for λ i.e. the contribution of the synthesis loss to the overall loss, a separate line search was performed for each dataset.

Finally, in all experiments model selection was performed on the validation set. The test set was not used during method development in any way.

Table 5.1: Quantitative comparison of reconstruction and synthesis performance of cardiac cine MRIs (ACDC dataset) in terms of SSIM, PSNR, and VIF between proposed model trained with *reconstruction* loss only ($ASI_{\lambda=0}$) and model trained with combination of *reconstruction* and *synthesis* loss ($ASI_{\lambda=0.05}$). A higher score indicates better performance. Measures (mean±standard deviation) are computed on cardiac short-axis slices. *Rec* denotes reconstructed and *Syn* synthesized slices. Synthesis performance was assessed on downsampled test volumes using a factor of 2 in through-plane direction. Best performance is indicated in bold.

Method	SSIM		PSNR		VIF	
	Rec	Syn	Rec	Syn	Rec	Syn
$ASI_{\lambda=0}$	**0.994** ±0.01	0.572 ±0.09	**41.34** ±1.66	17.94 ±2.01	**0.960** ±0.01	**0.815** ±0.01
$ASI_{\lambda=0.05}$	0.968 ±0.01	**0.650** ±0.07	32.83 ±1.31	**19.01** ±1.89	0.891 ±0.01	0.810 ±0.01

[3] https://github.com/richzhang/PerceptualSimilarity



(a) Original (b) Reconstruction (c) Differences

Figure 5.3: Qualitative evaluation of reconstruction performance of our method on cardiac cine MRI (ACDC dataset). (a) Original cardiac MRI scan; (b) Its reconstruction and (c) Differences between original (minuend) and corresponding reconstructed (subtrahend) slice. Note that to reconstruct a slice x_n the mixing coefficient α in Equation 5.1 is set to zero. Blue corresponds to negative and red to positive differences. Image intensities are scaled to a $[0, 1]$ range. All difference images use the same color scale $[-1, 1]$.

5.5 Evaluation

Performance of the method was quantitatively evaluated in terms of Structural Similarity Index Measure (SSIM), Peak Signal-to-Noise Ratio (PSNR) and Visual Information Fidelity[51] (VIF). Recent work of Mason *et al.*[52] conveyed that Visual Information Fidelity demonstrates a high correlation with radiologists' opinions of MRI quality. The proposed method was compared against performance of cubic B-spline interpolation which is known to outperform methods like Nearest-Neighbor or Linear interpolation.[53,54] Statistical significance of performance differences between evaluated methods was tested using the one-sided Wilcoxon signed-rank test.

In addition, upsampling performance was qualitatively evaluated by visually inspecting the reconstructed and synthesized slices. Visual inspection mainly focused on *anatomical plausibility* and *semantic similarity* of synthesized slices compared to corresponding reference slices. Furthermore, generated images were visually examined for smoothness of interpolation.

5.6 Experiments and Results

5.6.1 Comparison of autoencoding approaches

Before applying the proposed approach to cardiac and brain MRI scans, several autoencoder approaches were investigated for latent space interpolation using MNIST data.[55] Given any digit and its 40 degree rotated variant, referred to as input images, intermediate rotations were synthesized by mixing the latent space encodings of the two input images. Results were compared with digits that were rotated in the image

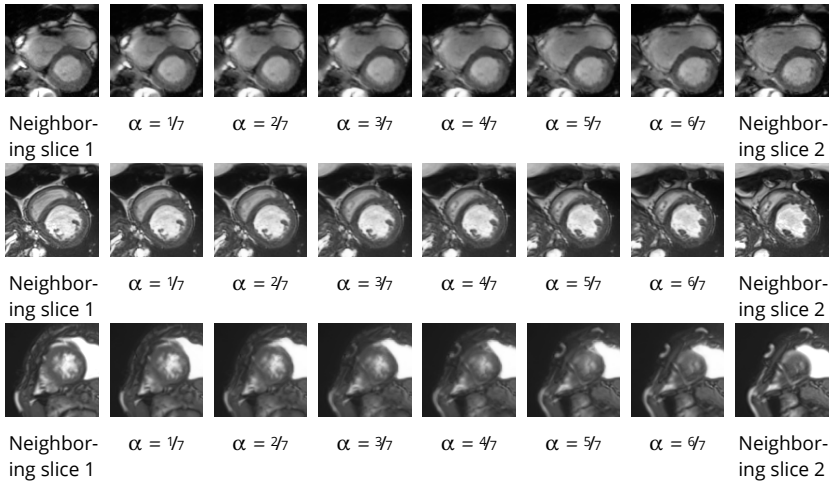


Figure 5.4: Qualitative evaluation of image synthesis performance of proposed method on cardiac cine MRI (ACDC dataset). Slice spacing was improved from 10 to 1.43 mm by synthesizing six intermediate slices (second to penultimate columns) using latent space encodings of the two neighboring slices (first and last column). α denotes the mixing coefficient as specified in Equation 5.1.

space. Three different approaches were compared: Variational Autoencoder[47,56] (VAE), Adversarially Constrained Autoencoder Interpolation[48] (ACAI) and the proposed approach (ASI). Interpolation performance was qualitatively evaluated using images of the MNIST dataset.

EXPERIMENTAL DETAILS The dataset was randomly split into training (60 000 images), validation (1000 images), and test sets (9000 images). To train the models, patches of 32×32 were randomly chosen from the training set in mini batches of 32 images. The training set was augmented by random rotations $\gamma \in [0, 2\pi]$ of the images. Models were trained for 100 epochs. The proposed model was implemented as described in section 5.3.2 except that 16 kernels were used for the latent space representation. Furthermore, λ as specified in Equation 5.3 was set to 10 after performing a line search ($\lambda \in \{0.05, 0.5, 1, 10, 100, 1000\}$).

To compute the synthesis loss as specified in Equation 5.3 each training image x_n was augmented with two *neighboring* images $\{x_{n-1}, x_{n+1}\}$. x_{n-1} denotes a 15 degree counterclockwise rotation of image x_n and x_{n+1} a 15 degree clockwise rotation of image x_n. This enabled synthesizing image $\hat{x}_n^{\alpha=0.5}$ in Equation 5.3 using a convex combination of the neighboring image encodings $\{z_{n-1}, z_{n+1}\}$ where α (the mixing coefficient) in Equation 5.1 was set to 0.5.

During evaluation, for each test image x three new images were synthesized by interpolating between the image and a 40 degree counterclockwise rotation of the same

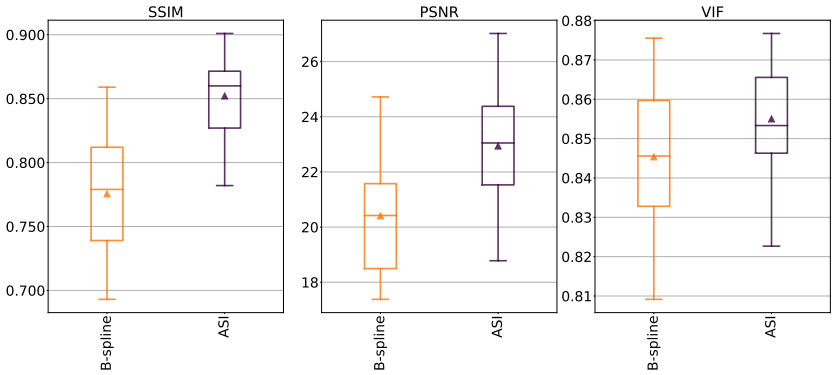


Figure 5.5: Boxplots comparing upsampling performance for cubic B-spline compared with proposed method (ASI) in terms of SSIM, PSNR, and VIF. Cardiac cine MRIs of 20 patients from the ACDC dataset were upsampled with factor 2 in through-plane direction. A higher score indicates better performance. Measures were computed on sagittal slices through short-axis volume. The proposed method achieved higher performance when evaluated by all measures compared with cubic B-spline interpolation. The differences between proposed and conventional method are statistically significant ($p < 0.0001$) using the one-sided Wilcoxon signed-rank test. Triangle indicates mean value.

image x_{rot40°. For this, the set of mixing coefficients $\mathscr{A}$ as specified in Equation 5.2 was set to $\{0.25, 0.5, 0.75\}$. As a result, the three synthesized in-between images should be rotated versions of the original image x in steps of 10 degree $\{x_{rot10^\circ}, x_{rot20^\circ}, x_{rot30^\circ}\}$. Three examples are shown in Figure 5.2.

Variational Autoencoder (VAE) To improve smoothness of the latent space of an autoencoder[47,56] proposed to model the latent representations as a random variable distributed according to a prior distribution. The latent distribution constraint is enforced by an additional loss term which measures the Kullback-Leibler divergence between approximate posterior, modelled by the encoder, and prior distribution. In this work the prior was equal to a Gaussian distribution with diagonal covariance matrix. Implementation of the autoencoder was identical to the proposed approach except for the encoder that was extended with two linear layers to model the mean and covariance matrix of the posterior Gaussian distribution.

Adversarially Constrained Autoencoder Interpolation (ACAI) To improve the ability of a convolutional autoencoder to interpolate in latent space Berthelot *et al.*[48] proposed to regularize the autoencoder by means of an adversarial training objective. Using a discriminator the autoencoder is encouraged to generate interpolated images that appear to be indistinguishable from reconstructions of real images. In this work, the approach was implemented following implementation details as described in Berthelot *et al.*[48]

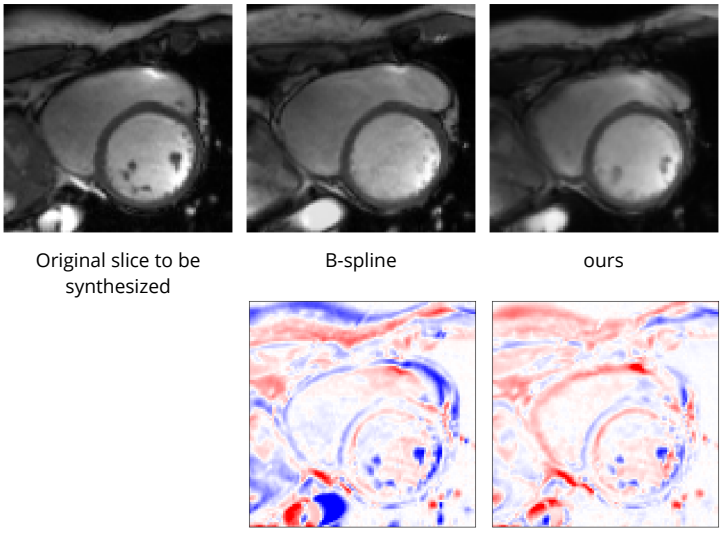
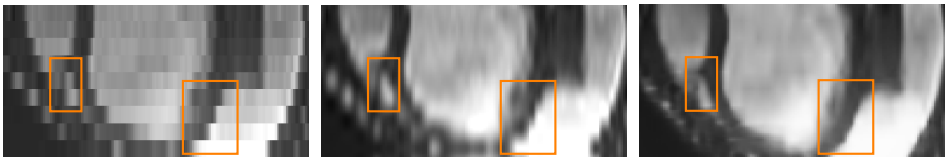


Figure 5.6: Qualitative comparison between cubic B-spline and proposed approach on cardiac MRI (ACDC dataset) for upsampling factor 2. First column shows slice from reference volume. Second column depicts image generated by conventional interpolation method. Last column shows synthesized slice using proposed method. Bottom row: Differences between original (minuend) and synthesized slice (subtrahend). Blue corresponds to negative and red to positive differences. Image intensities are scaled to a $[0, 1]$ range. All difference images use the same color scale $[-1, 1]$.

RESULTS Qualitative comparison of autoencoding approaches shown in Figure 5.2 demonstrates that our proposed model achieved the best interpolation performance. Performance differences become most apparent for interpolated images using a mixing coefficient equal to 0.5. Additionally, one can observe that linear steps taken in latent space using the set of mixing coefficients can approximate *rotation steps* in image space.

5.6.2 Semantic Interpolation of Cardiac Cine MRI

Short-axis cardiac cine MRIs are acquired to primarily investigate cardiac function. These images have a high temporal and in-plane resolution at the cost of lower through-plane resolution. The functional parameters extracted from these images, such as ejection fraction, may show high variability, which can be explained by the high anisotropic resolution, that may heavily influence volume measurements. These measurements may improve when extracted from upsampled images with smooth cardiac structures in through-plane direction. Therefore, the proposed approach was evaluated on highly anisotropic cardiac cine MRI using the ACDC dataset.



(a) Original sagittal view with 10mm slice spacing (b) B-spline (c) ours

Figure 5.7: Qualitative comparison between cubic B-spline and proposed method for upsampled cardiac short-axis stacks (ACDC dataset). Volumes were upsampled with factor 10 from 10mm (original) to 1mm slice spacing. Shown are sagittal views through upsampled short-axis stacks. The proposed method can generate new slices that result in smoother anatomical transitions compared to slices generated by conventional interpolation method. The performance difference is more pronounced for the structures of the left ventricle myocardium. Note that a feature located at the apex of the right ventricle seems to appear more accurate in the upsampled image using cubic B-spline interpolation method compared to the proposed approach.

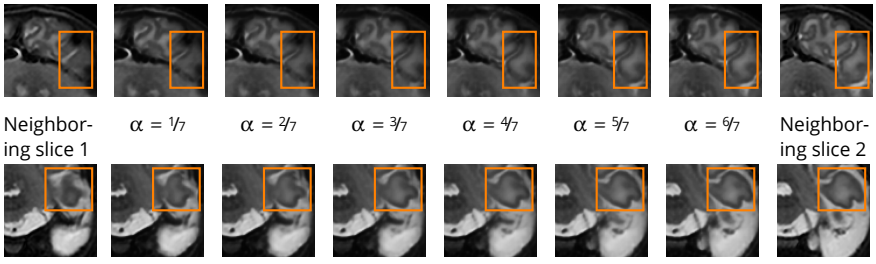


Figure 5.8: Zoomed-in images of synthesis results in scans of two different patients (one example per row) using neonatal brain MRIs from the dHCP dataset. Slice spacing was improved from 2 to 0.29 mm by synthesizing six intermediate slices (second to penultimate columns). For this, latent space encodings of the two neighboring slices (first and last column) were combined using their convex combination. α denotes the mixing coefficient as specified in Equation 5.1. Bounding boxes focus on anatomical variations between images in a row. Note that a small *cross-fade* artifact appears in the bright area of the fourth ($\alpha = 3/7$) and fifth ($\alpha = 4/7$) image in the second row.

Table 5.2: Distribution of patients in ACDC dataset over training, validation and test sets. First column specifies the slice spacing (mm) of cardiac MRI volumes. The ACDC dataset specifies slice spacing for each image volume while slice thickness is only specified as a range for all volumes.

Slice spacing (mm)	Training	Validation	Test
5	6	-	6
6.5	-	1	-
7	1	-	-
10	63	9	14

EXPERIMENTAL DETAILS The dataset was randomly split into training (70 patients), validation (10 patients), and test sets (20 patients). Table 5.2 specifies slice spacing of volumes over the three sets. The test set was only used for the final quantitative as well as qualitative evaluation.

To train a model, patches of 128×128 pixels were randomly chosen from the training set in mini-batches of 12 slice pairs i.e. 24 image slices. A model was trained for 900 epochs. Furthermore, λ in Equation 5.3 was set to 0.05.

Test images were center-cropped to 140×140 pixels covering all cardiac structures of interest. To quantitatively evaluate our method, lower through-plane resolution was mimicked by excluding every other slice in the test images i.e. by increasing slice spacing. These excluded slices were subsequently recovered by synthesizing them using the proposed approach. For this, the upsampling factor K was set to 2 and $\mathscr{A}$, the set of mixing coefficients was equal to 0.5. Downsampled test volumes had a slice spacing of 10 mm and 20 mm while slice thickness remained unchanged.

Comparison With Conventional Interpolation Method: Upsampling performance of the proposed unsupervised approach was quantitatively and qualitatively evaluated and compared with cubic B-spline interpolation.

RESULTS The primary goal of the proposed method is to *synthesize* new slices located in-between two spatially adjacent slices. However, the method's capacity to synthesize new slices depends on the ability of the autoencoder to *reconstruct* existing slices. Therefore, we report reconstruction and synthesis performance of the trained autoencoder separately.

Slice Reconstruction: Results for reconstructed and synthesized slices listed in Table 5.1 convey that the proposed approach achieved high reconstruction performance especially in terms of SSIM and PSNR. Figure 5.3 depicts qualitative results of reconstruction performance for the proposed method on cardiac MRI. The results show that the trained autoencoder can reconstruct high-quality images i.e. input slices. Nevertheless, difference image shown in Figure 5.3c depicts that some high spatial frequency details of the input slice are lacking in the reconstructed slice.

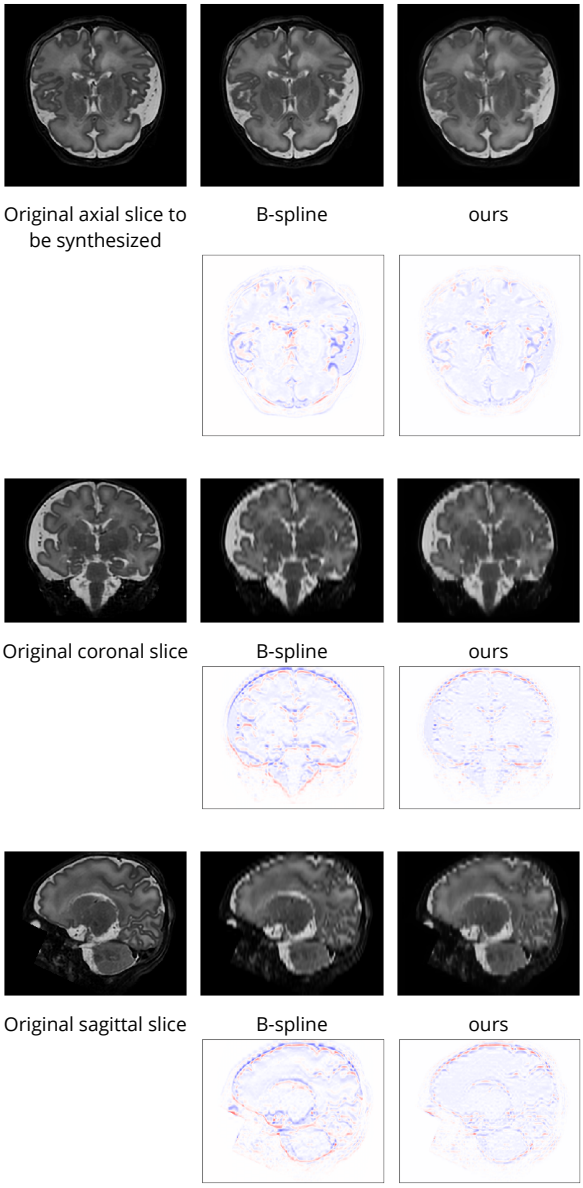


Figure 5.9: Qualitative comparison of image synthesis performance on neonatal brain MRI (dHCP dataset) between conventional interpolation methods and proposed approach. Original volumes with slice thickness and spacing of 0.5 mm were downsampled to 2.5 mm by applying a Gaussian blur before including every fifth slice in the test volume. Differences between reference (minuend) and synthesized slice (subtrahend). Blue corresponds to negative and red to positive differences. Image intensities are scaled to a $[0, 1]$ range. All difference images use the same color scale $[-1, 1]$.

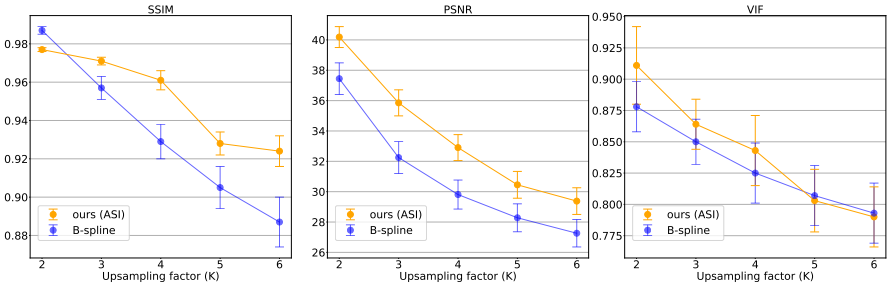


Figure 5.10: Comparison of upsampling performance for cubic B-spline interpolation compared with proposed method (ASI) in terms of SSIM, PSNR, and VIF. Neonatal brain MRIs of 20 subjects from the dHCP dataset were upsampled with factor $K \in \{2, 3, 4, 5, 6\}$ in through-plane direction. A higher score indicates better performance. The proposed method achieved higher performance in terms of SSIM and PSNR compared with conventional interpolation method. The differences between proposed and cubic B-spline interpolation approach in terms of SSIM and PSNR are statistically significant ($p < 0.001$) using the one-sided Wilcoxon signed-rank test.

Slice Synthesis: Qualitative evaluation of synthesis performance of the proposed method conveys that synthesized slices, i.e. those that are generated using a convex combination of the neighboring slice encodings, show an anatomically and semantically meaningful transition between the two neighboring slices. Moreover, despite large anatomical variations between the neighboring slices for the right ventricle, left ventricle and trabecular structures of the left ventricle, the proposed method can generate slices that depict an anatomically smooth transition between the neighboring slices. Figure 5.4 illustrates three example evaluations for basal, mid-ventricular and apical MRI slices, respectively, where upsampling factor K was set to 7 and $\mathscr{A}$, the set of mixing coefficients was equal to $\{1/7, 2/7, 3/7, 4/7, 5/7, 6/7\}$. Furthermore, quantitative evaluation of synthesis performance assessed on downsampled cardiac cine MRI scans listed in Table 5.1 reveals that synthesis performance is lower than reconstruction performance.

Comparison With Conventional Interpolation Method: Upsampling performance was compared with cubic B-spline in terms of SSIM, PSNR and VIF. For this, the methods synthesized cardiac slices that were excluded from the test volumes (see Section 5.6.2). Results for upsampling factor 2 are shown in Figure 5.5. We observe that the proposed method achieved better performance when evaluated by all measures compared with the conventional interpolation method. These differences are statistically significant ($p < 0.0001$) using the one-sided Wilcoxon signed-rank test.

Moreover, qualitative comparison of the methods reveals that synthesized images using the proposed method contain fewer errors than images generated by conventional interpolation method. Results shown in Figure 5.6 depict that performance differences are especially pronounced for the left ventricle papillary muscles and right ventricle myocardium.

Table 5.3: Comparison of upsampling performance in terms of SSIM and PSNR of proposed *unsupervised* method compared to *supervised* super-resolution approaches on neonatal brain MRI (dHCP dataset) as reported by Pham *et al.*[41,57] for upsampling factor 3. Approaches of Pham *et al.*[41,57] were evaluated on a subset of 20 scans taken from the dHCP dataset. Best performance is indicated in bold.

Method	SSIM	PSNR
Pham *et al.*, 2019[41] (supervised)	0.962	31.75
Pham *et al.*, 2017[57] (supervised)	**0.977**	35.84
ours	0.971	**35.85**

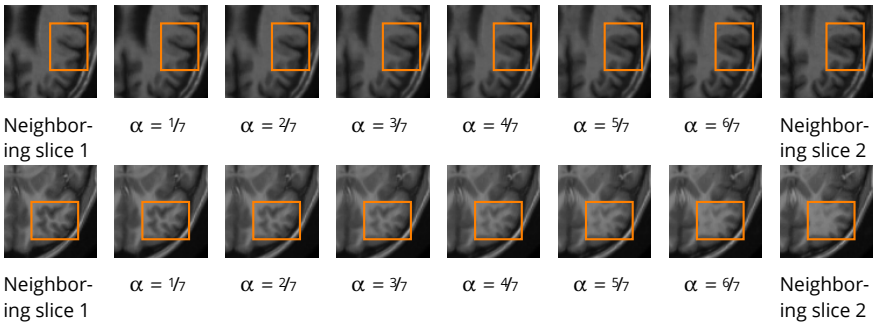


Figure 5.11: Zoomed-in images of synthesis results in scans of adult brain MRIs of two different patients (one example per row) from the OASIS dataset. Slice spacing was improved from 4 to 0.57 mm by synthesizing six intermediate slices (second to penultimate columns). For this, latent space encodings of the two neighboring slices (first and last column) were combined using their convex combination. α denotes the mixing coefficient as specified in Equation 5.1. Bounding boxes focus on anatomical variations between images in a row.

Finally, methods were qualitatively compared for a through-plane upsampling factor of ten using the original cardiac volumes. Visual inspection of the results discloses that proposed method can generate volumes with a higher image quality than cubic B-spline interpolation. Qualitative comparison shown in Figure 5.7 reveals that performance differences are most pronounced for the myocardial structures of the left ventricle. These structures show smoother transitions between adjacent axial slices when generated by proposed method compared to cubic B-spline interpolation. The latter method generates volumes that suffer from aliasing artifacts while the proposed method can mostly suppress such artifacts.

