



Deep learning for robust segmentation and explainable analysis of 3d and dynamic cardiac images

Qiao Zheng

► To cite this version:

Qiao Zheng. Deep learning for robust segmentation and explainable analysis of 3d and dynamic cardiac images. Artificial Intelligence [cs.AI]. COMUE Université Côte d'Azur (2015 - 2019), 2019. English. NNT : 2019AZUR4013 . tel-02083415v2

HAL Id: tel-02083415

<https://theses.hal.science/tel-02083415v2>

Submitted on 14 Feb 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



THÈSE DE DOCTORAT

Apprentissage Profond pour la Segmentation Robuste et
l'Analyse Explicable des Images Cardiaques Volumiques
et Dynamiques

Qiao ZHENG

INRIA Sophia Antipolis, Equipe EPIONE

Présentée en vue de l'obtention du grade de Docteur en Science
d'Université Côte d'Azur

Dirigée par : Nicholas AYACHE, Hervé DELINGETTE

Soutenue le : 27 mars 2019

Devant le jury, composé de :

Nicholas AYACHE	Dr.	INRIA (Equipe Epione)
Patrick CLARYSSE	Dr.	CNRS (CREATIS)
Hervé DELINGETTE	Dr.	INRIA (Equipe Epione)
Nicolas DUCHATEAU	Prof.	CNRS (CREATIS)
Alistair YOUNG	Prof.	KCL (Londres)

Apprentissage Profond pour la Segmentation Robuste et l'Analyse Explicable des Images Cardiaques Volumiques et Dynamiques

Jury :

<i>Président du jury :</i>	Patrick CLARYSSE	-	CNRS (CREATIS)
<i>Rapporteurs :</i>	Patrick CLARYSSE	-	CNRS (CREATIS)
	Alistair YOUNG	-	KCL (Londres)
<i>Examineurs :</i>	Nicolas DUCHATEAU	-	CNRS (CREATIS)
<i>Superviseurs :</i>	Nicholas AYACHE	-	INRIA (Equipe Epione)
	Hervé DELINGETTE	-	INRIA (Equipe Epione)

Apprentissage Profond pour la Segmentation Robuste et l'Analyse Explicable des Images Cardiaques Volumiques et Dynamiques

Résumé: L'IRM cardiaque est largement utilisée par les cardiologues car elle permet d'extraire des informations riches dans les images. Toutefois, si cela est fait manuellement, le processus d'extraction des informations est fastidieux et prend beaucoup de temps. Compte tenu des progrès de l'intelligence artificielle, je développe des méthodes d'apprentissage profond pour traiter l'automatisation de plusieurs tâches essentielles de l'analyse de l'IRM cardiaque. Tout d'abord, je propose une méthode basée sur les réseaux de neurones convolutifs pour effectuer la segmentation cardiaque sur des ensembles d'images IRM petit axe. Dans cette méthode, étant donné que la prédiction d'une segmentation d'une coupe dépend de celle d'une coupe adjacente, la cohérence 3D et la robustesse sont explicitement imposées. De plus, je propose une méthode de classification de plusieurs pathologies cardiaques, avec une nouvelle approche d'apprentissage profond pour extraire des attributs dérivés des images afin de caractériser la forme et le mouvement du cœur. En particulier, le modèle de classification est explicable, simple et flexible. Enfin et surtout, la même méthode d'extraction d'éléments est appliquée à un ensemble de données exceptionnellement volumineux (UK Biobank). La classification non supervisée des données est ensuite effectuée sur les attributs extraits pour caractériser ces pathologies cardiaques. Pour conclure, je discute de plusieurs prolongements possibles de mes recherches.

Mots clés: Apprentissage profond, segmentation cardiaque, analyse cardiaque, ciné-IRM

Deep Learning for Robust Segmentation and Explainable Analysis of 3D and Dynamic Cardiac Images

Abstract: Cardiac MRI is widely used by cardiologists as it allows extracting rich information from images. However, if done manually, the information extraction process is tedious and time-consuming. Given the advance of artificial intelligence, I develop deep learning methods to address the automation of several essential tasks on cardiac MRI analysis. First, I propose a method based on convolutional neural networks to perform cardiac segmentation on short axis MRI image stacks. In this method, since the prediction of a segmentation of a slice is dependent upon the already existing segmentation of an adjacent slice, 3D-consistency and robustness is explicitly enforced. Second, I develop a method to classify cardiac pathologies, with a novel deep learning approach to extract image-derived features to characterize the shape and motion of the heart. In particular, the classification model is explainable, simple and flexible. Last but not least, the same feature extraction method is applied to an exceptionally large dataset (UK Biobank).

Unsupervised cluster analysis is then performed on the extracted features in search of their further relation with cardiac pathology characterization. To conclude, I discuss several possible extensions of my research.

Keywords: Deep learning, cardiac segmentation, cardiac analysis, cine MRI

Acknowledgments

First, I want to thank my supervisors, Nicholas and Hervé, for the great supervision they provided during my Ph.D. Nicholas offered me this wonderful opportunity to conduct scientific research as a Ph.D. student in this amazing lab. In addition to carefully mentoring me in specific research projects, he also showed me the ways to become a scientist and academic. What I have learned from him is more than I can write down in this thesis. Hervé has been very supportive since the very beginning of my Ph.D. His ideas often inspired me deeply and enabled me to find interesting new ways in my research. Moreover, talking to him is a pleasure as he has broad knowledge in various domains.

Then, I want to thank the reviewers of my thesis, Dr. Patrick Clarysse, and Prof. Alistair Young, for kindly reading and reviewing my manuscript, as well as for attending my thesis defense. A special ‘thanks’ goes to Prof. Nicolas Duchateau, not only for being part of the jury but also for guiding me and working with me patiently in the first two years of my Ph.D.

Also, I want to thank Prof. Steffen E. Petersen and his team for kindly giving me access to the large UK Biobank dataset and working with me on it. I want to thank Prof. Daniel Rueckert and his team for generously sharing their tool to help my research.

Moreover, I want to say ‘thank you very much’ to all the other members of the lab. Xavier, Maxime, and Marco, even if they are busy, are always willing to share their great ideas and insights with me. I enjoy a lot talking to them. Isabelle is so warm-hearted and efficient that I can always solve all kinds of administration problems with her help. All the other colleagues in the lab, including those who have left the lab, are extremely friendly and helpful to me. Even though the work and life of Ph.D. candidates are challenging in general, I think I am very lucky here as I have these people around me, including Alain, Charles, Hervé L., Elena, Marco M., Fanny, Sara, Jan, Chloé, Rocio, Loïc L., Mehdi, Bisheh, Mathieu, Nina, Marc-Michel, Sophie, Thomas, Roch, Raphael, Loïc D., Pawel, Shuman, Wen, Julian, Nicolas C., Luigui, Clément, Jaume, Tania, Zihao, Nicolas G., Yann, Bastien and Santiago. This lab certainly has some of the nicest people in the world.

Furthermore, I want to thank all the teachers and staff and classmates who taught me, advised me and assisted me at Lycée Louis-le-Grand, Ecole Polytechnique and MIT. Without what I learned and got from these people in these great schools, I would not be able to become who I am today. Also, I want to thank all my supervisors and colleagues who supported me when I was doing the jobs and internships in Paris, Hong Kong, Boston, and New York City. The skills and experiences I obtained by working with them proved to be highly valuable, even for my research in academia.

In particular, let me thank my parents and family. I am aware that if I have seen further, it is by standing on their shoulders.

Finally, I acknowledge the financial support from the European Research Council (MedYMA ERC-AdG-2011-291080).

Contents

1	Introduction	1
1.1	Context	1
1.1.1	Medical Image Analysis for Fighting Cardiovascular Diseases: Big Challenges	1
1.1.2	Rise of Deep Learning	2
1.2	Main Objectives	2
1.3	Structure of the Thesis	3
2	Segmentation Using Simulation for Data Augmentation	5
2.1	Introduction	5
2.2	Approach	7
2.2.1	Data preprocessing.	7
2.2.2	Initial segmentation.	8
2.2.3	Spatial segmentation propagation.	8
2.3	Networks	8
2.3.1	Multi-scale coarse-to-fine prediction	8
2.3.2	Loss function.	11
2.3.3	Convolution layer group.	11
2.3.4	Data augmentation inside network.	11
2.4	Experiments	12
2.4.1	Training.	12
2.4.2	Testing.	12
2.5	Conclusion and Perspectives	15
3	Consistent and Robust Segmentation with Spatial Propagation	17
3.1	Introduction	18
3.2	Data	22
3.2.1	Datasets	22
3.2.2	Notation and Terminology	22
3.2.3	Adaptation of the UK Biobank Ground-Truth	23
3.3	Methods	24
3.3.1	Region of Interest (ROI) Determination: ROI-net	24
3.3.2	Segmentation with Propagation: LVRV-net and LV-net	27
3.3.3	Image Preprocessing	28
3.3.4	Loss Functions	29
3.4	Experiments and Results	30
3.4.1	Technical Details about Training the Three Networks	30
3.4.2	Experiments on UK Biobank & Contribution of the Propagation	30
3.4.3	Generalization Ability to Other Datasets	36
3.5	Conclusion and Discussion	45

3.6	Appendix	46
3.6.1	Datasets	46
3.6.2	Metrics	47
4	Explainable Pathology Classification with Motion Characterization	49
4.1	Introduction	50
4.2	Data	53
4.2.1	Dataset	53
4.2.2	Notation	53
4.3	Methods	54
4.3.1	Preprocessing: Region of Interest (ROI) Determination	54
4.3.2	Feature Extraction Step 1: Apparent Flow Generation	55
4.3.3	Feature Extraction Step 2: Segmentation	58
4.3.4	Feature Extraction Step 3: Shape-Related Features	58
4.3.5	Feature Extraction Step 4: Motion-Characteristic Features	59
4.3.6	Classification	63
4.4	Experiments and Results	65
4.4.1	Training ApparentFlow-net	65
4.4.2	Finetuning LVRV-net	67
4.4.3	Proposed Classification Model	67
4.4.4	Variants of the Proposed Classification Model	69
4.5	Conclusion and Discussion	72
4.6	Appendix	75
4.6.1	Loss Function for Training ApparentFlow-Net	75
4.6.2	Variants of the Proposed Classification Model with Different Values of Parameter C	76
4.6.3	Variants of the Proposed Classification Model with Different Classifiers and Input Features	76
4.6.4	Examples of Apparent Flow Generated by the ApparentFlow-net	77
5	Cluster Analysis of Image-Derived Features	83
5.1	Introduction	83
5.2	Data	85
5.2.1	UK Biobank	85
5.2.2	ACDC	85
5.3	Methods	86
5.3.1	Feature Extraction	86
5.3.2	Feature Selection	87
5.3.3	Cluster Analysis	87
5.4	Experiments and Results	88
5.4.1	Feature Extraction	88
5.4.2	Feature Selection	88

5.4.3	Cluster Analysis	89
5.4.4	Further Analysis for Confirmation	92
5.5	Conclusion and Discussion	96
6	Conclusion and Perspectives	103
6.1	Main Contributions	103
6.1.1	Segmentation Using Simulation for Data Augmentation	103
6.1.2	Consistent and Robust Segmentation with Spatial Propagation	104
6.1.3	Explainable Pathology Classification with Motion Characterization	104
6.1.4	Cluster Analysis of Image-Derived Features	104
6.2	Publications	105
6.3	Software	105
6.4	Perspectives	105
6.4.1	Cardiac Mesh Simulation and Image Synthesis for Deep Learning	105
6.4.2	Temporal Consistency of Segmentation	106
6.4.3	Semi-Supervised Learning and Unsupervised Learning	106
6.4.4	More Explainable Models	107
	Bibliography	109

Introduction

Contents

1.1 Context	1
1.1.1 Medical Image Analysis for Fighting Cardiovascular Diseases: Big Challenges	1
1.1.2 Rise of Deep Learning	2
1.2 Main Objectives	2
1.3 Structure of the Thesis	3

1.1 Context

1.1.1 Medical Image Analysis for Fighting Cardiovascular Diseases: Big Challenges

According to the World Health Organization¹, cardiovascular diseases are the first cause of death in the world. It is estimated that about 18 million people died from cardiovascular diseases in 2016, which is 31% of all global deaths. While the problems caused by cardiovascular diseases have been noticed in high-income countries for a long time, the social and economic impact is particularly heavy in low- and middle-income countries these days, in which over three-quarters of the deaths caused by cardiovascular diseases take place.

In order to diagnose and treat cardiovascular diseases, clinicians routinely rely on medical imaging. Medical images hence play an indispensable role in preventing and treating cardiovascular diseases. While medical images are the source of many valuable data in general, how to interpret them and extract the most useful and relevant information from them remain a challenge for many reasons. First, medical images are quite different from ordinary images. The expertise in medicine, which can only be obtained by medical professionals after years of study and practice, is required to understand and interpret medical images. However, the medical professionals' shortage has remained a concern worldwide ([Al-Shamshi 2017]). And the situation is unlikely to be much better in the near and medium-term future. So there are often not enough medical professionals to interpret the medical images.

¹[https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds))
(accessed January 30, 2019)

Second, manually processing and interpreting medical images, even if by experienced medical professionals, is a tedious and time-consuming job. Medical-image-related methods and tools to help medical professionals to relieve their burden are always in demand ([Adelman 2014]). Third, as we are already in the era of big data, more and more medical data, in particular, medical image data, are available in recent years. How to make good use of them is still largely an open question [Lee 2017].

1.1.2 Rise of Deep Learning

Over the last few years, deep learning has achieved great success in various areas of artificial intelligence. As shown in [LeCun 2015], deep learning has dramatically improved the state-of-the-art in various domains, such as image processing, speech recognition, visual object recognition, drug discovery, and genomics. The fact that deep learning is very good at discovering intricate structures in data makes it extremely versatile. Furthermore, based on modern hardware and software, deep learning models are very fast even when applied to large images.

Being aware of the various advantages of deep learning, people wonder what it can bring to medical image analysis. Nowadays, many researchers in medical image analysis are trying to apply deep learning to cope with various challenges in the domain. According to [Litjens 2017], the use of deep learning is rather successful in performing medical image analysis for different tasks (e.g. image classification, object detection, segmentation, registration) and application areas (e.g. neuro ([Chen 2016]), retinal ([Zilly 2017]), pulmonary ([Cheng 2016]), digital pathology ([Han 2016]), breast ([Dalmis 2017]), cardiac [Oktay 2018], abdominal ([Ravishankar 2016]), musculoskeletal ([Forsberg 2017])). Therefore, deep learning is being considered as a very powerful and promising method in medical image analysis. Yet the study of deep learning in medical image analysis is just at its beginning. While some encouraging accomplishments have been achieved, the research in this direction still has much to explore ([Zhou 2017]).

1.2 Main Objectives

Given the context above, this thesis focuses on the development and application of deep learning methods to tackle cardiac-related problems in medical image analysis. More specifically, the main questions we investigate are:

- Deep learning models usually learn from large amounts of data. In order to better train deep learning models, can we apply existing simulation and synthesis methods in medical image analysis for the data augmentation of cardiac images?
- Compared with accuracy, the consistency and robustness of cardiac segmentation have been less explored topics. How to make the segmentation of cardiac images more consistent and robust?
- While clinicians often rely on image-derived features for cardiac diagnosis, how can we extract useful features from cardiac images for pathology classification in an explainable manner using deep learning?

- With image-derived features, is it feasible to identify cases of cardiac pathologies in a large general population even without labels for training (i.e in an unsupervised way)?

1.3 Structure of the Thesis

The thesis is presented in chronological order.

Chapter 2 shows that using an existing method of cardiac mesh simulation and image synthesis, data augmentation can be effectively done for training a deep learning model of cardiac segmentation. This chapter is adapted from [Zheng 2018c].

Chapter 3 aims at developing and evaluating a deep-learning-based cardiac segmentation model with consistency and robustness. For this purpose, a technique called spatial propagation of segmentation is proposed. This chapter is based on the publication [Zheng 2018b].

