Imperial College London

Imperial College London

Department of Computing

Improving the Domain Generalization and Robustness of Neural Networks for Medical Imaging

Chen Chen

Main supervisor
Dr Daniel Rueckert

Second supervisor
Dr Wenjia Bai

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computing of Imperial College London

December 2021

Declaration of Originality

I, Chen Chen, hereby declare that the work described in this thesis is my own, except where specifically acknowledged.

Copyright Statement

Abstract

Deep neural networks are powerful tools to process medical images, with great potential to accelerate clinical workflows and facilitate large-scale studies. However, in order to achieve satisfactory performance at deployment, these networks generally require massive labeled data collected from various domains (e.g., hospitals, scanners), which is rarely available in practice. The main goal of this work is to improve the domain generalization and robustness of neural networks for medical imaging when labeled data is limited.

First, we develop multi-task learning methods to exploit auxiliary data to enhance networks. We first present a multi-task U-net that performs image classification and MR atrial segmentation simultaneously. We then present a shape-aware multi-view autoencoder together with a multi-view U-net, which enables extracting useful shape priors from complementary long-axis views and short-axis views in order to assist the left ventricular myocardium segmentation task on the short-axis MR images. Experimental results show that the proposed networks successfully leverage complementary information from auxiliary tasks to improve model generalization on the main segmentation task.

Second, we consider utilizing unlabeled data. We first present an adversarial data augmentation method with bias fields to improve semi-supervised learning for general medical image segmentation tasks. We further explore a more challenging setting where the source and the target images are from different data distributions. We demonstrate that an unsupervised image style transfer method can bridge the domain gap, successfully transferring the knowledge learned from labeled balanced Steady-State Free Precession (bSSFP) images to unlabeled Late Gadolinium Enhancement (LGE) images, achieving state-of-the-art performance on a public multi-sequence cardiac MR segmentation challenge.

For scenarios with limited training data from a single domain, we first propose a general training and testing pipeline to improve cardiac image segmentation across various unseen domains. We then present a latent space data augmentation method with a cooperative training framework to further enhance model robustness against unseen domains and imaging artifacts.

Acknowledgements

During the thesis writing, it kept bringing me so much joy especially when I recalled lovely people at Imperial that have been with me, provided and have been providing me huge help along this long journey. To me, they are stars shining brightly in the sky, lighting up my path.

First and foremost, I would like to express my huge gratitude to my esteemed supervisor, Prof. Daniel Rueckert, who gave me the opportunity to enter this fantastic research world and provided me with enormous invaluable advice, continuous support, and encouragement during my Ph.D. study. Thank you for sharing your immense knowledge and plentiful experience with me on a professional and on a more personal level when needed. Thank you for encouraging me to pursue my research interests with great freedom. I am also deeply impressed that you always make time for every student, post-doc, and staff, providing various levels of guidance or assistance in their work no matter how busy you are. Your values and attitude towards science and your philosophy of life have been inspirational for me. Danke!

I would also like to thank my second advisor, another wonderful person, Dr Wenjia Bai. It is indeed my privilege to work with him, who is always supportive, providing me with guidance and counsel whenever I need. He is also very generous in giving insightful comments and suggestions and practical tips on the design of experiment methods, academic writing, and presentation. He is not only my supervisor but also my mentor, friend, and role model! 謝謝!

The work in this thesis was supported by the EPSRC SmartHeart project (EP/P001009/1). Special gratitude also goes to my close collaborators: Cheng Ouyang, Chen Qin, Giacomo Tarroni, Kerstin Hammernik, Shuo Wang, Huaqi Qiu, Carlo Biffi in this project and friends in our BioMedIA group: Zeju Li, Jeremy Tan, Gavin Seegoolam and others. Their kind help and support have made my study and life in London a wonderful time. I would also like to thank my boyfriend Zehong Zhang and all my friends and beloved ones who supported me during good and bad times throughout this journey, especially during the COVID-19 lockdowns.

Last but not least, I would like to express my deep appreciation to my parents for their unconditional love and support. Without their understanding and financial aid, it is impossible for me to study abroad and pursue a Ph.D. degree. This Ph.D. thesis is dedicated solely to them.

‘The good life is one inspired by love and guided by knowledge. Neither love without knowledge, nor knowledge without love can produce a good life.’

– *Bertrand Russell*

‘To know that we know what we know, and to know that we do not know what we do not know, that is true knowledge.’

– *Nicolaus Copernicus*

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

In recent years, deep learning (DL) has gained significant attention and popularity both in the research and industry community and has been gradually developed as a state-of-the-art technique in various areas, including computer vision, natural language processing, and healthcare. Different from traditional machine learning (ML) algorithms which heavily rely on handcrafted feature engineering, DL algorithms, in general, adopt neural networks to *automatically* extract a set of complex hierarchical features from data. These features unveil the intricate structure in large raw data, which are essential for pattern recognition, decision-making, and inference.

In the field of medical data analysis, such an ability is highly desirable, allowing one to automatically extract, analyze, and interpret information from medical imaging data. For example, neural networks can be used to perform tedious tasks like segmenting anatomical structures and performing volume measurement from medical images (e.g., magnetic resonance imaging (MRI), computed tomography (CT), ultrasound) [1]. Fig. 1.1 presents an overview of typical cardiac segmentation tasks in the three most commonly used modalities where deep learning methods have been applied to. These applications include the segmentation of cardiac substructures such as the left ventricle (LV), right ventricle (RV), left atrium (LA), right atrium (RA), and coronary arteries, as well as the segmentation of tissues (e.g., scar) and other abnor-

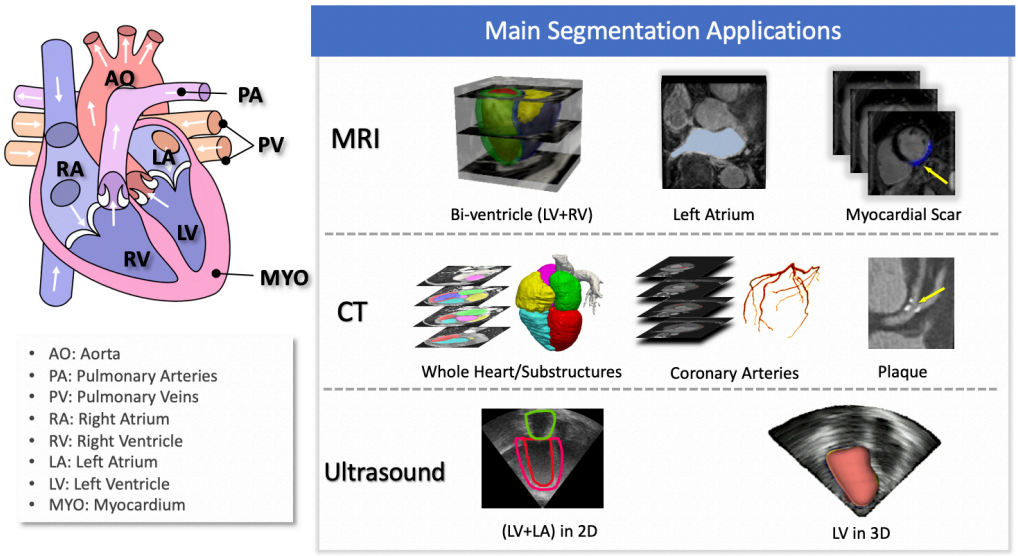


Figure 1.1: **Overview of cardiac image segmentation tasks for the three most common imaging modalities in which deep learning techniques have been applied.** Figure source: [1], reproduced under the terms of the Creative Commons Attribution License (CC BY 4.0).

malities such as plaque. This indicates DL's wide applicability to various segmentation tasks. Meanwhile, with the support of advanced hardware such as graphical processing units (GPUs) and tensor processing units (TPUs), neural networks can perform prediction very fast (e.g., less than a second). They can greatly reduce physicians, clinicians, and radiologists' workload and potentially improve healthcare with higher efficiency. Since 2015, neural networks have become the leading technique for automated medical image analysis, thanks to their impressive accuracy and speed in many vision tasks, such as anatomical structure segmentation, landmark detection, lesion detection, and segmentation, as well as image registration, image reconstruction, and computer-aided diagnosis/prognosis [2, 3].

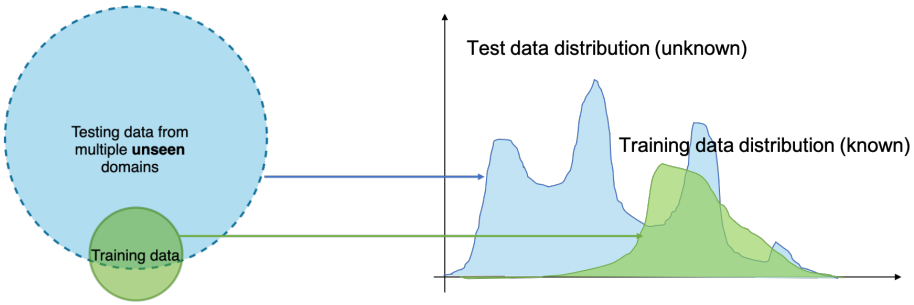


Figure 1.2: **Illustrative diagram of the distributional shift between the training and testing data in real-world applications.**

However, adopting DL into real-world medical imaging applications is still at an early stage.

One major obstacle that restricts the applicability of DL is that it, in general, requires large-scale labeled data from various scanners and sites to achieve satisfactory performance. Collecting and labeling such large-scale datasets for training can be expensive and even prohibitively impossible due to privacy concerns. As a result, it is common to have a limited training dataset, which fails to cover the full spectrum of test data in real-world clinical environments, as illustrated in Fig 1.2. Such a discrepancy between training and test data is termed as 'distributional shift' or 'domain shift', attributing to the model's significant performance drop at deployment time. For example, when Bai *et al.* applied a neural network-based segmentation model trained from a dataset from UK Biobank [5] to a public benchmark dataset collected from France: ACDC dataset [6], the left ventricle segmentation accuracy score dropped by 20 percent.

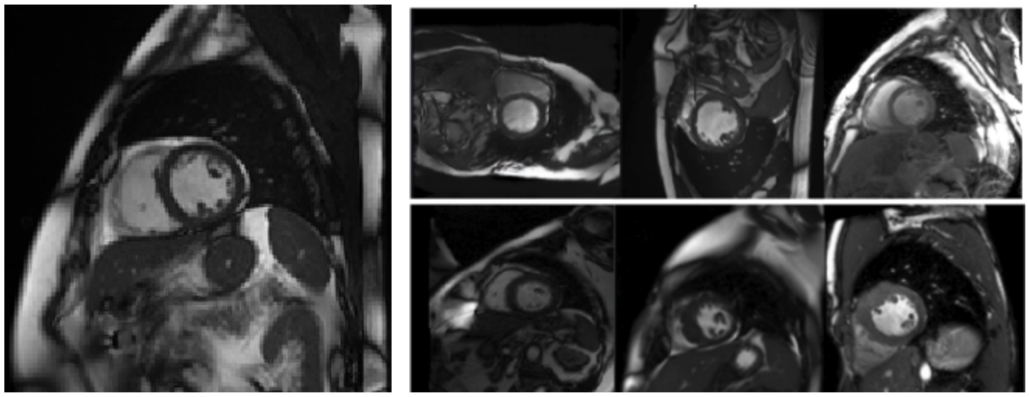


Figure 1.3: **Illustrative diagram of the domain shift problem in cardiac MR segmentation.** Variations in terms of cardiac structural differences as well as image appearances and contrast differences can be observed between the training distribution and unseen test distributions. Image source: [7].

As shown in Fig. 1.3, the domain shift problem or dataset bias problem in cardiac MR segmentation is mainly caused by two factors:

- Population shifts such as cardiac structural differences across different populations. Different datasets from other sites often comprise different populations regarding age, sex, race, and pathology. Among these datasets, a great of biological variability in heart size, orientation in the thorax, and cardiac structure deformations (not only in the diseased subjects with pathological deformations related to cardiac disease but also in the healthy subjects) can be observed;

- Image quality, appearance, and contrast variations resulted from differences in scanners (e.g., different vendors, different magnetic strengths), protocols, and image planning. For example, images from scanners with a 1.5T magnetic field often contain a higher noise level than those from 3T scanners, whereas 3T images show higher image contrast but are higher likely to suffer from imaging artifacts [8], see Fig. 1.4. Even with the same scanner, the quality of imaging can degrade significantly due to improper image acquisitions and abnormal patient conditions, e.g., very rapid heart rates, difficulty in holding their breath for a few seconds [8].

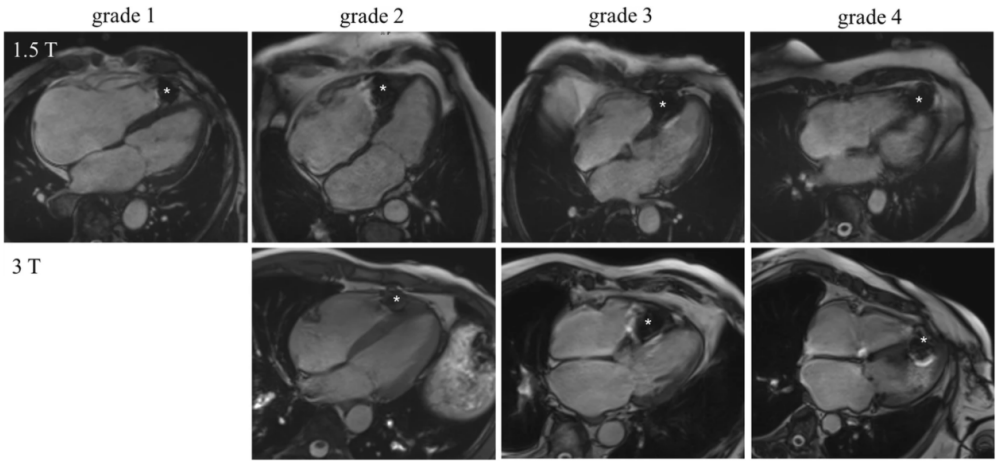


Figure 1.4: **Visualization of cardiac MR images scanned using different magnetic fields (1.5T vs 3T).** Here, images have been classified into four groups: grade 1 (excellent image quality), grade 2 (good), grade 3 (poor), and grade 4 (non-diagnostic). No patient in the 3 Tesla group showed a grade 1 in the examined datasets [9]. We can observe evident artifacts (bright stripes) on the 3T image with grade 3. Image source: [9], licensed under CC BY 4.0[a].

[a]http://creativecommons.org/licenses/by/4.0/

These biological differences and image appearance variations among different datasets pose challenges to the deployment of a DL-based model at scale. In this thesis, we focus on investigating techniques to enhance the generalization and robustness of neural networks without acquiring vast amounts of training data from new domains (e.g., hospitals, scanners). Specifically, we focus on improving:

- intra-domain generalization, which is used to describe a model's performance on unseen test data drawn from the *same* distribution as the training data, e.g., data from the same scanner or from the same population;

- out-of-domain generalization, which is used to describe a model's performance on out-of-distribution (OOD) data [1] where domain shift is presented between training and test datasets, e.g., data from different scanners or populations. Out-of-domain generalization [10] is very close to model robustness, which quantifies the model's stability against specific types of data shifts or corruptions, such as changes in vendors, image acquisition protocols, image quality, or population.

In Chapter 2, we will introduce DL basics as well as existing theories and common practices for improving model generalization. All these form the basis of our works presented in Chapters 3-5, which exploit different methods to improve model intra- and/or out-of-domain generalization under different data settings for cardiac image segmentation. More details can be found below.

1.2 Thesis outline and contributions

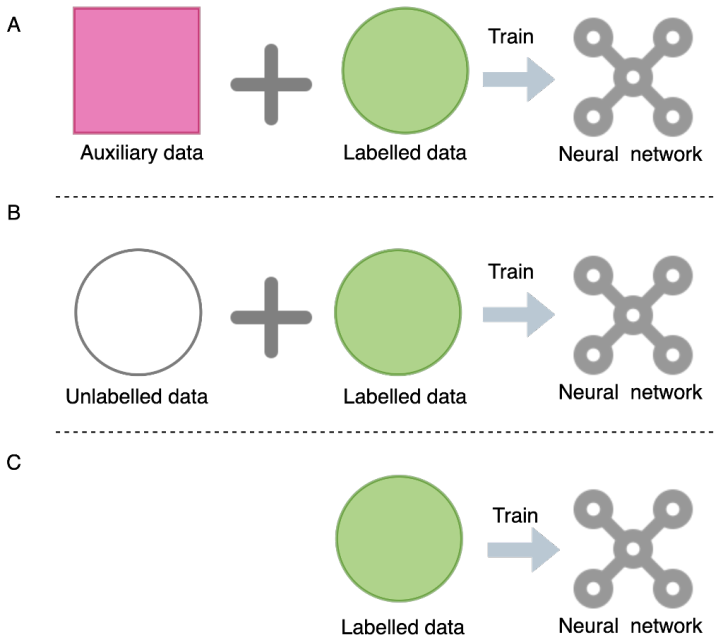


Figure 1.5: **Illustrative diagram of the three main topics covered in this thesis.** They are: A) learning with auxiliary data; B) learning with unlabeled data; C) learning with a limited labeled dataset without any additional data.

[1] In our thesis, the term 'out-of-domain' and 'out-of-distribution' are used interchangeably throughout the thesis. Here, the term 'domain' is referred to as 'dataset distribution'.

This thesis is mainly based on a list of works done in my Ph.D. study that has been published in top-tier conferences and peer-reviewed journals. A list of these works is provided in Sec.1.3. The remaining of this thesis is organized as follows: each chapter starts with a box recalling the publication(s) whose content is reproduced or adapted therein. Chapter 2 introduces fundamental concepts of deep learning and neural networks, followed by a literature review on recent developments for deep learning-based cardiac segmentation applications. We then discuss the limitations of deep learning approaches, as well as theory and practice to understand and to improve model generalization, which form the basis of the works that have been conducted.

In Chapter 3-5, we present our works for improving model generalization and robustness of neural networks. As illustrated in Fig. 1.2, our works can be summarized into three topics: a) learning with auxiliary data (Chapter 3), b) learning with unlabeled data (Chapter 4), c) learning with a limited labeled dataset without any additional data sources (Chapter 5). The last topic is of the greatest practical value but is the most challenging one due to the limitation of training data. A more detailed introduction for the three chapters is given below.

In Chapter 3 we present two different methods to exploit the value of auxiliary data for improving model intra-domain generalization. Specifically,

- We first introduce a multi-task learning network, which conducts image classification and segmentation tasks simultaneously. The network is constructed to exploit additional non-imaging patient information (i.e., whether this patient has undergone atrial ablation or not) to guide the representation learning process for segmentation. This method has been applied to segmenting left atrial from contrast-enhanced MR images and has achieved very promising results, ranking the 4^{th} in an international challenge;
- We then introduce a novel framework that can extract anatomical shape priors from multiple 2D standard views and leverage these anatomical priors to segment the left ventricular myocardium from short-axis MR image stacks. The proposed segmentation method has the advantage of being a 2D network but at the same time incorporates spatial context from multiple, complementary views that span a 3D space. We demonstrated

that our method achieves accurate and robust myocardium segmentation across different short-axis slices, especially on the most challenging slices: apical and basal slices.

In Chapter 4, we exploit unlabeled images for improving model performance for a particular domain. The first work is on semi-supervised learning for medical image segmentation, where a small labeled dataset and a relatively large unlabeled dataset draw from the same data distribution are available for training. The second work focuses on transferring knowledge learned from a domain with annotated training examples (source domain) to a different domain with unlabeled images only (target domain). Specifically,

- In the first part of the chapter, we present an adversarial data augmentation method for training neural networks for medical image segmentation. The proposed method is capable of generating adversarial images with plausible and realistic signal corruptions to supplement the training data. One of the main advantages of this adversarial data augmentation is that it does not require labeled data. Thus, it can be applied to both labeled and unlabeled data for semi-supervised learning. By continuously generating these realistic, 'hard' examples, we prevent the network from overfitting and, more importantly, encourage the network to defend itself from intensity perturbations by learning robust semantic features for the segmentation task. We demonstrate the efficacy of the proposed method on a public cardiac MR segmentation dataset in challenging low-data settings;

- In the second part, we present a fully automatic method to segment cardiac structures from late gadolinium enhancement (LGE) images without using labeled LGE data for training, but instead by transferring the anatomical knowledge and features learned on annotated balanced steady state free precession (bSSFP) images, which are easier to acquire. Specifically, we employ a multi-modal image translation network for style transfer and a cascaded segmentation network for image segmentation. The multi-modal image translation network generates realistic and diverse synthetic LGE images conditioned on a single annotated bSSFP image, forming a synthetic LGE training set. This set is then utilized to fine-tune the segmentation network pre-trained on labeled bSSFP images, achieving the goal of unsupervised LGE image segmentation. This method is evaluated on

the cardiac multi-sequence segmentation task and was ranked the 1^{st} in an international challenge [11].

In Chapter 5, we focus on improving model out-of-domain generalization without significantly sacrificing intra-domain performance. A very challenging but realistic data setting is considered: only labeled data from a single domain is available for training a neural network, which is then tested on multiple unseen test datasets. Specifically, we present two works:

- First, we present a simple yet effective way to improve network generalization ability by carefully designing data normalization and augmentation strategies to accommodate common scenarios in multi-site, multi-scanner clinical imaging data sets. We demonstrate that a neural network trained on a *single-site, single-scanner* dataset from the UK Biobank study[2] can be successfully applied to segmenting cardiac MR images across different unseen sites and different scanners without substantial loss of accuracy;

- In the second part, we present a cooperative framework for training image segmentation models and a latent space augmentation method for generating hard examples. Both contributions improve model generalization and robustness with limited data. The cooperative training framework consists of a fast-thinking network (FTN) and a slow-thinking network (STN). The FTN learns decoupled image features and shape features for image reconstruction and segmentation tasks. The STN learns shape priors for segmentation correction and refinement. The two networks are trained in a cooperative manner. The latent space augmentation generates challenging examples for training by masking the decoupled latent space in both channel-wise and spatial-wise manners. The network is trained on one dataset from one hospital and then evaluated on multiple different datasets acquired from different sources. We performed extensive experiments on public cardiac imaging datasets and demonstrated improved cross-site segmentation performance and particularly increased robustness against various unforeseen imaging artifacts compared to strong baseline methods.

[2] https://www.ukbiobank.ac.uk/

Finally, Chapter 6 concludes the work presented in this thesis and discusses potential future work.

1.3 List of publications

A list of published works is given below in chronological order:

1. **C. Chen**, W. Bai, and D. Rueckert, **Multi-task Learning for Left Atrial Segmentation on GE-MRI**, in Statistical Atlases and Computational Models of the Heart, Atrial Segmentation and LV Quantification Challenges - 9th International Workshop, STACOM 2018, Held in Conjunction with MICCAI 2018, 2018, pp. 292–301 [12].

2. **C. Chen**, C. Biffi, G. Tarroni, S. Petersen, W. Bai, and D. Rueckert, **Learning Shape Priors for Robust Cardiac MR Segmentation from Multi-view Images**, in Medical Image Computing and Computer Assisted Intervention - MICCAI 2019, 2019, pp. 523–531 [13].

3. **C. Chen**, C. Ouyang, G. Tarroni, J. Schlemper, H. Qiu, W. Bai, and D. Rueckert, **Unsupervised Multi-modal Style Transfer for Cardiac MR Segmentation**, in Statistical Atlases and Computational Models of the Heart - STACOM 2019, Held in Conjunction with MICCAI 2019, 2019, pp. 209–219 [14].

4. **C. Chen**, W. Bai, R. H. Davies, A. N. Bhuva, C. H. Manisty, J. B. Augusto, J. C. Moon, N. Aung, A. M. Lee, M. M. Sanghvi, K. Fung, J. M. Paiva, S. E. Petersen, E. Lukaschuk, S. K. Piechnik, S. Neubauer, and D. Rueckert, **Improving the Generalizability of Convolutional Neural Network-Based Segmentation on CMR Images**, Frontiers in Cardiovascular Medicine, vol. 7, p. 105, 2020 [7].

5. **C. Chen**, C. Qin, H. Qiu, G. Tarroni, J. Duan, W. Bai, and D. Rueckert, **Realistic Adversarial Data Augmentation for MR Image Segmentation**, in Medical Image Computing and Computer Assisted Intervention - MICCAI 2020, 2020, pp. 667–677 [15].

6. **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, p. 25, 2020 [1].

7. **C. Chen**, K. Hammernik, C. Ouyang, C. Qin, W. Bai, and D. Rueckert, **Cooperative Training and Latent Space Data Augmentation for Robust Segmentation**, in Medical Image Computing and Computer Assisted Intervention - MICCAI 2021, 2021 [16].

Other works and collaborations

Here, an extended list of publications is provided, in which I was involved as a collaborator during my Ph.D. study. These works also relate to the topics and aims of this thesis, though they are not included in this thesis.

1. W. Bai, **C. Chen**, G. Tarroni, J. Duan, F. Guitton, S. E. Petersen, Y. Guo, P. M. Matthews, and D. Rueckert, **Self-Supervised Learning for Cardiac MR Image Segmentation by Anatomical Position Prediction**, in Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, 2019, pp. 541-549 [17].

2. E. P. V. Le, N. R. Evans, J. M. Tarkin, M. M. Chowdhury, F. Zaccagna, C. Wall, Y. Huang, J. R. Weir-Mccall, **C. Chen**, E. A. Warburton, C. B. Schonlieb, E. Sala, and J. H. F. Rudd, **Contrast CT Classification of Asymptomatic and Symptomatic carotids in Stroke and Transient Ischaemic Attack with Deep Learning and Interpretability**, European Heart Journal, vol. 41, 2020 [18].

3. C. Ouyang, C. Biffi, **C. Chen**, T. Kart, H. Qiu, and D. Rueckert, **Self-supervision with Superpixels: Training Few-Shot Medical Image Segmentation Without Annotation**, in European Conference on Computer Vision, 2020, pp. 762-780 [19].

4. E. Puyol-Antón, **C. Chen**, J. R. Clough, and B. Ruijsink, **Interpretable Deep Models for Cardiac Resynchronisation Therapy Response Prediction**, in Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, 2020, pp. 284-293 [20].

5. C. Qin, S. Wang, **C. Chen**, H. Qiu, W. Bai, and D. Rueckert, **Biomechanics-Informed Neural Networks for Myocardial Motion Tracking in MRI**, in Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, 2020, pp. 296–306 [21].

6. S. Wang, G. Tarroni, C. Qin, Y. Mo, C. Dai, **C. Chen**, B. Glocker, Y. Guo, D. Rueckert, and W. Bai, **Deep Generative Model-Based Quality Control for Cardiac MRI Segmentation**, in Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, 2020, pp. 88–97 [22].

7. Z. Xiong, Q. Xia, Z. Hu, N. Huang, C. Bian, Y. Zheng, S. Vesal, N. Ravikumar, A. Maier, X. Yang, P.-A. Heng, D. Ni, C. Li, Q. Tong, W. Si, E. Puybareau, Y. Khoudli, T. Géraud, **C. Chen**, W. Bai, D. Rueckert, L. Xu, X. Zhuang, X. Luo, S. Jia, M. Sermesant, Y. Liu, K. Wang, D. Borra, A. Masci, C. Corsi, C. de Vente, M. Veta, R. Karim, C. J. Preetha, S. Engelhardt, M. Qiao, Y. Wang, Q. Tao, M. Nuñez-Garcia, O. Camara, N. Savioli, P. Lamata, and J. Zhao, **A Global Benchmark of Algorithms for Segmenting the Left Atrium from Late Gadolinium-enhanced Cardiac Magnetic Resonance Imaging**, Medical Image Analysis, vol. 67, p. 101832, 2021 [23].

8. S. Wang, C. Qin, N. Savioli, **C. Chen**, D. O'Regan, S. Cook, Y. Guo, D. Rueckert, and W. Bai, **Joint Motion Correction and Super Resolution for Cardiac Segmentation via Latent Optimisation**, in Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, 2021 [24].

9. X. Zhuang, J. Xu, X. Luo, **C. Chen**, C. Ouyang, D. Rueckert, V. M. Campello, K. Lekadir, S. Vesal, N. RaviKumar, Y. Liu, G. Luo, J. Chen, H. Li, B. Ly, M. Sermesant, H. Roth, W. Zhu, J. Wang, X. Ding, X. Wang, S. Yang, and L. Li, **Cardiac Segmentation on Late Gadolinium Enhancement MRI: A Benchmark Study from Multi-Sequence Cardiac MR Segmentation Challenge**, arXiv Preprint, 2020, submitted to Medical Image Analysis [11].

Chapter 2

Background

This chapter contains material from

1. C. Chen, C. Qin, H. Qiu, G. Tarroni, J. Duan, W. Bai *et al.*, 'Deep learning for cardiac image segmentation: A review,' *Frontiers in Cardiovascular Medicine*, vol. 7, p. 25, Mar. 2020, ISSN: 2297-055X. DOI: 10.3389/fcvm.2020.00025 [1]

2.1 Fundamentals of deep learning

Deep learning models are essentially deep artificial neural networks. Each neural network consists of an input layer, an output layer, and multiple hidden layers. In the following section, we will review several common deep learning networks and key techniques that have been commonly used in state-of-the-art DL-based medical imaging applications. We then will briefly review the recent developments of deep learning for cardiac MR segmentation. Finally, we will discuss the limitations of deep learning as well as theories and practical techniques to improve model generalization.

2.1.1 Deep neural networks

A deep neural network is an artificial neural network (ANN) with multiple layers ($n > 2$) between the input and output layers, which allows itself to model complex non-linear relationships in data. In this section, we first introduce basic building blocks in neural networks and then introduce several commonly used deep neural networks in image analysis. The basic building blocks of deep neural networks are:

- Convolution layers: A convolutional layer consists of a set of small filters with learnable weights and biases. Each filter in a convolutional layer is only connected to a small region of the input volume each time. By sliding across the whole input volume along the width and height and computing the dot product between the filter weights and the input volume plus bias offsets, a convolution layer produces a set of feature maps (activation maps). These activation maps correspond to the response of the convolutional filters at each spatial position of the input. For example, given a convolutional layer with k_{out} 2D $n \times n$ convolution kernels and an input image $\mathbf{x}_{in} \in \mathbb{R}^{H \times W \times k_{in}}$, the computation can be formulated as:

$$\forall i \in (1, k_{out}),\ \mathbf{x}_{out}^{(i)} = \mathbf{w}^{(i)} \circ \mathbf{x}_{in} + b^{(i)}, \tag{2.1}$$

  where $\mathbf{w}^{(i)} \in \mathbb{R}^{n \times n \times k_{in}}, b^{(i)} \in \mathbb{R}$ represent the weights and bias parameters in the i^{th} convolution kernel respectively, $\circ$ represents the convolution operation (i.e. dot products between the filters and local regions of the input), $\mathbf{x}_{out} \in \mathbb{R}^{H' \times W' \times k_{out}}$ represents the output feature maps. H', W' are determined by the size of the kernel n, the stride s, the amount of zero padding p and the input height H and width W respectively: $H' = (H - n + 2p)/s + 1, W' = (W - n + 2p)/s + 1$.

- Activation layers: Activation layers are nonlinear transformation functions, which transform input values to fall within an acceptable and useful range. In deep learning, the most commonly used activation function is the rectified linear unit (ReLU) function, which preserves the value of non-negative inputs and assigns zeros to negative inputs. The rectifier

function is given below:

$$f(x) = \begin{cases} x & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases} . \tag{2.2}$$

Two main advantages of ReLU are its simplicity and its computational efficiency. Compared to the other two commonly used sigmoid and tanh activation functions, the gradient calculation for ReLU is much simple. For non-negative inputs, the gradients are all 1s, whereas for negative inputs, the gradients are all 0s. It can, therefore, significantly reduce computational time at network training.

- Pooling layer: Pooling layers are used to reduce the spatial size of features and, more importantly, remove/suppress redundant features for improved generalization. One of the most commonly used pooling layers is Max Pooling. Max Pooling partitions the input into a set of non-overlapping regions and then returns the maximum value for each sub-region.

- Fully connected layers: A fully connected layer contains a set of neurons where each of them has *full* connections to its inputs. Given a set of features, it performs matrix multiplication plus bias offsets to compute activation maps.

Apart from the above basic layers, there is another family of layers called normalization layers, which are used to standardize the statistics of inputs to layers. A normalization layer is generally inserted between a convolution layer and its subsequent activation layer. By gently restricting the distributions of inputs to layers in a deep network, it can help the network to produce better gradients for weight update, thus alleviating the gradient explosion and vanishing problems during the network optimization [25]. Without normalization layers, training deep neural networks with tens of layers is challenging and time-consuming as networks can be very sensitive to the initial random weights and the change in the distribution of network activations during training. Several commonly used normalization layers includes batch normalization [25], layer normalization [26], and instance normalization [27], which normalize inputs batch-wise, layer-wise, and instance-wise, respectively.

2.1.1.1 Convolutional neural networks (CNNs)

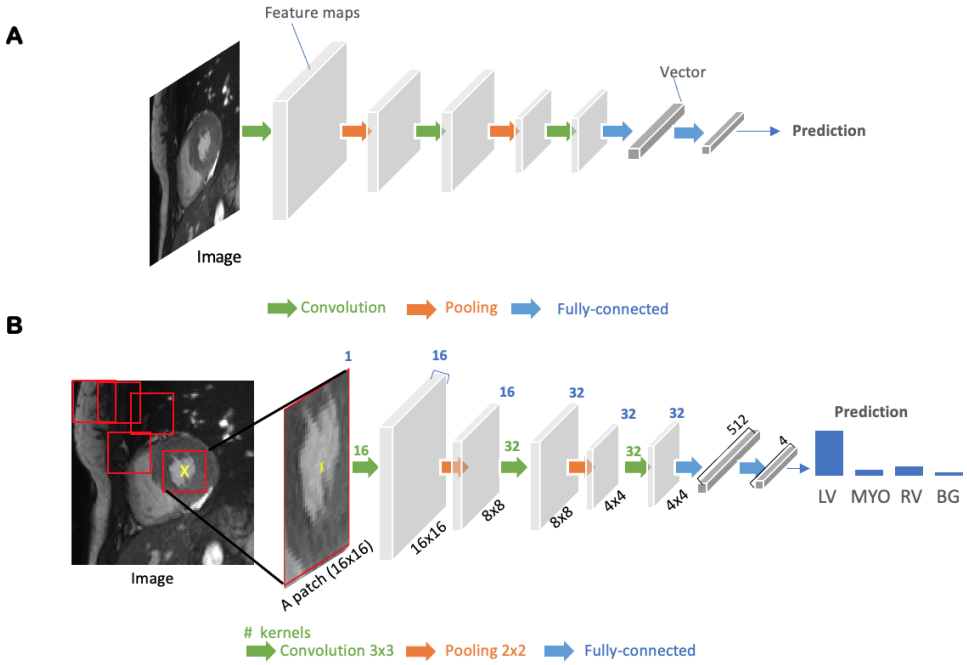


Figure 2.1: **(A) Generic architecture of convolutional neural networks.** A CNN takes a cardiac MR image as input, learning hierarchical features through a stack of convolutions and pooling operations. These spatial feature maps are then flattened and reduced into a vector through fully connected layers. This vector can be in many forms, depending on the specific task. It can be probabilities for a set of classes (image classification) or coordinates of a bounding box (object localization) or a predicted label for the center pixel of the input (patch-based segmentation), or a scalar for regression tasks, e.g., left ventricular volume estimation. **(B) Patch-based segmentation method based on a CNN classifier.** The CNN takes a patch as input and outputs the probabilities for four classes. The class with the highest score is the prediction for the center pixel (see the yellow cross) in this patch. By repeatedly forwarding patches located at different locations into the CNN for classification, one can finally get a pixel-wise segmentation map for the whole image. LV: left ventricle cavity; RV: right ventricle cavity; BG: Background; MYO: left ventricular myocardium. The blue number at the top indicates the number of channels of the feature maps. Each convolution kernel is a 3x3 kernel (stride=1, padding=1), producing an output feature map with the same height and width as the input.

In this part, we will introduce convolutional neural networks (CNNs), which are the most common type of deep neural networks for image analysis. CNNs have been successfully applied to advance the state-of-the-art on many image classification, object detection and segmentation tasks. As shown in Fig. 2.1A, a standard CNN consists of an input layer, an output layer, and a stack of functional layers in between that transform an input into an output in a specific form, e.g., vectors. These functional layers often contain convolutional layers, pooling layers, and/or fully connected layers. In general, a convolutional layer CONV_l contains k_l convolution kernels/filters, which is followed by a normalization layer, (e.g., batch normalization [25]), and

a nonlinear activation function (e.g., ReLU) to extract k_l feature maps from the input. These feature maps are then down-sampled by pooling layers, typically by a factor of 2, which remove redundant features to improve the statistical efficiency and model generalization. After that, fully connected layers are applied to reduce the dimension of features from its previous layer and find the most task-relevant features for inference. The output of the network is a fix-sized vector where each element can be a probabilistic score for each category (for image classification), a real value for a regression task, e.g., the left ventricular volume estimation, or a set of values, e.g., the coordinates of a bounding box for object detection and localization.

A key component of CNNs is the convolutional layer. Each convolutional layer has k_l convolution kernels to extract k_l feature maps and the size of each kernel n is chosen to be small in general, e.g., $n = 3$ for a 2D 3×3 kernel, to reduce the number of parameters[1]. While the kernels are small, one can increase the receptive field [2] by increasing the number of convolutional layers. For example, a convolutional layer with large 7×7 kernels can be replaced by three layers with small 3×3 kernels [28]. The number of weights is reduced by a factor of $7^2/(3 \times (3^2)) \approx 2$ while the receptive field remains the same (7×7). In general, increasing the depth of convolution neural networks (the number of hidden layers) to enlarge the receptive field can lead to improved model performance, e.g., classification accuracy [28].

CNNs for image classification can also be employed for image segmentation applications without major adaptations to the network architecture [29], as shown in Fig. 2.1B. However, this requires an additional step to divide each image into patches and then train a CNN to predict the class label of the center pixel for every patch. One major disadvantage of this patch-based approach is that, at inference time, the network has to be deployed for every patch individually despite the fact that there is a lot of redundancy due to multiple overlapping patches in the image. As a result of this inefficiency, the main application of CNNs with fully connected layers is object localization, which aims to estimate the bounding box of the object of interest in an image. This bounding box is then used to crop the image, forming an image

[1] In a convolution layer l with k_l 2D $n \times n$ convolution kernels and a l_{in}-channel input, the number of parameters in a convolutional layer is $k_l \times (n^2 \times l_{in} + 1)$. For a convolutional layer with 16 3×3 filters where the input is a $28 \times 28 \times 1$ 2D gray image, the number of parameters in this layer is $16 \times (3^2 \times 1 + 1) = 160$.

[2] The receptive field is the input image area that potentially impacts the activation of a particular convolutional kernel/neuron.

pre-processing step to reduce the computational cost for segmentation [30]. For efficient, end-to-end pixel-wise segmentation, a variant of CNNs called fully convolutional neural network (FCN) is more commonly used, which will be discussed in the next section.

2.1.1.2 Fully convolutional neural networks (FCNs)

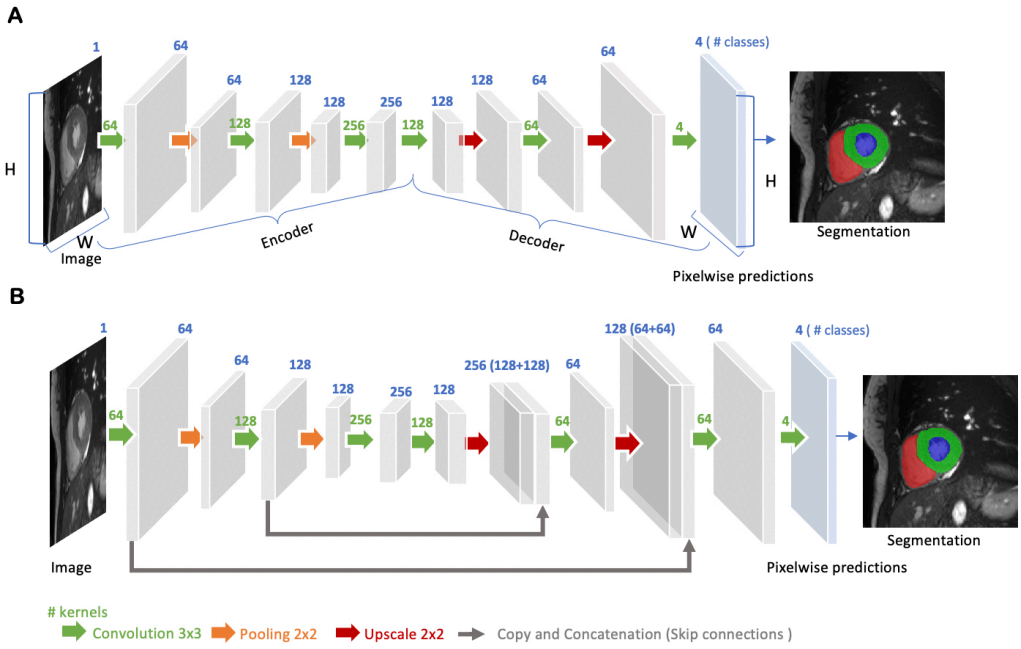


Figure 2.2: **(A) Architecture of a fully convolutional neural network (FCN).** The FCN first takes the whole image as input, learns image features through the encoder, gradually recovers the spatial dimension by a series of upscaling layers (e.g., transposed convolution layers, unpooling layers) in the decoder, and then produces pixel-wise probabilistic maps to predict regions of the left ventricle cavity (blue region), the left ventricular myocardium (green region) and the right ventricle cavity (red region). The final segmentation map is obtained by assigning each pixel with the class of the highest probability. One use case of this FCN-based cardiac segmentation can be found in [31]. **(B) Architecture of a U-net**. On the basis of FCN, U-net adds 'skip connections' (gray arrows) to aggregate feature maps from coarse to fine through concatenation and convolution operations. For simplicity, we reduce the number of downsampling and upsampling blocks in the diagram. For detailed information, we refer readers to the original paper [32].

The idea of FCN was first introduced by [33] for image segmentation. FCNs are a special type of CNNs that do not have any fully connected layers. In general, as shown in Fig. 2.2A, FCNs are designed to have an encoder-decoder structure such that they can take inputs of arbitrary size and produce an output with the same size. Given an input image, the encoder first transforms the input into a high-level feature representation, whereas the decoder interprets the feature maps and recovers spatial details back into the image space for pixel-wise prediction

through a series of upsampling and convolution operations. Here, upsampling can be achieved by applying transposed convolutions, e.g., 3×3 transposed convolutional kernels with a stride of 2 to up-scale feature maps by a factor of 2. These transposed convolutions can also be replaced by unpooling layers and upsampling layers. Compared to a patch-based CNN for segmentation, FCN is trained and applied to the entire images, removing the need for patch selection [34].

FCNs with the simple encoder-decoder structure in Fig. 2.2A may be limited in their ability to capture detailed contextual information in an image for precise segmentation as some features may be eliminated by the pooling layers in the encoder. Several variants of FCNs have been proposed to propagate features from the encoder to the decoder in order to boost the segmentation accuracy. The most well-known and most widespread variant of FCNs for biomedical image segmentation is the U-net [32]. On the basis of the vanilla FCN [33], the U-net employs skip connections between the encoder and decoder to recover spatial context loss in the down-sampling path, yielding more precise segmentation (see Fig. 2.2B). Several state-of-the-art medical image segmentation methods have adopted the U-net or its 3D variants, the 3D U-net [35] and the 3D V-net [36], as their backbone networks, achieving promising segmentation accuracy [37–39].

2.1.1.3 Recurrent neural networks (RNNs)

Recurrent neural networks (RNNs) are another type of neural networks which are used for sequential data, such as cine magnetic resonance imaging (MRI) and ultrasound image sequences. An RNN can 'remember' the past and use the knowledge learned from the past to make its present decision, see Fig 2.3A and B. For example, given a sequence of images, an RNN takes the first image as input, captures the information to make a prediction, and then memorize this information which is then utilized to make a prediction for the next image. The two most widely used architectures in the family of RNNs are long-short term memory (LSTM) [41] and gated recurrent unit (GRU) [42], which are capable of modeling long-term memory. A use case for cardiac segmentation is to combine an RNN with a 2D FCN so that the combined network is capable of capturing information from adjacent slices to improve the inter-slice coherence of

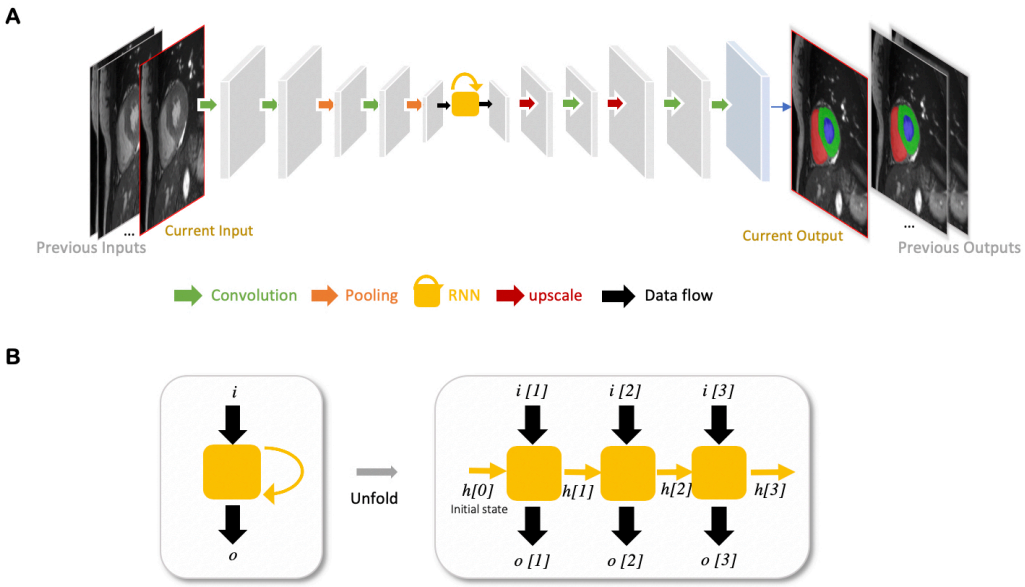


Figure 2.3: **(A) Example of an FCN with an RNN for cardiac image segmentation.** The yellow block with a curved arrow represents an RNN module, which utilizes the knowledge learned from the past to make the current decision. In this example, the network is used to segment cardiac ventricles from a stack of 2D cardiac MR slices, which allows the propagation of contextual information from adjacent slices for better inter-slice coherence [40]. This type of RNN is also suitable for sequential data such as cine MR images and ultrasound movies to learn temporal coherence. **(B) Unfolded schema of the RNN module for visualizing the inner process when the input is a sequence of three images.** Each time, this RNN module will receive an input $i[t]$ at time step t, and produce an output $o[t]$, considering not only the input information but also the hidden state ('memory') $h[t-1]$ from the previous time step $t-1$.

segmentation results [40].

2.1.1.4 Autoencoders (AE)

Autoencoders (AEs) are a type of neural networks that are designed to learn compact latent representations from data without supervision. A typical architecture of an autoencoder consists of two networks: an encoder network and a decoder network for the reconstruction of the input, see Fig. 2.4. Since the learned representations contain generally useful information in the original data, many researchers have employed autoencoders to extract general semantic features or shape information from input images or labels and then use those features to guide the medical image segmentation [43, 47, 48].

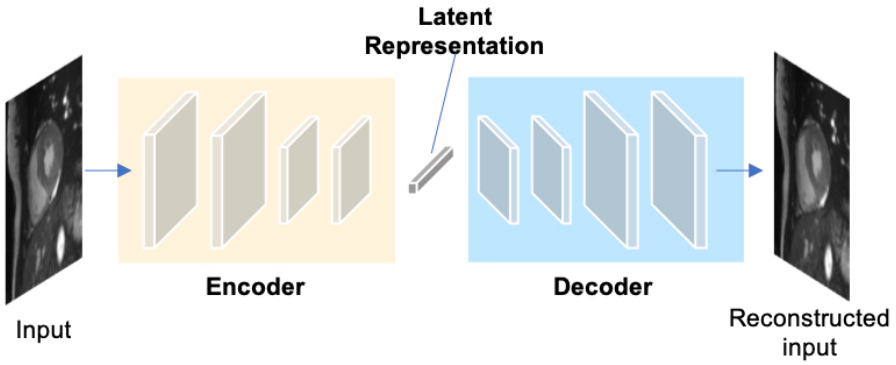


Figure 2.4: **Generic architecture of an autoencoder.** An autoencoder employs an encoder-decoder structure. The encoder maps the input data to a low-dimensional latent representation. The decoder interprets the code and reconstructs the input. The learned latent representation has been found effective for cardiac image segmentation [43, 44], cardiac shape modeling [45] and cardiac segmentation correction [46].

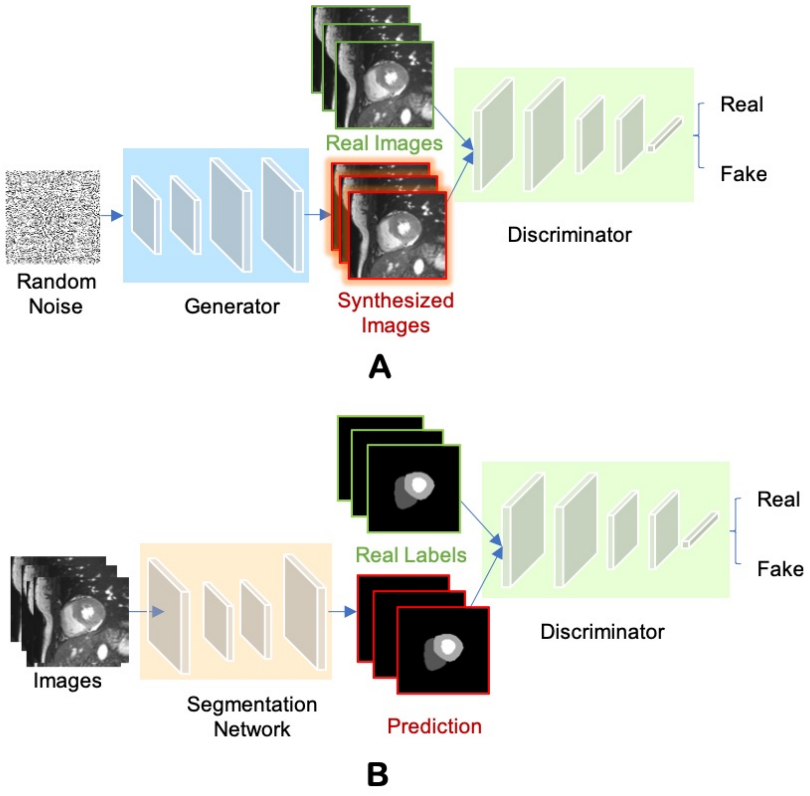


Figure 2.5: **GAN and adversarial training**. (A) Overview of GAN for image synthesis; (B) Overview of adversarial training for image segmentation.

2.1.1.5 Generative adversarial networks (GAN)

The concept of generative adversarial network (GAN) was proposed by [49] for image synthesis from noise. GANs are a type of generative models that learn to model the data distribution of real data and thus are able to create new image examples. As shown in Fig. 2.5A, a GAN

consists of two networks: a generator network and a discriminator network. During training, the two networks are trained to compete against each other: the generator produces fake images aimed at fooling the discriminator, whereas the discriminator tries to distinguish real images from fake ones. This type of training is referred to as 'adversarial training', since the two models are both set to win the competition. This training scheme can also be used for training a segmentation network. As shown in Fig. 2.5B, the generator is replaced by a segmentation network and the discriminator is required to distinguish the generated segmentation maps from the ground truth ones (the target segmentation maps). In this way, the segmentation network is encouraged to produce more anatomically plausible segmentation maps [50, 51].

2.1.1.6 Advanced building blocks for improved segmentation

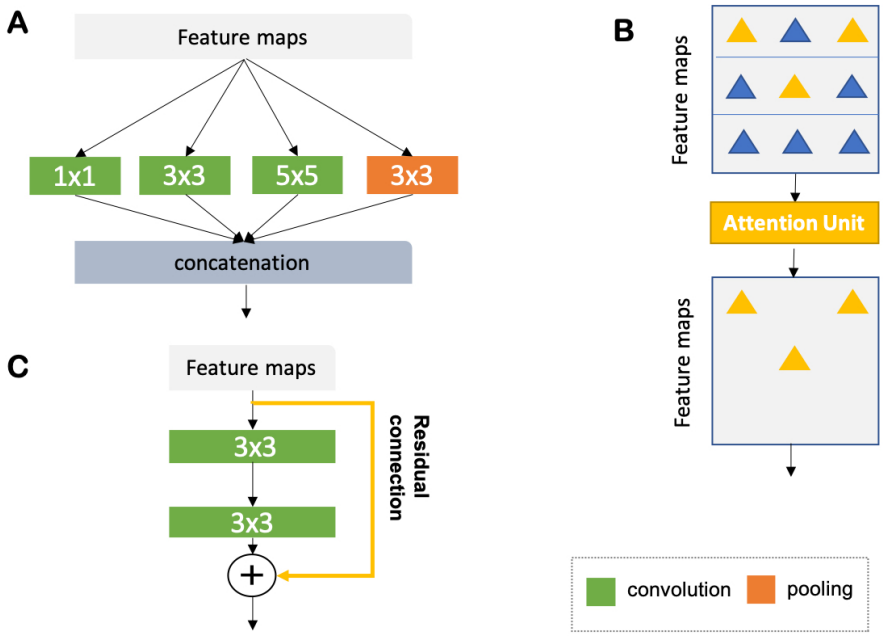


Figure 2.6: **(A) Naive version of the inception module [25].** In this module, convolutional kernels with varying sizes are applied to the same input for multi-scale feature fusion. On the basis of the naive structure, a family of advanced inception modules with more complex structures have been developed [52, 53]. **(B) Schematic diagram of the attention module [54, 55].** The attention module teaches the network to pay attention to important features (e.g., features relevant to anatomy) and ignore redundant features. **(C) Schematic diagram of a residual unit [56].** The yellow arrow represents a residual connection, which is applied to reusing the features from a previous layer. The numbers in the green and orange blocks denote the sizes of corresponding convolutional or pooling kernels. Here, for simplicity, all diagrams have been reproduced based on the illustration in the original papers.