5.6.3 Semantic Interpolation of Neonatal Brain MRI

Acquisition of high-resolution neonatal brain MRIs is typically hampered by uncontrollable full-term infant motion and their small size of the brain.[58] As a result, acquired neonatal brain MRIs are often anisotropic and poorly capture the 3D brain structures.[5]

Super-resolution of neonatal brain MRI may enhance the capacity of image analysis on the dynamics of brain maturation[59] and brain development.[60,61] Therefore, our proposed approach was evaluated using 240 randomly selected T_2-weighted (T2w) neonatal brain MRIs from the developing Human Connectome Project[39] (dHCP) hereafter referred to as dHCP dataset.

EXPERIMENTAL DETAILS The dataset was randomly split into training (200), validation (20) and test set (20). The images with isotropic resolution of $0.5\times0.5\times0.5\,\text{mm}^3$ served as ground truth high-resolution (HR) images. To simulate low through-plane resolution (LR) images were downsampled by factor $K \in \{2, 3, 4, 5, 6\}$ in the z-axis. More specifically, low-resolution images were generated by using a Gaussian blur with the full-width-at-half-maximum (FWHM) set to the desired slice thickness.[4] Subsequently, volumes were downsampled with factor K by including every K^{th} slice in the test images to obtain $0.5\times0.5\times K*0.5\,\text{mm}^3$ resolution. To assess upsampling performance of the proposed method downsampled test volumes were upsampled in through-plane direction by synthesizing $K-1$ new slices between each pair of neighboring slices using the set of mixing coefficients as defined in Equation 5.2. Resulting volumes were then compared with high-resolution ground truth data.

To train a model patches of 64×64 voxels were randomly chosen from the training set using mini-batches of 8 randomly selected slice pairs (16 slices) originating from the same volume as described in Section 5.4. To balance loss terms in Equation 5.3, λ was set to 0.001. A model was trained in 1300 epochs and the best performing model on the validation set was selected for final evaluation on the test volumes.

RESULTS **Slice Synthesis:** Qualitative evaluation of the proposed approach on neonatal brain MRI with reveals that generated slices using a convex combination of neighboring slice encodings, comprise a smooth anatomical transition between adjacent slices. Examples depicted in Figure 5.8 show upsampling performance of proposed method for neighboring slices with large anatomical variations. In the depicted figures slice spacing was improved from 2 to 0.29 mm by synthesizing six intermediate slices using latent space encodings of the two neighboring slices.

Visual inspection of Figure 5.9 conveys that our proposed approach was able to synthesize excluded high-resolution axial slices more accurately than cubic B-spline interpolation. These results are corroborated by the coronal and sagittal views revealing that volumes generated by the proposed method are less blurry and contain smoother transitions between slices compared to volumes generated by the conventional interpolation method.

Moreover, quantitative comparison shown in Figure 5.10 depicts that the proposed unsupervised method outperformed cubic B-spline interpolation in terms of SSIM and PSNR for all evaluated upsampling factors. Furthermore, the performance differences

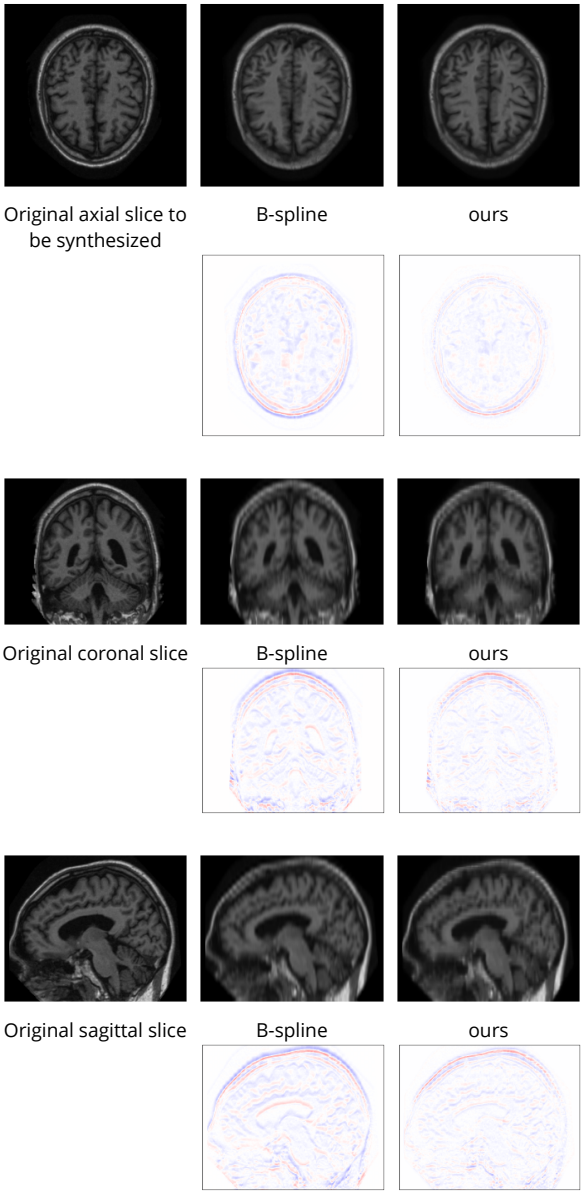


Figure 5.12: Comparing upsampling performance between cubic B-spline interpolation and proposed approach using T_1-weighted adult brain MRI of the OASIS dataset. Original volumes with slice thickness and spacing of 1 mm were downsampled to 5 mm by applying a Gaussian blur before including every fifth slice in the test volume. Differences between reference (minuend) and synthesized slice (subtrahend). Blue corresponds to negative and red to positive differences. Image intensities are scaled to a $[0, 1]$ range. All difference images use the same color scale $[-1, 1]$.

are statistically significant ($p < 0.001$) using the one-sided Wilcoxon signed-rank test.

Comparison With Supervised Super-Resolution Methods: Supervised deep-learning super-resolution methods developed by Pham *et al.*[41,57] were evaluated on a subset of 20 neonatal brain MRIs from the dHCP dataset. Table 5.3 lists results as reported by Pham *et al.*[41] together with quantitative evaluation of our proposed approach on the same dataset (see Section 5.6.2). Although methods of Pham *et al.*[41,57] used high-resolution ground-truth data for model training their results can be used to put results reported in this work into perspective. One may carefully conclude that our proposed unsupervised deep-learning approach is on par with supervised super-resolution approaches by Pham *et al.*[41,57] when evaluated on neonatal brain MRI.

5.6.4 Semantic Interpolation of Adult Brain MRI

Upsampling performance of the proposed method was also evaluated using T_1-weighted (T1w) adult brain MRIs of the OASIS project.[40] The images with isotropic resolution of $1.0 \times 1.0 \times 1.0\,\text{mm}^3$ served as ground truth high-resolution (HR) images.

EXPERIMENTAL DETAILS Model was trained on the first 200 unique patients, validated on the subsequent 20 patients, and tested on 50 randomly selected scans from the remaining set. To ensure that test images were divisible by four, slices were zero-padded to 220×220 voxels. Other experimental settings were identical to experiments performed on neonatal brain MRI described in section 5.6.3.

RESULTS **Slice Synthesis:** Qualitative evaluation of proposed method on adult brain MRI with reveals that synthesized slices constitute a smooth anatomical transition between neighboring slices. The proposed method is able to bridge large anatomical variations between adjacent slices. These findings are depicted in Figure 5.11.

Comparison With Conventional Interpolation Method: Qualitative comparison of generated axial brain MRI slices between cubic B-spline interpolation and proposed approach shown in Figure 5.12 reveals that the proposed method can synthesize excluded axial slices with higher image quality than conventional interpolation method. Moreover, visual inspection of coronal and sagittal slices shown in Figure 5.12 conveys that images generated by cubic B-spline interpolation more frequently suffer from aliasing artifacts than images generated by our proposed method.

In line with results reported for cardiac cine and neonatal brain MRI in Sections 5.6.2 and 5.6.3, respectively, quantitative evaluation in terms of SSIM, PSNR, and VIF depicted in Figure 5.13 corroborate the qualitative findings. Measures were computed on sagittal slices through volume. The proposed method outperformed cubic B-spline interpolation and the differences are statistically significant ($p < 0.0001$) in terms of SSIM and PSNR for all upsampling factors ($K \in \{2, 3, 4, 5, 6\}$).

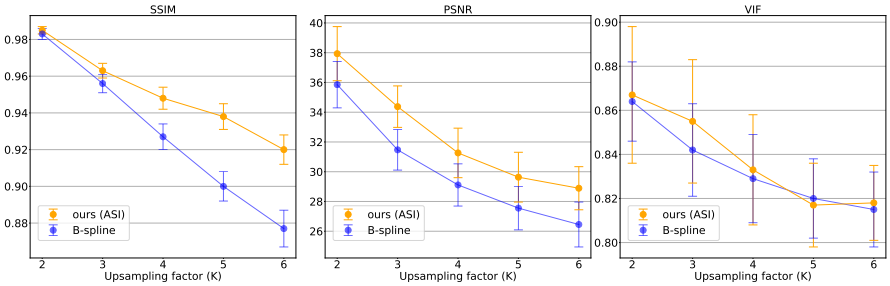


Figure 5.13: Quantitative comparison of upsampling performance for proposed method (ASI) and cubic B-spline interpolation in terms of SSIM, PSNR, and VIF. T_1-weighted adult brain MRIs of 50 subjects from the OASIS dataset were upsampled with factor $K \in \{2, 3, 4, 5, 6\}$. The differences between proposed and cubic B-spline interpolation approach in terms of SSIM and PSNR are statistically significant ($p < 0.001$) using the one-sided Wilcoxon signed-rank test.

Comparison With Unsupervised Super-Resolution Methods: Previously developed unsupervised super-resolution methods by Jog *et al.*[28] and Zhao *et al.*[29] were evaluated on T_1-weighted adult brain MRIs from the Neuromorphometrics dataset. The dataset is not publicly available. From the 114 brain scans in the Neuromorphometrics dataset a subset of 60 scans (30 patients) was taken from the OASIS project. Therefore, reported results of previously developed methods for images from the Neuromorphometrics dataset can be used as estimates for a careful comparison. Quantitative results in terms of PSNR and SSIM listed in Table 5.4 show that for upsampling factor 2 our proposed method is on par or better than previously developed super-resolution approaches. For upsampling factor 3 the method of Zhao *et al.*[29] achieved the best results and our proposed approach is on par with method of Jog *et al.*[28]

Generative image synthesis approach of Dalca *et al.*[32] was evaluated on 50 T_1-weighted brain MRIs of the ADNI dataset.[62] To compare performances, our model was trained (100 scans) and evaluated (50 scans) on the ADNI dataset using identical experimental settings as described in section 5.6.4. Table 5.5 lists quantitative results for both methods in terms of mean squared error (MSE). One can observe that our method achieved a lower mean squared error compared with approach of Dalca *et al.*[32]

5.6.5 Ablation Study

To investigate the effect of the synthesis loss on upsampling performance, the proposed model was trained and evaluated on cardiac cine MRIs by minimizing the reconstruction loss only. For this, λ in Equation 5.3 is set to zero (referred to as $ASI_{\lambda=0}$). All other experimental conditions were held constant.

This setting enables performance comparison between model $ASI_{\lambda=0}$ and a model trained with the combined reconstruction and synthesis loss ($ASI_{\lambda=0.05}$). Figure 5.14 depicts qualitative comparison between the two models for image synthesis of cardiac

Table 5.4: Comparison between proposed method (ASI) and *unsupervised* super-resolution approaches of Jog *et al.*[28] and Zhao *et al.*[29] in terms of SSIM and PSNR. Approaches of Jog *et al.*[28] and Zhao *et al.*[29] were evaluated on T_1-weighted adult brain MRIs with 1.0 mm^3 isotropic resolution from the Neuromorphometrics dataset. Slice spacing was improved from 2 to 1 mm (factor 2) and 3 to 1 mm (factor 3). Results reported here are taken from the original work by Jog *et al.*[28] and Zhao *et al.*.[29]

Method	**Factor 2**		**Factor 3**	
	SSIM	PSNR	SSIM	PSNR
Jog *et al.*[28]	0.983	37.98	0.968	33.49
Zhao *et al.*[29]	0.976	35.14	**0.977**	**34.44**
ours	**0.985**	**38.01**	0.967	34.24

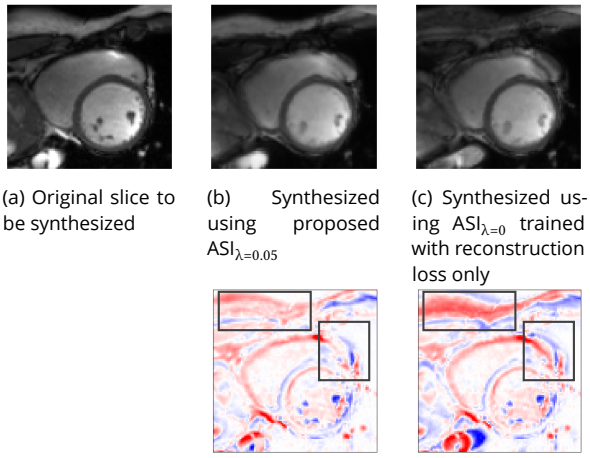
Table 5.5: Quantitative comparison of upsampling performance of proposed unsupervised method (ASI) compared with approach of Dalca *et al.*[32] in terms of mean squared error (MSE). Both approaches were evaluated on 50 randomly selected T_1-weighted adult brain MRIs with 1.0 mm^3 isotropic resolution from the ADNI dataset. Slice spacing was improved from 6 to 1 mm. Result listed here was reported in the work by Dalca *et al.*[32] Values represent medians over 50 scans.

Method	**MSE**
Dalca *et al.*[32]	2.1×10^{-3}
ours	1.5×10^{-3}

MRI. The results reveal that performance decreased for a model trained with the reconstruction loss only. The performance difference is more pronounced for larger anatomical variations between neighboring slices e.g. the shape of the right ventricle. Furthermore, Figure 5.15 demonstrates synthesis performance of the two models. The model trained with just the reconstruction loss ($ASI_{\lambda=0}$) generates *cross-fade* artifacts between the intensities of the two neighboring slices[63–65] whereas the model trained with a combination of reconstruction and synthesis loss ($ASI_{\lambda=0.05}$) can substantially suppress such artifacts.

These results are corroborated by quantitative evaluation in terms of SSIM, PSNR, and VIF depicted in Figure 5.16. One can observe that the model trained with the combined reconstruction and synthesis loss achieved better performance when evaluated by SSIM and PSNR compared to a model trained with the reconstruction loss only ($ASI_{\lambda=0}$). For the VIF measure the performance achievements are reversed. All differences are statistically significant ($p < 0.0001$) using the one-sided Wilcoxon signed-rank test.

To further investigate the effect of the synthesis loss on upsampling performance, separate quantitative evaluations were performed on reconstructed and synthesized short-axis slices, respectively. The results listed in Table 5.1 reveal that a model trained with the reconstruction loss only ($ASI_{\lambda=0}$) achieved better performance for the reconstruction task compared to a model trained with the combined reconstruction



(a) Original slice to be synthesized

(b) Synthesized using proposed $\text{ASI}_{\lambda=0.05}$

(c) Synthesized using $\text{ASI}_{\lambda=0}$ trained with reconstruction loss only

Figure 5.14: Comparison of synthesis performance between proposed model trained on cardiac MRI (ACDC dataset) using a combination of reconstruction and synthesis loss (b) compared to model trained with reconstruction loss only (denoted $\text{ASI}_{\lambda=0}$) (c). Bottom row: Differences between reference (minuend) and synthesized slice (subtrahend). Blue corresponds to negative and red to positive differences. Image intensities are scaled to a $[0, 1]$ range. All difference images use the same color scale $[-1, 1]$.

Figure 5.15: Comparison of synthesis performance between proposed model trained with (top row) reconstruction loss only (denoted $\text{ASI}_{\lambda=0}$) and (bottom row) model trained with a combination of reconstruction and synthesis loss $\text{ASI}_{\lambda=0.05}$. Second to penultimate columns show six synthesized intermediate slices using latent space encodings of the two neighboring slices (first and last column). Model trained with just reconstruction loss ($\text{ASI}_{\lambda=0}$) generates *cross-fade* artifacts (e.g. columns four to six) between the intensities of the two neighboring slices. Such artifacts are mostly suppressed by the $\text{ASI}_{\lambda=0.05}$ model. α denotes the mixing coefficient as specified in Equation 5.1.

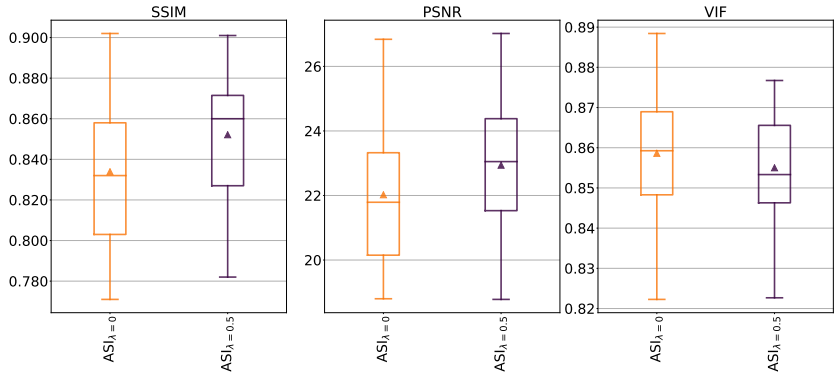


Figure 5.16: Boxplots compare performance of the proposed model trained with (a) reconstruction loss only (denoted $\text{ASI}_{\lambda=0}$) compared with model trained with (b) reconstruction and synthesis loss $\text{ASI}_{\lambda=0.05}$ in terms of SSIM, PSNR, and VIF. Cardiac cine MRIs of 20 patients from the ACDC dataset were upsampled with factor 2 in through-plane direction. A higher score indicates better performance. Measures were computed on sagittal slices through volume. Triangle indicates mean value.

and synthesis loss ($\text{ASI}_{\lambda=0.05}$). In contrast, the latter model performed better in terms of SSIM and PSNR when synthesizing the excluded slices. These results indicate that the additional synthesis loss resulted in increased interpolation performance but at the same time impacted reconstruction performance of the autoencoder.

Finally, Figure 5.17 shows the effect of different values of λ on reconstruction and synthesis performance of the proposed approach. Increasing the contribution of the synthesis loss, i.e. using larger values for λ, increases synthesis and lowers reconstruction performance of the model in terms of SSIM and PSNR.

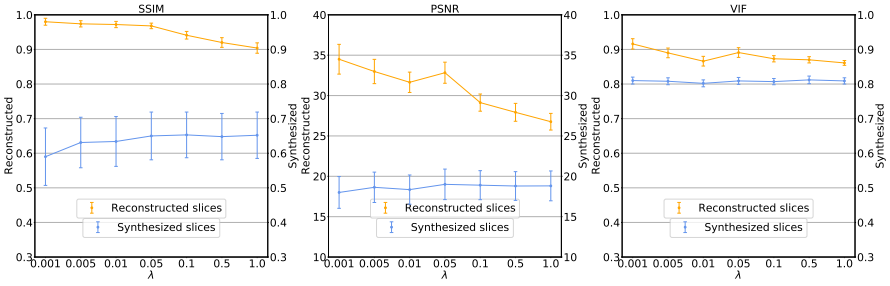


Figure 5.17: Quantitative comparison of reconstruction and synthesis performance in terms of SSIM, PSNR, and VIF between proposed model trained on cardiac cine MRIs (ACDC dataset) using different values for the hyperparameter $\lambda \in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1\}$ as specified in Equation 5.3. y-axis shows performance for reconstructed and synthesized slices, respectively.

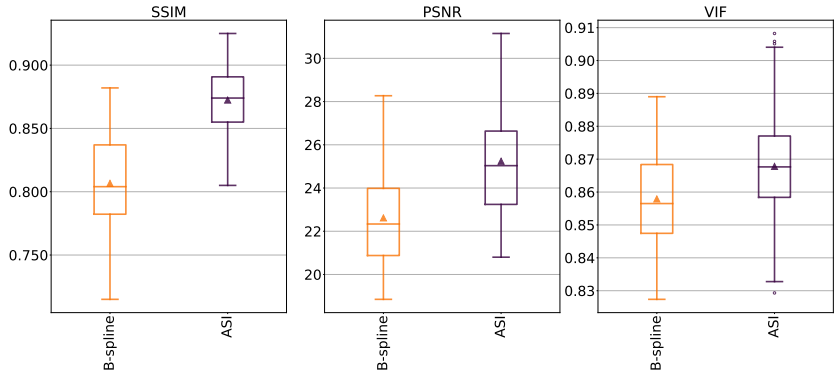


Figure 5.18: Boxplots showing results for upsampling of 90 cardiac MRIs from *Sunnybrook* dataset with upsampling factor 2 in through-plane direction for conventional interpolation method (cubic B-spline) compared to proposed method (ASI) in terms of SSIM, PSNR, and Visual Information Fidelity (VIF). Proposed approach was trained on CMR scans from *ACDC* dataset. Hence, depicted results demonstrate generalization performance of proposed method. A higher score indicates better performance. Measures were computed on sagittal slices through short-axis volume. Performance differences are statistically significant ($p < 0.0001$) using the one-sided Wilcoxon signed-rank test. Triangle indicates mean value.

5.6.6 Evaluation on Cardiac Cine MRIs from Sunnybrook dataset

To examine generalization performance of our proposed method a model trained on cardiac cine MRIs from the ACDC dataset was evaluated on cardiac MRI scans from the Sunnybrook dataset.[42] Figure 5.18 shows quantitative comparison for cubic B-spline compared with proposed method (ASI) in terms of SSIM, PSNR, and VIF. One can observe that the proposed approach trained on ACDC images outperformed cubic B-spline interpolation for all measures on Sunnybrook dataset. The latter methods do not require training. These differences are statistically significant ($p < 0.0001$) using the one-sided Wilcoxon signed-rank test. Furthermore, the results depict that relative performance differences between methods are nearly identical to those observed on ACDC dataset depicted in Figure 5.5. Finally, achieved performance on cardiac MRI scans of Sunnybrook dataset is higher for all measures compared with performance reported for evaluation on ACDC dataset depicted in Figure 5.5. This might have been caused by scans of Sunnybrook dataset having more consistent image quality, better alignment of adjacent slices, and higher bit depth compared with scans from ACDC dataset.

5.7 Discussion

A method for unsupervised deep-learning image synthesis of medical images has been presented. To synthesize new intermediate slices and thereby recovering spatial in-

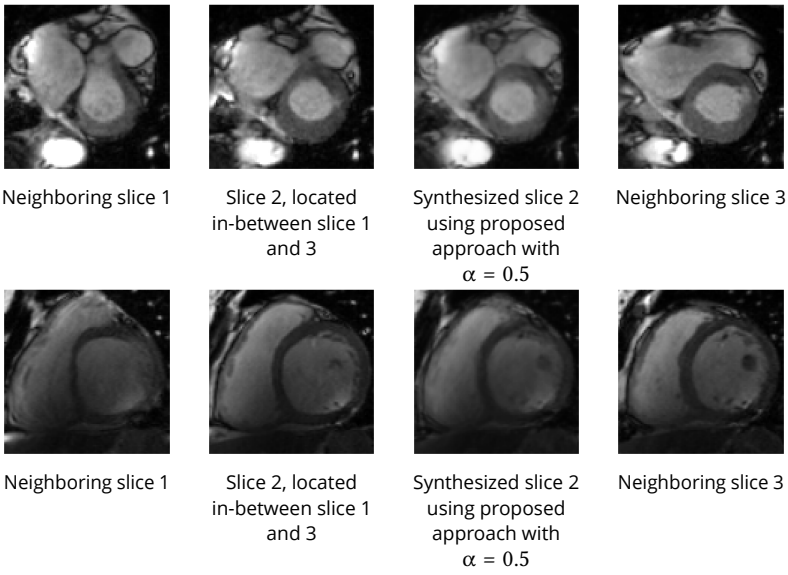


Figure 5.19: Qualitative comparison for upsampling factor 2 on cardiac MRI (ACDC dataset). Each row depicts an example of synthesizing intermediate slice 2 (second column) using latent space encodings of the two neighboring slices 1 and 3 (first and last column). The synthesized slice is shown in penultimate column. First row shows basal slices with large anatomical variations compared to second row that depicts mid-ventricular slices with mild anatomical variations. α denotes the mixing coefficient as specified in Equation 5.1. Scans have a slice spacing of 10 mm.