Chapter 4 describes how shape-related and motion-derived features can be extracted from cardiac images based on deep learning, and how these features enable classification of cardiac pathologies in an explainable way. The work presented in this chapter is published in [Zheng 2018a].

Chapter 5 highlights an example of unsupervised cluster analysis of shape-related and motion-characteristic features extracted from the large UK Biobank dataset. This study will be submitted to a journal soon for publication.

Chapter 6 summarizes the main contributions of the thesis and discusses the perspectives.

Segmentation Using Simulation for Data Augmentation

Contents

2.1	Introduction	5
2.2	Approach	7
2.2.1	Data preprocessing.	7
2.2.2	Initial segmentation.	8
2.2.3	Spatial segmentation propagation.	8
2.3	Networks	8
2.3.1	Multi-scale coarse-to-fine prediction	8
2.3.2	Loss function.	11
2.3.3	Convolution layer group.	11
2.3.4	Data augmentation inside network.	11
2.4	Experiments	12
2.4.1	Training.	12
2.4.2	Testing.	12
2.5	Conclusion and Perspectives	15

Part of this chapter corresponds to the following scientific article:

- [Zheng 2018c] **3D Consistent Biventricular Myocardial Segmentation Using Deep Learning for Mesh Generation**
Qiao Zheng, Hervé Delingette, Nicolas Duchateau and Nicholas Ayache. arXiv preprint, 2018

2.1 Introduction

While most research about myocardial segmentation focuses on either left ventricle (LV) [Avendi 2016a] or right ventricle (RV) [Avendi 2017] segmentation on 2D slices, there is a great need for 3D-consistent biventricular (BV) segmentation, in which LV and RV together are segmented. 3D-consistent BV segmentation provides not only consistency, but also robustness on poor-quality slices (e.g. near apex). Its output may then be used to generate complete meshes. These advantages are missing from most other methods. For example the model of [Tran 2016] is not very capable

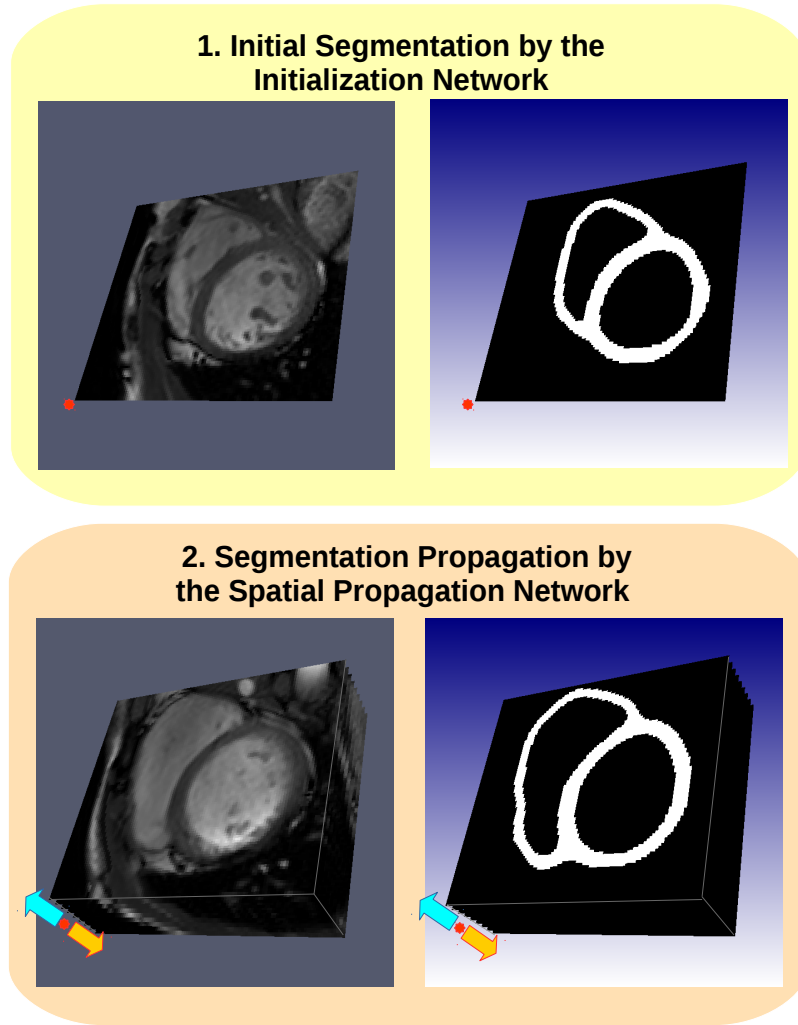


Figure 2.1: Overview of the proposed method. (1) Segmentation of a single slice at the middle of the volume by the initialization network. (2) Propagation of predicted mask towards base and apex by the spatial propagation network.

of segmenting the slices near apex. The lack of publicly available automatic BV mesh generation tool and the success of deep learning on medical image analysis [Zhou 2017] motivate us to develop this method based on two neural networks : the former segments a single slice in the volume and the latter propagates the segmentation to the other slices.

The most competitive BV myocardial segmentation methods are the automatic ones. Shape-constrained deformable models are applied on a dataset of 28 CT volumes in [Ecabert 2008]. The chain takes 22s for a volume. In [Zheng 2008] 4-chamber segmentation is performed using steerable features learned on a dataset of

457 annotated CT volumes. The speed is 4s per volume. The authors of [Wang 2013] apply marginal space learning and probabilistic boosting-tree on a dataset of 100 annotated MRI volumes to learn to jointly delineate LV and RV. It spends about 3s on each volume. We cannot compare directly with these methods on their reported error measures due to differences in datasets. In this paper we propose an effective pipeline based on 2 neural networks combining the assets of 2D (speed) and 3D (consistency). An original loss function is also applied for training. Our approach has the following advantages:

- Unlike the above-mentioned papers our networks are trained on a publicly available dataset STACOM [Tobon-Gomez 2013].
- The dataset we use of 15 annotated volumes is much smaller than the above-mentioned datasets. Our method is data-efficient as its generalize even with small training sets.
- Unlike the above methods, our method is model-free. Anyone familiar with deep learning may implement our networks without difficulty.
- Our method takes about 3s to segment a volume. This is the same as the fastest one in the above methods.
- Compared to MRI images, CT images usually have much better resolution and stronger heart/background contrast. Working with MRI images, we actually solve a more challenging version of segmentation problems.

2.2 Approach

The overview of our approach is shown in Figure 2.1.

2.2.1 Data preprocessing.

2.2.1.1 Cropping ROI.

Usually on MRI images there is more information than we need for myocardium segmentation. If this is the case, getting rid of irrelevant background information can simplify the job of segmentation. For any cardiac MRI short axis image set of a subject to be used as input to our model, we process 3D stacks of 2D slices, cropped around the heart. First, 3D image volumes are constructed by arranging the 2D MRI slices. Then a 2D ROI is manually determined for each volume such that the myocardium on all slices is included. We usually locate the borders of ROI such that there is a 10mm to 20mm margin between them and the largest myocardium on slices. Finally the sub-volume defined by the ROI is cropped.

2.2.1.2 Resampling.

As standardization the cropped volume is resampled into an isotropic volume by linear interpolation. The goal of this step is to standardize the input volume data to our model as well as to generate more slices of which the appearance gradually changes along the large axis. This smoothness of change is helpful for the propagation of segmentation to be accurate.

2.2.1.3 Identifying basal slice.

Most MRI image volumes include cardiac structures above the base. Furthermore, since our method aims to segment the myocardium up to the base, we manually identify the base slice near mitral annulus. The propagation of myocardium segmentation, as we will present in detail later, will stop once the slice of base is reached. For images of reasonably good quality (e.g. STACOM) the segmentation can be initialized from any slice around the middle of the volume. So in this paper, for testing on the 3 testing cases of STACOM in each fold, the initialization slice is automatically chosen as the one in the middle of the volume.

2.2.2 Initial segmentation.

We then apply the initialization network to segment the selected slice. The output is a mask of which each pixel is a probability (0 for background and 1 for myocardium).

2.2.3 Spatial segmentation propagation.

Then we apply the spatial propagation network to propagate segmentation masks. During upward (towards base) propagation, we suppose the slices up to that of index z are already segmented. Taking this slice, its predicted mask and the next 4 slices ($z+1$ to $z+4$) as input, the spatial propagation network predicts the next 4 segmentation masks. The iteration stops at the base slice. Similarly, we downward (towards apex) propagate the segmentation masks. Thereby we complete the segmentation.

2.3 Networks

2.3.1 Multi-scale coarse-to-fine prediction

The two neural networks we use for this paper are characterized by multi-scale coarse-to-fine predictions (Figure 2.2 and Figure 2.3). As presented in Figure 2.2, the main body of the initialization network is separated into 3 sub-networks with input/output sizes 32, 64 and 128 respectively. SubNet32, taking a downsampled slice of size 32 as input, outputs a predicted mask of the same size. Then SubNet64 takes a downsampled slice of size 64 and incorporates the predicted mask of size 32 to make a prediction of size 64. Similarly, SubNet128 outputs the final predicted mask of size 128. During training, 3 loss functions which compare the outputs of

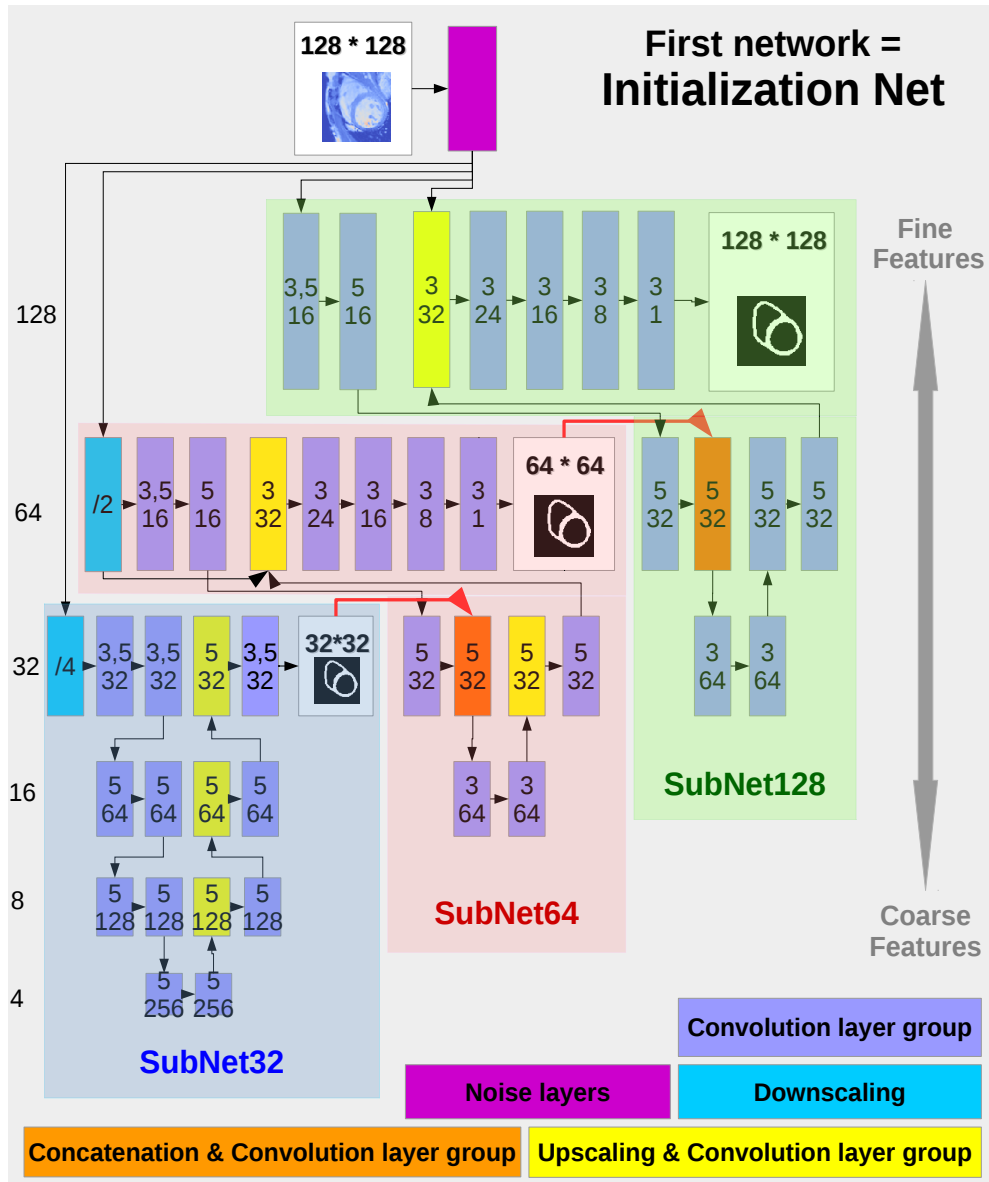


Figure 2.2: The initialization network: the number marked to the left of each row is the size of output feature maps in the row; the upper number(s) in the rectangle of convolution layer is the filter size while the lower number indicates output channels; the number in rectangle of downscaling layer is the applied scale.

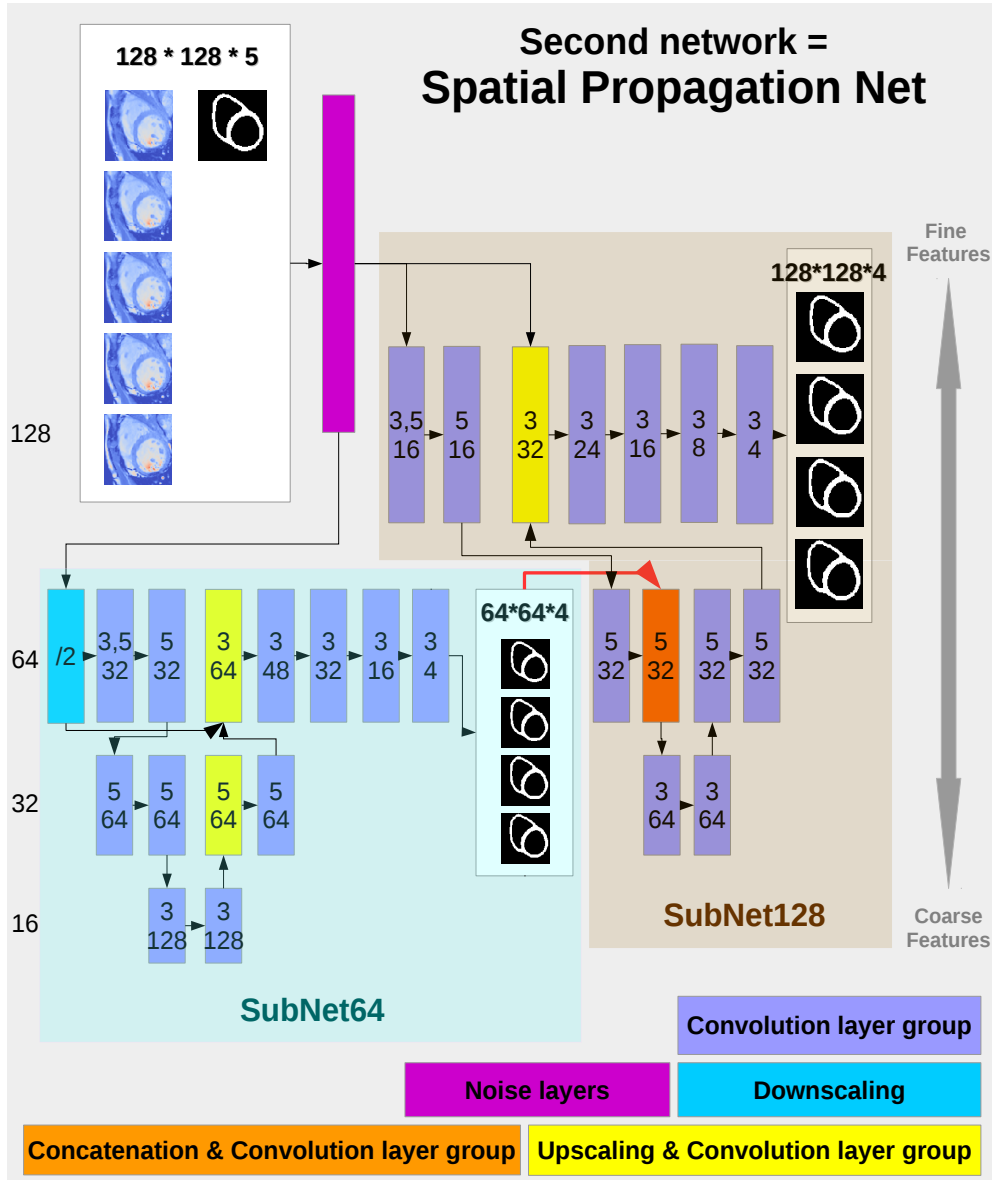


Figure 2.3: The spatial propagation network: the number marked to the left of each row is the size of output feature maps in the row; the upper number(s) in the rectangle of convolution layer is the filter size while the lower number indicates output channels; the number in rectangle of downscaling layer is the applied scale.