Medical image segmentation, as an important step for quantitative analysis and clinical re-

search, requires pixel-level accuracy. Over the past years, many researchers have developed advanced building blocks to learn robust, representative features for precise segmentation. These techniques have been widely applied to state-of-the-art neural networks (e.g., U-net) to improve medical image segmentation performance. Therefore, we identified several important techniques reported in the literature and present them with corresponding references for further reading. These techniques are:

1. Advanced convolutional modules for multi-scale feature aggregation:
   - Inception modules [25, 52, 53], which concatenate multiple convolutional filter banks with different kernel sizes to extract multi-scale features in parallel, see Fig. 2.6A;
   - Dilated convolutional kernels [57], which are modified convolution kernels with the same kernel size but different kernel strides to process input feature maps at different scales;
   - Deep supervision [58], which utilizes the outputs from multiple intermediate hidden layers for multi-scale prediction;
   - Atrous spatial pyramid pooling [59], which applies spatial pyramid pooling [60] with various kernel strides to input feature maps for multi-scale feature fusion;
2. Adaptive convolutional kernels designed to pay attention to important features:
   - Attention units [54, 55, 61], which learn to adaptively recalibrate features spatially, see Fig. 2.6B;
   - Squeeze-and-excitation blocks [62], which are used to recalibrate features with learnable weights across channels;
3. Interlayer connections designed to reuse features from previous layers:
   - Residual connections [56], which add outputs from a previous layer to the feature maps learned from the current layer, see Fig. 2.6C;
   - Dense connections [63], which concatenate outputs from all preceding layers to the feature maps learned from the current layer.

2.1.2 Training Neural Networks

Before being able to perform inference, neural networks must be trained. The standard training process requires a dataset that contains paired images and labels for training and testing, an optimizer (e.g., stochastic gradient descent (SGD) [64], adaptive moment estimation (Adam) [65]) and a loss function to update the model parameters. This function accounts for the error of the network prediction in each iteration during training, providing signals for the optimizer to update the network parameters through back-propagation [66]. The goal of training is to find proper values of the network parameters that minimize the loss function.

Mathematically, we can formulate it as a minimization problem. Given a neural network f with a set of learnable parameters θ (e.g., weights $\boldsymbol{w}$ and biases $\boldsymbol{b}$ in convolutional layers), the learning goal is to find optimal θ^*, so that the expected loss over the joint distribution $\mathrm{P}(\mathbf{X}, \mathbf{Y})$ is minimized:

$$\theta^* = \arg\min_{\theta} \mathcal{L}_{exp} = \arg\min_{\theta} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim \mathrm{P}(\mathbf{X},\mathbf{Y})}\, \mathcal{L}[\boldsymbol{y}, f(\boldsymbol{x};\theta)], \tag{2.3}$$

where $\boldsymbol{x} \in \mathbf{X}$ is an input image and $\boldsymbol{y} \in \mathbf{Y}$ is a corresponding target label.

Since the true joint distribution of $\mathrm{P}(\mathbf{X}, \mathbf{Y})$ is unknown, in practice, we instead find θ that minimizes the empirical loss/risk computed on a given dataset (e.g., training set) $\mathcal{D}_{tr}$ to find an approximate solution $\hat{\theta}$ of θ^*:

$$\hat{\theta} = \arg\min_{\theta} \mathcal{L}_{emp} = \arg\min_{\theta} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim \mathcal{D}_{tr}}\, \mathcal{L}[\boldsymbol{y}, f(\boldsymbol{x};\theta)]. \tag{2.4}$$

The above learning objective is also known as *empirical risk minimization (ERM)* [67] in the statistical learning theory, which states that the learning algorithm should choose a hypothesis that minimizes the empirical risk.

2.1.2.1 Back-propagation

A core element in the network learning process is the backpropagation (BP) algorithm [66], which adjusts the parameters to minimize the training loss during network training. At a high

level, BP computes the gradients from the very last layer to the earlier layers layer by layer and then employs gradient descent to update the associated weights in the direction to minimize the error between the actual outputs from the network and the desired outputs functions. Specifically, the learning process consists of four steps:

1. forward the input data $\boldsymbol{x}$ to the network $f(\cdot;\theta)$ parameterised by θ, and then compute predictions $f(\boldsymbol{x};\theta)$;

2. compute the errors $\mathcal{L}[\boldsymbol{y}, f(\boldsymbol{x};\theta)]$ between the desired outputs $\boldsymbol{y}$ and the network outputs $f(\boldsymbol{x};\theta)$;

3. backpropagate the errors from the final layers to previous layers by repeatedly applying chain rule to computing the gradients of the loss/errors with respect to the trainable parameters $\nabla_\theta \mathcal{L}$ layer by layer;

4. choose a gradient descent algorithm, e.g., SGD to update those parameters θ: $\theta \leftarrow \theta - \lambda \nabla_\theta \mathcal{L}$ where λ is the step size.

2.1.2.2 Common loss functions

There are several different common loss functions $\mathcal{L}$ to choose from. For regression tasks (e.g., heart localization, calcium scoring, landmark detection, image reconstruction), the simplest loss function is the mean squared error (MSE):

$$\mathcal{L}_{\text{MSE}} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i)^2, \tag{2.5}$$

where $\boldsymbol{y}$ is the vector of target values and $\hat{\boldsymbol{y}} = f(\boldsymbol{x};\theta)$ is the vector of the predicted values. The subscript i specifies the i-th element in the corresponding vector, and n is the total length of each vector.

Cross-entropy is the most common loss for both image classification and segmentation tasks

where the network produces the probability for each class rather than class labels[3]. In particular, the cross-entropy loss for segmentation summarizes the pixel-wise probability errors between the predicted probabilistic output from the network after softmax $\boldsymbol{p}^{(c)} = \frac{e^{f(\boldsymbol{x};\theta)^{(c)}}}{\sum_{d=1}^{C} e^{f(\boldsymbol{x};\theta)^{(d)}}}$ and its corresponding target one-hot segmentation map $\boldsymbol{y}^{(c)}$ for each class c:

$$\mathcal{L}_{\text{CE(segmentation)}} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{C} \boldsymbol{y}_i^{(c)} \log(\boldsymbol{p}_i^{(c)}), \tag{2.6}$$

where C is the number of all classes and n is the number of pixels in the corresponding image. For image-level classification tasks, the loss can be simplified by removing the pixel-wise summation: $\mathcal{L}_{\text{CE (classification)}} = -\sum_{c=1}^{C} \boldsymbol{y}^{(c)} \log(\boldsymbol{p}^{(c)})$.

Another loss function which is specifically designed for object segmentation is called soft-Dice loss function [36], which penalizes the mismatch between a predicted segmentation map and its target map at pixel-level:

$$\mathcal{L}_{Dice} = 1 - \frac{2\sum_{i=1}^{n}\sum_{c=1}^{C} \boldsymbol{y}_i^{(c)} \boldsymbol{p}_i^{(c)}}{\sum_{i=1}^{n}\sum_{c=1}^{C} (\boldsymbol{y}_i^{(c)} + \boldsymbol{p}_i^{(c)})}. \tag{2.7}$$

In addition, there are several variants of the cross-entropy and soft-Dice loss such as the weighted cross-entropy loss [68–70] and weighted soft-Dice loss [71], which are used to address potential class imbalance problem in medical image segmentation tasks where the loss term is weighted to account for rare classes or small objects. Specifically, the weighted cross-entropy loss is defined as:

$$\mathcal{L}_{\text{weighted CE}} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{C} w^{(c)} \boldsymbol{y}_i^{(c)} \log(\boldsymbol{p}_i^{(c)}), \tag{2.8}$$

where $w^{(c)}$ is a scalar, specifying the weight for the loss term associated with the class c. In practice $w^{(c)}$ for a rare class is set to a higher value than the one for the majority class.

[3] At inference time, the predicted segmentation map for each image is obtained by assigning each pixel with the class of the highest probability: $\hat{\boldsymbol{y}}_i = \arg\max_c \boldsymbol{p}_i^c$.

2.2 Applications of deep learning: cardiac MR image segmentation

In this section, we provide a literature review on recent developments of deep learning-based applications for medical imaging, with a particular focus on one of the most commonly used main imaging modalities: cardiac MRI. Cardiac MRI is a non-invasive imaging technique that can visualize the structures within and around the heart. Compared to computed tomography (CT), it does not require ionizing radiation. Instead, it relies on the magnetic field in conjunction with radio-frequency waves to excite hydrogen nuclei in the heart and then generates an image by measuring their response. By utilizing different imaging sequences, cardiac MRI allows accurate quantification of both cardiac anatomy and function (e.g., using cine imaging) and pathological tissues such as scars (e.g., using LGE imaging). Accordingly, cardiac MRI is currently regarded as the gold standard for quantitative cardiac analysis [72].

In the following, we summarize the recent developments of cardiac image segmentation in magnetic resonance (MR) imaging, with a particular focus on cardiac ventricle segmentation, where the deep learning techniques have been heavily adopted in.

2.2.1 Vanilla FCN-based segmentation

Tran was among the first ones to apply a FCN [34] to segment the left ventricle, myocardium, and right ventricle directly on short-axis cardiac MR images. Their end-to-end approach based on FCN achieved competitive segmentation performance, significantly outperforming traditional methods in terms of both speed and accuracy. In the following years, a number of works based on FCNs have been proposed, aiming at achieving further improvements in segmentation performance. In this regard, one stream of work focuses on optimizing the network structure to enhance the feature learning capacity for segmentation [38, 68, 71, 73–77]. For example, Isensee *et al.* developed a residual U-net to combine multi-scale features for robust segmentation across images with large anatomical variability, see Fig. 2.7. Several works [68, 70, 78, 79] investigated different loss functions such as weighted cross-entropy, weighted Dice loss, deep supervision loss

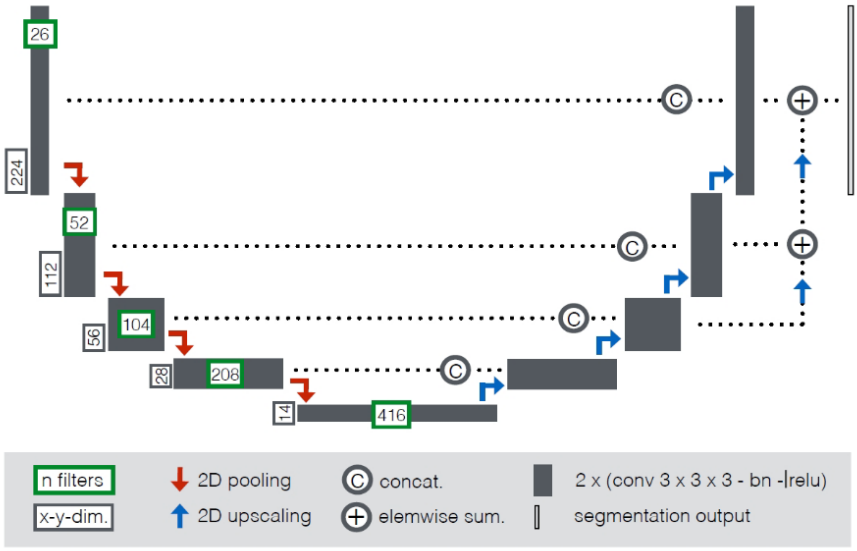


Figure 2.7: **Architecture of a residual U-Net with long-range concatenations and short-range residual connections.** Image source: [38], reproduced with permission from Springer Nature.

and focal loss to improve the segmentation performance. Among these FCN-based methods, the majority of approaches use 2D networks rather than 3D networks for segmentation. This preference is mainly due to the typical low through-plane resolution and motion artifacts of most cardiac MR scans, which limits the applicability of 3D networks [69].

2.2.2 Introducing spatial or temporal context

One drawback of using 2D networks for cardiac segmentation is that these networks work slice by slice, and thus they do not leverage any inter-slice dependencies. As a result, 2D networks can fail to locate and segment the heart on challenging slices such as apical and basal slices where the contours of the ventricles are not well defined. To address this problem, several works have attempted to introduce additional contextual information to guide 2D FCN. This contextual information can include shape priors learned from labels or multi-view images [13, 80, 81]. Others extract spatial information from adjacent slices to assist the segmentation, using recurrent units (RNNs) or multi-slice networks (2.5D networks) [40, 82–84]. These networks can also be applied to leveraging information across different time frames in the cardiac cycle to improve spatial and temporal consistency of segmentation results [83, 85–88].

2.2.3 Applying anatomical constraints

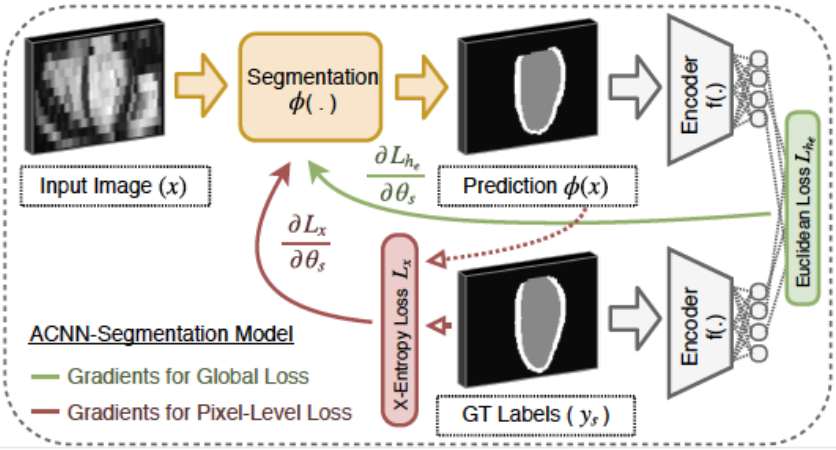


Figure 2.8: **Architecture of the anatomical constrained CNN (ACNN) described in [44].** The segmentation network $\Phi(;\theta)$ is trained to minimize a cross-entropy loss $\mathcal{L}_x$ and a Euclidean distance loss $\mathcal{L}_{h_e}$ measured on the latent spaces of the predicted label and ground truth. Image source [44], licensed under CC BY 4.0[a].

[a]http://creativecommons.org/licenses/by/4.0/

Another problem that may limit the segmentation performance of both 2D and 3D FCNs is that they are typically trained with pixel-wise loss functions only (e.g., cross-entropy or soft-Dice losses). These pixel-wise loss functions may not be sufficient to learn features that represent the underlying anatomical structures. Therefore, several approaches focus on designing and applying anatomical constraints to train the network to improve its prediction accuracy and robustness. These constraints are represented as regularization terms, which take into account the topology [89], contour and region information [90] or shape information [44, 48], as a way to encourage the network to generate more anatomically plausible segmentations.

For example, Oktay *et al.* proposed a network called anatomically constrained neural networks (ACNN) to improve cardiac segmentation performance, see Figure. 2.8. An auto-encoder is introduced in their network to embed the labels and predicted segmentations into latent space. This design allows one to quantify the dissimilarity of the global shape structures between the labels and predictions. The network is trained to minimize a dissimilarity loss computed on the latent space and a cross-entropy loss. Their experimental results suggest that learning global anatomical properties of the underlying anatomy could improve the prediction accuracy of state-of-the-art models. In addition to regularizing networks at training time, Painchaud *et*

al. proposed a variational AE to correct inaccurate segmentations at post-processing.

2.2.4 Multi-task learning

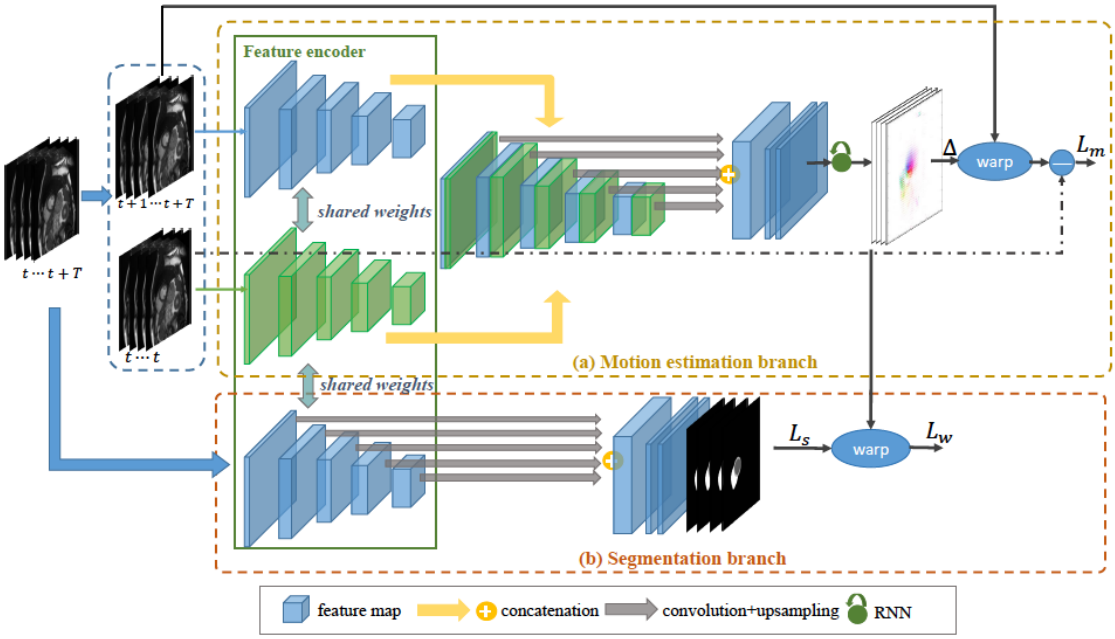


Figure 2.9: **Architecture of the multi-task learning network for joint estimation of cardiac motion and segmentation network.** It consists of two branches: a motion estimation branch and a segmentation branch. The two branches share the same feature encoder and are trained jointly. The motion estimation branch employs a Siamese-style recurrent multi-scale spatial transformer network to estimate the motion fields given MR image sequences. The segmentation branch is employed to predict segmentation simultaneously. The predicted segmentation for an unlabeled frame is wrapped to the labeled target frame using the motion fields estimated from the motion estimation branch for supervised learning[87]. Image source: [87], reproduced with permission of the rights holder, Springer Nature.

Multi-task learning has also been explored to regularize FCN-based cardiac ventricle segmentation during training by performing auxiliary tasks that are relevant to the main segmentation task, such as motion estimation [91], estimation of cardiac function [92], ventricle size classification [93] and image reconstruction [94–96]. Training a network for multiple tasks simultaneously encourages the network to extract features that are useful across these tasks, resulting in improved learning efficiency and prediction accuracy.

For example, Qin *et al.* proposed a joint learning method to estimate motion and segment-

ation for cardiac MR image sequences simultaneously (see figure 2.9). The motion estimation network and the segmentation network are jointly optimized by minimizing a composite loss function. This composite loss consists of an image dissimilarity loss, a smoothness penalty of motion fields, and pixel-wise cross-entropy segmentation losses. Motion information extracted from a large number of unlabeled images is used to improve their estimated segmentation results by encouraging their spatial-temporal smoothness in the same sequence. Their experimental results showed that with additional motion constraints, their segmentation accuracy was marginally improved in terms of dice (left ventricle blood pool (LV): from 0.92 to 0.93, left ventricular myocardium (MYO): from 0.84 to 0.86, right ventricular blood pool (RV): from 0.87 to 0.89) [87].

2.2.5 Multi-stage networks

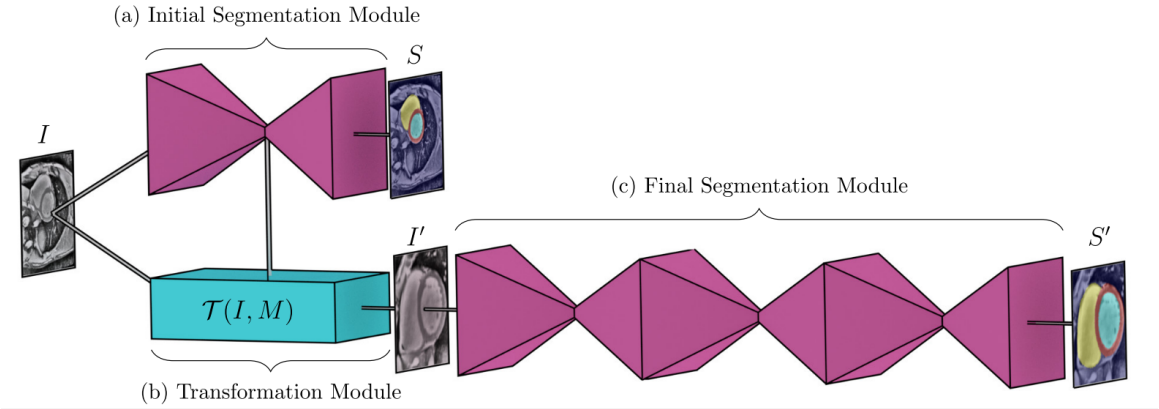


Figure 2.10: **Architecture of the Omega network.** Image source: [97], reproduced with permission of the rights holder, Elsevier.

Recently, there is a growing interest in applying neural networks in a multi-stage pipeline which breaks down the segmentation problem into subtasks [84, 97–100]. For example, Zheng *et al.*, Li *et al.* proposed a region-of-interest (ROI) localization network followed by a segmentation network. Likewise, Vigneault *et al.* proposed a network called Omega-Net, which consists of a U-net for cardiac chamber localization, a learnable transformation module to normalize image orientation, and a series of U-nets for fine-grained segmentation, see Fig. 2.10. By explicitly localizing the ROI and by rotating the input image into a canonical orientation, the proposed

method better generalizes to images with varying sizes and orientations.

2.2.6 Hybrid segmentation methods

Another stream of work aims to combine neural networks with classical segmentation approaches, e.g., level-sets [101, 102], deformable models [30, 103, 104], atlas-based methods [105, 106] and graph-cut based methods [107]. Here, neural networks are applied in the feature extraction and model initialization stages, reducing the dependency on manual interactions and improving the segmentation accuracy of the conventional segmentation methods deployed afterward. For example, Avendi *et al.* proposed one of the first deep learning (DL)-based methods for LV segmentation in cardiac short-axis MR images. The authors first applied a CNN to detect the LV automatically and then used an AE to estimate the shape of the LV. The estimated shape was then used to initialize follow-up deformable models for shape refinement. As a result, the proposed integrated deformable model converges faster than conventional deformable models, and the segmentation achieves higher accuracy. In their later work, the authors extended this approach to segment RV [103]. While these hybrid methods demonstrated better segmentation accuracy than previous non-deep learning methods, most of them still require an iterative optimization for shape refinement. Furthermore, these methods are often designed for one particular anatomical structure. As noted in the recent benchmark study [6], most state-of-the-art segmentation algorithms for bi-ventricle segmentation are based on end-to-end FCNs, which allows the simultaneous segmentation of the LV and RV.

2.2.7 Achievements that deep learning based approaches made for cardiac ventricle segmentation

To better illustrate these developments for cardiac ventricle segmentation from cardiac MR images, we collate a list of bi-ventricle segmentation methods that have been trained and tested on the Automated Cardiac Diagnosis Challenge (ACDC) dataset, reported in Table 2.1. For ease of comparison, we only consider those methods that were trained on the given training set

Table 2.1: **Segmentation accuracy of state-of-the-art segmentation methods verified on the cardiac bi-ventricular segmentation challenge dataset [6].** Bold numbers are the highest mean Dice values for the corresponding structure. LV: left ventricle cavity, RV: right ventricle cavity, MYO: left ventricular myocardium; ED: end-diastolic; ES: end-systolic.

Methods	Description	LV	MYO	RV
Isensee *et al.* [38]	2D U-net+3D U-net (ensemble)	**0.950**	**0.911**	0.923
Li *et al.* [98]	Two 2D FCNs for ROI detection and segmentation respectively;	0.944	**0.911**	**0.926**
Zotti *et al.* [81]	2D GridNet-MD with registered shape prior	0.938	0.894	0.910
Khened *et al.* [71]	2D Dense U-net with inception modules	0.941	0.894	0.907
Baumgartner *et al.* [69]	2D U-net with a cross-entropy loss	0.937	0.897	0.908
Zotti *et al.* [80]	2D GridNet with registered shape priors	0.931	0.890	0.912
Jang *et al.* [68]	2D M-Net with a weighted cross-entropy loss	0.940	0.885	0.907
Painchaud *et al.* [46]	FCN followed by an AE for shape correction	0.936	0.889	0.909
Wolterink *et al.* [88]	Multi-input 2D dilated FCN, segmenting paired ED and ES frames simultaneously	0.940	0.885	0.900
Patravali *et al.* [82]	2D U-net with a Dice loss	0.920	0.890	0.865
Rohé *et al.* [106]	Multi-atlas based method combined with 3D CNN for registration	0.929	0.868	0.881
Tziritas *et al.* [108]	Level-set+markov random field (MRF); *Non-deep learning method*	0.907	0.798	0.803
Yang *et al.* [70]	3D FCN with deep supervision	0.820	N/A	0.780

All the methods were evaluated on the same test set (50 subjects). Note that for simplicity, we report the average Dice scores for each structure over ED and ES phases. More detailed comparison for different phases can be found on the public leaderboard in the post testing part (`https://acdc.creatis.insa-lyon.fr`) as well as corresponding published works in this table. Last update: 2019.8.1.

(100 subjects) and have been evaluated on the same online test set (50 subjects). As the ACDC challenge organizers keep the online evaluation platform open to the public, our comparison includes not only the methods from the original challenge participants (summarized in the benchmark study paper from Bernard *et al.*[6]) but also three segmentation algorithms that have been proposed after the challenge (i.e. [46, 81, 98]). The Dice metric is used for comparison. The Dice score measures the ratio of overlap between two results (e.g., automatic segmentation vs. manual segmentation), ranging from 0 (mismatch) to 1 (perfect match). It is important to note that the segmentation accuracy of different methods cannot be directly comparable in general unless these methods are evaluated on the same dataset. This is because, even for the same segmentation task, different datasets can have different imaging modalities, different patient populations, and different methods of image acquisition, which will affect the task complexities and result in different segmentation performances.

From the comparison of results shown in Table 2.1, one can see those top algorithms are the ensemble method proposed by Isensee *et al.* and the two-stage method proposed by Li *et al.*, both of which are based on FCNs. In particular, compared to the traditional level-set method [108], both methods achieved considerably higher accuracy even for the more challenging segmentation of the left ventricular myocardium (MYO), indicating the power of deep learning-based approaches. One should note that the success of deep learning models on this

benchmark dataset comes not only from the emergence of advanced network architectures but also from the increased size of public datasets [6]. When it comes to deploying deep learning methods to real-world applications, the current literature suggests that there is still a long way to go due to several significant limitations. We summarize them in the next section.

2.3 Limitations of deep learning

> Even though such machines might do some things as well as we do them, or perhaps even better, they would inevitably fail in others, which would reveal they were acting not through understanding, but only from the disposition of their organs.
>
> *Discourse on the Method*
>
> Rene Descartes, 1637

In this section, we will discuss about main limitations of deep learning that hinder its deployment in real-world applications:

- Limitation 1: requirement of massive labeled data for training,
- Limitation 2: sensitivity to small changes in inputs, e.g., adversarial noise in images,
- Limitation 3: lack of explainability and interpretability.

2.3.1 Requirement of massive labeled data for training

Deep learning models are essentially deep artificial neural networks with millions of parameters and complex structures, which require large labeled data sets to avoid over-fitting. Fig. 2.11 illustrates the idea of under-fitting, optimal fitting, and over-fitting for a classification model.

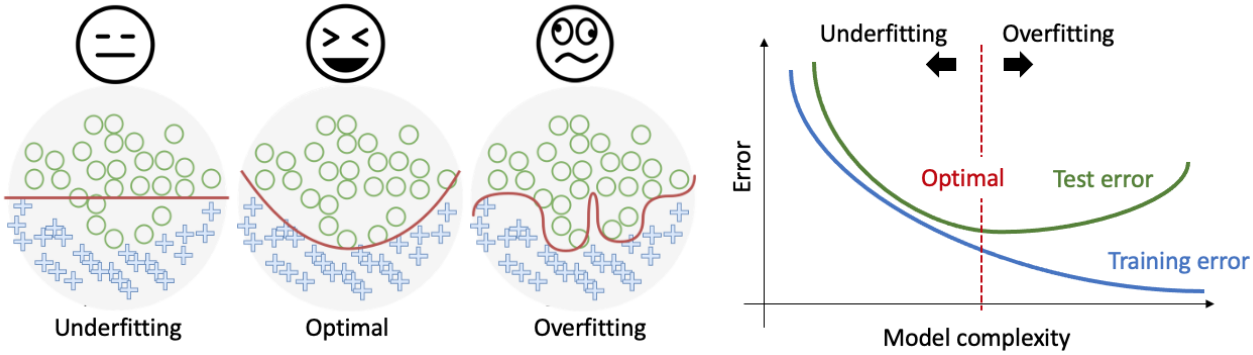


Figure 2.11: **Visual demonstration of under-fitting, optimal fitting, over-fitting and how they affect prediction accuracy.** Red lines in the left three diagrams represent decision boundaries.

Under-fitting happens when we have a model with limited capacity, which fails to capture the underlying structure of the data. For example, as shown in the left-most diagram in Fig. 2.11, under-fitting occurs when we fit a linear model (see the red line) to complex data with a non-linear structure. In practice, under-fitting can happen when the task complexity is much higher than the model complexity. Over-fitting happens when a model fits the training data too 'perfectly', learning irrelevant detail and noise in the training dataset, as shown in the third diagram in Fig. 2.11. In general, an over-fitted model tends to 'memorize' training data rather than 'learning' generalized concepts. In this case, the performance on unseen test data becomes worse while the error on the training examples still decreases when the model complexity is much larger than desired, see the right-most figure in Fig. 2.11. For deep learning models with many parameters and complex structures, over-fitting is more likely to happen. This is because in the real world it is very difficult to obtain a large-scale, representative labeled dataset.

In medical imaging applications, the lack of large labelled dataset is prevalent and extremely severe due to several reasons. First, manually labeling images can be prohibitively expensive and time-consuming. Taking cardiac MR image segmentation as an example, it generally takes a trained expert 20 minutes to analyse a single subject and to delineate the cardiac structures from images with manual annotations. Second, collecting and labeling datasets from multiple, different sites to form a large-scale dataset to cover the variety of real-world data is time-consuming and sometimes infeasible due to data privacy issues.

2.3.2 Sensitivity to small changes in images

There has been substantial evidence that modern CNNs can be surprisingly brittle even when changes to the input are nearly imperceptible. These networks are not only sensitive to carefully constructed adversarial noise, which introduces imperceptible changes to images [109] but are also sensitive to geometric changes, such as image translation, rotation [110, 111], and other contextual changes, e.g., adding a carefully constructed 'adversarial patch' [112]. This has raised concerns about the safety of AI when deploying deep neural networks in safety-critical applications such as autonomous driving, face recognition, and medical diagnosis [113, 114]. The fragility of neural networks against those small image transformations stems from their outstanding capacity to learn complex, salient, and non-salient features in deep layers. This also brings risks that a tiny change in the input can change their intermediate features and affect the final decision.

To strengthen neural networks' robustness, adversarial training has emerged as a principled approach, which augments training data with adversarial examples or other challenging data that may alter the network's prediction [109]. However, optimizing neural networks against one particular form of attacks can weaken them against others [115]. Another direction is instead to detect and report those outlier inputs before providing network predictions [114]. So far, there is no golden remedy to solve the brittleness of neural networks completely.

2.3.3 Lack of explainability and interpretability

Another issue with deep neural networks is their opacity, as most of them are 'black boxes' in nature. Traditional symbolic AI such as decision trees can reason about their decision-making. By contrast, deep neural networks, which have millions of operations with their complex structures, are extremely difficult for a human to follow the exact mapping from data input to prediction. The lack of transparency makes deep neural networks unpredictable, and hence untrustworthy. In recent years, many efforts have been made to make neural networks more explainable and interpretable. This includes the visualization and analysis of intermediate

features [116] and input attribution, e.g., highlight the pixels that were relevant for image classification by a neural network [117].

2.4 Theories and practices for model generalization

In this thesis, we particularly consider the limitation 1 and 2, focusing on alleviating the need of massive labeled data for domain generalization, and improving model robustness against realistic imaging corruptions. Below we introduce existing theories and common practices for improving model generalization, which form the basis of our work that will be introduced latter.

2.4.1 Generalization theory

To guide the model selection and alleviate the over-fitting problem, different theories have been proposed with different measures of model complexity. Central to these theories is 'simplicity'. As suggested by the *Occam's Razor* principle proposed in the 14^{th} century, the simplest one is the most preferable among all candidate solutions. However, formulating Occam's razor in machine learning is not trivial. Let *generalization error* of a model be the error rate on unseen data, and the *empirical/training error* be its error rate on the training examples that it was learned from. A formulation of the razor that may be the closest to Occam's original intent is: *Given two models with the same generalization error, the simpler one should be preferred.*

However, the *generalization error* is often not feasible to compute and how to effectively quantify the 'simplicity' or 'complexity' of different learning models for model selection is still an open question. These two problems pose challenges to model generalization.

A group of studies focus on establishing theories and utilizing them to quantify the model complexity for specific types of models. One group of studies are based on information theory. Two of the most representative works in this regard are the minimum description length (MDL) principle [118] and Solomonoff's inference theory [119]. Viewing learning models as data compressors, MDL suggests that the one that permits the greatest compression of the

data should be selected. Solomonoff's inference theory of universal inductive inference uses the Kolmogorov complexity [120], a.k.a. algorithmic complexity, to quantify the model complexity, which is determined by the length of the shortest binary computer program that describes the object. Similar to the MDL principle, this theory favors models with the 'shortest program' to produce the training data. However, in practice, the information-oriented theoretic minimum description length cannot be easily computed, as it can be very time consuming especially when the dataset itself is extremely large.

Later on, several works proposed different model complexity measures based on statistical learning theories to obtain the upperbound of generalization error. Two of the most well-known measures are the Vapnik–Chervonenkis (VC) dimension (d_{VC}) [121] [4] and Rademacher complexity (d_R) [122]. A key assumption behind both of them is that a training set $\mathcal{D}_{tr}$ is generated by an *unknown* distribution $\mathcal{D}$, where each data point is I.I.D (independently, identically distributed). A learning algorithm chooses a function/hypothesis $f_i : \mathbf{X} \to \mathbf{Y}$ from a hypothesis space $\mathcal{H}$:$\{f_1, f_2, f_3, ..., f_k\}$ based on the training dataset $\mathcal{D}_{tr}$, and then performs prediction with this hypothesis on unseen data (e.g., test data) from the same distribution. The upperbound of the generalization error for the class of hypothesis ($Err_{\text{generalization}}$) can then be estimated as:

$$Err_{\text{generalization}} \leq Err_{\text{train}} + g(d(\mathcal{H}), m), \tag{2.9}$$

where $d(\cdot)$ is a complexity measure of the hypothesis class, e.g., d_{VC} or d_R. And $g(\cdot)$ is a function that approaches 0 when the training data size $m = |\mathcal{D}_{tr}|$ approaches infinity; $g(\cdot)$ approaches infinity when the measured complexity explodes to infinity. This formula suggests that the generalization error is dependent on both model complexity and the size of training data and help to quantify how much data is needed as a function of a particular complexity measure.

While the above theoretical analysis provides nice formal guarantees, they can be difficult to apply in practice, especially in deep neural networks where the model complexity is difficult

[4] VC dimension is the maximum cardinality of the largest set that an algorithm can shatter. In practice, VC dimension is mainly used for statistical binary classification algorithms (e.g., linear classifiers), and is highly correlated with the number of parameters in most cases.

to quantity and certain assumptions may not hold [28, 123–125]. In fact, both the VC dimension [121] and Rademacher complexity[122] fail to estimate *tight* generalization bounds for deep learning models. And it has been reported that over-parameterized networks whose model capacity greatly exceeds the training set size can still have good intra-domain generalization performance on the hold-out test set [124].

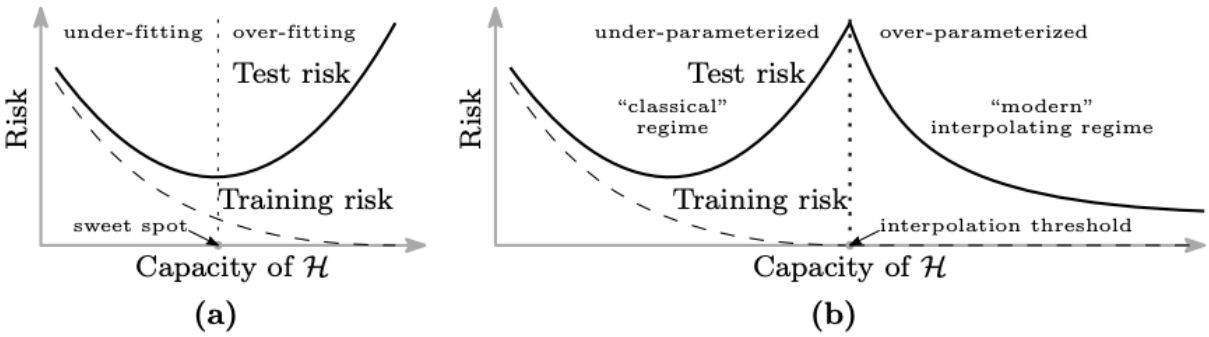


Figure 2.12: **Risk curves for classical models and modern deep learning models.** (a) The classical U-shaped risk curve in the bias-variance trade-off for conventional statistical models. (b) The double descent risk curve for modern neural networks, which incorporates the classical U-shaped risk curve and the observed behavior from deep neural networks with high capacity. The interpolation threshold is the critical point with zero training error. After this point, the test error begins to decrease with increased model capacity, where the traditional bias-variance trade-off fails to predict. Image source: [126]. Image reproduced with permission from the Proceedings of the National Academy of Sciences USA (PNAS) for noncommercial use.

To explain this phenomenon, Belkin *et al.* proposed a new double U-shaped risk curve for deep neural networks (see Fig. 2.12), which challenges the traditional bias-variance trade-off theory in classical statistical learning theory. It suggests that once the number of network parameters is high enough, the risk curve enters into the second regime, where the higher the capacity of networks, the lower the generalization error. However, this double descent risk curve is largely empirically observed and can be tricky to reproduce [127]. Other works try to tighten the generalization bounds by establishing new theory to measure the model complexity of deep neural networks [128, 129], such as intrinsic dimension [128] and the lottery ticket hypothesis [129]. Both works suggest that the complexity of deep learning models are significantly smaller than what they might appear to be. For example, the lottery ticket hypothesis states that in a dense, feed-forward network there exists a pool of sub-networks ('winning tickets') which can achieve good accuracy that is comparable to the performance of the original net-

work. This suggests that while there is a huge number of parameters in the network, which gives the network freedom to discover and to model the data structure, the final solution after training only occupies a smaller set of 'active' parameters. However, finding those sub-networks requires significant computational resources since models must be trained with a full structure and retrained many times with pruned networks. Yet, most of these theories are mostly verified on specific types of neural networks with a specific task (e.g., supervised learning for image classification), and their prescriptive and descriptive value is still uncertain [124, 125]. So far, the generalization theory in deep learning is still an under-explored domain.

2.4.2 Practical techniques to avoid over-fitting in deep learning

While it is difficult to theoretically quantify model complexity of deep neural networks, there are several practical techniques proposed to improve model generalization and reduce over-fitting.

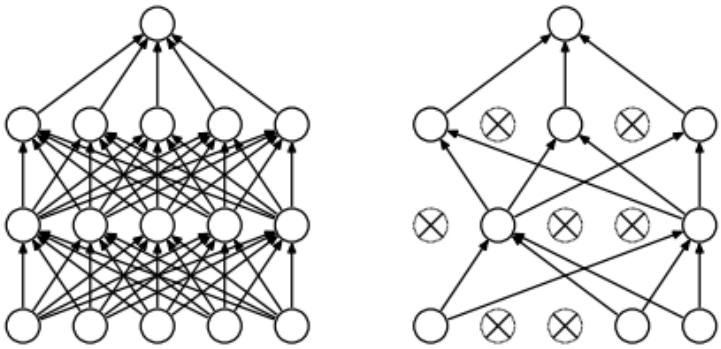


Figure 2.13: **A standard neural network and its variant with dropout.** An example of a two-layer net without and with dropout. Image source: [130], license: Creative Commons license (CC BY 4.0).

Several commonly used techniques are:

- **Train-val split strategy**: This is a training strategy to estimate unseen the test error for model selection since the true test set is not accessible during training. Specifically, it splits the training set into a training subset and a validation subset without overlapping. Then a model is trained on the training subset and evaluated on the validation set

throughout learning. The validation error is used to estimate the unseen test error. The model with the lowest validation error is chosen as the 'optimal' model;

- **Weight regularization**: Weight regularization is a type of regularization techniques that add weight penalties $R(\boldsymbol{w})$ to the empirical loss function $\mathcal{L}_{emp}$: Weight regularization encourages small or zero weights $\boldsymbol{w}$ for less relevant or irrelevant inputs. Common methods to constrain the weights include L1 regularization: $R(\boldsymbol{w}) = \|\boldsymbol{w}\|_1 = \sum_j |w_j|$ which penalizes the sum of the absolute weights; and L2 regularization:$R(\boldsymbol{w}) = \|\boldsymbol{w}\|_2^2 = \sum_j w_j^2$ which penalizes the sum of the squared weights;

- **Dropout** [130]: Dropout is a regularization method that randomly drops some units/neurons from the neural network during training (see Fig 2.13), encouraging the network to learn a sparse representation. On the basis of the vanilla Dropout, there are also several variants developed to further enhance the regularization effect for specific vision tasks, such as Spatial Dropout [131] which drops out entire feature maps rather than individual neurons;

- **Data augmentation**: Data augmentation is a training strategy that artificially generates more training samples to increase the diversity of the training data. This can be done by applying different transformations to each input sample, such as injecting random noise, applying affine transformations (e.g., rotation, scaling), flipping, or cropping to the original labeled sample. Recently, there is a growing interest in learning-based data augmentation to improve the diversity and effectiveness of augmented samples, including adversarial data augmentation [132], and generative model-based data augmentation such as GAN-based approaches [133];

- **Ensemble learning**: Ensemble learning is a type of machine learning algorithms that combine multiple trained models to obtain better predictive performance than individual models, which has been shown effective for medical image segmentation [134]. By averaging predictions from different learners, individuals mistakes can be potentially dismissed. There have been many different ways to construct diverse models, including training the same network with different hyper-parameters [134], constructing different training subsets [135], or training different networks with the same data [38];

- **Transfer learning**: Transfer learning aims to transfer knowledge from one task to another related but different target task. This is often achieved by reusing the weights of a pre-trained model to initialize the weights in a new model for the target task. Transfer learning can help to decrease the training time and achieve lower generalization error [136].

The above strategies are independent of network architectures, and they have been widely adopted for improved model generalization. These techniques can help to control the model complexity and alleviate over-fitting without explicitly modifying the architecture of networks. So far, these techniques have been widely adopted in modern CNN-based methodologies. In this thesis, these techniques such as train-val split, weight regularization, data augmentation, ensemble learning, transfer learning have been employed and/or investigated in our works to enhance model generalization.

2.5 Conclusion

In this chapter, we have introduced some representative deep learning networks together with advanced techniques for improved representation learning. We also provided a brief review of their applications in cardiac MR segmentation, giving a glimpse into the superior capacities of deep learning against non-deep learning models for medical image analysis. We then discussed the limitations of deep learning models and the theory and existing common practical techniques for improving model generalization. The following chapters will focus more on our recent works on improving model generalization and robustness for specific applications, particularly cardiac MR segmentation.

Chapter 3

Learning with Auxiliary Data

This chapter contains material from

1. C. Chen, W. Bai and D. Rueckert, 'Multi-task learning for left atrial segmentation on GE-MRI,' in *Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges - 9th International Workshop, STACOM 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers*, vol. 11395, Springer International Publishing, 2019, pp. 292–301. DOI: 10.1007/978-3-030-12029-0_32 [12]

2. C. Chen, C. Biffi, G. Tarroni, S. Petersen, W. Bai and D. Rueckert, 'Learning shape priors for robust cardiac MR segmentation from multi-view images,' in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2019 - 22nd International Conference, Shenzhen, China*, vol. 11765, Springer, Jul. 2019, pp. 523–531. DOI: 10.1007/978-3-030-32245-8_58 [13]

In this chapter, we aim to alleviate data scarcity and improve model generalization by utilizing auxiliary data. In practice, while there is limited labeled data for a particular task, there are often auxiliary data available for other tasks. For example, in our modern digitized healthcare environment, there are images of the same modality taken from different views, and images of different modalities (e.g., CT, MR) to visualize different parts of a body (e.g.,

bones, soft tissues). Aside from medical images, there is also non-imaging data such as clinical history, physical examination, and other laboratory results available to gain a comprehensive understanding of the patient's condition. Inspired by the fact that clinicians can process data for multiple sources and apply the learned knowledge to improve decision making, we would like to develop multi-task learning algorithms that extract useful contexts from auxiliary data and leverage them to help the main task. In Sec. 3.1, we present a multi-task learning framework for atrial segmentation, which utilizes non-imaging patient information as auxiliary data to help our atrial segmentation from gadolinium enhancement MR images. In Sec. 3.2, we present a novel segmentation framework that utilizes auxiliary data from multiple views for learning the shape prior and guiding the segmentation. Experiments show that utilizing auxiliary data for multi-task learning can relax the constraints of massive labeled data and improve the model generalization.

3.1 Multi-task learning for left atrial segmentation on GE-MRI

3.1.1 Introduction

Atrial fibrillation (AF) is a condition of the heart that causes an irregular and often abnormally fast heart rate [137]. This can cause blood clots to form, which can restrict blood supply to vital organs, and further leads to a stroke and heart failure [138]. One of the most common treatments for AF is called ablation which can isolate the pulmonary veins (PVs) from the left atrium (LA) electrically by inducing circumferential lesion and destroying abnormal tissues. During this procedure, a good understanding of the patient atrial anatomy is very vital for planning and guiding the surgery, and further improving the patient outcome [138].

A good way to learn the anatomical structure of the LA is by performing LA segmentation on medical images, such as CT scans and MR images. With the development of imaging techniques and computer science, many automatic or semi-automatic algorithms [139] have

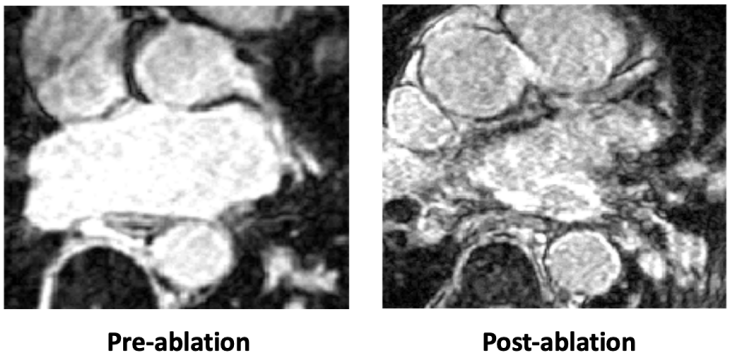


Figure 3.1: **Visualization of pre-ablation and post-ablation GE-MRI images**

been proposed for atrial segmentation. However, this is still a challenging problem and many traditional methods may fail to segment due to several reasons. For example, intensity-based methods such as region growing may fail to segment those atria with extremely thin myocardial walls, especially when their surroundings have very similar intensity to their blood pool [139]. In addition, there is large shape variation among the LA of different individuals, such as atrial sizes and pulmonary vein structures [139]. These variations will make it too complex for model-based segmentation methods to impose shape prior. An alternative way is to use atlas-based methods that can be robust to the LA with high anatomical variations. However, this kind of approach is time-consuming which typically takes 8 minutes around [140]. Most recently, with the increase of computing hardware performance and more data becoming available, deep learning has become the state-of-the-art method due to its efficiency and effectiveness on computer vision tasks, and has been widely used in the medical domain [141].

In this work, we focus on the segmentation of the LA from gadolinium enhancement MR (GE-MRI) images. These images can be taken either before or after ablation treatment. Noticing that there might be contextual difference between the pre-ablation and post-ablation images, e.g., ablation will cause scars in the LA [142] and may influence the quality of images as shown in Fig. 3.1, we propose a multi-task CNN that could segment a patient left atrium from GE-MRI images and detect whether this patient is pre- or post-ablation. In this way, our network could not only learn structural information from segmentation masks, but also retrieve contextual information through the auxiliary classification task. Our network is trained simultaneously for the two tasks, using a stack of 2D slices extracted from each MRI scan along

with its corresponding segmentation masks and a pre/post ablation label. In addition, in order to improve the robustness of segmentation on images with various image contrast and sizes, we employ a contrast augmentation method to augment our training set and trained our network with images in different sizes. In order to produce a fixed-length vector to classify input images in multiple sizes, spatial pyramid pooling [143] is adopted in this network.

The proposed framework was trained and evaluated on the data set of the Atrial Segmentation Challenge 2018[1]. Our experimental results show that by sharing features between related tasks, our network can achieve better segmentation performance compared to a variant of U-net trained with a single task. During the test phase, our network can directly inference the segmentation mask from a scan of MR images without taking extra pre-processing steps for image contrast enhancement. In total, our method is very efficient as one 3D segmentation result for each individual was obtained in 6 seconds on a Nvidia Titan Xp GPU using our model, plus 3 or 4 seconds for post-processing on the whole volume, which is far more faster than general atlas-based methods that usually take minutes.

Related work. There has been several works on automating the segmentation of atrial segmentation. Traditional methods such as region growing [144] and atlas-based label fusion methods [145], and image registration-based methods [146] have been applied. However, the accuracy of these methods highly rely on good initialization and ad-hoc pre-processing methods, which limits the widespread adoption in the clinic. Recently, it has been shown that 2D fully convolutional neural networks can be very effective techniques for segmenting the left atrium from standard 2D long-axis images, i.e., 2-chamber (2CH), 4-chamber (4CH) views [4, 97]. Different from previous single-task learning framework, in this work, we would like to investigate the benefit of multi-task learning for GE-MRI image segmentation, which has not been fully explored before.

[1] https://atriaseg2018.cardiacatlas.org/

3.1.2 Methodology

In this section, we present the architecture of our proposed multi-task network and how we post-process the network output to get the final 3D segmentation mask. Our proposed network is adapted from a commonly used fully convolutional network, i.e., U-net architecture [32] where we increase the depth of the network and add a classification branch. The input to the network is a stack of 2D images. The output are predictions of the atrial segmentation mask for this stack and pre/post ablation classification scores.

3.1.2.1 Network architecture

In order to explore the benefit of multi-task learning, the proposed network is designed to conduct both the atrial segmentation task and an auxiliary pre/post ablation classification task with images of multiple sizes.

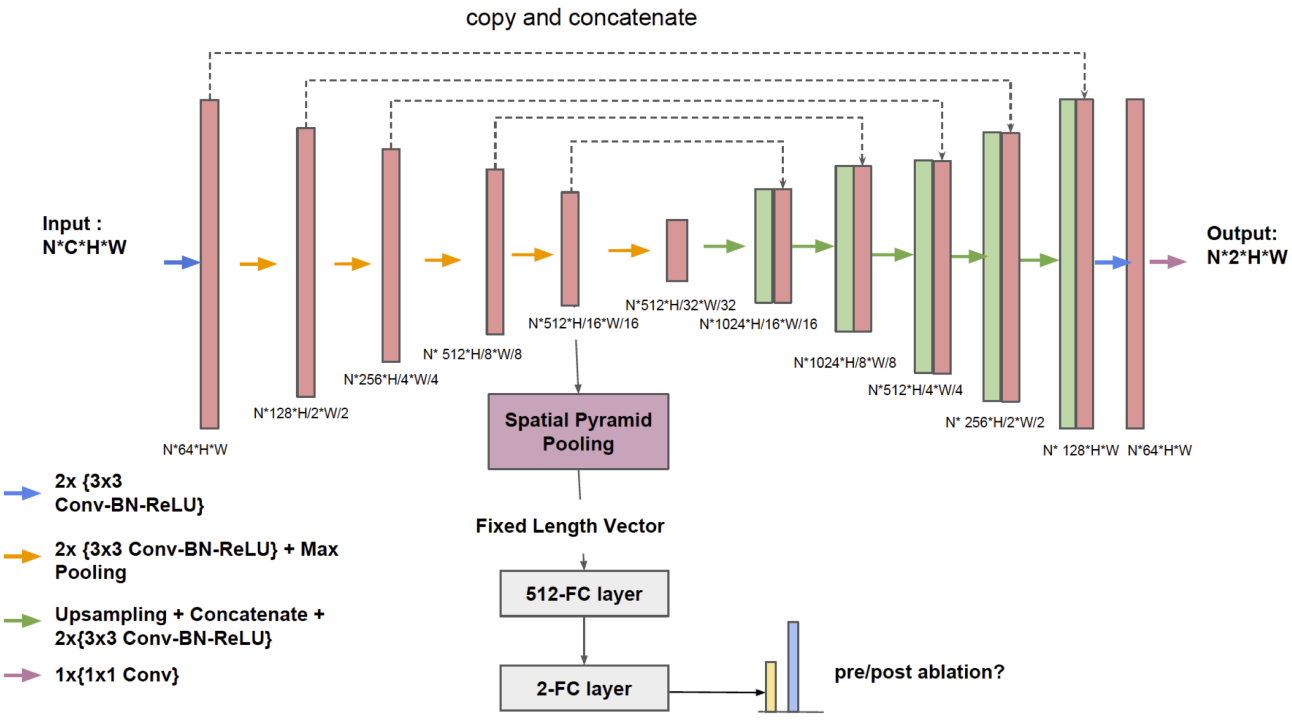


Figure 3.2: **Architecture of the proposed multi-task Deep U-net.** Conv: convolutional layers; BN: batch normalization layers; FC: fully connected layers. N: number of input 2D slices. H, W: image height and width. Best viewed in color.