.

formation, the method exploits the latent space interpolation ability of autoencoders. New intermediate slices are generated by mixing the latent space encodings of two spatially adjacent slices. High-resolution ground-truth images are not required to train the approach. Results of our preliminary experiments using MNIST data demonstrated that our proposed approach outperformed a variational autoencoder (VAE) and Adversarially Constrained Autoencoder Interpolation (ACAI) approach[48] for interpolating rotations of handwritten digits. Evaluation of the approach on cardiac and brain structures using four publicly available MRI datasets revealed that the method can outperform cubic B-spline interpolation. Performance differences between evaluated methods become more apparent when the upsampling task becomes more difficult i.e. for highly anisotropic volumes e.g. cardiac MRI or larger through-plane upsampling factors. This might indicate that the model can infer the missing information from contextual and anatomical information captured in the latent space. Furthermore, the experimental results revealed that our proposed approach can compete with related unsupervised[28,29] and supervised[41] super-resolution approaches. Compared with unsupervised super-resolution methods of Jog *et al.*,[28] Zhao *et al.*,[29–31] Dalca *et al.*[32] and Xia *et al.*,[34] our approach can be applied with any desired upsampling factor and uses a single encoder-decoder structure. Moreover, applying methods of Jog *et al.*[28] and Zhao *et al.*[29–31] requires optimization during inference for every image at hand and therefore, inference requires several minutes of GPU processing time.[31] In contrast, at test time our method can synthesize multiple intermediate slices between each pair of adjacent slices in an MRI scan in less than a second on a GPU. Furthermore, using the method of Dalca *et al.*[32] necessitates creation of a common atlas space to which each image must be transformed. Finally, it is fair to note that the approach of Zhao *et al.*[30,31] performs explicitly anti-aliasing by using an additional CNN.

Even though autoencoders are designed to learn a lower-dimensional representation of the input while minimizing information loss, in this work, we used them to perform semantic interpolation and thereby recover spatial information in anisotropic medical images. Specifically, we used an over-complete autoencoder that can potentially retain all information contained in the input. Theoretically, such an approach could learn an identity to minimize the reconstruction loss. Nonetheless, in line with previous research,[66] the results presented in this work seem to indicate that such a model can learn a useful representation of its input. Furthermore, the experimental results revealed that interpolation between image representations is feasible to approximate information orthogonal to the input images. This suggests that the proposed approach learns to extract contextual and high level conceptual information from the input images. Moreover, the results demonstrate that the decoder learns to exploit this information to instantiate semantically meaningful intermediate slices. While previously developed shape-based interpolation approaches of Raya et al.[67] and Grevera *et al.*[68] exploit anatomical shape information to achieve high-order interpolation between cross sections of 3D anatomical structures, we argue that our approach performs

semantic interpolation between two spatially adjacent slices.

Synthesized images are not guaranteed to be semantically meaningful. However, the solution space of the proposed approach is constrained using encodings of two spatially adjacent slices. Moreover, compared with an adversarial training objective[48] training the autoencoder with the proposed synthesis loss enforces an explicit constraint on the solution space because the synthesized image is evaluated against its reference. Nevertheless, image interpolation in latent space can still result in *cross-fade* artifacts between the intensities of the two images i.e. neighboring slices.[63–65] Appearance of such an artifact can be observed in Figure 5.8 (second row, columns four to six). However, results presented in Figure 5.15 demonstrated that the proposed model can synthesize images with substantially less artifacts compared with a standard autoencoding approach. Although adversarial approaches are extremely difficult to train,[69] adding a critic to our approach could further constrain the model and improve synthesis performance. Moreover, to further constrain the model an additional synthesis loss in latent space could have been proposed.[65] We deliberately leave the challenge of defining a useful distance metric in latent space for future work.

Our approach assumes that a linear spacing in latent space corresponds to slice spacing in image space. Although, alternatives such as spherical latent space interpolation[70,71] or an enforced Riemannian latent space[63,72] can be used, linear interpolation in the latent space of an autoencoder trained with the proposed synthesis loss showed excellent results. Nevertheless, human anatomy does not change linearly along spatial dimensions. We conjecture that the model can learn such a nonlinearity from the training data. Furthermore, we presume that the synthesis loss encourages the model to learn the nonlinear mapping between *distances* in latent and image space. For example, our experiments on cardiac MRI revealed that the model has learned that structural changes at the base of the heart are substantially different than at the apex. Finally, our experiments on neonatal and adult brain MRI apply mixing coefficients (α) unequal to 0.5. Results of these experiments corroborate our assumption that linear steps taken in latent space can approximate anatomical distances in image space, however, the approach does not guarantee such a relationship to be exact.

Performance of the proposed method is affected by large anatomical variations between adjacent slices as shown in Figure 5.19. Furthermore, quality of synthesized images is decreased when adjacent slices within the original volume are misaligned. This is a known problem in cardiac cine MR imaging caused by surrounding organ motion during breath-hold acquisition. In these cases intermediate points along the interpolation path spanned by two adjacent slice encodings result in anatomically implausible images i.e. cross-fades. This might indicate that the latent space between the two endpoints is too sparsely populated. For cardiac cine MRI this could be alleviated by choosing additional interpolation endpoints e.g. from other time frames. This direction will be investigated in future work.

Training the autoencoder with a combination of reconstruction and synthesis

loss slightly hampered reconstruction performance when compared to training with reconstruction loss only as shown in Table 5.1. However, quality of synthesized slices considerably improved when adding the synthesis loss during training. To render 3D high-resolution images the method only relies on the quality of newly synthesized slices. Hence, all existing slices do not need to be reconstructed but can be taken from the original anisotropic 3D volumes.

The here presented approach was developed in parallel with the super-resolution method proposed by Xia *et al.*[34] In the current work quantitative results are presented in terms of SSIM, PSNR and VIF for all cardiac MRI slices of test patients from the ACDC dataset and 45 subjects from the Sunnybrook dataset. In comparison, work of Xia *et al.*[34] depicts results in terms of PSNR and correlation coefficient for a limited selection of two cardiac MRI slices (mid-ventricular and basal) from the UK Biobank. Note that intensity statistics of images from different datasets may be very different and hence, PSNR measurements might be inaccurate.[18] Therefore, thorough quantitative comparison of the two methods is hardly feasible.

To conclude, we presented a method for unsupervised semantic interpolation of anisotropic 3D medical images achieving anatomically smooth transitions in through-plane direction. New intermediate slices are generated by mixing the latent space encodings of two spatially adjacent slices. The experiments using cardiac cine and brain MRIs demonstrated that the proposed approach outperforms cubic B-spline interpolation on cardiac cine and brain MRIs. Given the unsupervised nature of the method, high-resolution training data is not required and hence, the method can be readily applied in clinical settings.

References

[1] M. Lustig, D. Donoho, and J. M. Pauly. "Sparse mri: the application of compressed sensing for rapid mr imaging," *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 58 (2007), pp. 1182–1195.

[2] R. M. Heidemann, Ö. Özsarlak, P. M. Parizel, J. Michiels, B. Kiefer, V. Jellus, M. Müller, F. Breuer, M. Blaimer, M. A. Griswold, et al. "A brief review of parallel magnetic resonance imaging," *European radiology*, vol. 13 (2003), pp. 2323–2337.

[3] C. E. Duchon. "Lanczos filtering in one and two dimensions," *Journal of Applied Meteorology and Climatology*, vol. 18 (1979), pp. 1016–1022.

[4] H. Greenspan. "Super-resolution in medical imaging," *The computer journal*, vol. 52 (2009), pp. 43–63.

[5] A. Gholipour, J. A. Estroff, and S. K. Warfield. "Robust super-resolution volume reconstruction from slice acquisitions: application to fetal brain mri," *IEEE transactions on medical imaging*, vol. 29 (2010), pp. 1739–1758.

[6] S. Peled and Y. Yeshurun. "Superresolution in MRI: application to human white matter fiber tract visualization by diffusion tensor imaging," *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 45 (2001), pp. 29–35.

[7] R. R. Peeters, P. Kornprobst, M. Nikolova, S. Sunaert, T. Vieville, G. Malandain, R. Deriche, O. Faugeras, M. Ng, and P. Van Hecke. "The use of super-resolution techniques to reduce slice thickness in functional MRI," *International Journal of Imaging Systems and Technology*, vol. 14 (2004), pp. 131–138.

[8] J. V. Manjón, P. Coupé, A. Buades, V. Fonov, D. L. Collins, and M. Robles. "Non-local mri upsampling," *Medical image analysis*, vol. 14 (2010), pp. 784–792.

[9] A. Rueda, N. Malpica, and E. Romero. "Single-image super-resolution of brain mr images using overcomplete dictionaries," *Medical image analysis*, vol. 17 (2013), pp. 113–132.

[10] W. Shi, J. Caballero, C. Ledig, X. Zhuang, W. Bai, K. Bhatia, A. M. S. M. de Marvao, T. Dawes, D. O'Regan, and D. Rueckert. "Cardiac image super-resolution with global correspondence using multi-atlas patchmatch," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. 2013, pp. 9–16.

[11] K. K. Bhatia, A. N. Price, W. Shi, J. V. Hajnal, and D. Rueckert. "Super-resolution reconstruction of cardiac mri using coupled dictionary learning," *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, IEEE. 2014, pp. 947–950.

[12] D. C. Alexander, D. Zikic, J. Zhang, H. Zhang, and A. Criminisi. "Image quality transfer via random forest regression: applications in diffusion mri," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. 2014, pp. 225–232.

[13] C. Dong, C. C. Loy, K. He, and X. Tang. "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38 (2015), pp. 295–307.

[14] J. Kim, J. Kwon Lee, and K. Mu Lee. "Deeply-recursive convolutional network for image super-resolution," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1637–1645.

[15] O. Oktay, W. Bai, M. Lee, R. Guerrero, K. Kamnitsas, J. Caballero, A. de Marvao, S. Cook, D. O'Regan, and D. Rueckert. "Multi-input cardiac image super-resolution using convolutional neural networks," *International conference on medical image computing and computer-assisted intervention*, Springer. 2016, pp. 246–254.

[16] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. "Photo-realistic single image super-resolution using a generative adversarial network," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[17] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee. "Enhanced deep residual networks for single image super-resolution," *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 136–144.

[18] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. De Marvao, T. Dawes, D. P. O'Regan, et al. "Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation," *IEEE transactions on medical imaging*, vol. 37 (2017), pp. 384–395.

[19] R. Timofte, S. Gu, J. Wu, and L. Van Gool. "Ntire 2018 challenge on single image super-resolution: methods and results," *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 852–863.

[20] Y. Chen, Y. Xie, Z. Zhou, F. Shi, A. G. Christodoulou, and D. Li. "Brain MRI super resolution using 3d deep densely connected neural networks," *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, IEEE. 2018, pp. 739–742.

[21] J. Shi, Q. Liu, C. Wang, Q. Zhang, S. Ying, and H. Xu. "Super-resolution reconstruction of mr image with a novel residual learning network algorithm," *Physics in Medicine & Biology*, vol. 63 (2018), p. 085011.

[22] N. Basty and V. Grau. "Super resolution of cardiac cine mri sequences using deep learning," *Image Analysis for Moving Organ, Breast, and Thoracic Images*, Springer, 2018, pp. 23–31.

[23] Y. Chen, F. Shi, A. G. Christodoulou, Y. Xie, Z. Zhou, and D. Li. "Efficient and accurate mri super-resolution using a generative adversarial network and 3d multi-level densely connected network," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. 2018, pp. 91–99.

[24] C.-H. Pham, C. Tor-Díez, H. Meunier, N. Bednarek, R. Fablet, N. Passat, and F. Rousseau. "Multiscale brain MRI super-resolution using deep 3d convolutional networks," *Computerized Medical Imaging and Graphics*, vol. 77 (2019), p. 101647.

[25] D. Mahapatra, B. Bozorgtabar, and R. Garnavi. "Image super-resolution using progressive generative adversarial networks for medical image analysis," *Computerized Medical Imaging and Graphics*, vol. 71 (2019), pp. 30–39.

[26] K. Xuan, D. Wei, D. Wu, Z. Xue, Y. Zhan, W. Yao, and Q. Wang. "Reconstruction of isotropic high-resolution mr image from multiple anisotropic scans using sparse fidelity loss and adversarial regularization," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. 2019, pp. 65–73.

[27] E. M. Masutani, N. Bahrami, and A. Hsiao. "Deep learning single-frame and multiframe super-resolution for cardiac mri," *Radiology* (2020), p. 192173.

[28] A. Jog, A. Carass, and J. L. Prince. "Self super-resolution for magnetic resonance images," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. 2016, pp. 553–560.

[29] C. Zhao, A. Carass, B. E. Dewey, and J. L. Prince. "Self super-resolution for magnetic resonance images using deep networks," *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, IEEE. 2018, pp. 365–368.

[30] C. Zhao, A. Carass, B. E. Dewey, J. Woo, J. Oh, P. A. Calabresi, D. S. Reich, P. Sati, D. L. Pham, and J. L. Prince. "A deep learning based anti-aliasing self super-resolution algorithm for mri," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. 2018, pp. 100–108.

[31] C. Zhao, B. E. Dewey, D. L. Pham, P. A. Calabresi, D. S. Reich, and J. L. Prince. "Smore: a self-supervised anti-aliasing and super-resolution algorithm for mri using deep learning," *IEEE transactions on medical imaging*, vol. 40 (2020), pp. 805–817.

[32] A. V. Dalca, K. L. Bouman, W. T. Freeman, N. S. Rost, M. R. Sabuncu, and P. Golland. "Medical image imputation from image collections," *IEEE transactions on medical imaging*, vol. 38 (2018), pp. 504–514.

[33] J. Sander, B. D. de Vos, and I. Išgum. "Unsupervised super-resolution: creating high-resolution medical images from low-resolution anisotropic examples," *Medical Imaging 2021: Image Processing*, vol. 11596 International Society for Optics and Photonics. (2021), 115960E.

[34] Y. Xia, N. Ravikumar, J. P. Greenwood, S. Neubauer, S. E. Petersen, and A. F. Frangi. "Super-resolution of cardiac MR cine imaging using conditional GANs and unsupervised transfer learning," *Medical Image Analysis* (2021), p. 102037.

[35] M. Delbracio and G. Sapiro. "Removing camera shake via weighted fourier burst accumulation," *IEEE Transactions on Image Processing*, vol. 24 (2015), pp. 3293–3307.

[36] H. Jiang, D. Sun, V. Jampani, M.-H. Yang, E. Learned-Miller, and J. Kautz. "Super slomo: high quality estimation of multiple intermediate frames for video interpolation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9000–9008.

[37] W. Bao, W.-S. Lai, C. Ma, X. Zhang, Z. Gao, and M.-H. Yang. "Depth-aware video frame interpolation," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3703–3712.

[38] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester, et al. "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?" *IEEE Transactions on Medical Imaging* (2018).

[39] E. Hughes, L. Cordero-Grande, M. Murgasova, J. Hutter, A. Price, A. D. S. Gomes, J. Allsop, J. Steinweg, N. Tusor, J. Wurie, et al. "The developing human connectome: announcing the first release of open access neonatal brain imaging," *23rd Annual Meeting of the Organization for Human Brain Mapping*, 2017.

[40] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner. "Open access series of imaging studies (oasis): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults," *Journal of cognitive neuroscience*, vol. 19 (2007), pp. 1498–1507.

[41] C.-H. Pham, C. Tor-Díez, H. Meunier, N. Bednarek, R. Fablet, N. Passat, and F. Rousseau. "Simultaneous super-resolution and segmentation using a generative adversarial network: application to neonatal brain MRI," *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, IEEE. 2019, pp. 991–994.

[42] P. Radau, Y. Lu, K. Connelly, G. Paul, A. Dick, and G. Wright. "Evaluation framework for algorithms segmenting short axis cardiac mri," *The MIDAS Journal-Cardiac MR Left Ventricle Segmentation Challenge*, vol. 49 (2009).

[43] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. "Learning internal representations by error propagation," tech. rep. California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[44] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber. "Stacked convolutional auto-encoders for hierarchical feature extraction," *International conference on artificial neural networks*, Springer. 2011, pp. 52–59.

[45] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. "The unreasonable effectiveness of deep features as a perceptual metric," *CVPR*, 2018.

[46] K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556* (2014).

[47] D. P. Kingma and M. Welling. "Stochastic gradient vb and the variational auto-encoder," *Second International Conference on Learning Representations, ICLR*, vol. 19 (2014), p. 121.

[48] D. Berthelot, C. Raffel, A. Roy, and I. Goodfellow. "Understanding and improving interpolation in autoencoders via an adversarial regularizer," *ICLR workshop poster*, 2019.

[49] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. "Automatic differentiation in PyTorch," *NIPS Autodiff Workshop*, 2017.

[50] D. Kingma and J. Ba. "Adam: a method for stochastic optimization," *ICLR*, vol. 5 (2015).

[51] H. R. Sheikh, A. C. Bovik, and G. De Veciana. "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Transactions on image processing*, vol. 14 (2005), pp. 2117–2128.

[52] A. Mason, J. Rioux, S. E. Clarke, A. Costa, M. Schmidt, V. Keough, T. Huynh, and S. Beyea. "Comparison of objective image quality metrics to expert radiologists' scoring of diagnostic quality of mr images," *IEEE transactions on medical imaging*, vol. 39 (2019), pp. 1064–1072.

[53] E. H. Meijering, W. J. Niessen, and M. A. Viergever. "Quantitative evaluation of convolution-based methods for medical image interpolation," *Medical image analysis*, vol. 5 (2001), pp. 111–126.

[54] T. M. Lehmann, C. Gonner, and K. Spitzer. "Addendum: b-spline interpolation in medical image processing," *IEEE Transactions on Medical Imaging*, vol. 20 (2001), pp. 660–665.

[55] Y. LeCun. "The mnist database of handwritten digits," *http://yann. lecun. com/exdb/mnist/* (1998).

[56] D. J. Rezende, S. Mohamed, and D. Wierstra. "Stochastic backpropagation and approximate inference in deep generative models," *International conference on machine learning*, PMLR. 2014, pp. 1278–1286.

[57] C.-H. Pham, A. Ducournau, R. Fablet, and F. Rousseau. "Brain MRI super-resolution using deep 3d convolutional networks," *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, IEEE. 2017, pp. 197–200.

[58] J. Dubois, M. Alison, S. J. Counsell, L. Hertz-Pannier, P. S. Hüppi, and M. J. Benders. "Mri of the neonatal brain: a review of methodological challenges and neuroscientific advances," *Journal of Magnetic Resonance Imaging*, vol. 53 (2021), pp. 1318–1343.

[59] O. Glenn and A. Barkovich. "Magnetic resonance imaging of the fetal brain and spine: an increasingly important tool in prenatal diagnosis, part 1," *American Journal of Neuroradiology*, vol. 27 (2006), pp. 1604–1611.

[60] M. Rutherford, S. Jiang, J. Allsop, L. Perkins, L. Srinivasan, T. Hayat, S. Kumar, and J. Hajnal. "Mr imaging methods for assessing fetal brain development," *Developmental neurobiology*, vol. 68 (2008), pp. 700–711.

[61] P. Moeskops, I. Išgum, K. Keunen, N. H. Claessens, I. C. van Haastert, F. Groenendaal, L. S. de Vries, M. A. Viergever, and M. J. Benders. "Prediction of cognitive and motor outcome of preterm infants based on automatic quantitative descriptors from neonatal mr brain images," *Scientific reports*, vol. 7 (2017), pp. 1–10.

[62] C. R. Jack Jr, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, J. L. Whitwell, C. Ward, et al. "The alzheimer's disease neuroimaging initiative (adni): MRI methods," *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 27 (2008), pp. 685–691.

[63] G. Arvanitidis, L. Hansen, and S. Hauberg. "Latent space oddity: on the curvature of deep generative models," *Proceedings of International Conference on Learning Representations (ICLR)*, 2018.

[64] S. Laine. "Feature-based metrics for exploring the latent space of generative models," *ICLR workshop poster*, 2018.

[65] A. Oring, Z. Yakhini, and Y. Hel-Or. "Autoencoder image interpolation by shaping the latent space," *Proceedings of the 38th International Conference on Machine Learning*, edited by M. Meila and T. Zhang. Vol. 139 *Proceedings of Machine Learning Research* (PMLR, 2021), pp. 8281–8290.

[66] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, et al. "Greedy layer-wise training of deep networks," *Advances in neural information processing systems*, vol. 19 (2007), p. 153.

[67] S. P. Raya and J. K. Udupa. "Shape-based interpolation of multidimensional objects," *IEEE transactions on medical imaging*, vol. 9 (1990), pp. 32–42.

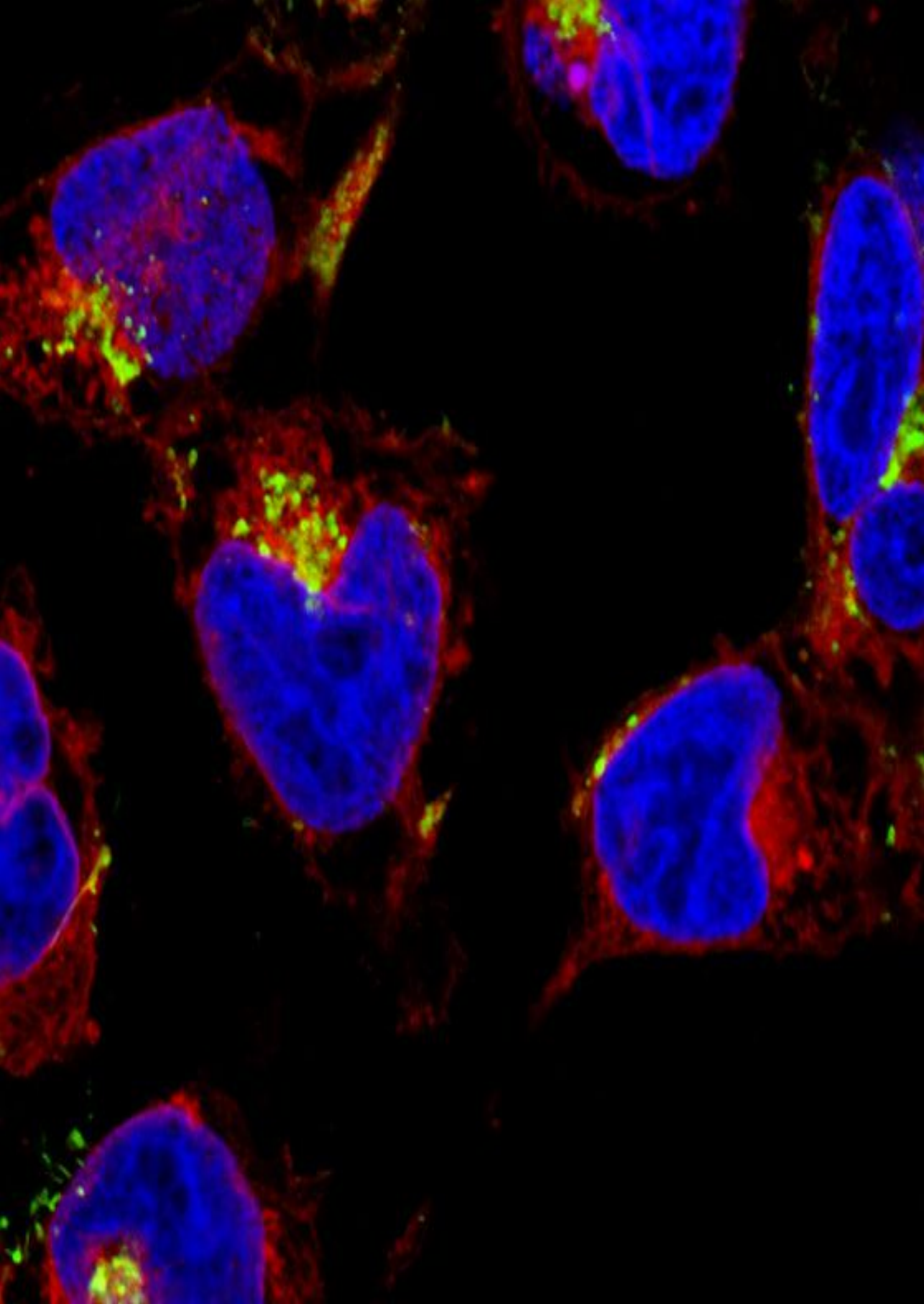
[68] G. J. Grevera and J. K. Udupa. "Shape-based interpolation of multidimensional grey-level images," *IEEE transactions on medical imaging*, vol. 15 (1996), pp. 881–892.

[69] M. Arjovsky and L. Bottou. "Towards principled methods for training generative adversarial networks," *arXiv preprint arXiv:1701.04862* (2017).

[70] A. Roberts, J. Engel, C. Raffel, C. Hawthorne, and D. Eck. "A hierarchical latent vector model for learning long-term structure in music," *International conference on machine learning*, PMLR. 2018, pp. 4364–4373.

[71] T. White. "Sampling generative networks: notes on a few effective techniques," *CoRR, abs/1609.04468*, vol. 7 (2016).

[72] H. Shao, A. Kumar, and P. Thomas Fletcher. "The riemannian geometry of deep generative models," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 315–323.