SubNet32, SubNet64 and SubNet128 with the ground truth masks of size 32, 64 and 128 respectively are applied. The spatial propagation network, consisting of SubNet64 and SubNet128, has analogous structure and loss functions for training.

2.3.2 Loss function.

The networks are trained by stochastic gradient descent. An original loss function is designed to overcome numerical instability and class imbalance during training. We call it *stabilized and class-balanced cross entropy loss*, where pixel-wise losses are added to work with a total loss. For each pixel, suppose the predicted probability is p and the ground truth is g . The pixel loss is

$$pixelLoss = \begin{cases} 0 & \text{if } |g - p'| < t, \\ -\log(p') & \text{if } g = 1 \text{ and } p' \leq 1 - t, \\ -\log(1 - p') & \text{if } g = 0 \text{ and } p' \geq t, \end{cases} \quad (2.1)$$

with

$$p' = ap + b \quad (2.2)$$

and a , b and t are parameters such that $a > 0$, $b > 0$, $a + 2b = 1$ and $0 < t < 1$. To roughly preserve the predicted probability, a is set close to 1, b and t are set near 0. In this paper we empirically pick $a = 0.999$, $b = 0.0005$ and $t = 0.02$.

The purpose of applying (2) is to avoid computation of logarithm on any value too close to 0 while roughly reserving the predicted probability. Without it, poorly predicted values of p may result in extremely large loss and gradient values, which may harm numerical stability of training and even cause overflow.

On the other hand, there is a strong imbalance between myocardial and background pixel. The latter represents around 80% of the image. With common loss functions, the overall training effect of background pixels dominates the effect of the myocardial pixels. It may hinder the network performance in recognizing the myocardium. Setting the loss to 0 in (1) whenever the prediction is close enough to ground truth reduces the effect. When the predicted probabilities for background are close enough to 0, our loss function stops “pulling” them further to 0 and instead focuses on “pushing” the probabilities on myocardium to 1.

2.3.3 Convolution layer group.

The two networks mainly consist of convolution layer groups. In each group, a convolution layer is followed by a batch normalization layer and a leaky ReLU layer of negative part coefficient 0.25.

2.3.4 Data augmentation inside network.

The first layers in the two networks are noise layers for data augmentation during training to make the networks more robust. They are removed in testing. Data

augmentation includes randomly rotating input slices together and adding Gaussian and pepper-and-salt noise.

2.4 Experiments

Our experiments involved an existing dataset with MRI image volume sequences: 15 subjects from STACOM [Tobon-Gomez 2013] (30 instants per cycle, with ground truth segmentation). After resampling to isotropic volume of voxel size 1.25mm there are about 60 slices below the base in each volume. We divide the 15 cases into 5 groups of 3 cases. In each fold of the 5-fold cross-validation, the 3 cases of one picked group are used as testing. And the training set consists of the 12 cases from the remaining 4 groups.

2.4.1 Training.

As data augmentation to generate a large database with ground truth from a small database, we combine a motion simulation method and an image synthesis algorithm to generate realistic volume sequence variants of the training cases. Infarcted mesh motion sequences were simulated according to the scheme depicted in [Duchateau 2016]. Then the original volume sequences were warped to generate synthesized image variants using an algorithm inspired from [Prakosa 2013]. For each of the training subjects, 31 (1 healthy and 30 infarcted) 30-instant volume sequence variants were generated. In total $12 \times 31 \times 30 = 11160$ volumes were used for training the spatial segmentation network in each fold of the 5-fold cross-validation. On the other hand, we observe better image/mesh coincidence around end-diastole than around end-systole in the synthesized sequences. Considering the trade-off between robustness (diversity in training set) and accuracy (image/mesh coincidence), we decide to train the initialization network only with volumes from 10 instants around end-diastole. Hence it is trained on $12 \times 31 \times 10 = 3720$ volumes. Please note the augmentations of a same case remain similar. Our synthesized database is not comparable to real ones of similar size in terms of diversity and richness.

We use only the slices below the base in synthesized volumes to train the spatial propagation network. For the initialization net we also use these slices except the top 1/6 and the bottom 1/6 of them. The potentially poor image quality of slices near base and apex may cause additional inaccuracies.

The networks are trained by stochastic gradient descent with batch size 1 and learning rate 0.0001. The initialization network is trained for 300000 iterations. It takes about 23 hours in total on GPU. The spatial propagation network is trained for 600000 iterations, which together take roughly 44 hours.

2.4.2 Testing.

The method is tested on the end-diastole (the only instant where ground truth is available) of testing cases. It takes about 3s to segment a volume using GPU. The

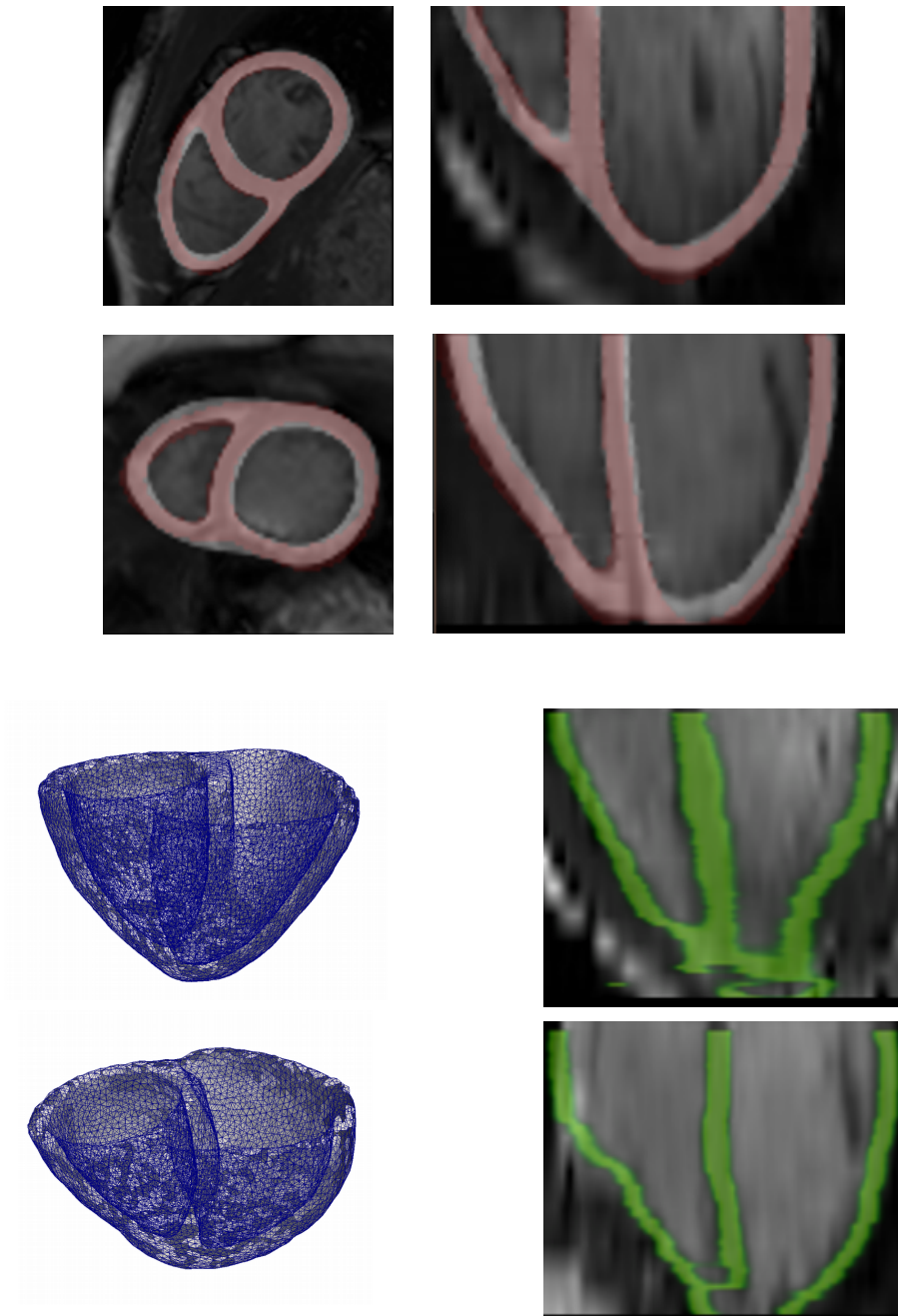


Figure 2.4: Upper: prediction (white) and ground truth (red) for the two tested cases from STACOM. Lower left: the generated meshes for the two tested cases from STACOM. Lower right: segmentation results (green) without application of propagation.

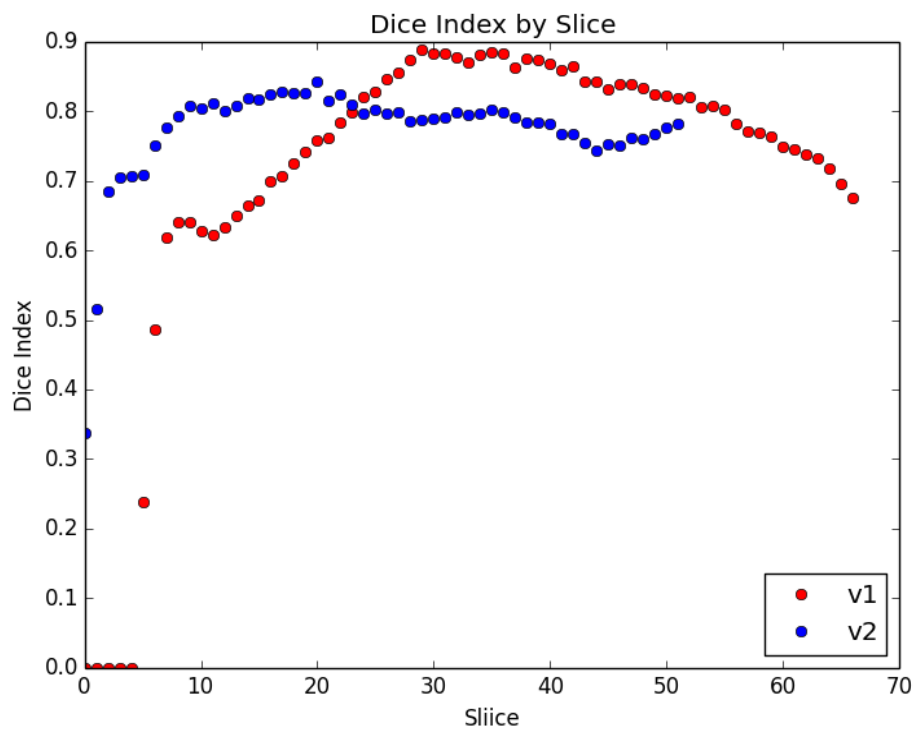


Figure 2.5: Slice-wise Dice Index on the test set.

output probabilities are binarized to obtain myocardium/background segmentation taking 0.5 as threshold. We use the Dice index to measure performance. The 3D Dice indices (considering all pixels of all slices below base) are 0.7851 for case v1 and 0.7817 for case v2. The predicted masks and the ground-truth (axial and coronal views) as well as the BV meshes generated directly from the predicted masks using CGAL¹ are presented in the lower left part of Figure 2.4. In Figure 2.5, 2D Dice indices for both subjects change smoothly across slices, confirming the spatial consistency of our method.

For comparison, if we use the initialization network to segment all the slices independently without propagation, the method not only loses the spatial consistency but also fails completely on the slices near apex (the lower right part of Figure 2.4). Our propagation method therefore appears crucial to maintain spatial consistency and reach accurate results even on the most difficult slices.

2.5 Conclusion and Perspectives

We demonstrate that our deep-learning-based automatic method for BV segmentation is robust, and combines the assets of 2D (speed) and 3D to provide spatially consistent meshes ready to be used for simulations. Besides, we proposed two original networks: (i) the initialization network predicts segmentation in a multi-scale coarse-to-fine manner; (ii) the second network propagates segmentation with spatial consistency. A novel loss function is also proposed to overcome class imbalance. For training, we use image synthesis as data augmentation. Meshes of high quality are generated. In the future, we will explore the capacity of neural networks in maintaining temporal consistency of segmentation. As we only use 15 subjects in this paper, significant improvement is expected if more data are added afterwards.

¹<http://www.cgal.org>

Consistent and Robust Segmentation with Spatial Propagation

Contents

3.1	Introduction	18
3.2	Data	22
3.2.1	Datasets	22
3.2.2	Notation and Terminology	22
3.2.3	Adaptation of the UK Biobank Ground-Truth	23
3.3	Methods	24
3.3.1	Region of Interest (ROI) Determination: ROI-net	24
3.3.2	Segmentation with Propagation: LVRV-net and LV-net	27
3.3.3	Image Preprocessing	28
3.3.4	Loss Functions	29
3.4	Experiments and Results	30
3.4.1	Technical Details about Training the Three Networks	30
3.4.2	Experiments on UK Biobank & Contribution of the Propagation	30
3.4.3	Generalization Ability to Other Datasets	36
3.5	Conclusion and Discussion	45
3.6	Appendix	46
3.6.1	Datasets	46
3.6.2	Metrics	47

Part of this chapter corresponds to the following scientific article:

- [Zheng 2018b] **3D Consistent and Robust Segmentation of Cardiac Images by Deep Learning with Spatial Propagation**
Qiao Zheng, Hervé Delingette, Nicolas Duchateau and Nicholas Ayache. IEEE Transactions on Medical Imaging, 2018

3.1 Introduction

The manual segmentation of cardiac images is tedious and time-consuming, which is even more critical given the new availability of huge databases (e.g. UK Biobank [Petersen 2016]). Magnetic resonance imaging (MRI) is widely used by cardiologists. Yet MRI is challenging to segment due to its anisotropic resolution with somewhat distant 2D slices which might be misaligned. There is hence a great need for automated and accurate cardiac MRI segmentation methods.

In recent years, many state-of-the-art cardiac segmentation methods are based on deep learning and substantially overcome the performance of previous methods. Currently, they dominate various cardiac segmentation challenges. For instance, in the Automatic Cardiac Diagnosis Challenge¹ (ACDC) of MICCAI 2017, 9 out of the 10 cardiac segmentation methods were based on deep learning. In particular, the 8 best-ranked methods were all deep learning ones. Deep learning methods can be roughly divided into to 2 classes: 2D methods, which segment each slice independently (i.e.[Winther 2017], [Tran 2016], [Baumgartner 2017]), and 3D methods, which segment multiple slices together as a volume (i.e.[Isensee 2017], [Baumgartner 2017]). 2D methods are popular because they are lightweight and require less data for training. But as no 3D context is taken into consideration, they might hardly maintain the 3D-consistency between the segmentation of different slices, and even fail on “difficult” slices. For example, the 2D method used in [Tran 2016] achieves state-of-the-art segmentation on several widely used datasets but makes the most prominent errors in apical slices and even fails to detect the presence of the heart.