The core of our method, named 'Deep U-net' and shown in Fig 3.2, is derived from the 2D U-net [32] for semantic segmentation. Since the largest size of images in our dataset is

640×640 in x-y planes, we increased the receptive field of U-net by adding more pooling layers. The modified network now consists of five down-sampling blocks and five up-sampling blocks. Each down-sampling layer contains two 3×3 convolutions, with Batch Normalizations [147] and Rectified Linear Unit activations, as well as a 2×2 max pooling operation with stride 2 for down-sampling. The up-sampling path is symmetric to the down-sampling path. By aggregating both coarse and fine features learned at different scales from the down-sampling path and up-sampling path, our network is supposed to achieve better segmentation performance than those networks without the aggregation operations.

Our classification task is performed by utilizing image features learned from the down-sampling path. Features after the 4^{th} max pooling layer are extracted for classification, which is a common practice for many existing classification networks [56, 148]. In order to generate fix-length feature vectors learned from input images with different sizes and scales, spatial pyramid pooling [143] is applied. These fixed-length vectors are then processed through fully connected layers followed by a softmax layer to calculate class probabilities (pre/post-ablation) for each image. Dropout [130] is applied to the output of fully connected layers with a probability of 0.5 during the training process, which functions as regularizer to encourage the sparsity of network for improved model generalization [130].

3.1.2.2 Loss function

Given a training set $\mathcal{D}_{tr}$: $\{(\boldsymbol{x}, \mathbf{y}_s, \mathbf{y}_c)^t\}_{t=1...m}$ consisting of a number of training images $\boldsymbol{x}$ and their corresponding segmentation $\mathbf{y}_s$ and classification labels $\mathbf{y}_c$ (pre/post ablation), the loss function $\mathcal{L}$ for our multi-task network is defined as follows:

$$\mathcal{L} = \mathbb{E}_{(\boldsymbol{x},\mathbf{y}_s,\mathbf{y}_c)\sim\mathcal{D}_{tr}} \mathcal{L}_S(\mathbf{p}_s, \mathbf{y}_s) + \lambda \mathcal{L}_C(\mathbf{p}_c, \mathbf{y}_c), \tag{3.1}$$

where $\mathcal{L}_S$ is the loss for the segmentation task measuring errors between the predicted segmentation maps $\mathbf{p}_s$ and the ground truth one-hot maps $\mathbf{y}_s$ for each training image $\boldsymbol{x}$, $\mathcal{L}_C$ is the loss for the classification task measuring classification errors given the predicted class scores $\mathbf{p}_c$

and the ground truth class labels $\mathbf{y}_c$, and λ is a coefficient to balance the two terms, which is empirically set to 1 in our experiments. For the segmentation part, pixel-wise cross-entropy loss is employed:

$$\mathcal{L}_S(\mathbf{p}_s, \mathbf{y}_s) = -1/n \sum_{j=1}^{n} \sum_{i=0,1} \mathbf{y}_s^{i,j} \log \mathbf{p}_s^{i,j}, \tag{3.2}$$

where $\mathbf{p}_s^{i,j}$ is the probabilistic prediction for the class i from the network (after softmax) for an given image $\boldsymbol{x}$ at pixel j: $\mathbf{p}_s^{i,j} = softmax(f_s(\boldsymbol{x})^{i,j})$ and $f_s(\boldsymbol{x})$ denotes the network output from the segmentation branch given $\boldsymbol{x}$, i indicates the segmentation class index (0: background, 1: left atrium); n is the total number of pixels in each training image. For classification part, we adopt the sigmoid cross-entropy to measure pre/post ablation classification loss:

$$\mathcal{L}_C(\mathbf{p}_c, \mathbf{y}_c) = -\sum_{k=0,1} \mathbf{y}_c^k \log \mathbf{p}_c^k, \tag{3.3}$$

where k indicates the image class index (0: pre-ablation, 1: post-ablation); $\mathbf{p}_c$ is the classification score, where $\mathbf{p}_c^k = \frac{1}{1+e^{-f_c(\boldsymbol{x})^k}}$ and $f_c(\boldsymbol{x})$ denotes the network output from the classification branch given $\boldsymbol{x}$. The classification ground truth of a 2D image is labeled as 1 if this slice is extracted from a post-ablation object. Otherwise, its ground truth is 0. The whole network is trained jointly on the combined loss, where the classification loss works as a regularization term, enabling the network to learn the high-level representation that generalizes well on both tasks.

3.1.2.3 Post-processing for shape refinement

During the inference time, axial slices extracted from a 3D image are fed into the network slice by slice. The segmentation branch predicts pixel-wise probability score for both background and atrium classes. A 2D segmentation mask is then generated by finding the class with the highest probability for each pixel on the slice. By concatenating these segmentation results slice by slice,a rough 3D mask for each patient is produced. In order to refine the boundary of those masks, we performed 3D morphological dilation and erosion, and kept the largest connected component for each volume.

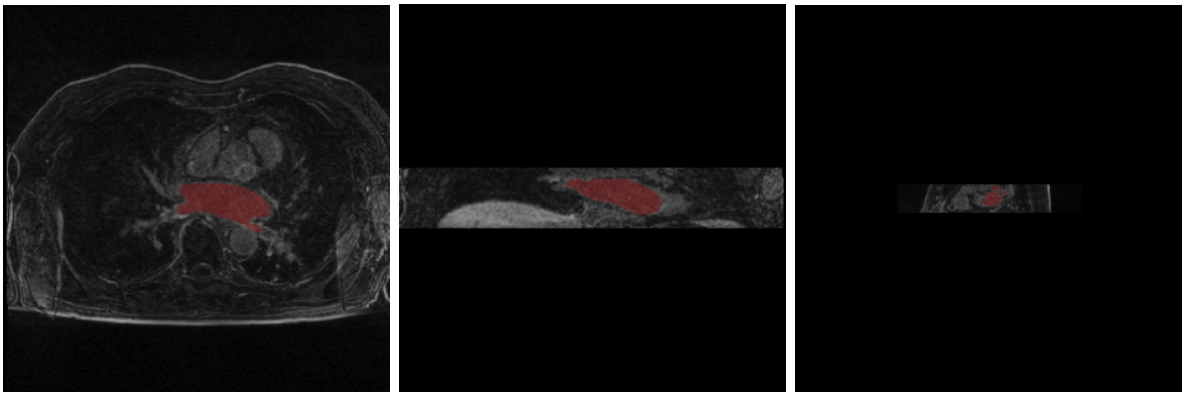
3.1.3 Experiments

Data. In this work, our algorithm was trained and evaluated on the dataset of the 2018 Atrial Segmentation Challenge [2]. This dataset contains a training set of 100 3D GE-MRI scans along with corresponding LA manual segmentation mask and pre/post ablation labels for training and validation. In addition, there is a set of 54 images without labels provided for testing. For model training and evaluation, we randomly splitted the training set into 80 : 20. We did not use any external data for training or pre-training of our network.

Images in this dataset have been resampled and preprocessed by the organizers. So there is no need to do re-sampling procedure in the pre-processing stage. Despite the consistency observed in the resolution of the data, this dataset exhibits large differences in images sizes and image contrast. For example, there are two sizes of images in this dataset: 576×576 and 640×640 on the axial planes. Apart from that, atria, in different images, can also have various shapes and sizes. These phenomena may arise due to the fact that these scans were collected from multiple sites which may have different scanners and imaging protocols. Hence, it is important to build a robust method for those images. Fig. 3.3 visualizes the difference in image contrast of different images in different views. In this work, we use data augmentation to increase data variety with the aim of improving the model's generalization ability on different images, which will be discussed in section 3.1.3.

Data pre-processing. In order to preserve the resolution of images, image re-scaling was not performed in the data pre-processing stage. Instead, multi-scale cropping was used to increase the data variety, so that network can analyze images with different contexts. More details will be described in the next section as it actually happens in the data augmentation process. For testing, images can be directly fed into the network provided that its length is a factor of 32 due to the architecture of the network. Otherwise, zero padding is required. The only necessary step in our pre-processing stage in both training and testing stage is to normalize image intensity to zero mean and unit variance, which has been widely accepted in common practice.

[2] http://atriaseg2018.cardiacatlas.org/



(a) Patient A

(b) Patient B

Figure 3.3: **Visualization of 2D raw slices at axial (left), coronal (middle), sagittal (right) views.** Despite the homogeneity in image resolution, there are significant differences in the image contrast and quality among different individuals, which can be challenging to segment the left atrium (red) from MR images. Best viewed in color.

Data augmentation. Our training data was augmented via a composition of image transformations, including random horizontal/vertical flip with a probability of 50%, random rotation with degree range from -10 to $+10$, random shifting along X and Y axis within the range of 10 percent of its original image size, and zooming with a factor between 0.7 and 1.3.

In order to process images at multiple sizes with objects at multiple scales, we centrally cropped 2D images at various image scales. The cropped sizes include $256\times256, 384\times384, 480\times480, 512\times512, 576\times576, 640\times640$. If the cropped size was larger than the image's original size, zero padding was performed instead. Motivated by Curriculum Learning [149], we trained our network firstly with cropped images where the left atrium taking a large portion of the image and then we gradually increased the image size. In this way, our network learns to segment from easy scenarios to hard scenarios and this helps the model to quickly converge in the beginning [150]. Despite the change in input images sizes, our network could still output

a fixed length feature vector for classification since we employed spatial pyramid pooling [143]. In practice, we found this could help the network focus on learning task-specific structural features for organ segmentation regardless of the contextual changes in sizes and scales. It is also beneficial for quantitative analysis based on medical image segmentation since we do not use rescaling nor resizing operations which have the risk of introducing scaling/shifting artifacts during prediction.

Contrast augmentation. We found that there exists a diversity of image contrast in the dataset, where low-contrast effects can reduce the visual quality of an image [151] and thus affect segmentation accuracy. To solve this issue, traditional machine learning methods often require image contrast enhancement methods during image pre-processing. Here, we proposed a contrast augmentation method based on gamma correction instead, to generate a variety of images with different levels of contrast during training. In this way, our CNN could gain the ability to segment images regardless of the difference of image contrast. And there is no need to do any contrast adjustment during testing. Therefore, our method is more efficient than those general traditional methods which require those adjustments.

The proposed contrast augmentation is based on a point-wise nonlinear transformation: $G(x, y) = F(x, y)^{1/\gamma}$ where $F(x, y)$ is the original value of each pixel in an image, and $G(x, y)$ is the transformed value for each pixel (x, y). The value of γ is randomly chosen from the range of $(0.8, 2.0)$ for each image. By applying gamma correction randomly, the variety of image contrast in the training set was significantly increased.

To show our contrast augmentation method is superior to the traditional contrast enhancement methods, we compared it with two image contrast enhancement methods: contrast limited adaptive histogram equalization (CLAHE) [152] and automatic gamma correction [151]. Both of them have been widely used in the pre-processing of CT image and MR image applications [153–155] in order to improve medical image quality for visual tasks. For CLAHE, we divided each image into 8×8 regions and performed contrast enhancement on each region by default.

The above experiments were performed based on a simple Deep U-net (without multi-task) for comparison. All networks was optimized using Stochastic Gradient Descent(SGD) [66] with the same setting: a momentum of 0.99 and weight decay of 0.0005. The initial learning rate is 0.001, which will be decreased at a rate of 0.5 after every 50 epochs. From Table 3.1, it can be seen that our proposed data augmentation method could significantly improve the robustness of our network for processing images with various image contrast and outperformed the traditional image pre-processing methods which may have the risk of amplifying noises and take extra processing time. Therefore, in the following sections, we would like to employ contrast augmentation as our default experimental setting.

Table 3.1: **Segmentation results of a single-task Deep U-net with different image contrast enhancement strategies.**

Base Model	Method	Need Extra Time	Dice ↑
Deep U-net	Baseline	No	0.847 (0.18)
Deep U-net	+ Automatic Gamma Correction	Yes	0.854 (0.15)
Deep U-net	+ CLAHE	Yes	0.876 (0.09)
Deep U-net	+ Gamma Augmentation	No	0.883 (0.08)

3.1.4 Results

To evaluate our segmentation accuracy for different experimental settings, we use four measurements: the Dice score (also known as Dice similarity coefficient score), the Jaccard Similarity Coefficient (JC) score, the Hausdorff Distance (HD) and the Average Symmetric Surface Distance (ASSD).

Table 3.2: **Segmentation results of different methods.**

	Dice ↑	JC ↑	HD ↓	ASSD ↓
Vanilla U-net	0.855 (0.11)	0.760 (0.14)	21.81 (19.35)	1.58 (1.07)
Deep U-net	0.883 (0.08)	0.798 (0.11)	21.18 (21.00)	1.20 (0.47)
Deep U-net + multi-task	0.896 (0.04)	0.815 (0.07)	15.40 (6.39)	1.11(0.35)
Deep U-net + multi-task + post-processing	**0.901 (0.03)**	**0.822 (0.06)**	**14.23 (4.83)**	**1.04 (0.32)**

To show the advancement of our deep network with additional pooling/max-pooling layers, we compared our modified networks with the vanilla 2D U-net [32]. The results are shown in

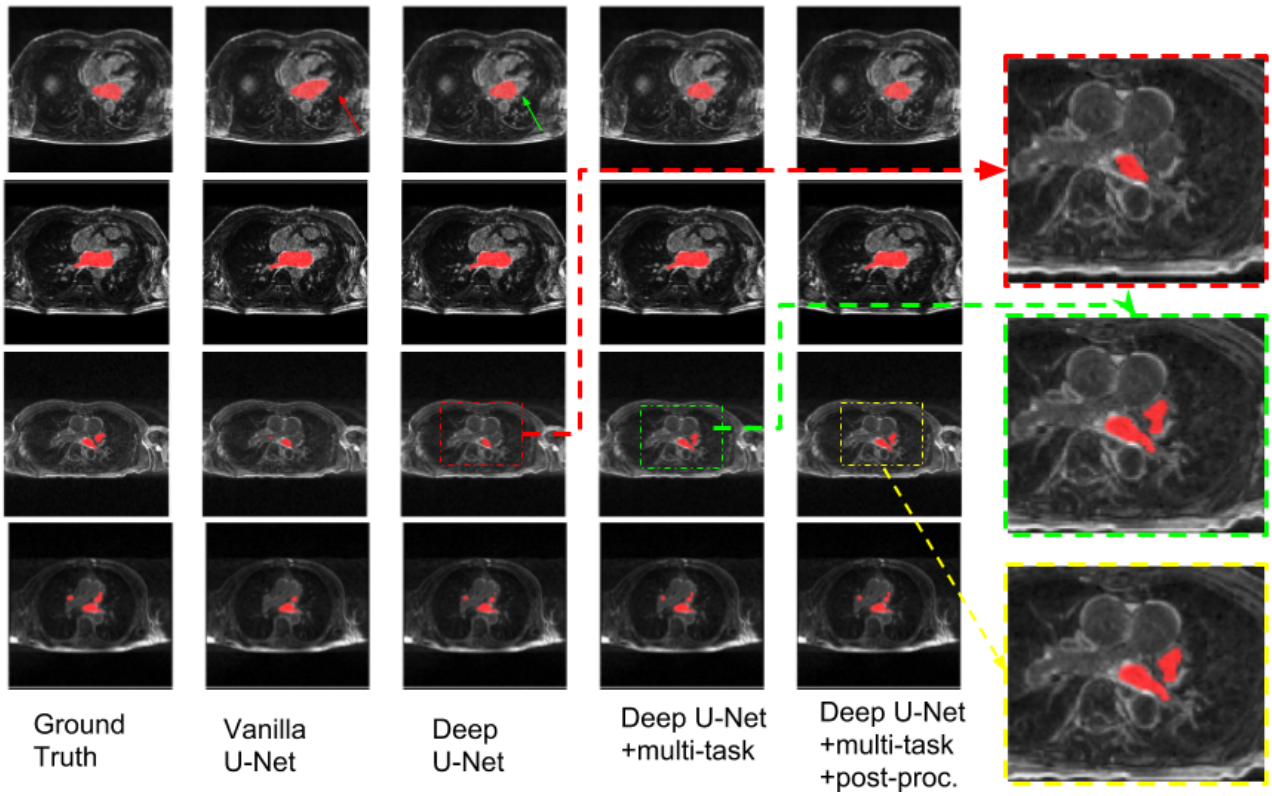


Figure 3.4: **Exemplar segmentations for axial slices using different methods. Each column shows different axial slices from the mitral to the PVs plane (top to bottom).**

Table 3.2. It can be seen that the segmentation performance was greatly improved by increasing the depth of the network, increasing the Dice score from 0.855 to 0.883. Our best results were achieved by using the multi-task Deep U-net followed by post-processing, producing a Dice score of 0.901. In particular, applying multi-task learning greatly reduces the Hausdorff distance from 21.18 to 15.40 *mm*. From the visualization plots in Fig. 3.4, we could see that our multi-task U-net is more robust than the other two with only one segmentation goal. One reason could be that by sharing features with segmentation and related pre/post ablation classification, the network is forced to learn better representation on images taken before the ablation treatment and those after the treatment, which could further improve segmentation performance. Fig. 3.5 shows that our model achieved high overlap ratio between our 3D segmentation result and the ground truth in different subjects. However, one significant failure mode can be observed around the region of pulmonary veins. One possible reason might be that the number and the length of pulmonary veins vary from person to person, making it too hard for the network to learn from limited cases.

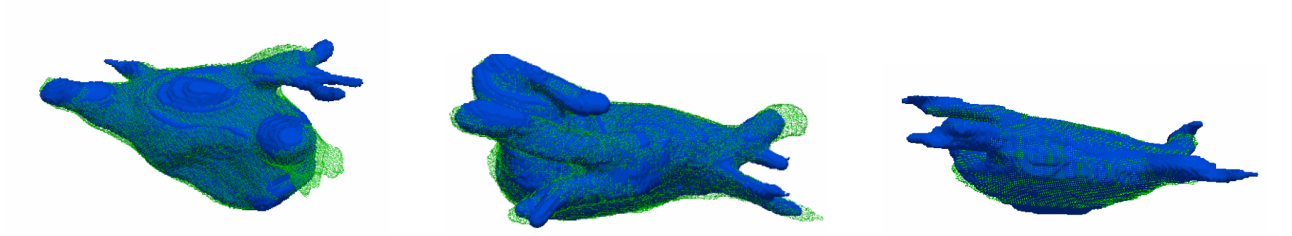


Figure 3.5: **3D visualization of three samples from the validation set.** Blue objects are the ground truth, and the green ones are the predicted segmentation of the proposed method.

For the Atrial Segmentation Challenge 2018, we adopted an ensemble method called Boostrap Aggregating (Bagging) [135] to improve our model's performance in the test phase. We noticed that samples in the dataset were collected from multiple sites while a large portion is from The University of Utah. In that case, domain shift or domain bias may exist when we use a model trained on one certain subset from a limited dataset to predict data from another subset as they may have different intensity distributions. Therefore, we trained the same model 5 times, each with a random subset and then averaged the class probabilities produced by these five models for prediction. Our ensembled results on a set of 54 test cases given by the organizers improved from an averaged Dice score of 0.9197 to 0.9206. Besides, the total processing procedure (inference + post-processing) for each whole 3D MRI predicted by our network took only approximately 10 seconds on average on one Nvidia Titan Xp GPU. It is therefore much more efficient than those atlas-based methods which typically take eight minutes [140].

3.1.5 Conclusion

In this section, we proposed a deep 2D fully convolutional neural network to automatically segment the left atrium from GE-MRI images. By applying multi-task learning to utilize the auxiliary non-imaging patient information, our network demonstrated improved segmentation accuracy on the unseen test set compared to a baseline U-net method. In addition, we showed that contrast augmentation is an efficient and effective way to enhance our model's robustness and efficiency when analyzing images with various image contrast. Yet, the proposed network can still fail to segment subtle structures, i.e., pulmonary veins where the number and the length of pulmonary veins vary from person to person. Since the network segments images

in the slice-by-slice fashion, it does not fully utilize the global spatial information across the volume. Extending the current 2D network to a 3D network could be one solution to help the network better understand the shape anatomy for each subject for precise segmentation.

3.2 Learning shape priors for robust cardiac MR segmentation from multi-view images

3.2.1 Introduction

Accurate segmentation of cardiovascular magnetic resonance (CMR) images is fundamental for assessing cardiac morphology and diagnosing heart conditions [5]. Manual segmentation of the anatomical structures is tedious, time-consuming and prone to subjective errors, which is not suitable for large-scale studies such as UK Biobank[3] [4]. Therefore, it is essential to develop automated, fast and accurate CMR segmentation techniques.

Recently, CNN based methods have achieved very good performance for cardiac image segmentation in terms of both speed and accuracy [4, 6, 37]. However, they may still produce sub-optimal segmentation results in some circumstances. For example, in the ACDC [6], the top segmentation methods (all CNN-based) achieve high overall segmentation scores for mid-ventricular short-axis slices. However, they sometimes produce poor results or even fail to locate the myocardium in basal slices (due to its more complex shape) and apical slices (due to its small size). This problem is not uncommon and has been reported in the related literature [6, 71, 84]. Methods based on 2D networks, trained in a slice-by-slice fashion, are particularly affected by this problem since they do not incorporate spatial context from neighboring short-axis (SAX) images or long-axis (LAX) views. On the other hand, 3D networks are capable of incorporating 3D spatial information to perform the segmentation task. Yet the 3D spatial context can be affected by potential inter-slice motion artifacts [156] and the low through-plane spatial resolution in cardiac SAX stacks, thus limiting their segmentation performance. Compared to 2D ones, 3D networks usually contain more parameter and are prone to over-fitting especially when the training set is limited in size since they use 3D volumes rather than 2D slices as input, significantly reducing the number of training samples.

Experienced clinicians are able to assess the cardiac morphology and function from multiple

[3] https://imaging.ukbiobank.ac.uk/

standard views, using both SAX and LAX images to form an understanding of the cardiac anatomy. Inspired by this, we propose a method which learns the anatomical prior knowledge across four standard views as auxiliary information and leverages this to assist the segmentation on 2D SAX images. The intuition behind our work is that the representation learned from multiple standard views is beneficial for the segmentation task on the SAX slices as different views should share the same representation of the 3D anatomy if they are from the same subject.

The main contributions of this work are the following: a) we developed a novel encoder-decoder architecture (Shape MAE) which learns latent representation of cardiac shapes from multiple standard views; b) we developed a segmentation network (multi-view U-net, adapted from [32]), which is capable of incorporating the anatomical shape priors learned from multi-view images to guide the segmentation on SAX images; c) we assessed the segmentation accuracy and the data efficiency of the proposed segmentation method against common 2D and 3D segmentation baselines by limiting the number of training images, demonstrating that the proposed method is more robust, and less dependent on the size of training data.

3.2.2 Related work

A large number of methods have been developed to improve the robustness of the cardiac segmentation. One approach is to learn an ensemble model where the predictions of a 2D and a 3D network are combined [38]. This method is capable of producing accurate results, but has a relatively high computational cost and requires an extra post-processing step to merge the predictions from the two networks. Another approach is to incorporate cardiac anatomical prior knowledge into segmentation networks [44, 157]. In [44], the learned representation of the 3D cardiac shape is employed to constrain the segmentation model to predict anatomically plausible shapes. The main bottleneck of this method is the requirement of fully annotated 3D high-resolution MR images which are free from inter-slice motion artifacts and have high through-plane spatial resolution. However, compared to the standard 2D imaging protocol, the 3D one requires the subjects to hold their breath for a relatively long time and therefore is often not feasible for patients with cardiovascular diseases. Instead of using 3D images, we

exploit *routinely acquired* 2D standard views to learn the shape representation of the cardiac structures. The learned representation is then injected into a segmentation network to improve its performance on SAX MR images. Of note, the approach in [158] also injects shape priors produced from an autoencoder into a segmentation network. However, the aim of that approach is to generate multiple segmentation hypotheses for ambiguous images, and cannot be readily employed to learn shape priors from different views to enhance cardiac segmentation.

3.2.3 Methodology

The proposed method consists of two novel architectures: 1) A **shape-aware multi-view convolutional neural network** (Shape MV-CNN) which aims at learning anatomical shape priors from standard cardiac acquisition planes incl. short-axis and long-axis views and 2) a **multi-view U-net** which performs cardiac short-axis image segmentation by incorporating anatomical priors learned by Shape MV-CNN into a modified U-net architecture.

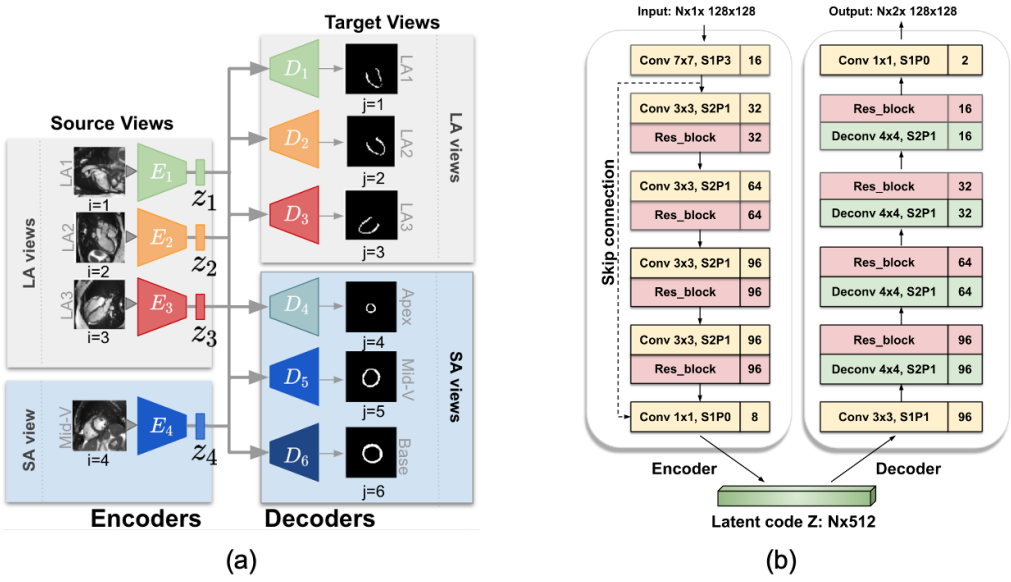


Figure 3.6: **(a) Overview of Shape MV-CNN. (b) Detailed architectures of each encoder and each decoder.** Each rectangle represents one or a series of convolutional (Conv) or transposed convolutional (Deconv) layers, where the number in the square box represents the number of filters for each layer. A 'Res_block' (pink rectangles) consists of two convolutional layers (3×3) with a residual connection that adds its input to the features from the second layer. Instance normalization and leaky ReLU activation are applied throughout the network. A sigmoid function is applied to the latent code z to bound its range.

3.2.3.1 Shape MV-CNN: Shape-aware multi-view convolutional neural network.

As illustrated in Fig. 3.6, we first present a novel architecture named shape-aware multi-view convolutional neural network (Shape MV-CNN) which learns anatomical shape priors from standard cardiac views through multi-task learning. Given a source view X_i, the network learns the low-dimensional representation z_i of X_i that best reconstructs all the j target views segmentations Y_j. In this work, we employ four source views X_i $(i = 1, \ldots, 4)$ which are three LAX images - the two-chamber view (LA1), three-chamber view (LA2), the four-chamber view (LA3) - and one mid-ventricular slice (Mid-V) from the SAX view. The target segmentations views Y_j $(j = 1, \ldots, 6)$ correspond to the four previous views plus two SAX slices: the apical one and the basal one. All encoders $E_i : z_i = E_i(X_i)$ and all decoders $D_j : Y_j = D_j(z_i)$ in the Shape MV-CNN share the same architecture (see Fig. 3.6 b).

The loss function $\mathcal{L}_{\text{Shape MAE}}$ for the whole network is defined as follows:

$$\mathcal{L}_{\text{Shape MAE}} = \mathcal{L}_{intra} + \alpha \mathcal{L}_{inter} + \beta \mathcal{L}_{reg} \tag{3.4}$$

The first two terms of Eq. 3.4 are defined as the cross entropy loss $\mathcal{F}_{ce}$ between the predicted myocardium segmentation $\hat{Y}_{i \to j} = D_j(E_i(X_i))$ for the target view j given a source image X_i of the same subject and its ground truth segmentation Y_j. $\mathcal{L}_{intra}$ denotes the segmentation loss when the source view X_i and the target view Y_j correspond to the same view: $\mathcal{L}_{intra} = \sum_{i=1, i=j}^{4} \mathcal{F}_{ce}(Y_j, \hat{Y}_{i \to j})$, whereas the second term $\mathcal{L}_{inter}$ denotes the loss when two views are different: $\mathcal{L}_{inter} = \sum_{i=1}^{4} \sum_{j=1, i \neq j}^{6} \mathcal{F}_{ce}(Y_j, \hat{Y}_{i \to j})$. The third term is a regularization term on the latent representations $z_i, z_i \in Z$: $\mathcal{L}_{reg} = \frac{1}{|Z|} \sum_{i=1}^{4} ||z_i - \bar{z}||^2$, which penalizes the L2 distance between z_i and $\bar{z}$, with $\bar{z} = \frac{1}{|z|} \sum_{i=1}^{4} z_i$ being the average z for a subject. Although the latent shape codes from different views of the same subject are not directly shared, this regularization term forces them to be close to each other. We use coefficients α and β to control the relative importance of $\mathcal{L}_{inter}$ and $\mathcal{L}_{reg}$.

The principle behind the proposed network is that different views require *independent* functions to map them to the latent space that describes **global** shape characteristics; whereas

translating this latent space to another view or plane also requires a *specific* projection function. Predicting the shape of the myocardium based on the six target views instead of a single view encourages the network to learn and exploit correlations between different views, resulting in a global, view-invariant shape representation rather than a local representation for a particular view. All the encoders and the decoders in this framework are trained jointly in a multi-task learning fashion, with the benefit of avoiding over-fitting and encouraging model generalisation [159].

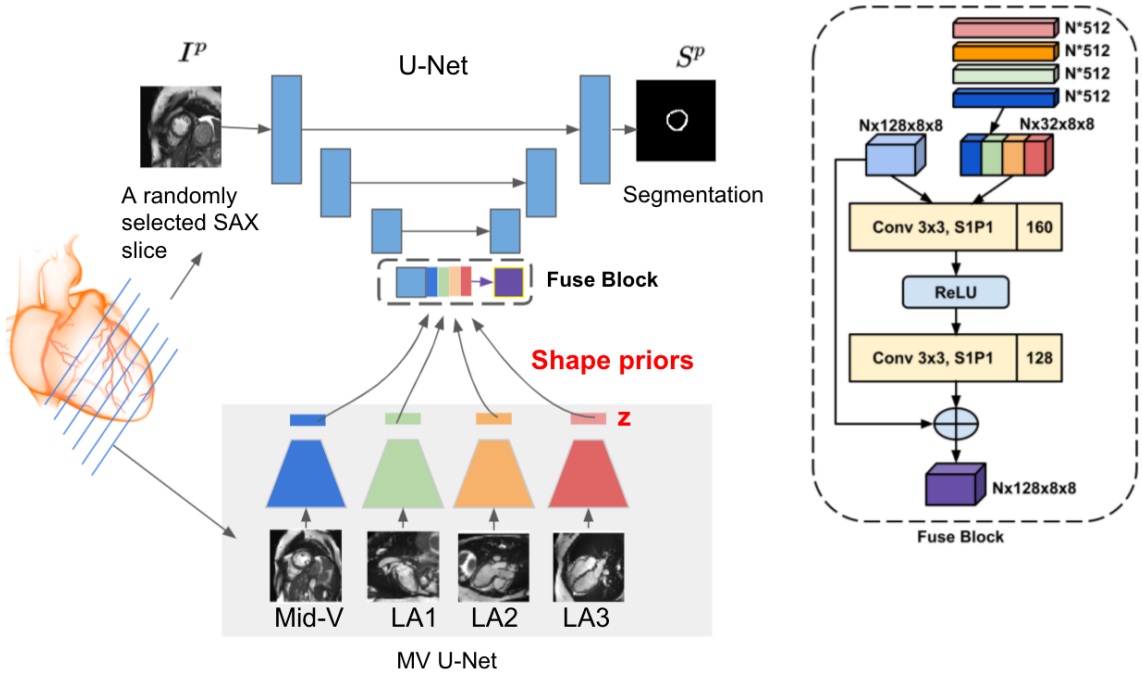


Figure 3.7: **(a) Overview of the proposed MV U-net.** (b) Architecture of the 'Fuse Block'. SAX: short-axis; LA: long-axis; Mid-V: mid-ventricle slice. The number of shown feature map blocks of the U-net is reduced for clarity of presentation. Batch normalization and ReLU activations are applied throughout the network. For each subject, the shape code of each view is reshaped to $1 \times 8 \times 8 \times 8$ and then concatenated with the other three along the second axis to form an input of $1 \times 32 \times 8 \times 8$ to the Fuse Block.

3.2.3.2 MV U-net: Multi-view U-net.

As shown in Fig. 3.7, we propose a segmentation network called multi-view U-net (MV U-net) based on the original U-net [32] for cardiac SAX image segmentation. The proposed network is capable of incorporating the anatomical shape priors learned by Shape MV-CNN. Similar to the original architecture, the proposed architecture comprises 4 down-sampling blocks and 4 up-

sampling blocks to learn multi-scale features. Differently from the original U-net, we reduced the number of filters at each level by four times to account for the fact that cardiac segmentation is simpler than the lesion segmentation (with multiple candidates) which was the task that the original U-net was applied to. In addition, a module called 'Fuse Block' is introduced in the bottleneck of the network (see Fig. 3.7 b) to inject the latent codes into the segmentation network. This fusing approach is different from that in [158] where the latent codes are simply concatenated with U-net activations. The proposed module consists of two convolutional kernels (3×3) and a residual connection to combine the shape representations from different views through learnable weights. Thanks to this module, given an arbitrary short-axis image slice I^p from a subject p and its correspondent shape representations $z_1^p, z_2^p, z_3^p, z_4^p$ obtained by Shape MV-CNN (one for each of the four standard views), the network can predict a segmentation $S^p = f_{\text{MV U-Net}}(I^p, z_1^p, z_2^p, z_3^p, z_4^p; \theta)$ by distilling the prior knowledge to the high-level features of the network, allowing it to efficiently refine the segmentations through multi-view information. The network is trained using standard training procedure with a cross entropy loss to optimise the parameters θ of the MV U-net.

3.2.4 Experiments

Experiments were performed on a dataset[4] acquired from 734 subjects from UK Biobank study [5]. For each subject, a stack of 2D SAX slices and three orthogonal 2D LAX images are available. All the LV myocardium were annotated on the SAX images as well as the LAX images at the end-diastolic (ED) frame using an automated method followed by manual quality control. All the images were acquired using one scanner. The spatial resolution of the images is $1.8 \times 1.8 \times 10\ mm$.

In our experiments, the dataset was randomly split into two subsets: a training set (570 cases), a test set (164 cases). All LAX images were registered to a template subject using rigid transformation with MIRTK toolkit[5]. All 2D SAX slices have been cropped to the size of 128×128 pixels where the left ventricle is roughly in the center of every image. Benefiting from

[4]The cardiac multi-view image dataset has been provided under UK Biobank Access Application 18545.
[5]https://mirtk.github.io/

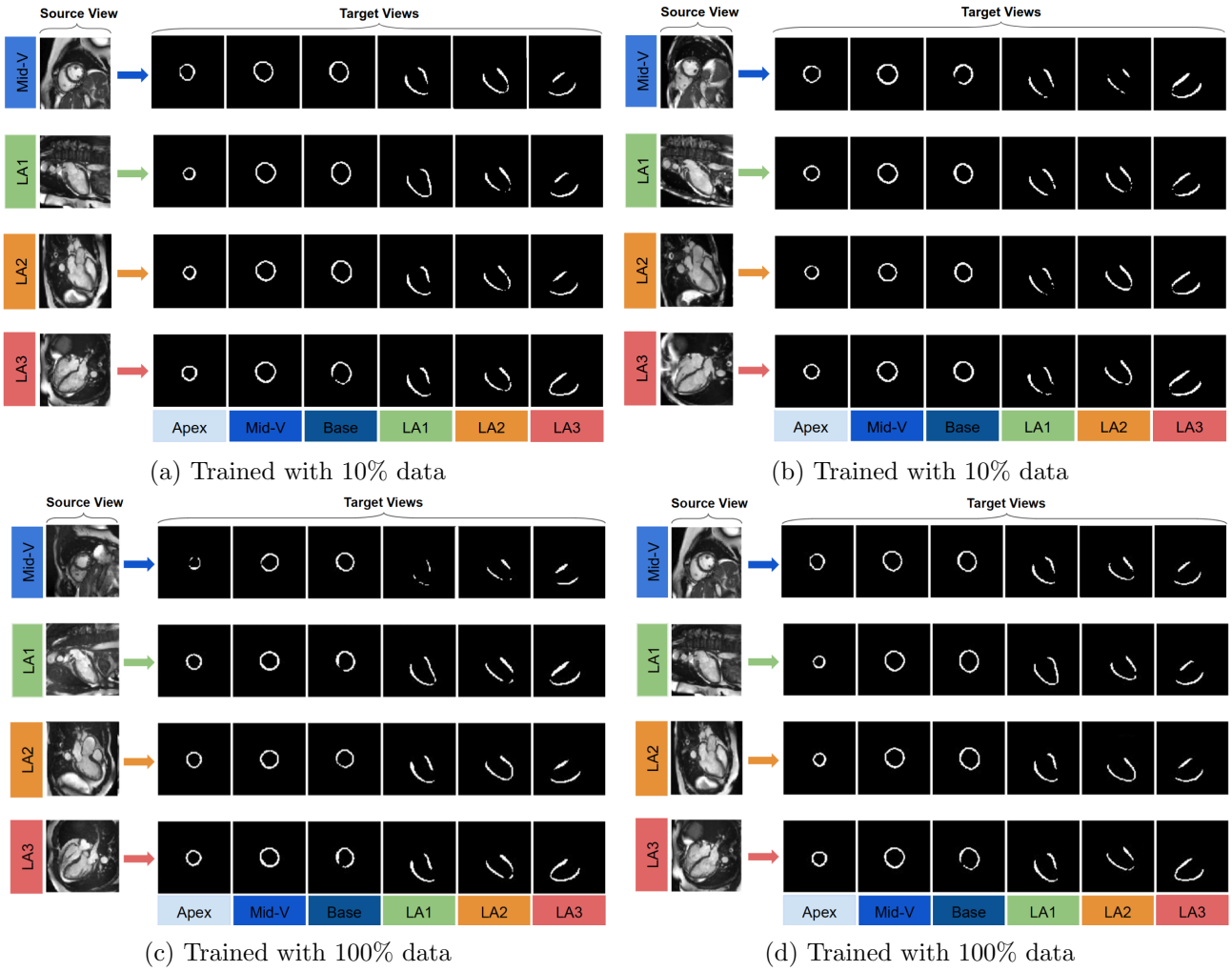
the view planning (which is a standard step during the cardiac image acquisition), we simply use the intersection point of the three orthogonal LAX images on every SAX slice to determine its center of the interest region. All the networks were trained for 200 epochs on an NVIDIA® GeForce® 2080 Ti, using an Adam optimizer with a batch size of 10. The learning rate for Shape MV-CNN was set to 0.0001 whereas the learning rate for the segmentation network was set to 0.001. In our experiments, α was empirically set to 0.5 and β to 0.001 in the $\mathcal{L}_{\text{Shape MAE}}$. The proposed algorithm was implemented in Pytorch.

3.2.5 Results

To evaluate the segmentation accuracy, we use two measurements: the Dice score and the Hausdorff distance (HD). The proposed method is compared against: a 2D U-net [32], a state-of-the-art 2D FCN for cardiac MR image segmentation [4], and a 3D U-net [35]. For fairness and ease of comparison, all models were set with the same number of filters at each level (starting with 16 filters in the first layer) and trained with the same pre-processing and training schedule. For the 3D network, we resampled SAX images to a voxel size of $1.8 \times 1.8 \times 1.8$ mm and cropped each to a size of $128 \times 128 \times 64$ during pre-processing. We trained MV U-net and the baseline networks with two settings: in one case we used **10%** of the training set, while in the other one we used **100%**. Of note, in each setting, we first trained a Shape MV-CNN and then trained a MV U-net where shape priors of four standard views were obtained using corresponding encoders in the Shape MV-CNN.

We visualize the output of the trained Shape MV-CNN network in Fig. 3.8 in the two settings. We can see that given only **one** source view (the first column) as input, the proposed Shape MV-CNN is able to predict the myocardium shapes on the **six** target views (column 2 to column 7). This indicates that the proposed approach has the potential to encode the global shape characteristics of the myocardium in the latent space instead of a local embedding for a particular view of a subject.

Quantitative results on the U-nets' segmentation on the test set are shown in Table 3.3.



(a) Trained with 10% data (b) Trained with 10% data

(c) Trained with 100% data (d) Trained with 100% data

Figure 3.8: **Exemplar results of the proposed shape MV-CNN**. Here, for each training setting, two examples are shown for illustration.

Table 3.3: **Segmentation performance of the baseline models and the proposed method.** Reported values are the mean and the standard deviation of Dice score and HD distance (mm) obtained on the test set (n=164). The comparison has been carried out separately for apical, mid-ventricular, and basal slices.

Method	# Training subjects	Dice↑			HD↓		
		Apex	Middle	Base	Apex	Middle	Base
2D U-net	57 (10%)	0.898 (0.090)	0.932 (0.035)	0.923 (0.077)	3.239 (6.918)	2.337 (2.913)	3.617 (9.058)
2D FCN	57 (10%)	0.873 (0.113)	0.926 (0.041)	0.919 (0.069)	3.088 (3.882)	2.317 (1.440)	2.948 (2.691)
3D U-net	57 (10%)	0.890 (0.083)	0.923 (0.043)	0.923 (0.043)	2.839 (3.980)	3.573 (9.05)	4.469 (10.02)
MV U-net	57 (10%)	**0.905 (0.076)**	**0.932 (0.025)**	**0.926 (0.088)**	**2.487 (3.022)**	**2.093 (0.577)**	**2.758 (3.697)**
2D U-net	570 (100%)	0.937 (0.029)	0.955 (0.016)	0.948 (0.071)	1.917 (0.294)	1.888 (0.178)	2.327 (2.566)
2D FCN	570 (100%)	0.934 (0.032)	0.958 (0.015)	0.949 (0.078)	1.913 (0.297)	1.890 (0.347)	2.161 (1.068)
3D U-net	570 (100%)	0.913 (0.112)	0.945 (0.078)	0.933 (0.093)	2.104 (1.24)	1.957 (0.68)	2.722 (3.57)
MV U-net	570 (100%)	**0.938 (0.027)**	**0.958 (0.013)**	**0.952 (0.079)**	**1.903 (0.345)**	**1.874 (0.142)**	**2.146 (1.004)**
Approx. # of conv weights (million)		2D U-net: 0.8	2D FCN: 1.0	3D U-net: 2.5	MV U-net: 1.2		

From the table, it can be observed that the proposed method outperforms the baseline models in both the low-data setting and the high-data setting, with improved Dice scores at the apex, middle, and base of the left ventricular myocardium. In particular, when only 10% data was

used, the proposed method reduces the mean HD from 3.24 to 2.49 *mm* on the apical slices, from 2.34 to 2.09 on the middle slices and from 3.62 to 2.76 on the basal slices, compared to the 2D U-net. Fig. 3.9 and Fig. 3.10 show examples of the segmentation results from all the networks in both low data setting and high data setting. We can observe that in both data settings, the proposed method not only produces more robust segmentation across slices compared to the results from the 2D networks, but also achieves more anatomically plausible results in comparison to the 3D one (see the red arrows in this figure).

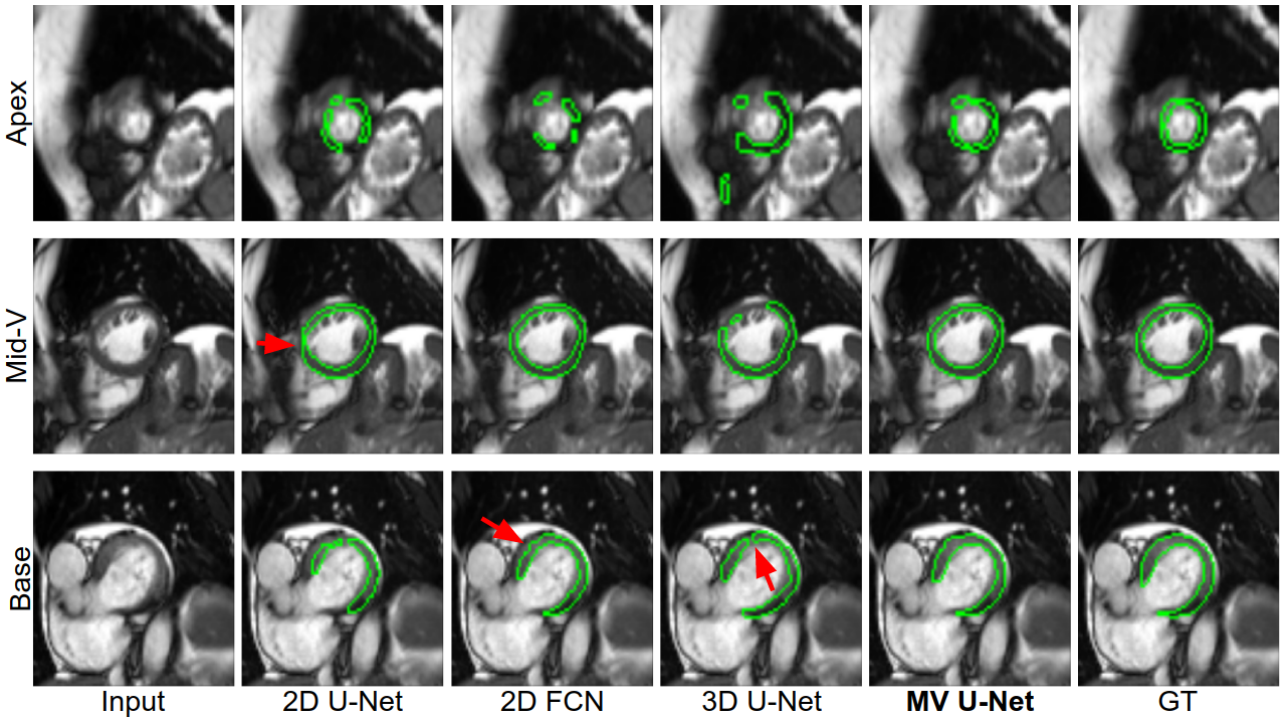


Figure 3.9: **Visualization of ground truth (GT) and corresponding predicted segmentations from the baseline models and MV U-net.** All methods were trained using *10%* training subjects. Here we present predicted segmentation and corresponding GT on an apical, a mid-ventricular, and a basal slice from one patient. Compared to the baseline models, MV U-net produces more accurate segmentation with stronger spatial coherence.

Ablation study. To further quantify the effectiveness of the learned shape priors, we compared the proposed method to its downgraded version by setting priors to be all zeros. Results of this ablation study in the low-data setting is shown in Table 3.4. It is clearly observed that the improvement of segmentation accuracy mainly comes from the learned shape knowledge from auxiliary multiple standard views rather than the increased network capacity.

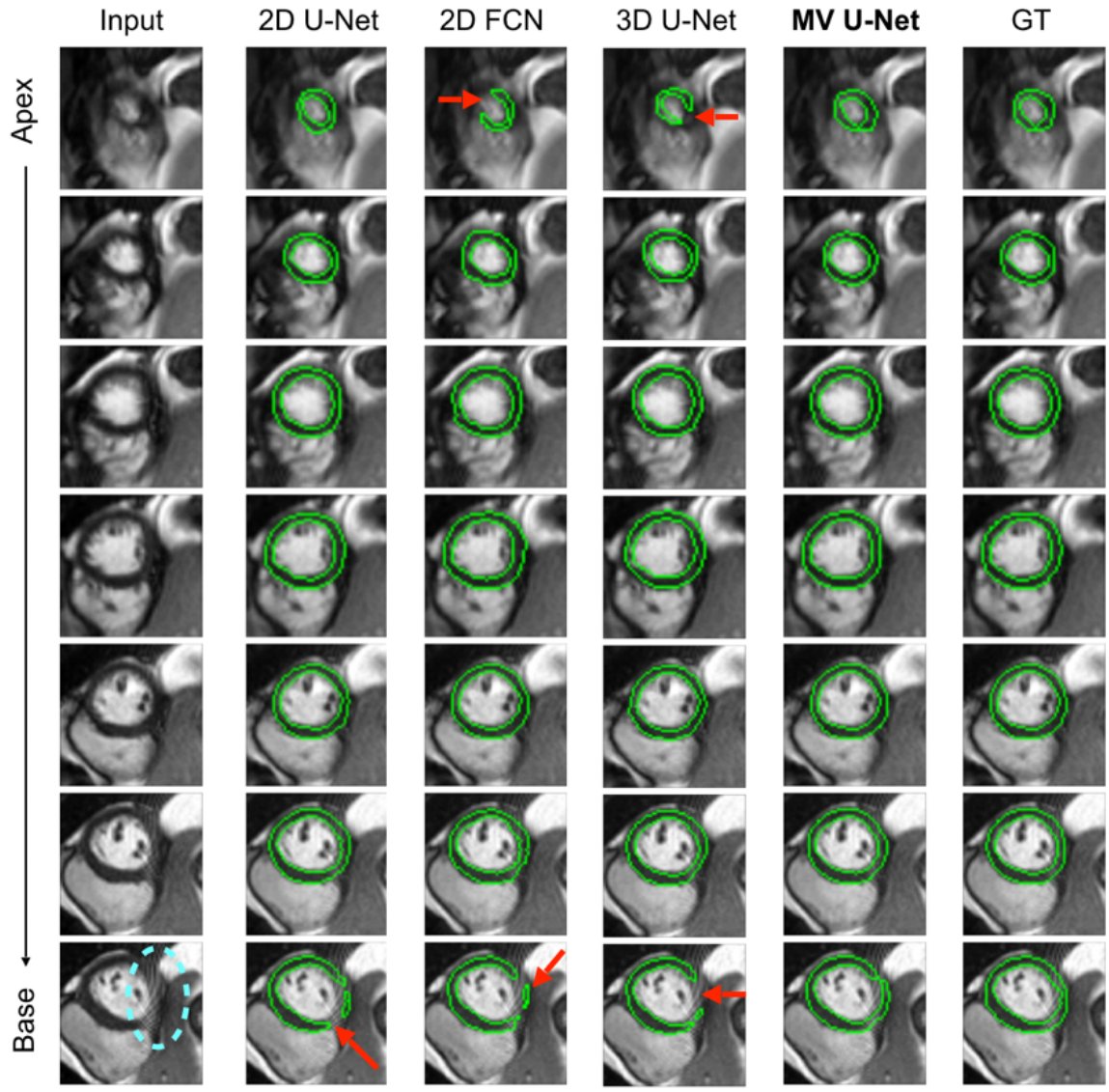


Figure 3.10: **Example results of the proposed segmentation method (MV U-net) and the baseline models.** All methods were trained using *100%* training data. Representative improvements for cardiac image segmentation can be observed when using the proposed method. For example, baseline models produce poor results when there are unexpected artifacts on the image (see the region inside the cyan ellipse). By contrast, the proposed method can properly identify the correct contours.

Table 3.4: **Ablation study results.** Incorporating shape priors into the network improves segmentation accuracy, especially on apical and basal slices. The student's t-test has been conducted to compute p-values for statistical significance analysis.