CHAPTER 6

High-resolution reconstruction and completion of anisotropic cardiac MRI segmentations using continuous implicit neural representations

---

This chapter is based on: J. Sander, B. D. de Vos, S. Bruns, N. Planken, M. A. Viergever, T. Leiner, and I. Išgum. "Reconstruction and completion of high-resolution 3d cardiac shapes using anisotropic cmri segmentations and continuous implicit neural representations," *Computers in Biology and Medicine* (2023), p. 107266.

Abstract

Since the onset of computer-aided diagnosis in medical imaging, voxel-based segmentation has emerged as the primary methodology for automatic analysis of left ventricle (LV) function and morphology in cardiac magnetic resonance images (CMRI). In standard clinical practice simultaneous multi-slice 2D cine short-axis MR imaging is performed under multiple breath-holds resulting in highly anisotropic 3D images. Furthermore, sparse-view CMRI often lacks whole heart coverage caused by large slice thickness and often suffers from inter-slice misalignment induced by respiratory motion. Therefore, these volumes only provide limited information about the true 3D cardiac anatomy which may hamper highly accurate assessment of functional and anatomical abnormalities. To address this, we propose a method that learns a continuous implicit function representing 3D LV shapes by training an auto-decoder. For training, high-resolution segmentations from cardiac CT angiography are used. The ability of our approach to reconstruct high-resolution shapes from sparse-view cardiac shape information is evaluated by using paired high- and low-resolution CMRI LV segmentations. The results show that the reconstructed LV shapes have an unconstrained subvoxel resolution and appear smooth and plausible in through-plane direction. Furthermore, these high-resolution reconstructed ventricle volumes are closer to the corresponding reference volumes than reference low-resolution volumes. Finally, the results demonstrate that the proposed approach allows recovering missing shape information and can correct motion artifacts.

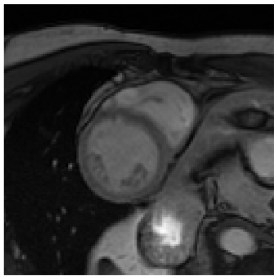
6.1 Introduction

Cardiovascular magnetic resonance (CMR) imaging is the reference modality for morphological and functional assessment of the heart.[1] Conventionally, to acquire stacks of short-axis 3D cine CMR images (CMRI) simultaneous multi-slice 2D cine CMR imaging is performed under multiple breath-holds. To mitigate the risk of motion artifacts and to sustain patient comfort fast scanning is often required. As a result, short-axis CMR scans with high temporal resolution (i) are often highly anisotropic (low through-plane resolution ranging between 5 and 10 mm), (ii) lack whole-heart coverage, and (iii) suffer from respiratory motion-induced inter-slice misalignment (example depicted in Figure 6.1c). As a consequence, left and right ventricular shapes obtained from manual or (semi-)automatic segmentations in short-axis CMRI only provide a sparse representation of the true 3D cardiac anatomy. Nevertheless, in current clinical practice these representations are used to compute volume-based imaging biomarkers, e.g., ejection fraction and stroke volume. Furthermore, volumetric indicators are limited to global cardiac function and therefore, cannot capture local functional and anatomical abnormalities.[2] Moreover, current shortcomings of anisotropic cardiac segmentations in CMRI hamper research progress in the field of computational cardiac physiology. The latter has recently shown the potential to improve and complement the management of cardiovascular diseases[3] by, e.g. accurately simulating cardiac electrophysiology and mechanics,[4] discovering new biomarkers for patient risk stratification,[5] and predicting cardiac outcomes.[6] Therefore, to increase accuracy and reproducibility of CMRI analysis and to enable advanced morphological and functional assessment of the heart, personalized high-resolution representation of cardiac anatomy is considered a prerequisite.[7,8]

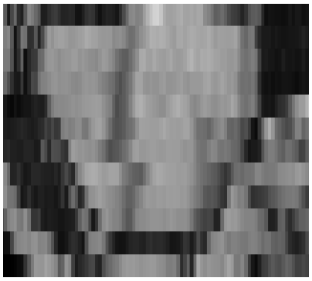
To obtain high-resolution 3D representations of the entire left and right ventricles from CMR images, previous approaches aim to increase through-plane resolution of anisotropic CMRI scans by imputing missing slices and/or correcting for motion artifacts after images are reconstructed. Methods are either applied to CMR images[9–15] or directly to the CMRI segmentations of the cardiac structures of interest.[16–21]

In this work, we focus on the latter, more specifically on the reconstruction and completion of high-resolution 3D cardiac shapes using anisotropic CMRI segmentations with inter-slice misalignment. In previous work, Oktay *et al.*[16] developed a deep learning approach that simultaneously performs voxel-based automatic CMRI segmentation and super-resolution of the obtained cardiac shapes. To correct for respiratory motion artifacts, an autoencoder is used to apply shape constraints during training. More recently, Duan *et al.*[17] designed a pipeline to generate high-resolution 3D bi-ventricular shapes from anisotropic short-axis CMR scans. To refine automatically obtained low-resolution short-axis segmentations they used atlas propagation.[17] A generative approach for high-resolution 3D segmentation of the left ventricle (LV) myocardium was proposed by Biffi *et al.*[18] In contrast with[16,17] the approach described by Biffi *et*

(a) Cardiac MRI long-axis view



(b) Cardiac MRI short-axis view



(c) Sagittal view through CMRI short-axis volume

Figure 6.1: Examples of cardiac MRI (CMRI) (a) long-axis; (b) short-axis and (c) low-resolution sagittal view through CMRI short-axis view with motion induced slice-misalignment.

al.[18] requires high-resolution long-axis segmentations during training and inference. This might hamper its applicability in clinical practice because such high-resolution long-axis segmentations are often not available. Recently, Wang *et al.*[20] proposed a latent optimization framework that jointly performs motion correction and super-resolution for short-axis CMRI segmentations. First, a latent space of high-resolution cardiac segmentations is learned by means of a generative model. Subsequently, given a low-resolution CMRI segmentation, the nearest high-resolution encoding is found by using an iterative optimization approach that downsamples and degrades the retrieved high-resolution shape to match the low-resolution CMRI segmentation. By contrast with the aforementioned methods that operate on a voxel-grid, Beetz *et al.*[19] proposed a method to reconstruct high-resolution bi-ventricular surfaces that uses point cloud data accumulated from sparse-view short- and long-axis CMRI segmentations. Recently, Beetz *et al.*[21] replaced the point completion network with a mesh deformation U-Net. To train previously developed approaches,[16–21] real or synthetic[22] high-resolution CMRI segmentations are required for training. Unfortunately, real high-resolution CMRI segmentations are in practice difficult to obtain. Furthermore, upsampling factor of the approaches depends on the resolution of the high-resolution CMRI reference segmentations and therefore, limits its applicability in case higher resolutions are required.

In contrast with the above approaches, our work exploits the high-resolution and fast acquisition of CT and uses segmentations from cardiac CT angiography (CCTA) to reconstruct and complete high-resolution 3D cardiac shapes from anisotropic incomplete CMRI segmentations. We adapt the probabilistic auto-decoder approach[23] that performs shape reconstruction using a latent space and continuous deep implicit function.[24] Instead of predicting voxelized representations of cardiac shapes at a fixed resolution, we use implicit neural representations that represent surfaces indepen-

dently of the resolution. Inasmuch as assessment of LV function and morphology is important for diagnosis of cardiovascular diseases, we focus on high-resolution shape reconstruction and completion of the LV.

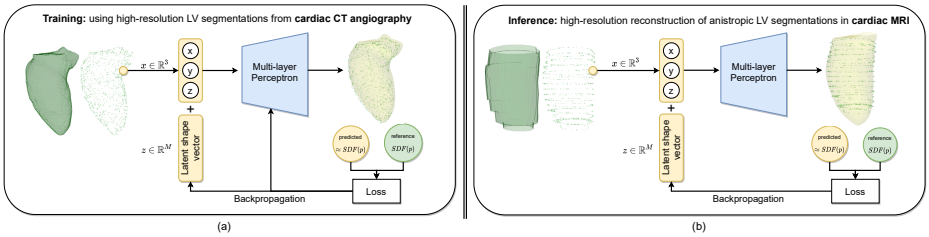
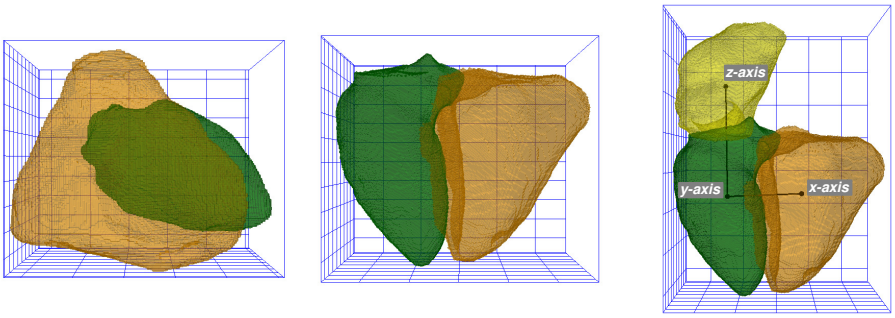


Figure 6.2: Visualization of the proposed method. To model the relationship between spatial points and a surface, a signed distance function (SDF) parametrized by a multi-layer perceptron (MLP) is used. A shape is implicitly represented by the zero iso-surface of its SDF. A single MLP can represent multiple shapes by conditioning the network output, i.e., a signed distance value on a spatial point x and a shape-specific latent vector z. During training, model weights and latent shape vectors are jointly optimized. To this end, a distance loss is computed between the reference and predicted signed distance values of the known shape coordinates. To encode a strong LV shape prior, the approach is trained using high-resolution segmentations in cardiac CT angiography (a). During inference (b), to perform high-resolution reconstruction of anisotropic cardiac MRI segmentations, a latent shape vector needs to be determined for each new shape. Best viewed in color.

To provide evidence that high-resolution 3D LV reconstructions approximate high-resolution LV reference segmentations better than the corresponding low-resolution LV reference segmentations, the method is evaluated on publicly available CMRI segmentations from the UK Digital Heart project. Moreover, we demonstrate that reconstruction performance of our approach can be improved by seamlessly integrating LV segmentations from different cardiac MRI views. Finally, to show that the proposed method can generalize to unseen LV shapes of patients suffering from a variety of cardiomyopathies, the approach is evaluated on publicly available segmentations from the MICCAI 2021 Multi-Disease, Multi-View & Multi-Center Right Ventricular Segmentation Challenge.[25]

6.2 Data

To reconstruct high-resolution LV shapes from low-resolution CMRI LV segmentations, the model is first trained on high-resolution CCTA segmentations of the LV (section 6.5.1). Subsequently, the model is evaluated on (i) 500 segmentations of the LV in paired 3D high- and low-resolution short-axis CMRI (section 6.5.2), and (ii) 360 LV segmentations in paired 3D low-resolution short-axis and 2D high-resolution 4-chamber long-axis CMRI (section 6.5.3).



(a) CCTA original orientation (b) CCTA in canonical orientation (c) CCTA with shape-specific coordinate frame

Figure 6.3: Visualization of (a) original, and (b) canonical orientation in reference coordinate system of CCTA segmentations of left (green) and right (orange) ventricles; (c) CCTA with shape-specific coordinate frame defined using three orthogonal unit vectors (x, y, z) where (i) x was defined as the vector pointing from the center of mass of the LV (green) to the center of mass of the right ventricle (orange), (ii) z was defined as the vector pointing from the center of mass of the LV to the center of mass of the left atrium (yellow), and (iii) y (pointing out of the Figure 6.3c towards the reader) was equal to the cross-product of x and z. Best viewed in color.

6.2.1 Cardiac CT angiography imaging

This study includes CCTA scans of 452 adult patient with acute ischemic stroke with median age of 72 years, 268 (59.3%) were men, and median baseline NIH Stroke Scale of 5. Prospective ECG-gated sequential CCTA (tube voltage 100 kVp and tube current 288 mAs) were acquired in end-diastole at the Amsterdam University Medical Center location UvA, The Netherlands.[26] In-plane resolutions of 3D CCTA images range from 0.27 to 0.52 mm while all scans have a slice thickness and increment of 0.6 mm and 0.4 mm, respectively. Reference annotations of the LV, right ventricle and left atrium were defined by automatic segmentation[27,28] and manual correction performed by two investigators if needed. For 434/452 (96%) of the patients CCTA LV reference segmentations were used. Since model training required high-resolution shapes that cover the complete LV, CCTA LV reference segmentations with incomplete coverage of the left ventricle or insufficient scan quality were excluded.

6.2.2 Cardiac cine MRI from UK Digital Heart Project

This work used publicly available cardiac CMRI segmentations of 1331 healthy adults from the UK Digital Heart Project (UKDHP) at Imperial College London.[22] The dataset is composed of pairs of high- and low-resolution short-axis CMRI segmentations of LV, right ventricle and LV myocardium for end-diastolic and end-systolic time

frames. In the experiments, only the end-diastolic time frame was considered. High-resolution segmentations were obtained from 3D balanced steady-state free precession (SSFP) cine sequences acquired in a single breath-hold. In contrast, low-resolution segmentations were obtained from 2D balanced SSFP cine sequences in different breath-holds and therefore may contain inter-slice motion artefacts. The provided reference segmentations were defined by automatic segmentation[29] and manually corrected by two expert clinicians if necessary.[22] High-resolution segmentation volumes are provided with a voxel resolution of $1.2\times1.2\times2\,\mathrm{mm}^3$ (average reconstruction matrix 281×281 voxels). In-plane resolution of low-resolution volumes range from 1.08 to 1.25 mm (average reconstruction matrix 287×287 voxels) and slice spacing between 8 to 10.12 mm. Each high- and low-resolution volume consists of on average 99 and 12 slices, respectively.

6.2.3 Cardiac cine MRI from M&Ms-2 challenge

This work included LV and right ventricle segmentations in paired 3D low-resolution short-axis and 2D high-resolution 4-chamber long-axis CMRI from the MICCAI 2021 Multi-Disease, Multi-View & Multi-Center Right Ventricular Segmentation Challenge in cardiac MRI (M&Ms-2).[25] The publicly available dataset contains images from 360 patients distributed over normal cardiac function and seven disease groups: hypertrophic cardiomyopathy, inter-atrial communication, arrhythmogenic cardiomyopathy, tetrology of fallot, dilated LV, dilated right ventricle, tricuspid regurgitation. Figures 6.1a and 6.1b show an example of a cardiac CMR long-axis and short-axis view, respectively. In-plane resolutions of short-axis stacks and long-axis slices range from 0.61 to 1.64 mm and 0.68 to 1.88 mm, with average reconstruction matrix 282×263 and 253×276 voxels. Slice thickness of short-axis volumes ranges from 5 to 19.2 mm while all long-axis slices have 1 mm slice thickness. Each short-axis volume consists of on average 11 slices covering the heart while each long-axis image consists of one slice. Expert manual reference segmentations are provided for the LV, right ventricle and LV myocardium for all short-axis and long-axis slices at end-diastolic and end-systolic time frames. In this work, only the end-diastolic time frame was considered. Consistency between short-axis and 4-chamber long-axis segmentations in basal and apical regions was guaranteed by two clinical experts who had to agree on the final annotation result. Moreover, the annotation protocol was identical to the one used for the MICCAI 2017 Automated Cardiac Diagnosis Challenge (ACDC) dataset.[30] For further details the reader is referred to Campello *et al.*[25] and the challenge website.

6.3 Method

We propose a method for reconstruction of high-resolution LV with complete shape from anisotropic incomplete CMRI segmentations. To this end, an LV shape is repre-

sented as zero iso-surface of its signed distance function (SDF). The SDF is parametrized using a multi-layer perceptron. A single multi-layer perceptron can represent multiple LV shapes by conditioning the model output on a learned latent shape vector. Therefore, the model takes a spatial point and a latent shape vector to predict the signed distance of a point to the closest surface. The approach is visualized in Figure 6.2.

6.3.1 Implicit neural representation of shapes

Our approach implicitly represents shapes as the zero iso-surface decision boundaries of a deep neural network trained to represent SDFs. An SDF is a continuous function that outputs the shortest distance ($s \in \mathbb{R}$) of a spatial point $x \in \mathbb{R}^3$ to the boundary of the shape. The sign of the distance encodes whether the point is inside (negative) or outside (positive) the surface. The set of SDF observations $\mathcal{O}_i$ of a shape i contains all coordinates and their corresponding signed distance values:

$$\mathcal{O}_i = \{(x_{ij}, s_{ij}) | s_{ij} = \mathrm{SDF}^i(x_{ij})\}\,, \tag{6.1}$$

where s_{ij} denotes the reference signed distance of a coordinate x_{ij} from shape i. To approximate the SDF a multi-layer perceptron f_ϕ with parameters ϕ is used. Using an SDF, the surface of a shape i is implicitly represented by the points on the decision boundary. The model can represent multiple shapes by conditioning the network output on a latent shape vector z_i. Given a spatial point x_{ij}, and a latent vector $z_i \in \mathbb{R}^m$, f_ϕ predicts the signed distance of the point to the surface:

$$f_\phi(z_i, x_{ij}) \approx \mathrm{SDF}^i(x_{ij})\,. \tag{6.2}$$

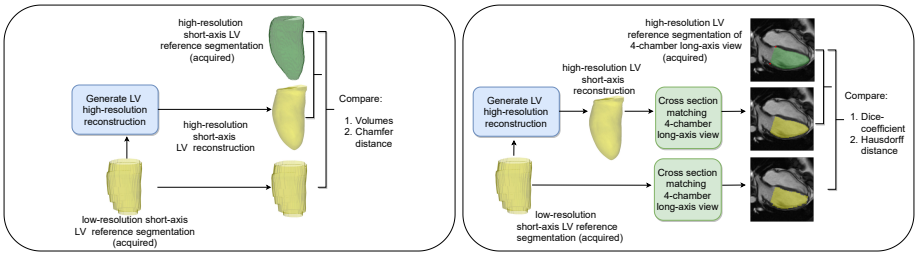
During training for each shape i a latent vector is randomly initialized $z_i \sim \mathcal{N}(0, \sigma_z)$ and optimized together with the model parameters using stochastic gradient descent. By contrast with[23] where an L1 loss was used during training, we found that a mean squared error between predicted and reference signed distance resulted in superior reconstruction performance:

$$\arg\min_{\phi, \{z_i\}_{i=1}^N} \frac{1}{N} \sum_{i=1}^{N} \frac{1}{K} \sum_{j=1}^{K} \left(f_\phi(z_i, x_{ij}) - \tanh(s_{ij}) \right)^2 + \lambda |z_i|_2^2\,, \tag{6.3}$$

where s_{ij} denotes the reference distance of a coordinate x_{ij} from shape i with corresponding latent shape vector z_i for a dataset of N shapes. Furthermore, K denotes the number of coordinates and their signed distance values sampled from the set $\mathcal{O}_i$ (see Equation 6.1). λ is a hyper-parameter weighting the contribution of the regularization loss (second term in Equation 6.3). Furthermore, tanh(.) denotes the hyperbolic tangent function. Previously, Park *et al.*[23] introduced a hyper-parameter (δ) to administer the distance from the surface over which the model should learn an SDF metric, i.e., to control the domain of the function f_ϕ. Hence, hyper-parameter δ limits the values of

the predicted and reference signed distances to an interval between $[-\delta, \delta]$. In the here proposed approach, the hyper-parameter δ was omitted, and instead, by applying a hyperbolic tangent function to the reference signed distance (see Equation 6.3), the network was implicitly forced to an output interval of $[-1, 1]$. Owning to the steepness of the hyperbolic tangent function around zero, the model is encouraged to focus on the ventricle's surface structure.

All latent shape vectors have 128 dimensions. Furthermore, the multi-layer perceptron consists of eight fully connected hidden layers. To speed up model training, weight-normalization was used, and the fully connected layers were implemented as one-dimensional convolutional layers with a kernel size of one. The first seven layers consist each of 512 kernels, followed by a ReLU nonlinearity. The final layer i.e., the output layer has one kernel representing the distance of a spatial point to the surface boundary. Unlike in the original work,[23] we found that adding a skip connection between the input and the fourth layer, and enabling dropout during training, did not increase model performance.



(a) Evaluation using paired high- and low-resolution segmentation volumes

(b) Evaluation using paired high- and low-resolution segmentation slices

Figure 6.4: Visualization of experimental evaluation of proposed approach on two publicly available cardiac MRI datasets (see section 6.2). (a) Evaluation using paired high- and low-resolution CMRI left ventricle (LV) volumes from the UKDHP dataset (see section 6.5.2). (b) Evaluation using paired 2D high- and low-resolution 4-chamber long-axis CMRI LV segmentations from the M&Ms-2 dataset. A low-resolution slice was obtained by taking a cross section from the 3D low-resolution short-axis LV CMRI segmentation. Orientation of the *sliced* low-resolution short-axis volume matches the 4-chamber long-axis view (as acquired). Best viewed in color.

6.3.2 High-resolution shape reconstruction

After training, to reconstruct a high-resolution LV shape, first, the nearest latent shape representation z_i needs to be determined for a shape i (hereafter referred to as optimal latent shape vector). For this, the model parameters ϕ are frozen, and the randomly initialized latent vector z_i is optimized using gradient descent, minimizing the following

loss as described by Park *et al.*:[23]

$$\arg\min_{z_i} \frac{1}{K} \sum_{j=1}^{K} \left| f_\phi(x_{ij}, z_i) - s_{ij} \right| + \lambda |z_i|_2^2 , \tag{6.4}$$

where K denotes the number of coordinates (x_{ij}) and their signed distance values (s_{ij}) sampled from the set $\mathcal{O}_i$ of SDF observations (see Equation 6.1). λ is a hyper-parameter weighting the contribution of the regularization loss (second term in Equation 6.4). We found empirically that during inference the L1 loss (see Equation 6.4) was superior over the proposed L2 training loss (see Equation. 6.3).

Subsequently, using the optimized latent shape vector z_i, a discretized voxel-based signed distance volume $\mathbf{V}_i$ with an unconstrained resolution is obtained by querying the network at continuous spatial locations. The high-resolution reconstructed surface is implicitly represented by the coordinates on the decision boundary, i.e., the zero iso-surface $\{x_{ij} \in \mathbb{R}^3 | f_\phi(z_i, x_{ij}) = 0\}$, where x_{ij} denotes a coordinate of shape i.

In this work, LV shapes are reconstructed using a volume $\mathbf{V}_i$ of size 200×200×200 (width, height, depth), i.e., for each shape the model is queried using 8×10^6 sample coordinates. Voxel-grid coordinates of $\mathbf{V}_i$ are transformed into coordinates of the reference frame by defining the spatial sampling bounds $\mathbf{B}_{\mathbf{V}_i}$ of the voxel-grid in the reference frame. In other words, $\mathbf{B}_{\mathbf{V}_i}$, a matrix of size 3×2, defines the space in which the high-resolution shape can be reconstructed. Using a fixed volume size for all reconstructed shapes, the spatial resolution of a shape $\mathbf{V}_i$ is determined by its spatial sampling bounds $\mathbf{B}_{\mathbf{V}_i}$. After obtaining a signed distance volume, a binary volume of the LV can be obtained by thresholding the signed distance volume.

$$\mathbb{1}(x_{ij}, z_i) = \begin{cases} 1, & f_\phi(x_{ij}, z_i) <= \theta \\ 0, & f_\phi(x_{ij}, z_i) \quad > \theta \end{cases} \tag{6.5}$$

For this, θ is set to zero. In this work, a voxelized binary volume of the LV is only generated for quantitative evaluation of the reconstructed shape. Additionally, a mesh is extracted from the same signed distance volume using the Lewiner Marching Cubes algorithm.[31]

6.3.3 Reference coordinate system and canonical orientation of cardiac shapes

To encourage the model to extract spatial regularities from the shapes' coordinates, all shapes need to be aligned to the *canonical orientation* of a shared 3D Cartesian reference coordinate system.[32,33] For this purpose, the center of mass of a shape is associated with the origin of the reference coordinate system (object-centered coordinate system). To accomplish alignment of cardiac shapes with the canonical orientation of the reference coordinate system, a common shape-specific coordinate frame needs to be defined.

Subsequently, for each shape, the shape-specific coordinate frame is rotated onto the reference coordinate frame. The chosen canonical orientation in this work matches the cardiac short-axis view of the heart. The axial plane of the short-axis view is perpendicular to the long-axis of the heart, which is considered the axis that aligns the heart's base and apex.

The reference coordinate system is identified by the unit vectors $\hat{x}$, $\hat{y}$, $\hat{z}$. It is assumed that the z-axis ($\hat{z}$) of the reference coordinate system is colinear with the long-axis of the heart. Furthermore, the x-axis ($\hat{x}$) is colinear to the vector pointing from the center of mass of the LV to the right ventricle. Finally, the y-axis ($\hat{y}$) of the reference framework is equal to the cross-product of $\hat{x}$ and $\hat{z}$. An example of the canonical orientation for a segmentation of the LV and right ventricle is depicted in Figure 6.3.