On the other hand, 3D methods should theoretically be robust to these issues. But in [Baumgartner 2017], with experiments on the ACDC dataset, the authors found that all the 2D approaches they proposed consistently outperformed the 3D method being considered. In fact, 3D methods have some significant shortcomings. First, using 3D data drastically reduces the number of training images. Second, 3D methods mostly rely on 3D convolution. Yet border effects from 3D convolution may compromise the information in intermediate representations of the neural networks. Third, 3D methods require far more GPU memory. Therefore, substantial downsampling of data is often necessary for training and prediction, which causes loss of information.

One possible way to combine the strengths of 2D and 3D methods is to use recurrent neural networks. In [Poudel 2016], the authors merge U-Net [Ronneberger 2015] and a recurrent unit into a neural network to process all slices in the same stack, arranging the slices from the base to the apex. Information from the slices already segmented in the stack is preserved in the recurrent unit and used as context while segmenting the current slice. Comparisons in [Poudel 2016] prove that this contextual information is helpful to achieve better segmentation. However, the approaches based on recurrent neural networks are still limited. On the one hand, as the slice

¹<https://www.creatis.insa-lyon.fr/Challenge/acdc/>. Accessed September 15 2017

thickness (usually 5 to 10mm) is often very large compared to the slice resolution (usually 1 to 2mm), the correlation between slices is low except for adjacent slices. Thus, considering all slices at once may not be optimal. On the other hand, the prediction on each slice made by a recurrent neural network does not depend on an existing prediction. With this setting, the automatic segmentation is remarkably different from the procedure of human experts. As presented in [Suinesiaputra 2015], human experts are very consistent in the sense that the intra-observer variability is low; yet the inter-observer variability is high, as segmentation bias varies remarkably between human experts. Hence in general, for given a slice, there is no unique correct segmentation. But human operators still maintain consistency in their predictions respectively. Being inspired by these facts, we adopt a novel perspective: we train our networks to explicitly maintain the consistency between the current segmentation and the already predicted segmentation on an adjacent slice. We do not assume that there is a unique correct segmentation. Instead, the prediction for the current slice explicitly depends on another previously predicted segmentation.

Another possible method to improve segmentation consistency is to incorporate anatomical prior knowledge into neural networks. In [Oktay 2018], the segmentation models are trained to follow the cardiac anatomical properties via a learned representation of the 3D shape. While adopting novel training procedure, this method is based on 3D convolution neural networks for segmentation. So the issues of 3D methods discussed above still exist.

In this paper, we propose a novel method based on deep learning to perform cardiac segmentation. Our main contribution is threefold:

- The spatial consistency in cardiac segmentation is barely addressed in general, while this is a remarkable aspect of human expertise. Our method explicitly provides spatially consistent results by propagating the segmentations across slices. This is a novel perspective, as we do not assume the existence of a unique correct segmentation, and the prediction of the current segmentation depends on the segmentation of the previous slice.
- After training our method with a large dataset, we demonstrate its robustness and generalization ability on a large number of unseen cases from the same cohort as well as from other reference databases. These aspects are crucial for the application of a segmentation model in general, yet have not yet been explored before.
- Most segmentation methods proceed in a 2D manner to benefit from more training samples and higher training speed in comparison with 3D methods. In contrast, we proposed an original approach that keeps the computational assets of 2D methods but still addresses key 3D issues.

We hence believe in its potential impact on the community².

²The code and the models are available in this repository: <https://github.com/julien-zheng/CardiacSegmentationPropagation>

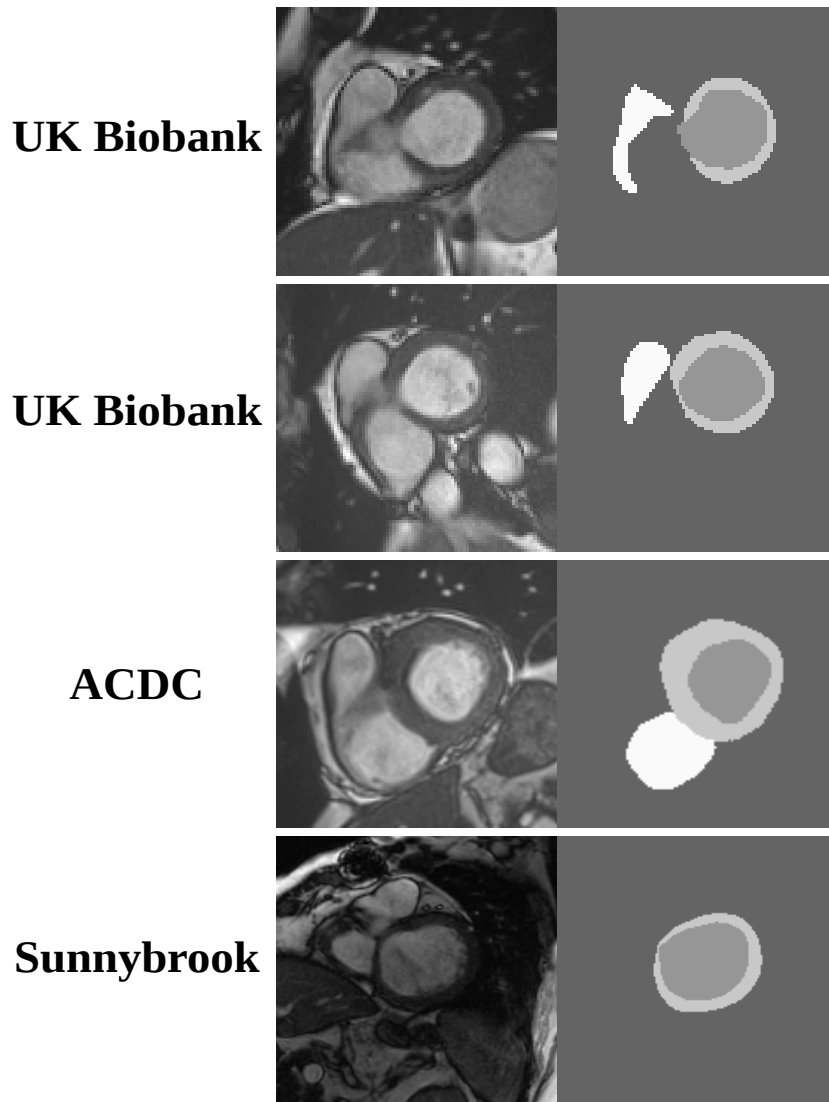


Figure 3.1: Intra- and inter-dataset inconsistencies of the basal slice ground-truth (RVSC contains no basal slice and is therefore not shown).

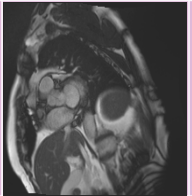
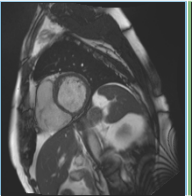
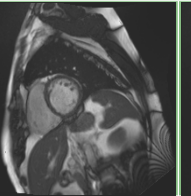
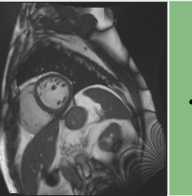

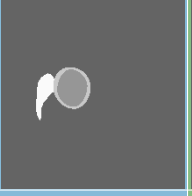
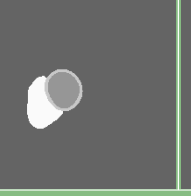
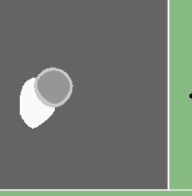
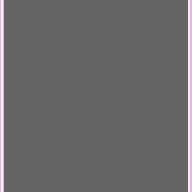
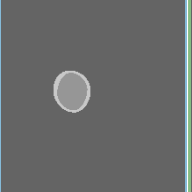
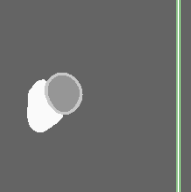

	Above the base		Basal slice	Below the base		
MRI slices in a stack (arranged in spatial order)	...					...
Original ground-truth	...					...
Adaptation	All labels are removed		Remove RVC	No operation		
Adapted ground-truth	...					...

Figure 3.2: Ground-truth adaptation proposed for UK Biobank. the basal slice is first identified (blue), then the RVC labels are removed in this slice, and the labels are removed from the slices above (pink).

3.2 Data

3.2.1 Datasets

The proposed method was trained using four datasets: the very large UK Biobank[Petersen 2016] dataset through our access application³, the ACDC challenge training dataset, the Sunnybrook dataset [Radau 2009] (made available for the MICCAI 2009 challenge on automated left ventricle (LV) segmentation), and the Right Ventricle Segmentation Challenge (RVSC) dataset [Petitjean 2015] (provided for the MICCAI 2012 challenge on automated right ventricle (RV) segmentation). Depending on the dataset, expert manual segmentation for different cardiac structures (e.g. the left and right ventricular cavity (LVC, RVC), the left ventricular myocardium (LVM)) is provided as ground-truth for all slices at end-diastole (ED) and/or end-systole (ES) phases. All other structures in the image are considered as background (BG). Training involved a subset (80%) of the UK Biobank dataset. Testing used the remaining 20% from the same dataset, as well as the whole three other datasets. Details about these datasets are provided in the appendix. We mainly adopt the metrics used in the three challenges above to measure segmentation performance. The exact definitions of the metrics used in this paper (e.g. Dice index, Hausdorff distance, presence rate) are provided in the appendix.

3.2.2 Notation and Terminology

In this paper, slices in image stacks are indexed in spatial order from the basal to the apical part of the heart. Given an image stack S , we denote N the number of its slices. Given two values a and b between 0 and N , we note $S[a, b]$ the sub-stack consisting of slices of indexes in the interval $[round(a), round(b)[$ ($round(a)$ is included while $round(b)$ is excluded) with $round$ the function rounding to nearest integer. For instance, if S is a stack of $N=10$ slices of indexes from 0 to 9, then $S[0.2N, 0.6N]$ is the stack consisting of slices number 2 to 5. Similarly, if the basal slice is defined in S , we denote $base$ its index. Then $S[base]$ and $S[base+1]$ are the basal slice and the first slice below the base.

Segmentation of slices above and below the base of the heart can be quite different. For convenience, in a stack with known base slice, we call the slices located above it the AB (above-the-base) slices, and the ones located below it BB (below-the-base) slices. In the remainder of this paper, we propose methods to determine the base slice for image stacks of UK Biobank using the provided ground-truth.

Finally, given a segmentation mask M , $edge(LVC, LVM)$ is the number of pairs of neighboring pixels (two pixels sharing an edge, defined using the 4-connectivity) on M such that one is labeled to LVM while the other is to LVC. Similarly we define $edge(LVC, BG)$ and $edge(LVC, RVC)$.

³Application Number 2964.

3.2.3 Adaptation of the UK Biobank Ground-Truth

Let's first compare the segmentation conventions followed by the ground-truth between datasets. For BB slices, the convention is roughly the same: if LV is segmented, LVC is well enclosed in LVM; if RVC is segmented, it is identified as the whole cardiac cavity zone next to the LV. But for AB slices, the variability of segmentation conventions within and between datasets can be significant. In Figure 3.1, we show examples of (base slice, ground-truth) pairs from UK Biobank (row-1 and row-2, two different cases), ACDC (row-3) and Sunnybrook (row-4). For better visualization, we crop out the heart regions from the original MRI images and ground-truths accordingly. The segmentation ground-truth on these similar images are significantly different. In particular, we notice the intra- and inter-dataset inconsistencies in the segmentation of (1) the RVC at the outflow tract level, (2) the LVM and LVC at the mitral valve level (some dataset seems to be segmented in a way such that the LVC mask is always fully surrounded by the LVM mask). In contrast, the convention seems roughly the same for the BB slices.

Hence we decided to adapt the ground-truth of UK Biobank to improve both consistency and generality. As presented in Figure 3.2, we i) set all pixels in all the slices above the base to BG; ii) relabel all the pixels in the basal slice originally labeled as RVC to BG while keeping the LVC and LVM pixels unchanged; iii) keep the ground-truth of all slices below the base unchanged.

Moreover, we propose a method to determine the basal slice automatically in the stacks of UK Biobank. While checking the ground-truth of the slices starting from the apex part, the basal slice is determined as the first one such that:

- the LVC mask is not fully surrounded by the LVM mask:

$$edge(LVC, BG) + edge(LVC, RVC) > 0 \quad (3.1)$$

- or the area of the RVC mask shrinks substantially comparing to that of the slice below:

$$\begin{cases} overlap(RVC_1, RVC_2)/RVC_2 \leq T_1 \\ RVC_1/RVC_2 \leq T_2 \end{cases} \quad (3.2)$$

$$\quad (3.3)$$

with RVC_1 and RVC_2 the RVC masks of the slice and the slice below it respectively, $T_1=0.75$ and $T_2=0.8$ thresholds. If the basal slice is not determined after examining all slices in the stack, we define that the index of the base slice is -1 (so $S[base+1]$ is the first slice in the stack).

According to the current international guidelines in [Schulz-Menger 2013], the “standard” basal slice is the topmost short-axis view slice that has more than 50% myocardium around the blood cavity. To test whether the UK Biobank basal slices determined above are close to the standard basal slices, we randomly picked 50 cases (50 ED stacks + 50 ES stacks) and estimated their standard basal slices at ED and ES visually according to the guidelines. Among the 100 pairs of standard basal slice and ground-truth-deduced basal slice, 59 pairs are exactly the same, 40 pairs are 1-slice away in stack, and only 1 pair is 2-slice away. The “adapted” ground-truth will stand as the ground-truth for the rest of this paper.

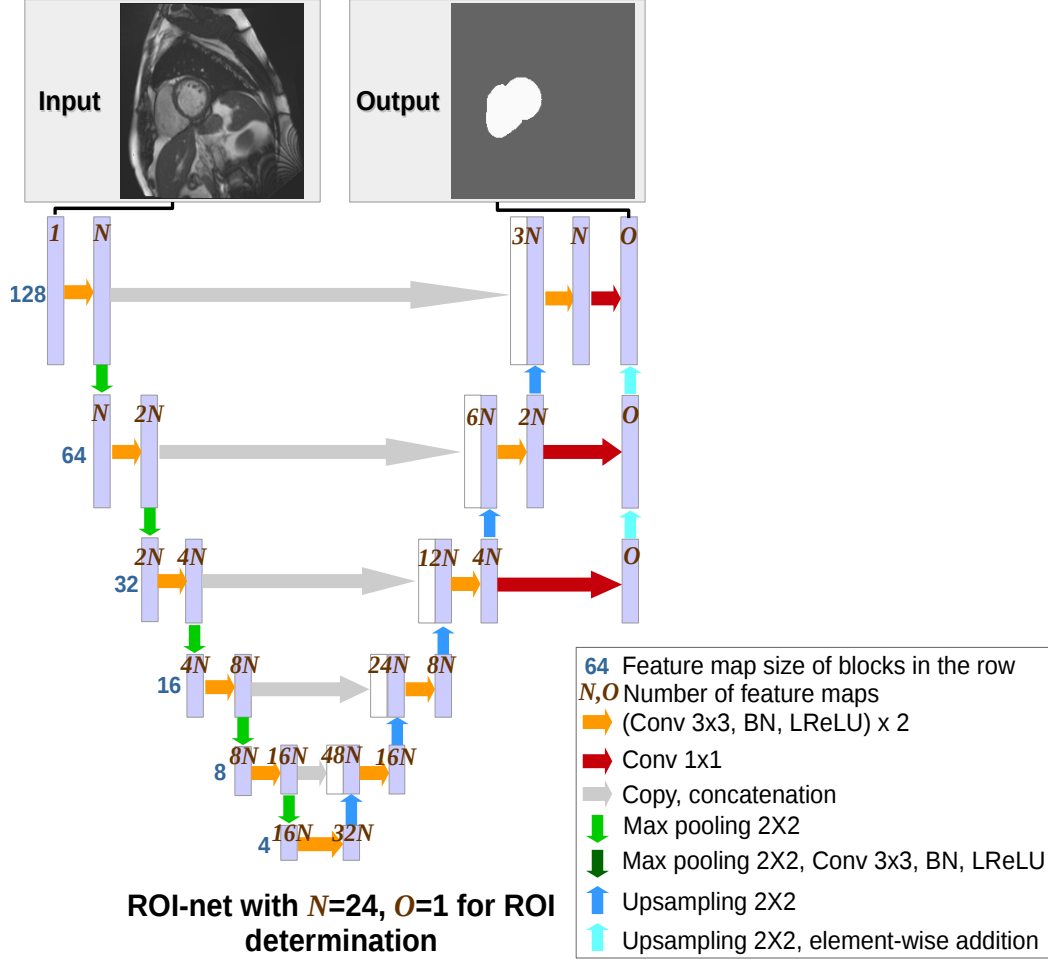


Figure 3.3: ROI-net: for ROI determination over image stack. A sigmoid function is applied to the output channel to generate pixel-wise probabilities.