	Dice↑			HD↓			Overall Dice↑
	Apex	Middle	Base	Apex	Middle	Base	
w/o shape priors	0.898 (0.086)	0.932 (0.032)	0.916 (0.101)	3.542 (10.269)	2.272 (1.844)	3.060 (3.896)	0.905 (0.026)
w/ shape priors	**0.905 (0.076)**	**0.932 (0.025)**	**0.926 (0.088)**	**2.487 (3.022)**	**2.093 (0.577)**	**2.758 (3.697)**	**0.913 (0.021)**
p-value	≤ 0.05	≤ 0.1	≤ 0.05	≤ 0.05	≤ 0.05	≤ 0.05	≤ 0.05

3.2.6 Conclusion

In this work, we presented a shape-aware multi-view CNN, a neural network capable of learning anatomical shape priors from multiple standard views, and a multi-view U-net, a modification of the original U-net architecture that incorporates the learned shape priors to improve the robustness of cardiac segmentation. In contrast to existing works which treat long-axis CMR segmentation and short-axis CMR segmentation as two separate tasks [4, 97], our approach, to the best of our knowledge, is the first trial in deep learning that exploits the spatial context from the long-axis images to guide the segmentation on the short-axis images. The reported experimental results show that the proposed segmentation method not only demonstrates superior segmentation accuracy over state-of-the-art 2D baseline methods [4, 32], but also outperforms a 3D U-net [35]. This improvement is particularly evident on the basal and apical slices in the low-data setting, as expected. When training data is limited, segmenting these challenging slices particularly benefits from the auxiliary anatomical information extracted from the LAX views and injected into the segmentation network.

Of note, our approach does not require a dedicated acquisition protocol since LAX images are routinely acquired in most CMR imaging schemes. Moreover, the proposed MV U-net maintains the computational advantage of a 2D network, using fewer parameters ($\sim$ 1.2 million weights) than the 3D U-net ($\sim$ 2.5 million weights) during training. This advantage also contributes to the data efficiency of our method, achieving high segmentation performance with limited training data. Importantly, our method could be extended in the future to multi-structure cardiac segmentation. The proposed approach could also be potentially adopted to other medical image segmentation tasks. It is also interesting to exploit and compare other approaches that learn and leverage the shape priors from multi-view images, such as a 2.5D network with a multi-branch encoder [160]. We will leave it for future work.

Chapter 4

Learning with Unlabeled Data

This chapter contains material from

1. C. Chen, C. Qin, H. Qiu, C. Ouyang, S. Wang, L. Chen *et al.*, 'Realistic adversarial data augmentation for MR image segmentation,' in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, Springer International Publishing, 2020, pp. 667–677. DOI: 10.1007/978-3-030-59710-8_65 [15]

2. [a] C. Chen, C. Ouyang, G. Tarroni, J. Schlemper, H. Qiu, W. Bai *et al.*, 'Unsupervised multi-modal style transfer for cardiac MR segmentation,' in *Statistical Atlases and Computational Models of the Heart. Multi-Sequence CMR Segmentation - 10th International Workshop, STACOM 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Revised Selected Papers*, ser. Lecture Notes in Computer Science, vol. 12009, Springer International Publishing, 2019, pp. 209–219. DOI: 10.1007/978-3-030-39074-7_22 [14]

[a] This is a joint work with Ouyang Cheng, and both authors contributed equally.

Labeling large datasets for network training requires a considerable amount of resources, time, and effort, limiting the adoption and application of neural networks. In this chapter, we aim to utilize both labeled and unlabeled images to improve model generalization on unseen test data. Compared to a labeled dataset which requires expertise to annotate, an unlabeled dataset

is easier and cheaper to acquire. Therefore it is of great practical value to develop learning algorithms that can utilize unlabeled data to regularize the model for improved performance. In Sec. 4.1, we present an adversarial data augmentation approach, which finds effective bias fields to augment labeled and *unlabeled* data of the same domain. We then utilize these augmented samples for consistency regularization, which encourages the network to produce consistent predictions over similar input images as a way of enhancing model generalization. In Sec. 4.2, we present a GAN-based image style transfer for unsupervised cardiac LGE segmentation, where we have a set of labeled bSSFP images and a set of *unlabeled* LGE images. We utilize the two sets to learn a GAN-based image style translator, so that labeled bSSFP images can be translated into LGE-like images automatically. In this way, we augment the data with a set of synthetic labeled LGE datasets to facilitate the segmentation network training.

4.1 Realistic adversarial data augmentation for MR image segmentation

4.1.1 Introduction

Deep learning-based approaches in general require a large-scale labeled dataset for training, in order to achieve good model generalization ability and robustness on unseen test cases. However, acquiring and manually labeling such large medical datasets is extremely challenging, due to the difficulties that lie in data collection and sharing, as well as to the high labeling costs [161]. To address the aforementioned problems, one of the commonly adopted strategies is data augmentation, which aims to increase the diversity of the available training data without collecting and manually labeling new data. Conventional data augmentation methods mainly focus on applying simple *random* transformations to labeled images. These random transformations include intensity transformations (e.g., pixel-wise noise, image brightness and contrast adjustment) and geometric transformations (e.g., affine, elastic transformations). Recently, there is a growing interest in developing generative network-based methods for data augment-

ation [162–165], which have been found effective for one-shot brain segmentation [162] and low-shot cardiac segmentation [164]. Unlike conventional data augmentation, which generates new examples in an uninformative fashion and does not account for complex variations in data, this generative network-based method is data-driven, learning optimal image transformations from the underlying labeled and unlabeled data distribution in the real world [164]. However, in practice, training generative networks is not trivial due to their sensitivity to hyper-parameters tuning [166] and it can suffer from the mode collapse problem.

In this work, we introduce an effective adversarial data augmentation method for medical imaging without resorting to generative networks. Specifically, we introduce a realistic intensity transformation function to amplify intensity non-uniformity in images, simulating potential image artifacts that may occur in clinical MR imaging (i.e. bias field). Our work is motivated by the observations that MR images often suffer from low-frequency intensity corruptions caused by inhomogeneities in the magnetic field. This artifact cannot be easily eliminated [167, 168] and can be regarded as a physical-world attack to neural networks, which have been reported to be sensitive to intensity perturbations [169, 170]. To efficiently improve the model generalization and robustness, we apply adversarial training to directly search for optimal intensity transformations that benefit model training. This optimization process can be applied to both labeled and unlabeled data. By continuously generating these realistic, 'hard' examples, we prevent the network from over-fitting and, more importantly, encourage the network to defend itself from intensity perturbations by learning robust semantic features for the segmentation task.

Our main contributions can be summarized as follows: (1) We introduce a realistic adversarial intensity transformation model for data augmentation in MRI, which simulates intensity inhomogeneities which are common artifacts in MR imaging. The proposed data augmentation is complementary to conventional data augmentation methods. (2) We present a simple yet effective framework based on adversarial training to learn adversarial transformations and to regularize the network for segmentation robustness, which can be used as a plug-in module in general segmentation networks, see Sec. 4.1.3.2. More importantly, unlike conventional adversarial example construction [109, 171], generating adversarial bias fields does

not require manual labels, which makes it applicable for both supervised and semi-supervised learning. (3) We demonstrate the efficacy of the proposed method on a public cardiac MR segmentation dataset in challenging low-data settings. In this scenario, the proposed method greatly outperforms competitive baseline methods, see Sec. 4.1.5.

4.1.2 Related work

Recent studies have shown that adversarial data augmentation, which generates adversarial data samples during training, is effective to improve model generalization and robustness[132, 171]. Most existing works are based on designing attacks with pixel-wise noise, i.e. by adding gradient-based adversarial noise [109, 169, 172–174]. More recently, there have been studies showing that neural networks can also be fragile to other, more natural form of transformations that can occur in images, such as affine transformations [110, 175, 176], illumination changes [176], and small deformations [170, 177]. In medical imaging, designing and constructing realistic adversarial perturbations, which can be used for improving medical image segmentation networks, has not been explored in depth.

4.1.3 Methodology

In this work, we aim at generating realistic adversarial examples to improve model generalization ability and robustness, given a limited number of training examples and a number of unlabeled images if applicable. To achieve the goal, we first introduce a physics-based intensity transformation model that can simulate intensity inhomogeneities in MR images. We then propose an adversarial training method, which finds effective adversarial transformation parameters to augment training data, and then regularizes the network with a distance loss function which penalizes network's sensitivity to such adversarial perturbations. Since our method is based on virtual adversarial training (VAT) [174], we will first briefly review VAT before introducing our method.

4.1.3.1 Virtual adversarial training

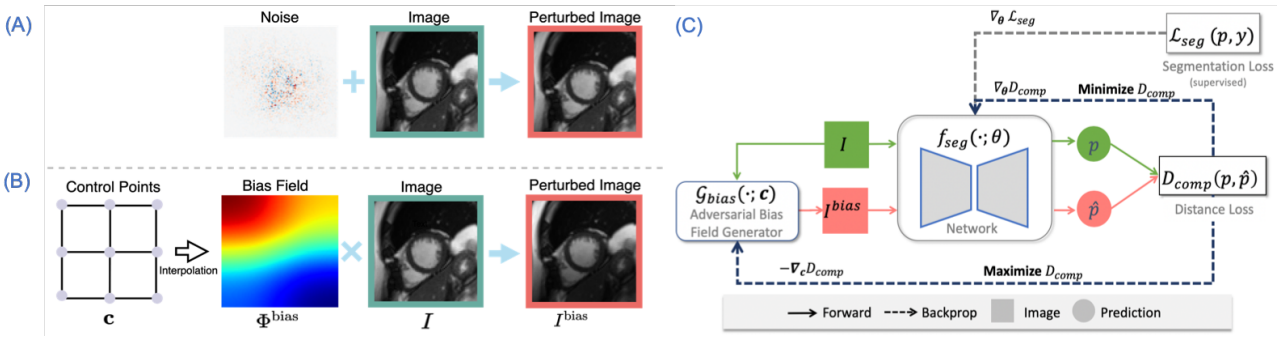


Figure 4.1: **Adversarial example construction and adversarial training.** (A) Adversarial example construction with additive gradient-based noise in VAT [174]; (B) Adversarial example construction with a multiplicative control point-based bias field (proposed); (C) Adversarial training with bias field perturbation.

VAT is a regularization method based on adversarial data augmentation, which can prevent the model from over-fitting and improve the generalization performance and robustness[174]. Given an input image $\boldsymbol{x} \in \mathbb{R}^{H\times W\times C}$ (H,W,C denote image height, width, and number of channels, respectively) and a classification network $f_{cls}(\cdot;\theta)$, VAT first finds a small adversarial noise $\mathbf{r}^{\text{adv}} \in \mathbb{R}^{H\times W\times C}$ to construct its adversarial example $\boldsymbol{x}^{adv} = \boldsymbol{x} + \mathbf{r}^{\text{adv}}$ (as shown in Fig.4.1A), with the goal of maximising the Kullback−Leibler (KL) divergence $\mathcal{D}_{\text{KL}}$ between an original probabilistic prediction $f_{cls}(\boldsymbol{x};\theta)$ and its perturbed prediction $f_{cls}(\boldsymbol{x}+\mathbf{r}^{\text{adv}};\theta)$. The adversarial example is then used to regularize the network for robust feature learning.

The adversarial noise can be generated by taking the gradient of $\mathcal{D}_{\text{KL}}$ with respect to a random noise vector: $\mathbf{r}^{\text{adv}} = \epsilon \cdot \frac{r'}{\|r'\|_2}$,$r' = \nabla_{\mathbf{r}}\mathcal{D}_{\text{KL}}[f(\boldsymbol{x};\theta) \parallel f(\boldsymbol{x}+\mathbf{r};\theta)]$. Here ϵ is a hyper-parameter that controls the strength of perturbation. After finding adversarial examples, one can utilize them for robust learning, which penalizes the network's sensitivity to local perturbations. This is achieved by adding $\mathcal{D}_{\text{KL}}$ to its main objective function.

4.1.3.2 Adversarial training by modeling intensity inhomogeneities

In this work, we extend the VAT approach by introducing a new type of adversarial attack, namely intensity inhomogeneities (bias field) that often occur in MR imaging. In MR imaging, a bias field is a low frequency field that smoothly varies across images, introducing intensity

non-uniformity across the anatomy being imaged. The model for the intensity non-uniformity can be defined as follows [167, 178]: $\boldsymbol{x}^{\text{bias}} = \mathcal{G}_{\text{bias}}(\boldsymbol{x}; \mathbf{c}) = \boldsymbol{x} \times \Phi^{\text{bias}}(\mathbf{c})$. Here, the intensity of the image $\boldsymbol{x}$ is perturbed with a multiplication with the bias field $\Phi^{\text{bias}} \in \mathbb{R}^{H \times W}$. As the bias field is typically composed of low frequencies and thus slowly varying across the image, it can be modelled using a set of uniformly distributed k by k points $\mathbf{c} = \{\mathbf{c}_{(i)}\}_{1 \ldots k \times k}$ [167], see Fig. 4.1B. A smooth bias field at the finest resolution is obtained by interpolating scattered control points with a third-order B-spline smoothing [179].

While one can repeatedly sample random bias fields for data augmentation, this might be computationally inefficient as it may generate images which are of no added value for model optimization. We therefore would like to construct adversarial examples (perturbed by bias field as described above) targeting the weakness of the network in an intelligent way. This allows the use of the generated adversarial examples to improve the model performance and robustness, which can be achieved via the following min-max game:

$$\begin{aligned} &\min_{\theta} \max_{\mathbf{c}} \quad \mathcal{D}_{\text{comp}}[f_{seg}(\boldsymbol{x}; \theta), f_{seg}(\mathcal{G}_{\text{bias}}(\boldsymbol{x}; \mathbf{c}); \theta)] \\ &\text{subject to} \quad \forall (x, y) \in \mathbb{R}^2,\ \Phi^{\text{bias}}_{(x,y)} > 0;\ |\Phi^{\text{bias}} - \mathbf{1}|_{\infty} \leq \alpha, 0 < \alpha < 1. \end{aligned} \tag{4.1}$$

As shown in Fig. 4.1C, given a segmentation network $f_{seg}(\cdot; \theta)$ and an input image $\boldsymbol{x}$, we first find optimal values for control points $\mathbf{c}$ in the search space to construct an adversarial bias field, so that it **maximizes** the distance measured by $\mathcal{D}_{\text{comp}}$ between the original prediction and the prediction after perturbation: $\mathbf{p} = f_{seg}(\boldsymbol{x}; \theta), \hat{\mathbf{p}} = f_{seg}(\mathcal{G}_{\text{bias}}(\boldsymbol{x}; \mathbf{c}); \theta)$, with θ fixed. We then optimize the parameters θ in the network to **minimize** the distance between the original prediction and the prediction after the generated adversarial bias attack $f_{seg}(\mathcal{G}_{\text{bias}}(\boldsymbol{x}; \mathbf{c}^{adv}); \theta)$.

4.1.3.3 Finding adversarial bias fields

To find the optimal values for the control points $\mathbf{c}$ for adversarial example construction, we use the gradient descent algorithm and search the values of control points in its log space for numerical stability [167, 178], which allows to produce positive bias fields. Specifically, similar to the projected gradient decent (PGD) attack construction in [171], we first randomly initialize

the values of control points and then apply a projected gradient ascent algorithm to iteratively update $\mathbf{c}$ with n steps: $\mathbf{c} \leftarrow \Pi(\mathbf{c}+\xi\cdot\mathbf{c}'/\|\mathbf{c}'\|_2)$ where $\mathbf{c}' = \nabla_{\mathbf{c}}\mathcal{D}_{\text{comp}}[f_{seg}(\boldsymbol{x};\theta), f_{seg}(\mathcal{G}_{\text{bias}}(\boldsymbol{x};\mathbf{c});\theta)]$. Π denotes the projection function which projects $\mathbf{c}$ onto the feasible set, and ξ is the step size. For neural networks, gradients $\mathbf{c}'$ can be efficiently computed with back-propagation. Φ^{bias} is updated by first interpolating the coarse-grid control points (log values at the current iteration) to its finest grid using B-spline convolution, and then taking the exponential function for value recovering. Finally, the generated bias field is rescaled to meet the magnitude constraint in Eq. 4.1.

4.1.3.4 Composite distance function

Here, we propose a composite distance function $\mathcal{D}_{\text{comp}}$ to enhance its discrimination ability between the original prediction $\mathbf{p}$ (short for $f_{seg}(\boldsymbol{x};\theta)$) and the prediction after perturbation $\hat{\mathbf{p}}$, for *semantic segmentation* tasks. This composite loss consists of (1) the original $\mathcal{D}_{\text{KL}}$ used in VAT, which measures the difference between distributions and (2) a contour-based loss function $\mathcal{D}_{\text{contour}}$ [14] which is specifically designed to capture mismatch between object boundaries: $\mathcal{D}_{\text{comp}}(\mathbf{p},\hat{\mathbf{p}}) = \mathcal{D}_{\text{KL}}[\mathbf{p} \,||\, \hat{\mathbf{p}}] + w\mathcal{D}_{\text{contour}}(\mathbf{p},\hat{\mathbf{p}})$; $\mathcal{D}_{\text{contour}}(\mathbf{p},\hat{\mathbf{p}}) = \sum_{m\in M}\sum_{S\in\{S_x,S_y\}} \|S(\mathbf{p}^m) - S(\hat{\mathbf{p}}^m)\|_2$. M denotes foreground channels, S_x, S_y denote two Sobel filters in x- and y-direction for edge extraction and w controls the relative importance of both terms.

4.1.3.5 Optimizing segmentation network

After constructing the adversarial examples, one can compute $\mathcal{D}_{\text{comp}}$ and apply it to regularizing the network, encouraging the network to be less sensitive to adversarial perturbations, and thus produce consistent predictions. Since this algorithm uses probabilistic predictions (produced by the network) rather than manual labels for adversary construction, it can be applied to both labeled (l) and unlabeled data (u) for supervised and semi-supervised learning [174]. The loss functions for the two scenarios are defined as: $\mathcal{L}_{\text{SU}} = \mathcal{L}_{seg}(\mathbf{p}^{(l)}, \mathbf{y}_{gt}^{(l)}) + \lambda_l\mathcal{D}_{\text{comp}}(\mathbf{p}^{(l)}, \hat{\mathbf{p}}^{(l)})$; $\mathcal{L}_{\text{SE}} = \mathcal{L}_{\text{SU}} + \lambda_u\mathcal{D}_{\text{comp}}(\mathbf{p}^{(u)}, \hat{\mathbf{p}}^{(u)})$. $\mathcal{L}_{seg}$ denotes a general task-related segmentation loss function for supervised learning (e.g., cross-entropy loss) and $\mathbf{y}_{gt}^{(l)}$ denotes ground truth.

4.1.4 Experiments

To test the efficacy of the proposed method, we applied it to training a segmentation network for the left ventricular myocardium from MR images in low-data settings. We compared the results with several competitive baseline methods.

ACDC dataset. Experiments were performed on a public benchmark dataset for cardiac MR image segmentation: The Automated Cardiac Diagnosis Challenge (ACDC) dataset [6] [1]. This dataset was collected from 100 subjects which were evenly classified into 5 groups: 1 normal group (NOR) and 4 pathological groups with cardiac abnormalities: dilated cardiomyopathy (DCM); hypertrophic cardiomyopathy (HCM); myocardial infarction with altered left ventricular ejection fraction (MINF); abnormal right ventricle (ARV). The left ventricular myocardium in end-diastolic and end-systolic frames were manually labeled.

Image pre-processing. We used the same image preprocessing as in [164]. In addition, all images were centrally cropped into 128×128, given that the heart is generally located in the center of the image. This saves computational costs.

Random data augmentation (Rand Aug). We applied a strong random data augmentation method to our training data as a basic setting. Random affine transformation (i.e. scaling, rotation, translation), random horizontal and vertical flipping, random global intensity transformation (brightness and contrast) [164] and elastic transformation were applied.

Training details. For ease of comparison, same as [164], we adopted the commonly-used 2D U-net as our segmentation network, which takes 2D image slices as input. The Adam optimizer with a batch size of 20 was used to update network parameters. For the proposed method, we first trained the network with the default data augmentation (Rand Aug) for 10,000 iterations (learning rate=$1e^{-3}$), and then finetuned the network by adding the proposed adversarial training using a smaller learning rate ($1e^{-5}$) for 2,000 iterations. The common standard cross-entropy loss function was used as $\mathcal{L}_{seg}$. For bias field construction, we adopted the B-spline convolution kernel (order=3) with 4×4 control points. The kernel was provided by AirLab library [180]. We empirically set: $\alpha = 0.3$, $w = 0.5$, $\lambda_l = 1$ and $\lambda_u = 0.1$. Besides,

[1] https://www.creatis.insa-lyon.fr/Challenge/acdc/databases.html

we found that in our experiments, one step searching in the inner loop produced sufficient improvement. Thus, we set $n = 1, \xi = 1$ to save computational cost. All the experiments were performed on an Nvidia® GeForce® 2080 Ti with Pytorch. Our code is available on the GitHub [2].

4.1.5 Results

4.1.5.1 Experiment 1: low-shot learning

In this experiment, the proposed method was evaluated in both *supervised* learning and *semi-supervised* learning scenarios, where only 1 or 3 labeled subjects are available. Specifically, we used the same data splitting setting as in [164]. The ACDC dataset was split into 4 subsets: a labeled set (where N_l images were sampled from for training), unlabeled training set (N=25), validation set (N=2), test set (N=20). N denotes the number of subjects. Details of the low-data setting can be found in [164]. For one-shot learning (N_l=1) and three-shot learning (N_l=3) in both supervised and semi-supervised settings, we trained the network for five times, each with a different labeled set.

We compared the proposed method (**Adv Bias**) with several competitive data augmentation methods including **VAT** [174], an effective data mixing-based method (**Mixup**) [181] for supervised learning and the state-of-the-art semi-supervised generative model-based method(**cGANs**) [164]. For VAT and Mixup, we used the set of hyperparameters that achieved the best performance on the validation set and applied the same training procedure. For cGANs, we report the results of one-shot and three-shot learning in their original paper for reference, which were tested on the same test set. Table 4.1 compares the segmentation accuracy obtained by different data augmentation methods.

In the supervised learning setting (no access to unlabeled images), when only one or three labeled subject was available, the proposed method clearly outperformed all baseline methods. For semi-supervised learning, the proposed methods outperformed VAT, especially when only

[2] https://github.com/cherise215/AdvBias

Table 4.1: **Segmentation performance of the segmentation network using the proposed method (Adv Bias) and other data augmentation methods.** Each reported value is the average Dice score of 20 test cases.

Setting	Method	# labeled subjects 1	3
Supervised	No Aug	0.293	0.544
	Rand Aug	0.560	0.796
	+Mixup[181]	0.575	0.801
	+VAT[174]	0.570	0.811
	+Adv Bias	**0.650**	**0.826**
Semi-supervised	+VAT[174]	0.625	0.826
	+Adv Bias	0.692	**0.830**
	cGANs[164]	**0.710**	0.823

Table 4.2: **Segmentation performance of the proposed method and baseline methods across five populations.** All were trained with NOR cases only. Reported values are the average Dice score of each test population.

Population	Rand Aug	+Mixup	+VAT	**+Adv Bias** (Proposed)
NOR	0.911	0.901	0.909	**0.912**
DCM	0.831	0.803	0.843	**0.871**
HCM	0.871	0.881	**0.891**	0.890
MINF	0.805	0.789	0.824	**0.847**
ARV	0.843	0.844	0.843	**0.853**
Average	0.841	0.833	0.853	**0.868**

one labeled subject is available (0.686 vs 0.625). The proposed method achieves competitive results compared to the semi-supervised GAN-based method (cGANs) as well. Of note, cGANs adopts two additional GANs to sample geometric transformations and intensity transformations from unlabeled images. This is why it was only compared in the semi-supervised learning setting here. On the contrary, our approach is applicable to both low-shot supervised learning and semi-supervised learning. In addition, cGANs contains more parameters than our method and thus it is less computationally efficient.

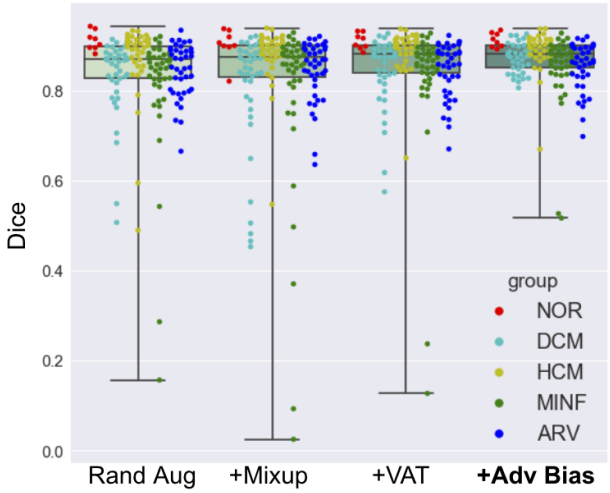


Figure 4.2: **Boxplots of the segmentation results across five different populations.** Each dot represents the Dice score for each test subject, and its color indicates its group. Our method (column 4) produces more accurate segmentation on **unseen** pathological cases than the baselines. This indicates that the proposed method can improve the model robustness for abnormal cases, even the network was only trained with normal cases (NOR).

4.1.5.2 Experiment 2: learning from limited population

In this experiment, we trained the network using only normal healthy subjects (NOR) and evaluated its performance on pathological cases (80 cases in total). 20 healthy subjects were split into 14/2/4 subjects for training, validation and test. This setting simulates a practical data scarcity problem, where pathological cases are rarer, compared to healthy data. As shown in Table 4.2 and corresponding box-plots in Fig. 4.2, while the conventional method (Rand Aug) achieved excellent performance on the test healthy subjects (NOR), its performance dropped on pathological cases. Interestingly, applying Mixup did not help to solve this population shift problem, but rather slightly reduced the average performance compared to the baseline, from 0.841 to 0.833. This might be due to the fact that Mixup generates unrealistic images through its linear combination of paired images, which may modify semantic features and affect representation learning for *precise* segmentation. By contrast, our method outperformed both Mixup and VAT, yielding substantial and consistent improvements across five different populations. Notably, we attained evident improvement on the most challenging MINF images (0.805 vs 0.847), where the shape of the myocardium is clearly irregular. As shown in Fig. 4.3, the proposed method not only generates adversarial examples during training, but also increases

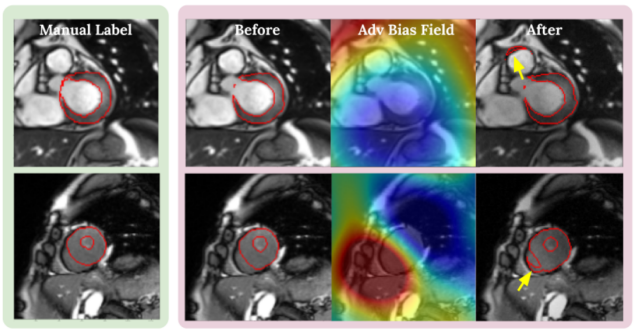


Figure 4.3: **Visualization of generated adversarial examples and failed network predictions.** Before/After: network prediction before/after bias field attack (Adv Bias Field).

the variety of image styles while preserving the shape information. Augmenting images with various styles can encourage the network to learn high-level shape-based representation instead of texture-based representation, leading to improved network robustness on unseen classes, as discussed in [182]. By contrast, VAT only introduces imperceptible noise, failing to model realistic image appearance variations.

4.1.5.3 Ablation study

Table 4.3: **Random bias field vs Adversarial bias field**

Method	Distance Loss	Dice↑	HD ↓	VolumeSim ↑
Rand Bias	$\mathcal{D}_{\text{comp}}$	0.852	6.25	0.941
Adv Bias	$\mathcal{D}_{\text{comp}}$	**0.868**	**5.91**	**0.957**

HD: Hausdorff distance; VolumeSim: Volume similarity index [183]. Reported values are average scores across all test subjects from five populations ($20 \times 4 + 4 = 84$ subjects). The same applies to Table 4.4.

To get a better understanding of the effectiveness of adversarial bias field, we compared it to data augmentation using random bias field, using experiment setting 2. Results clearly showed that training with adversarial bias field improved the model generalization ability, increasing the Dice score from 0.852 to 0.868, (see Table 4.3). As visualized in Fig. 4.4, while the difference between the random and the adversarial bias field is mild, the proposed method is stronger at attacking the network. Therefore, adding these adversarial examples during training will encourage the network to learn more robust features for precise segmentation.

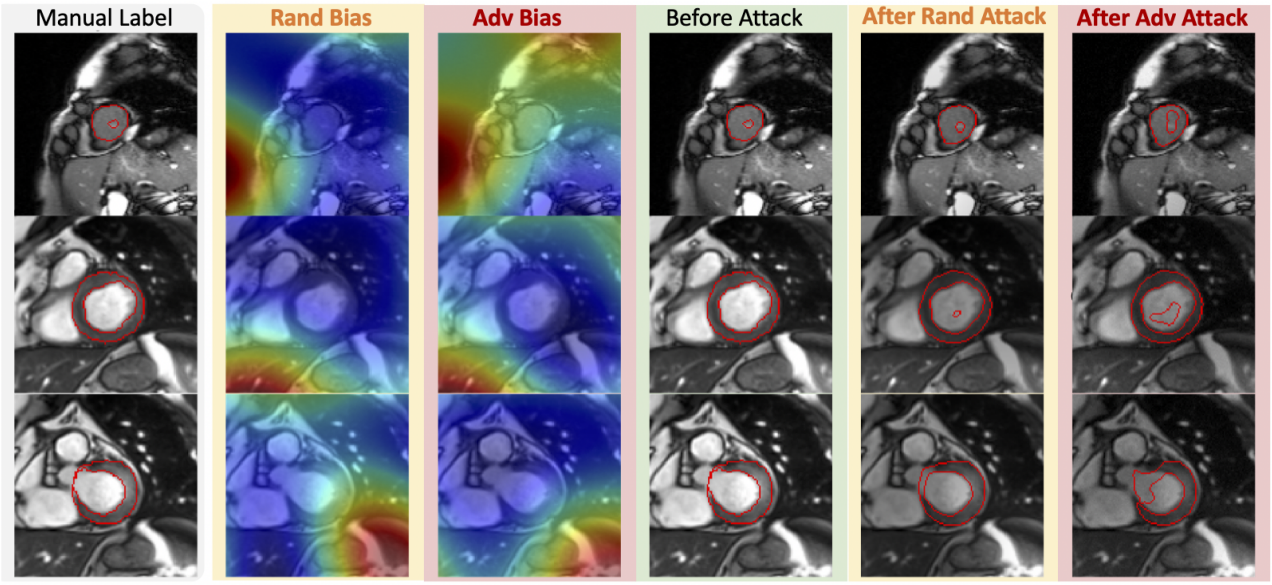


Figure 4.4: **Performance of adversarial bias field attack (Adv Bias) vs random bias field attack (Rand Bias).**

On the other hand, applying $\mathcal{D}_{\text{comp}}$ to regularize the network improved the average Dice score from 0.859 to 0.868, compared to the one trained with only $\mathcal{D}_{\text{KL}}$ (see Table 4.4). Unlike random-based approach, constructing adversarial attacks considers both the posterior probability information estimated by the model and semantic information from images. In experiments, we found these attacks focused on attacking challenging images on which the network was uncertain, e.g., object boundary is not clear or there is another similar structure presented, see Fig. 4.3. In the same spirit of online hard example mining, utilizing these borderline examples during training helps the network to improve its generalization and robustness ability.

Table 4.4: $\mathcal{D}_{\text{comp}}$ **vs** $\mathcal{D}_{\text{KL}}$**.**

Method	Distance Loss	Dice↑	HD↓	VolumeSim ↑
VAT	$\mathcal{D}_{\text{KL}}$	0.853	6.678	0.949
VAT	$\mathcal{D}_{\text{comp}}$	0.856	6.331	0.946
Adv Bias	$\mathcal{D}_{\text{KL}}$	0.859	6.330	0.949
Adv Bias	$\mathcal{D}_{\text{comp}}$	**0.868**	**5.912**	**0.957**

4.1.6 Discussion and conclusion

In this work, we presented a realistic adversarial data augmentation method to improve the generalization and robustness of neural network-based medical image segmentation methods.

We demonstrated that by modeling the bias field and introducing adversarial learning, the proposed method could promote learning robust semantic features for cardiac image segmentation. This method can be used in both supervised and semi-supervised settings, leveraging unlabeled data to improve generalization. It can also alleviate the data scarcity problem, as demonstrated in the low-data setting and cross-population experiments. The proposed method does not rely on generative networks but instead employs a small set of explainable and controllable parameters to augment data with image appearance variations that are realistic for MR. It can be easily extended for multi-class segmentation and used in general segmentation networks for improving model generalization and robustness. In this work, we only consider constructing bias fields. It is also worthwhile to model other domain-specific intensity artifacts to increase the variety of data augmentation, such as motion artifacts [184].

4.2 Unsupervised multi-modal style transfer for cardiac MR segmentation

4.2.1 Introduction

Cardiac segmentation from late-gadolinium enhanced (LGE) cardiac magnetic resonance (CMR) images which highlights myocardial infarcted tissue is of great clinical importance, enabling quantitative measurements useful for treatment planning and patient management. To this end, the segmentation of the myocardium is an important first step for myocardial infarction analysis.

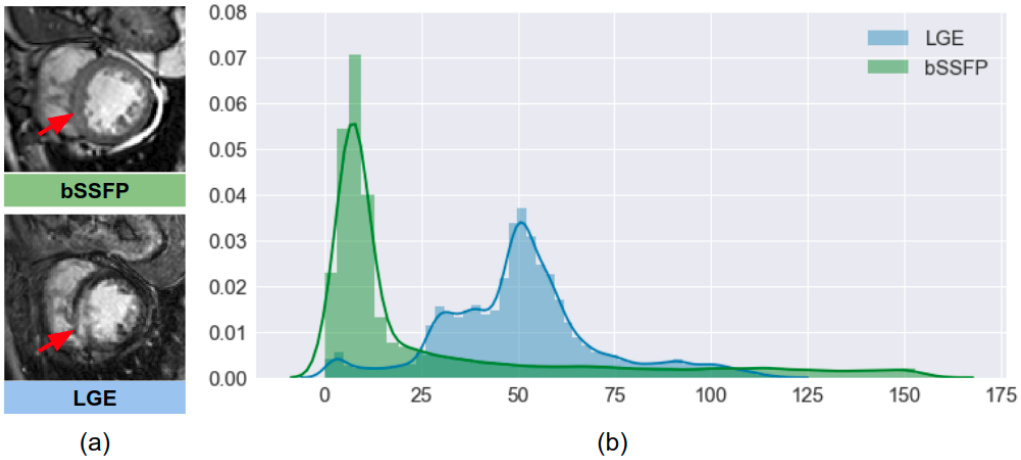


Figure 4.5: **The differences of (a) image appearance and (b) intensity distributions in the cardiac region between LGE images and bSSFP images.** Here the cardiac region covers the union of LV, MYO, and RV.

Since manual segmentation is tedious and likely to suffer from inter-observer variability, it is of great interest to develop an accurate automated segmentation method. However, this is a challenging task due to the fact that 1) the infarcted myocardium presents an enhanced and heterogeneous intensity distribution different from the normal myocardium region and 2) the border between infarcted myocardium and blood pool appears blurry and ambiguous [185]. While the borders of the myocardium can be difficult to delineate on LGE images, they are clear and easy to identify on the balanced steady-state free precession (bSSFP) CMR images, which have high signal-to-noise ratio and whose contrast is less sensitive to pathology (see red arrows in Fig. 4.5 (a)). In clinical practice, it is common to acquire both bSSFP images

and LGE images for patients that suffer from myocardial infarction, where bSSFP imaging captures cardiac motions with clear cardiac structures and LGE imaging highlights the infarcts over the cardiac region. Conventional methods [186, 187] use the segmentation result from the bSSFP CMR of the same patient as prior knowledge to assist the segmentation on LGE CMR images. These methods generally require accurate registration between the bSSFP and LGE images, which can be challenging as the imaging field-of-view (FOV), image contrast and resolution between the two acquisitions can vary significantly [185, 188]. Fig. 4.5 (b) visualizes the discrepancy between the intensity distributions of the two imaging modalities in the cardiac structures (specifically, left ventricle (LV), myocardium (MYO), and right ventricle (RV)).

Most recently, a deep neural network-based method has been proposed to segment the three cardiac structures directly from LGE images [48], reporting superior performance. However, this supervised segmentation method requires a large amount of labeled LGE data. Because of the heterogeneous intensity distribution of the myocardium in LGE images and the scarcity of experienced image analysts, it is difficult to perform accurate manual segmentations on LGE images and collect a large training set, compared to that on bSSFP images.

In this work, we present a fully automatic framework that addresses the above mentioned issues by training a segmentation model without using manual annotations on LGE images. This is achieved by transferring the anatomical knowledge and features learned on annotated bSSFP images, which are easier to acquire. Specifically, given a set of labeled bSSFP images, and a set of unlabeled LGE images, a generative image style translation network is trained to model the conditional image distribution, so that labeled bSSFP images can be translated into LGE-like images automatically. We then use these synthetic LGE images to train a network for LGE image segmentation. Our framework mainly consists of two neural networks:

- A GAN-based multi-modal image translation network: this network is used for translating annotated bSSFP images into LGE images through style transfer. Of note, the network is trained in an unsupervised fashion where the training bSSFP images and LGE images are **unpaired**. In addition, unlike common one-to-one translation networks, this network allows the generation of **multiple** synthetic LGE images conditioned on a single bSSFP

image;

- A cascaded segmentation network for LGE images consisting of two U-net [32] models (Cascaded U-net): Inspired by curriculum learning [149], the segmentation network is first trained using the labeled bSSFP images and then fine-tuned using the synthetic LGE data generated by the image translation network. This allows the network to transfer the learned shape knowledge from the easy task to the hard task for improved model generalization.

The main contributions of our work are the following: 1) we employ a translation network that can generate **realistic** and **diverse** synthetic LGE images given a single bSSFP image. This network enables generative model-based data augmentation for unsupervised domain adaptation, which not only closes the domain gap between the two modalities, but also improves the generalization properties of the following segmentation network by increasing data variety; 2) we demonstrate that the proposed two-stage cascaded network, which takes both anatomical **shape** information and image **appearance** information into account, produces accurate segmentation on LGE images, greatly outperforming baseline methods; 3) the proposed framework can be easily extended to other unsupervised cross-modality domain adaptation applications where labels of one modality are not available.

4.2.2 Methodology

The proposed method aims at learning an LGE image segmentation model using labeled bSSFP $\{(\mathbf{x}_b, \mathbf{y}_b)\}$ and unlabeled LGE $\{\mathbf{x}_l\}$ only. Specifically, the proposed method is a two-stage framework. In the first stage, an unsupervised **image translation** network is trained to translate each bSSFP image $\mathbf{x}_b$ into multiple instances of LGE-like images, noted as $\{\mathbf{x}_{bl}\}$. In the second stage, these LGE-stylized bSSFP images are used together with their original labels $\{(\mathbf{x}_{bl}, \mathbf{y}_b)\}$ to adapt an **image segmentation** network pre-trained on labeled bSSFP images to segment LGE images.

4.2.2.1 Image translation

We employ the state-of-the-art multi-modal unsupervised image-to-image translation network (MUNIT) [189] as our multi-modal image translator. Let $\{\mathbf{x}_l\}$ and $\{\mathbf{x}_b\}$ denote unpaired images from the two different imaging modalities (domains): LGE and bSSFP, given an image drawn from one domain as input, the network is able to change the appearance (i.e. image style) of the image to that of the other domain while preserving the underlying anatomical structure [190]. This is achieved by learning disentangled image representations.

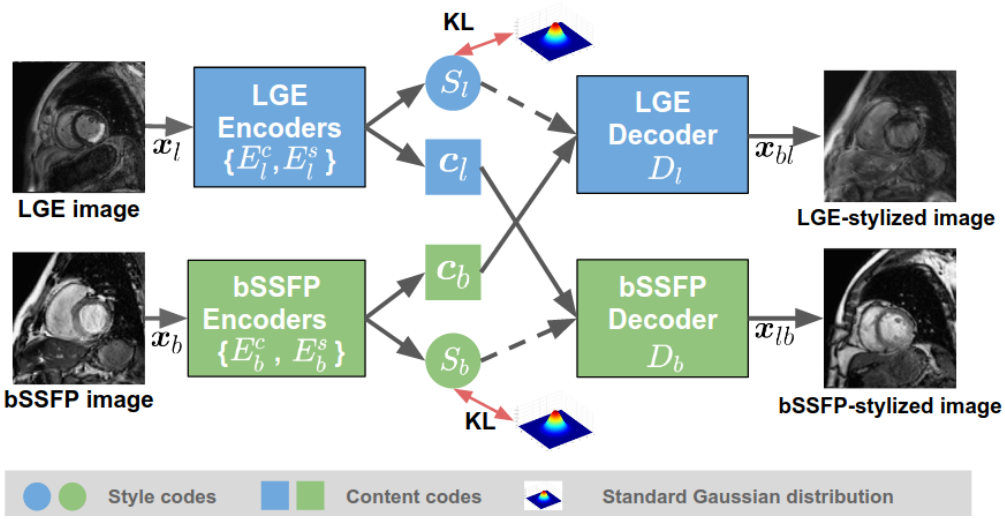


Figure 4.6: **Overview of the multi-modal image translation network.** The network employs the structure of MUNIT [189], which consists of two encoder-decoder pairs for the two domains: bSSFP and LGE, respectively.

As shown in Fig. 4.6, each image $\mathbf{x}$ is disentangled into (a) a domain-invariant content code $\boldsymbol{c}$: $\boldsymbol{c} = E^c(\mathbf{x})$ and (b) a domain-specific style code $\boldsymbol{s}$: $\boldsymbol{s} = E^s(\mathbf{x})$ using the content encoder E^c and the style encoder E^s relative to its domain where the content code captures the anatomical structure and the style code carries the information for rendering the structure which is determined by the imaging modality. The image-to-image translation from one domain to the other is achieved by swapping latent codes in two domains. For example, translating a bSSFP image $\boldsymbol{x_b}$ to be stylized as LGE, is achieved by feeding the content code $\boldsymbol{c_b}$ for the bSSFP image and the style code $\boldsymbol{s_l}$ into the LGE decoder D_l: $\mathbf{x}_{bl} = D_l(\mathbf{c}_b, \mathbf{s}_l)$.

We apply a bidirectional reconstruction loss to train the image translation network, con-

sisting of an image reconstruction loss computed on the image space $\boldsymbol{x}$:

$$\mathcal{L}_{recon}^{\boldsymbol{x}} = \mathbb{E}_{\boldsymbol{x}_l \in \{\boldsymbol{x}_l\}} \left[\| D_l \left(E_l^c \left(\boldsymbol{x}_l \right), E_l^s \left(\boldsymbol{x}_l \right) \right) - \boldsymbol{x}_l \|_1 \right] + \mathbb{E}_{\boldsymbol{x}_b \in \{\boldsymbol{x}_b\}} \left[\| D_b \left(E_b^c \left(\boldsymbol{x}_b \right), E_b^s \left(\boldsymbol{x}_b \right) \right) - \boldsymbol{x}_b \|_1 \right], \quad (4.2)$$

and two latent code reconstruction losses computed on $\mathbf{c}$ and $\mathbf{s}$, respectively. They are:

$$\mathcal{L}_{\text{recon}}^{\mathbf{c}} = \mathbb{E}_{\mathbf{c}_l \sim p(\mathbf{c}_l), \mathbf{s}_b \sim q(\mathbf{s}_b)} \left[\| E_b^c \left(D_b \left(\mathbf{c}_l, \mathbf{s}_b \right) \right) - \mathbf{c}_l \|_1 \right] + \mathbb{E}_{\mathbf{c}_b \sim p(\mathbf{c}_b), \mathbf{s}_l \sim q(\mathbf{s}_l)} \left[\| E_l^c \left(D_l \left(\mathbf{c}_b, \mathbf{s}_l \right) \right) - \mathbf{c}_b \|_1 \right], \quad (4.3)$$

$$\mathcal{L}_{\text{recon}}^{\mathbf{s}} = \mathbb{E}_{\mathbf{c}_l \sim p(\mathbf{c}_l), \mathbf{s}_b \sim q(\mathbf{s}_b)} \left[\| E_b^s \left(D_b \left(\mathbf{c}_l, \mathbf{s}_b \right) \right) - \mathbf{s}_b \|_1 \right] + \mathbb{E}_{\mathbf{c}_b \sim p(\mathbf{c}_b), \mathbf{s}_l \sim q(\mathbf{s}_l)} \left[\| E_l^s \left(D_l \left(\mathbf{c}_b, \mathbf{s}_l \right) \right) - \mathbf{s}_l \|_1 \right]. \quad (4.4)$$

Minimizing $\mathcal{L}_{\text{recon}}^{\mathbf{c}}$ and $\mathcal{L}_{\text{recon}}^{\mathbf{s}}$ forces the network to produce the same latent code (style and content) on reconstructed images to the one that is used for image reconstruction, respectively.

We further employ GANs [49] to ensure that the distribution of translated images matches the distribution of target domain. This is achieved by employing two discriminator networks $\mathcal{F}_b$, $\mathcal{F}_l$ that learn to distinguish between translated images and real images in their corresponding domains. We apply the adversarial training to the translation network and the two discriminators wherein the translation network is optimized to minimize the following two adversarial losses. The two discriminators are optimized to maximize their corresponding losses:

$$\mathcal{L}_{\text{GAN}}^{\boldsymbol{x}_{lb}} = \mathbb{E}_{\mathbf{c}_l \sim p(\mathbf{c}_l), \mathbf{s}_b \sim q(\mathbf{s}_b)} \left[\log \left(1 - \mathcal{F}_b \left(D_b \left(\mathbf{c}_l, \mathbf{s}_b \right) \right) \right) \right] + \mathbb{E}_{\boldsymbol{x}_b \sim p(\boldsymbol{x}_b)} \left[\log \mathcal{F}_b \left(\boldsymbol{x}_b \right) \right], \quad (4.5)$$

$$\mathcal{L}_{\text{GAN}}^{\boldsymbol{x}_{bl}} = \mathbb{E}_{\mathbf{c}_b \sim p(\mathbf{c}_b), \mathbf{s}_l \sim q(\mathbf{s}_l)} \left[\log \left(1 - \mathcal{F}_l \left(D_l \left(\mathbf{c}_b, \mathbf{s}_l \right) \right) \right) \right] + \mathbb{E}_{\boldsymbol{x}_l \sim p(\boldsymbol{x}_l)} \left[\log \mathcal{F}_l \left(\boldsymbol{x}_l \right) \right]. \quad (4.6)$$

The total loss is a weighted sum of the adversarial losses and the bidirectional reconstruction losses. The encoders and decoders in the translation network and the two discriminators are jointly trained via the min-max optimization:

$$\min_{E_l, E_b, D_l, D_b} \max_{\mathcal{F}_l, \mathcal{F}_b} \mathcal{L} \left(E_l, E_b, D_l, D_b, \mathcal{F}_l, \mathcal{F}_b \right) = \mathcal{L}_{\text{GAN}}^{\boldsymbol{x}_{lb}} + \mathcal{L}_{\text{GAN}}^{\boldsymbol{x}_{bl}} + \lambda_x \mathcal{L}_{recon}^{\boldsymbol{x}} + \lambda_c \mathcal{L}_{recon}^{\mathbf{c}} + \lambda_s \mathcal{L}_{recon}^{\mathbf{s}}. \quad (4.7)$$

Here, $\lambda_x, \lambda_c, \lambda_s$ are coefficients that control the importance of corresponding reconstruction terms.

Of note, during training, each style encoder is trained to embed images into a latent space that matches the standard Gaussian distribution $\mathcal{N}(0, I)$, minimizing the Kullback-Leibler (KL)

divergence between the two. This allows to generate an arbitrary number of synthetic LGE images $\boldsymbol{x}_{bl}$ given a single bSSFP image during inference. Although this prior distribution is unimodal, the distribution of translated images in the output space is multi-modal thanks to the nonlinearity of the decoder[189]. For more details about training the translation network, readers are referred to the original work by Huang *et al.* [189].

4.2.2.2 Image segmentation

Let $\boldsymbol{x}_l$ be an observed LGE image, the aim of the segmentation task is to estimate label maps $\boldsymbol{y}_l$ having observed $\boldsymbol{x}_l$ by modeling the posterior $p(\boldsymbol{y}_l|\boldsymbol{x}_l)$. Inspired by curriculum learning [149] and transfer learning, we first train a segmentation network using annotated bSSFP images (source domain; easy examples) and then fine-tune it to segment LGE images (target domain; hard examples). Since labeled LGE images $\{(\mathbf{x}_l, \mathbf{y}_l)\}$ are not available for finetuning, we use a synthetic dataset $\mathcal{X}_{bl} : \{(\mathbf{x}_{bl}, \mathbf{y}_b)\}_{1..N}$ generated by the aforementioned multi-modal image translator. Specifically, given the labelled bSSFP set, we can generate a synthetic labeled LGE dataset $\mathcal{X}_{bl} : \{(\mathbf{x}_{bl}, \mathbf{y}_b)\}_{1..N}$, where $\mathbf{x}_{bl}$ is a reconstructed image using the original content code $\mathbf{c}_b$ from a labeled bSSFP image $\boldsymbol{x}_b$: $\mathbf{c}_b = E_b(\mathbf{x}_b)$ and a randomly sampled style code s_l drawn from $\mathcal{N}(0, I)$:$\mathbf{x}_{bl} = D_l(\mathbf{c}_b, \mathbf{s}_l)$, $\mathbf{s}_l \sim \mathcal{N}(0, I)$. Here, N is the number of sampling. Ideally, the posterior modeled by the network $p(\boldsymbol{y}_b|\boldsymbol{x}_{bl})$ matches $p(\boldsymbol{y}_l|\boldsymbol{x}_l)$ when image space and label space are shared. For simplicity, we use $\boldsymbol{x}$ and $\boldsymbol{y}$ to denote an image and its corresponding label map from the synthetic dataset in the following paragraphs.

The segmentation network is a two-stage cascaded network which consists of two U-nets [32], see Fig. 4.7. Specifically, given an image $\boldsymbol{x}$ as input, the first U-net (U-net 1) aims at predicting four-class pixel-wise probabilistic maps $\boldsymbol{p_1} = f^1_{\text{U-net}}(\boldsymbol{x}; \theta)$ for the three cardiac structures (i.e. LV, MYO, RV) and the background class (BG). Inspired by the auto-context architecture [191], we combine these learned probabilistic maps $\boldsymbol{p_1}$ from the first network with the raw image $\boldsymbol{x}$ to form a 5-channel input to train the second U-net (U-net 2) for fine-grained segmentation: $\boldsymbol{p_2} = f^2_{\text{U-net}}(\boldsymbol{x}, \boldsymbol{p_1}; \phi)$. By combining the appearance information from the image $\mathbf{x}$ with the shape prior information from the initial segmentation $\boldsymbol{p_1}$ as input, the cascaded network has

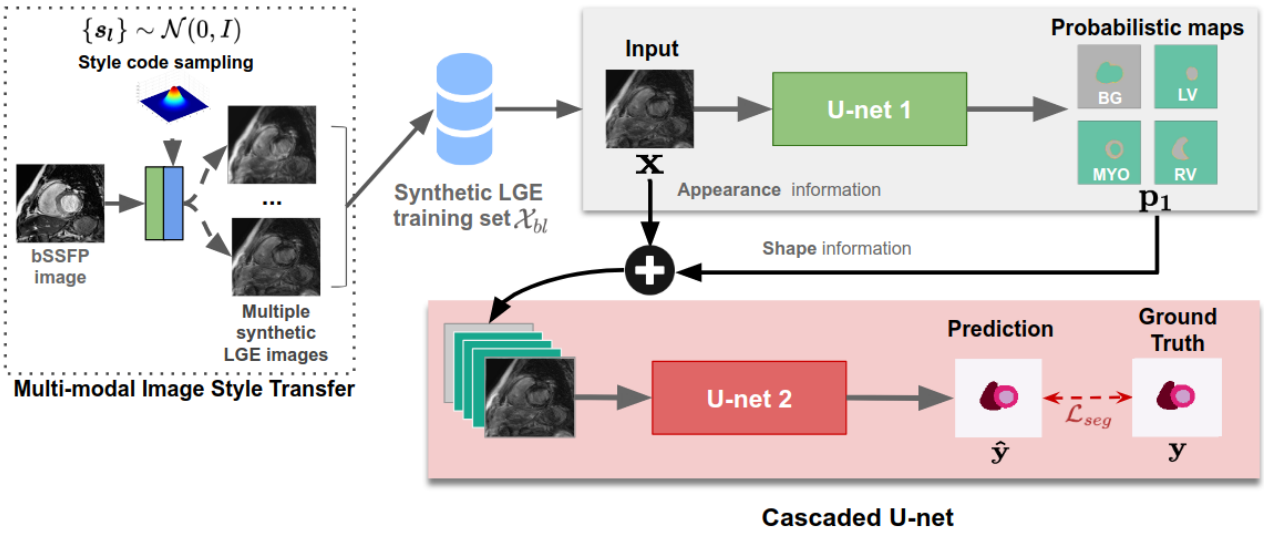


Figure 4.7: **Overview of the two-stage cascaded segmentation network.** The architecture of each U-net is the same as the one of the vanilla U-net [32], except for two main differences: (1) batch normalization is applied after each convolutional layer; (2) a dropout layer (dropout rate=0.1) is applied after each concatenation operation in the network's expanding path to encourage model generalizability. Of note, in this diagram, we simplify the training procedure by omitting the pre-training procedure using labeled bSSFP images.

the potential to produce more precise and robust segmentations even in the presence of unclear boundaries for the different cardiac structures.