6.4 Evaluation

To quantitatively assess our method's ability to reconstruct high-resolution LV shapes from sparse-view short-axis CMRI LV reference segmentations, paired 3D high- and low-resolution short-axis LV reference shapes from the UKDHP dataset were used (section 6.2.2). To evaluate whether high-resolution LV reconstructions approximate high-resolution LV reference volumes better than the corresponding low-resolution LV reference volumes, overlap (3D Dice similarity coefficient) and boundary distances (3D Hausdorff distance, 95^{th} percentile Hausdorff distance, Average symmetric surface distance) were computed between (i) paired 3D high- and low-resolution short-axis LV reference shapes, and (ii) high-resolution reference and reconstructed LV shapes using sparse-view short-axis CMRI LV reference segmentations. Furthermore, to assess performance in terms of clinical metrics and to enable indirect performance comparison with high-resolution reconstruction approach by Beetz *et al.*[21] (section 6.7), the LV end-diastolic volume (in mL) was computed on a population level for high- and low-resolution references and reconstructed high-resolution LV shapes. Figure 6.4a visualizes the aforementioned evaluation.

In addition, to evaluate the method's potential to complete missing shape information, overlap (2D Dice similarity coefficient) and boundary distances (2D Hausdorff distance, 95^{th} percentile Hausdorff distance, Average symmetric surface distance) were assessed between reference and reconstructed LV shapes using paired 2D high-resolution 4-chamber long-axis and 3D low-resolution short-axis CMRI LV segmentations from the M&Ms-2 dataset (see section 6.2.3). While sparse-view 3D short-axis CMRI LV shapes often lack whole heart coverage, 2D high-resolution 4-chamber long-axis views do cover the complete LV anatomy from apex to base. The evaluation approach was previously described by Wang *et al.*[20] and is visualized in Figure 6.4b.

Furthermore, statistical significance of performance differences between (i) paired high- and low-resolution reference segmentations, and (ii) high-resolution reference and reconstructed shapes were tested using the one-sided Wilcoxon signed-rank test.

Table 6.1: Quantitative evaluation of high-resolution reconstruction performance using high-resolution left ventricle (LV) shapes from CCTA scans of 77 test patients. Comparing high-resolution LV reference meshes (reference) obtained from LV segmentations in CCTA (section 6.2.1) with reconstructed high-resolution LV meshes (reconstructed) in terms of (a) Hausdorff distance (HD), 95th percentile Hausdorff distance (HD95) and Average symmetric surface distance (ASSD) and (b) LV end-diastolic volume (LV_{EDV}) (mean±standard deviation).

Surface distances (mm)		
HD↓	**HD95↓**	**ASSD↓**
5.80±3.20	2.41±0.65	1.04±0.33

(a)

LV_{EDV} (mL)	
reference	**reconstructed**
121±43	120±42

(b)

Finally, performance of the method was qualitatively evaluated by visually inspecting the reconstructed surfaces or voxel-based segmentations. Visual assessment focused on: smoothness of the reconstructed surface, anatomical shape completion, and correction of motion-induced inter-slice misalignment.

6.5 Experiments and Results

A latent space and deep implicit neural function (multi-layer perceptron) for the reconstruction of high-resolution LV shapes were trained and evaluated on high-resolution CCTA segmentations of the LV (section 6.5.1). Subsequently, to demonstrate the ability of the approach to reconstruct and complete high-resolution LV shapes from sparse-view incomplete short-axis CMRI segmentations, the approach was evaluated on (i) 500 segmentations of the LV in paired 3D high- and low-resolution short-axis CMRI from the UKDHP dataset (section 6.5.2), and (ii) 360 LV segmentations in paired 3D low-resolution short-axis and 2D high-resolution 4-chamber long-axis CMRI from the M&Ms-2 dataset (section 6.5.3). Figures 6.4a and 6.4b visualize the aforementioned experimental evaluations.

6.5.1 Learning a LV shape prior

To learn a LV shape prior for the representation of plausible high-resolution LV shapes, our approach was trained on high-resolution LV meshes obtained from 3D cardiac CTA LV reference segmentations (section 6.2.1).

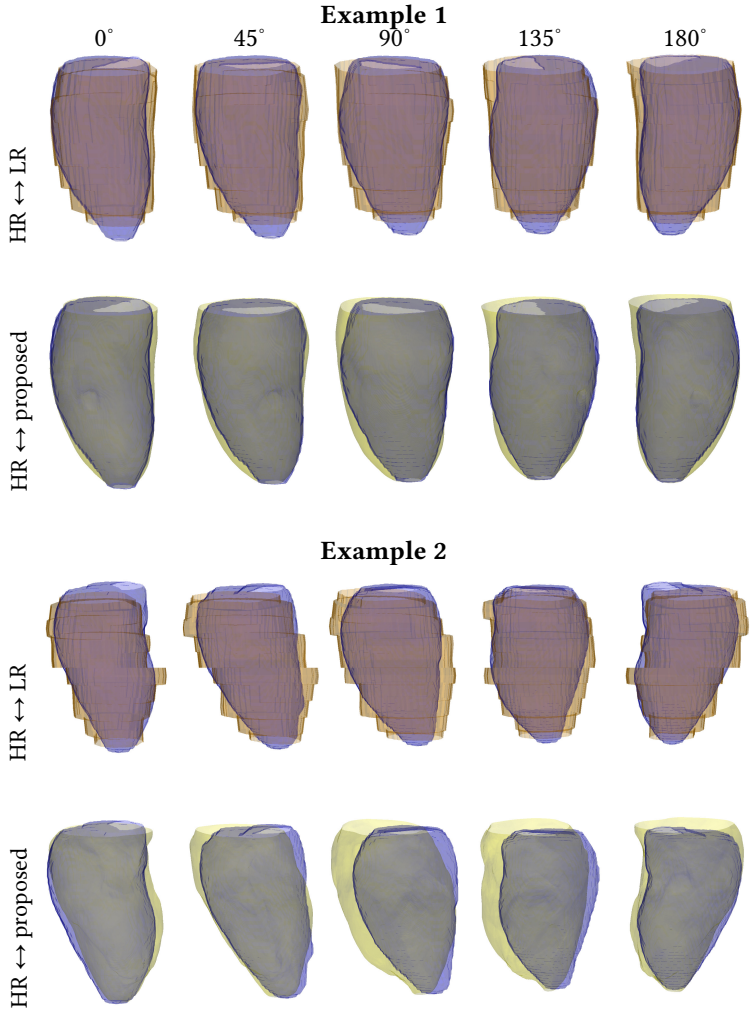


Figure 6.5: Two examples of qualitative evaluation using paired high (HR) and low-resolution (LR) CMRI short-axis left ventricle (LV) segmentations from UKDHP dataset (see section 6.2.2). Example 1, first row shows comparison between reference high (blue) and low-resolution (orange) LV reference segmentations (HR ↔ LR). Second row depicts comparison between LV reference high-resolution segmentation (blue) and high-resolution reconstruction (yellow) (HR ↔ proposed). Example 2 illustrates the same comparisons in row three (HR ↔ LR) and four (HR ↔ proposed), respectively. Furthermore, example 2 (third row, all columns) depicts a low-resolution short-axis shape (orange) with severe motion-induced slice misalignment.

EXPERIMENTAL DETAILS Our approach requires that all LV shapes are represented in a shared reference coordinate system (see section 6.3.3). Therefore, in this experiment, LV segmentations in CCTA were first linearly transformed to align with the canonical orientation of the reference coordinate system. To accomplish this, for each volume, a shape-specific coordinate frame was defined (an example is depicted in Figure 6.3c). To define the shape-specific coordinate frame, reference segmentations of the LV, right ventricular and left atrium were required. Computing the Euler angles between the reference coordinate frame and the shape-specific coordinate frame allows for rotation of the LV segmentation into the canonical orientation of the reference coordinate system. Subsequently, rotated LV segmentations were aligned with the origin of the reference coordinate system (using the center of mass of the LV). To obtain a compressed representation of LV shapes that facilitates fast training, shapes were converted to triangular meshes using the Lewiner Marching Cubes[1] algorithm.[31] Moreover, meshes were smoothed and reduced to contain 10,000 points each. Meshes of 434 patients were randomly split into training (341/434 patients, 78%), validation (16/434 patients, 4%) and test (77/434 patients, 18%) sets. We deliberately chose a small validation set to retain a large training set.

The model was trained using mini-batch stochastic gradient descent and a batch size of eight. Meshes of the training set were provided once per epoch to the model in random order. For each mesh, 1,024 coordinates were randomly sampled from the surface and another 1,024 were uniformly sampled within the boundaries of the mesh. For this, mesh boundaries were extended in each direction by 10 mm. Reference signed distance values for the sampled coordinates were obtained by computing the 3D Euclidean distance of each sampled coordinate to the mesh surface[2]. The learning rate was set to 0.001 and decayed over 5,000 iterations using a cosine annealing learning rate schedule. Network parameters (ϕ) and latent shape vectors were optimized jointly using the Adam optimizer,[34] minimizing the loss specified in Equation 6.3. Furthermore, following Park *et al.*[23] λ in Equation 6.3 was set to 0.0001, and latent shape vectors were initialized using a univariate normal with σ equal to 0.01.

The model was trained for 2,000 epochs and validated every 100^{th} epoch. To reconstruct a validation shape, a latent shape vector was optimized for 1,000 iterations using the Adam optimizer with a learning rate of 0.001, minimizing the loss specified in Equation 6.4. Moreover, validation loss was computed for each shape by randomly sampling 4,096 pairs of coordinates and signed distances from the set of SDF observations (as defined in Equation 6.1). The model with the lowest validation loss was used for performance evaluation of high-resolution reconstruction using anisotropic CMRI LV segmentations. The method was implemented using the PyTorch framework[35] and trained on one Nvidia RTX 2080 GPU with 11 GB memory. The model was trained in approximately 6 hours.

[1] Lewiner marching cubes implementation of *scikit-image*

[2] using pyvista's `DataSetFilters.compute_implicit_distance` function

RESULTS Quantitative test results in terms of surface distances and LV end-diastolic volume listed in Table 6.1 demonstrate that the proposed approach can accurately reconstruct high-resolution LV shapes from high-resolution LV reference segmentations in CCTA images. The average surface distance is within two (average) voxel spacings.

Table 6.2: Quantitative evaluation using left ventricle (LV) segmentations in paired high- and low-resolution CMRI from the UKDHP dataset. Comparing high-resolution (HR) CMRI LV reference segmentations with (i) low-resolution (LR) LV reference segmentations (LR LV), and (ii) high-resolution LV reconstructions (proposed). The evaluation is performed using (a) Dice similarity coefficient (DSC), Hausdorff distance (HD), 95th percentile Hausdorff distance (HD95) and Average symmetric surface distance (ASSD) and (b) LV end-diastolic volume (LV_{EDV}) (mean±standard deviation). Differences between (i) and (ii) are statistically significant ($p < 0.001$) using the one-sided Wilcoxon signed-rank test. Best performance is indicated in bold.

	DSC↑	**HD↓** (mm)	**HD95↓** (mm)	**ASSD↓** (mm)
LR LV ↔ HR LV	0.87 ±0.03	10.72 ±2.71	5.77 ±1.54	2.28 ±0.59
proposed ↔ HR LV	***0.92** ±0.02	***8.29** ±3.16	***3.90** ±0.88	***1.61** ±0.35

(a)

LV_{EDV} (mL)		
reference LR LV	reference HR LV	proposed
133±12	146±36	150±37

(b)

6.5.2 High-resolution shape reconstruction using segmentations from the UKDHP dataset

To assess model performance, paired LV segmentations in 3D high- and low-resolution short-axis CMRI at end-diastole were taken from the first 500 patients of the UKDHP dataset (section 6.2.2). Using the trained model as described in section 6.5.1, high-resolution LV shapes were reconstructed from low-resolution LV shapes. High-resolution LV reconstructions were then quantitatively and qualitatively compared with reference high-resolution LV shapes. Figure 6.4a visualizes the comparison.

EXPERIMENTAL DETAILS The approach requires that reference high- and low-resolution LV segmentations are aligned to the canonical orientation of the reference coordinate system as described in section 6.3.3. To accomplish this, for each volume, a shape-specific coordinate frame was defined as described in section 6.5.1. Because the short-axis acquisition plane in CMRI is orthogonal to the z-axis of the reference coordinate

Table 6.3: Quantitative comparison between high-resolution 2D CMRI 4-chamber long-axis (LAX) left ventricle (LV) reference segmentations (HR LAX) with cross sections taken from (i) low-resolution (LR) 3D CMRI short-axis (SAX) LV reference segmentations (LR-SAX), and (ii) high-resolution (HR) 3D SAX LV reconstructions (proposed). Orientation of cross sections matches orientation of 4-chamber LAX view. The evaluation is performed using Dice similarity coefficient (DSC), Hausdorff distance (HD), 95th percentile Hausdorff distance (HD95) and Average symmetric surface distance (ASSD) (mean±standard deviation). Differences between (i) and (ii) are statistically significant ($p < 0.001$) using the one-sided Wilcoxon signed-rank test. Best performance is indicated in bold.

	DSC↑ ±σ	**HD↓** ±σ	**HD95↓** ±σ	**ASSD↓** ±σ
LR SAX ↔ HR LAX	0.89 ±0.05	10.56 ±4.08	7.67 ±3.40	3.08 ±1.48
proposed ↔ HR LAX	***0.91** ±0.05	***8.16** ±3.86	***6.51** ±3.38	***2.79** ±1.57

system (see Figure 6.3c), alignment of short-axis CMRI segmentations with the canonical orientation of the reference coordinate frame only required rotation in the xy-plane. Furthermore, rotation of a CMRI shape was performed as described in section 6.5.1. Finally, to align CMRI segmentations with the origin of the reference coordinate system, voxel coordinates of each volume were centered using the center of mass of the LV.

Using the trained model (see section 6.5.1), high-resolution LV reconstructions were obtained from low-resolution CMRI references segmentations following the steps described in section 6.3.2. Latent shape optimization requires pairs of coordinates and their corresponding signed distance values (Equation 6.4). To obtain the latter, for all segmentation voxels, the 3D Euclidean distance transform was computed [3] taking the volume's voxel spacing into account.

Furthermore, to accelerate inference, for each shape, latent optimization was performed with a filtered subset $\mathcal{S}_i$ of SDF observations $\mathcal{O}_i$ as defined in Equation 6.1:

$$\mathcal{S}_i = \left\{(x_{ij}, s_{ij}) \in \mathcal{O}_i \mid |s_{ij}| <= \gamma \sqrt{(v_i^x)^2 + (v_i^y)^2}\right\}, \tag{6.6}$$

where s_{ij} denotes the signed distance of a coordinate x_{ij} from shape i. Furthermore, $\gamma \in \mathbb{N}^+$ and v_i^x and v_i^y denote the in-plane resolution (x and y-direction, respectively) of the original low-resolution voxel-based shape representation. $\mathcal{S}_i$ contains all on-surface points and, depending on γ and the segmentation's in-plane resolution, an additional number of (off-surface) coordinates. Hereafter, we refer to $\mathcal{S}_i$ as the filtered set of SDF observations. To determine the optimal value for γ a line search was performed for $\gamma \in \{1, 2, 5, 10, 15\}$. Trading off between computation time and model performance (in terms of overlap measure and surface distance), γ was set to 2 in all experiments.

Furthermore, using the Adam optimizer with a learning rate of 0.001, optimization

[3] using `scipy.ndimage.distance_transform_edt` function

of a latent shape vector was performed for 1,000 iterations minimizing the loss as specified in Equation 6.4. λ in Equation 6.4 was set to 0.0001. Subsequently, high-resolution binary volumes and meshes were reconstructed as described in section 6.3.2. For this, spatial sampling bounds $\mathbf{B}_{\mathbf{V}_i}$ of the high-resolution signed distance volume $\mathbf{V}_i$ (see section 6.3.2) were set to the spatial bounds of the high-resolution reference shape.

RESULTS To perform latent shape optimization, the filtered set of SDF observations $\mathcal{S}_i$ as defined in Equation 6.6 contained on average 7,150 (σ=1,100) pairs of coordinates and signed distance values. Quantitative evaluation shown in Table 6.2b demonstrates that high-resolution LV reconstructions can approximate LV end-diastolic volume more accurately compared with low-resolution reference volumes. Moreover, quantitative results listed in Table 6.2a show that high-resolution LV reconstructions have a substantially higher overlap in terms of Dice similarity coefficient and lower surface distance to the high-resolution reference surfaces compared with low-resolution reference volumes. These results indicate improved accuracy and completeness of the high-resolution reconstructions compared with the reference low-resolution volumes.

In addition, qualitative evaluation of the approach shown in Figure 6.5 reveals that high-resolution reconstruction of low-resolution CMRI LV segmentations results in plausible and smooth LV surfaces. Furthermore, surfaces of the high-resolution LV reconstructions follow the surfaces of high-resolution reference LV shapes more closely compared with low-resolution reference shapes. Moreover, the results demonstrate that LV shapes with missing apical and basal slices can be completed using the proposed approach. Finally, qualitative comparison depicted in Figure 6.5 also illustrates that reconstruction performance of the approach is affected by low-resolution reference segmentations with severe inter-slice misalignment. Furthermore, one can observe in Figure 6.5 that reconstructed LV shapes do not contain the complete LV outflow tract. This could be surprising because the model was trained on high-resolution CCTA LV reference segmentations that do contain the LV outflow tract (see Figure 6.2). However, the spatial sampling bounds (see section 6.3.2) of the high-resolution reconstruction were set to the spatial bounds of the high-resolution CMRI LV reference shape. CMRI LV shapes typically only include a small part of the LV outflow tract and hence, spatial bounds between apex and base are smaller compared with LV shapes in CCTA.

6.5.3 High-resolution shape reconstruction using segmentations from the M&Ms-2 dataset

To demonstrate that our proposed approach can generalize to unseen LV shapes of patients suffering from a variety of cardiomyopathies, the approach is evaluated using paired 2D high-resolution 4-chamber long-axis and 3D low-resolution short-axis CMRI LV segmentations from 360 patients of the M&Ms-2 dataset (section 6.2.3). In clinical practice, 2D 4-chamber long-axis CMR images (examples shown in first column of

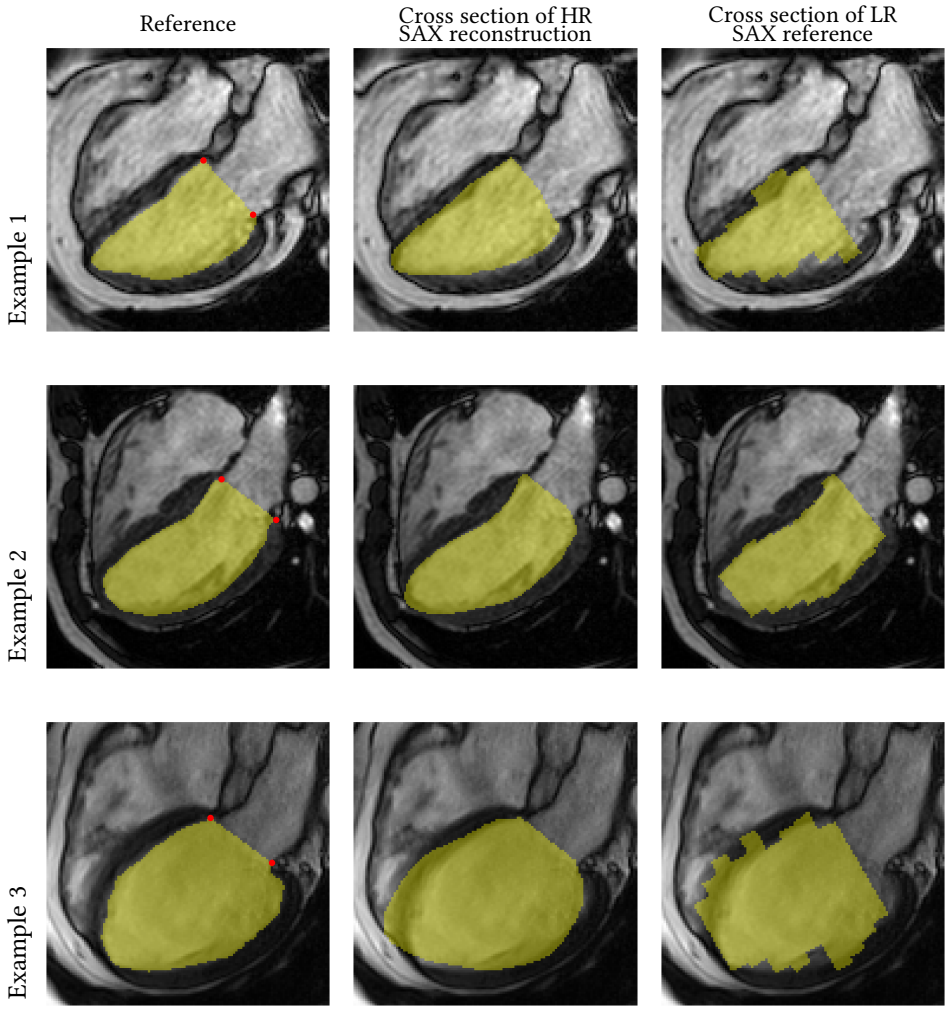


Figure 6.6: Qualitative comparison between high-resolution (HR) left ventricle (LV) reference segmentation of 4-chamber long-axis view (LAX, first column) with (second column) cross section taken from 3D high-resolution short-axis (SAX) reconstruction (proposed) and (third column) cross section taken from 3D low-resolution (LR) SAX LV reference segmentation. Orientation of cross section matches orientation of 4-chamber LAX view. Each row shows a different patient. Red dots in images of first column indicate manually drawn landmarks for mitral valve plane. Cross sections (second and third columns) were clipped if exceeding the mitral valve plane. Best viewed in color.

Figure 6.6) are routinely acquired together with short-axis CMRIs. To allow evaluation, cross sections were obtained from 3D low-resolution short-axis LV reference segmentations and 3D high-resolution LV reconstructions using image metadata provided in the headers of short- and long-axis images. Orientation of cross sections matches the 4-chamber long-axis acquisition plane. Example images are shown in second and third column of Figure 6.6. Cross sections were then compared with 2D high-resolution 4-chamber long-axis LV reference segmentations (see Figure 6.4b).

EXPERIMENTAL DETAILS To generate 3D high-resolution short-axis LV reconstructions, first, low-resolution short-axis LV segmentations were aligned to the reference coordinate system using the steps described in section 6.5.2. Second, for each low-resolution short-axis LV shape, an optimal latent shape vector was determined following the steps described in section 6.5.2. Third, because the approach was trained using coordinates in the reference coordinate system, coordinates of the 2D high-resolution long-axis LV reference segmentations were transformed to the reference coordinate system. Finally, using the optimal latent shape vector and the transformed 4-chamber long-axis coordinates, the trained model (section 6.5.1) can be queried to obtain a cross section from the 3D high-resolution short-axis LV reconstruction matching the orientation of the long-axis view.

RESULTS Quantitative results presented in Table 6.3 convey that cross sections from 3D high-resolution short-axis LV reconstructions show improved overlap and reduced surface distance with 2D high-resolution long-axis LV reference segmentations compared with cross sections from 3D short-axis low-resolution LV reference segmentations. Differences between the high-resolution LV reconstructions (proposed) and low-resolution short-axis LV reference segmentations in terms of Dice similarity coefficient, Hausdorff distance, 95^{th} Hausdorff distance and Average symmetric surface distance are statistically significant ($p < 0.001$) determined by the one-sided Wilcoxon signed-rank test. Qualitative results depicted in Figure 6.6 corroborate this finding. Surfaces that were reconstructed with proposed approach appear smoother than cross sections taken from 3D short-axis LV reference segmentations. Furthermore, Figure 6.6 illustrates the potential of the approach to complete missing shape information especially at the apex and base of the heart. Finally, one can observe that the proposed approach can correct for limited inter-slice misalignment.

6.6 Additional experiments

To obtain the optimal latent shape vector for a low-resolution CMRI LV segmentation i, in previous experiments (sections 6.5.2 and 6.5.3), the filtered set of SDF observations $\mathcal{S}_{\text{i}}$ was used (as specified in Equation 6.6). To scrutinize the shape completion capabilities of the method, additional experiments were performed (section 6.6.1) using sparse

Table 6.4: Quantitative comparison between high-resolution (HR) 3D CMRI short-axis (SAX) reference left ventricle (LV) shapes and high-resolution LV reconstructions using corresponding low-resolution CMRI SAX LV reference segmentations. High-resolution reconstruction was performed using (i) SDF observations taken from low-resolution reference SAX volumes (# of LR SAX SDF observ.), and (ii) a combination of (i) and SDF observations taken from two cross sections of 3D high-resolution LV reference shapes (HR-cross). The evaluation is performed using Dice similarity coefficient (DSC), Hausdorff distance (HD), 95th percentile Hausdorff distance (HD95), Average symmetric surface distance (ASSD) and LV end-diastolic volume (LV_{EDV}). Best performance is indicated in (i) blue bold using only LR reference coordinates, and (ii) black bold using a combination of LR reference coordinates and coordinates of HR cross sections. Using *All* SDF observations resulted on average in 7,150 (σ=1,100) pairs of coordinates and signed distance values. Best viewed in color.