3.3 Methods

Our method mainly consists of two steps: region of interest (ROI) determination and segmentation with propagation. The first step is either based on a trained neural network (the ROI-net) or on center cropping, depending on the dataset. The second step is based on either the LVRV-net or the LV-net (originally designed by us and inspired from U-net [Ronneberger 2015]), depending on whether the RVC must be segmented. This section will also present the image preprocessing methods and the loss functions we used.

3.3.1 Region of Interest (ROI) Determination: ROI-net

On cardiac MRI images, defining an ROI is useful to save memory usage and to increase the speed of segmentation methods. There are many different ROI deter-

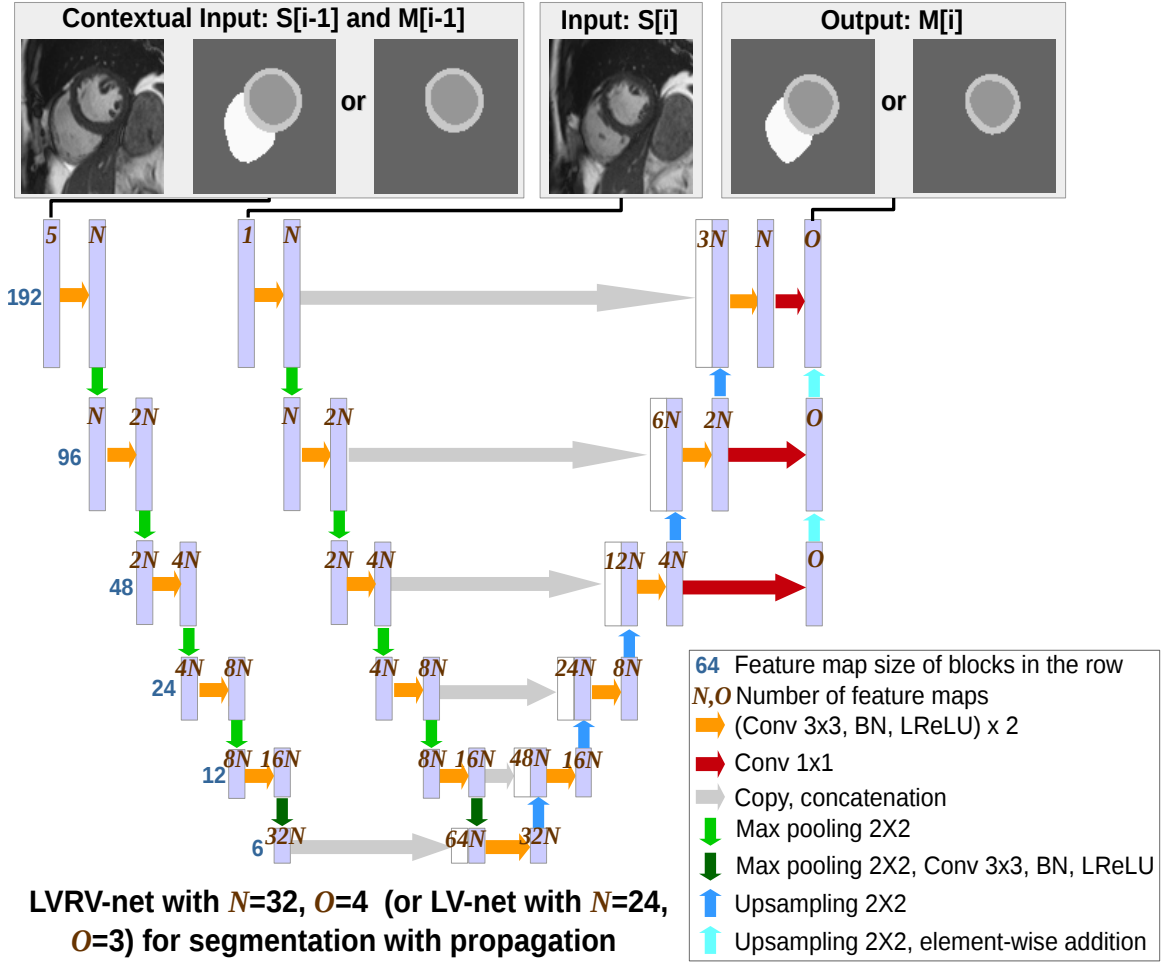


Figure 3.4: LVRV-net and LV-net: for cardiac segmentation on ROIs. $S[i]$ is the slice to be segmented and $M[i]$ is the predicted mask. A softmax function is applied to the output channels to generate pixel-wise 4- or 3-class probabilities.

mination methods available in the community. But for most of them, the robustness remains a question, as the training and the evaluation are done with cases from the same cohort of limited size. We propose a robust approach as follows. With a large number of available cases from UK Biobank, a deep learning based method becomes a natural choice. In particular, we design the ROI-net (Figure 3.3) to perform heart segmentation on MRI images.

Notice that for some datasets (Sunnybrook and RVSC), the images are already centered around the heart. Similar to what was done in [Tran 2016], in such cases, images are simply cropped. However this is not valid for most datasets (here UK Biobank and ACDC), and an ROI needs to be determined specifically for each stack based on the predictions of ROI-net, as explained in the following. ROI-net is a variant of U-net with a combination of convolutions, batch normalizations (BN) and leaky ReLUs (LReLU) [Maas 2013] as building blocks. In leaky ReLU the gradient parameter when the unit is not active is set to 0.1. Furthermore, we implement deep supervision as in [Kayalibay 2017] to generate low resolution (of size 32 and 64) segmentation outputs, and then upsample and add them to the final segmentation. A sigmoid function is applied to the output channel of ROI-net to generate pixel-wise probabilities.

In brief, ROI-net takes one original MRI image as input and predicts pixel-wise probabilities as a way of heart/background binary segmentation (0 for background, 1 for the heart, and the threshold is 0.5 in inference). The heart to be segmented is defined as the union of LVC, LVM, and RVC. The ROI determination takes only the ED stack slices into account. In practice, an ROI containing the heart with some margin at ED also contains well the heart at other instants including ES.

3.3.1.1 Training

The network is trained with slices in $S[(base+1), (base+1)+0.4N]$ (the 40% of slices just below the base) of the ED stack S from the UK Biobank training cases. The purpose of using only slices in $S[(base+1), (base+1)+0.4N]$ is to avoid the top slices around the base on which RVC ground-truth shrinks (Figure 3.1), and the bottom slices around the apex on which the heart is small and almost does not affect the ROI determination.

3.3.1.2 Prediction

To confirm the robustness of ROI-net for inference, we apply it to the sub-stacks roughly covering the largest cross-section of the hearts in a dataset (the position of the base is supposed to be unknown for individual cases). The slice indexes of these sub-stacks are determined based on visual observation for a given dataset. More specifically, the trained ROI-net is used to segment slices in $S[0.2N, 0.6N]$ of the ED stack S of all the UK Biobank cases, and slices in $S[0.1N, 0.5N]$ of the ED stack S of all the ACDC cases. For noise reduction and as post-processing for the ROI net, for each image, only the largest connected component of the output heart mask is kept for prediction.

3.3.1.3 ROI Determination

For each ED stack, the union of all predicted heart masks, as well as the minimum square M covering their union, is determined. We add to it a padding of width 0.3 times the size of M to generate a larger square bounding box, which is defined as the ROI for the case.

After ROI determination on an ED stack, the same ROI applies to both the ED and ES stacks of the same case. Then the ED and ES stacks are cropped out according to this ROI and used as inputs for the LVRV-net and the LV-net in the second step. Hence in the remainder of this paper, we refer to the cropped version of the images, slices or stacks.

3.3.2 Segmentation with Propagation: LVRV-net and LV-net

The second step is segmenting the cropped images (the ROIs). Depending on whether we segment RVC or not, we proposed two networks: LVRV-net and LV-net. They share the same structure template as depicted in Figure 3.4. Both perform slice segmentation of $S[i]$ taking $S[i-1]$, the adjacent slice above, and $M[i-1]$ its segmentation mask, as contextual input. In the contextual input, there are five channels in total: $S[i-1]$ takes one, while $M[i-1]$, being converted to pixel-wise one-hot channels (BG, LVC, LVM, RVC), takes four. In case $S[i]$ is the first slice to be segmented in a stack, $M[i-1]$ does not exist and is set to a null mask; in case $S[i]$ is the top slice in a stack, $S[i-1]$ does not exist and is set to a null image. The main body of LVRV-net and LV-net is also a variant of U-net with convolution, BN, LReLU and deep supervision, very similar to that of ROI-net. In addition to the main body, an extra encoding branch encodes the contextual input. Information extracted by the main body encoding branch and the extra encoding branch are combined at the bottom of the network, before being decoded in the decoding branch. Finally, a softmax function is applied to the output channels to generate pixel-wise 4- or 3-class probabilities. For inference, each pixel is labeled to the class with the highest probability.

3.3.2.1 Training

LVRV-net and LV-net are trained to segment slices $S[i]$ in $S[(base+1), N]$ (the BB slices, the green column in Figure 3.2) and $S[base, N]$ (the basal slice and the BB slices, the blue column and the green columns in Figure 3.2) respectively of the stack S at ED and ES of the UK Biobank training set. Regarding the contextual input, $S[i-1]$ and $M[i-1]$ are set to a null image or a null mask if they are not available as described above; otherwise $M[i-1]$ is set to the corresponding ground-truth mask.

3.3.2.2 Testing

The trained LVRV-net and LV-net are used to segment the cases in the UK Biobank testing set and the other datasets (ACDC, Sunnybrook, RVSC). Let us note S' the

sub-stack to be segmented and M' the corresponding predicted mask stack. Notice that for UK Biobank, S' is $S[(base+1), N]$ for LVRV-net, and $S[base, N]$ for LV-net; for the other datasets, S' is the whole stack. LVRV-net or LV-net iteratively segments $S'[i]$ by predicting $M'[i]$, taking $S'[i-1]$ and $M'[i-1]$ as contextual input, for $i = 0, 1, 2$, etc.. In other words, the segmentation prediction of a slice is used as contextual information while segmenting the slice immediately below it in the next iteration. The segmentation prediction is iteratively “propagated” from top to bottom (or roughly speaking from base to apex) slices in S' .

3.3.2.3 Post-processing

We post-process the predictions at each iteration while segmenting a stack (hence the post-processed mask will be used as the contextual mask in the next iteration if it exists). A predicted mask is considered as successful if the two conditions below are satisfied:

- LVM is present on the mask;
- LVC is mostly surrounded by LVM:

$$\begin{aligned} & (edge(LVC, BG) + edge(LVC, RVC)) \\ & \leq 0.5 \times edge(LVC, LVM) \end{aligned} \tag{3.4}$$

The parameter 0.5 above is determined empirically. If the predicted mask is successful, for LVRV-net only, we further process the mask by preserving only the largest connected component of RVC and turning all the other RVC connected components (if they exist) to background; otherwise, the predicted mask is reset to a null mask.

3.3.3 Image Preprocessing

Each input image or mask of the three networks in this paper (ROI-net, LVRV-net, and LV-net) is preprocessed as follows:

3.3.3.1 Extreme Pixel Value Cutting and Contrast Limited Adaptive Histogram Equalization (CLAHE) for ROI-net only

Input images to ROI-net are thresholded to the 5th and 95th percentiles of gray levels. Then we apply CLAHE as implemented in OpenCV⁴ to perform histogram equalization and improve the contrast of the image with the parameters *clipLimit* = 3 and *tileGridSize* = (8, 8).

3.3.3.2 Padding to Square and Resize

The input image or mask is zero-padded to a square if needed. Then it is resampled using nearest-neighbor interpolation to 128×128 for ROI-net or 192×192 for LVRV-net and LV-net.

⁴https://docs.opencv.org/3.1.0/d5/daf/tutorial_py_histogram_equalization.html

3.3.3.3 Normalization

Finally, for each input image of all networks, the mean and standard deviation of the slice intensity histogram cropped between the 5th and 95th percentiles are computed. The image is then normalized by subtracting this mean and dividing by this standard deviation.

3.3.4 Loss Functions

We use the two Dice loss (DL) functions below to train the three neural networks mentioned above. As suggested in [Wolterink 2017], loss functions based on Dice index help overcoming difficulties in training caused by class imbalance.

3.3.4.1 DL_1 for ROI-net Training

Given an input image I of N pixels, let's note p_n the pixel-wise probability predicted by ROI-net and g_n the pixel-wise ground-truth value (g_n is either 0 or 1). DL_1 is defined as

$$DL_1 = -\frac{2 \sum_{n=1}^N p_n g_n + \varepsilon}{\sum_{n=1}^N p_n + \sum_{n=1}^N g_n + \varepsilon} \quad (3.5)$$

where ε is used to improve the training stability by avoiding division by 0, i.e. when p_n and g_n are 0 for each pixel n . Empirically we take $\varepsilon = 1$. The value of DL_1 varies between 0 and -1. Good performance of ROI-net corresponds to DL_1 close to -1.

3.3.4.2 DL_2 for LVRV-net Training

For the segmentation of a N -pixel input image, the outputs are four probabilities $p_{n,c}$ with $c = 0, 1, 2, 3$ (BG, LVC, LVM and RVC) such that $\sum_c p_{n,c} = 1$ for each pixel. Let's note $g_{n,c}$ the corresponding one-hot ground-truth ($g_{n,c}$ is 1 if the pixel is labeled with the class corresponding to c ; otherwise $g_{n,c}$ is 0). Then we define

$$DL_2 = -\frac{1}{4} \sum_{c=0}^3 \left(\frac{2 \sum_{n=1}^N p_{n,c} g_{n,c} + \varepsilon}{\sum_{n=1}^N p_{n,c} + \sum_{n=1}^N g_{n,c} + \varepsilon} \right) \quad (3.6)$$

The role of ε here is similar to that in DL_1 . Empirically we use $\varepsilon = 1$.

3.3.4.3 DL_3 for LV-net Training

Its formula is very similar to that of DL_2 . The only difference is, instead of calculating the average of the 4 Dice index terms with c ranges from 0 to 3, DL_3 sums up the 3 Dice index terms with c ranges from 0 to 2 (BG, LVC, LVM) and computes their average.

3.4 Experiments and Results

The three networks (ROI-net, LVRV-net, LV-net) are implemented using TensorFlow⁵ and trained with 3078 UK Biobank cases as described in the “Methods” section. Then they are applied to the other datasets (ACDC, Sunnybrook, RVSC) without any fine-tuning or further training.

3.4.1 Technical Details about Training the Three Networks

ROI-net is trained for 50 epochs and applied to these cases to determine the ROIs. The cropped ROI volumes are then used to train LVRV-net and LV-net for 80 epochs. For each of the three networks, weights are initialized randomly, Adam optimizer is used with initial learning rate 0.0001, batch size is set to 16, and data augmentation is applied (the input images are randomly rotated, shifted and zoomed in/out along the row/column dimension independently, flipped horizontally and flipped vertically).

3.4.2 Experiments on UK Biobank & Contribution of the Propagation

The three trained networks are evaluated on the 756 evaluation cases of UK Biobank.

3.4.2.1 ROI Determination by ROI-net

The trained ROI-net is applied to determine and crop ROIs (prediction on the ED sub-stack $S[0.2N, 0.6N]$, the minimum square to cover the union of the predicted masks in the sub-stack, etc.). For all the cases, the determined ROI is successful in the sense that the heart (defined as the union of the pixels labeled to LVC, LVM or RVC in the ground-truth) is fully located inside the ROI, at both ED and ES. Furthermore, all the ROIs are small: the heart and the ROI are distant from 18 ± 3 pixels in average, for image and ROI sizes of 209 ± 1 and 91 ± 8 pixels respectively.