To train the network, we use a **composite segmentation loss** function $\mathcal{L}_{seg}$ which consists of two loss terms:

$$\mathcal{L}_{seg} = \mathcal{L}_{wce} + \lambda \mathcal{L}_{edge}. \tag{4.8}$$

The first term $\mathcal{L}_{wce}$ is a weighted cross entropy loss:

$$\mathcal{L}_{wce} = -\sum_{m} \omega^m \boldsymbol{y}^m \log\left(\boldsymbol{p}^m\right) \tag{4.9}$$

where w^m denotes the weight for class m and $\boldsymbol{p}^m$ is the corresponding predicted probability map. We set the weight for myocardium ω^{MYO} to be higher than the weights for the other three classes to address class imbalance problem since there is a lower percentage of pixels that corresponds to the myocardium class in CMR images. The second term $\mathcal{L}_{edge}$ is an edge-based loss which penalizes the disagreement on the contours of the cardiac structures. Specifically, we apply two 2D 3×3 Sobel filters [192] S_k (k=1,2) to the soft prediction maps $\boldsymbol{p}$ as well as

the one-hot heatmaps $\boldsymbol{y}$ of the ground truth to extract edge information along horizontal and vertical directions.

The edge loss is then computed by calculating the l_2 distance between the predicted edge maps and the ground truth edge maps: $\mathcal{L}_{edge} = \sum_{m,m\neq BG} \sum_{k=1,2} \|f_{S_k}(\boldsymbol{p}^m) - f_{S_k}(\boldsymbol{y}^m)\|_2^2$, where $f_{S_k}(\boldsymbol{p}^m)$ is the edge map extracted by applying the sobel filter S_k to the predicted probabilistic map $\boldsymbol{p}^m$ for foreground class m.

By using the edge loss together with the weighted cross entropy for optimization, the network is encouraged to focus more on the contours of the three structures and the myocardium, which are usually more difficult to delineate. In our experiments, we set $\lambda = 0.5$ to balance the contribution of the two losses.

4.2.2.3 Post-processing

At inference time, each slice from a previously unseen LGE stack is fed to the cascaded network to get the probabilistic maps for the four classes. Dense conditional random field (CRF) [193] is then applied to refine the 2D predicted segmentation mask slice by slice. After that, 3D morphological dilation and erosion operations are applied to the whole segmentation stack to further improve the global smoothness. In particular, we perform the operations in a hierarchical order: first we apply them to the binary map covering all the three structures, then to the MYO and the LV labels, separately.

4.2.3 Experiments

4.2.3.1 Data

The framework was trained and evaluated on the 2019 Multi-sequence Cardiac MR Segmentation Challenge (MS-CMRSeg) dataset[3]. We used a subset of 40 bSSFP and 40 LGE images to train the image translation network. Then, we created a synthetic dataset by applying the

[3] https://zmiclab.github.io/mscmrseg19/

learned translation network to 30 labeled bSSFP images. Specifically, for each bSSFP image, we randomly sampled the style code from $\mathcal{N}(0, I)$ five times ($N = 5$), resulting in a set of 150 synthetic LGE images in total. This synthetic dataset and the original 30 bSSFP images with corresponding labels formed the training set for the segmentation network. Exemplar results of these synthetic LGE images are provided in the supplemental material. For validation, we used a subset of 5 annotated LGE images provided by the challenge organizers. Our method was finally tested on a hold-out test set with 40 cases.

4.2.3.2 Implementation details

Image preprocessing. To deal with the different image size and heterogeneous pixel spacing between different imaging modalities, all images were resampled to a pixel spacing of $1.25\ mm \times 1.25\ mm$ and then cropped to 192×192 pixels, with the heart roughly at the center of each image. This spatial normalization would reduce the computational cost and task complexity in the following training procedure of image translation and segmentation, making the networks focus on the relevant regions. To identify the heart, we trained a localization network based on U-net using the 30 annotated bSSFP images in the training set to produce rough segmentations for the three structures. The localization network employs instance normalization layers which perform style normalization [194], encouraging the network invariance to image style changes (e.g., image contrast). As a result, the network is able to produce coarse masks localizing the heart on all bSSFP images and most LGE images even though it was trained on bSSFP images only. In case that this network might fail to locate the heart on certain LGE slices, we summed the segmentation masks across slices in each volume and then cropped them according to the center of the aggregated mask. After cropping, each image was intensity normalized.

Network training. (1) For the image translation network, we used the official implementation[4] of [189]. Network configuration and hyper-parameters were kept the same as in [189] except the

[4] https://github.com/NVlabs/MUNIT

input and output images are 2D, single-channel. It was trained for 20k iterations with a batch size of 1. (2) For the segmentation network, we first trained the first U-net with the labeled bSSFP images and then fine-tuned it with synthetic LGE images. This procedure was replicated to train the second U-net with the parameters of the first U-net being fixed. Both networks were optimized using the composite loss $\mathcal{L}_{seg}$ where the Adam optimization algorithm [65] was used for stochastic gradient descent. The learning rate was initially set to 0.001 and was then decreased to 1×10^{-5} for fine-tuning. The weights for BG, LV, MYO, and RV in $\mathcal{L}_{wce}$ were empirically set to $0.2 : 0.25 : 0.3 : 0.25$. During training, we applied data augmentation on the fly. Specifically, elastic deformations, random scaling and random rotations as well as gamma augmentation [12] were used. The algorithm was implemented using python and PyTorch and was trained for 1000 epochs in total on an NVIDIA® Tesla P40 GPU.

4.2.4 Results and discussion

To evaluate the accuracy of segmentation results, the Dice metric and the average surface distance (ASD) between the automatic segmentation and the corresponding manual segmentation for each volume were calculated.

We compare the proposed method with two baseline methods: (1) a registration-based method and (2) a single U-net. Specifically, for the registration-based method, each LGE segmentation result was obtained by directly registering the corresponding bSSFP labels to the LGE image using MIRTK toolkit [5] for ease of comparison. The transformation matrix was learned by applying mutual information-based registration (Rigid+Affine+FFD) between the two images. For U-net, we trained it with two settings: a) **U-net**: trained on labeled bSSFP images only; b) **U-net with fine-tuning (FT)**: trained on labeled bSSFP images and then fine-tuned using the synthetic LGE data, which is the same training procedure of the proposed method. Quantitative and qualitative results are shown in Table 4.5 and Fig. 4.9.

While the registration-based method (MIRTK) outperforms the U-net (see row 1 and row 2 in Table 4.5), it still fails to produce accurate segmentation on the myocardium (see the red

[5] https://mirtk.github.io/

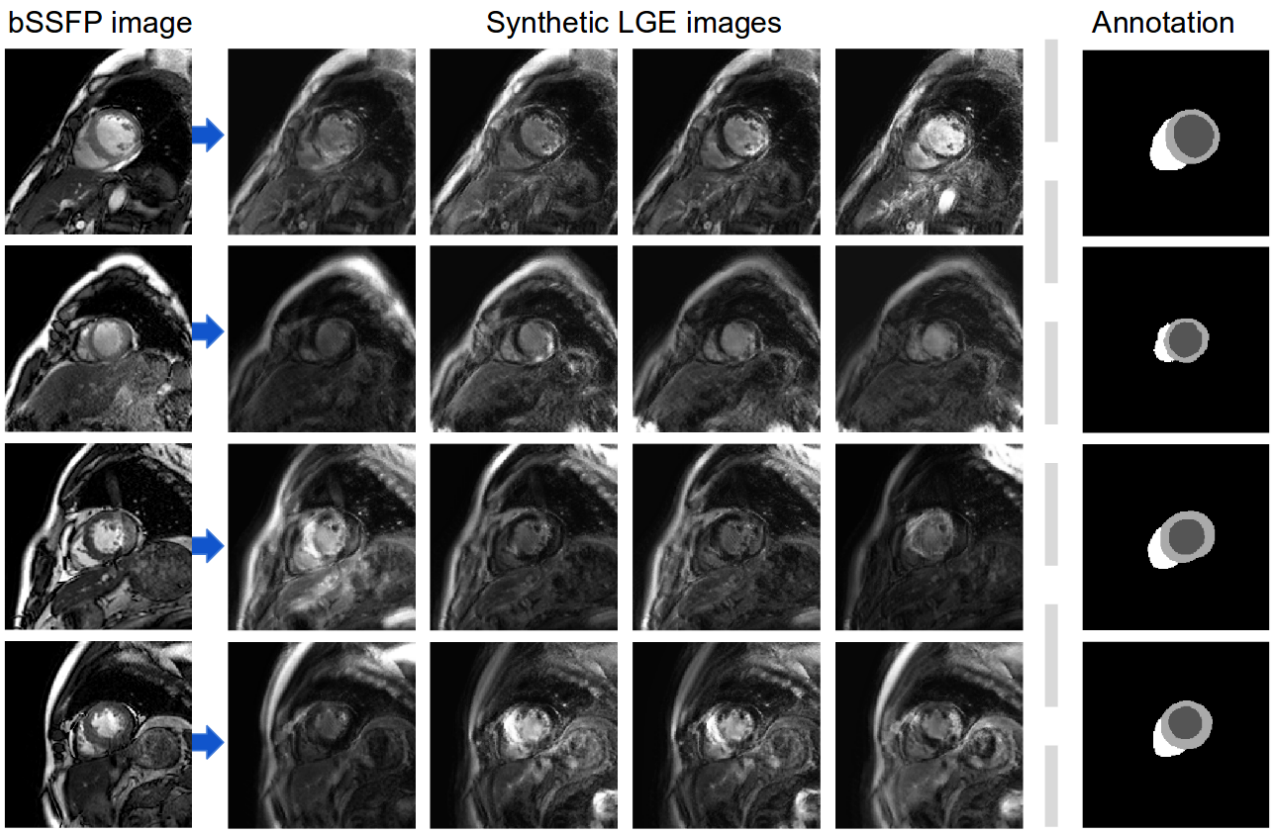


Figure 4.8: **Exemplar synthetic LGE images generated by the multi-modal image translation network This is achieved by reconstructing images using the same content code from the given labeled bSSFP images and a style code repeatedly sampled from the uniform Gaussian distribution.** Given **one** bSSFP image (column 1), the translation network translates the image into *multi-modal* LGE-like images (column 2 to 4). These translated images differ in image brightness, contrast, and intensity distribution in the cardiac region while preserving the same cardiac anatomy. Together with the annotations on the original bSSFP images (the last column), these synthetic images contribute to the synthetic dataset used to fine-tune the proposed segmentation network.

number in row 1), indicating the limitation of this registration-based method. However, by contrast, neural network-based methods (row 3-5) fine-tuned using the *synthetic LGE dataset* significantly improves the segmentation accuracy, increasing the Dice score for MYO by $\sim 15\%$. This improvement demonstrates the learned translation network is capable of generating realistic LGE images while preserving the domain-invariant structural information that is informative to optimize the segmentation network, see Fig. 4.8. In particular, compared to U-net (FT), the proposed **Cascaded U-net** (FT) achieves more accurate segmentation performance with improvement in terms of both Dice and ASD (see blue numbers). The model even produces robust segmentation results on the challenging apical and basal slices (please see the last column in Fig. 4.9). This demonstrates the benefit of integrating the high-level

shape knowledge and low-level image appearance to guide the segmentation procedure. In addition, the proposed post-processing further refines the segmentation results through smoothing, reducing the average ASD from 1.37 to 1.26 (see the last row in Table 4.5).

Table 4.5: **Segmentation performance of the proposed segmentation method (Cascaded U-net) and baseline methods on the validation set.** Reported values are the mean Dice scores and ASD (mm). Blue numbers indicate the best scores among the results obtained by those methods before post-processing (PP) whereas red numbers are those mean Dice scores under 0.700. FT: fine-tuning using the synthetic LGE dataset. N/A means that the ASD value cannot be calculated due to missing predictions for that cardiac structure.

Method	**Dice↑**				**ASD↓**			
	LV	MYO	RV	AVG*	LV	MYO	RV	AVG*
MIRTK	0.819	0.665	0.831	0.772	2.56	1.65	2.11	2.11
U-net	0.624	0.441	0.577	0.547	10.03	6.07	N/A	N/A
U-net (FT)	0.874	0.781	0.896	0.850	1.78	1.50	1.28	1.52
Cascaded U-net (FT)	0.895	0.812	0.898	0.868	1.41	1.46	1.23	1.37
Cascaded U-net (FT) + PP	0.897	0.816	0.895	0.869	1.17	1.42	1.18	1.26

* For ease of comparison, we calculate the average (AVG) Dice score and the average ASD score over the three structures for each method.

Finally, we applied ensemble learning to improve our model's performance in the test phase. Specifically, we trained the proposed segmentation network for multiple times, each time regenerating a new synthetic LGE dataset for fine-tuning. We trained four models in total. Our final submission result for each test image was obtained by averaging the probabilistic maps from these models and then assigning to each pixel the class with the highest score. In the testing stage of the competition, the method achieves very promising segmentation performance on a relative large test set (40 subjects), with an average Dice score of 0.92 for LV, 0.83 for MYO, and 0.88 for RV; an ASD of 1.66 for LV, 1.76 for MYO, and 2.16 for RV.

4.2.5 Conclusion

In this work, we utilized labeled bSSFP images and *unlabeled* LGE images to learn a multi-modal image translation network for data augmentation. We showed that synthesizing multi-modal LGE images from labeled bSSFP images to fine-tune a pre-trained segmentation network shows impressive segmentation performance on LGE images even though the network has not seen *real* labeled LGE images before. We also demonstrated that the proposed segmentation

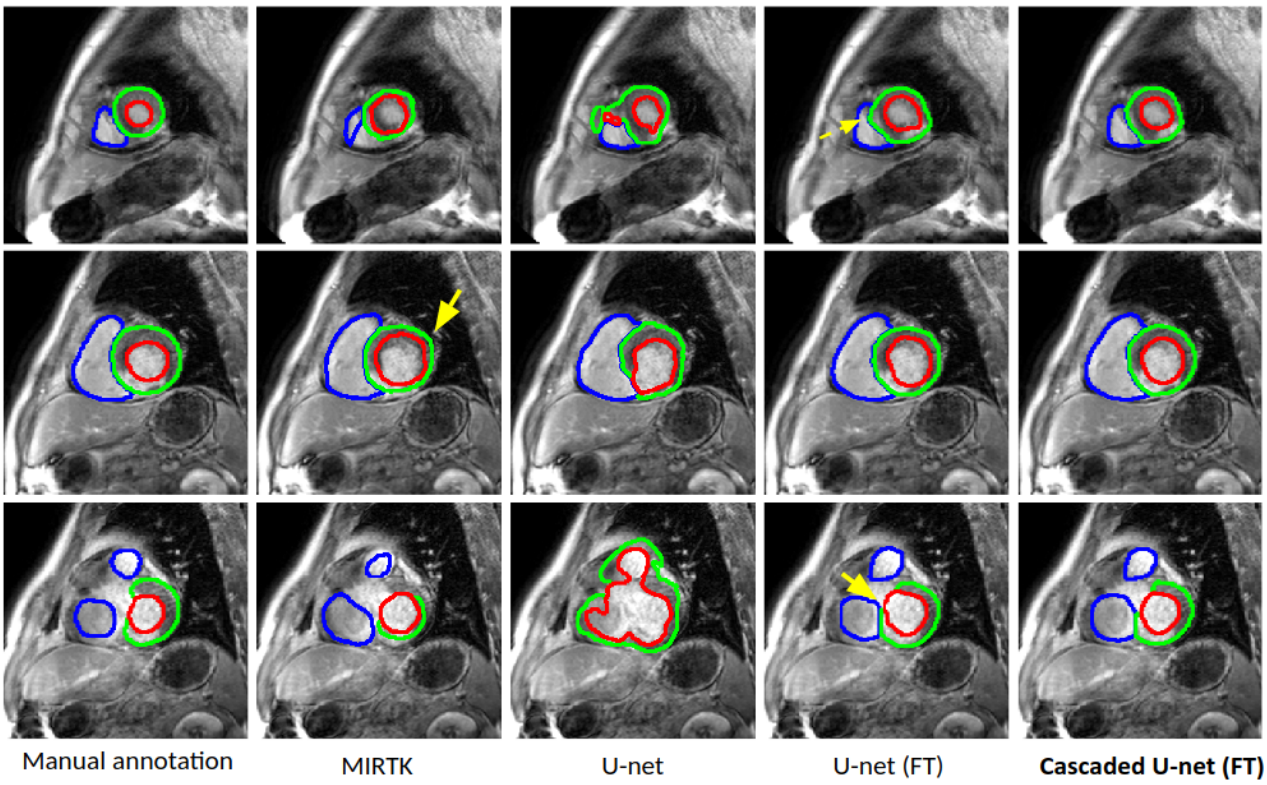


Figure 4.9: **Visualization of segmentation results produced by the proposed Cascaded U-net and the baseline approaches.** Our proposed method (the right-most column) produces more anatomically plausible segmentation results on the images, greatly outperforming the baseline methods, especially in the challenging cases: the apical (the top row) and the basal slices (the bottom row).

network (Cascaded U-net) outperformed the baseline methods by a significant margin, suggesting the benefit of integrating the high-level shape knowledge and low-level image appearance to guide the segmentation procedure. More importantly, our cascaded segmentation network is independent of the particular architecture of underlying convolutional neural networks. In other words, the basic neural network (U-net) in our work can be replaced with any state-of-the-art segmentation network to improve prediction accuracy and robustness potentially. Moreover, the proposed solution based on unsupervised multi-modal style transfer is not only limited to the cardiac image segmentation but can be extended to other multi-modal image analysis tasks where the manual annotations of one modality are not available. Future work will focus on the application of the method to the problems such as domain adaptation for multi-modality brain segmentation. The current limitation of the proposed method is that it still requires sufficient bSSFP images and LGE images (unlabelled) in order to avoid the discriminator overfitting and allow the generator to reconstruct images with diverse image appearance. In our case, we used

40 training subjects of each sequence. Training GAN with limited data (e.g. 5 subjects) is not easy and is still an active research area. Potential solutions such as incorporating advanced regularization on the discriminator [195] in together with effective data augmentation [196] could be adopted in the current framework to alleviate this problem.

Chapter 5

Learning From Limited Data

This chapter contains material from

1. C. Chen, W. Bai, R. H. Davies, A. N. Bhuva, C. H. Manisty, J. B. Augusto *et al.*, 'Improving the generalizability of convolutional neural Network-Based segmentation on CMR images,' *Frontiers in Cardiovascular Medicine*, vol. 7, p. 105, Jun. 2020, ISSN: 2297-055X. DOI: 10.3389/fcvm.2020.00105 [7]

2. C. Chen, K. Hammernik, C. Ouyang, Q. Chen, W. Bai and D. Rueckert, 'Co-operative training and latent space data augmentation for robust segmentation,' in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, arXiv:2107.01079, Springer International Publishing, 2021 [16]

In this chapter, we are concerned with a worst-case scenario in model generalization, where a model aims to perform well on many unseen domains while there is only one dataset collected from a single domain (e.g., one scanner, one hospital) for training. Learning robust networks from single-domain data and limited data is of great practical value for medical imaging research. Due to data privacy issues, as well as the high data storing and collection costs, it is likely that only data from a single site is available for training. To alleviate the data scarcity problem, a natural solution is to apply data augmentation as a way of increasing the size and diversity of training data. In this chapter, we explore two ways of data augmentation: image

space data augmentation and latent space data augmentation. In Sec. 5.1, we present a general training/testing pipeline with a proper design of data normalization and image space data augmentation to improve a CNN-based cardiac segmentation generalization ability. We demonstrated that this method can achieve good segmentation accuracy across images from various unseen scanners at different sites despite the training images being collected from only one scanner. However, this method still has some limitations, such as high sensitivity to images with poor quality, e.g., images with artifacts. In Sec. 5.2, we present a latent space data augmentation approach to enhance cross-domain generalization further, and in particular, demonstrate its improved model robustness against unseen imaging artifacts.

5.1 Improving the generalizability of convolutional neural network-based segmentation on CMR images

5.1.1 Introduction

Automatic cardiac segmentation algorithms provide an efficient way for clinicians to assess the structure and function of the heart from CMR images for the diagnosis and management of a wide range of abnormal heart conditions [6]. Recently, CNN-based methods have become state-of-the-art techniques for automated cardiac image segmentation [2, 6]. However, related work [4] has shown that the segmentation accuracy of a CNN may degrade if the network is directly applied to images collected from different scanners or sites. For instance, CMR images from different scanners using different acquisition protocols can exhibit differences in terms of noise levels, image contrast, and resolution [197–199]. Moreover, images coming from different sites may comprise different population demographics in terms of cardiovascular diseases, resulting in the clinically appreciable difference not only in cardiac morphology but also in image quality (e.g., irregular heartbeat can affect image quality) [197, 200, 201]. Thus, a CNN learned from a limited dataset may not be able to generalize over subjects with heart conditions outside of the training set. All these differences pose challenges for deploying CNN-based image segmentation

algorithms in real-world practice.

In general, a straightforward way to address this problem is to fine-tune a CNN learned from one dataset (source domain) with additional labeled data from another dataset (target domain). Nevertheless, collecting sufficient pixel-wise labeled medical data for every scenario can be difficult, since it requires domain-specific knowledge and intensive labor to perform manual annotation. To alleviate the labeling cost, unsupervised deep domain adaptation (UDDA) approaches have been proposed [202]. Compared to fine-tuning, UDDA does not require labeled data from the target domain. Instead, it only uses either feature-level information [203–205] or image-level information [205] to optimize the network performance on the target domain. However, these methods usually require hand-crafted hyper-parameter tuning for each scenario, which may be difficult to scale to highly heterogeneous datasets. Therefore, it is of great interest to explore how to learn a network that can be successfully applied to other datasets without the requirement of additional model tuning.

In this work, we investigate the possibility of building a generalisable model for cardiac MR image segmentation, given a training set from only one scanner in a single site. Instead of fine-tuning or adapting to get a new model for each particular scenario, our goal is to find a generalizable solution that can analyze 'real-world' test images collected from multiple sites and scanners. These images consist of various pathologies and cardiac morphologies that may not be present in the training set, reflecting the complexity of a real-world clinical setting. To achieve this goal, we choose the U-net [32] as the fundamental CNN architecture, which is the most popular network for medical image segmentation. We apply this network to segment the cardiac anatomy from CMR images (short-axis view), including the left ventricle (LV), the myocardium (MYO), and the right ventricle (RV). An image pre-processing pipeline is proposed to normalize images across sites before feeding them to the network in both training and testing stages. Data augmentation is employed in the pipeline during the training to improve the generalization ability of the network. Although there has been a number of works [31, 69] which have already applied data normalization and data augmentation in their pipelines, these methods are particularly designed for one specific dataset and the importance of applying data augmentation for model generalization ability across datasets is less explored. Here we demon-

strate that the proposed data normalization and augmentation strategies can greatly improve the model performance in the cross-dataset setting (section 5.1.5.2). The main contributions of the work are as follows:

- To the best of our knowledge, this is the first work to explore the generalizability of CNN-based methods for cardiac MR image multi-structure segmentation, where the training data is collected from a **single scanner** but the test data comes from **multiple scanners** and **multiple sites**.

- The proposed pipeline which employs data normalization and data augmentation (section 5.1.3.4) is simple yet efficient and can be applied to training and testing of many state-of-the-art CNN architectures to improve the model segmentation accuracy across domains without necessarily sacrificing the accuracy in the original domain. Experiment results show that the proposed segmentation method is capable of segmenting multi-scanner, multi-vendor and multi-site datasets (section 5.1.5.3 and 5.1.5.4).

- Our work reveals that significant cardiac shape deformation caused by cardiac pathologies (section 5.1.5.5), low image quality (section 5.1.5.5), and inconsistent labeling protocols among different datasets (section 5.1.6) are still major challenges for generalizing deep learning-based cardiac image segmentation algorithms to images collected across different sites, which deserve further study.

5.1.2 Related work

Table 5.1: **Related work that applies CNN-based CMR image segmentation models across multiple datasets.**

Methods	Target domain ≠ Source domain	Need Finetuning	Test on	Total size of test set(s)
Tran [4]	Yes	Yes	LV/MYO/RV separately	<200
Bai *et al.* [4]	Yes	Yes	LV+MYO+RV	<100
Khened *et al.*[71]	Yes	No	MYO	<200
Our work	Yes	No	LV+MYO+RV	699

There have been a great number of works which develop sophisticated deep learning ap-

proaches to perform CMR image segmentation tasks on a specific dataset [4, 6, 31, 69]. While these models can achieve overall high accuracy over the samples from the same dataset, only a few have been validated in cross-dataset settings. Table 5.1 shows a list of related works that demonstrate the segmentation performance of their proposed method by first training a model from one set (source domain) and then testing it on other datasets (target domain). However, these approaches requires re-training or fine-tuning to improve the performance on the target domain in a fully supervised fashion. To the best of our knowledge, when we conducted this study, there were few studies reported in the literature which investigate the generalization ability of the cardiac segmentation networks that can directly work across various sites.

One work [37] in this line of research has been recently presented, which integrates training samples from multiple sites and multiple vendors [37] to improve segmentation performance across sites. Their results show that the best segmentation performance on their multi-scanner test set was achieved when the data used for training and testing are from the same scanners. Nevertheless, their solution requires collecting annotated data from multiple vendors and sites. For deployment, this may not always be practical because of the high data collection and labelling costs as well as data privacy issues.

Another direction to improve model generalization is to optimise the CNN architecture. In the work of [71], the authors proposed a novel network structure with residual connections to improve the network generalizability. They pointed out that networks with a large number of parameters may easily suffer from over-fitting problem with limited data [71]. They demonstrated that their light-weight network trained on a limited dataset outperformed the U-net [32], achieving higher accuracy on LV, myocardium, and RV. Moreover, model generalization was demonstrated by directly testing this network (without any re-training or fine-tuning) on the LV-2011 dataset [206]. As a result, this model produced comparable results to the results from a network that had been trained on the LV-2011, achieving a high mean Dice score for the myocardium (0.84). However, because of the lack of RV labels in their test set, their network's generalization ability for the RV segmentation task is unclear. In fact, segmenting the RV is considered to be harder than segmenting the LV because the RV has a more complex shape with higher variability across individuals, and its walls are thinner, making it harder to delin-

Table 5.2: **General descriptions of the three datasets used in this study.**

Name	Number of Subjects	Cohort	Sites	Scanners	Image Spatial Resolution
UKBB	4875	General population	1	1.5 T, Aera, Siemens (100%)	in-plane resolution: 1.8 mm^2 /pixel; slice thickness: 8 mm
ACDC	100	Without cardiac disease (20%); Dilated cardiomyopathy (20%); Hypertrophic cardiomyopathy (20%); Myocardial infarction with altered left ventricular ejection (20%); Abnormal right ventricle (20%)	1	1.5 T, Area, Siemens (67%) 3 T, Trio Tim, Siemens (33%)	in-plane resolution: 1.34 - 1.68 mm^2 /pixel; slice thickness: 5 -10 mm
BSCMR-AS	599	Aortic stenosis	6	1.5 T, Ingenia, Philips (5.2%); 1.5 T, Intera, Philips (17.9%); 1.5 T, Sonata, Siemens (6.2%); 1.5 T, Aera, Siemens (0.5%); 1.5 T, Avanto, Siemens (56.6%); 3 T, Achieva, Philips (0.7%); 3 T, Skyra, Siemens (3.8%); 3 T, Verio, Siemens (5.0%); 3 T, TrioTim, Siemens (4.2%);	in-plane resolution: 0.78 - 2.3 mm^2; slice thickness: 5 - 10 mm

eate from its surroundings. Because of the high shapes variability and complexity, it is more difficult to generalize a model to segment the RV across domains.

In this study, we evaluate the generalizability of the proposed method not only on the cardiac left ventricle segmentation but also on the right ventricle segmentation. Different from the works in [37, 71], the proposed method demonstrates model generalizability in a more challenging but realistic setting: our training data was collected from only one scanner (most of them are healthy subjects) while test data was collected from various unseen sites and scanners, which covers a wide range of pathologies, reflecting the spectrum of clinical practice.

5.1.3 Methodology

5.1.3.1 Data

Three datasets are used in this study and the general descriptions of them are summarised in Table 5.2.

UK Biobank dataset. The UK Biobank (UKBB) is a large-scale data set that is open to researchers worldwide who wish to conduct a prospective epidemiological study. The UKBB

study covers a large population, which consists of over half a million voluntary participants aged between 40 and 69 from across the UK. The UKBB study performs comprehensive MR imaging for nearly 100,000 participants, including brain, cardiac and whole-body MR imaging. An overview of the cohort characteristics can be found on the UK Biobank's website[1]. All CMR images we used in this study are balanced steady-state free precession (bSSFP) sequences, which were collected from one 1.5 Tesla scanner (MAGNETOM Aera, syngo MR D13A, Siemens, Erlangen, Germany). Detailed information about the imaging protocol can be found in [207]. Pixel-wise segmentations of three essential structures (LV, MYO and RV) for both end-diastolic (ED) frames and end-systolic (ES) frames are provided as ground truth [5]. Subjects in this dataset were annotated by a group of eight observers and each subject was annotated only once by one observer. After that, visual quality control was performed on a subset of data to assure acceptable inter-observer agreement.

ACDC dataset. The Automated Cardiac Diagnosis Challenge (ACDC) dataset is part of the MICCAI 2017 benchmark dataset for CMR image segmentation[2]. This dataset is composed of 100 CMR images, acquired using bSSFP imaging in breath hold with a retrospective or prospective gating [6]. The patients covered in this study have been divided into 5 groups: dilated cardiomyopathy (DCM), hypertrophic cardiomyopathy (HCM), myocardial infarction with altered left ventricular ejection fraction (MINF), abnormal right ventricle (ARV) and patients without cardiac disease (NOR). Each group has 20 patients. Detailed information about the classification rules and the characteristics of each group can be found in the benchmark study [6] as well as its website (see footnote 2). All images were collected from one hospital in France. The LV, MYO and RV in this dataset have been manually segmented for both ED frames and ES frames. Images in this dataset were labelled by two cardiologists with more than 10 years of experience [3].

BSCMR-AS dataset. The British Society of Cardiovascular Magnetic Resonance Aortic

[1] http://imaging.ukbiobank.ac.uk/
[2] https://www.creatis.insa-lyon.fr/Challenge/acdc/
[3] https://www.creatis.insa-lyon.fr/Challenge/acdc/evaluation.html

Stenosis (BSCMR-AS) dataset [208] consists of CMR images of 599 patients with severe aortic stenosis (AS), who had been listed for surgery. Images were collected from *six* hospitals across the UK with 9 types of scanners, (see Table 5.2). Specifically, these images are bSSFP sequences, which were acquired using standard imaging protocols [208]. Although the primary pathology is AS, several other pathologies coexist in these patients (e.g., coronary artery disease, amyloid) and have led to a variety of cardiac phenotypes including left ventricular hypertrophy, left ventricular dilatation and regional infarction [208]. A more detailed report on patients characteristics can be found in [208]. In this dataset, no subjects were excluded due to arrhythmi. A significant amount of diversity in image appearance and image contrast can be observed in this dataset. Different from the above two data sets, images in this dataset are partially labelled. Only the left ventricle in ED frames and ES frames, as well as the myocardium in ED frames, have been annotated manually. The contours on each slice were refined by an expert.

Ethics approval and consent to participate. The UK Biobank data has approval from the North West Research Ethics Committee (REC reference: 11/NW/0382). The ACDC data is a publicly available dataset for cardiac MR image analysis which has approval from the local ethics committee of Hospital of Dijon (France)[4]. The BSCMR-AS data has approval from the UK National Research Ethics Service (REC reference:13/NW/0832), and has been conformed to the principles of the Declaration of Helsinki. All patients gave written informed consent.

5.1.3.2 Training set and test sets

In this study, we use the UKBB dataset for training and intra-domain testing, and use the ACDC data and BSCMR-AS dataset for cross-domain testing. Following the same data splitting strategy in [4], we split the UKBB dataset into three subsets, containing 3975, 300 and 600 subjects for each set. Specifically, 3975 subjects were used to train the neural network while 300 validation subjects were used for tracking the training progress and avoid over-fitting. The subset consisting of remaining 600 subjects was used for evaluating models' performance in

[4] https://acdc.creatis.insa-lyon.fr/description/databases.html

the intra-domain setting. In addition, we directly tested this trained network on the other two unseen cross-domain datasets: ACDC and BSCMR-AS datasets *without any further re-training or fine-tuning process*. The diversity of pathology observed in the ACDC dataset and the diversity of scanners and cardiac morphologies in the BSCMR-AS set make them ideal test sets for evaluating the proposed method's segmentation performance across sites.

5.1.3.3 Network architecture

In this work, the U-net architecture [32] is adopted to perform the cardiac multi-structure segmentation task since it is the most successful and commonly used architecture for biomedical segmentation. The structure of our network is illustrated in Fig. 5.1A. The network structure is as same as the one proposed in the original paper [32], except for two main differences: (1) we apply batch normalization (BN) [147] after each hidden convolutional layer to stabilise the training; (2) we apply dropout regularization [130] after each concatenating operation to avoid over-fitting and encourage generalization.

While both 2D U-net and 3D U-net architectures can be used to solve volumetric segmentation tasks [38, 69], we opt for 2D U-net for several reasons. Firstly, performing segmentation tasks in a 2D fashion allows the network to work with images even if they have different slice thickness or have severe respiratory motion artefacts between the slices (which is not uncommon). Secondly, 3D networks require much more parameters than 2D networks. Therefore, it is more memory-consuming and time-consuming to train a 3D network than a 2D one. Thirdly, the manual annotation for images in the three datasets were done in 2D (slice-by-slice) rather than 3D. Thus, it is natural to employ a 2D network rather than a 3D network to learn segmentation from those 2D labels.

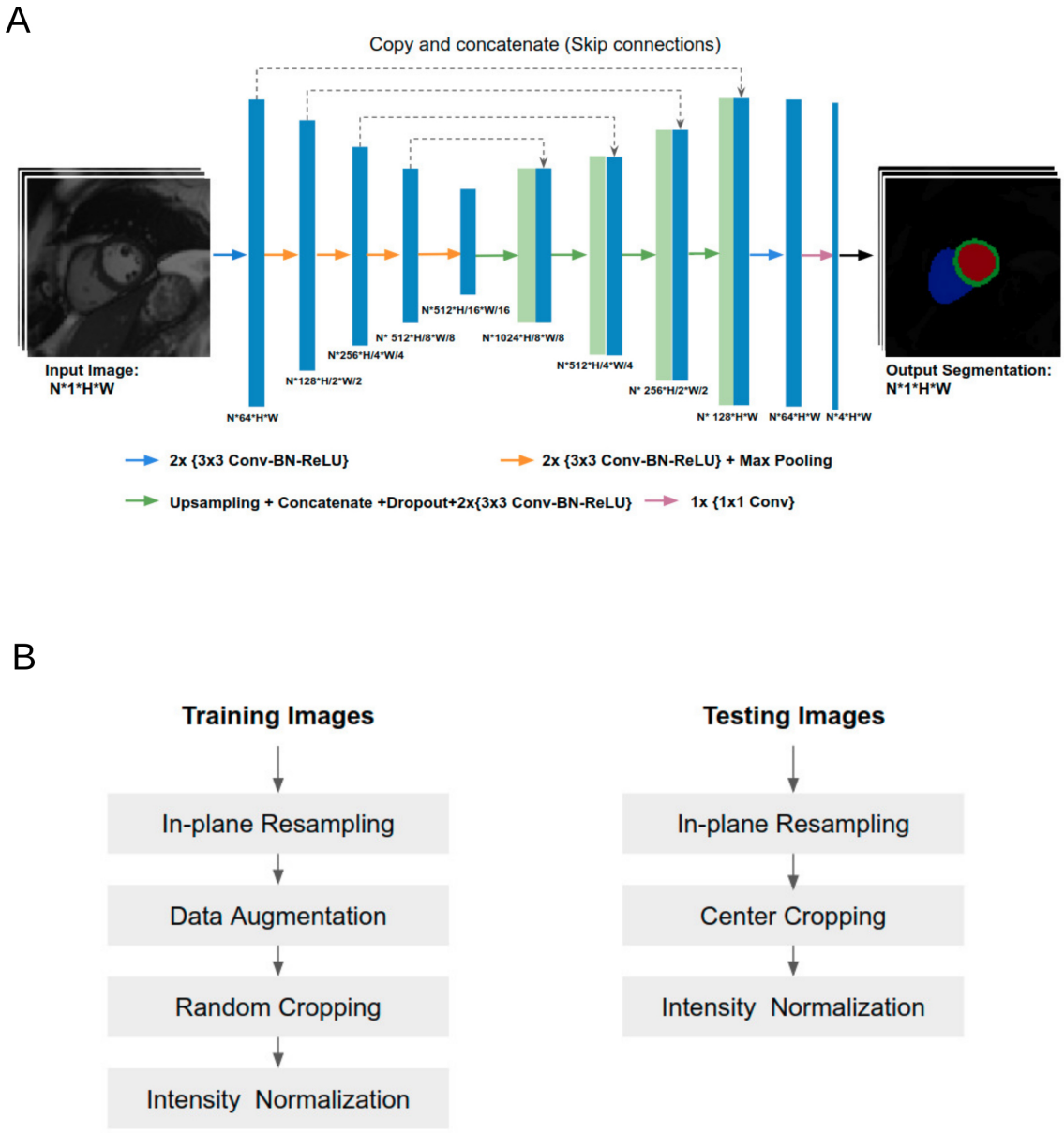


Figure 5.1: **Overview of the network structure and image pre-processing pipeline at training and testing.** (A) Overview of the network structure. Conv: Convolutional layer. BN: Batch normalization. ReLU: Rectified linear unit. The U-Net takes a batch size of N 2D CMR images as input at each iteration, learning multi-scale features through a series of convolutional layers, max-pooling operations. These features are then combined through upsampling and convolutional layers from coarse to fine scales to generate pixel-wise predictions for the 4 classes (background, LV, MYO, RV) on each slice. (B) Image pre-processing at training and testing.

5.1.3.4 Training and testing pipeline

Since training images and testing images in this study were collected from various scanners, it is vital to normalise the input images before feeding them into the network. Fig. 5.1B shows an

overview of the pipeline for image pre-processing during training and testing. Specifically, we employ image resampling and intensity normalization to normalise images in both the training and testing stages while online data augmentation is applied for improving the model generalization ability during the training process.

Image resampling. Observing that the size of the heart in images with different resolution can vary significantly, we propose to perform image resampling both in the training and testing phases before cropping. The main advantage is that after image resampling, the proportion of the heart and the background is relatively consistent, which can help to reduce the task complexity of the follow-up segmentation. However, image re-sampling is not a lossless operation, and different interpolation kernels can also affect the quality of reconstructed images [209]. In the experiments, we resampled all the images to a standard resolution of $1.25 \times 1.25\ mm^2$, which is a median value of the pixel spacings in our datasets. Following [38], images are resampled using the bilinear interpolation and the label maps are resampled using nearest-neighbor interpolation.

Here we only perform image resampling within the short-axis plane, without changing the slice thickness along the z-axis. This is consistent with the preprocessing step in other existing 2D CNN-based approaches for cardiac image segmentation [6, 38, 69]. Also, in our experiments, we found that the slice thickness does not have a significant impact on the model performance. The model performs consistently well across test images of different slice thicknesses (see Table in the appendix), while it was only trained using images of 8 mm slice thickness.

Data augmentation. Data augmentation has been widely used when training convolutional neural networks for computer vision tasks on natural images. While different tasks may have different domain-specific augmentation strategies, the common idea is to enhance model's generalization by artificially increasing the variety of training images so that the training set distribution is more close to the test set population in the real world.

In this study, the training dataset is augmented in order to cover a wide range of geometrical variations in terms of the heart pose and size. To achieve this goal, we apply:

- random horizontal and vertical flips with a probability of 0.5 to increase the variety of image orientation;
- random rotation to increase the diversity of the heart pose. The range of rotation is determined by a hyper-parameter search process. As a result, each time, the angle for augmentation is randomly selected from $[-30, +30]$;
- random image scaling with a scale factor s: $s \in [0.7, 1.4]$ to increase variations of the heart size;
- random image cropping. The random cropping crops images to acceptable sizes required by the network structure while implicitly performing random shifting to augment data context variety without black borders. Note that cropping is done after all other image augmentations. As a consequence, all images are cropped to the same size of 256×256 before being sent to the network.

We also experimented with contrast augmentation [12] (random gamma correction where the gamma value is randomly chosen from a certain range) to increase image contrast variety, but only minor improvements were found in the experiments. Therefore, it is not included in the pipeline. For each cropped image, intensity normalization with a mean of 0 and a standard deviation of 1 is performed, which is a common practice for training deep neural networks.

Training. After pre-processing, batches of images are fed to the network for training. To track the training progress, we also use a subset (validation set) from the same dataset to validate the performance of the segmentation and to identify possible over-fitting. Specifically, we apply the same data augmentation strategy on both the training and validation sets and record the average accuracy (mean intersection of union between predicted results and ground truth) on the validation set for each epoch. The model with the highest accuracy is selected as the best model. This selection criterion works as early stopping and has the benefit of allowing the network to explore if there is further opportunity to generalise better before it reaches to the final epoch.

Testing. For testing, 2D images extracted from volume data are first re-sampled and centrally cropped to the same size as the one of the training images. Again, intensity normalization is performed on each image slice which is then passed into the network for inference. After that, bilinear up-sampling or down-sampling is performed on the outputs of the network to recover the resolution back to the original one. Finally, each pixel of the original image is assigned to the class that has the highest probability among the four classes (background, LV, myocardium, RV). As a result, a final segmentation map for one input image is generated.

5.1.4 Experiments

During training, a random batch of 20 2D short-axis slices were fed into the network for each iteration after data pre-processing. The dropout rate for each dropout layer is set to be 0.2. In every iteration, cross entropy loss was calculated to optimize the network parameters through back-propagation. Specifically, the stochastic gradient descent (SGD) method was used during the optimization, with an initial learning rate of 0.001. The learning rate was decreased by a factor of 0.5 every 50 epochs. The method was implemented using Python and PyTorch. We trained the U-net for 1,000 epochs in total which took about 60 hours on one NVIDIA Tesla P40 GPU using our proposed training strategy. During testing, the computation time for segmenting one subject is less than a second.

Evaluation metrics. The performance of the proposed method was evaluated using the Dice score (3D version) which was also used in the ACDC benchmark study [6] and [4]. The Dice score evaluates the overlap between automated segmentation A and manual segmentation B, which is defined as: $\text{Dice} = \frac{2|A \cap B|}{|A|+|B|}$. The value of a Dice score ranges from 0 (no overlap between the predicted segmentation and its ground truth) to 1 (perfect match).

We also compared the volumetric measures derived from our automatic segmentation results

and those from manual ones (see section 5.1.5.6), since they are essential for cardiac function assessment. Specifically, for each manual ground truth mask and its corresponding automatic segmentation mask, we calculated the volumes of LV and RV at ED frames and ES frames, as well as the mass of myocardium estimated at ED frames. The myocardium mass around the LV is estimated by multiplying the LV myocardial volume with a density of 1.05 g/mL. After that, Bland-Altman analysis and correlation analysis for each pair were conducted. Of note, for Bland-Altman analysis, we removed the outlying mean values that fall outside the range of 1.5 × IQR (interquartile range) in order to avoid the standard deviation of mean difference being biased by extremely large values. These outliers are often associated with poor image quality. As a result, $< 3\%$ subjects were removed in each comparison.

The statistical analysis was performed using python with public packages: *pandas*[5], *scipy.stats* [6], and *statsmodel* [7].

5.1.5 Results analysis

To demonstrate the improvement of model generalization performance, we directly tested the proposed segmentation method across three sets: the UKBB test set, the ACDC set, and the BSCMR-AS set, and compared the segmentation accuracy to the performance of the segmentation method in our previous work [4]. Specifically, in [4], a fully convolutional neural network (FCN) was proposed, which was specifically designed to automatically segment a large scale of scans for the same cohort study (i.e. UKBB study) with maximum accuracy whereas the proposed method in our study focuses on improving the robustness of the neural network-based segmentation method (using the same UKBB training set as training data) for data from different domains (e.g., non-UKBB data). The comparison results are shown in Table 5.3. While both methods achieve very similar Dice scores on the intra-domain UKBB test set with high accuracy, the proposed method significantly outperforms the previous approach on the two

[5] https://pandas.pydata.org/
[6] https://docs.scipy.org/doc/scipy/reference/tutorial/stats.html
[7] https://www.statsmodels.org/stable/index.html

Table 5.3: **Comparison results of segmentation performance between a baseline method and the proposed method across three test sets.** Both methods were trained using the same UKBB training set where images were all collected from a *single* scanner. The results were evaluated on three sets from *multiple scanners* at *different sites*. Numbers listed in the table are the means and standard deviation of Dice scores.

		UKBB Test set (n=600)			**ACDC set (n=100)**			**BSCMR-AS set (n=599)**	
Method	Training set	LV	MYO	RV	LV	MYO	RV	LV	MYO*
Bai *et al.*[4]	UKBB training set	0.94 (0.04)	0.88 (0.03)	0.90 (0.05)	0.81 (0.22)	0.70 (0.20)	0.68 (0.31)	0.82 (0.21)	0.74 (0.17)
Ours	UKBB training set	0.94 (0.04)	0.88 (0.03)	0.90 (0.05)	0.90 (0.10)	0.81 (0.07)	0.82 (0.13)	0.89 (0.09)	0.83 (0.07)

*: The myocardium segmentation performance on the BSCMR-AS set was only evaluated on ED frames because of the lack of annotation at ES frames, whereas the performance on the other two datasets was evaluated on both ED and ES frames. For simplicity, Dice scores for the myocardium on the BSCMR-AS in the following tables were calculated in the same way without further illustration.

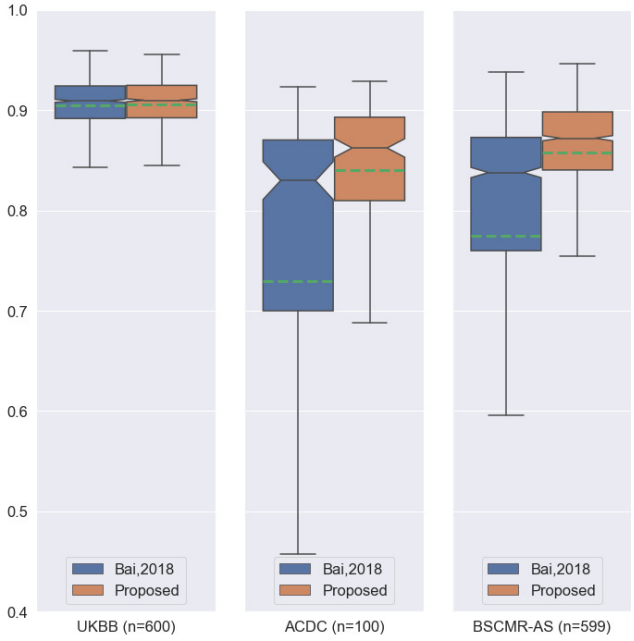


Figure 5.2: **Boxplots of the average Dice scores between the results of our previous work (Bai *et al.*, 2018 [4]) and the results of the proposed method on the three datasets.** For simplicity, we calculate the average Dice score over the three structures (LV, MYO, RV) for each image in the three datasets. The boxplots in orange are the results of the proposed method whereas the boxplots in blue are the results of the previous work. The green dashed line in each boxplot shows the mean value of the Dice scores for the segmentation results on one dataset.

cross-domain datasets: ACDC set and BSCMR-AS set. Compared to the results predicted using the method in [4] on the ACDC data, the proposed one achieves higher mean Dice scores for all of the three structures: LV (0.90 vs 0.81), myocardium (0.81 vs 0.70), and RV (0.82 vs 0.68). On the BSCMR-AS dataset, the proposed method also yields higher average Dice scores for the LV cavity (0.89 vs 0.82) and the myocardium (0.83 vs 0.74). Fig. 5.2 compares the distributions of Dice scores for the results obtained by the proposed method and the previous

work. From the results, the boxplots of the proposed method are shorter than those of the previous method and have higher mean values, which suggests that the proposed method achieves comparatively higher overall segmentation accuracy with lower variance on the three datasets.

In order to identify what contributes to the improved performance, we further compare the proposed method with [4] in terms of methodology. Two main differences are spotted:

- **Network structure and capacity**. Compared to the U-net we used in this study, FCN in [4] has a smaller number of filters at each level. For example, the number of convolutional kernels (filters) in the first layer of FCN is 16 whereas the one in the U-net is 64. In addition, in the decoder part, FCN directly upsamples the feature map from each scale to the finest resolution and concatenates all of them, whereas the U-net adopts a hierarchical structure for feature aggregation.
- **Training strategy in terms of data normalization and data augmentation**. Compared to the image pre-processing pipeline in the previous work, the proposed pipeline adopts image resampling and random image flip augmentation in addition to the general data augmentation based on affine transformations.

In order to study the influence of the network structure as well as the data normalization and augmentation settings on model generalizability, extensive experiments were carried out and the results are shown in the next two sections.

5.1.5.1 The influence of network structure and capacity

To investigate the influence of network structure on model generalization, we trained three additional networks:

- FCN-16: the FCN network presented in [4] which has 16 filters in the first convolutional layer.
- FCN-64: a wider version of FCN where the number of filters in each convolutional layer is increased by 4 times.

- UNet-16: a smaller version of U-net where the number of filters in each convolutional layer is reduced by four times. Same as FCN-16, it has 16 filters in the first layer.

All of them were trained using the same UKBB training set and with the same training hyper-parameters. These networks were then compared to the proposed network (UNet-64). Table 5.4

Table 5.4: **Cross-dataset segmentation performances of four different network architectures.** All the networks were trained using the same UKBB training set with the proposed data normalization and augmentation strategy for *1,000* epochs. Results listed in the table are the means and standard deviation of the Dice scores evaluated on the three sets. Numbers in red denote mean Dice scores below 0.70, whereas numbers in the **bold** font style denote the highest mean Dice scores among the results of the four networks.

Network Structure	num of conv weights (aprox.)	UKBB Test set (n=600)			ACDC set (n=100)			BSCMR-AS set (n=599)	
		LV	MYO	RV	LV	MYO	RV	LV	MYO
FCN-16	0.98 million	0.92 (0.04)	0.84(0.04)	0.88(0.05)	0.80(0.20)	0.67(0.19)	0.68(0.27)	0.84(0.14)	0.77(0.11)
FCN-64	15.6 million	0.94 (0.04)	0.87(0.03)	0.89(0.05)	0.87(0.12)	0.78(0.11)	0.77(0.17)	0.85(0.12)	0.79(0.10)
UNet-16	0.84 million	0.92 (0.04)	0.83(0.04)	0.87(0.05)	0.87(0.12)	0.66(0.14)	0.67(0.22)	0.85(0.11)	0.73(0.11)
Ours (UNet-64)	13.4 million	**0.94 (0.04)**	**0.88(0.03)**	**0.90(0.05)**	**0.90(0.10)**	**0.81(0.07)**	**0.82(0.13)**	**0.88(0.09)**	**0.83(0.07)**

compares the performances of the four different networks over the three different test sets. It can be seen that while there is no significant performance difference among the four networks on the UKBB test set, small networks: UNet-16 and FCN-16 perform much more poorly than their wider versions: UNet-64 and FCN-64, on the ACDC set (see red numbers in Table 5.4). This may indicate that in order to accommodate more variety of data augmentation for generalization, the network requires a larger capacity. It is also worth noticing that UNet-64 outperforms FCN-64 on all of the three test sets, while UNet-64 contains fewer parameters than FCN-64. This improvement may result from U-net's special architecture: skip connections with its step-by-step feature upsampling and aggregation. The results indicate that the network structure and capacity can affect the segmentation model generalizability across datasets.

5.1.5.2 The influence of different data normalization and data augmentation techniques

In this section, we investigate the influence of different data normalization and augmentation techniques on the generalizability of the network, including image resampling (data normaliz-

Table 5.5: **Cross-dataset segmentation performances of U-Nets with different training configurations.** All experiments were performed with the standard U-Net architecture: UNet-64. Each U-Net was trained using the same UKBB training set for *200* epochs to save computation. Statistics listed in the table are the means and standard deviation of the Dice scores evaluated on the three sets. Numbers in red are those mean Dice scores below 0.70.