# of LR SAX SDF observ.	HR-cross	**Dice**↑	**HD**↓ (mm)	**HD95**↓ (mm)	**ASSD**↓ (mm)	**LV_{EDV}** (mL)
None	☒	**0.92** ±0.02	**7.88** ±2.78	**3.81** ±0.78	**1.58** ±0.33	**148** ±36
100		0.89 ±0.03	11.70 ±4.39	5.43 ±1.92	2.10 ±0.64	135 ±31
100	☒	0.89 ±0.01	8.60 ±2.10	4.79 ±0.68	2.00 ±0.26	129 ±33
1000		0.91 ±0.02	8.43 ±2.63	3.97 ±0.89	1.63 ±0.35	**147** ±36
1000	☒	0.91 ±0.01	8.42 ±2.34	4.41 ±0.62	1.78 ±0.27	135 ±33
2000		0.91 ±0.02	8.31 ±2.70	3.93 ±0.90	1.62 ±0.36	149 ±36
2000	☒	0.91 ±0.02	8.43 ±2.62	4.21 ±0.67	1.69 ±0.28	138 ±33
3000		**0.92** ±0.02	**8.00** ±2.36	**3.86** ±0.85	**1.60** ±0.34	149 ±36
3000	☒	0.91 ±0.02	8.53 ±2.99	4.07 ±0.72	1.65 ±0.30	142 ±34
4000		0.91 ±0.02	8.25 ±2.83	3.92 ±0.90	1.62 ±0.36	149 ±36
4000	☒	**0.92** ±0.02	8.28 ±3.06	3.95 ±0.83	1.61 ±0.32	144 ±34
All		**0.92** ±0.02	8.29 ±3.16	3.90 ±0.88	1.61 ±0.35	150 ±37
All	☒	**0.92** ±0.02	8.03 ±3.05	3.82 ±0.81	**1.58** ±0.33	**148** ±36
					LV_{EDV} reference	**146** (±36)

subsets of $\mathcal{S}_{\mathrm{i}}$. Furthermore, experiments described in section 6.6.2 illustrate how SDF observations from different cardiac views can be seamlessly integrated. Moreover, the latter experiments investigate the effect of such an integration on high-resolution reconstruction performance. Other experimental settings in these experiments, including the evaluation approach (see Figure 6.4 and section 6.4), were identical to experiments described in previous sections.

6.6.1 Effect of number of SDF observations on reconstruction performance

To investigate the shape completion capabilities of the approach, for each low-resolution LV segmentation in CMRI, latent shape optimization used different numbers of SDF observations. For this, random subsets of observations of size K as defined in Equation 6.4 ($\mathrm{K} \in [100, 4000]$) were taken from the set of filtered SDF observations ($\mathcal{S}_{\mathrm{i}}$ as defined in Equation 6.6).

Table 6.4 lists results for quantitative comparison between high-resolution reference and high-resolution reconstructed LV shapes. Reconstruction was performed using corresponding low-resolution LV reference shapes (see Figure 6.4a). Furthermore, complementary results listed in Table 6.5 show quantitative comparison between 2D high-resolution LV reference segmentations in 4-chamber long-axis CMRI with cross sections taken from 3D high-resolution short-axis LV reconstructions. Orientation of the short-axis cross section matches the 4-chamber long-axis acquisition plane. Results of both evaluations convey that reconstruction performance of the approach improved when more SDF observations were sampled from the low-resolution shapes, i.e., when K as defined in Equation 6.4 increased. Furthermore, one can notice that performance, in terms of overlap and surface distance converges when 3,000 SDF observations were used for the latent shape optimization. Qualitative results depicted in Figure 6.7 corroborate this finding. This demonstrates the strong LV shape prior that is encoded in the multi-layer perceptron parameters by training the model on high-resolution CCTA LV segmentations.

6.6.2 Effect of integrating observations from different cardiac views on reconstruction performance

Additional experiments were performed to demonstrate the potential of the method to improve reconstruction performance by seamlessly integrating observations from different cardiac views. Experimental settings were identical to settings described in previous section 6.6.1, except that, for each shape, latent shape optimization was performed with additional SDF observations from different cardiac views.

To perform high-resolution reconstruction experiments using paired high- and low-resolution short-axis LV reference segmentations in CMRI (Figure 6.4a), SDF observations taken from low-resolution short-axis LV segmentations in CMRI were

combined with SDF observations taken from two orthogonal cross sections of 3D high-resolution short-axis LV reference segmentations (example depicted in Figure 6.8a). To find the nearest latent shape representation, for each low-resolution LV shape, the combined set of SDF observations is used. Subsequently, high-resolution LV shapes were reconstructed using the steps described in sections 6.3.2 and 6.5.2).

Quantitative comparison between high-resolution reference and reconstructed LV shapes listed in Table 6.4 shows that reconstruction performance is superior when using only SDF observations of the high-resolution cross sections. The latter contain on average 450 SDF observations. The result illustrates the positive effect of SDF observations in through-plane direction on high-resolution reconstruction performance. Furthermore, one can observe that performance drops substantially if only a 100 SDF observations from the low-resolution short-axis view were included. Nevertheless, reconstruction performance of the approach recovers when an increasing number of SDF observations from the low-resolution short-axis LV segmentations were included.

A similar experiment was performed for high-resolution reconstruction using paired high- and low-resolution 4-chamber long-axis CMRI segmentation slices (Figure 6.4b). For this, SDF observations taken from low-resolution short-axis LV segmentations in CMRI were combined with SDF observations taken from 2D high-resolution 4-chamber long-axis LV reference segmentations (example shown in Figure 6.8b). Since coordinates of all LV segmentations in CMRI were already aligned with the reference coordinate system (see section 6.3.2), SDF observations of the different cardiac views could be combined without additional effort. Using the combined set of SDF observations high-resolution LV shapes were reconstructed using the steps described in sections 6.3.2 and 6.5.2. To allow evaluation, cross sections were taken from 3D high-resolution LV reconstructions matching the 4-chamber long-axis acquisition plane (see Figure 6.4b).

Quantitative comparison between 2D high-resolution 4-chamber long-axis LV reference segmentations and cross sections obtained from 3D high-resolution short-axis LV reconstructions listed in Table 6.5 are in line with previous results, i.e., reconstruction performance is superior when using only SDF observations of the high-resolution 4-chamber long-axis LV reference segmentations (on average 360 SDF observations). Furthermore, performance decreases as more SDF observations of the low-resolution short-axis view were included. Results listed in Tables 6.4 and 6.5 also reveal that for the same number of low-resolution short-axis LV observations performance increases if latent optimization included additional SDF observations taken from different cardiac views. However, the effect softens with increased number of SDF observations taken from the low-resolution short-axis LV segmentation.

6.7 Comparison with other work

Closest to our work are previously developed high-resolution reconstruction methods for anisotropic CMRI bi-ventricular segmentations by Wang *et al.*[20] and Beetz *et al.*[19,21]

Table 6.5: Evaluation of shape completion performance in terms of Dice similarity coefficient (DSC), Hausdorff distance (HD), 95th percentile HD (HD95) and Average symmetric surface distance (ASSD) (mean±standard deviation) using paired 2D high-resolution (HR) 4-chamber long-axis (LAX) left ventricle (LV) reference segmentations and cross sections taken from 3D high-resolution short-axis (SAX) LV reconstructions (proposed). High-resolution reconstruction of SAX LV shape is performed using (i) SDF observations taken from low-resolution (LR) SAX LV reference segmentations (# of LR SAX SDF observ.), and (ii) a combination of (i) and SDF observations from HR LAX LV reference segmentations in CMRI (include HR LAX observ.). Using *All* SDF observations resulted on average in 6,900 (σ=2,200) pairs of coordinates and signed distance values. Best performance is indicated in blue bold (LR reference coordinates only) and black bold. Best viewed in color.

# of LR SAX SDF observ.	include HR LAX observ.	**DSC↑**	**HD↓ (mm)**	**HD95↓ (mm)**	**ASSD↓ (mm)**
None	☒	**0.96** ±0.03	**5.37** ±5.83	**3.41** ±3.99	**1.16** ±0.69
100		0.90 ±0.05	9.88 ±4.20	7.84 ±3.49	3.25 ±1.55
100	☒	**0.96** ±0.04	5.71 ±5.31	3.93 ±3.76	1.35 ±0.96
1000		**0.91** ±0.05	8.42 ±4.16	6.65 ±3.48	2.83 ±1.59
1000	☒	0.94 ±0.05	7.00 ±4.53	5.30 ±3.70	1.99 ±1.52
2000		**0.91** ±0.05	8.09 ±3.78	**6.45** ±3.27	**2.77** ±1.57
2000	☒	0.93 ±0.05	7.47 ±4.42	5.76 ±3.58	2.28 ±1.59
3000		**0.91** ±0.05	**8.08** ±3.74	6.46 ±3.29	**2.77** ±1.58
3000	☒	0.92 ±0.05	7.56 ±4.04	5.98 ±3.50	2.45 ±1.59
4000		**0.91** ±0.05	8.24 ±4.06	6.53 ±3.51	2.78 ±1.57
4000	☒	0.92 ±0.05	7.79 ±4.10	6.10 ±3.44	2.53 ±1.58
All		**0.91** ±0.05	8.16 ±4.04	6.53 ±3.50	2.78 ±1.57
All	☒	0.92 ±0.05	7.99 ±4.28	6.32 ±3.63	2.63 ±1.57

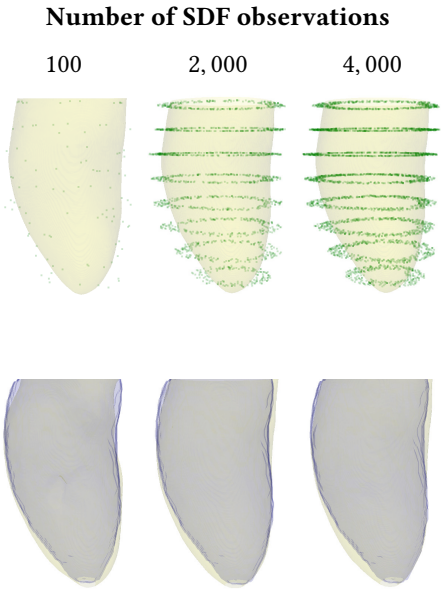


Figure 6.7: Three examples (columns one to three) of high-resolution left ventricle (LV) reconstruction (proposed) using three ($\{100, 2000, 4000\}$) different number of SDF observations taken from the same low-resolution LV reference segmentation (UKDHP dataset, section 6.2.2). First row depicts SDF observations taken from low-resolution LV reference shape (green) and high-resolution LV reconstruction (yellow). Second row shows comparison between reference high-resolution LV shape (blue) and high-resolution LV reconstruction (yellow, same as in first row) using our proposed approach. Best viewed in color.

Table 6.6: Indirect comparison of proposed method with high-resolution reconstruction approach by Wang *et al.*[20] Results were directly taken from Wang *et al.*,[20] Table 3. The evaluation in Wang *et al.*[20] using CMRI segmentations of the UK Biobank dataset was identical to ours on the M&Ms-2 dataset (see Figure 6.4b). Dice similarity coefficient (DSC) specifies overlap between high-resolution 2D CMRI 4-chamber long-axis (LAX) left ventricle (LV) reference segmentation and cross section taken from high-resolution 3D short-axis (SAX) LV reconstruction. In addition, both approaches were evaluated using a combination of short- and long-axis segmentations for the reconstruction task (include LAX column).

Method	Test set	DSC	include LAX
ours	M&Ms-2	0.91±0.05	
		0.96±0.04	☒
Wang *et al.* (2021)[20]	UK Biobank	0.91±0.08	
		0.92±0.05	☒

Table 6.7: Indirect comparison of proposed method with bi-ventricular high-resolution reconstruction approaches by Beetz *et al.*[21] Results were taken from Beetz *et al.*,[21] Table 3. The evaluation in Beetz *et al.*[21] on CMRI segmentations of the UK Biobank was identical to ours on the UKDHP dataset (see Figure 6.4a). Left ventricle (LV) end-diastolic volume (LV_{EDV}) was computed for reference LV shapes (reference) and high-resolution LV reconstructions (recon.). Furthermore, Beetz *et al.*[21] report LV_{EDV} separately for females and males. Values indicate mean±standard deviation.

Method	Test set	Reference LV_{EDV} (mL)	Recon. LV_{EDV} (mL)	Remark
ours	UKDHP	146±36	150±37	
Beetz *et al.* (2022)[21]	UK Biobank	124±21	129±21	Female
		166±32	166±38	Male

(see section 6.1). The approach described in Wang *et al.*[20] was trained on high-resolution CMRI bi-ventricular reference segmentations of the UKDHP dataset. The evaluation procedure described in Wang *et al.*[20] was identical to our experimental evaluation depicted in Figure 6.4b, section 6.5.3. Therefore, high-resolution reconstruction results of our proposed approach listed in Table 6.3 can be indirectly compared with results reported by Wang *et al.*[20] using anisotropic CMRI segmentations of the UK Biobank dataset[36] (original work Table 3). Results of the indirect comparison listed in Table 6.6 reveal that performance of the method proposed in Wang *et al.*[20] is comparable to ours in terms of overlap measure. During testing, both approaches can combine information from multiple cardiac views to find an optimal latent shape vector. Results listed in Table 6.6 show that our approach clearly outperforms the approach by Wang *et al.*[20] when high-resolution CMRI LV long-axis segmentations were incorporated in the reconstruction task. Furthermore, one should note that compared with the M&Ms-2 dataset, subjects in the UK Biobank dataset represent mainly healthy subjects and therefore, one might expect that cardiac shape variability is larger in the M&Ms-2 dataset compared with cardiac shapes of the UK Biobank dataset.

In addition, quantitative results listed in Table 6.7 show an indirect performance comparison between our method and the approach of Beetz *et al.*[21] While the latter approach was trained on synthetic data using the statistical shape model dataset,[22] high-resolution reconstruction performance of the method was assessed using anisotropic CMRI bi-ventricular segmentations of the UK Biobank dataset. These results can be carefully compared with performance assessment of our method using sparse-view CMRI LV shapes from the UKDHP dataset (section 6.5.2 and 6.5.2). Hence, quantitative results in Table 6.7 were taken directly from the work of Beetz *et al.*,[21] Table 3. Because high-resolution CMRI short-axis reference segmentations are not available in the UK Biobank dataset, Beetz *et al.*[21] report high-resolution reconstruction performance only in terms of LV end-diastolic volume. Results in Table 6.7 reveal that high-resolution

reconstruction performance of our approach is on par with the method described in Beetz *et al.*[21]

It is fair to note that approaches of Wang *et al.*[20] and Beetz *et al.*[21] perform high-resolution reconstruction for both ventricles and the LV epi-cardial structure. However, the upsampling factor of the approaches[20,21] depends on the resolution of the training set while our approach performs high-resolution reconstruction using any desired spatial resolution.

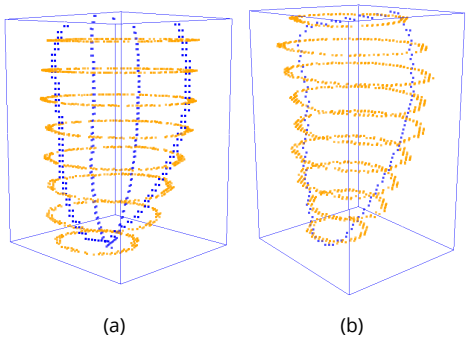


Figure 6.8: Examples of combining coordinates of low-resolution short-axis left ventricle (LV) reference shape (orange) with (a) coordinates of two cross sections of high-resolution short-axis LV reference shape (blue), and (b) surface coordinates of high-resolution 4-chamber long-axis LV reference shape. Best viewed in color.

6.8 Discussion

To mitigate shortcomings of highly anisotropic short-axis LV segmentations in CMRI, we proposed a deep learning-based method for high-resolution reconstruction and completion of LV shapes. Since high-resolution CMR images and segmentations are impracticable or even impossible to acquire, the approach is trained on high-resolution CCTA segmentations of the LV. The presented results demonstrate that the method can exploit properties of high-resolution cardiac segmentations in CCTA to infer missing shape information and improve spatial resolution of typically incomplete low-resolution segmentations in short-axis cardiac MRI.

Evaluation of the proposed approach on two publicly available cardiac MRI datasets revealed that, compared with low-resolution LV volumes, high-resolution LV reconstructions improved in terms of smoothness and anatomical plausibility. Furthermore, the results illustrate that the approach can recover missing shape information and consequently, global cardiac function can be approximated more accurately with reconstructed high-resolution shapes compared with low-resolution LV reference volumes.

Therefore, applying the proposed approach to anisotropic short-axis LV segmentations in CMRI might improve the analysis of cardiac function, morphology and motion. This could be especially beneficial for assessing cardiac motion in longitudinal direction, which is currently hampered by the low through-plane resolution of short-axis CMRI. This direction will be investigated in future work.

Experiments conducted in this work illustrated that our approach can represent multiple (unseen) LV shapes using a single deep implicit function. Therefore, the representation is compact and can reduce the computational cost in simulations of cardiac electrophysiology and mechanics.[3] However, because the auto-decoder approach[23] lacks an explicit shape encoder, high-resolution shape reconstruction requires latent shape optimization during inference. It has been argued that this limits the efficiency and capability of the approach.[37] Nevertheless, we conjecture that, compared with the low-resolution LV reference segmentations, the improved quality of the high-resolution reconstructed LV shapes justifies the additional computational effort. In this study, latent shape optimization took on average seven seconds of GPU processing time (using 1,000 iterations and 2,000 SDF observations).

Furthermore, in contrast to a standard encoder-decoder the chosen auto-decoder approach[23] can handle any form of partial observations such as sparse-view short-axis LV segmentations in CMRI. This is a major advantage over the auto-encoder framework whose encoder expects a test input similar to the training data.[23] However, the auto-decoder approach lacks an explicit encoder and therefore, it does not have local shape features at its disposal. Instead, high-resolution reconstruction is performed by conditioning the model's output on coordinates and a learned global shape embedding. Therefore, to extract spatial regularities from the shapes' coordinates, the approach requires that each shape is represented in a shared reference coordinate system (section 6.3.3). Consequently, reconstruction of a shape might fail if it cannot be correctly aligned with the shared reference coordinate frame. For example, our approach requires that the center of mass of the LV is aligned with the origin of the reference coordinate system. For incomplete or rare anatomical LV shapes this might be infeasible. Future work could investigate whether it is feasible to predict affine transformation parameters that improve shape alignment with the reference coordinate system, e.g., by adopting a deep implicit template[38] or deformed implicit field.[39]

Quantitative results of the additional experiments listed in Tables 6.4 and 6.5 demonstrate that our approach can reconstruct a high-resolution LV shape from a sparse-view input. Determining the optimal latent shape vector with 100 SDF observations, the deep implicit function can instantiate a complete and plausible short-axis LV shape (see first column Figure 6.7). This finding illustrates the strong LV shape prior that is encoded in the multi-layer perceptron. Results presented in Tables 6.4 and 6.5 also show that excellent high-resolution reconstruction performance can be achieved using sparse SDF observations originating (i) from two orthogonal 2D high-resolution short-axis cross sections (Figure 6.8a), and (ii) from the 2D high-resolution 4-chamber long-axis

segmentations (Figure 6.8b). However, performance starts to decrease when including SDF observations taken from the low-resolution short-axis LV shapes. On the one hand such results illustrate that, to achieve high reconstruction performance, samples in through-plane direction are favourable. On the other hand it demonstrates that SDF observations taken from the low-resolution LV representations do not accurately describe the LV in through-plane direction. This is potentially caused by large slice spacing and motion-induced inter-slice misalignment. Nevertheless, the additional experiments exemplify the method's ability to improve reconstruction by seamlessly integrating SDF observations from different cardiac views.

Although the results presented in this work demonstrate the method's potential to correct motion artifacts, Example 2 in Figure 6.5 reveals that severe motion artifacts hamper model performance. In those cases, high-resolution reconstructions still appear smooth in through-plane direction, but might deviate from the high-resolution reference shapes in terms of morphology and volume. Future work could extend the proposed training approach by encouraging explicit correction of simulated inter-slice motion artifacts added to the high-resolution reference CCTA segmentations during training.

Previous work[33,40] has argued that deep implicit functions with fixed-length latent shape vectors have a limited capacity to represent complex shapes, e.g., human shapes. In Figure 6.1 (Examples 1 and 2) one can indeed observe that reconstructed shapes sometimes appear excessively smoothed. As a result, reconstructions can lack fine anatomical details that are present in the LV reference segmentations (first column Figure 6.1). Future work could use local implicit functions and structured latent codes[33,40–43] to potentially improve the performance of implicit representations and high-resolution reconstructions of cardiac shapes. Furthermore, in this work, a deep implicit function was used to represent one cardiac structure. Future work could extend the model to simultaneously predict signed distance values of multiple structures,[44] e.g., LV endo- and epicardial structures. Finally, we surmise that our approach is likely not limited to LV segmentations in CMRI, but can be used for high-resolution reconstruction of other cardiac chambers and time frames of the cardiac cycle e.g., at end-diastole, if high-resolution examples are available during training.

To conclude, our proposed method can reconstruct high-resolution short-axis LV shapes from low-resolution incomplete CMRI segmentations. A single continuous deep implicit function can encode multiple LV shapes and can interpolate and extrapolate LV shapes. Finally, using the method for high-resolution reconstruction of anisotropic CMRI short-axis segmentations has the potential to improve assessment of LV function, morphology and motion.

References

[1] T. Leiner, J. Bogaert, M. G. Friedrich, R. Mohiaddin, V. Muthurangu, S. Myerson, A. J. Powell, S. V. Raman, and D. J. Pennell. "SCMR position paper (2020) on clinical

indications for cardiovascular magnetic resonance," *Journal of Cardiovascular Magnetic Resonance*, vol. 22 (2020), pp. 1–37.

[2] J. N. Cohn, R. Ferrari, N. Sharpe, and an International Forum on Cardiac Remodeling. "Cardiac remodeling—concepts and clinical implications: a consensus paper from an international forum on cardiac remodeling," *Journal of the American College of Cardiology*, vol. 35 (2000), pp. 569–582.

[3] P. Lamata, M. Sinclair, E. Kerfoot, A. Lee, A. Crozier, B. Blazevic, S. Land, A. J. Lewandowski, D. Barber, S. Niederer, et al. "An automatic service for the personalization of ventricular cardiac meshes," *Journal of The Royal Society Interface*, vol. 11 (2014), p. 20131023.

[4] N. A. Trayanova. "Whole-heart modeling: applications to cardiac electrophysiology and electromechanics," *Circulation research*, vol. 108 (2011), pp. 113–128.

[5] J. Xi, P. Lamata, S. Niederer, S. Land, W. Shi, X. Zhuang, S. Ourselin, S. G. Duckett, A. K. Shetty, C. A. Rinaldi, et al. "The estimation of patient-specific cardiac diastolic functions from clinical measurements," *Medical image analysis*, vol. 17 (2013), pp. 133–146.

[6] M. Sermesant, R. Chabiniok, P. Chinchapatnam, T. Mansi, F. Billet, P. Moireau, J.-M. Peyrat, K. Wong, J. Relan, K. Rhode, et al. "Patient-specific electromechanical models of the heart for the prediction of pacing acute effects in CRT: a preliminary clinical validation," *Medical image analysis*, vol. 16 (2012), pp. 201–215.

[7] A. Suinesiaputra, P. Ablin, X. Alba, M. Alessandrini, J. Allen, W. Bai, S. Cimen, P. Claes, B. R. Cowan, J. D'hooge, et al. "Statistical shape modeling of the left ventricle: myocardial infarct classification challenge," *IEEE journal of biomedical and health informatics*, vol. 22 (2017), pp. 503–515.

[8] J. Corral Acero, A. Schuster, E. Zacur, T. Lange, T. Stiermaier, S. J. Backhaus, H. Thiele, A. Bueno-Orovio, P. Lamata, I. Eitel, et al. "Understanding and improving risk assessment after myocardial infarction using automated left ventricular shape analysis," *Cardiovascular Imaging*, vol. 15 (2022), pp. 1563–1574.

[9] K. K. Bhatia, A. N. Price, W. Shi, J. V. Hajnal, and D. Rueckert. "Super-resolution reconstruction of cardiac mri using coupled dictionary learning," *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, IEEE. 2014, pp. 947–950.

[10] O. Oktay, W. Bai, M. Lee, R. Guerrero, K. Kamnitsas, J. Caballero, A. de Marvao, S. Cook, D. O'Regan, and D. Rueckert. "Multi-input cardiac image super-resolution using convolutional neural networks," *International conference on medical image computing and computer-assisted intervention*, Springer. 2016, pp. 246–254.

[11] N. Basty and V. Grau. "Super resolution of cardiac cine mri sequences using deep learning," *Image Analysis for Moving Organ, Breast, and Thoracic Images*, Springer, 2018, pp. 23–31.