3.4.2.2 LVRV-net and LV-net

We report the segmentation performance in terms of Dice index and Hausdorff distance in Table 3.1. The mean values are reported along with the standard deviation in parentheses. LV-epi is defined as the union of the LVC and LVM.

We notice that the Dice index of the LVM is significantly lower than that of the other parts. We believe that this is partly due to the variability of the ground-truth in UK Biobank as presented in Figure 3.5. This kind of variability influences both the learning and the evaluation of our method. The Dice index of LVM is most heavily affected. Indeed, on the one hand, LVM is more difficult to segment than LVC due to its shape. Ambiguity on the ground-truth makes the learning of the LVM segmentation even harder. On the other hand, LVM represents a small volume.

⁵<https://www.tensorflow.org/>

Table 3.1: Segmentation Results (Mean and Standard Deviation) on the UK Biobank Testing Cases

	Dice				Hausdorff (mm)			
	LVM	LVC	LV-epi	RVC	LVM	LVC	LV-epi	RVC
proposed	0.769	0.903	0.932	0.881	7.66	5.94	7.13	10.39
LVRV-net	(0.06)	(0.03)	(0.01)	(0.04)	(4.55)	(2.26)	(4.32)	(4.71)
LVRV-mid-starting-net	0.767	0.904	0.931	0.886	8.96	5.87	8.46	9.90
	(0.06)	(0.03)	(0.01)	(0.04)	(10.94)	(2.90)	(10.98)	(4.01)
LVRV-no-propagation-net	0.793	0.915	0.939	0.896	9.86	6.66	9.40	10.32
	(0.05)	(0.03)	(0.01)	(0.03)	(12.03)	(7.74)	(12.03)	(5.32)
proposed	0.752	0.896	0.923	-	9.78	6.97	8.72	-
LV-net	(0.06)	(0.04)	(0.02)	(-)	(9.22)	(3.43)	(9.22)	(-)

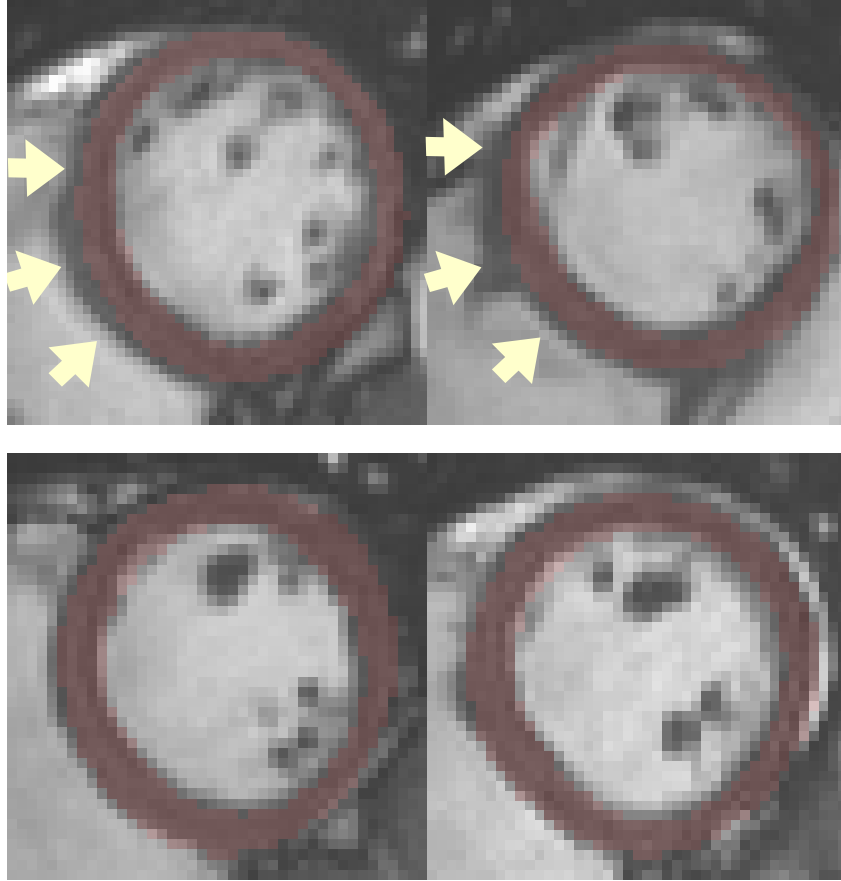


Figure 3.5: UK Biobank ground-truth variability: These slices are extracted from 4 different cases in UK Biobank. Compared to the ground-truth of the slices in the second row, the ground-truth of the slices in the first row clearly under-segments the portion of myocardium between LV and RV (indicated by the arrows).

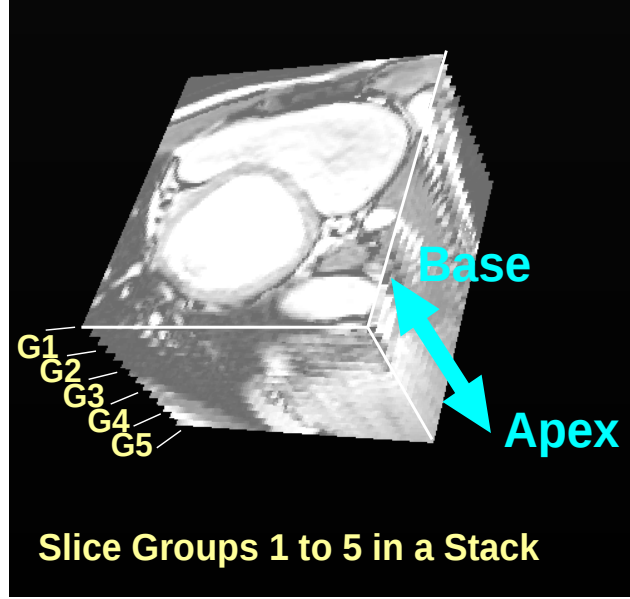


Figure 3.6: An example of slice group division (G1 to G5) in a stack.

The Dice index is hence more sensitive to errors in this structure. In general, not only for LVM, the variability in UK Biobank ground-truth reduces the performance for all structures in terms of Dice index. In contrast, the Hausdorff distance is much less sensitive to this variability, which also explains the better performance of our model.

Notice that the results reported in Table 3.1 are based on 3D volumes. To evaluate the performance of LVRV-net across different slices, given a structure (e.g. LVM), we also provide results for 5 evenly distributed levels from the slice $S[base+1]$ to the last slice on which the structure is present (Figure 3.6). Group 1 (G1) is on top of the sub-stack and close to the base. Group 5 (G5) is close to the apex. Then we evaluate the segmentation performance of LVRV-net in terms of heart (defined as the union of LVC, LVM, and RVC) presence rate (Figure 3.7), and 2D Hausdorff distance for 4 different structures (Figure 3.8).

The evaluation results of LV-net are reported in Table 3.1. Note that although the results of LVRV-net and LV-net are in the same table, LV-net is applied to the basal slice $S[base]$ (the adapted ground-truth of which has no RVC mask) while LVRV-net is not. The higher ground-truth variability on $S[base]$, what we observe in UK Biobank, may explain the slightly lower performance measures of LV-net.

We notice that in general, the performances of the networks are better on ED stacks than on ES stacks. Since the heart is larger at ED than at ES, maybe it is also easier to be segmented at ED.

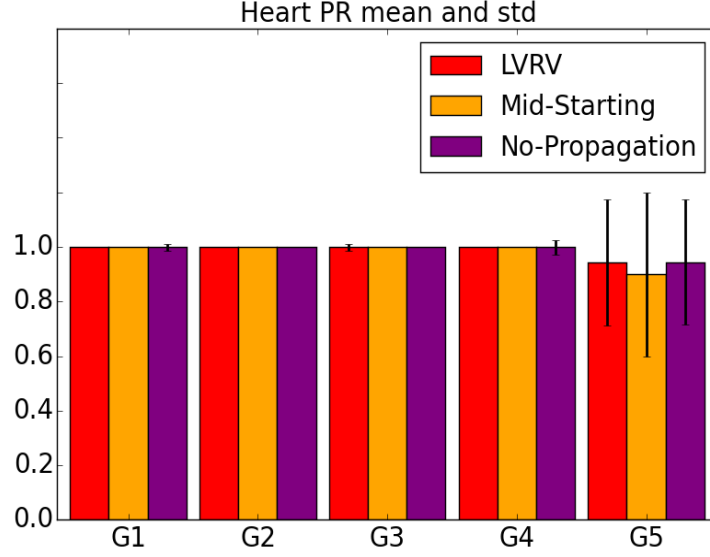


Figure 3.7: Performance measured by heart presence rate (PR) of the LVRV-net, the LVRV-mid-starting-net and the LVRV-no-propagation-net on UK Biobank.

3.4.2.3 LVRV-net vs. Its Variants: Justification of the Top-Starting-Propagation Procedure

To justify our designs of propagation and of starting propagation from the top slice in the proposed method, we compare LVRV-net with two variants of it, which are considered as baselines. The first baseline is the LVRV-no-propagation-net. Its structure is obtained by removing the extra propagation branch from LVRV-net. So LVRV-no-propagation-net takes an image as its only input and outputs the predicted segmentation mask. LVRV-no-propagation-net is trained and evaluated in the same way as LVRV-net. The evaluation results are reported in Table 3.1, Figure 3.7 and Figure 3.8. Another baseline is the LVRV-mid-starting-net. Its structure is identical to that of LVRV-net. But it is trained and then evaluated to segment the middle slice (determined from slice index) in $S[(base+1), N]$ with a null contextual input mask, and to propagate the segmentation results upward to the top and down to the bottom of $S[(base+1), N]$ using the prediction of the already segmented adjacent slice as the contextual input mask. The results are reported in Table 3.1, Figure 3.7 and Figure 3.8.

In Table 3.1, we can see that in terms of Dice index, LVRV-net and LVRV-mid-starting-net are almost the same while LVRV-no-propagation-net is slightly (0.01 to 0.02) higher. Yet in terms of Hausdorff distance, LVRV-net is clearly the best with low values of both mean and standard deviation. Regarding the PR by groups in Figure 3.7, we find that LVRV-net and LVRV-no-propagation-net detect the presence of the heart slightly better than LVRV-mid-starting-net in G5. In Figure 3.8 we can see that the differences on mean values of Hausdorff distance are pretty small

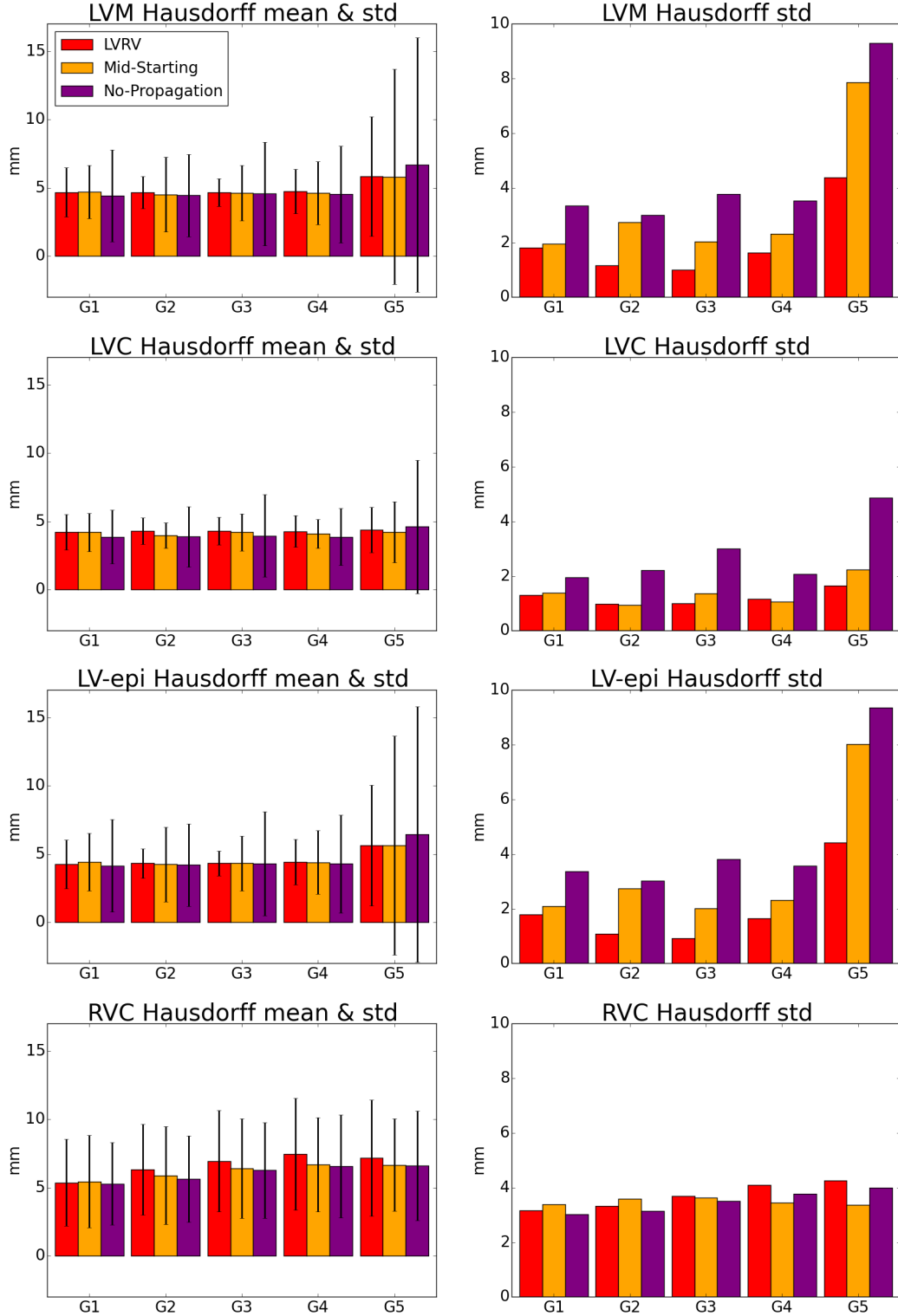


Figure 3.8: Performance measured by Hausdorff distance of the LVRV-net, the LVRV-mid-starting-net and the LVRV-no-propagation-net on UK Biobank. The first column indicates both the mean and the standard deviation values, while the second column depicts the standard deviation values only. The four rows stand for LVM, LVC, LV-epi and RVC respectively.

(within 1mm) for the three networks; but on standard deviation, especially for the LV structures, LVRV-net largely outperforms its variants, sometimes by several mm. Furthermore, we performed the Mann-Whitney U test to prove that the contribution of the propagation is statistically significant. Under the null hypothesis that the LVRV-net and LVRV-no-propagation-net predictions have the same distribution in terms of 3D Hausdorff distance, with the results on the UK Biobank testing set as samples, we obtain p -values of <0.001 , <0.001 , 0.001 , and 0.042 for the LVC, LVM, LV-epi, and RVC respectively, which are small enough (≤ 0.05) to conclude on the significance of the results. LVRV-net is clearly more robust than its variants.