Configurations				UKBB Test set (n=600)			ACDC set (n=100)			BSCMR-AS set (n=599)	
Image Resample	Rotation Aug	Flip Aug	Scale Aug	LV	MYO	RV	LV	MYO	RV	LV	MYO
✓	✓	✓	✓	0.923 (0.041)	0.847 (0.038)	0.878 (0.048)	0.873 (0.101)	0.744 (0.104)	0.750 (0.187)	0.851 (0.113)	0.783 (0.095)
	✓	✓	✓	0.916 (0.046)	0.836 (0.041)	0.864 (0.053)	0.811 (0.179)	0.614 (0.186)	0.575 (0.270)	0.798 (0.172)	0.673 (0.162)
✓		✓	✓	0.922 (0.042)	0.848 (0.038)	0.878 (0.050)	0.869 (0.117)	0.733 (0.117)	0.722 (0.210)	0.853 (0.118)	0.784 (0.093)
✓	✓		✓	0.924 (0.041)	0.849 (0.037)	0.881 (0.049)	0.858 (0.115)	0.705 (0.142)	0.681 (0.266)	0.862 (0.110)	0.779 (0.092)
✓	✓	✓		0.921 (0.047)	0.845 (0.039)	0.876 (0.050)	0.785 (0.188)	0.640 (0.187)	0.596 (0.279)	0.834 (0.148)	0.752 (0.125)

ation), scale, flip and rotation augmentation (data augmentation). We focus on these four operations because convolutional neural networks are designed to be translation-equivariant [210] but they are not rotation-equivariant, nor scale and flip-equivariant [211, 212]. This means that if we rotate the input, the networks cannot be guaranteed to produce the same predictions with the corresponding rotation, indicating that they are not robust to geometrical transformations on images. Current methods to improve these networks' ability to deal with rotation/flip/scale variations still heavily rely on data augmentation while intensity-level difference might be addressed by further doing domain adaptation techniques such as style transfer or adaptive batch normalization [213].

To investigate the influence of these four operations on model generalization, we trained additional three U-nets using the UKBB training set, each of them was trained with the same settings except that only one operation was removed. To save the computational time for this ablation study, each network was trained for 200 epochs, which still took 10 hours for each network since the training set from the UKBB dataset was considerably large (3,975 subjects). The test results on the UKBB test set, the ACDC dataset, and the BSCMR-AS dataset are shown in Table 5.5. It can be observed that while the results on the test data from the same domain (UKBB) with different settings do not vary much, there are significant differences on the other two test sets, demonstrating the importance of the four data augmentation operations. For example, image resampling increases the averaged Dice score from 0.673 to 0.783 for the RV segmentation on the BSCMR-AS set, whereas augmentation by scaling improves the mean Dice score from 0.596 to 0.750 for the RV on the ACDC set. The best segmentation performance

over the three sets is achieved by combining all the four operations.

These results suggest that increasing variations regarding pixel spacing (image scale augmentation), image orientation (flip augmentation), heart pose (rotation augmentation) as well as data normalization (image resampling) can be beneficial to improve model generalisabilty over unseen cardiac datasets. While one may argue that there is no need to do image resampling if scale augmentation is performed properly during training, we found that image resampling can significantly reduce the complexity of real-world data introduced by heterogeneous image pixel spacings, such that training and testing data are more similar to each other, bringing benefits to both model learning and prediction. In the following sections, for the sake of simplicity, we will use 'UKBB model' to refer to our best model (the U-net which was trained using the UKBB training set with our proposed training strategy).

5.1.5.3 Segmentation performance on images from different types of scanners

Table 5.6: **Segmentation performance of the UKBB model across different scanners.** Tests were performed on the BSCMR-AS dataset and ACDC dataset. This table presents the mean and standard deviation (numbers in the brackets) of the Dice score.

Dataset	MRI Scanner Attributes	Scanners	# of subjects	LV	MYO	RV
BSCMR-AS	Manufactures	Philips	142	0.89 (0.07)	0.85 (0.04)	-
		Siemens	457	0.88 (0.10)	0.83 (0.08)	-
	Magnetic Field Strengths	1.5T	517	0.88 (0.09)	0.83 (0.09)	-
		3 T	82	0.88 (0.09)	0.84 (0.09)	-
ACDC	Magnetic Field Strengths	1.5T	65	0.89 (0.09)	0.81 (0.06)	0.80 (0.09)
		3 T	29	0.91 (0.06)	0.82 (0.05)	0.80 (0.08)

In this section, UKBB model's segmentation performance is analysed according to different manufacturers (Philips and Siemens) and different magnetic field strengths (1.5 Telsa and 3 Telsa). The results on the two datasets (BSCMR-AS and ACDC) are listed in Table 5.6. For ACDC data, only the results regarding scans imaged using different magnetic strengths are reported since these scans are all from Siemens. Furthermore, results in the ACDC dataset with Dice scores below 0.50 are not taken into account for this evaluation. This is because the number of subjects from a 3T scanner in the ACDC is so small (33 subjects) that the averaged performance can be easily affected given only a few cases with extreme low Dice

scores. Here, six subjects were excluded. The final results show that the model trained only using 1.5T Siemens data (UKBB data) could still produce similar segmentation performance on other Siemens and Philips data (top two rows in Table 5.6). Similar results are found on those images acquired from 1.5T scanners and those acquired from 3T scanners (see the bottom four rows in Table 5.6). This indicates that the proposed method has the potential to train a model capable of segmenting images across **various scanners** even if the training images are only from **one** scanner.

5.1.5.4 Segmentation performance on images from different sites

Table 5.7: **Segmentation performance of the UKBB model across different sites.** This table presents the mean and the standard deviation (numbers in the brackets) of Dice scores for each site.

Dataset	Site	# of subjects	LV	MYO	RV
ACDC	site A	100	0.91 (0.07)	0.81 (0.08)	0.82 (0.11)
BSCMR-AS	site B	28	0.88 (0.09)	0.83 (0.04)	-
	site C	74	0.88 (0.09)	0.83 (0.04)	-
	site D	150	0.89 (0.07)	0.85 (0.04)	-
	site E	122	0.86 (0.11)	0.81 (0.08)	-
	site F	64	0.88 (0.09)	0.84 (0.08)	-
	site G	160	0.89 (0.09)	0.85 (0.08)	-

We also evaluate the performance of the UKBB model across seven sites: one from ACDC data, six sites from BSCMR-AS data. Results are shown in Table 5.7. From the results, no significant difference is found when evaluating the LV and the myocardium segmentation performances among the seven sites (A-G) while the generalization performance for RV segmentation still needs further investigation when more data with annotated RV becomes available for evaluation.

5.1.5.5 Segmentation performance on images belonging to different pathologies

We further report the segmentation performance of the proposed method on five groups of pathological data and the group of normal subjects (NOR), see Table 5.8. Surprisingly, the UKBB model achieves satisfying segmentation accuracy over the healthy group as well as DCM

Table 5.8: **Segmentation performance of the UKBB model across the five groups of pathological cases and normal cases (NOR).** This table presents the mean and standard deviation of the Dice score. Red numbers are those mean Dice scores below 0.80.

Dataset	Group	# of subjects	LV	MYO	RV
ACDC	NOR	20	0.91 (0.05)	0.83 (0.04)	0.85 (0.14)
	DCM	20	0.94 (0.04)	0.81 (0.05)	0.82 (0.11)
	HCM	20	0.84 (0.12)	0.84 (0.03)	0.84 (0.08)
	MINF	20	0.92 (0.05)	0.81 (0.04)	0.78 (0.13)
	ARV	20	0.86 (0.13)	0.74 (0.11)	0.79 (0.16)
BSCMR-AS	AS	599	0.88 (0.09)	0.83 (0.07)	-

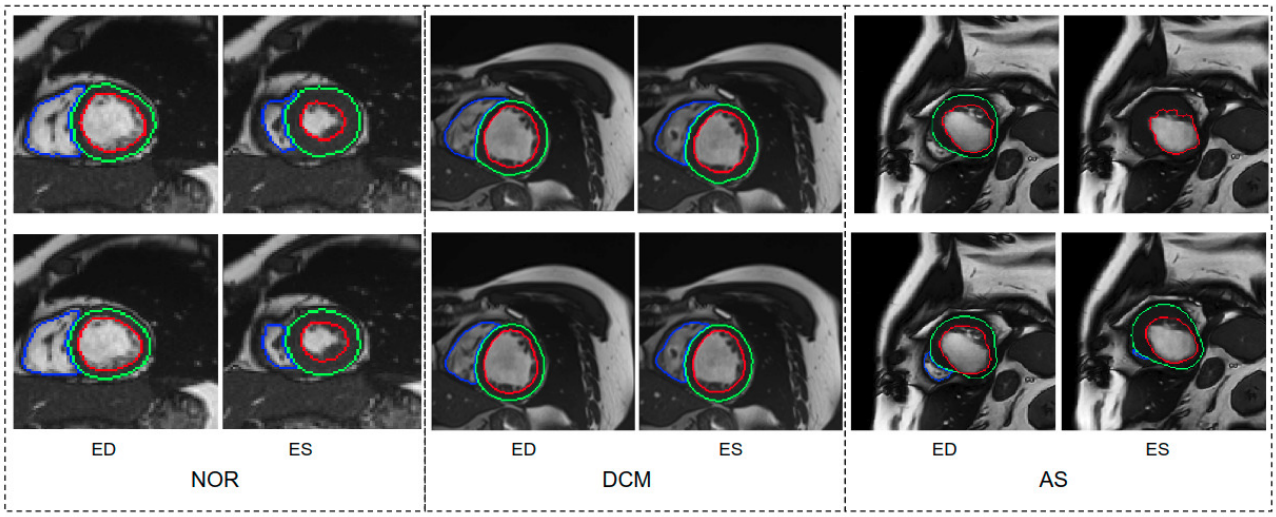


Figure 5.3: **Visualization of good segmentation examples selected from 3 patient groups.** Three groups are NOR (without cardiac disease), DCM (dilated cardiomyopathy), AS (aortic stenosis). Row 1: Ground truth (manual annotations); row 2: predicted results by the UKBB model. Each block contains a slice from ED frame and its corresponding ES one for the same subject. This figure shows that the UKBB model produced satisfying segmentation results not only on healthy subjects but also on those DCM and AS cases with abnormal cardiac morphology. The AS example in this figure is a patient with aortic stenosis who previously had a myocardial infarction. Note that this AS case is from BSCMR-AS dataset where the MYO and RV on ES frames were not annotated by experts.

images and those images diagnosed with AS, indicating the model is capable of segmenting not only those with normal cardiac structures but also some abnormal cases with the cardiac morphological variations in those HCM images and AS images, see Fig. 5.3. However, the model fails to segment some of the other pathological images, especially those in the HCM, MINF, and ARV pathology groups where lower Dice scores are observed. For example, the mean Dice score for LV segmentation on HCM images is the lowest (0.84). Fig. 5.4 demonstrates some of the worst cases produced by the proposed method. The first column in Fig. 5.4, shows a failure case where the UKBB model underestimated the myocardium and overestimated the LV

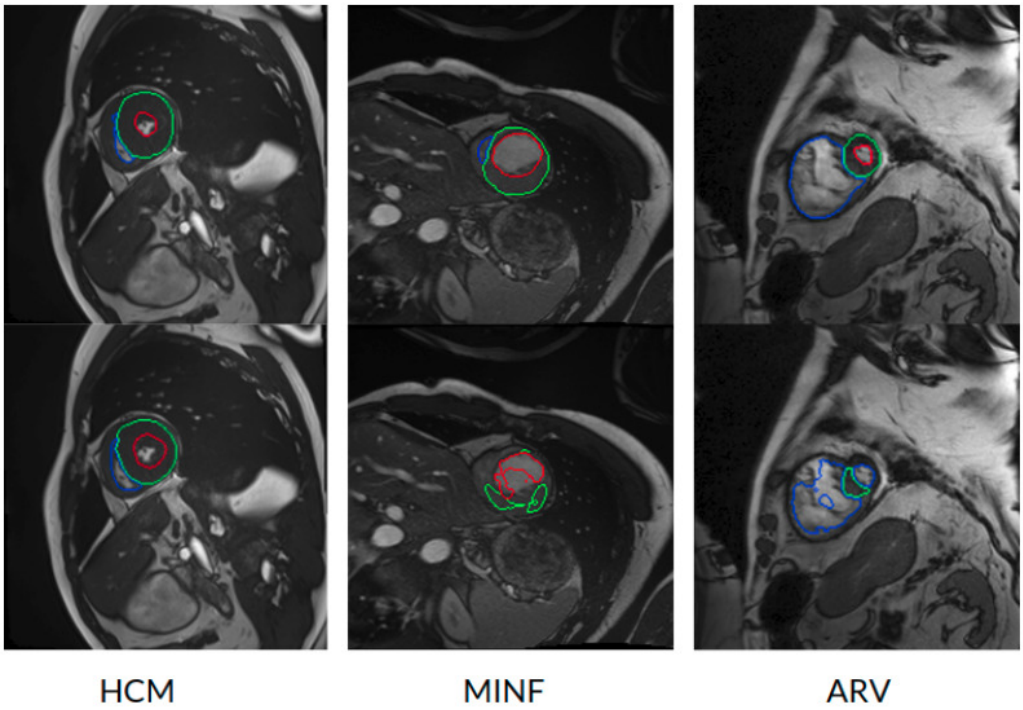


Figure 5.4: **Examples of the worst cases that have pathological deformations.** Row 1: Ground truth; row 2: predicted results by the UKBB model. HCM: hypertrophic cardiomyopathy; MINF: myocardial infarction with altered left ventricular ejection fraction; ARV: abnormal right ventricle. Column 1 shows that the UKBB model underestimates the myocardium in patients with HCM. Column 2 shows that the model struggles to predict the cardiac structure when certain sections of the myocardium are extremely thin. Column 3 shows a failure case where an extremely large right ventricle is shown in the image. All these images are from the ACDC dataset.

when a thickened myocardial wall is present in a patient with HCM. Also, the model struggles to segment cardiac structure on a patient with MINF which contains the abnormal myocardial wall with non-uniform thickness (the second column in Fig. 5.4). Compared to images in the other four groups with pathology, images from patients with ARV seem to be more difficult for the model to segment as the model not only achieves a low mean Dice score on the RV (0.79) but also a low averaged value on the myocardium (0.74).

One possible reason for these unsatisfactory segmentation results might be the lack of pathological data in the current training set. In fact, the UKBB data only consists of a small amount of subjects with self-reported cardiovascular diseases, and the majority of the data are healthy subjects in middle and later life [4, 5, 214]. This indicates that the network may not be able to 'learn' the range of those pathologies that are seen in everyday clinical practice, especially those abnormalities which are not currently represented in the UKBB dataset.

Failure mode analysis. We also visually inspected the images where the UKBB model produces poor segmentation masks. In general, there are two main failure modes we identified, apart from the failure found on the abnormal pathological cases which we have discussed above:

- **Apical and basal slices**. These slices are more error-prone than mid-ventricle slices, which has also been reported in [6]. Segmenting these slices is difficult because apical slices have extremely tiny objects which can be hard to locate and segment (see Fig. 5.5A) whereas basal slices with complex structures increase the difficulty of identifying the contour of the LV (see Fig. 5.5B);

- **Low image quality**. Images with poor quality are found both in 1.5T and 3T images (see Fig. 5.5C and 5.5D). As reported in [197, 198], 1.5T images are more likely to have low image contrast than 3T images due to the low signal-to-noise (SNR) limits, whereas 3T images can have more severe imaging artefact issues than 1.5T images. These artefacts and noise can greatly affect the segmentation performance.

5.1.5.6 Statistical analysis on clinical parameters

We further compare the proposed automatic method with manual approach on five clinical parameters, including the end-diastolic volume of LV (LV_{EDV}), the end-systolic volume of LV (LV_{ESV}), the left ventricular mass (LVM), the end-diastolic volume of right ventricle (RV_{EDV}), and the end-systolic volume of RV (RV_{ESV}).

Figure 5.6 shows the Bland-Altman plots for the five clinical parameters on the three datasets. The Bland-Altman plot is commonly used for analysing agreement and bias between two measurements. Here, each column shows the comparison results between automated measurements and manual measurements for one particular parameter, including the mean differences (MD) with corresponding standard deviation (SD) and the limits of agreement (LOA). In addition, we also conducted the Bland-Altman analysis for the automatic method (FCN) in our previous work [4], for comparison.

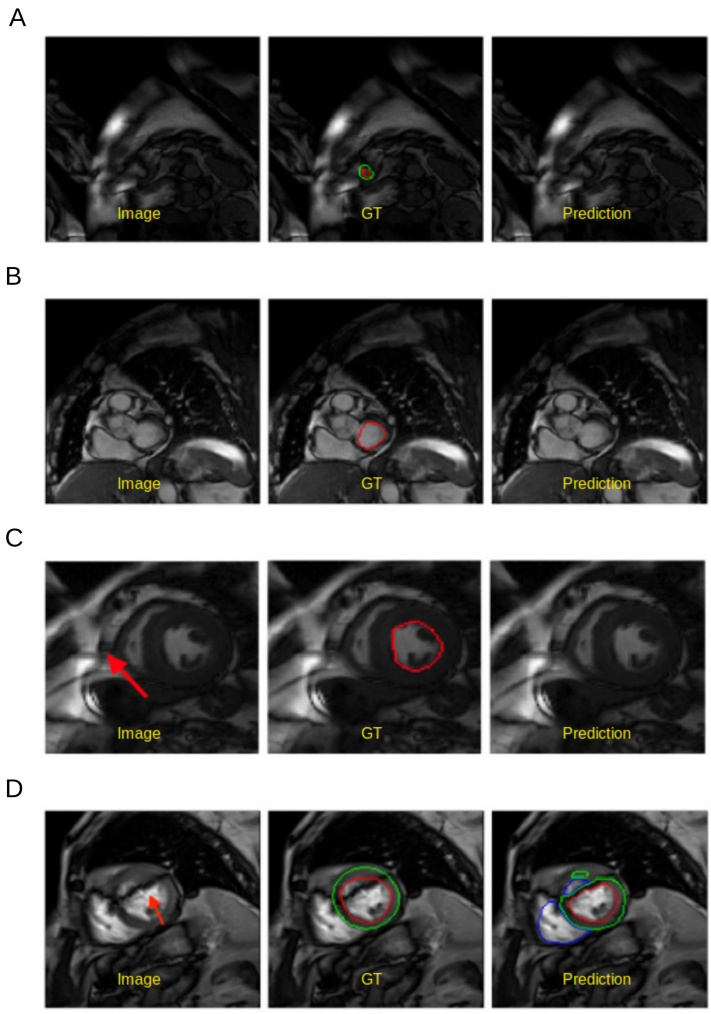


Figure 5.5: **Examples of worst segmentation results found on challenging slices.** Left: Image, middle: ground truth (GT), right: prediction from the UKBB model. (A) Failure to predict LV when the apical slice has a very small LV. (B) LV segmentation missing on the basal slice (ES frame). This sample is from the BSCMR-AS dataset where only the LV endocardial annotation is available. (C) Failure to recognize the LV due to a stripe of high-intensity noise around the cardiac chambers in this 1.5T image. This sample is an ES frame image from the BSCMR-AS dataset. (D) Failure to estimate the LV structure when unexpected strong dark artifacts disrupt the shape of the LV in this 3T image. Note that this image is an ED frame image from the BSCMR-AS dataset where RV was not annotated by experts.

From the first two columns in the Fig. 5.6, one can see that both FCN and the proposed method achieve excellent agreements with human observers on the UKBB dataset, indicating both of them can be used interchangeably with manual measurements. For the other two datasets, by contrast, the proposed method achieves much better agreement than FCN, as the LOA between the proposed method and manual results is narrower. For example, for LVM on the ACDC dataset, the LOA between the proposed method and the manual approach is from

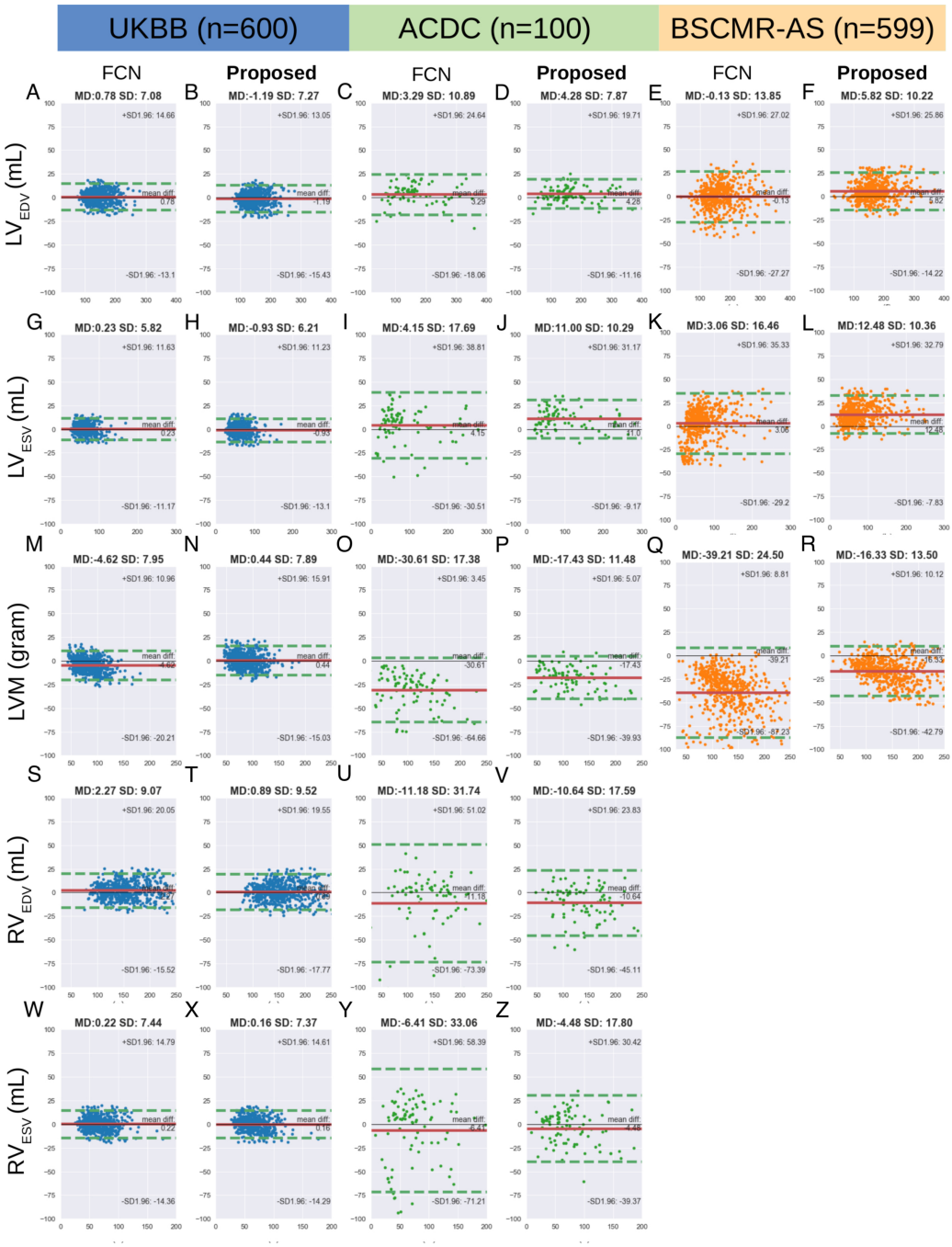


Figure 5.6: **Agreement of clinical measurement from automatic and manual segmentation.** Figures A-Z are Bland Altman plots (automatic - manual) on the three sets. In each Bland-Altman plot, the x-axis denotes the average of two measurements, whereas the y-axis denotes the difference between them. The solid line in red denotes the mean difference (bias) and the two dashed lines in green denote ±1.96 standard deviations from the mean. The title of each plot shows the mean difference (MD) and its standard deviation (SD) for each pair of measurements. FCN: the automatic method in our previous work [4], LV/RV: left/right ventricle, EDV/ESV: end-diastolic/systolic volume, LVM: left ventricular mass. Best viewed in color and zoom in.

5.07 to -39.93 (MD =-17.43) while the LOA between the FCN and the manual method is from 3.45 to -64.66 (MD = -30.61), see Fig. 5.6O and Fig. 5.6P, respectively.

Table 5.9: **Spearman's rank correlation coefficients of clinical parameters derived from the automatic measurements and the manual measurements on the three sets.** All segmentations are produced by the U-Net trained with the UKBB training set.

Comparison	Test set	$\mathbf{LV_{EDV}}$	$\mathbf{LV_{ESV}}$	LVM	$\mathbf{RV_{EDV}}$	$\mathbf{RV_{ESV}}$
Automatic vs Manual	UKBB (n=600)	0.97	0.91	0.93	0.96	0.91
Automatic vs Manual	ACDC (n=100)	0.97	0.94	0.96	0.79	0.83
Automatic vs Manual	BSCMR-AS (n=599)	0.94	0.92	0.92	-	-

Note: Each coefficient reported in this table has a P-value below 0.0001.

Finally, we calculate the Spearman's rank correlation coefficients (r^2) of the five clinical parameters derived from the automatic segmentation using the proposed method and the manual segmentation, which are reported in Table 5.9. From the results, it can be observed that the clinical measurements based on the LV segmentation and the myocardium segmentation derived by our automatic model are highly positively correlated with the manual analysis (≥ 0.91), although the RV correlation coefficients on the ACDC dataset are relatively lower.

5.1.6 Discussion

In this work, we developed a general training/testing pipeline based on data normalization and augmentation for improving the generalizability of neural network-based CMR image segmentation methods. We also highlighted the importance of the network structure and capacity (section 5.1.5.1) as well as the data normalisation and augmentation strategies (section 5.1.5.2) for model generalizability. Extensive experiments on multiple test sets were conducted to validate the effectiveness of the proposed method. The proposed method achieves promising results on a large number of test images from various scanners and sites even though the training set is from one scanner, one site (section 5.1.5.3, 5.1.5.4). Besides, the network is capable of segmenting healthy subjects as well as a group of pathological cases from multiple sources although it had only been trained with a *small* portion of pathological cases.

The limitation of the current method (the UKBB model) is that it still tends to underestimate the myocardium especially when the size of the myocardium becomes larger (see points in the right part of Fig. 5.6R. Again, we conclude this limitation is mainly due to the lack of pathological cases in the training set.

Besides, we found that the difference (bias) between the automatic measurements and the manual measurements in the cross-domain test sets: ACDC and BSCMR-AS, are more significant than the difference in the intra-domain set: UKBB test set. The larger bias may be caused by not only those challenging pathological cases we have discussed above, but also inter-observer bias and the inconsistent labelling protocols used in the three datasets. The evident inter-observer variability when delineating myocardial boundaries on apical and basal slices in a single dataset has been reported in [215]. In this study, however, there are three datasets which were labelled by three *different* groups of observers. Each group followed an independent labelling protocol. As a result, significant variations of RV labels and MYO labels on the basal planes among the three datasets are found. This inter-dataset inconsistency of the RV labels on basal planes has been reported in [84]. The mismatch of RV labels can partially account for the negative MD values for the RV measurements in the ACDC dataset (see Fig. 5.6 V). The differences in the labelling protocols together with inter-observer variability in different datasets pose challenges to evaluate the model generalizability across domains accurately.

In the future, we will focus on improving the segmentation performance of the neural network by increasing the diversity of the training data in terms of pathology. A promising way of doing it, instead of collecting more labelled data, is to synthesize pathological cases by transforming existing healthy subjects with pathological deformations. A pioneering work [216] in this direction has successfully transported pathological deformations from certain pathological subjects (i.e. HCM, DCM) to healthy subjects, which can help to increase the number of pathological cases. Similarly, one can also adopt other types of learning-based data augmentation approaches (e.g., generative adversarial network based data augmentation [164], adversarial data augmentation [217]) to improve the model robustness on challenging cases, generating more realistic and challenging images (e.g., apical/basal slices, images with different types of artefacts) for the network to learn. Another direction, is to add a post-processing module to

correct those failed predictions with anatomical constraints [46, 218]. Both of these approaches can be easily integrated in the proposed training pipeline without significant modifications. Last but not least, for clinical deployment, it is necessary to alert users when failure happens. In this regard, future work can be integrating the segmentation approach with an automatic quality control module, providing automatic segmentation assessment (e.g., estimated segmentation scores [219], model uncertainty maps [220]) to clinicians for further verification and refinement.

5.1.7 Conclusion

In this work, we proposed a general training/testing pipeline for neural network-based cardiac segmentation methods and revealed that a proper design of data normalization and augmentation, as well as network structure, play essential roles in improving its generalization ability across images from various domains. We have shown that a neural network (U-net) trained with CMR images from a **single** scanner has the potential to produce competitive segmentation results on **multi-scanner** data across domains. Besides, experimental results have shown that the network is capable of segmenting healthy subjects as well as a group of pathological cases from multiple sources, although it had only been trained with the UK Biobank data, which has only a *small* portion of pathological cases. Although it might still have limitations in segmenting images with low quality and some images with significant pathological deformations, higher segmentation accuracy for these subjects could be further achieved by increasing the diversity of training data regarding image quality and pathology in the future. Also, for simplicity, our current data normalization step consists of a image in-plane re-sampling and a standard intensity normalization step to harmonize images from unseen sites. More advanced techniques on data harmonization could be considered to improve the model performance on unseen test images across different scanners and sites. These include image correction methods for reduced imaging artifacts [167], and other learning-based image normalization methods to unify the spatial resolutions [221, 222], and to adjust intensity distributions for unified image appearance with improved image quality [223–225]. One should note that these approaches

generally require an iterative process at test time [167] or training a model (e.g., an intensity normalization network) to optimize its parameters before deployment [221, 222, 224].

5.2 Cooperative training and latent space data augmentation for robust segmentation

5.2.1 Introduction

Segmenting anatomical structures from medical images is an important step for diagnosis, treatment planning and clinical research. In recent years, deep convolutional neural networks (CNNs) have been widely adopted to automate the segmentation procedure [3, 226]. However, a major obstacle for deploying deep learning-based methods to real-world applications is domain shift during clinical deployment, which includes changes of image appearance and contrasts across medical centers and scanners as well as various imaging artifacts. Recent works on domain generalization provide a promising direction to address this issue [227–231]. A majority of them require training data from *multiple* domains to learn domain-invariant features for segmentation. Multi-domain datasets, however, may not always be feasible due to data privacy concerns and collection costs. Learning robust networks from single-domain data and limited data is of great practical value for medical imaging research.

In this work, we propose a novel cooperative training framework for learning a robust segmentation network from *single-domain* data. We make the following contributions. (1) First, to improve model performance on unseen domains, we design a cooperative training framework where two networks collaborate in both training and testing. This is inspired by the two-system model in human behavior sciences [232], where a fast-thinking system makes intuitive judgment and a slow-thinking system corrects it with logical inference. Such a collaboration is essential for humans to deal with unfamiliar situations. In our framework, a fast-thinking network (FTN) aims to understand the context of images and extracts task-related image and shape features for an initial segmentation. Subsequently, a slow-thinking network (STN) refines the initial segmentation according to a learned shape prior. (2) We introduce a latent space data augmentation method, which performs channel-wise and spatial-wise masking for the latent code learned from FTN in random and targeted fashions. Reconstructing images with masked latent codes generates a diverse set of challenging images and corrupted segmentation maps to

reinforce the training of both networks. Experimental results on cardiac imaging datasets show that the cooperative training mechanism with generated challenging examples can effectively enhance FTN's segmentation capacity and STN's shape correction ability, leading to more robust segmentation. (3) The proposed method alleviates the need for multi-domain data, making it applicable to a wide range of applications.

5.2.2 Related work

Our work is conceptually related to data augmentation, multi-task learning (MTL) and multi-stage learning. *a) data augmentation* applies transformations or perturbations to improve the diversity of training data, which is effective for improving model generalization [133]. A large number of the works focuses on image-space data augmentation, including both intensity and geometric transformation functions [7, 233] and patch-wise perturbations [234–237]. Adversarial data augmentation has also been explored, which takes the segmentation network into account and generates adversarial examples that can fool the network [15, 174, 238, 239]. A major novelty of our work is that we perform data augmentation in the latent space. The latent space contains abstract representation of both image and shape features and challenging examples can be generated by manipulating this space. Different from existing latent DA methods used in metric learning [240], our method is based on feature masking rather than feature interpolation (i.e. linear combination) and thus does not require paired images from the same/different categories to generate synthetic data. To the best of our knowledge, our work is the first to explore latent space DA for robust segmentation with single domain data. *b) MTL* is extremely beneficial when training data is limited [236, 237, 241]. MTL enhances network capacity by encouraging the learning of common semantic features across various tasks. c) Our work is also related to multi-stage learning, which consists of two stages of segmentation: a first network for coarse segmentation from images and a second network for refinement [46, 242]. For example, in [242], manually designed functions are used to generate poor segmentation and a denoising autoencoder is independently trained for segmentation refinement. Another novelty of our work is that we seek the mutual benefits of a segmentation network and a denoising network

by training them cooperatively, using hard examples constructed from latent space.

5.2.3 Methodology

Given a training dataset from *one, single* domain $\mathcal{D}_{tr} = \{(\boldsymbol{x}_i, \mathbf{y}_i)\}_{i=1}^{n}$, with pairs of images $\boldsymbol{x}_i \in \mathbb{R}^{H \times W}$ and one-hot encoded C-class label maps $\mathbf{y}_i \in \{0,1\}^{H \times W \times C}$ as ground truth (GT), our goal is to learn a robust segmentation network across various 'unseen' domains with different image appearance and/or quality. Here, H, W denote image height and width, respectively.

5.2.3.1 Overview of the framework

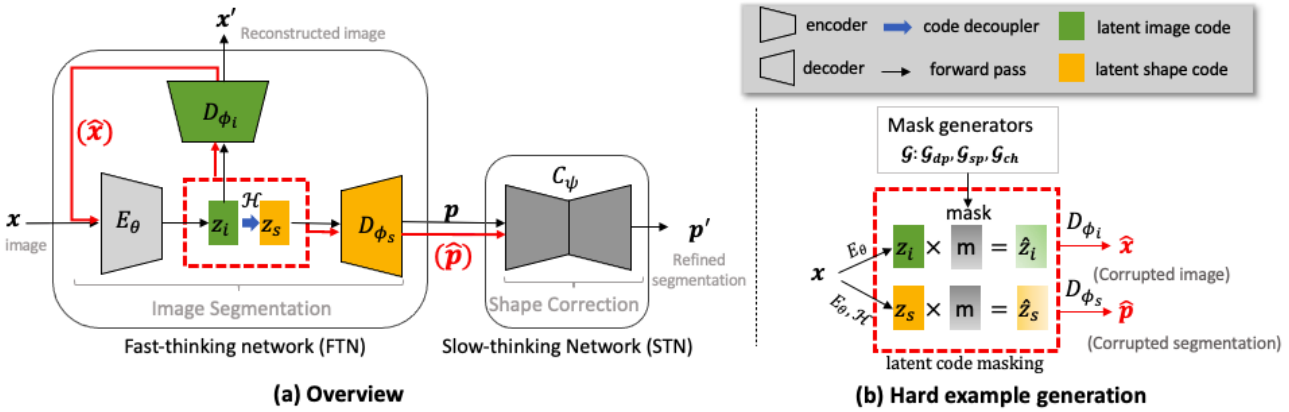


Figure 5.7: **Visual demonstration of the proposed cooperative training framework and latent space data augmentation.** (a) The proposed cooperative training framework, which consists of a fast-thinking network (FTN) and a slow-thinking network (STN). (b) Hard example generation in the latent space. Latent code masking is performed for generating both corrupted images and segmentations for cooperative training.

An overview of the proposed framework is illustrated in Fig. 5.7 (a). At a high level, our framework consists of a fast-thinking network (FTN) and a slow-thinking network (STN). Given an image $\boldsymbol{x}$, the FTN extracts task-specific shape features $\mathbf{z}_s$ to perform the segmentation task and image contextual features $\mathbf{z}_i$ to perform the image reconstruction task. This network consists of a shared encoder E_θ, a feature decoupler $\mathcal{H}$ and two task-specific decoders $\mathcal{D}_{\phi_s}$ and $\mathcal{D}_{\phi_i}$ for image segmentation and reconstruction tasks. We apply the latent code decoupler $\mathcal{H}$ to $\mathbf{z}_i$, so that task-unrelated information (e.g., image texture information, brightness) is deactivated in $\mathbf{z}_s$. This encourages a sparse latent code $\mathbf{z}_s$, which is beneficial for model robustness [243]. $\mathcal{H}$

employs a stack of two convolutional layers followed by a ReLU activation function. STN is a denoising autoencoder network $\mathcal{C}_\psi$, which corrects the segmentation predicted by FTN by using a learned shape prior encoded in $\mathcal{C}_\psi$. At inference time, we first employ FTN to perform fast segmentation for a given image $\boldsymbol{x}$: $\mathbf{p} = \mathcal{D}_{\phi_s}(\mathcal{H}(E_\theta(\boldsymbol{x})))$, and then STN to refine the prediction for improved segmentation quality: $\mathbf{p}' = \mathcal{C}_\psi(\mathbf{p})$.

5.2.3.2 Standard training

To train the two networks, we propose a standard approach which jointly trains the three encoder-decoder pairs with a supervised multi-task loss function for image reconstruction $\mathcal{L}_{rec}$, image segmentation $\mathcal{L}_{seg}$ and shape correction $\mathcal{L}_{shp}$. The loss is defined as:

$$\mathcal{L}_{std} = \mathbb{E}_{(\boldsymbol{x},\mathbf{y})\in D_{tr}}[\mathcal{L}_{rec}(\boldsymbol{x}', \boldsymbol{x}) + \mathcal{L}_{seg}(\mathbf{p}, \mathbf{y}) + \mathcal{L}_{shp}(\mathbf{p}', \mathbf{y}) + \mathcal{L}_{shp}(\mathbf{y}', \mathbf{y})], \tag{5.1}$$

where $\mathcal{L}_{rec}$ is the mean squared error (MSE) between the original input image $\boldsymbol{x}$ and the reconstructed image $\boldsymbol{x}' = \mathcal{D}_{\phi_i}(E_\theta(\boldsymbol{x}))$, $\mathcal{L}_{seg}$ and $\mathcal{L}_{shp}$ are cross-entropy loss functions between ground truth $\mathbf{y}$ and predicted segmentation. The predicted segmentation can be the initial prediction $\mathbf{p} = \mathcal{D}_{\phi_s}(\mathcal{H}(E_\theta(\boldsymbol{x})))$, or the reconstructed prediction $\mathbf{p}' = \mathcal{C}_\psi(\mathbf{p})$ or reconstructed ground-truth map $\mathbf{y}' = \mathcal{C}_\psi(\mathbf{y})$. Different from $\mathcal{L}_{seg}$, optimizing $\mathcal{L}_{shp}(\mathbf{p}', \mathbf{y})$ will trigger gradient flows from STN to FTN. This allows STN to transfer shape knowledge to FTN to improve model generalizability.

5.2.3.3 Latent space data augmentation for hard example generation

Standard training is likely to suffer from over-fitting when training data is limited. To solve this problem, a novel latent space data augmentation method is proposed which allows FTN to automatically construct hard examples. As shown in Fig. 5.7(b), the proposed method requires a mask generator $\mathcal{G}$ to produce a mask $\mathbf{m}$ on the latent code $\mathbf{z}$. The masked latent code $\hat{\mathbf{z}} = \mathbf{z} \cdot \mathbf{m}$ is then fed to the decoders to reconstruct a corrupted image $\hat{\boldsymbol{x}} = \mathcal{D}_{\phi_i}(\hat{\mathbf{z}}_i)$ and segmentation $\hat{\mathbf{p}} = \mathcal{D}_{\phi_s}(\hat{\mathbf{z}}_s)$. Here, $\cdot$ denotes element-wise multiplication. In our work, we use latent

code masking for data augmentation. This differs from existing latent code dropout techniques for explicit regularization [131, 244]. By dynamically masking the latent code, the proposed method can generate samples with a wide diversity of image appearances and segmentations, which are not bound to specific image transformation or corruption functions. Below we introduce three latent-code masking schemes: random dropout $\mathcal{G}_{dp}$, and two targeted masking schemes, channel-wise targeted mask generation $\mathcal{G}_{ch}$ and spatial-wise targeted mask generation $\mathcal{G}_{sp}$.

(1) Random Masking with Dropout A naïve approach for latent code masking is random channel-wise dropout [131], which is an enhanced version of the original dropout method. An entire channel of the latent code can be masked with all zeros at a probability of p at training. Mathematically, this can be viewed as sampling a mask from a Bernoulli distribution:

$$\mathcal{G}_{dp}(\mathbf{m}^{(i)}; p) = \begin{cases} p & \mathbf{m}^{(i)} = \mathbf{0} \in \mathbb{R}^{h \times w} \\ 1-p & \mathbf{m}^{(i)} = \mathbf{1} \in \mathbb{R}^{h \times w}; \end{cases} \quad \forall i \in 1, ..., c. \tag{5.2}$$

The masked code at i-th channel is obtained via $\hat{\mathbf{z}}^{(i)} = \mathbf{z}^{(i)} \cdot \mathbf{m}^{(i)}$. In the following, we will use i-j-k to denote the three coordinates of latent code $\mathbf{z} \in \mathbb{R}^{c \times h \times w}$.

(2) Targeted Masking Inspired by the recent success on latent code masking for domain generalized image classification algorithm [244], we propose targeted latent code masking schemes which takes gradients as a clue to identify 'salient' features to mask. Following the common practice in adversarial data augmentation [109, 171], we take task-specific losses (image reconstruction loss and image segmentation loss) to calculate the gradients $\mathbf{g}_{\mathbf{z}_i}$, $\mathbf{g}_{\mathbf{z}_s}$ for $\mathbf{z}_i$ and $\mathbf{z}_s$ respectively, formulated as: $\mathbf{g}_{\mathbf{z}_i} = \nabla_{\mathbf{z}_i} \mathcal{L}_{rec}(\mathcal{D}_{\phi_i}(\mathbf{z}_i), \boldsymbol{x})$, $\mathbf{g}_{\mathbf{z}_s} = \nabla_{\mathbf{z}_s} \mathcal{L}_{seg}(\mathcal{D}_{\phi_s}(\mathbf{z}_s), \mathbf{y})$. By ranking the values of task-specific gradients, we can identify most predictive elements in the latent space to attack. We hypothesize that the elements with high response to task-specific loss functions are leading causes to performance drop under unforeseen domain shifts. We therefore focus on attacking these primary elements to simulate strong data distribution shifts. Two types of targeted masking are implemented, which mask features in latent code $\mathbf{z}$ along the channel

dimension and spatial dimension. They are:

a) channel-wise mask generator:

$$\mathcal{G}_{ch}(\mathbf{m}^{(i)}; \mathbf{g}_\mathbf{z}, p) = \begin{cases} \mathbf{m}^{(i)} = a\mathbf{1} \in \mathbb{R}^{h \times w} & \text{if } \mathbb{E}[\mathbf{g}_\mathbf{z}^{(i)}] \geq z_p^{ch} \\ \mathbf{m}^{(i)} = \mathbf{1} \in \mathbb{R}^{h \times w} & \text{if } \mathbb{E}[\mathbf{g}_\mathbf{z}^{(i)}] < z_p^{ch}; \end{cases} \quad \forall i \in 1, ..., c, \tag{5.3}$$

b) spatial-wise mask generator:

$$\mathcal{G}_{sp}(\mathbf{m}^{(j,k)}; \mathbf{g}_\mathbf{z}, p) = \begin{cases} \mathbf{m}^{(j,k)} = a\mathbf{1} \in \mathbb{R}^{c} & \text{if } \mathbb{E}[\mathbf{g}_\mathbf{z}^{(j,k)}] \geq z_p^{sp} \\ \mathbf{m}^{(j,k)} = \mathbf{1} \in \mathbb{R}^{c} & \text{if } \mathbb{E}[\mathbf{g}_\mathbf{z}^{(j,k)}] < z_p^{sp}; \end{cases} \quad \forall j \in [1, h], \forall k \in [1, w]. \tag{5.4}$$

Thresholds $z_p^{ch}, z_p^{sp} \in \mathbb{R}$ are top p-th value across the channel means and spatial means. a is an annealing factor randomly sampled from (0,0.5) to create soft masks. Compared to hard-masking (a=0), soft-masking generates more diverse corrupted data (see Fig. 5.8 and 5.9). Channel-wise masked code at i-th channel is obtained via $\hat{\mathbf{z}}^{(i)} = \mathbf{z}^{(i)} \cdot \mathbf{m}^{(i)}$. Spatial-wise masked code at (j, k) position is obtained via $\hat{\mathbf{z}}^{(j,k)} = \mathbf{z}^{(j,k)} \cdot \mathbf{m}^{(j,k)}$.

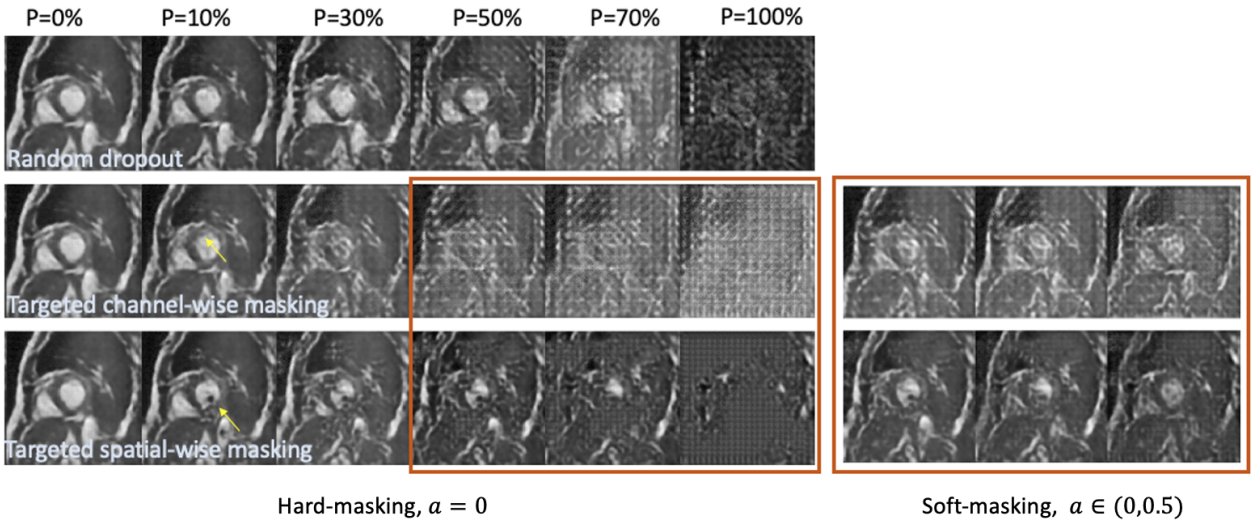


Figure 5.8: **Visualization of generated corrupted images.** Three types of latent code masking schemes generate a *diverse* set of challenging images with *unseen* mixed artifacts, e.g. 'dark dots', 'checkerboard artifacts', 'blurring'. a: the annealing factor in Eq. 5.3 and Eq. 5.4.

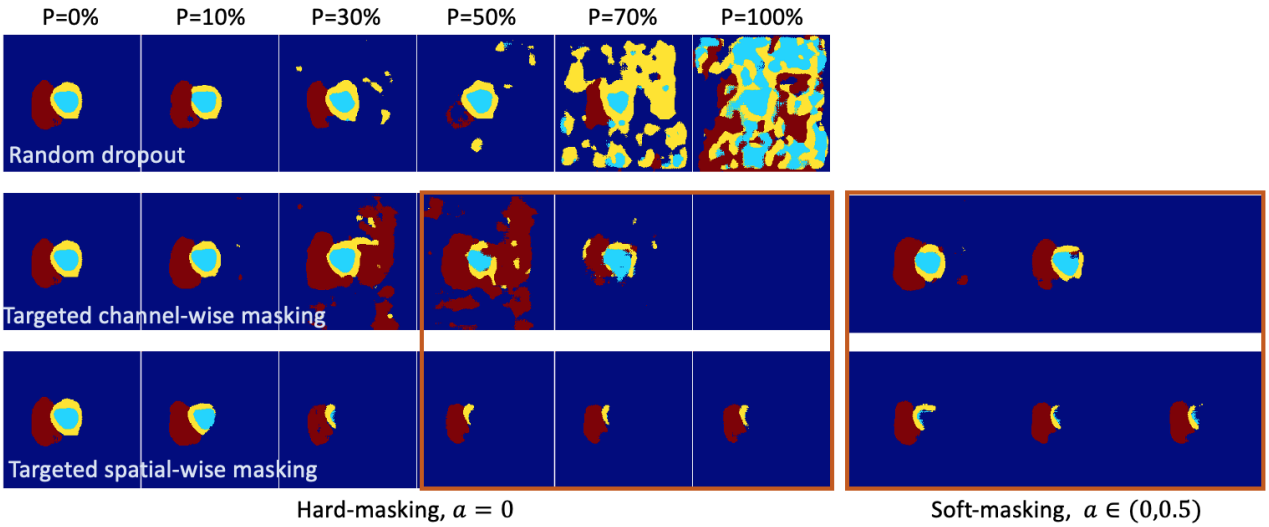


Figure 5.9: **Visualization of generated corrupted segmentation maps.** Three types of latent code masking schemes generate various over-segmented and under-segmented predictions at different thresholds p. Compared to hard-masking, soft-masking produces milder but more *diverse* corrupted images and segmentation maps. a: the annealing factor in Eq. 5.3 and Eq. 5.4

5.2.3.4 Cooperative training

During training, we randomly apply one of the three mask generators to both $\mathbf{z}_i$, $\mathbf{z}_s$. This process generates a rich set of corrupted images $\hat{\boldsymbol{x}}$ and segmentations $\hat{\mathbf{p}}$ on-the-fly. It allows us to train our dual-network on three hard example pairs, i.e. corrupted images-clean images $(\hat{\boldsymbol{x}}, \boldsymbol{x})$, corrupted images-GT $(\hat{\boldsymbol{x}}, \mathbf{y})$, corrupted prediction-GT $(\hat{\mathbf{p}}, \mathbf{y})$. The final loss for the proposed cooperative training method is a combination of losses defined on easy examples and hard examples: $\mathcal{L}_{cooperative} = \mathcal{L}_{std} + \mathcal{L}_{hard}$, where $\mathcal{L}_{hard}$ is defined as:

$$\mathcal{L}_{hard} = \mathbb{E}_{\hat{\boldsymbol{x}},\hat{\mathbf{p}},\boldsymbol{x},\mathbf{y}}[\mathcal{L}_{rec}(\mathcal{D}_{\phi_i}(E_\theta(\hat{\boldsymbol{x}})), \boldsymbol{x}) + \mathcal{L}_{seg}(\bar{\mathbf{p}}, \mathbf{y}) + \mathcal{L}_{shp}(\mathcal{C}_\psi(\hat{\mathbf{p}}), \mathbf{y}) + \mathcal{L}_{shp}(\mathcal{C}_\psi(\bar{\mathbf{p}}), \mathbf{y})]. \tag{5.5}$$

Here, $\bar{\mathbf{p}} = \mathcal{D}_{\phi_i}(\mathcal{H}(E_\theta(\hat{\boldsymbol{x}})))$ is FTN's predicted segmentation on $\hat{\boldsymbol{x}}$.

5.2.4 Experiments

To evaluate the efficacy of the proposed method, we apply it to the cardiac image segmentation task to segment the left ventricle cavity, left ventricular myocardium and right ventricle from MR images. Three datasets are used: the Automated Cardiac Diagnosis Challenge

dataset (ACDC)[8] [6], Multi-centre, Multi-vendor & Multi-disease Cardiac Image Segmentation Challenge (M&Ms) dataset [9] [245] and corrupted ACDC, named as ACDC-C. For all experiments, the training set is a **single-site** set of only 10 subjects from ACDC. 10 and 20 subjects from ACDC are used for validation and intra-domain test. The multi-site M&Ms dataset (150 subjects from 5 different sites) is used for cross-domain test. The ACDC-C dataset is used for evaluating the robustness of the method for corrupted images. Challenging scenarios are simulated, where 20 ACDC test subjects are augmented three times with four different types of MR artefacts: bias field, ghosting, motion and spike artifacts [184] using the TorchIO[10] toolkit. This produces 4 subsets with 60 subjects, named as *RandBias*, *RandGhosting*, *RandMotion*, *RandSpike* in experiments.

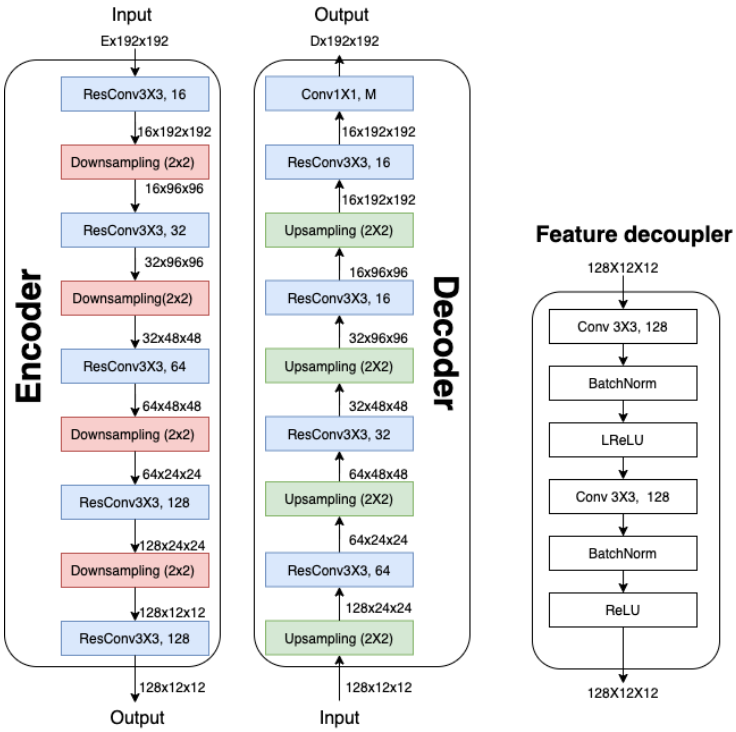


Figure 5.10: **Structures of the Unet-like encoder-decoder pairs, and the feature decoupler used in our experiments.** We used the same structures for encoders and decoders accordingly. E: # of input channel(s), D: # of output channel(s). ResConv: Convolutional Block with residual connections [246]. Conv: Standard convolutional kernels. Of note, our framework is **generic**, other encoders and decoders can also be used.