[12] E. M. Masutani, N. Bahrami, and A. Hsiao. "Deep learning single-frame and multiframe super-resolution for cardiac mri," *Radiology* (2020), p. 192173.

[13] Y. Xia, N. Ravikumar, J. P. Greenwood, S. Neubauer, S. E. Petersen, and A. F. Frangi. "Super-resolution of cardiac MR cine imaging using conditional GANs and unsupervised transfer learning," *Medical Image Analysis* (2021), p. 102037.

[14] N. Savioli, A. de Marvao, W. Bai, S. Wang, S. A. Cook, C. W. Chin, D. Rueckert, and D. P. O'Regan. "Joint semi-supervised 3d super-resolution and segmentation with mixed adversarial gaussian domain adaptation," *arXiv preprint arXiv:2107.07975* (2021).

[15] J. Sander, B. D. de Vos, and I. Išgum. "Autoencoding low-resolution MRI for semantically smooth interpolation of anisotropic MRI," *Medical Image Analysis*, vol. 78 (2022), p. 102393.

[16] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. De Marvao, T. Dawes, D. P. O'Regan, et al. "Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation," *IEEE transactions on medical imaging*, vol. 37 (2017), pp. 384–395.

[17] J. Duan, G. Bello, J. Schlemper, W. Bai, T. J. Dawes, C. Biffi, A. de Marvao, G. Doumou, D. P. O'Regan, and D. Rueckert. "Automatic 3d bi-ventricular segmentation of cardiac images by a shape-refined multi-task deep learning approach," *IEEE transactions on medical imaging* (2019).

[18] C. Biffi, J. J. Cerrolaza, G. Tarroni, A. de Marvao, S. A. Cook, D. P. O'Regan, and D. Rueckert. "3d high-resolution cardiac segmentation reconstruction from 2d views using conditional variational autoencoders," *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, IEEE. 2019, pp. 1643–1646.

[19] M. Beetz, A. Banerjee, and V. Grau. "Biventricular surface reconstruction from cine MRI contours using point completion networks," *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, IEEE. 2021, pp. 105–109.

[20] S. Wang, C. Qin, N. Savioli, C. Chen, D. P. O'Regan, S. Cook, Y. Guo, D. Rueckert, and W. Bai. "Joint motion correction and super resolution for cardiac segmentation via latent optimisation," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. 2021, pp. 14–24.

[21] M. Beetz, A. Banerjee, and V. Grau. "Reconstructing 3d cardiac anatomies from misaligned multi-view magnetic resonance images with mesh deformation u-nets," *Geometric Deep Learning in Medical Image Analysis*, PMLR. 2022, pp. 3–14.

[22] W. Bai, W. Shi, A. de Marvao, T. J. Dawes, D. P. O'Regan, S. A. Cook, and D. Rueckert. "A bi-ventricular cardiac atlas built from 1000+ high resolution mr images of healthy subjects and an analysis of shape and motion," *Medical image analysis*, vol. 26 (2015), pp. 133–145.

[23] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove. "Deepsdf: learning continuous signed distance functions for shape representation," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 165–174.

[24] Z. Chen and H. Zhang. "Learning implicit fields for generative shape modeling," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5939–5948.

[25] V. M. Campello, P. Gkontra, C. Izquierdo, C. Martin-Isla, A. Sojoudi, P. M. Full, K. Maier-Hein, Y. Zhang, Z. He, J. Ma, et al. "Multi-centre, multi-vendor and multi-disease cardiac segmentation: the m&ms challenge," *IEEE Transactions on Medical Imaging*, vol. 40 (2021), pp. 3543–3554.

[26] L. A. Rinkel, V. Guglielmi, C. F. Beemsterboer, N.-S. Groeneveld, N. H. Lobé, S. M. Boekholdt, B. J. Bouma, F. F. Muller, L. F. Beenen, H. Marquering, et al. "Diagnostic yield of ecg-gated cardiac ct in the acute phase of ischemic stroke vs transthoracic echocardiography," *Neurology* (2022).

[27] S. Bruns, J. M. Wolterink, T. P. Van Den Boogert, J. P. Henriques, J. Baan, R. N. Planken, and I. Išgum. "Automatic whole-heart segmentation in 4d tavi treatment planning CT," *Medical Imaging 2021: Image Processing*, vol. 11596 SPIE. (2021), pp. 55–62.

[28] S. Bruns, J. M. Wolterink, T. P. van den Boogert, J. H. Runge, B. J. Bouma, J. P. Henriques, J. Baan, M. A. Viergever, R. N. Planken, and I. Išgum. "Deep learning-based whole-heart segmentation in 4d contrast-enhanced cardiac CT," *Computers in biology and medicine*, vol. 142 (2022), p. 105191.

[29] A. de Marvao, T. J. Dawes, W. Shi, C. Minas, N. G. Keenan, T. Diamond, G. Durighel, G. Montana, D. Rueckert, S. A. Cook, et al. "Population-based studies of myocardial hypertrophy: high resolution cardiovascular magnetic resonance atlases improve statistical power," *Journal of cardiovascular magnetic resonance*, vol. 16 (2014), pp. 1–10.

[30] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester, et al. "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?" *IEEE Transactions on Medical Imaging* (2018).

[31] T. Lewiner, H. Lopes, A. W. Vieira, and G. Tavares. "Efficient implementation of marching cubes' cases with topological guarantees," *Journal of graphics tools*, vol. 8 (2003), pp. 1–15.

[32] M. Tatarchenko, S. R. Richter, R. Ranftl, Z. Li, V. Koltun, and T. Brox. "What do single-view 3d reconstruction networks learn?" *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3405–3414.

[33] J. Chibane, T. Alldieck, and G. Pons-Moll. "Implicit functions in feature space for 3d shape reconstruction and completion," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6970–6981.

[34] D. Kingma and J. Ba. "Adam: a method for stochastic optimization," *ICLR*, vol. 5 (2015).

[35] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. "Automatic differentiation in PyTorch," *NIPS Autodiff Workshop*, 2017.

[36] S. E. Petersen, P. M. Matthews, F. Bamberg, D. A. Bluemke, J. M. Francis, M. G. Friedrich, P. Leeson, E. Nagel, S. Plein, F. E. Rademakers, et al. "Imaging in population science: cardiovascular magnetic resonance in 100,000 participants of uk biobank-rationale, challenges and approaches," *Journal of Cardiovascular Magnetic Resonance*, vol. 15 (2013), pp. 1–10.

[37] Q. Xu, W. Wang, D. Ceylan, R. Mech, and U. Neumann. "Disn: deep implicit surface network for high-quality single-view 3d reconstruction," *Advances in Neural Information Processing Systems*, vol. 32 (2019).

[38] Z. Zheng, T. Yu, Q. Dai, and Y. Liu. "Deep implicit templates for 3d shape representation," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1429–1439.

[39] Y. Deng, J. Yang, and X. Tong. "Deformed implicit field: modeling 3d shapes with learned dense correspondence," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10286–10296.

[40] K. Genova, F. Cole, A. Sud, A. Sarna, and T. Funkhouser. "Local deep implicit functions for 3d shape," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4857–4866.

[41] R. Chabra, J. E. Lenssen, E. Ilg, T. Schmidt, J. Straub, S. Lovegrove, and R. Newcombe. "Deep local shapes: learning local sdf priors for detailed 3d reconstruction," *European Conference on Computer Vision*, Springer. 2020, pp. 608–625.

[42] C. Jiang, A. Sud, A. Makadia, J. Huang, M. Nießner, T. Funkhouser, et al. "Local implicit grid representations for 3d scenes," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6001–6010.

[43] S. Peng, Y. Zhang, Y. Xu, Q. Wang, Q. Shuai, H. Bao, and X. Zhou. "Neural body: implicit neural representations with structured latent codes for novel view synthesis of dynamic humans," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9054–9063.

[44] J. M. Wolterink. "Going off-grid: continuous implicit neural representations for 3d vascular modeling," *Statistical Atlases and Computational Models of the Heart. Regular and CMRxMotion Challenge Papers: 13th International Workshop, STACOM 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Revised Selected Papers*, vol. 13593 Springer Nature. (2023), p. 79.

CHAPTER 7

Discussion and future perspectives

7.1 They are just not that into us

I would like to make a modest attempt, and link the work in chapters 2 and 3 to the societal discussion about the impact of information technology and digitization on human life. At the time of writing the Dutch philosopher of law, Maxim Februari, published an essay[1] that describes how especially the raise of artificial intelligence influences our democratic constitutional state. Februari warns himself and us that until now we have built machines *that do not care about us* and points us to the fact that *we do not form a moral community with them (the machines).* Hence, if we ask a neural network to make a decision *it* will do so without any moral concerns. This is what the title of this section refers to (with courtesy to Februaris work). In the same vein, the famous American cultural anthropologist and writer Mary Catherine Bateson stated[2] that until now we have not (yet) built one of our most essential elements of human wisdom into our devices, i.e., *humility.* I would like to claim that this is what chapters 2 and 3 in this thesis are about, but I am afraid, that would dishonor human humility. Nevertheless, I agree with Februari and Bateson and hope that future efforts succeed in building modest machine learning approaches that *augment human intellect.*[3] One aspect of modesty, I think, is aiming to *know what we do not know* and reasoning carefully if we are uncertain. Therefore, machine learning algorithms that make predictions should have a notion of uncertainty that can be trusted, i.e., is calibrated.

Research presented in chapters 2 and 3 made an attempt to increase trustworthiness, i.e., reliability of automatic cardiac segmentation methods, by exploiting predictive uncertainties. By separating the obvious uncertainties, e.g., at the borders of the cardiac structures, from the not-so-obvious uncertainties, we established a human-in-the-loop, i.e., a semi-automatic approach to increase accuracy of cardiac segmentations and therefore, accuracy of, e.g., global cardiac functional indices like left and right ventricular stroke volumes.

Since then, the community was able to develop even more accurate and robust approaches for automatic biomedical image segmentation.[4] However, automatic segmentation followed by manual correction is the common practice in clinical settings.[5] To reduce manual effort and make automatic segmentation approaches more reliable, it is still desired to equip our models with a mechanism to warn us humans when it (the software) might have failed. A recent benchmark study[6] confirmed that CMRI segmentation uncertainty is correlated with segmentation accuracy and therefore, can be exploited for segmentation quality control. Furthermore, the authors conclude that to quantify segmentation uncertainties, currently, Deep Ensembles[7] outperform competing methods.[8–14] By hindsight, it might have been valuable and satisfying to spend more of my PhD time on the development of a human-in-the-loop framework that enables uncertainty-guided semi-automatic cardiac image segmentation.[15] Such a software product is also very useful when segmentation models are deployed in

environments that are different from the training environments, i.e., when training and *real-life* data originate from different data distributions (out-of-distribution data). Generalization to out-of-distribution data is a capability natural to humans yet challenging for machines to reproduce.[16] Although, I am convinced, one day, we can trust fully automatic segmentation algorithms in the same way, we currently rely on anti-lock braking systems. However, I agree with many other researchers in the community that we are not so far yet. More training data with a larger variety of cardiac anatomies and shapes is a well known way forward. Therefore, collecting publicly available CMRI datasets through e.g., large scale cohort studies[17,18] is valuable. In addition, federated learning approaches[19–22] that enable training of our methods on distributed data without data sharing, is another way to improve robustness and accuracy of any deep learning based image analysis approach.

7.2 Simplicity sometimes works

In Chapter 4 we applied our previously developed automatic deep learning CMRI segmentation approach to assess right ventricular (RV) function in subjects suspected of Arrhythmogenic right ventricular cardiomyopathy (ARVC). We reveal that a fully automated approach is not good enough because our method struggles to accurately segment the RV in the most basal slices. Nevertheless, RV function can be accurately assessed by our method if the automatic segmentation of the most basal slice is replaced with the corresponding manual reference. Hence, despite our effort in chapters 2 and 3 to develop an uncertainty-guided semi-automatic CMRI segmentation approach, we show in Chapter 4 that a simple quality control and correction step can achieve significant segmentation improvement. Furthermore, this result also indicates that current state-of-the-art CMRI segmentation methods could be used in a clinical setting if combined with a straightforward manual quality control step. However, based on our findings from Chapter 4 one cannot conclude that our uncertainty-guided semi-automatic approach (Chapter 3) is inferior compared with the simple quality control step i.e., to always manually correct only the last basal slices of the automatic segmentations. Although, such an approach can be sufficient to assess global cardiac function, it neglects automatic segmentation errors in mid-ventricular or apical slices that might hamper assessment of functional and anatomical abnormalities.

7.3 Prerequisites for accurate assessment of cardiac anatomy and function

Ideally, to accurately assess cardiac anatomy and function, one would like to have high-resolution 4D cardiac MRI volumes at one's disposal. Moreover, to enable advanced morphological and functional assessment of the heart, personalized high-resolution representation of cardiac anatomy is considered a prerequisite [23, 24]. Unfortunately,

until now acquisition of high fidelity CMRI volumes with high spatial and temporal resolution is not yet feasible. Moreover, currently, short-axis CMR scans with high temporal resolution are often highly anisotropic, lack whole-heart coverage, and suffer from respiratory motion-induced inter-slice misalignment. Therefore, in Chapter 5 we developed an unsupervised deep learning method to increase spatial resolution of short-axis CMR volumes in through-plane direction. Our methodological contribution was driven by the fact that although clinical practice lacks 3D high-resolution CMRI volumes, a large amount of anisotropic CMR images is typically produced in daily clinical workflow. Using the latter images to train our approach, i.e., given the unsupervised nature of the method, high-resolution training data is not required and hence, the method can be readily applied in clinical settings. While the method can synthesize new CMR slices in-between two adjacent slices, the approach cannot infer missing shape information at the apex or base of the heart, if the acquisition does not include a slice below the apex or above the base, respectively. Furthermore, synthesized slices only contain anatomical structures that are present in at least one of the two adjacent slices and therefore, the model does not *hallucinate* new content.

Moreover, the approach does not correct for respiratory motion induced inter-slice misalignment. Although, the method can synthesize semantically meaningful intermediate slices for CMR volumes with inter-slice misalignment, 3D geometry of ventricle shapes still exhibits unrealistic deformations. To accurately assess cardiac anatomy and function, slice alignment is an essential prerequisite because motion induced inter-slice misalignment (i) hinders CMRI short-axis segmentation using state-of-the-art 3D convolutional neural networks (CNN), (ii) complicates super-resolution of short-axis CMRIs, and (iii) hampers accurate registration of cardiac cine MRIs to assess cardiac motion. Certainly, other researchers have addressed the issue of CMR slice misalignment.[25–28] Nevertheless, they all require additional CMRI LV long-axis segmentations that are often not available in clinical settings.

To tackle some of the above mentioned shortcomings of our CMRI super-resolution approach, in Chapter 6, we described a method for high-resolution 3D LV reconstruction and anatomical shape completion using anisotropic CMRI segmentations. Unlike the previous approach (Chapter 5) that operates in image space, the latter method performs super-resolution in shape space. Furthermore, the approach (chapter 6) can complete anatomical shape information missing in the sparse-view CMRI segmentations, especially at the apex and base of the heart, and can correct for some of the motion induced inter-slice misalignment. This is accomplished by training our approach on high-resolution LV segmentations from cardiac computed tomography (CT) angiography (CCTA), and hence, exploiting the high-resolution and fast acquisition of CT. In addition, compared with automatically obtained low-resolution CMRI LV segmentations, reconstructed high-resolution LV shapes are geometrically smooth and might be less susceptible to anatomical implausibilities, i.e., topological errors, due to the strong LV shape prior that was learned during model training. This might be of

advantage in computational cardiac electrophysiological simulations which require geometrical smoothness and topological correctness of cardiac shapes.[29]

7.4 Clinically useful reconstructions of cardiac anatomy

Bi-ventricular shapes obtained from deep learning-based voxel-wise CMRI segmentations often contain anatomical inconsistencies, e.g., fragmented cardiac structures or holes.[30,31] Whilst such automatically obtained bi-ventricular shapes often suffice to derive volumetric measurements of the structures of interest (e.g., stroke volume), they might preclude assessment of local functional and anatomical abnormalities.[32] Moreover, topological correctness of acquired cardiac shapes are potentially important for applications like motion and strain analysis[33–35] or cardiac electrophysiological simulation.[29] However, topological correctness of automatically obtained cardiac anatomy from CMRI is typically not evaluated. Instead, performance evaluation of automatic segmentation and/or high-resolution reconstruction methods predominantly comprises overlap and surface distance metrics computed between automatically obtained cardiac shapes and their corresponding references. What follows is not new but reverberates what has been proclaimed before. To potentially reduce the gap between research and practical use, it might be beneficial if we, the research community, would sometimes manage to be less fixated on improving specific metrics. Instead, evaluation approaches for the aforementioned methods should depend on the requirements of the clinical task. Finally, sometimes novel promising methods should be granted some space on the *publication podium* although, they may not yet surpass current state-of-the-art approaches.

References

[1] M. Februari. "Doe zelf normaal," Prometheus Amsterdam, 2023.

[2] M. C. Bateson. "How to be a systems thinker," `https://www.edge.org/conversation/mary_catherine_bateson-how-to-be-a-systems-thinker`, Edge.org, 2018.

[3] D. C. Engelbart. "Augmenting human intellect: a conceptual framework," *Menlo Park, CA*, vol. 21 (1962).

[4] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18 (2021), pp. 203–211.

[5] T. A. Retson, E. M. Masutani, D. Golden, and A. Hsiao. "Clinical performance and role of expert supervision of deep learning for cardiac ventricular volumetry: a validation study," *Radiology: Artificial Intelligence* (2020), p. e190064.

[6] M. Ng, F. Guo, L. Biswas, S. E. Petersen, S. K. Piechnik, S. Neubauer, and G. Wright. "Estimating uncertainty in neural networks for cardiac MRI segmentation: a benchmark study," *IEEE Transactions on Biomedical Engineering* (2022).

[7] B. Lakshminarayanan, A. Pritzel, and C. Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in Neural Information Processing Systems*, 2017, pp. 6402–6413.

[8] D. P. Kingma, T. Salimans, and M. Welling. "Variational dropout and the local reparameterization trick," *Advances in neural information processing systems*, 2015, pp. 2575–2583.

[9] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. "Weight uncertainty in neural network," *International conference on machine learning*, PMLR. 2015, pp. 1613–1622.

[10] Y. Gal and Z. Ghahramani. "Dropout as a bayesian approximation: representing model uncertainty in deep learning," edited by M. F. Balcan and K. Q. Weinberger. Vol. 48 (PMLR, 2016), pp. 1050–1059.

[11] C. Louizos and M. Welling. "Multiplicative normalizing flows for variational bayesian neural networks," *International Conference on Machine Learning*, PMLR. 2017, pp. 2218–2227.

[12] Y. Wen, P. Vicol, J. Ba, D. Tran, and R. Grosse. "Flipout: efficient pseudo-independent weight perturbations on mini-batches," *arXiv preprint arXiv:1803.04386* (2018).

[13] B. Efron and R. Tibshirani. "[bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy]: rejoinder," *Statistical science*, vol. 1 (1986), pp. 77–77.

[14] P. Schulam and S. Saria. "Can you trust this prediction? auditing pointwise reliability after learning," *The 22nd international conference on artificial intelligence and statistics*, PMLR. 2019, pp. 1022–1031.

[15] M. Huellebrand, M. Ivantsits, L. Tautz, S. Kelle, and A. Hennemuth. "A collaborative approach for the development and application of machine learning solutions for CMR-based cardiac disease classification," *Frontiers in Cardiovascular Medicine*, vol. 9 (2022).

[16] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy. "Domain generalization: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45 (2023), pp. 4396–4415.

[17] D. E. Bild, D. A. Bluemke, G. L. Burke, R. Detrano, A. V. Diez Roux, A. R. Folsom, P. Greenland, D. R. JacobsJr, R. Kronmal, K. Liu, et al. "Multi-ethnic study of atherosclerosis: objectives and design," *American journal of epidemiology*, vol. 156 (2002), pp. 871–881.

[18] S. E. Petersen, P. M. Matthews, F. Bamberg, D. A. Bluemke, J. M. Francis, M. G. Friedrich, P. Leeson, E. Nagel, S. Plein, F. E. Rademakers, et al. "Imaging in population science: cardiovascular magnetic resonance in 100,000 participants of uk biobank-rationale, challenges and approaches," *Journal of Cardiovascular Magnetic Resonance*, vol. 15 (2013), pp. 1–10.

[19] J. Scherer, M. Nolden, J. Kleesiek, J. Metzger, K. Kades, V. Schneider, M. Bach, O. Sedlaczek, A. M. Bucher, T. J. Vogl, et al. "Joint imaging platform for federated clinical data analytics," *JCO clinical cancer informatics*, vol. 4 (2020), pp. 1027–1038.

[20] J. Klein, M. Wenzel, D. Romberg, A. Köhn, P. Kohlmann, F. Link, A. Hänsch, V. Dicken, R. Stein, J. Haase, et al. "Quantmed: component-based deep learning platform for translational research," *Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications*, vol. 11318 SPIE. (2020), pp. 229–236.

[21] Y. Wu, D. Zeng, Z. Wang, Y. Shi, and J. Hu. "Federated contrastive learning for volumetric medical image segmentation," *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, Springer. 2021, pp. 367–377.

[22] S. Goto, D. Solanki, J. E. John, R. Yagi, M. Homilius, G. Ichihara, Y. Katsumata, H. K. Gaggin, Y. Itabashi, C. A. MacRae, et al. "Multinational federated learning approach to train ECG and echocardiogram models for hypertrophic cardiomyopathy detection," *Circulation*, vol. 146 (2022), pp. 755–769.

[23] A. Suinesiaputra, P. Ablin, X. Alba, M. Alessandrini, J. Allen, W. Bai, S. Cimen, P. Claes, B. R. Cowan, J. D'hooge, et al. "Statistical shape modeling of the left ventricle: myocardial infarct classification challenge," *IEEE journal of biomedical and health informatics*, vol. 22 (2017), pp. 503–515.

[24] J. Corral Acero, A. Schuster, E. Zacur, T. Lange, T. Stiermaier, S. J. Backhaus, H. Thiele, A. Bueno-Orovio, P. Lamata, I. Eitel, et al. "Understanding and improving risk assessment after myocardial infarction using automated left ventricular shape analysis," *Cardiovascular Imaging*, vol. 15 (2022), pp. 1563–1574.

[25] J. Lötjönen, M. Pollari, S. Kivistö, and K. Lauerma. "Correction of movement artifacts from 4-d cardiac short-and long-axis MR data," *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2004: 7th International Conference, Saint-Malo, France, September 26-29, 2004. Proceedings, Part II 7*, Springer. 2004, pp. 405–412.

[26] C. Zakkaroff, A. Radjenovic, J. Greenwood, and D. Magee. "Stack alignment transform for misalignment correction in cardiac MR cine series," tech. rep. Citeseer, 2012.

[27] B. Villard, E. Zacur, E. Dall'Armellina, and V. Grau. "Correction of slice misalignment in multi-breath-hold cardiac MRI scans," *Statistical Atlases and Computational Models of the Heart. Imaging and Modelling Challenges: 7th International Workshop, STACOM 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 17, 2016, Revised Selected Papers 7*, Springer. 2017, pp. 30–38.

[28] A. Banerjee, E. Zacur, R. P. Choudhury, and V. Grau. "Optimised misalignment correction from cine mr slices using statistical shape model," *Annual Conference on Medical Image Understanding and Analysis*, Springer. 2021, pp. 201–209.

[29] P. Lamata, M. Sinclair, E. Kerfoot, A. Lee, A. Crozier, B. Blazevic, S. Land, A. J. Lewandowski, D. Barber, S. Niederer, et al. "An automatic service for the personalization of ventricular cardiac meshes," *Journal of The Royal Society Interface*, vol. 11 (2014), p. 20131023.

[30] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester, et al. "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?" *IEEE Transactions on Medical Imaging* (2018).

[31] S. Bohlender, I. Oksuz, and A. Mukhopadhyay. "A survey on shape-constraint deep learning for medical image segmentation," *IEEE Reviews in Biomedical Engineering* (2021).

[32] J. N. Cohn, R. Ferrari, N. Sharpe, and an International Forum on Cardiac Remodeling. "Cardiac remodeling—concepts and clinical implications: a consensus paper from an international forum on cardiac remodeling," *Journal of the American College of Cardiology*, vol. 35 (2000), pp. 569–582.

[33] M. A. Morales, M. Van den Boomen, C. Nguyen, J. Kalpathy-Cramer, B. R. Rosen, C. M. Stultz, D. Izquierdo-Garcia, and C. Catana. "Deepstrain: a deep learning workflow for the automated characterization of cardiac mechanics," *Frontiers in Cardiovascular Medicine* (2021), p. 1041.