To better understand the role of propagation as well as the robustness achieved by the LVRV-net, we look at the cases for which different methods have extremely contrasting performances, and define that the LVRV-no-propagation-net fails while the LVRV-net succeeds on a stack, if the latter outperforms the former on Hausdorff distance by a large value S , for any of the 4 structures (LVM, LVC, LV-epi, and RVC). And vice versa. For illustration, we use $S=30\text{mm}$, but similar interpretations can stand for other values of S . In Figure 3.9 to 3.11, we present three typical examples out of the 73 stacks in the UK Biobank testing set for which the LVRV-no-propagation-net fails while the LVRV-net succeeds. In the first example, on the one hand, the apex is so faint on the apical slice that it is barely possible to determine its size precisely. The ground-truth apex seems to be somewhat too large while the LVRV-net prediction looks a little bit too small (in a way learned from the training set with ground-truth variability). But the LVRV-net prediction still well determines the location of the apex using the contextual information. On the other hand, there is a structure on the slice of appearance very similar to the heart. The LVRV-no-propagation-net is confused by it and hence makes a completely wrong prediction. If we reconstruct the anatomical mesh of the heart based on the segmentation (to overcome the problem of large slice thickness, we apply interpolation to generate the segmentation on the intermediate slices between two adjacent slices), the mesh reconstructed from LVRV-no-propagation-net is clearly wrong on the apex. Similarly, the LVRV-no-propagation-net misses the right structure and/or makes a false positive prediction. In contrary, LVRV-net fails while LVRV-no-propagation-net succeeds on only 10 stacks. For 7 of them, LVRV-net predicts a tiny false positive component on a slice either below the apex or around the base. These failures may be simply fixed via the removal of all but the largest connected components. For the other 3 stacks, the errors are caused by image quality problems including large artifact on image and serious misalignment between adjacent slices.

The authors of [Bai 2017b] propose a method achieving human-level MRI analysis on UK Biobank. They aim at segmenting as accurately as possible each slice, in contrast with our method, which focuses on the consistency of segmentation across slices. Though the results of their method and that of ours are not directly comparable due to the differences on metrics (e.g. 2D Hausdorff Distance vs 3D Hausdorff Distance), training/testing datasets, preprocessing methods, etc., [Bai 2017b] inspired us to compare the performance of our method with that of human experts

Table 3.2: Segmentation Results on the ACDC Dataset, Compared to the Performance from the State-of-the-art Methods

	Dice				Hausdorff (mm)			
	LVM		LVC		LVM		LVC	
	mean	std	mean	std	mean	std	mean	std
proposed LV-net	0.715	0.07	0.862	0.08	9.76	3.31	8.74	3.76
Isensee et al. [Isensee 2017]	0.873	-	0.930	-	9.668	-	8.416	-
Jang et al. [Jang 2017]	0.879	0.04	0.938	0.05	9.76	6.02	7.27	4.83
Wolterink et al. [Wolterink 2017]	0.87	0.04	0.93	0.05	11.31	5.62	8.68	4.51

in terms of 3D consistency, which is the main focus of our method. In Figure 3.9, Figure 3.10 and Figure 3.11, for each example, we present a slice of the long-axis view (the last row) for both meshes reconstructed from the ground-truth and the LVRV-net prediction. As indicated by the arrows, with qualitative comparison we find that among these pairs of meshes the ground-truth reconstruction meshes are less regular and less smooth. It suggests that our method maintains 3D consistency even better than human experts.

3.4.3 Generalization Ability to Other Datasets

All the 3 trained networks are applied to the other 3 datasets without finetuning for two reasons. First, we do so to demonstrate their strong generalization ability. Second, as the 3 networks are designed to be big to learn from the large UK Biobank dataset of thousands of cases, finetuning them on small datasets of tens of cases easily results in overfitting. In fact, we have tried to finetune LVRV-net on ACDC. While a certain level (e.g. 10 epochs) of finetuning is beneficial, overfitting happens very soon afterward (obviously since the 50th epoch).

3.4.3.1 Experiments on ACDC

The trained ROI-net is applied to the ED sub-stacks $S[0.1N, 0.5N]$ of the 100 ACDC cases. Again as we found in the experiments on UK Biobank, the ROI determination is successful on 100% of the cases, as all the ROIs contain the heart completely on the one hand, and are very reasonably small on the other hand.

As we pointed out in the “Data” section, the RVC is segmented in ACDC with conventions quite different from that of UK Biobank. So we only try to segment the LV with the trained LV-net. Some slices to be segmented in ACDC are located well above the base. They are quite different from all the slices used to train LV-net so LV-net can predict some false positives. To deal with this challenge, for the application on ACDC only, we add three more points to the LV-net postprocessing

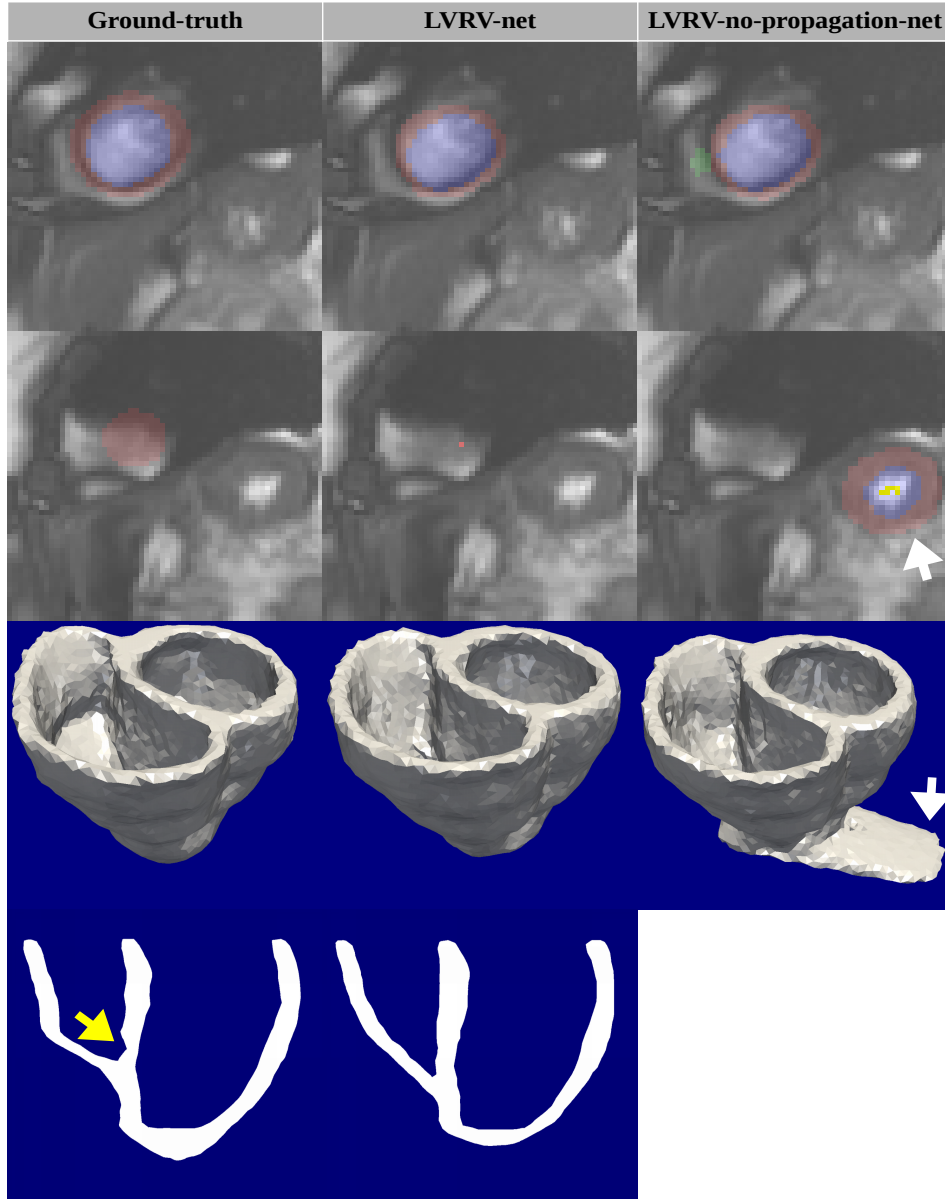


Figure 3.9: An example of the segmentation on difficult slices (zoomed-in versions of ROIs for better visualization) and the reconstructed meshes with the ground-truth, the prediction of LVRV-net and that of the LVRV-no-propagation-net. The last row shows a slice of the long-axis view of the meshes reconstructed with the ground-truth and the LVRV-net prediction (irregularities of the ground-truth reconstruction meshes are indicated by the arrows). The large-spread abnormal structures on the meshes in the third column are due to the interpolation of the wrong segmentation (indicated by the arrows). The first two rows are the segmentation on the last two slices of the stack. The apex is faint and there is another structure very similar to the heart. The LVRV-net correctly predicts the location of the apex, while the LVRV-no-propagation-net prediction is completely wrong.

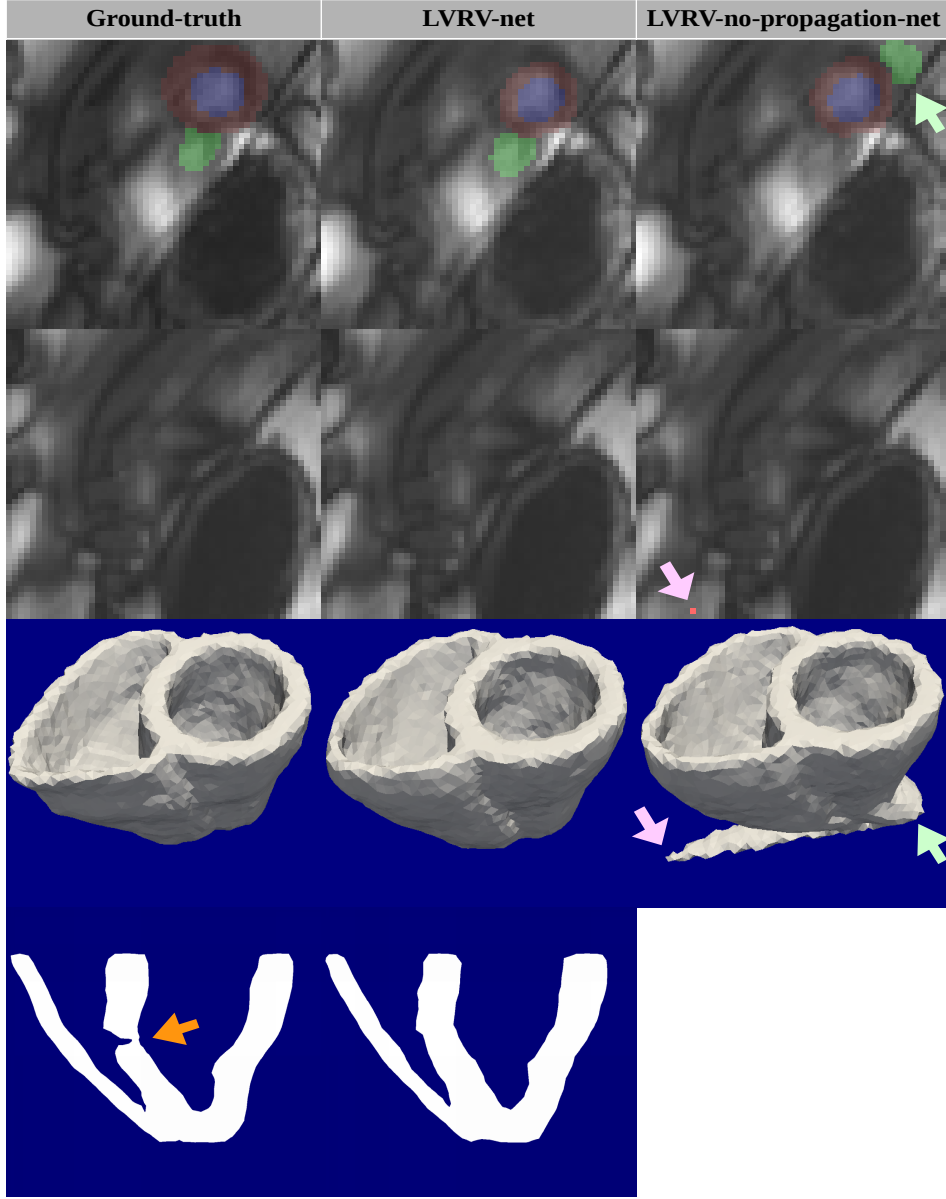


Figure 3.10: An example of the segmentation on difficult slices (zoomed-in versions of ROIs for better visualization) and the reconstructed meshes with the ground-truth, the prediction of LVRV-net and that of the LVRV-no-propagation-net. The last row shows a slice of the long-axis view of the meshes reconstructed with the ground-truth and the LVRV-net prediction (irregularities of the ground-truth reconstruction meshes are indicated by the arrows). The large-spread abnormal structures on the meshes in the third column are due to the interpolation of the wrong segmentation (indicated by the arrows). The first two rows are the segmentation on the last two slices of the stack. The LVRV-no-propagation-net predicts RVC incorrectly on the slice just above the apex and makes a false positive prediction of LVM on the other slice.

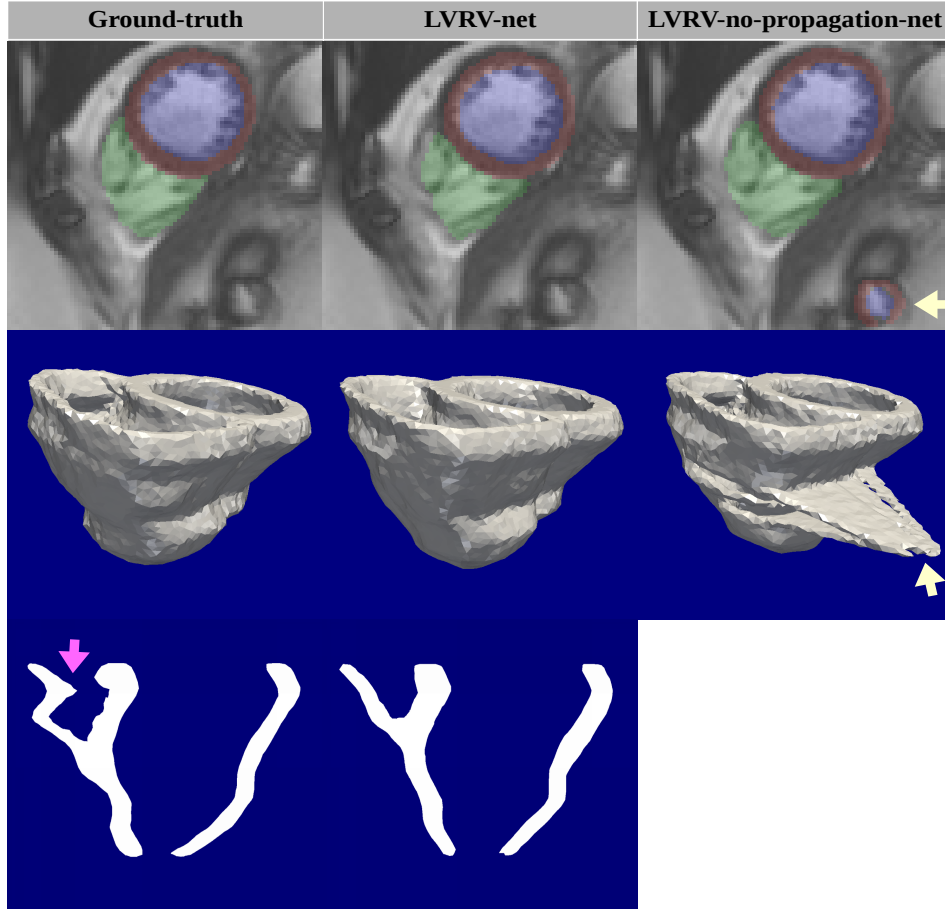


Figure 3.11: An example of the segmentation on difficult slices (zoomed-in versions of ROIs for better visualization) and the reconstructed meshes with the ground-truth, the prediction of LVRV-net and that of the LVRV-no-propagation-net. The last row shows a slice of the long-axis view of the meshes reconstructed with the ground-truth and the LVRV-net prediction (irregularities of the ground-truth reconstruction meshes are indicated by the arrows). The large-spread abnormal structures on the meshes in the third column are due to the interpolation of the wrong segmentation (indicated by the arrows). The LVRV-no-propagation-net makes a false positive prediction of LV on an intermediate slice.