[8] https://www.creatis.insa-lyon.fr/Challenge/acdc/databases.html
[9] https://www.ub.edu/mnms/
[10] https://github.com/fepegar/torchio

5.2.4.0.1 Implementation and evaluation. We employed the image pre-processing and default data augmentation pipeline described in [15], including common photo-metric and geometric image transformations. Our encoder and decoder pairs support general structures. Without loss of generality, we used a U-net like structure[32]. Fig.5.10 visualizes detailed structures of encoder-decoder pairs as well as the latent space decoupler. For mask generation, we randomly select one type of the masking scheme described above at training, where p is randomly selected from [0% , 50%]. We use the Adam optimizer with a batch size of 20 to update network parameters, with a learning rate=$1e^{-4}$. Our code is available on the Github[11]. For all methods, we trained the same network *three* times using a set of randomly selected 10 ACDC subjects (600 epochs each run, on an Nvidia®, using Pytorch). The average Dice score is reported for segmentation performance evaluation.

5.2.5 Results and discussion

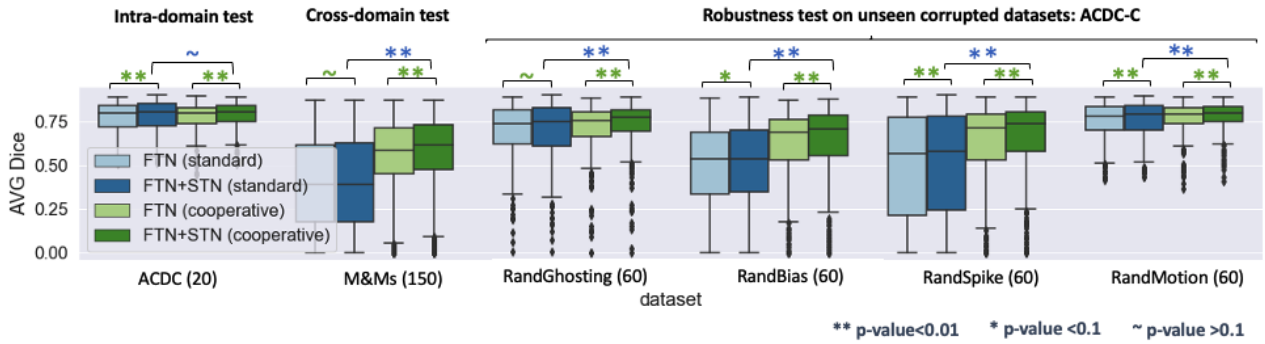


Figure 5.11: **Boxplots of average Dice scores on the intra-domain test set, cross-domain test set, and unseen corrupted testsets.** All networks were trained using only 10 subjects. Compared to standard training, cooperative training with self-generating hard examples greatly improves the segmentation performance on various *unseen, challenging* domains (p-value$<$ 0.01, average improvement: 15%).

5.2.5.1 Experiment 1: standard training vs cooperative training

We compared the proposed cooperative training method with the standard training method (using $\mathcal{L}_{standard}$ only) using the same backbone network structure. Fig. 5.11 shows the box-plots

[11] https://github.com/cherise215/Cooperative_Training_and_Latent_Space_Data_Augmentation

for each method. While both methods achieve comparable performance on the intra-domain test set (p-value> 0.1), it is clear that cooperative training with dual-network (FTN+STN) yields the best performance across out-of domain test sets (see dark green boxes). Consistent improvements made by STN can be clearly observed across all domains. By contrast, STN with
2495 standard training fails to provide significant improvements on some datasets (p-value> 0.1). This indicates the superiority of cooperative training with latent space data augmentation.

5.2.5.2 Experiment 2: latent space data augmentation vs image space data augmentation

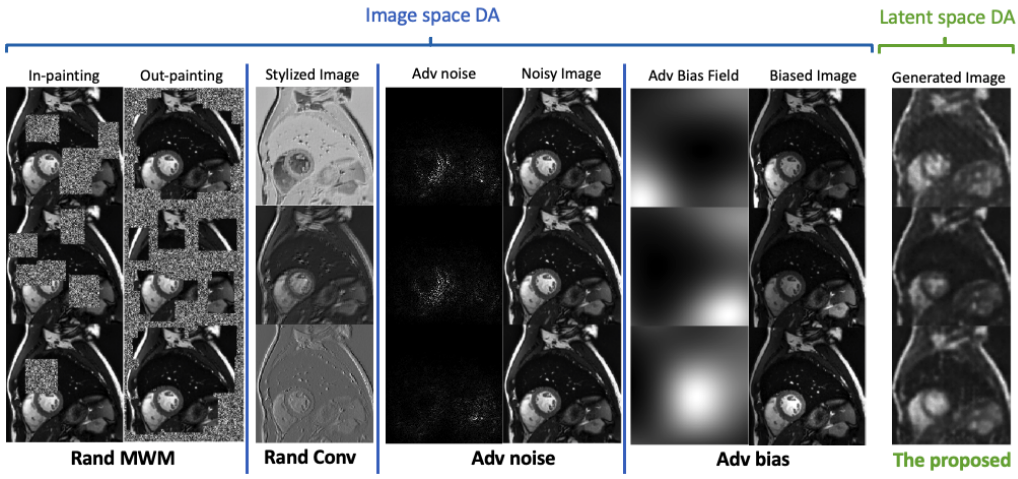


Figure 5.12: **Visualization of augmented images using input space data augmentation and the proposed latent space data augmentation.** DA: data augmentation. Adv: Adversarial.

Table 5.10: **Comparison results of segmentation performances of the proposed latent space data augmentation and competitive image space data augmentation methods for domain generalization.** The proposed latent space augmentation method improves the performance on out-of-domain datasets compared to image space data augmentation methods. AVG: average Dice scores across six datasets. Red numbers are average Dice scores under 0.5.

Method	ACDC	M&Ms	RandBias	RandGhosting	RandMotion	RandSpike	AVG (FTN)	AVG (FTN+STN)
Standard training	0.7681	0.3909	0.4889	0.6964	0.7494	0.4901	0.5970	0.6018
Rand MWM [236]	0.7515	0.3984	0.4914	0.6685	0.7336	0.5713	0.6024	0.6131
Rand Conv [247]	0.7604	0.4544	0.5538	0.6891	0.7493	0.4902	0.6162	0.6404
Adv Noise [174]	0.7678	0.3873	0.4903	0.6829	0.7543	0.6244	0.6178	0.6276
Adv Bias [15]	0.7573	**0.6013**	**0.6709**	0.6773	0.7348	0.3840	0.6376	0.6604
Proposed w. $\hat{\boldsymbol{x}}$	0.7497	0.5154	0.5921	0.6921	0.7417	**0.6633**	0.6591	0.6709
Proposed w. $\hat{\boldsymbol{x}}, \hat{\mathbf{p}}$	**0.7696**	0.5454	0.6174	**0.7073**	**0.7643**	0.6226	**0.6711**	**0.6901**

We compared the proposed latent space based method to other competitive image space data augmentation methods: a) *random multi-window in-and-out masking (Rand MWM)* [236, 237], which uses an enhanced variant of Cutout [234] and Patch Gaussian [235] to introduce patch-wise perturbation to images; b) *random convolutional kernels (Rand Conv)* [247], which applies various random convolutional kernels to augment image texture and appearance variations; c) *adversarial noise (Adv Noise)* [174]; d) *adversarial bias field (Adv Bias)* [15], which augments image styles by adding realistic intensity inhomogeneities. We visualize augmented images using above methods in Fig. 5.12. For methods under comparison, we used their official code implementation if available and ran experiments using the same backbone network for fairness. Results are shown in Table 5.10.

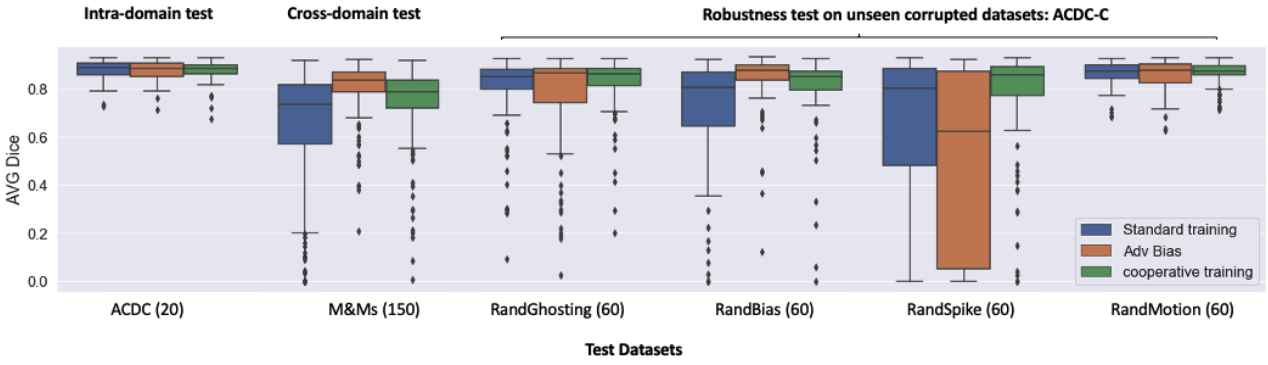


Figure 5.13: **Boxplots of segmentation results in the large training data setting.** In the large training data setting (70 ACDC subjects for training), when compared to the baseline method (standard training), our cooperative training method can further improve not only intra-domain segmentation accuracy (with reduced variance) but also robustness against various domain shifts. Adv bias, by contrast, fails to provide consistent improvement. This reveals our method's great potential to be applied to a wide range of scenarios for both improved generalization and robustness.

Surprisingly, with limited training data, both random and adversarial data augmentation methods do not necessarily improve the network generalization on all datasets. While *AdvBias* achieves the best performance on *M&Ms* dataset and *RandBias*, this method has a side effect, making it more sensitive to the spiking artifacts (Dice score 0.4901 vs 0.3840). By contrast, the proposed latent space data augmentation achieves the top average performance across six datasets, without any dramatic failures (Dice score < 0.5). Similar results can be found in a large training setting, see Fig. 5.13. Our method can generate not only perturbed images but also realistically corrupted segmentations with increased uncertainty (Fig. 5.14). These corrupted segmentations attribute to the increased model generalization (AVG Dice: 0.6709 vs

0.6901). While one may argue that characterizing and combining various image-space DAs and corruptions together could be an interesting direction to improve cross-domain performance, it is time-consuming and computationally inefficient to find the optimal data augmentation policy [248], and has the risk of sacrificing intra-domain performance [249].

5.2.5.3 Experiment 3: ablation study

Table 5.11: **Effectiveness of the targeted masking, latent code decoupler $\mathcal{H}$ and cooperative training.**

Methods	FTN	FTN+STN
w.o. $\mathcal{G}_{ch}, \mathcal{G}_{sp}$	0.6344	0.6584
share code (a) ($\mathbf{z}_i = \mathbf{z}_i, \mathbf{z}_s = \mathbf{z}_i$)	0.6625	0. 6868
share code (b) ($\mathbf{z}_i = \mathbf{z}_s, \mathbf{z}_s = \mathbf{z}_s$)	0.6343	0.6587
Separate Training [242]	0.6020	0.6077
Proposed	**0.6711**	**0.6901**

We further investigate three key contributions: 1) the proposed targeted masking; 2) latent code decoupler $\mathcal{H}$; 3) cooperative training. Results are shown in Table 5.11. We can see that disabling $\mathcal{G}_{ch}, \mathcal{G}_{sp}$ drops the average Dice score from 0.6901 to 0.6584, highlighting the effectiveness of targeted masking. Fig. 5.8 and 5.9 shows that targeted masking focuses more on attacking cardiac structures, resulting in more challenging images with mixed artifacts and under or over-segmented predictions. We compared the proposed network architecture to its two variants, where $\mathbf{z}_i$ and $\mathbf{z}_s$ are shared in two different ways. Both variants lead to inferior performance. This suggests the benefit of $\mathcal{H}$ for a more sparse $\mathbf{z}_s$ code. Image reconstruction requires low-level information, whereas image segmentation relies on more concentrated high-level information. Introducing $\mathcal{H}$ explicitly defines a hierarchical feature structure to improve model generalization. Lastly, we compared our method to the state-of-the-art denoising auto-encoder-based shape refinement method (Separate Training) [242] where FTN and STN are trained independently. It has been shown that this learning-based method can outperform the commonly used non-learning-based condition random field-based refinement method [250]. Results show that our method can greatly outperform this advanced method by a large margin

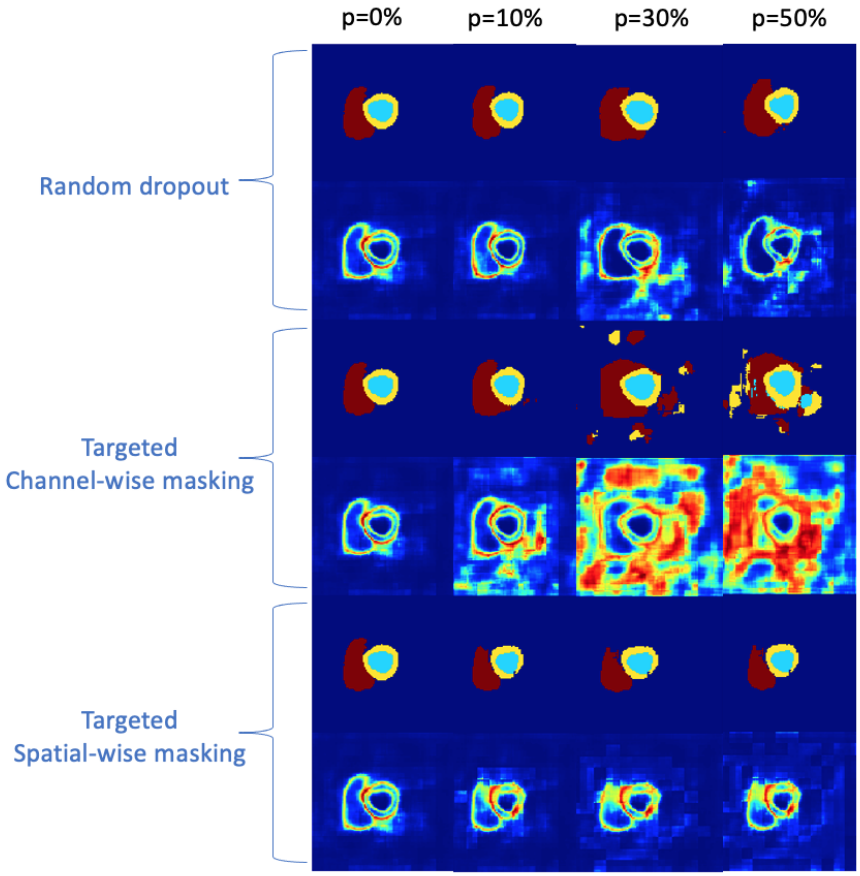


Figure 5.14: **Visualization of corrupted segmentations and corresponding entropy maps generated with the proposed three latent space masking schemes.** The first row and the second row in each block display the examples of corrupted segmentation and corresponding entropy maps, respectively. Latent masking schemes generate *realistic* poor segmentation with *increased* entropy, which is beneficial to train our denoising autoencoder (STN) for shape correction. Latent masking schemes generate *realistic* poor segmentation with *increased* entropy, which is beneficial to train our denoising autoencoder (STN) for shape correction.

(Dice score 0.6901 vs. 0.6077), highlighting the benefits of the cooperative training strategy for enhancing learning-based shape refinement and correction.

5.2.6 Conclusion

We present a novel cooperative training framework together with a latent space masking-based data augmentation method. Experiments show that it greatly improves model generalizability and robustness against unforeseen domain shifts, despite the training data being collected from only one domain. Unlike existing methods which require multi-domain datasets or domain

knowledge to specify particular forms of image transformation and corruption functions, our latent space data augmentation method requires *little* human effort, and it has the potential to be applied to other data-driven applications. Although we only demonstrate the performance for cardiac image segmentation, our *generic* framework has the potential to be extended to a wide range of data-driven applications. However, as also noted by [251], one limitation of the two-stage approach with shape refinement is that it may not be suitable for abnormality segmentation (e.g., tumor, lesion segmentation) where the region-of-interests are of higher shape complexity and wider shape variety across different subjects compared to anatomical structures. In that case, STN could fail to learn a generalized shape prior knowledge for shape correction and refinement.

Chapter 6

Conclusion

In this thesis, we have presented three ways to improve the generalization performance of deep learning models with limited labeled data: a) exploiting auxiliary data using multi-task learning, b) exploiting unlabeled data for semi-supervised learning and unsupervised domain adaptation, c) maximizing the value of limited labeled data by designing effective data augmentation. We have designed and validated our methods on medical image segmentation applications, wherein the scarcity of labels is a significant problem due to the high labeling costs. A summary of thesis achievements and some potential future works are presented in the following.

6.1 Summary of thesis achievements

Multi-task learning with *auxiliary data* from related tasks improves model generalization.

In real-world applications, while there is limited labeled data for a particular task, there are also auxiliary data available for other related tasks that can provide complementary information to each other. In Chapter 3, we introduced two works that successfully extract useful contexts from auxiliary data and leverage them to help the main image segmentation task. In the first part of Chapter 3, we presented a multi-task U-net for left atrial MR segmentation (Sec. 3.1), which performs image classification (pre-ablation/post-ablation) and segmentation simultan-

eously, using a shared feature encoder and two different task-specific decoders. We compared the proposed multi-task U-net to a single-task U-net without the classification branch. Results show that multi-task learning improved model generalization, as it can encourage the network to learn shared representations across the two tasks. In the second work, we presented a multi-view shape prior aware segmentation network for cardiac myocardium segmentation from SAX images (Sec. 3.2). In particular, we introduced a novel shape-aware multi-view convolution neural network that learns latent cardiac shape priors from multiple standard views by performing the cross-view shape prediction task. We then presented a multi-view U-Net where we introduced a 'fuse block' to the bottleneck of the network so that it can automatically incorporate the learned anatomical shape priors in the latent feature space. In this way, the learned features (shape priors) from other tasks can be explicitly shared with the main segmentation tasks to improve the segmentation robustness. Experimental results show that adding shape priors is especially useful when segmenting challenging slices where the image contrast is low or boundaries of anatomical structures are unclear.

To summarize, we have presented two ways to enable knowledge sharing across multiple related tasks with auxiliary data, either by parameter sharing or feature sharing. Our results suggest that knowledge sharing across multiple related tasks can help the network obtain higher accuracy on unseen test data when compared to the standard approach learned from a single task.

Utilizing *unlabeled data* for semi-supervised learning/unsupervised domain adaptation The second contribution we made is developing two learning frameworks to utilize unlabeled data for enhancing model generalization. Labeling medical images requires expertise, and can be super expensive and time consuming. It is more economical to just label a small set of images and then utilize a large number of unlabeled images for enhancing neural networks. In Chapter 4 Sec 4.1, we developed an adversarial data augmentation method, which can be applied to both labeled and unlabeled images to facilitate semi-supervised learning. This method takes segmentation network and image information into account, simulating effective intensity homogeneity (bias fields) to perturb images so that the neural network is fooled to produce inconsistent predictions. By forcing the network to produce consistent predictions on clean

images and perturbed images, we enhance the network robustness against bias fields, and more importantly, utilize unlabeled data to improve the accuracy on unseen test data in the same domain.

The above method works under the assumption where the labeled and unlabeled images are from the same domain. In Sec. 4.2, we also demonstrated a learning framework that can transfer knowledge learned from one domain with a set of labeled images to a different domain with unlabeled images only. The two domains consist of images from two different imaging sequences (bSSFP vs LGE imaging) where large differences in terms of image appearances can be observed. We demonstrated that an image style translation network based on a generative model (i.e.GAN) is capable of modeling the conditional image distribution so that labeled bSSFP images can be translated into LGE-like images automatically. These synthetic LGE-like labeled images make it easy to train a segmentation network for LGE images, even without any manually labeled LGE images. In addition, we proposed a cascaded network, which consists of two U-nets where the second U-net utilizes the predicted probabilistic maps produced by the first U-net as shape information to assist the segmentation. The proposed method greatly outperformed several baseline methods and other unsupervised learning methods, achieving the state-of-the-art segmentation accuracy on the target domain in the public multi-sequence cardiac MR segmentation challenge [11][1].

Constructing effective data augmentation for *limited single-domain data* to improve cross-domain generalization.

In the worst case of data scarcity, there is only one single domain data with limited data diversity for training. In order to improve model performance across various unseen domains, we developed a general training/testing pipeline for improving the generalization of neural network-based CMR image segmentation methods. With cardiac imaging, we highlighted that it is important to perform data normalization and augmentation (section 5.1.5.2) to align and expand training data distribution for effective training. We also highlighted that the network structure and capacity also matter for model generalization (section 5.1.5.1). The proposed method achieves promising results on a large number of test images from various scanners and

[1] https://zmiclab.github.io/projects/mscmrseg19/index.html

sites even though the training set is from one scanner, one site. Besides, the trained network is capable of segmenting healthy subjects as well as a group of pathological cases from multiple sources while it had only been trained with a small portion of pathological cases. This proposed method has been adopted in the winner algorithm in the public multi-center, multi-vendor and multi-disease cardiac image segmentation challenge (M&Ms[2]), as a strong base to improve cross-domain segmentation performance [245, 252]. However, one limitation of our method is that it requires domain knowledge, expertise to design the data augmentation strategy in the training and testing pipeline. Also, results show that the trained segmentation network still has some limitations, such as high sensitivity to images with poor quality, e.g., images with artifacts.

To further enhance model robustness against unseen domain shifts and imaging artifacts, we presented a novel cooperative training framework in together with a latent space masking-based data augmentation method in Sec. 5.2. The latent space data augmentation method performs channel-wise and spacial-wise masking in self-discovering image content and shape-related latent code space. Specifically, we developed methods to mask latent codes in both random and adversarial fashions. Images are reconstructed with those masked latent codes to form a diverse set of challenging images and corrupted segmentation maps, which are used to reinforce neural networks' training. By training a cardiac segmentation using training data from only *one* hospital and evaluating the network on *multiple* different datasets from different sources, we demonstrated that the proposed method could improve cross-site segmentation performance and particularly increased robustness against various unforeseen imaging artifacts compared to strong baseline methods.

A more generalized, robust cardiac segmentation model for cardiac imaging applications Last but not least, our works provide several fully automated cardiac MR segmentation frameworks based on CNN, with improved model generalization and robustness against unseen domain shifts. Such a segmentation model can be used as a tool to accelerate the analysis of massive images collected at various unseen sites, providing the visualization of cardiac anatomy, volume quantification, and valuable functional information such as ejection fraction measure-

[2] https://www.ub.edu/mnms/

ment and assisting the diagnosis of cardiovascular diseases. The provided segmentation results can also be used to support follow-up clinical research studies, such as shape modeling and analysis [45], cardiac motion analysis [253], treatment planning and therapy response prediction [20], as well as survival prediction [253].

6.2 Future work

In this section, we will discuss potential research directions that build upon our work, as well as the limitations and open challenges of deep learning that could be considered as future research topics.

Beyond medical image segmentation So far, all works we presented have only been applied to medical imaging segmentation tasks. However, most of our works have the potential to be applied to other medical imaging tasks, such as medical image classification, detection, and image registration, and reconstruction for improved model generalization. For example, in this thesis, we have presented four different ways to augment data to alleviate data scarcity, which are based on:

- *Hand-crafted transformations with random sampling*: applying a stack of traditional image transformation functions to increase the variation of image appearances and geometry, including gamma correction-based image contrast augmentation, affine transformation, Sec. 3.1, Sec. 5.1;

- *Gradient-based data augmentation*: taking the gradients of the network to optimize image transformations parameters so that augmented images can challenge the network to produce inconsistent predictions. These challenging images can then used to regularize the network better, Sec. 4.1;

- *Generative model-based data augmentation*: employing a generative model (i.e., GAN) to translate images across different image sequences, Sec. 4.2;

- *Latent space-based data augmentation*: applying random/adversarial-based masking schemes in the latent spaces and then using masked latent codes to generate corrupted images and predictions as hard examples to inform network training, Sec. 5.2.

The above data augmentation frameworks are generic and thus have the potential to be adapted to other data-driven methods for improved generalization. For example, our gradient-based data augmentation (i.e., adversarial bias field data augmentation) can also be applied to image registration, which can encourage the network to produce robust deformation fields regardless of the presence of intensity inhomogeneity in fixed and moving images. Also, it is interesting to extend the adversarial bias field data augmentation with other forms of image transformations (e.g., affine transformation, diffeomorphic transformation), to increase the data diversity of augmented data for enhanced regularization [3].

Learning from heterogeneous labeled datasets In this thesis, all labeled datasets for training are labeled by the same group of observers/physicians using a consistent labeling protocol. In this case, label inconsistency is minimized as expected. However, in real-world applications, training and testing images are often gathered from different sources in order to have substantial data diversity to reflect the spectrum of real-world diversity. These images are often labeled by different groups of observers for various reasons (e.g., save time). As a result, it is likely to have large inter-observer variability and inconsistent label quality with such a large volume. A number of works have reported the existence of missing labels and inconsistent labeling protocols across different cardiac image datasets [7, 84]. These inconsistencies can be a major obstacle for transferring, evaluating, and deploying deep learning models trained from one domain (e.g., hospital) to another. Therefore, it is of great interest to develop an automated tool to combine existing public datasets from multiple sources and then to harmonize them to a unified, high-quality dataset for training and evaluation. This tool can not only open the door for crowd-sourcing but also enable the rapid deployment of those DL-based applications.

Improving model interpretability and explainablity As introduced in Sec. 2.3.3, most deep learning systems have poor interpretability and explainability, as they are 'black-box' in

[3] A work built upon this idea has been submitted to Medical Image Analysis [254] recently.

nature. Compared to traditional symbolic machine learning systems, deep learning systems are in general difficult to interpret their predictions, i.e., why certain decisions or predictions have been made. This issue makes the model intractable for model verification and ultimately untrustworthy. In the future, we will look into developing deep learning algorithms with improved interpretability and explainability to support the development of safety-critical medical imaging applications. Theoretically, we could build causality into neural networks to understand cause and effect, e.g., knowing why a model might fail [255]. There are several emerging fields on this topic, such as explainable artificial intelligence (XAI), and causal artificial intelligence [256]. Three main streams for enhancing model explainability and interpret-ability are feature importance estimation [117], causal effects of model components [256, 257], and counterfactual explanation [256, 258]. Yet, many works on this topic are still conceptual, as it is not easy to verify causal interpretability. Another direction is to add a failure awareness module into the deployed networks. This can be achieved by providing users with quantified measures, such as prediction quality scores [219], uncertainty maps [78] and attention maps [259].

Bibliography

[1] 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, p. 25, 2020, ISSN: 2297-055X. DOI: 10.3389/fcvm.2020.00025.

[2] H. Greenspan, B. Van Ginneken and R. M. Summers, 'Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique,' *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1153–1159, 2016. DOI: 10.1109/TMI.2016.2553401.

[3] D. Shen, G. Wu and H.-I. Suk, 'Deep learning in medical image analysis,' *Annual Review of Biomedical Engineering*, vol. 19, pp. 221–248, 2017. DOI: 0.1146/annurev-bioeng-071516-044442.

[4] W. Bai, M. Sinclair, G. Tarroni, O. Oktay, M. Rajchl, G. Vaillant, A. Lee, N. Aung, E. Lukaschuk, M. Sanghvi, F. Zemrak, K. Fung, J. Paiva, V. Carapella, Y. Kim, H. Suzuki, B. Kainz, P. Matthews, S. Petersen and D. Rueckert, 'Automated cardiovascular magnetic resonance image analysis with fully convolutional networks,' *Journal of Cardiovascular Magnetic Resonance*, vol. 20, no. 1, p. 65, 2018. DOI: 10.1186/s12968-018-0471-x.

[5] S. E. Petersen, N. Aung, M. M. Sanghvi, F. Zemrak, K. Fung, J. M. Paiva, J. M. Francis, M. Y. Khanji, E. Lukaschuk, A. M. Lee, V. Carapella, Y. J. Kim, P. Leeson, S. K. Piechnik and S. Neubauer, 'Reference ranges for cardiac structure and function using cardiovascular magnetic resonance (CMR) in caucasians from the UK biobank

population cohort,' *Journal of Cardiovascular Magnetic Resonance*, vol. 19, no. 1, p. 18, 2017. DOI: 10.1186/s12968-017-0327-9.

[6] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. Gonzalez 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, I. Isgum, Y. Jang, Y. Hong, J. Patravali, S. Jain, O. Humbert and P.-M. Jodoin, 'Deep learning techniques for automatic MRI cardiac Multi-Structures segmentation and diagnosis: Is the problem solved?' *IEEE Transactions on Medical Imaging*, vol. 37, no. 11, pp. 2514–2525, 2018, Data source: https://acdc.creatis.insa-lyon.fr/ (Accessed September 1, 2019)., ISSN: 0278-0062, 1558-254X. DOI: 10.1109/TMI.2018.2837502.

[7] C. Chen, W. Bai, R. H. Davies, A. N. Bhuva, C. H. Manisty, J. B. Augusto, J. C. Moon, N. Aung, A. M. Lee, M. M. Sanghvi, K. Fung, J. M. Paiva, S. E. Petersen, E. Lukaschuk, S. K. Piechnik, S. Neubauer and D. Rueckert, 'Improving the generalizability of convolutional neural Network-Based segmentation on CMR images,' *Frontiers in Cardiovascular Medicine*, vol. 7, p. 105, 2020, ISSN: 2297-055X. DOI: 10.3389/fcvm.2020.00105.

[8] P. F. Ferreira, P. D. Gatehouse, R. H. Mohiaddin and D. N. Firmin, 'Cardiovascular magnetic resonance artefacts,' *Journal of Cardiovascular Magnetic Resonance*, vol. 15, no. 1, pp. 1–39, 2013. DOI: 10.1186/1532-429X-15-41.

[9] D. Kiblboeck, C. Reiter, J. Kammler, P. Schmit, H. Blessberger, J. Kellermair, F. Fellner and C. Steinwender, 'Artefacts in 1.5 tesla and 3 tesla cardiovascular magnetic resonance imaging in patients with leadless cardiac pacemakers,' *Journal of Cardiovascular Magnetic Resonance*, vol. 20, no. 1, p. 47, 2018. DOI: 10.1186/s12968-018-0469-4.

[10] H. Xu and S. Mannor, 'Robustness and generalization,' in *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010*, A. T. Kalai and M. Mohri, Eds., Omnipress, 2010, pp. 503–515. DOI: 10.1007/978-3-030-32245-8_58.

[11] X. Zhuang, J. Xu, X. Luo, C. Chen, C. Ouyang, D. Rueckert, V. M. Campello, K. Lekadir, S. Vesal, N. RaviKumar, Y. Liu, G. Luo, J. Chen, H. Li, B. Ly, M. Sermesant, H. Roth, W. Zhu, J. Wang, X. Ding, X. Wang, S. Yang and L. Li, 'Cardiac segmentation on late gadolinium enhancement MRI: A benchmark study from Multi-Sequence cardiac MR segmentation challenge,' *arXiv Preprint*, 2020, arXiv:2006.12434.

[12] C. Chen, W. Bai and D. Rueckert, 'Multi-task learning for left atrial segmentation on GE-MRI,' in *Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges - 9th International Workshop, STACOM 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers*, vol. 11395, Springer International Publishing, 2019, pp. 292–301. DOI: 10.1007/978-3-030-12029-0_32.

[13] C. Chen, C. Biffi, G. Tarroni, S. Petersen, W. Bai and D. Rueckert, 'Learning shape priors for robust cardiac MR segmentation from multi-view images,' in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2019 - 22nd International Conference, Shenzhen, China*, vol. 11765, Springer, 2019, pp. 523–531. DOI: 10.1007/978-3-030-32245-8_58.

[14] C. Chen, C. Ouyang, G. Tarroni, J. Schlemper, H. Qiu, W. Bai and D. Rueckert, 'Unsupervised multi-modal style transfer for cardiac MR segmentation,' in *Statistical Atlases and Computational Models of the Heart. Multi-Sequence CMR Segmentation - 10th International Workshop, STACOM 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Revised Selected Papers*, ser. Lecture Notes in Computer Science, vol. 12009, Springer International Publishing, 2019, pp. 209–219. DOI: 10.1007/978-3-030-39074-7_22.

[15] C. Chen, C. Qin, H. Qiu, C. Ouyang, S. Wang, L. Chen, G. Tarroni, W. Bai and D. Rueckert, 'Realistic adversarial data augmentation for MR image segmentation,' in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, Springer International Publishing, 2020, pp. 667–677. DOI: 10.1007/978-3-030-59710-8_65.

[16] C. Chen, K. Hammernik, C. Ouyang, Q. Chen, W. Bai and D. Rueckert, 'Cooperative training and latent space data augmentation for robust segmentation,' in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, arXiv:2107.01079, Springer International Publishing, 2021.

[17] W. Bai, C. Chen, G. Tarroni, J. Duan, F. Guitton, S. E. Petersen, Y. Guo, P. M. Matthews and D. Rueckert, 'Self-Supervised learning for cardiac MR image segmentation by anatomical position prediction,' in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2019 - 22nd International Conference, Shenzhen, China, October 13-17, 2019, Proceedings, Part II*, Springer International Publishing, 2019, pp. 541–549. DOI: 10.1007/978-3-030-32245-8_60.

[18] E. P. V. Le, N. R. Evans, J. M. Tarkin, M. M. Chowdhury, F. Zaccagna, C. Wall, Y. Huang, J. R. Weir-Mccall, C. Chen, E. A. Warburton, C. B. Schonlieb, E. Sala and J. H. F. Rudd, 'Contrast CT classification of asymptomatic and symptomatic carotids in stroke and transient ischaemic attack with deep learning and interpretability,' *European Heart Journal*, vol. 41, 2020, ISSN: 0195-668X, 1522-9645. DOI: 10.1093/ehjci/ehaa946.2418.

[19] C. Ouyang, C. Biffi, C. Chen, T. Kart, H. Qiu and D. Rueckert, 'Self-supervision with superpixels: Training few-shot medical image segmentation without annotation,' in *European Conference on Computer Vision - ECCV 2020, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIX*, A. Vedaldi *et al.*, Eds., vol. 12374, Springer, 2020, pp. 762–780. DOI: 10.1007/978-3-030-58526-6_45.

[20] E. Puyol-Antón, C. Chen, J. R. Clough, B. Ruijsink *et al.*, 'Interpretable deep models for cardiac resynchronisation therapy response prediction,' in *Medical Image Computing and Computer Assisted Intervention*, Springer, 2020. DOI: 10.1007/978-3-030-59710-8_28.

[21] C. Qin, S. Wang, C. Chen, H. Qiu, W. Bai and D. Rueckert, 'Biomechanics-Informed neural networks for myocardial motion tracking in MRI,' in *Medical Image Computing*

and Computer Assisted Intervention – MICCAI 2020, Springer International Publishing, 2020, pp. 296–306. DOI: 10.1007/978-3-030-59716-0_29.

[22] S. Wang, G. Tarroni, C. Qin, Y. Mo, C. Dai, C. Chen, B. Glocker, Y. Guo, D. Rueckert and W. Bai, 'Deep generative Model-Based quality control for cardiac MRI segmentation,' in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, Springer International Publishing, 2020, pp. 88–97. DOI: 10.1007/978-3-030-59719-1_9.

[23] Z. Xiong, Q. Xia, Z. Hu, N. Huang, C. Bian, Y. Zheng, S. Vesal, N. Ravikumar, A. Maier, X. Yang, P.-A. Heng, D. Ni, C. Li, Q. Tong, W. Si, E. Puybareau, Y. Khoudli, T. Géraud, C. Chen, W. Bai, D. Rueckert, L. Xu, X. Zhuang, X. Luo, S. Jia, M. Sermesant, Y. Liu, K. Wang, D. Borra, A. Masci, C. Corsi, C. de Vente, M. Veta, R. Karim, C. J. Preetha, S. Engelhardt, M. Qiao, Y. Wang, Q. Tao, M. Nuñez-Garcia, O. Camara, N. Savioli, P. Lamata and J. Zhao, 'A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging,' *Medical Image Analysis*, vol. 67, p. 101 832, 2021, ISSN: 1361-8415, 1361-8423. DOI: 10.1016/j.media.2020.101832.

[24] S. Wang, C. Qin, N. Savioli, C. Chen, D. O'Regan, S. Cook, Y. Guo, D. Rueckert and W. Bai, 'Joint motion correction and super resolution for cardiac segmentation via latent optimisation,' in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, arXiv:2107.03887, 2021.

[25] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, 'Going deeper with convolutions,' in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 2015, pp. 1–9. DOI: 10.1109/CVPR.2015.7298594.

[26] L. J. Ba, J. R. Kiros and G. E. Hinton, 'Layer normalization,' *arXiv Preprint*, 2016, arXiv:1607.06450.

[27] D. Ulyanov, A. Vedaldi and V. S. Lempitsky, 'Instance normalization: The missing ingredient for fast stylization,' *arXiv Preprint*, 2016, arXiv:1607.08022.

[28] K. Simonyan and A. Zisserman, 'Very deep convolutional networks for large-scale image recognition,' in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015, p. 14.

[29] D. C. Ciresan, A. Giusti, L. M. Gambardella and J. Schmidhuber, 'Deep neural networks segment neuronal membranes in electron microscopy images,' in *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger, Eds., 2012, pp. 2852–2860.

[30] M. R. Avendi, A. Kheradvar and H. Jafarkhani, 'A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI,' *Medical Image Analysis*, vol. 30, pp. 108–119, 2016. DOI: 10.1016/j.media.2016.01.005.

[31] P. V. Tran, 'A fully convolutional neural network for cardiac segmentation in Short-Axis MRI,' *arXiv Preprint*, 2016, arXiv:1604.00494.

[32] O. Ronneberger, P. Fischer and T. Brox, 'U-Net: Convolutional networks for biomedical image segmentation,' in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, N. Navab, J. Hornegger, W. M. W. III and A. F. Frangi, Eds., Springer International Publishing, 2015, pp. 234–241.

[33] J. Long, E. Shelhamer and T. Darrell, 'Fully convolutional networks for semantic segmentation,' in *Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440. DOI: 10.1109/CVPR.2015.7298965.

[34] E. Shelhamer, J. Long and T. Darrell, 'Fully convolutional networks for semantic segmentation,' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2017, ISSN: 0162-8828. DOI: 10.1109/TPAMI.2016.2572683.

[35] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox and O. Ronneberger, '3D U-Net: Learning dense volumetric segmentation from sparse annotation,' in *Medical Image Com-*

puting and Computer-Assisted Intervention - MICCAI 2016 - 19th International Conference, Athens, Greece, S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. B. Ünal and W. Wells, Eds., Springer International Publishing, 2016, pp. 424–432. DOI: 10.1007/978-3-319-46723-8_49.

[36] F. Milletari, N. Navab and S. Ahmadi, 'V-Net: Fully convolutional neural networks for volumetric medical image segmentation,' in *Fourth International Conference on 3D Vision, 3DV 2016, Stanford, CA, USA*, IEEE Computer Society, 2016, pp. 565–571. DOI: 10.1109/3DV.2016.79.

[37] 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, no. 1, p. 180 513, 2019, ISSN: 0033-8419, 1527-1315. DOI: 10.1148/radiol.2018180513.

[38] 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,' in *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges - 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017, Quebec City, Canada, September 10-14, 2017, Revised Selected Papers*, M. Pop, M. Sermesant, P.-M. Jodoin, A. Lalande, X. Zhuang, G. Yang, A. A. Young and O. Bernard, Eds., Springer International Publishing, 2017, pp. 120–129. DOI: 10.1007/978-3-319-75541-0_13.

[39] Q. Xia, Y. Yao, Z. Hu and A. Hao, 'Automatic 3D atrial segmentation from GE-MRIs using volumetric fully convolutional networks,' in *Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges - 9th International Workshop, STACOM 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers*, M. Pop, M. Sermesant, J. Zhao, S. Li, K. McLeod, A. A. Young, K. S. Rhode and T. Mansi, Eds., Springer International Publishing, 2018, pp. 211–220. DOI: 10.1007/978-3-030-12029-0_23.

[40] R. P. K. Poudel, P. Lamata and G. Montana, 'Recurrent fully convolutional neural networks for multi-slice MRI cardiac segmentation,' in *1st International Workshops on Reconstruction and Analysis of Moving Body Organs, RAMBO 2016 and 1st International Workshops on Whole-Heart and Great Vessel Segmentation from 3D Cardiovascular MRI in Congenital Heart Disease, HVSMR 2016*, Data source: http://segchd.csail.mit.edu/(Accessed September 1, 2019)., 2016, pp. 83–94. DOI: 10.1007/978-3-319-52280-7_8.

[41] S. Hochreiter and J. Schmidhuber, 'Long short-term memory,' *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997, ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735.

[42] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, 'Learning phrase representations using RNN Encoder-Decoder for statistical machine translation,' in *Conference on Empirical Methods in Natural Language Processing*, ACL, 2014, pp. 1724–1734. DOI: 10.3115/v1/d14-1179.

[43] J. Schlemper, O. Oktay, W. Bai, D. C. Castro, J. Duan, C. Qin, J. V. Hajnal and D. Rueckert, 'Cardiac MR segmentation from undersampled k-space using deep latent representation learning,' in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018 - 21st International Conference, Granada, Spain, September 16-20, 2018*, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López and G. Fichtinger, Eds., Springer International Publishing, 2018, pp. 259–267. DOI: 10.1007/978-3-030-00928-1_30.

[44] O. Oktay, E. Ferrante, K. Kamnitsas, M. P. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. de Marvao, T. Dawes, D. P. O'Regan, B. Kainz, B. Glocker and D. Rueckert, 'Anatomically constrained neural networks (acnns): Application to cardiac image enhancement and segmentation,' *IEEE Transactions on Medical Imaging*, vol. 37, no. 2, pp. 384–395, 2018. DOI: 10.1109/TMI.2017.2743464.

[45] C. Biffi, O. Oktay, G. Tarroni, W. Bai, A. De Marvao, G. Doumou, M. Rajchl, R. Bedair, S. Prasad, S. Cook, D. O'Regan and D. Rueckert, 'Learning interpretable anatomical features through deep generative models: Application to cardiac remodeling,' in *Medical*

Image Computing and Computer Assisted Intervention - MICCAI 2018 - 21st International Conference, Granada, Spain, September 16-20, 2018, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López and G. Fichtinger, Eds., vol. 11071 LNCS, Springer International Publishing, 2018, pp. 464–471. DOI: 10.1007/978-3-030-00934-2_52.

[46] N. Painchaud, Y. Skandarani, T. Judge, O. Bernard, A. Lalande and P.-M. Jodoin, 'Cardiac MRI segmentation with strong anatomical guarantees,' in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2019 - 22nd International Conference, Shenzhen, China, October 13-17, 2019, Proceedings, Part II*, D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap and A. Khan, Eds., Springer International Publishing, 2019, pp. 632–640. DOI: 10.1007/978-3-030-32245-8_70.

[47] 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,' in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016 - 19th International Conference, Athens, Greece*, S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. B. Ünal and W. Wells, Eds., vol. 9902 LNCS, Springer International Publishing, 2016, pp. 246–254. DOI: 10.1007/978-3-319-46726-9_29.

[48] Q. Yue, X. Luo, Q. Ye, L. Xu and X. Zhuang, 'Cardiac segmentation from LGE MRI using deep neural network incorporating shape and spatial priors,' in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap and A. Khan, Eds., Cham: Springer International Publishing, 2019, pp. 559–567. DOI: 10.1007/978-3-030-32245-8_62.

[49] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, 'Generative adversarial nets,' in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence and K. Q. Weinberger, Eds., vol. 27, Curran Associates, Inc., 2014.

[50] P. Luc, C. Couprie, S. Chintala and J. Verbeek, 'Semantic segmentation using adversarial networks,' in *NIPS Workshop on Adversarial Training*, 2016, pp. 1–12.

[51] N. Savioli, M. S. Vieira, P. Lamata and G. Montana, 'A generative adversarial model for right ventricle segmentation,' *arXiv Preprint*, 2018, arXiv:1810.03969.

[52] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, 'Rethinking the inception architecture for computer vision,' in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, IEEE Computer Society, 2016, pp. 2818–2826. DOI: 10.1109/CVPR.2016.308.

[53] C. Szegedy, S. Ioffe, V. Vanhoucke and A. A. Alemi, 'Inception-v4, inception-resnet and the impact of residual connections on learning,' in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, 2017, pp. 4278–4284.

[54] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. U. Kaiser and I. Polosukhin, 'Attention is all you need,' in *Conference on Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, Eds., Curran Associates, Inc., 2017, pp. 5998–6008.

[55] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker and D. Rueckert, 'Attention U-Net: Learning where to look for the pancreas,' in *Medical Imaging with Deep Learning*, 2018, p. 1804.03999.

[56] K. He, X. Zhang, S. Ren and J. Sun, 'Deep residual learning for image recognition,' in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, IEEE Computer Society, 2016, pp. 770–778.

[57] F. Yu and V. Koltun, 'Multi-Scale context aggregation by dilated convolutions,' in *International Conference on Learning Representations*, 2016, pp. 1–13.

[58] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang and Z. Tu, 'Deeply-Supervised nets,' in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, G. Lebanon and S. V. N. Vishwanathan, Eds., ser. Proceedings of Machine Learning Research, vol. 38, San Diego, California, USA: PMLR, 2015, pp. 562–570.

[59] L.-C. Chen, G. Papandreou, F. Schroff and H. Adam, 'Rethinking atrous convolution for semantic image segmentation,' *arXiv Preprint*, 2017, arXiv:1706.05587.

[60] K. He, X. Zhang, S. Ren and J. Sun, 'Spatial pyramid pooling in deep convolutional networks for visual recognition,' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, ISSN: 0162-8828. DOI: 10.1109/TPAMI.2015.2389824.

[61] S. Jetley, N. A. Lord, N. Lee and P. H. S. Torr, 'Learn to pay attention,' in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

[62] J. Hu, L. Shen and G. Sun, 'Squeeze-and-Excitation networks,' in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, IEEE Computer Society, 2018, pp. 7132–7141. DOI: 10.1109/CVPR.2018.00745.

[63] G. Huang, Z. Liu, L. van der Maaten and K. Q. Weinberger, 'Densely connected convolutional networks,' in *Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2261–2269.

[64] H. Robinds and S. Monro, 'Adam: A method for stochastic optimization,' *A stochastic approximation method," Annals of Mathematical Statistics*, vol. 22, pp. 400–407, 1951.

[65] D. P. Kingma and J. Ba, 'Adam: A method for stochastic optimization,' in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.

[66] D. E. Rumelhart, G. E. Hinton and R. J. Williams, 'Learning representations by back-propagating errors,' *Nature*, vol. 323, no. 6088, pp. 533–536, 1986, ISSN: 0028-0836. DOI: 10.1038/323533a0.

[67] V. N. Vapnik, 'An overview of statistical learning theory,' *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 988–999, 1999.

[68] Y. Jang, Y. Hong, S. Ha, S. Kim and H.-J. Chang, 'Automatic segmentation of LV and RV in cardiac MRI,' in *International Workshop on Statistical Atlases and Computational*

Models of the Heart, M. Pop, M. Sermesant, P.-M. Jodoin, A. Lalande, X. Zhuang, G. Yang, A. A. Young and O. Bernard, Eds., Springer, 2017, pp. 161–169.

[69] C. F. Baumgartner, L. M. Koch, M. Pollefeys and E. Konukoglu, 'An exploration of 2D and 3D deep learning techniques for cardiac MR image segmentation,' in *International Workshop on Statistical Atlases and Computational Models of the Heart*, M. Pop, M. Sermesant, P.-M. Jodoin, A. Lalande, X. Zhuang, G. Yang, A. A. Young and O. Bernard, Eds., vol. 10663, Springer, 2017, pp. 1–8.

[70] X. Yang, C. Bian, L. Yu, D. Ni and P.-A. Heng, 'Class-Balanced deep neural network for automatic ventricular structure segmentation,' in *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges - 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017, Quebec City, Canada, September 10-14, 2017, Revised Selected Papers*, M. Pop, M. Sermesant, P.-M. Jodoin, A. Lalande, X. Zhuang, G. Yang, A. A. Young and O. Bernard, Eds., Springer International Publishing, 2017, pp. 152–160. DOI: 10.1007/978-3-319-75541-0_16.

[71] M. Khened, V. A. Kollerathu and G. Krishnamurthi, 'Fully convolutional multi-scale residual DenseNets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers,' *Medical Image Analysis*, vol. 51, pp. 21–45, 2019, ISSN: 1361-8415, 1361-8423. DOI: 10.1016/j.media.2018.10.004.

[72] R. J. Van Der Geest and J. H. Reiber, 'Quantification in cardiac MRI,' *Journal of Magnetic Resonance Imaging*, vol. 10, no. 5, pp. 602–608, 1999, ISSN: 1053-1807.

[73] J. Li, Z. Yu, Z. Gu, H. Liu and Y. Li, 'Dilated-Inception net: Multi-Scale feature aggregation for cardiac right ventricle segmentation,' *IEEE Transactions on Biomedical Engineering*, pp. 1–1, 2019, ISSN: 0018-9294, 1558-2531. DOI: 10.1109/TBME.2019.2906667.

[74] X.-Y. Zhou and G.-Z. Yang, 'Normalization in training U-Net for 2D biomedical semantic segmentation,' *IEEE Robotics and Automation Letters*, pp. 1–1, 2019. DOI: 10.1109/LRA.2019.2896518.

[75] J. Zhang, J. Du, H. Liu, X. Hou, Y. Zhao and M. Ding, 'LU-NET: An improved U-Net for ventricular segmentation,' *IEEE Access*, vol. 7, pp. 92 539–92 546, 2019, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2019.2925060.

[76] C. Cong and H. Zhang, 'Invert-U-Net DNN segmentation model for MRI cardiac left ventricle segmentation,' *The Journal of Engineering*, vol. 2018, no. 16, pp. 1463–1467, 2018, ISSN: 2051-3305. DOI: 10.1049/joe.2018.8302.

[77] A. S. Fahmy, H. El-Rewaidy, M. Nezafat, S. Nakamori and R. Nezafat, 'Automated analysis of cardiovascular magnetic resonance myocardial native T1 mapping images using fully convolutional neural networks,' *Journal of Cardiovascular Magnetic Resonance*, vol. 21, no. 1, pp. 1–12, 2019, ISSN: 1097-6647. DOI: 10.1186/s12968-018-0516-1.

[78] J. Sander, B. D. de Vos, J. M. Wolterink and I. Išgum, 'Towards increased trustworthiness of deep learning segmentation methods on cardiac MRI,' in *Medical Imaging 2019: Image Processing*, vol. 10949, International Society for Optics and Photonics, 2019, p. 1 094 919. DOI: 10.1117/12.2511699.

[79] M. Chen, L. Fang and H. Liu, 'FR-NET: Focal loss constrained deep residual networks for segmentation of cardiac MRI,' in *16th IEEE International Symposium on Biomedical Imaging, ISBI 2019, Venice, Italy, April 8-11, 2019*, IEEE, 2019, pp. 764–767. DOI: 10.1109/ISBI.2019.8759556.

[80] C. Zotti, Z. Luo, A. Lalande, O. Humbert and P.-M. Jodoin, 'GridNet with automatic shape prior registration for automatic MRI cardiac segmentation,' in *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges - 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017, Quebec City, Canada*, M. Pop, M. Sermesant, P.-M. Jodoin, A. Lalande, X. Zhuang, G. Yang, A. A. Young and O. Bernard, Eds., vol. 10663, Springer, 2017, pp. 73–81.

[81] C. Zotti, Z. Luo, A. Lalande and P.-M. Jodoin, 'Convolutional neural network with shape prior applied to cardiac MRI segmentation,' *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 3, pp. 1119–1128, 2019, ISSN: 2168-2208, 2168-2194. DOI: 10.1109/JBHI.2018.2865450.

[82] J. Patravali, S. Jain and S. Chilamkurthy, '2D-3D fully convolutional neural networks for cardiac mr segmentation,' in *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges - 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017, Quebec City, Canada, September 10-14, 2017, Revised Selected Papers*, M. Pop, M. Sermesant, P.-M. Jodoin, A. Lalande, X. Zhuang, G. Yang, A. A. Young and O. Bernard, Eds., 2017, pp. 130–139.

[83] X. Du, S. Yin, R. Tang, Y. Zhang and S. Li, 'Cardiac-DeepIED: Automatic pixel-level deep segmentation for cardiac bi-ventricle using improved end-to-end encoder-decoder network,' *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 7, no. March, pp. 1–10, 2019, ISSN: 2168-2372. DOI: 10.1109/JTEHM.2019.2900628.

[84] 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, no. 9, pp. 2137–2148, 2018, ISSN: 0278-0062, 1558-254X. DOI: 10.1109/TMI.2018.2820742.