[34] Q. Meng, C. Qin, W. Bai, T. Liu, A. de Marvao, D. P. O'Regan, and D. Rueckert. "Mulvimotion: shape-aware 3d myocardial motion tracking from multi-view cardiac MRI," *IEEE transactions on medical imaging*, vol. 41 (2022), pp. 1961–1974.

[35] C. Qin, S. Wang, C. Chen, W. Bai, and D. Rueckert. "Generative myocardial motion tracking via latent space exploration with biomechanics-informed prior," *Medical Image Analysis*, vol. 83 (2023), p. 102682.

Summary

Cardiovascular magnetic resonance (CMR) imaging is the reference modality for morphological and functional assessment of the heart. Typically, obtaining a high-resolution (HR) image and shape of the cardiac anatomy is crucial for such assessment. Short-axis CMR imaging, covering the entire left and right ventricles is routinely used to determine quantitative parameters of both ventricles' function. For this, manual or (semi-)automatic segmentation of the left and right endo- and epicardial structures in short-axis CMR images (CMRI) for at least end-diastole and end-systole is a key task. Manual segmentation of CMRIs is laborious (≈20 minutes for both time points) and prone to intra- and inter-observer variability. Moreover, to quantify parameters of cardiac motion requires segmentation across a complete cardiac cycle, comprising 20 to 40 phases per patient. Due to the required workload, this is practically infeasible and hence, precludes comprehensive routine analysis. Over the last few years many state-of-the-art deep learning segmentation approaches for short-axis CMRI have been developed. For automatic left ventricle segmentation such methods can achieve performance level of human experts. However, even the best performing methods generate anatomically implausible segmentations, mainly, in the most basal and apical slices. Therefore, existing semi-automated or automated segmentation methods for CMRIs regularly require (substantial) manual intervention. Furthermore, conventionally, to acquire stacks of short-axis 3D cine CMR images simultaneous multi-slice 2D cine CMR imaging is performed under multiple breath-holds. To mitigate the risk for motion artifacts and to sustain patient comfort fast scanning is often required. As a result, short-axis CMR scans with high temporal resolution are often highly anisotropic and suffer from respiratory motion induced inter-slice misalignment (example depicted in Fig 6.1). Moreover, caused by the low through-plane resolution ranging between 5 and 10 mm, short-axis CMR volumes often lack whole-heart coverage predominately at the apex and base of the heart. These shortcomings may hamper correct assessment of cardiac anatomy and subsequently hinder accurate analysis of cardiac function. This thesis presents approaches to tackle the aforementioned challenges.

CHAPTER 2 presents a method for automatic segmentation of cardiac anatomical structures in cardiac magnetic resonance images (CMRI). The Bayesian dilated convolutional network generates segmentation masks and spatial uncertainty maps for the input image at hand. Combining segmentations and uncertainty maps, we observed that image areas indicated as highly uncertain, regarding the obtained segmentation, almost entirely cover regions of incorrect segmentations. Furthermore, we found

that a model trained with the soft-Dice loss produces inferior calibrated probabilities compared to Brier and cross-entropy loss functions.

Based on these findings, in **CHAPTER 3** we present an approach that combines automatic CMRI segmentation with detection of image regions containing local segmentation failures. To predict regions in the automatic segmentation mask that potentially contain local segmentation failures, a detection network takes a cardiac MR image together with the corresponding spatial uncertainty map as input. We found that (simulated) manual correction of detected segmentation failures resulted in increased segmentation performance.

In **CHAPTER 4**, we applied our previously developed automatic deep learning CMRI segmentation approach to assess right ventricular function in subjects suspected of Arrhythmogenic right ventricular cardiomyopathy (ARVC). ARVC is diagnosed according to the Task Force Criteria (TFC) in which assessment of right ventricular function using CMRI segmentations plays an important role. We found that automatic segmentation of CMRIs in combination with correction of the most basal slice results in accurate CMR TFC classification of subjects suspected of ARVC.

CHAPTER 5 presents an deep learning semantic interpolation approach to increase through-plane resolution of anisotropic CMR short-axis images. Accurate analysis of cardiac function using CMRI is typically hampered by low through-plane resolution of CMR short-axis images. The approach synthesizes new intermediate slices from encoded low-resolution examples. Evaluation on cardiac, neonatal and adult brain MRI revealed that the approach outperforms cubic B-spline interpolation in terms of Peak Signal-to-Noise Ratio and Structural Similarity Index Measure.

CHAPTER 6 describes a deep learning approach to learn a continuous implicit function representing 3D left ventricle shapes. The model is trained using high-resolution segmentations from cardiac CT angiography. We found that such a model can be used to perform high-resolution reconstruction and anatomical shape completion of anisotropic incomplete cardiac MRI segmentations. Furthermore, the evaluation on segmentations in CMR short-axis images revealed that the approach can correct motion artifacts.

Finally, **CHAPTER 7** provides a general discussion of the presented approaches and discusses possible future directions.

Nederlandse samenvatting

Cardiovasculaire *magnetic resonance imaging* (CMRI), in het Nederlands soms aangeduid met kernspintomografie, is een non-invasieve beeldvormingstechniek die in de kliniek vaak wordt toegepast indien het vermoeden bestaat dat een patiënt een aandoening heeft aan de hartspier. In vergelijking met andere modaliteiten beschikt MRI beeldvorming over een superieure weke-delen contrast. Verder bestaat de mogelijkheid om in een enkel onderzoek zowel anatomie als functie te evalueren. Dit heeft ertoe geleid dat MRI beeldvorming van het hart inmiddels de referentiemodaliteit is geworden om de morfologie en functie van het hart accuraat te beoordelen. Hiervoor is delineatie, d.w.z. segmentatie, van linker (LV) en rechter ventrikel (RV) in driedimensionale (3D) CMR beelden een vereiste. Manuele segmentatie van deze structuren is zeer bewerkelijk en tijdrovend. Bovendien leidt manuele segmentatie tot grote intra- en inter-waarnemer variabiliteit. Op grond hiervan zijn er in het verleden automatische segmentatie methoden ontwikkeld. De meeste methoden maken daarbij gebruik van *deep learning* met zogenaamde convolutionele neurale netwerken (CNN). De best presterende CNNs voor automatische segmentatie van het linker ventrikel in CMR beelden bereiken inmiddels het prestatieniveau van menselijke experts. Desalniettemin produceren dezelfde methoden vaak vormen van hartstructuren die anatomisch gezien niet plausibel zijn, vooral in de meest basale en apicale segmenten. Daarom vereisen bestaande (semi-)automatische segmentatiemethoden voor CMR beelden regelmatig (aanzienlijke) handmatige controle en interventie.

Voor een betrouwbare en accurate diagnose is een hoge spatiële en temporele resolutie van de 3D MR beelden van belang. In de praktijk is het verkrijgen van dergelijke beelden meestal onmogelijk, voornamelijk, omdat het CMR beeldvormingsproces te langzaam verloopt. De meest gebruikte scanprotocollen zijn zo opgebouwd dat er een serie afzonderlijke 2D slices van het hart wordt geacquireerd in verschillende anatomische oriëntaties.[1] Deze tijdrovende 2D-acquisitietechniek is deels nodig omdat een 3D-acquisitie te veel tijd kost en niet compatibel is met een korte ademstilstand van de patiënt.[1]

Als gevolg hiervan zijn CMR beelden hoog anisotroop, d.w.z., de spatiële resolutie in de richting van de z-as ligt tussen de 5 en 10 mm (voorbeeld in Figuur 6.1). Bovendien worden de linker en rechter ventrikel vaak niet volledig door de beelden afgedekt. Doordat de patiënt en het hart gedurende de beeldopname kunnen bewegen, vertonen cardiale korte-as MR beelden regelmatig bewegingsartefacten. Deze tekortkomingen kunnen een correcte beoordeling van de anatomie en functie van het hart belemmeren.

Desalniettemin worden deze beelden in de dagelijkse klinische praktijk gebruikt om meetbare indicatoren van de hartfunctie te bepalen (b.v. het slagvolume en de ejectie fractie). Dit proefschrift presenteert benaderingen om de bovengenoemde uitdagingen aan te gaan.

HOOFDSTUK 2 presenteert een methode voor automatische segmentatie van cardiale anatomische structuren in cardiale magnetische resonantiebeelden (CMRI). Het *Bayesian dilated convolutional network* genereert segmentatiemaskers en ruimtelijke onzekerheidskaarten voor het ingevoerde beeld. Door segmentaties en onzekerheidskaarten te combineren, hebben wij vastgesteld dat beeldgebieden die als zeer onzeker zijn aangeduid, met betrekking tot de verkregen segmentatie, bijna volledig gebieden van onjuiste segmentaties omvatten. Voorts hebben wij vastgesteld dat een model dat is getraind met de *soft-Dice* verliesfunctie inferieur gekalibreerde waarschijnlijkheden oplevert in vergelijking met de *Brier* en *cross-entropy* verliesfuncties.

Op basis van deze bevindingen presenteren wij in **HOOFDSTUK 3** een benadering die automatische CMRI segmentatie combineert met detectie van beeld regionen die lokale segmentatiefouten bevatten. Om regionen in het automatische segmentatiemasker te voorspellen die mogelijk lokale segmentatiefouten bevatten, neemt een detectienetwerk een cardiaal MR beeld en bijbehorende ruimtelijke onzekerheidskaart als input. De resultaten hebben aangetoond dat (gesimuleerde) handmatige correctie van gedetecteerde segmentatiefouten heeft geresulteerd in een betere segmentatieprestatie.

In **HOOFDSTUK 4** hebben we onze eerder ontwikkelde automatische deep learning CMRI segmentatie methode toegepast om de functie van de rechter ventrikel te beoordelen bij personen waar het vermoeden bestaat van aritmogene rechter ventrikel cardiomyopathie (ARVC). ARVC wordt gediagnosticeerd volgens de Task Force Criteria (TFC) waarbij beoordeling van de rechter ventrikel functie met behulp van CMRI segmentaties een belangrijke rol speelt. De resultaten van ons onderzoek toonden aan dat voor de CMR TFC classificatie van personen met aanwijzingen voor ARVC, automatische CMRI segmentatie gebruikt kan worden, indien de meest basale slice manueel ingetekend wordt.

HOOFDSTUK 5 presenteert een op deep learning gebaseerde semantische interpolatie methode om de transversale spatiële resolutie van anisotrope cardiale MR beelden te verhogen. Nauwkeurige analyse van de hartfunctie met behulp van cardiale korte-as MR beelden wordt doorgaans belemmerd door de lage transversale resolutie van deze beelden. De methode genereert nieuwe tussenliggende slices uit gecodeerde beelden met lage resolutie. We hebben de methode op cardiale MR beelden, neonatale en volwassen MR hersens-scans geëvalueerd. De resultaten tonen aan dat de voorgestelde methode beter presteert als cubic B-spline interpolatie in termen van *Peak Signal-to-Noise Ratio* en *Structural Similarity Index Measure*.

In **HOOFDSTUK 6** beschrijven we een deep learning-benadering die leert om bestaande vormen van de linker hartkamer te representeren en met hoge spatiële resolutie te reconstrueren. Het model is getraind met behulp van hoge-resolutie segmentaties

van cardiale computertomografie (CT) angiografie scans. Wij ontdekten dat een dergelijk model gebruikt kan worden voor hoge-resolutie reconstructie en anatomische vormaanvulling van anisotrope onvolledige cardiale MRI segmentaties. Bovendien bleek uit de evaluatie van segmentaties in cardiale korte-as MR beelden dat de aanpak bewegingsartefacten kan corrigeren.

Ten slotte geeft **HOOFDSTUK 7** een algemene bespreking van de gepresenteerde benaderingen en de belangrijkste bevindingen, inclusief beperkingen, mogelijke klinische toepassingen en toekomstige richtingen.

Portfolio

Name PhD student:	Jörg Sander
PhD period:	January 2018–March 2023
Names PhD supervisors:	prof. dr. I. Išgum
	prof. dr. T. Leiner
Names PhD co-supervisors:	dr. B. D. de Vos
	prof. dr. ir. M. A. Viergever

PhD portfolio

	Year	ECTS
General courses		
Career development	2022	1.5
Specific courses		
Cardiovascular Epidemiology	2018	1.5
Medical Image Formation	2018	4.0
Front-End Vision and Multiscale Image Analysis	2018	2.5
Computer Vision by Learning	2019	1.5
Seminars, workshops and master classes		
Workshop data visualization	2020	0.1
Workshop negotiating	2020	0.1
Workshop focus like a pro	2022	0.1
Presentations		
SPIE Medical Imaging San Diego (oral)	2019	0.5
SPIE Medical Imaging online (oral)	2021	0.5
(Inter)national conferences		
MIDL Amsterdam	2018	0.9
SPIE Medical Imaging San Diego	2019	1.5
MIDL London	2019	0.9
SPIE Medical Imaging (online)	2021	1.5
Supervising		
Douwe van der Wal, Master thesis	2021	1.5
Wenli Xue, Master internship	2021	1.0
Roel Klein, Master thesis assessment	2022	0.1

Reviewed manuscripts		
Journal of Computers in Biology and Medicine	2019	
Journal of Medical Image Analysis	2022	
Journal of Medical Imaging	2022	
Other		
Weekly ISI meeting	2018–2019	2.0
Weekly qia/qurAI meeting	2018–2022	4.0
DLMedIA Hackathon	2019	0.5
Bi-weekly Cardiology AI meeting	2019–2020	0.5
Weekly machine learning meeting	2018–2022	1.5

List of publications

Journal publications

J. Sander, B. D. de Vos, and I. Išgum. "Automatic segmentation with detection of local segmentation failures in cardiac MRI," *Scientific Reports*, vol. 10 (2020), pp. 1–19.

J. Sander, B. D. de Vos, and I. Išgum. "Autoencoding low-resolution MRI for semantically smooth interpolation of anisotropic MRI," *Medical Image Analysis*, vol. 78 (2022), p. 102393.

M. Bourfiss, **J. Sander**, B. D. de Vos, A. S. Te Riele, F. W. Asselbergs, I. Išgum, and B. K. Velthuis. "Towards automatic classification of cardiovascular magnetic resonance task force criteria for diagnosis of arrhythmogenic right ventricular cardiomyopathy," *Clinical Research in Cardiology* (2022), pp. 1–16.

J. Sander, B. D. de Vos, S. Bruns, N. Planken, M. A. Viergever, T. Leiner, and I. Išgum. "Reconstruction and completion of high-resolution 3d cardiac shapes using anisotropic cmri segmentations and continuous implicit neural representations," *Computers in Biology and Medicine* (2023), p. 107266.

D. van der Wal, I. Jhun, I. Laklouk, J. Nirschl, L. Richer, R. Rojansky, T. Theparee, J. Wheeler, **J. Sander**, F. Feng, et al. "Biological data annotation via a human-augmenting AI-based labeling system," *NPJ digital medicine*, vol. 4 (2021), pp. 1–7.

Conference proceedings

J. Sander, B. D. de Vos, J. M. Wolterink, and I. Išgum. "Towards increased trustworthiness of deep learning segmentation methods on cardiac MRI," *Medical Imaging 2019: Image Processing*, vol. 10949 International Society for Optics and Photonics. (2019), p. 1094919.

J. Sander, B. D. de Vos, and I. Išgum. "Unsupervised super-resolution: creating high-resolution medical images from low-resolution anisotropic examples," *Medical Imaging 2021: Image Processing*, vol. 11596 International Society for Optics and Photonics. (2021), 115960E.

B. D. de Vos, B. H. van der Velden, **J. Sander**, K. G. Gilhuijs, M. Staring, and I. Išgum. "Mutual information for unsupervised deep learning image registration," *Medical Imaging 2020: Image Processing*, vol. 11313 SPIE. (2020), pp. 155–161.

Author contributions

Different types of author contributions

1	Study conception and design
2	Acquisition of data
3	Analysis and interpretation of data
4	Drafting the manuscript
5	Revision of the manuscript

CHAPTER 2: Towards increased trustworthiness of deep learning segmentation methods on cardiac MRI

Author	Contribution
J. Sander	1,2,3,4
Bob D. de Vos, J.M. Wolterink, I. Išgum	1,3,4,5

CHAPTER 3: Automatic segmentation with detection of local segmentation failures in cardiac MRI

Author	Contribution
J. Sander	1,2,3,4
Bob D. de Vos, I. Išgum	1,3,4,5

CHAPTER 4: **Towards automatic classification of cardiovascular magnetic resonance task force criteria for diagnosis of arrhythmogenic right ventricular cardiomyopathy**

Author	Contribution
J. Sander, Mimount Bourfiss	1,2,3,4
Bob D. de Vos, I. Išgum, Birgitta K. Velthuis	2,3,5
Anneline S.J.M. te Riele, Folkert W. Asselbergs	5

CHAPTER 5: **Autoencoding Low-Resolution MRI for Semantically Smooth Interpolation of Anisotropic MRI**

Author	Contribution
J. Sander	1,2,3,4
Bob D. de Vos, I. Išgum	1,3,4,5

CHAPTER 6: **High-resolution reconstruction and completion of anisotropic cardiac MRI segmentations using continuous implicit neural representations**

Author	Contribution
J. Sander	1,2,3,4
Bob D. de Vos, I. Išgum	1,3,4,5
S. Bruns	2
R.N. Planken	2,5
T. Leiner, M.A. Viergever	5

Acknowledgments

During my PhD I came across a Japanese self-reflection method called *Naikan*. The word Naikan translates to *inside looking*. It's a form of introspection, designed to help us understand ourselves and the human relationships we have. The method is structured around three questions, of which the first begins: What have I received from person **X**? I consider this a great question for writing this last chapter.

I claim to remember each step towards the beginning of my PhD, so, I also recall our first encounter Ivana, which was during my job interview at the Images Sciences Institute in Utrecht. In German we have this strong idiom *doctoral mother/father*, meaning doctoral advisor. This is exactly what I thought after I had spoken to you for about half an hour. Finally, when Bob joined the conversation for the last 15 minutes, he remarked that they often said to you, when heading home from work, now she is going to her real children. I was looking for an advisor who was dedicated to help me during my PhD. I had the strong feeling that I would not succeed without such substantial support. I was not disappointed by your support. You were always willing to help, no matter what. Thank you for the hundreds of hours you have dedicated to meeting, discussing our work, and particularly for your assistance in crafting conference and journal articles. Your bar is high and helped me to push my limits. Now years later, I respect the patience you have with each of us again and again. Unfortunately, we did not have many chances to visit conferences together. But I vividly remember the flight back from San Diego (SPIE conference) to Amsterdam in February 2019. At a certain moment, an elderly Dutch couple approached us. They were tired of our constantly ongoing conversation. You were friendly, and we kept on talking :). Finally, thank you for your ongoing trust and support although I needed more time than the average PhD student. No PhD for me without you!

My buddy Bob! For you, I regularly developed feelings of brotherhood. And although age-wise the reverse applies; I was the younger sibling. Your enthusiasm and inexhaustible creative ideas often blew me meters in the right direction and sometimes out of an energy sink. Goethe's poetic words apply without any doubt to you, "Himmelhoch jauchzend, zum Tode betrübt, glücklich allein ist die Seele die liebt"[1]. We spend hundreds of hours talking about our work and matters beyond. And I hope, you also could sometimes benefit from those conversations. Especially in the beginning of my PhD, we could fight like boys. And although Jelmer with whom you were sharing an office at that time, was sometimes questioning what was going on between us, in my

[1]Now on top of the world, now in the depths of despair, only the soul that loves is happy beyond compare

memories, we could always settle our dispute within minutes. Your scientific rigor and intuition for the weak spots in our and others scientific work impressed me. I learned so many things from you. Thank you so much for your immense effort to support and motivate me. Again, no PhD for me without you!

Thank you, Tim, for providing me with valuable perspectives that broadened my understanding of cardiac MRI and its clinical applications. Even though our interactions were relatively few, the moments we did share, you provided me with abundant clinical research insights and new research directions.

Max, unfortunately and fortunate for yourself, you officially retired (I know you don't :) as director of the Images Sciences Institute shortly after I started my PhD. Nevertheless, when we occasionally met, you were always interested in a personal story. Thank you for your support and trust to admit me as PhD student, although, I was already age 48.

I would also like to thank my reading committee. Prof. dr. ir. Jan-Jakob Sonke, prof. dr. Cees Snoek, prof. dr. ir. Marcel Breeuwer, prof. dr. Hildo Lamb, prof. dr. Rozemarijn Vliegenhart, and prof. dr. Henk Marquering, thank you for critically evaluating my thesis.

Furthermore, I would like to thank everybody with whom I have been collaborating during my PhD, and, in particular, Mimount. Our collaboration already started in the first month of my PhD. Although it took quite a while before we managed to produce a fruitful journal publication which is part of this thesis, I am very proud that we showed so much persistence. You are a wonderful collaborator, and I enjoyed working with you all the way. Birgitta, thank you for having patience with us, and supporting us to make the collaboration a success.

Unfortunately, Laura, we only had a couple of months of PhD overlap. It was such a pleasure to collaborate with you on your first and my last project. Thank you, for your willingness to carry on the cardiac strain torch and turn it into a successful conference publication.

Moreover, I would like to thank everybody who was part of the qurAI research team during my PhD: Bob, Caroline, Clarisa, Coen, Gino, Jelmer, Julia, Jurica, Laura, Louis, Majd, Maria, Marinka, Matthijs, Michiel, Mohammad, Nadieh, Navchetan, Nikolas, Nils, Riaan, Roel, Sanne, Steffen and Zhiwei. Thank you for all the valuable discussions and feedback.

Special thanks to my favourite *Toren C Pubermeisjes* and office mates Julia and Sanne. I was intimidated by both of your strong acting skills.

Moreover, special thanks to Lucas, Nils and Roel, my great last roommates. After Corona had impaired social PhD life to the fullest, I never could have imagined that in the short time left, we would develop such a great bonding! Thank you for distracting me from my work in such a valuable way.

Dear Fenghua! I am very glad that we could develop a friendship by genuinely sharing our PhD difficulties. I hope you can find your turn back home in China.

Dear Steffen, you deserve a special remark here. You did not succeed in protecting Julia from my incredible poetic, kauderwelsh, gibberish German, the reminiscence of 30 years living in the Netherlands. But you impressed me with your ironing skills when we shared a hotel room in San Diego. Thank you for your inspiration.

Also special thanks to my UvA AI *matties* and friends: Bas, Joost, Maartje, Maurits, Thijs and Ties. Although, you are starting to "fly out of the Mokum nest", I still enjoy our regular chill and cook sessions. Thank you for not asking too many questions about our PhDs during the last years.

Dear Amsterdam zen.nl community, dear Arthur and Raoul. Mentally, I ran into a concrete wall in the early stages of my PhD. As a result, I was lacking concentration and experienced strong anxieties. Our weekly lessons, sesshins in St. Adelbert, and the daily meditations brought me back on track. I am very grateful for your inspiration, presence and attention.

Lieve koorleden van Slavuj en in het bijzonder Ivo! Was ik maar eerder begonnen bij Slavuj te zingen, dan was ik door mijn PhD gevlogen uiteraard ;) Ik geniet iedere dinsdag van onze muziek en voel me zeer verbonden met jullie. Heel veel dank hiervoor.

Julia and Nils, thank you so much for carrying the burden of being my paranymphs.

I am very grateful to all my dear friends and acquaintances. For more than five years you all had to listen to my long PhD stories that I always intended to make short ;) Thanks for being patient with me.

My only and dearest sister, Stephanie. It often seems as if your subjective experience of *this world* resembles mine. I am so happy and grateful to have you as my sister.

Liebe Elsa und Karli! Mama und Papa, ohne Euch kein Jörg :) Ohne Eure beharrliche Unterstützung während der ersten 19 Jahre, (es müssen tausende Stunden gewesen sein!) hätte ich es wohl nie zur Universität geschafft (wäre das besser gewesen? ;). Aber viel wichtiger, Ihr habt mir so viel Liebe geschenkt. Dafür werde ich Euch ewig dankbar sein.

Finally, liebe Manja! My wife and dearest friend. I am so lucky to have your daily sunshine around. What a great life do we have together, fantastic! More than I could have hoped for. Let's get really old together :)

Biography

Jörg Sander was born on May 18th, 1969 in Münster, Germany. He received his Master of Arts in Experimental Psychology at the University of Leiden in 1997. After having worked for more than 15 years as an IT consultant, he decided in 2015 to study Artificial Intelligence at the University of Amsterdam. In his Master thesis he proposed a meta-learning approach for optimizing loss functions of base-learners by combining adaptive-computation-time and learning-to-learn approaches. In January 2018, Jörg started his PhD at the Image Sciences Institute, UMC Utrecht as part of Prof. dr. Ivana Išgum's Quantitative Medical Image Analysis group. In 2019, the group transferred to the Department of Biomedical Engineering and Physics at the Amsterdam UMC where Jörg continued his PhD under the supervision of Prof. dr. Ivana Išgum, Dr. ir. Bob D. de Vos, Prof. dr. Tim Leiner and Prof. dr. ir. Max Viergever. His work on assessment of anatomy and function of the heart using 4D cardiac MRI and deep learning resulted in this thesis.