[85] W. Yan, Y. Wang, Z. Li, R. J. van der Geest and Q. Tao, 'Left ventricle segmentation via Optical-Flow- net from Short-Axis cine MRI : Preserving the temporal coherence of cardiac motion,' in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018 - 21st International Conference, Granada, Spain*, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López and G. Fichtinger, Eds., vol. 11073 LNCS, Springer International Publishing, 2018, pp. 613–621, ISBN: 9783030009373. DOI: 10.1007/978-3-030-00937-3_70.

[86] N. Savioli, M. S. Vieira, P. Lamata and G. Montana, 'Automated segmentation on the entire cardiac cycle using a deep learning work - flow,' in *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 2018, pp. 153–158. DOI: 10.1109/SNAMS.2018.8554962.

[87] C. Qin, W. Bai, J. Schlemper, S. E. Petersen, S. K. Piechnik, S. Neubauer and D. Rueckert, 'Joint learning of motion estimation and segmentation for cardiac mr image sequences,' in *Medical Image Computing and Computer Assisted Intervention - MICCAI*

2018 - 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López and G. Fichtinger, Eds., Springer International Publishing, 2018, pp. 472–480.

[88] J. M. Wolterink, T. Leiner, M. A. Viergever and I. Išgum, 'Automatic segmentation and disease classification using cardiac cine mr images,' in *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges - 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017, Quebec City, Canada, September 10-14, 2017, Revised Selected Papers*, M. Pop, M. Sermesant, P.-M. Jodoin, A. Lalande, X. Zhuang, G. Yang, A. A. Young and O. Bernard, Eds., ser. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10663 LNCS, Springer International Publishing, 2017, pp. 101–110.

[89] J. R. Clough, I. Oksuz, N. Byrne, J. A. Schnabel and A. P. King, 'Explicit topological priors for deep-learning based image segmentation using persistent homology,' in *Information Processing in Medical Imaging - 26th International Conference,IPMI 2019, Hong Kong, China, June 2-7, 2019, Proceedings*, A. C. S. Chung, J. C. Gee, P. A. Yushkevich and S. Bao, Eds., vol. 11492 LNCS, 2019, pp. 16–28. DOI: 10.1007/978-3-030-20351-1_2.

[90] X. Chen, B. M. Williams, S. R. Vallabhaneni, G. Czanner, R. Williams and Y. Zheng, 'Learning active contour models for medical image segmentation,' in *Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 632–11 640.

[91] C. Qin, W. Bai, J. Schlemper, S. E. Petersen, S. K. Piechnik, S. Neubauer and D. Rueckert, 'Joint motion estimation and segmentation from undersampled cardiac MR image,' in *Machine Learning for Medical Image Reconstruction - First International Workshop, MLMIR 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings*, F. Knoll, A. K. Maier and D. Rueckert, Eds., Springer International Publishing, 2018, pp. 55–63. DOI: 10.1007/978-3-030-00129-2_7.

[92] S. Dangi, Z. Yaniv and C. A. Linte, 'Left ventricle segmentation and quantification from cardiac cine MR images via multi-task learning,' in *Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges - 9th International Workshop, STACOM 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers*, M. Pop, M. Sermesant, J. Zhao, S. Li, K. McLeod, A. A. Young, K. S. Rhode and T. Mansi, Eds., Springer, 2018, pp. 21–31. DOI: 10.1007/978-3-030-12029-0_3.

[93] L. Zhang, G. V. Karanikolas, M. Akçakaya and G. B. Giannakis, 'Fully automatic segmentation of the right ventricle via Multi-Task deep neural networks,' in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, IEEE, 2018, pp. 6677–6681. DOI: 10.1109/ICASSP.2018.8461556.

[94] A. Chartsias, T. Joyce, G. Papanastasiou, S. Semple, M. Williams, D. Newby, R. Dharmakumar and S. A. Tsaftaris, 'Factorised spatial representation learning: Application in semi-supervised myocardial segmentation,' in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018 - 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II*, vol. 11071 LNCS, Springer International Publishing, 2018, pp. 490–498.

[95] A. Chartsias, T. Joyce, G. Papanastasiou, S. Semple, M. Williams, D. E. Newby, R. Dharmakumar and S. A. Tsaftaris, 'Disentangled representation learning in cardiac image analysis,' *Medical Image Analysis*, vol. 58, p. 101 535, 2019, ISSN: 1361-8415, 1361-8423. DOI: 10.1016/j.media.2019.101535.

[96] Q. Huang, D. Yang, J. Yi, L. Axel and D. Metaxas, 'FR-Net: Joint reconstruction and segmentation in compressed sensing cardiac MRI,' in *Functional Imaging and Modeling of the Heart - 10th International Conference, FIMH 2019, Bordeaux, France*, Y. Coudière, V. Ozenne, E. J. Vigmond and N. Zemzemi, Eds., Springer International Publishing, 2019, pp. 352–360. DOI: 10.1007/978-3-030-21949-9_38.

[97] D. M. Vigneault, W. Xie, C. Y. Ho, D. A. Bluemke and J. A. Noble, 'Ω-Net (Omega-Net): Fully automatic, multi-view cardiac MR detection, orientation, and segmentation with deep neural networks,' *Medical Image Analysis*, vol. 48, pp. 95–106, 2018, ISSN: 1361-8415, 1361-8423. DOI: 10.1016/j.media.2018.05.008.

[98] C. Li, Q. Tong, X. Liao, W. Si, S. Chen, Q. Wang and Z. Yuan, 'APCP-NET: Aggregated parallel Cross-Scale pyramid network for CMR segmentation,' in *16th IEEE International Symposium on Biomedical Imaging, ISBI 2019, Venice, Italy, April 8-11, 2019*, 2019, pp. 784–788. DOI: 10.1109/ISBI.2019.8759147.

[99] L. K. Tan, Y. M. Liew, E. Lim and R. A. McLaughlin, 'Convolutional neural network regression for short-axis left ventricle segmentation in cardiac cine mr sequences,' *Medical Image Analysis*, vol. 39, pp. 78–86, 2017. DOI: 10.1016/j.media.2017.04.002.

[100] F. Liao, X. Chen, X. Hu and S. Song, 'Estimation of the volume of the left ventricle from MRI images using deep neural networks,' *IEEE Transactions on Cybernetics*, vol. 49, no. 2, pp. 495–504, 2019, ISSN: 2168-2275, 2168-2267. DOI: 10.1109/TCYB.2017.2778799.

[101] T. A. Ngo, Z. Lu and G. Carneiro, 'Combining deep learning and level set for the automated segmentation of the left ventricle of the heart from cardiac cine magnetic resonance,' *Medical Image Analysis*, vol. 35, pp. 159–171, 2017. DOI: 10.1016/j.media.2016.05.009.

[102] J. Duan, J. Schlemper, W. Bai, T. J. W. Dawes, G. Bello, G. Doumou, A. De Marvao, D. P. O'Regan and D. Rueckert, 'Deep nested level sets: Fully automated segmentation of cardiac MR images in patients with pulmonary hypertension,' in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018 - 21st International Conference, Granada, Spain*, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López and G. Fichtinger, Eds., Springer International Publishing, 2018, pp. 595–603.

[103] M. R. Avendi, A. Kheradvar and H. Jafarkhani, 'Automatic segmentation of the right ventricle from cardiac MRI using a learning-based approach,' *Magnetic resonance in medicine: official journal of the Society of Magnetic Resonance in Medicine / Society of*

Magnetic Resonance in Medicine, vol. 78, no. 6, pp. 2439–2448, 2017, ISSN: 0740-3194, 1522-2594. DOI: 10.1002/mrm.26631.

[104] D. O. Medley, C. Santiago and J. C. Nascimento, 'Segmenting the left ventricle in cardiac in cardiac MRI: From handcrafted to deep region based descriptors,' in *16th IEEE International Symposium on Biomedical Imaging, ISBI 2019, Venice, Italy, April 8-11, 2019*, IEEE, 2019, pp. 644–648. DOI: 10.1109/ISBI.2019.8759179.

[105] H. Yang, J. Sun, H. Li, L. Wang and Z. Xu, 'Deep fusion net for multi-atlas segmentation: Application to cardiac MR images,' in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016 - 19th International Conference, Athens, Greece*, S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. B. Ünal and W. Wells, Eds., Springer International Publishing, 2016, pp. 521–528. DOI: 10.1007/978-3-319-46723-8_60.

[106] M.-M. Rohé, M. Sermesant and X. Pennec, 'Automatic Multi-Atlas segmentation of myocardium with SVF-Net,' in *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges - 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017, Quebec City, Canada, September 10-14, 2017, Revised Selected Papers*, M. Pop, M. Sermesant, P.-M. Jodoin, A. Lalande, X. Zhuang, G. Yang, A. A. Young and O. Bernard, Eds., vol. 10663, Springer International Publishing, 2017, pp. 170–177. DOI: 10.1007/978-3-319-75541-0_18.

[107] X. Lu, X. Chen, W. Li and Y. Qiao, 'Graph cut segmentation of the right ventricle in cardiac MRI using multi-scale feature learning,' in *Proceedings of the 3rd International Conference on Cryptography, Security and Privacy,ICCSP 2019, Kuala Lumpur, Malaysia,*, Y. Wang and C.-C. Chang, Eds., ACM, 2019, pp. 231–235, ISBN: 9781450366182. DOI: 10.1145/3309074.3309117.

[108] G. Tziritas and E. Grinias, 'Fast fully-automatic localization of left ventricle and myocardium in MRI using MRF model optimization, substructures tracking and b-spline smoothing,' in *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges - 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017, Quebec City, Canada, September 10-14, 2017, Revised Selected Papers*,

M. Pop, M. Sermesant, P.-M. Jodoin, A. Lalande, X. Zhuang, G. Yang, A. A. Young and O. Bernard, Eds., Springer International Publishing, 2017, pp. 91–100.

[109] I. J. Goodfellow, J. Shlens and C. Szegedy, 'Explaining and harnessing adversarial examples,' in *International Conference on Learning Representations*, 2015.

[110] L. Engstrom, B. Tran, D. Tsipras, L. Schmidt and A. Madry, 'Exploring the landscape of spatial robustness,' in *ICML*, K. Chaudhuri and R. Salakhutdinov, Eds., ser. Proceedings of Machine Learning Research, vol. 97, Long Beach, California, USA: PMLR, 2019, pp. 1802–1811.

[111] A. Azulay and Y. Weiss, 'Why do deep convolutional networks generalize so poorly to small image transformations?' *Journal of Machine Learning Research: JMLR*, vol. 20, no. 184, pp. 1–25, 2019.

[112] T. B. Brown, D. Mané, A. Roy, M. Abadi and J. Gilmer, 'Adversarial patch,' *arXiv Preprint*, 2017.

[113] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam and I. S. Kohane, 'Adversarial attacks on medical machine learning,' *Science*, vol. 363, no. 6433, pp. 1287–1289, 2019, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aaw4399.

[114] X. Ma, Y. Niu, L. Gu, Y. Wang, Y. Zhao, J. Bailey and F. Lu, 'Understanding adversarial attacks on deep learning based medical image analysis systems,' 2019.

[115] D. Heaven, 'Why deep-learning AIs are so easy to fool,' *Nature*, vol. 574, no. 7777, pp. 163–166, 2019.

[116] D. Bau, B. Zhou, A. Khosla, A. Oliva and A. Torralba, 'Network dissection: Quantifying interpretability of deep visual representations,' in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, IEEE Computer Society, 2017, pp. 3319–3327. DOI: 10.1109/CVPR.2017.354.

[117] S. M. Lundberg and S.-I. Lee, 'A unified approach to interpreting model predictions,' in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, I.

Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan and R. Garnett, Eds., 2017, pp. 4765–4774.

[118] P. D. Grünwald, *The Minimum Description Length Principle*, ser. MIT Press Books 0262072815. The MIT Press, 2007, vol. 1, ISBN: 0x399ef9e8.

[119] R. J. Solomonoff, 'A formal theory of inductive inference. part I,' *Information and Control*, vol. 7, no. 1, pp. 1–22, 1964, ISSN: 0019-9958. DOI: 10.1016/S0019-9958(64)90223-2.

[120] Wikipedia contributors, *Kolmogorov complexity*, https://en.wikipedia.org/w/index.php?title=Kolmogorov_complexity&oldid=1022018769, Accessed: 2021-5-27, 2021.

[121] V. Cherkassky, 'The nature of statistical learning theory,' *IEEE Transactions on Neural Networks*, vol. 8, no. 6, p. 1564, 1997. DOI: 10.1109/TNN.1997.641482.

[122] Wikipedia contributors, *Rademacher complexity*, https://en.wikipedia.org/w/index.php?title=Rademacher_complexity&oldid=1014690670, Accessed: 2021-5-27, 2021.

[123] K. Kawaguchi, L. P. Kaelbling and Y. Bengio, 'Generalization in deep learning,' *Mathematics of Deep Learning*, 2017, arXiv:1710.05468.

[124] C. Zhang, S. Bengio, M. Hardt, B. Recht and O. Vinyals, 'Understanding deep learning requires rethinking generalization,' in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, OpenReview.net, 2017.

[125] Zhang, Chiyuan and Bengio, Samy and Hardt, Moritz and Recht, Benjamin and Vinyals, Oriol, 'Understanding deep learning (still) requires rethinking generalization,' *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.

[126] M. Belkin, D. Hsu, S. Ma and S. Mandal, 'Reconciling modern machine-learning practice and the classical bias–variance trade-off,' *Proceedings of the National Academy of Sciences*, vol. 116, no. 32, pp. 15 849–15 854, 2019, ISSN: 0027-8424. DOI: 10.1073/pnas.1903070116.

[127] L. Weng, 'Are deep neural networks dramatically overfitted?' *lilianweng.github.io/lil-log*, 2019.

[128] C. Li, H. Farkhoor, R. Liu and J. Yosinski, 'Measuring the intrinsic dimension of objective landscapes,' in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, OpenReview.net, 2018.

[129] J. Frankle and M. Carbin, 'The lottery ticket hypothesis: Finding sparse, trainable neural networks,' in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, OpenReview.net, 2019.

[130] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, 'Dropout: A simple way to prevent neural networks from overfitting,' *Journal of Machine Learning Research: JMLR*, vol. 15, pp. 1929–1958, 2014, ISSN: 1532-4435.

[131] J. Tompson, R. Goroshin, A. Jain, Y. LeCun and C. Bregler, 'Efficient object localization using convolutional networks,' in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, IEEE Computer Society, 2015, pp. 648–656. DOI: 10.1109/CVPR.2015.7298664.

[132] R. Volpi, H. Namkoong, O. Sener, J. C. Duchi, V. Murino and S. Savarese, 'Generalizing to unseen domains via adversarial data augmentation,' in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett, Eds., vol. 31, Curran Associates, Inc., 2018.

[133] C. Shorten and T. M. Khoshgoftaar, 'A survey on image data augmentation for deep learning,' *Journal of Big Data*, vol. 6, no. 1, p. 60, 2019, ISSN: 2196-1115. DOI: 10.1186/s40537-019-0197-0.

[134] K. Kamnitsas, W. Bai, E. Ferrante, S. Mcdonagh and M. Sinclair, 'Ensembles of multiple models and architectures for robust brain tumour segmentation,' in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers*, 2017, pp. 450–462.

[135] L. Breiman, 'Bagging predictors,' *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[136] S. Chen, K. Ma and Y. Zheng, 'Med3d: Transfer learning for 3D medical image analysis,' *arXiv Preprint*, 2019, arXiv:1904.00625.

[137] NHS, *Atrial fibrillation*, https://www.nhs.uk/conditions/atrial-fibrillation/, Last visited on 2018-06-25, 2018.

[138] H. Calkins, K. H. Kuck, R. Cappato, J. Brugada, A. J. Camm, S.-A. Chen, H. J. G. Crijns, R. J. Damiano Jr, D. W. Davies, J. DiMarco, J. Edgerton, K. Ellenbogen, M. D. Ezekowitz, D. E. Haines, M. Haissaguerre, G. Hindricks, Y. Iesaka, W. Jackman, J. Jalife, P. Jais, J. Kalman, D. Keane, Y.-H. Kim, P. Kirchhof, G. Klein, H. Kottkamp, K. Kumagai, B. D. Lindsay, M. Mansour, F. E. Marchlinski, P. M. McCarthy, J. L. Mont, F. Morady, K. Nademanee, H. Nakagawa, A. Natale, S. Nattel, D. L. Packer, C. Pappone, E. Prystowsky, A. Raviele, V. Reddy, J. N. Ruskin, R. J. Shemin, H.-M. Tsao, D. Wilber and Heart Rhythm Society Task Force on Catheter and Surgical Ablation of Atrial Fibrillation, '2012 hrs/ehra/ecas expert consensus statement on catheter and surgical ablation of atrial fibrillation: Recommendations for patient selection, procedural techniques, patient management and follow-up, definitions, endpoints, and research trial design,' *Heart rhythm: the official journal of the Heart Rhythm Society*, vol. 9, no. 4, 632–696.e21, 2012.

[139] C. Tobon-Gomez, A. J. Geers, J. Peters, J. Weese, K. Pinto, R. Karim, M. Ammar, A. Daoudi, J. Margeta, Z. Sandoval, B. Stender, Yefeng Zheng, M. A. Zuluaga, J. Betancur, N. Ayache, M. Amine Chikh, J.-L. Dillenseger, B. M. Kelm, S. Mahmoudi, S. Ourselin, A. Schlaefer, T. Schaeffter, R. Razavi and K. S. Rhode, 'Benchmark for algorithms segmenting the left atrium from 3D CT and MRI datasets,' *IEEE Transactions on Medical Imaging*, vol. 34, no. 7, pp. 1460–1473, 2015.

[140] M. Depa, M. R. Sabuncu, G. Holmvang, R. Nezafat, E. J. Schmidt and P. Golland, 'Robust Atlas-Based segmentation of highly variable anatomy: Left atrium segmenta-

tion,' in *Statistical atlases and computational models of the heart. STACOM (Workshop)*, vol. 6364, 2010, pp. 85–94.

[141] 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, pp. 60–88, 2017. DOI: 10.1016/j.media.2021.102062.

[142] L. Malcolme-Lawes, C. Juli, R. Karim, W. Bai, R. Quest, P. B. Lim, S. Jamil-Copley, P. Kojodjojo, B. Ariff, D. W. Davies, D. Rueckert, D. Francis, R. Hunter, D. Jones, R. Boubertakh, S. Petersen, R. Schilling, P. Kanagaratnam and N. Peters, 'Automated analysis of atrial late gadolinium enhancement imaging that correlates with endocardial voltage and clinical outcomes: A 2-center study,' *Heart rhythm : the official journal of the Heart Rhythm Society*, vol. 10, 2013.

[143] K. He, X. Zhang, S. Ren and J. Sun, 'Spatial pyramid pooling in deep convolutional networks for visual recognition,' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015. DOI: 10.1109/TPAMI.2015.2389824.

[144] R. Karim, R. Mohiaddin and D. Rueckert, 'Left atrium segmentation for atrial fibrillation ablation,' in *Medical Imaging 2008: Visualization, Image-Guided Procedures, and Modeling, San Diego, California, United States, 16-21 February 2008*, SPIE, 2008, 69182U. DOI: 10.1117/12.771023.

[145] Q. Tao, R. Shahzad, E. G. Ipek, F. F. Berendsen, S. Nazarian and R. J. van der Geest, 'Fully automatic segmentation of left atrium and pulmonary veins in late gadolinium-enhanced MRI: Towards objective atrial scar assessment,' *Journal of Magnetic Resonance Imaging*, vol. 44, no. 2, pp. 346–354, 2016, ISSN: 1053-1807, 1522-2586. DOI: 10.1002/jmri.25148.

[146] X. Zhuang, K. S. Rhode, R. S. Razavi, D. J. Hawkes and S. Ourselin, 'A registration-based propagation framework for automatic whole heart segmentation of cardiac MRI,' *IEEE Transactions on Medical Imaging*, vol. 29, no. 9, pp. 1612–1625, 2010, ISSN: 0278-0062, 1558-254X. DOI: 10.1109/TMI.2010.2047112.

[147] S. Ioffe and C. Szegedy, 'Batch normalization: Accelerating deep network training by reducing internal covariate shift,' in *International Conference on Machine Learning*, JMLR.org, 2015, pp. 448–456.

[148] A. Krizhevsky, I. Sutskever and G. E. Hinton, 'Alexnet,' *Advances In Neural Information Processing Systems*, pp. 1–9, 2012, ISSN: 10495258. DOI: http://dx.doi.org/10.1016/j.protcy.2014.09.007.

[149] Y. Bengio, J. Louradour, R. Collobert and J. Weston, 'Curriculum learning,' in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09, Montreal, Quebec, Canada: ACM, 2009, pp. 41–48, ISBN: 9781605585161. DOI: 10.1145/1553374.1553380.

[150] J. Lieman-Sifry, M. Le, F. Lau, S. Sall and D. Golden, 'Fastventricle: Cardiac segmentation with ENet,' in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10263 LNCS, 2017, pp. 127–138, ISBN: 9783319594477. DOI: 10.1007/978-3-319-59448-4_13.

[151] K. Somasundaram and P. Kalavathi, 'Medical image contrast enhancement based on gamma correction,' *Knowledge Management & E-Learning*, vol. 3, no. 1, pp. 15–18, 2011.

[152] K. J. Zuiderveld, 'Contrast limited adaptive histogram equalization,' in *Graphics Gems*, P. S. Heckbert, Ed., Elsevier, 1994, pp. 474–485. DOI: 10.1016/b978-0-12-336156-1.50061-6.

[153] S. M. Pizer, R. E. Johnston, J. P. Ericksen, B. C. Yankaskas and K. E. Muller, 'Contrast-limited adaptive histogram equalization: Speed and effectiveness,' in *Visualization in Biomedical Computing, 1990., Proceedings of the First Conference on*, IEEE, 1990, pp. 337–345.

[154] P. T. Selvy, V. Palanisamy and M. S. Radhai, 'A proficient clustering technique to detect csf level in mri brain images using pso algorithm,' *WSEAS Transactions on Computers*, vol. 7, pp. 298–308, 2013.

[155] S. Ghose, J. Mitra, A. Oliver, R. Marti, X. Lladó, J. Freixenet, J. C. Vilanova, D. Sidibé and F. Meriaudeau, 'A random forest based classification approach to prostate segmentation in MRI,' *MICCAI Grand Challenge: Prostate MR Image Segmentation*, vol. 2012, 2012.

[156] G. Tarroni, O. Oktay, W. Bai, A. Schuh, H. Suzuki, J. Passerat-Palmbach, A. de Marvao, D. P. O'Regan, S. Cook, B. Glocker, P. M. Matthews and D. Rueckert, 'Learning-Based quality control for cardiac MR images,' *IEEE Transactions on Medical Imaging*, vol. 38, no. 5, pp. 1127–1138, 2019, ISSN: 0278-0062, 1558-254X. DOI: 10.1109/TMI.2018.2878509.

[157] J. Duan, G. Bello, J. Schlemper, W. Bai, T. J. W. 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-constrained multi-task deep learning approach,' *IEEE Transactions on Medical Imaging*, vol. PP, no. c, p. 1, 2019, ISSN: 0278-0062, 1558-254X. DOI: 10.1109/TMI.2019.2894322.

[158] S. Kohl, B. Romera-Paredes, C. Meyer, J. D. Fauw, J. R. Ledsam, K. H. Maier-Hein, S. M. A. Eslami, D. J. Rezende and O. Ronneberger, 'A probabilistic U-Net for segmentation of ambiguous images,' in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett, Eds., 2018, pp. 6965–6975.

[159] R. Caruana, 'Multitask learning,' *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1997.

[160] K. Kushibar, S. Valverde, S. González-Villà, J. Bernal, M. Cabezas, A. Oliver and X. Lladó, 'Automated sub-cortical brain structure segmentation combining spatial and deep convolutional features,' *Medical image analysis*, vol. 48, pp. 177–186, 2018.

[161] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu and X. Ding, 'Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation,' *Medical Image Analysis*, vol. 63, p. 101 693, 2020.

[162] A. Zhao, G. Balakrishnan, F. Durand, J. V. Guttag and A. V. Dalca, 'Data augmentation using learned transformations for one-shot medical image segmentation,' in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA: IEEE, 2019.

[163] J. Liu, C. Shen, T. Liu, N. Aguilera and J. Tam, 'Active appearance model induced generative adversarial network for controlled data augmentation,' in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, Springer International Publishing, 2019, pp. 201–208.

[164] K. Chaitanya, N. Karani, C. Baumgartner, O. Donati, A. Becker and E. Konukoglu, 'Semi-Supervised and Task-Driven data augmentation,' in *Information Processing in Medical Imaging - 26th International Conference, IPMI 2019, Hong Kong, China, June 2-7, 2019, Proceedings*, A. C. S. Chung, J. C. Gee, P. A. Yushkevich and S. Bao, Eds., 2019, pp. 29–41. DOI: 10.1007/978-3-030-20351-1_3.

[165] Y. Xing, Z. Ge, R. Zeng, D. Mahapatra, J. Seah, M. Law and T. Drummond, 'Adversarial pulmonary pathology translation for pairwise chest x-ray data augmentation,' in *Medical Image Computing and Computer Assisted Intervention*, 2019.

[166] N. Lei, D. An, Y. Guo, K. Su, S. Liu, Z. Luo, S.-T. Yau and X. Gu, 'A geometric understanding of deep learning,' *Proceedings of the Estonian Academy of Sciences: Engineering*, vol. 6, no. 3, pp. 361–374, 2020.

[167] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich and J. C. Gee, 'N4ITK: Improved N3 bias correction,' *IEEE Transactions on Medical Imaging*, vol. 29, no. 6, pp. 1310–1320, 2010.

[168] N. Khalili, N. Lessmann, E. Turk, N. Claessens, R. d. Heus, T. Kolk, M. A. Viergever, M. J. N. L. Benders and I. Išgum, 'Automatic brain tissue segmentation in fetal MRI using convolutional neural networks,' *Journal of Magnetic Resonance Imaging*, vol. 64, pp. 77–89, 2019.

[169] M. Paschali, S. Conjeti, F. Navarro and N. Navab, 'Generalizability vs. robustness: Investigating medical imaging networks using adversarial examples,' in *Medical Image*

Computing and Computer Assisted Intervention – MICCAI 2018, Springer International Publishing, 2018, pp. 493–501.

[170] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara and D. Rueckert, 'Intelligent image synthesis to attack a segmentation CNN using adversarial learning,' in *Simulation and Synthesis in Medical Imaging - 4th International Workshop, SASHIMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings*, N. Burgos, A. Gooya and D. Svoboda, Eds., ser. Lecture Notes in Computer Science, vol. 11827, Springer, 2019, pp. 90–99.

[171] A. Madry, A. Makelov, L. Schmidt, D. Tsipras and A. Vladu, 'Towards deep learning models resistant to adversarial attacks,' in *International Conference on Learning Representations*, 2017.

[172] N. Carlini and D. A. Wagner, 'Towards evaluating the robustness of neural networks,' in *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, 2017, pp. 39–57. DOI: 10.1109/SP.2017.49.

[173] F. Tramèr and D. Boneh, 'Adversarial training and robustness for multiple perturbations,' in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox and R. Garnett, Eds., 2019, pp. 5858–5868.

[174] T. Miyato, S.-I. Maeda, M. Koyama and S. Ishii, 'Virtual adversarial training: A regularization method for supervised and semi-supervised learning,' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, ISSN: 0162-8828.

[175] C. Kanbak, S.-M. Moosavi-Dezfooli and P. Frossard, 'Geometric robustness of deep networks: Analysis and improvement,' in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 4441–4449. DOI: 10.1109/CVPR.2018.00467.

[176] X. Zeng, C. Liu, Y.-S. Wang, W. Qiu, L. Xie, Y.-W. Tai, C.-K. Tang and A. L. Yuille, 'Adversarial attacks beyond the image space,' in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA: IEEE, 2019.

[177] R. Alaifari, G. S. Alberti and T. Gauksson, 'Adef: An iterative algorithm to construct adversarial deformations,' in *International Conference on Learning Representations*, 2019.

[178] J. Sled, A. Zijdenbos and A. Evans, 'A nonparametric method for automatic correction of intensity nonuniformity in MRI data,' *IEEE Transactions on Medical Imaging*, vol. 17, no. 1, pp. 87–97, 1998. DOI: 10.1109/42.668698.

[179] J. Gallier, *Curves and surfaces in geometric modeling: theory and algorithms*. Morgan Kaufmann, 2000.

[180] R. Sandkühler, C. Jud, S. Andermatt and P. C. Cattin, 'AirLab: Autograd image registration laboratory,' *arXiv Preprint*, 2018, arXiv:1806.09907.

[181] H. Zhang, M. Cisse, Y. N. Dauphin and D. Lopez-Paz, 'Mixup: Beyond empirical risk minimization,' in *International Conference on Learning Representations*, 2018, pp. 1–13.

[182] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann and W. Brendel, 'ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness,' in *International Conference on Learning Representations*, 2018, pp. 1–20.

[183] A. A. Taha and A. Hanbury, 'Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool,' *BMC Medical Imaging*, vol. 15, p. 29, 2015.

[184] F. Pérez-García, R. Sparks and S. Ourselin, 'Torchio: A python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning,' *Computer Methods and Programs in Biomedicine*, p. 106 236, 2021, ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2021.106236.

[185] X. Zhuang, 'Multivariate mixture model for cardiac segmentation from multi-sequence MRI,' in *Medical Image Computing and Computer Assisted Intervention*, 2016, pp. 581–588.

[186] Y. Lu, G. Wright and P. E. Radau, 'Automatic myocardium segmentation of LGE MRI by deformable models with prior shape data,' *Journal of Cardiovascular Magnetic Resonance*, vol. 15, no. 1, pp. 1–2, 2013.

[187] Q. Tao, S. R. D. Piers, H. J. Lamb and R. J. van der Geest, 'Automated left ventricle segmentation in late gadolinium-enhanced MRI for objective myocardial scar assessment,' *Journal of Magnetic Resonance Imaging*, vol. 42, no. 2, pp. 390–399, 2015.

[188] X. Zhuang, 'Multivariate mixture model for myocardium segmentation combining multi-source images,' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[189] X. Huang, M.-Y. Liu, S. Belongie and J. Kautz, 'Multimodal unsupervised Image-to-Image translation,' in *Computer Vision – ECCV 2018*, Springer International Publishing, 2018, pp. 179–196.

[190] C. Qin, B. Shi, R. Liao, T. Mansi, D. Rueckert and A. Kamen, 'Unsupervised deformable registration for multi-modal images via disentangled representations,' in *Information Processing in Medical Imaging*, Springer, Cham, 2019, pp. 249–261. DOI: 10.1007/978-3-030-20351-1_19.

[191] Z. Tu and X. Bai, 'Auto-context and its application to high-level vision tasks and 3D brain image segmentation,' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 10, pp. 1744–1757, 2010, ISSN: 01628828. DOI: 10.1109/TPAMI.2009.186.

[192] I. Sobel and G. Feldman, 'A 3x3 isotropic gradient operator for image processing,' *Pattern Classification and Scene Analysis*, pp. 271–272, 1973.

[193] P. Krähenbühl and V. Koltun, 'Efficient inference in fully connected CRFs with gaussian edge potentials,' in *Advances in Neural Information Processing Systems*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira and K. Q. Weinberger, Eds., Curran Associates, Inc., 2011, pp. 109–117.

[194] X. Huang and S. J. Belongie, 'Arbitrary style transfer in real-time with adaptive instance normalization,' in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2017, pp. 1510–1519. DOI: 10.1109/ICCV.2017.167.

[195] H.-Y. Tseng, L. Jiang, C. Liu, M.-H. Yang and W. Yang, 'Regularing generative adversarial networks under limited data,' in *CVPR*, 2021.

[196] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen and T. Aila, 'Training generative adversarial networks with limited data,' *arXiv preprint arXiv:2006.06676*, 2020.

[197] P. Rajiah and M. A. Bolen, 'Cardiovascular mr imaging at 3 t: Opportunities, challenges, and solutions,' *RadioGraphics*, vol. 34, no. 6, pp. 1612–1635, 2014, PMID: 25310420. DOI: 10.1148/rg.346140048.

[198] K. Alfudhili, P. G. Masci, J. Delacoste, J.-B. Ledoux, G. Berchier, V. Dunet, S. D. Qanadli, J. Schwitter and C. Beigelman-Aubry, 'Current artefacts in cardiac and chest magnetic resonance imaging: Tips and tricks,' *British Journal of Radiology*, vol. 89, no. 1062, p. 20 150 987, 2016. DOI: 10.1259/bjr.20150987.

[199] M. Gutberlet, R. Noeske, K. Schwinge, P. Freyhardt, R. Felix and T. Niendorf, 'Comprehensive cardiac magnetic resonance imaging at 3.0 tesla: Feasibility and implications for clinical applications,' *Investigative Radiology*, vol. 41, no. 2, pp. 154–167, 2006.

[200] P. Medrano-Gracia, B. R. Cowan, B. Ambale-Venkatesh, D. A. Bluemke, J. Eng, J. P. Finn, C. G. Fonseca, J. A. Lima, A. Suinesiaputra and A. A. Young, 'Left ventricular shape variation in asymptomatic populations: the multi-ethnic study of atherosclerosis,' *Journal of Cardiovascular Magnetic Resonance*, vol. 16, no. 1, p. 56, 2014.

[201] C. Petitjean and J.-N. Dacher, 'A review of segmentation methods in short axis cardiac MR images,' *Medical Image Analysis*, vol. 15, no. 2, pp. 169–184, 2011. DOI: 10.1016/j.media.2010.12.004.

[202] M. Wang and W. Deng, 'Deep visual domain adaptation: A survey,' *Neurocomputing*, vol. 312, pp. 135–153, 2018.

[203] B. Sun and K. Saenko, 'Deep CORAL: Correlation alignment for deep domain adaptation,' in *Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III*, 2016, pp. 443–450. DOI: 10.1007/978-3-319-49409-8_35.

[204] M. Long, Y. Cao, J. Wang and M. I. Jordan, 'Learning transferable features with deep adaptation networks,' in *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, 2015, pp. 97–105.

[205] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros and T. Darrell, 'Cycada: Cycle-consistent adversarial domain adaptation,' in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, 2018, pp. 1994–2003.

[206] A. Suinesiaputra, B. R. Cowan, A. O. Al-Agamy, M. A. Elattar, N. Ayache, A. S. Fahmy, A. M. Khalifa, P. Medrano-Gracia, M.-P. Jolly, A. H. Kadish, D. C. Lee, J. Margeta, S. K. Warfield and A. A. Young, 'A collaborative resource to build consensus for automated left ventricular segmentation of cardiac MR images,' *Medical Image Analysis*, vol. 18, no. 1, pp. 50–62, 2014, Data source: http://www.cardiacatlas.org/challenges/lv-segmentation-challenge/ (Accessed September 1, 2019)., ISSN: 1361-8415, 1361-8423. DOI: 10.1016/j.media.2013.09.001.

[207] S. E. Petersen, P. M. Matthews, J. M. Francis, M. D. Robson, F. Zemrak, R. Boubertakh, A. A. Young, S. Hudson, P. Weale, S. Garratt, R. Collins, S. Piechnik and S. Neubauer, 'UK Biobank's cardiovascular magnetic resonance protocol,' *Journal of Cardiovascular Magnetic Resonance*, vol. 18, no. 1, p. 8, 2016.

[208] T. A. Musa, T. A. Treibel, V. S. Vassiliou, G. Captur, A. Singh, C. Chin, L. E. Dobson, S. Pica, M. Loudon, T. Malley, M. Rigolli, J. R. J. Foley, P. Bijsterveld, G. R. Law, M. R. Dweck, S. G. Myerson, G. P. McCann, S. K. Prasad, J. C. Moon and J. P. Greenwood, 'Myocardial scar and mortality in severe aortic stenosis,' *Circulation*, vol. 138, no. 18, pp. 1935–1947, 2018.

[209] D. Dumitrescu and C.-A. Boiangiu, 'A study of image upsampling and downsampling filters,' *Computers*, vol. 8, no. 2, p. 30, 2019.

[210] I. Goodfellow, *Deep learning*, ser. Adaptive computation and machine learning. Cambridge, Massachusetts; London, England: The MIT Press, 2016, ISBN: 9780262035613.

[211] E. J. Bekkers, M. W. Lafarge, M. Veta, K. A. Eppenhof, J. P. Pluim and R. Duits, 'Roto-translation covariant convolutional networks for medical image analysis,' in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018 - 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, Springer, 2018, pp. 440–448.

[212] S. Dieleman, K. Willett and J. Dambre, 'Rotation-invariant convolutional neural networks for galaxy morphology prediction,' *Monthly Notices of the Royal Astronomical Society*, vol. 450, no. 2, pp. 1441–1459, 2015.

[213] Y. Li, N. Wang, J. Shi, X. Hou and J. Liu, 'Adaptive Batch Normalization for practical domain adaptation,' *Pattern Recognition*, vol. 80, pp. 109–117, 2018.

[214] A. Fry, T. J. Littlejohns, C. Sudlow, N. Doherty, L. Adamska, T. Sprosen, R. Collins and N. E. Allen, 'Comparison of sociodemographic and health-related characteristics of uk biobank participants with those of the general population,' *American Journal of Epidemiology*, vol. 186, no. 9, pp. 1026–1034, 2017.

[215] A. Suinesiaputra, B. R. Cowan, A. O. Al-Agamy, M. A. Elattar, N. Ayache, A. S. Fahmy, A. M. Khalifa, P. Medrano-Gracia, M.-P. Jolly, A. H. Kadish, D. C. Lee, J. Margeta, S. K. Warfield and A. A. Young, 'A collaborative resource to build consensus for automated left ventricular segmentation of cardiac MR images,' *Medical Image Anal.*, vol. 18, no. 1, pp. 50–62, 2014. DOI: 10.1016/j.media.2013.09.001.

[216] J. Krebs, H. e Delingette, B. Mailhé, N. Ayache and T. Mansi, 'Learning a probabilistic model for diffeomorphic registration,' *IEEE Transactions on Medical Imaging*, vol. 38, no. 9, pp. 2165–2176, 2019, ISSN: 0278-0062. DOI: 10.1109/TMI.2019.2897112.

[217] R. Volpi, H. Namkoong, O. Sener, J. C. Duchi, V. Murino and S. Savarese, 'Generalizing to unseen domains via adversarial data augmentation,' in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, 2018, pp. 5339–5349.

[218] H. Wei, W. Xue and D. Ni, 'Left ventricle segmentation and quantification with attention-enhanced segmentation and shape correction,' in *Proceedings of the Third International Symposium on Image Computing and Digital Medicine, ISICDM 2019, Xi'an, China, August 24-26, 2019*, 2019, pp. 226–230.

[219] R. Robinson, V. V. Valindria, W. Bai, O. Oktay, B. Kainz, H. Suzuki, M. M. Sanghvi, N. Aung, J. M. Paiva, F. Zemrak, K. Fung, E. Lukaschuk, A. M. Lee, V. Carapella, Y. J. Kim, S. K. Piechnik, S. Neubauer, S. E. Petersen, C. Page, P. M. Matthews, D. Rueckert and B. Glocker, 'Automated quality control in image segmentation: Application to the UK biobank cardiovascular magnetic resonance imaging study,' *Journal of Cardiovascular Magnetic Resonance*, vol. 21, no. 1, p. 18, 2019, ISSN: 1097-6647, 1532-429X. DOI: 10.1186/s12968-019-0523-x.

[220] 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, pp. 129–141, 2018, ISSN: 1361-8415, 1361-8423. DOI: 10.1016/j.media.2017.10.001.

[221] J. Yang, Y. He, X. Huang, J. Xu, X. Ye, G. Tao and B. Ni, 'Alignshift: Bridging the gap of imaging thickness in 3d anisotropic volumes,' in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu and L. Joskowicz, Eds., Cham: Springer International Publishing, 2020, pp. 562–572.

[222] C. M. Tax, F. Grussu, E. Kaden, L. Ning, U. Rudrapatna, C. John Evans, S. St-Jean, A. Leemans, S. Koppers, D. Merhof, A. Ghosh, R. Tanno, D. C. Alexander, S. Zappalà, C. Charron, S. Kusmia, D. E. Linden, D. K. Jones and J. Veraart, 'Cross-scanner and cross-protocol diffusion mri data harmonisation: A benchmark database and evaluation of algorithms,' *NeuroImage*, vol. 195, pp. 285–299, 2019, ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2019.01.077.

[223] P.-L. Delisle, B. Anctil-Robitaille, C. Desrosiers and H. Lombaert, 'Realistic image normalization for multi-domain segmentation,' *CoRR*, vol. abs/2009.14024, 2020.

[224] B. E. Dewey, C. Zhao, J. C. Reinhold, A. Carass, K. C. Fitzgerald, E. S. Sotirchos, S. Saidha, J. Oh, D. L. Pham, P. A. Calabresi, P. C. van Zijl and J. L. Prince, 'Deepharmony: A deep learning approach to contrast harmonization across scanner changes,' *Magnetic Resonance Imaging*, vol. 64, pp. 160–170, 2019, Artificial Intelligence in MRI, ISSN: 0730-725X. DOI: https://doi.org/10.1016/j.mri.2019.05.041.

[225] N. Karani, E. Erdil, K. Chaitanya and E. Konukoglu, 'Test-time adaptable neural networks for robust medical image segmentation,' *Medical Image Analysis*, vol. 68, p. 101 907, 2021.

[226] 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, no. 1995, pp. 60–88, 2017, ISSN: 1361-8415.

[227] Q. Dou, D. C. Castro, K. Kamnitsas and B. Glocker, 'Domain generalization via Model-Agnostic learning of semantic features,' in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, 2019, pp. 6447–6458.

[228] I. Albuquerque, N. Naik, J. Li, N. S. Keskar and R. Socher, 'Improving out-of-distribution generalization via multi-task self-supervised pretraining,' *arXiv Preprint*, 2020.

[229] P. Chattopadhyay, Y. Balaji and J. Hoffman, 'Learning to balance specificity and invariance for in and out of domain generalization,' in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IX*, Springer International Publishing, 2020, pp. 301–318.

[230] S. Wang, L. Yu, C. Li, C.-W. Fu and P.-A. Heng, 'Learning from extrinsic and intrinsic supervisions for domain generalization,' in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IX*, A. Vedaldi, H. Bischof, T. Brox and J.-M. Frahm, Eds., ser. Lecture Notes in Computer Science, vol. 12354, Springer, 2020, pp. 159–176.

[231] S. Shankar, V. Piratla, S. Chakrabarti, S. Chaudhuri, P. Jyothi and S. Sarawagi, 'Generalizing across domains via Cross-Gradient training,' in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, OpenReview.net, 2018.

[232] K. Daniel, *Thinking, fast and slow*. 2017.

[233] L. Zhang, X. Wang, D. Yang, T. Sanford, S. Harmon, B. Turkbey, B. J. Wood, H. Roth, A. Myronenko, D. Xu and Z. Xu, 'Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation,' *IEEE Transactions on Medical Imaging*, vol. 39, no. 7, pp. 2531–2540, 2020.

[234] T. Devries and G. W. Taylor, 'Improved regularization of convolutional neural networks with cutout,' *arXiv Preprint*, 2017, arXiv:1708.04552.

[235] R. G. Lopes, D. Yin, B. Poole, J. Gilmer and E. D. Cubuk, 'Improving robustness without sacrificing accuracy with patch gaussian augmentation,' *arXiv Preprint*, 2019, arXiv:1906.02611.

[236] Z. Zhou, V. Sodha, M. M. Rahman Siddiquee, R. Feng, N. Tajbakhsh, M. B. Gotway and J. Liang, 'Models genesis: Generic autodidactic models for 3D medical image analysis,' in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, Springer International Publishing, 2019, pp. 384–393. DOI: 10.1007/978-3-030-32251-9_42.

[237] Z. Zhou, V. Sodha, J. Pang, M. B. Gotway and J. Liang, 'Models genesis,' *Medical Image Analysis*, vol. 67, p. 101 840, 2021. DOI: 10.1016/j.media.2020.101840.

[238] X. Zhang, Z. Wang, D. Liu, Q. Lin and Q. Ling, 'Deep adversarial data augmentation for extremely low data regimes,' *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 1, pp. 15–28, 2021. DOI: 10.1109/TCSVT.2020.2967419.

[239] L. Zhao, T. Liu, X. Peng and D. N. Metaxas, 'Maximum-entropy adversarial data augmentation for improved generalization and robustness,' in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Sys-*

tems 2020, NeurIPS 2020, December 6-12, 2020, virtual, H. Larochelle, M. Ranzato, R. Hadsell, M.-F. Balcan and H.-T. Lin, Eds., 2020.

[240] W. Zheng, Z. Chen, J. Lu and J. Zhou, 'Hardness-aware deep metric learning,' in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 2019, pp. 72–81.

[241] Y. Zhang and Q. Yang, 'A survey on multi-task learning,' *arXiv Preprint*, 2017, arXiv:1707.08114.

[242] A. J. Larrazabal, C. Martinez and E. Ferrante, 'Anatomical priors for image segmentation via post-processing with denoising autoencoders,' in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, Springer International Publishing, 2019, pp. 585–593.

[243] N. Tishby and N. Zaslavsky, 'Deep learning and the information bottleneck principle,' in *2015 IEEE Information Theory Workshop, ITW 2015, Jerusalem, Israel, April 26 - May 1, 2015*, IEEE, 2015, pp. 1–5. DOI: 10.1109/ITW.2015.7133169.

[244] Z. Huang, H. Wang, E. P. Xing and D. Huang, 'Self-challenging improves Cross-Domain generalization,' in *Computer Vision – ECCV 2020*, Springer International Publishing, 2020, pp. 124–140.

[245] 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, L. Li, X. Zhuang, D. Vilades, M. L. Descalzo, A. Guala, L. La Mura, M. G. Friedrich, R. Garg, J. Lebel, F. Henriques, M. Karakas, E. Cavus, S. E. Petersen, S. Escalera, S. Segui, J. F. Rodriguez-Palomares and K. Lekadir, 'Multi-Centre, Multi-Vendor and Multi-Disease cardiac segmentation: The M&Ms challenge,' *IEEE Transactions on Medical Imaging*, 2021. DOI: 10.1109/TMI.2021.3090082.

[246] K. He, X. Zhang, S. Ren and J. Sun, 'Identity mappings in deep residual networks,' in *Computer Vision – ECCV 2016*, Springer International Publishing, 2016, pp. 630–645.

[247] Z. Xu, D. Liu, J. Yang, C. Raffel and M. Niethammer, 'Robust and generalizable visual representation learning via random convolutions,' in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, OpenReview.net, 2021.

[248] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan and Q. V. Le, 'Autoaugment: Learning augmentation strategies from data,' in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 113–123.

[249] R. Shaw, C. H. Sudre, S. Ourselin and M. J. Cardoso, 'MRI k-space motion artefact augmentation: Model robustness and task-specific uncertainty,' in *International Conference on Medical Imaging with Deep Learning, MIDL 2019, 8-10 July 2019, London, United Kingdom*, M. J. Cardoso, A. Feragen, B. Glocker, E. Konukoglu, I. Oguz, G. B. Unal and T. Vercauteren, Eds., ser. Proceedings of Machine Learning Research, vol. 102, PMLR, 2019, pp. 427–436.

[250] P. F. Christ, M. E. A. Elshaer, F. Ettlinger, S. Tatavarty, M. Bickel, P. Bilic, M. Rempfler, M. Armbruster, F. Hofmann, M. D'Anastasi, W. H. Sommer, S.-A. Ahmadi and B. H. Menze, 'Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields,' in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, Springer International Publishing, 2016, pp. 415–423.

[251] A. J. Larrazabal, C. Martinez, B. Glocker and E. Ferrante, 'Post-DAE: Anatomically plausible segmentation via Post-Processing with denoising autoencoders,' *IEEE transactions on medical imaging*, vol. 39, no. 12, pp. 3813–3820, 2020.

[252] P. M. Full, F. Isensee, P. F. Jäger and K. H. Maier-Hein, 'Studying robustness of semantic segmentation under domain shift in cardiac MRI,' in *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges - 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers*, E. Puyol-Antón, M. Pop, M. Sermesant, V. M. Campello, A. Lalande, K. Lekadir, A. Suinesiaputra, O. Camara and A. A. Young, Eds.,

ser. Lecture Notes in Computer Science, vol. 12592, Springer, 2020, pp. 238–249. DOI: 10.1007/978-3-030-68107-4_24.

[253] G. A. Bello, T. J. W. Dawes, J. Duan, C. Biffi, A. de Marvao, L. S. G. E. Howard, J. S. R. Gibbs, M. R. Wilkins, S. A. Cook, D. Rueckert and D. P. O'Regan, 'Deep learning cardiac motion analysis for human survival prediction,' *Nature Machine Intelligence*, vol. 1, pp. 95–104, 2019, ISSN: 2522-5839. DOI: 10.1038/s42256-019-0019-2.

[254] C. Chen, C. Qin, C. Ouyang, S. Wang, H. Qiu, L. Chen, G. Tarroni, W. Bai and D. Rueckert, 'Enhancing MR image segmentation with realistic adversarial data augmentation,' *arXiv Preprint*, 2021, arXiv:2108.03429.

[255] D. C. Castro, I. Walker and B. Glocker, 'Causality matters in medical imaging,' *Nature Communications*, vol. 11, no. 1, p. 3673, 2020. DOI: 10.1038/s41467-020-17478-w.

[256] G. Xu, T. D. Duong, Q. Li, S. Liu and X. Wang, 'Causality learning: A new perspective for interpretable machine learning,' *arXiv Preprint*, 2020, arXiv:2006.16789.

[257] T. Narendra, A. Sankaran, D. Vijaykeerthy and S. Mani, 'Explaining deep learning models using causal inference,' *arXiv Preprint*, 2018, arXiv:1811.04376.

[258] S. Barocas, A. D. Selbst and M. Raghavan, 'The hidden assumptions behind counterfactual explanations and principal reasons,' in *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, M. Hildebrandt, C. Castillo, E. Celis, S. Ruggieri, L. Taylor and G. Zanfir-Fortuna, Eds., ACM, 2020, pp. 80–89. DOI: 10.1145/3351095.3372830.

[259] J. Heo, H. B. Lee, S. Kim, J. Lee, K. J. Kim, E. Yang and S. J. Hwang, 'Uncertainty-Aware attention for reliable interpretation and prediction,' in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett, Eds., Curran Associates Inc., 2018, pp. 909–918.

Appendices

Supplementary material

Below is the supplementary material for the work presented in Sec. 5.1.

Table A1: **Segmentation performance across images of different slice thicknesses.** The average Dice scores and standard deviation are reported for each group. While the segmentation network was trained only using images of 8 mm slice thickness, this network produces satisfactory performances on images of different slice thicknesses.

Test dataset	Slice thickness (mm)	Number of images	LV		MYO		RV	
			mean	std	mean	std	mean	std
ACDC	5	24	0.89	0.06	0.76	0.09	0.82	0.09
	6.5	2	0.93	0.03	0.81	0.00	0.83	0.03
	7	2	0.85	0.09	0.69	0.05	0.82	0.06
	10	172	0.90	0.10	0.82	0.06	0.82	0.14
BSCMR-AS	5	4	0.89	0.08	0.85	0.04	-	-
	6	94	0.86	0.12	0.83	0.09	-	-
	7	486	0.88	0.1	0.83	0.07	-	-
	8	294	0.89	0.09	0.83	0.07	-	-
	10	318	0.89	0.07	0.85	0.04	-	-

Permissions for content reuse

Springer

Springer allows authors to reuse their article's Version of Record, in whole or in part, in their own thesis[4]. This applies to the contents presented in Sec. 3.1, Sec. 3.2, Sec. 4.1, Sec. 4.2, and Sec. 5.2, in which I reused my own published works. These published works have been cited in the beginning of corresponding chapters.

Frontiers in Cardiovascular Medicine

Parts of contents in Chapter. 2 and Sec. 5.1 reused my own works published in Frontiers. Both of them have been granted with *Open access* under the terms of the Creative Commons Attribution License (CC BY 4.0)[5]. These published works have been cited in the beginning of corresponding chapters.

Permissions for reproduced figures

All reproduced figures from others work in this thesis are licensed under the terms of the Creative Commons Attribution License (CC BY 3.0 or 4.0), except the figures listed in Table A2. For these figures, permission for noncommercial use has been granted by corresponding rights holders, see below.

Table A2: **A list of reproduced figures with granted license.**

Type of work	Name of work	Source of work	Licensed content publisher	Permission requested on	I have permission yes /no	Licence number
Figure	Fig. 2.7	[38]	Springer Nature	20-Jul-2021	yes	5113091203298
Figure	Fig. 2.9	[87]	Springer Nature	20-Jul-2021	yes	5113100502868
Figure	Fig. 2.10	[97]	Elsevier	20-Jul-2021	yes	5113100802783
Figure	Fig. 2.12	[126]	Proceedings of the National Academy of Sciences USA (PNAS)	-	yes	N/A [a]

[a] Permission is not required to use original figures or tables for noncommercial and educational use, providing the source of figure and tables is properly cited (https://www.pnas.org/page/about/rights-permissions).

[4] https://www.springer.com/gp/rights-permissions/obtaining-permissions/882

[5] https://creativecommons.org/licenses/by/4.0/