Table 3.3: Segmentation Results by Pathological Group on the ACDC Dataset

	Dice				Hausdorff (mm)			
	LVM		LVC		LVM		LVC	
	mean	std	mean	std	mean	std	mean	std
Dilated cardiomyopathy	0.705	0.04	0.916	0.02	8.50	2.31	7.19	1.81
Hypertrophic cardiomyopathy	0.773	0.05	0.792	0.12	12.02	3.74	11.41	5.48
Myocardial infarction	0.708	0.06	0.890	0.03	9.83	3.51	8.35	2.40
Abnormal right ventricle	0.666	0.07	0.850	0.06	9.54	2.83	8.08	2.77
Normal	0.721	0.06	0.863	0.05	8.93	2.76	8.67	3.72

procedure:

- The first condition for a successful predicted mask becomes “both LVC and LVM are present” (instead of “LVM is present”).
- If the predicted mask is successful, only the largest components of LVC and LVM are respectively reserved as predicted masks.
- If the predicted LVC mask has any neighboring background pixels, we reset the prediction of those pixels to LVC (indicated by the arrows in Figure 3.12). We do so to follow the ACDC convention that LVC is almost always enclosed by LVM.

Among the methods in the ACDC challenge, [Isensee 2017] (ranked 1st), [Jang 2017] (ranked 4th) and [Wolterink 2017] (ranked 5th) report their performances on the 100 training cases. The performances on these cases of LV-net and these methods are presented in Table 3.2. Due to the variability of the UK Biobank training set ground-truth, as well as the difference between UK Biobank and ACDC images, LV-net is not as good as the state-of-the-art methods on Dice index. But it is rather comparable to them in terms of the mean of Hausdorff distance, and even better in terms of the standard deviation. This confirms the robustness of our method. In Figure 3.12, we also show some examples of LV-net prediction along with the ACDC ground-truth and the UK Biobank ground-truth on similar slices. It is clear that LV-net learns the segmentation “pattern” of the ground-truth from UK Biobank, which is different from that of ACDC.

We also find that the difference between the performances of our method on the 5 pathological groups remains limited as presented in Table 3.3. The pathological group seems to have less influence than the image quality of individual stack on the segmentation performance. Being trained with cases from the general population, our method generalizes well to the cases with pathology.

Table 3.4: Segmentation Results (Mean and Standard Deviation) on the Sunnybrook Dataset, Compared to the Performance from the State-of-the-art Methods

	Dice		APD (mm)		PGC (%)	
	LVC	LV-epi	LVC	LV-epi	LVC	LV-epi
proposed	0.88	0.94	2.11	1.95	97.08	99.21
LV-net	(0.07)	(0.03)	(0.49)	(0.42)	(6.04)	(2.95)
Tran	0.92	0.96	1.73	1.65	98.48	99.17
[Tran 2016]	(0.03)	(0.01)	(0.35)	(0.31)	(4.06)	(2.20)
Winther et al.	0.94	0.95	-	-	-	-
[Winther 2017]	(0.03)	(0.03)	(-)	(-)	(-)	(-)
Avendi et al.	0.94	-	1.81	-	96.69	-
[Avendi 2016a]	(0.02)	(-)	(0.44)	(-)	(5.7)	(-)
Queiros et al.	0.90	0.94	1.76	1.80	92.70	95.40
[Queiros 2014]	(0.05)	(0.02)	(0.45)	(0.41)	(9.5)	(9.6)
Poudel et al.	0.90	-	2.05	-	95.34	-
[Poudel 2016]	(0.04)	(-)	(0.29)	(-)	(7.2)	(-)

3.4.3.2 Experiments on Sunnybrook

The slices to be segmented of the 30 cases in Sunnybrook are well located on or below the base of the heart. We segment them with the trained LV-net. In a way similar to the practice in [Tran 2016], 160×160 central zones are cropped out as ROIs, which are then used as inputs to LV-net. Comparison of the performance of LV-net and up-to-date state-of-the-art research is presented in Table 3.4. LV-net is somewhat less accurate on Dice index and on average perpendicular distance (APD). But its robustness makes it comparable or even better than the state-of-the-art on the percentage of good contours (PGC). Examples of predicted masks and ground-truth are shown in Figure 3.13.

3.4.3.3 Experiments on RVSC

The slices to be segmented for the 16 cases in RVSC are all located below the base and above the apex. Similar to [Tran 2016], 216×216 central zones are cropped out as ROIs. We then apply the trained LVRV-net on these ROIs and evaluate the predicted RVC masks. Comparison with the up-to-date state-of-the-art research is presented in Table 3.5. In terms of Hausdorff distance, our method not only achieves better mean value but also generates much smaller standard deviation value compared the to state-of-the-art. Examples of predicted masks and ground-truth are presented in Figure 3.14.

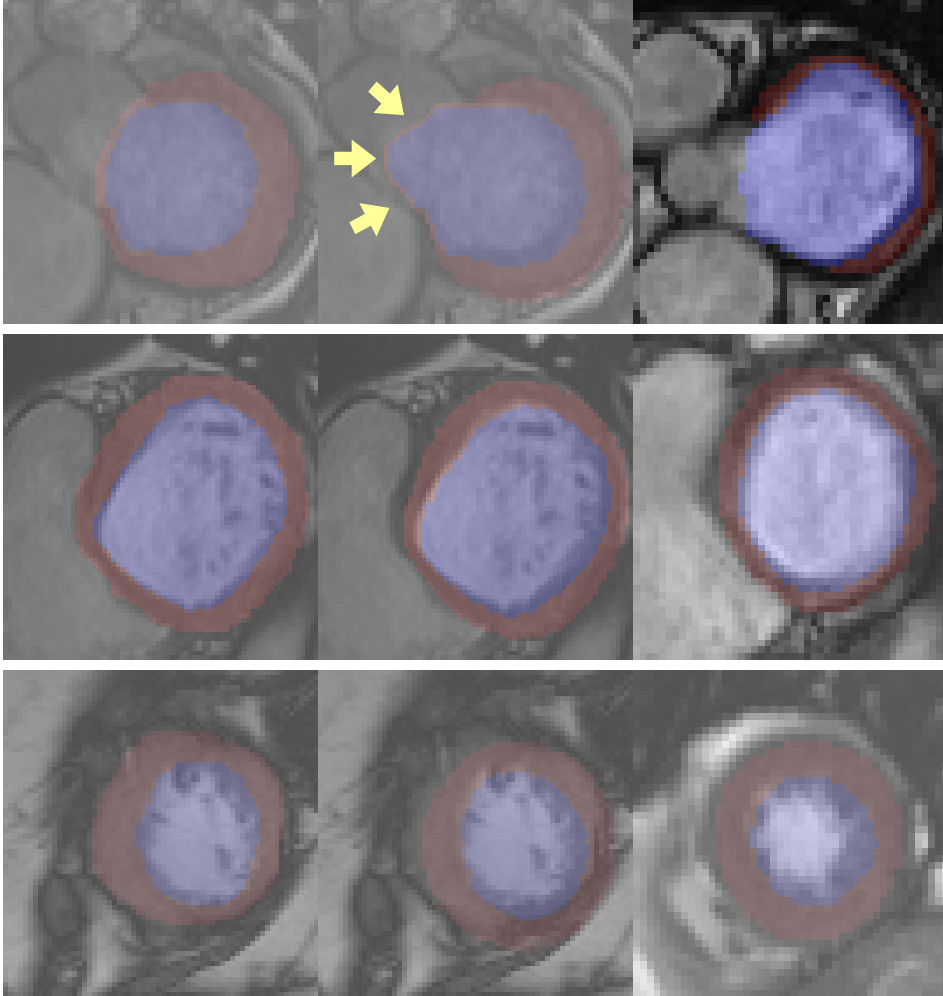


Figure 3.12: Examples of ground-truth (first column) vs prediction (second column) on ACDC dataset (the arrows indicate the pixel labels reset to LVM). We also add similar slices with the ground-truth in UK Biobank (third column). The 3 rows correspond to slices roughly on the top (around the base), in the middle and at the bottom (around the apex) of image stacks. LVC and LVM are marked as purple and brown respectively. Note that these images are zoomed-in versions of ROIs for the sake of better visualization. They are not the ROIs which LVRV-net and LV-net take as inputs.

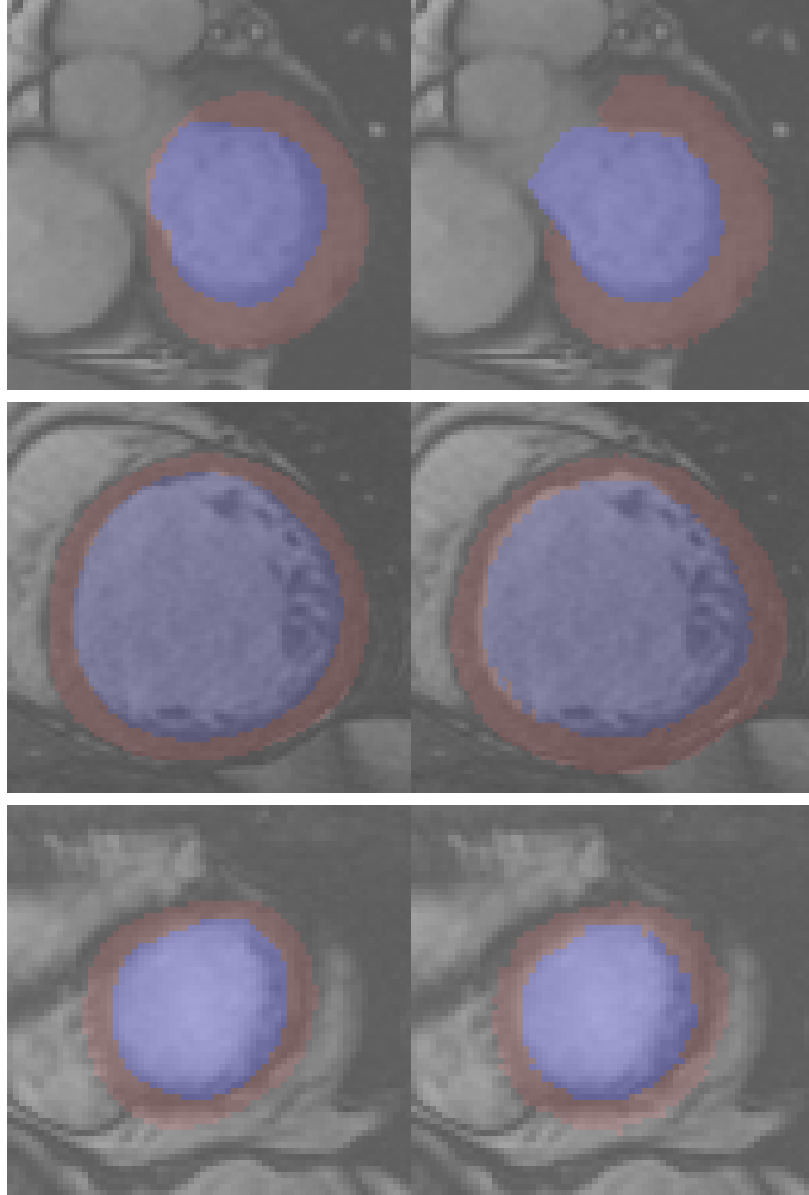


Figure 3.13: Examples of ground-truth (first column) vs prediction (second column) on Sunnybrook dataset. The 3 rows correspond to slices roughly on the top (around the base), in the middle and at the bottom (around the apex) of image stacks. LVC and LVM are marked as purple and brown respectively. Note that these images are zoomed-in versions of ROIs for the sake of better visualization. They are not the ROIs which LVRV-net and LV-net take as inputs.

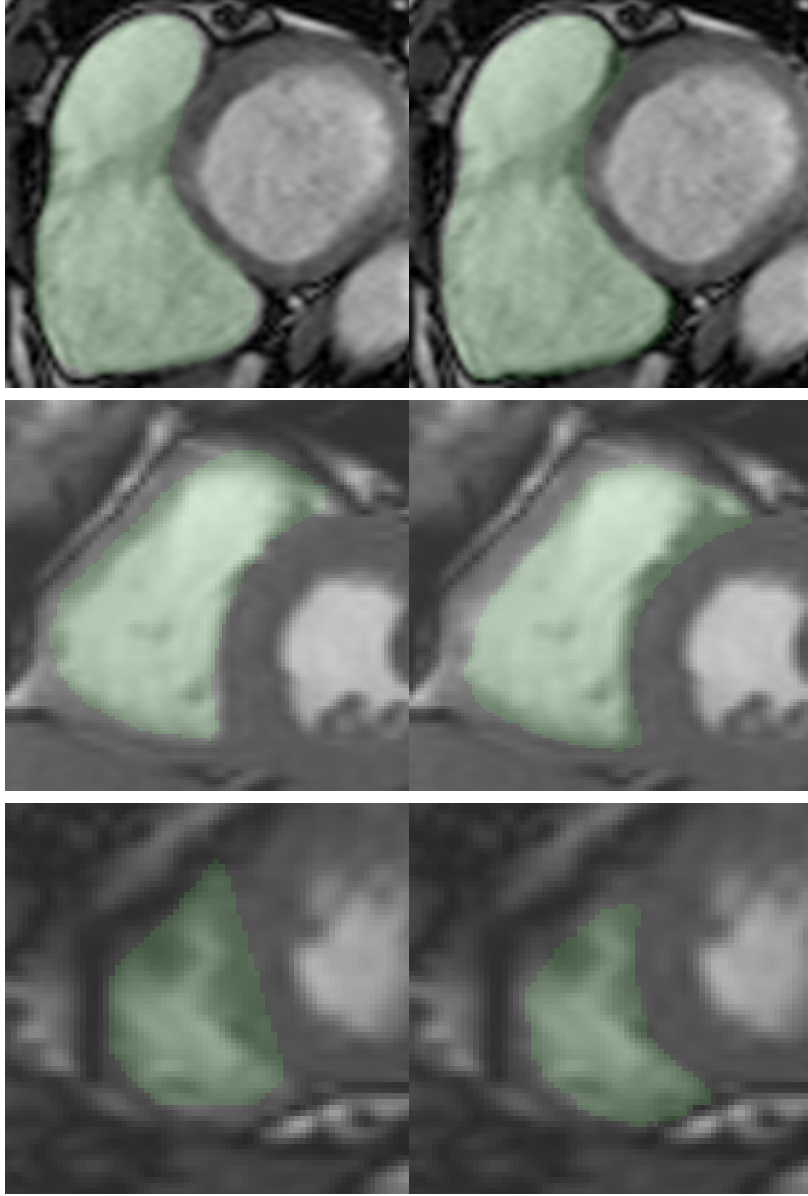


Figure 3.14: Examples of ground-truth (first column) vs prediction (second column) on RVSC dataset. The 3 rows correspond to slices roughly on the top (around the base), in the middle and at the bottom (around the apex) of image stacks. RVC is marked as green. Note that these images are zoomed-in versions of ROIs for the sake of better visualization. They are not the ROIs which LVRV-net and LV-net take as inputs.

Table 3.5: Segmentation Results on the RVSC Dataset, Compared to the Performance from the State-of-the-art Methods

	Dice		Hausdorff (mm)	
	RVC		RVC	
	mean	std	mean	std
proposed LVRV-net	0.82	0.07	7.56	3.50
Tran [Tran 2016]	0.84	0.21	8.86	11.27
Winther et al. [Winther 2017]	0.85	0.07	-	-
Avendi et al. [Avendi 2016b]	0.81	0.21	7.79	5.91
Zuluaga et al. [Zuluaga 2013]	0.76	0.25	11.51	10.06

3.5 Conclusion and Discussion

We propose a method of segmentation with spatial propagation that is based on originally designed neural networks. By taking the contextual input into account, the spatial consistency of segmentation is enforced. Also, we conduct thorough and unprecedented testing to evaluate the generalization ability of our model and achieve performance better than or comparable to the state-of-the-art. Furthermore, an exceptionally large dataset (UK Biobank) collected from the general population is used for training and evaluation.

Given the experiments in this paper, we notice that our method is very robust in terms of distance measures (e.g. Hausdorff distance) but less precise than the state-of-the-art in terms of Dice index. The variability of ground-truth in the UK Biobank training set is one important reason for that. For instance, the high ground-truth variability on the basal slice, which is included in the testing sub-stacks for LV-net but not for LVRV-net, explains the slightly lower performance measures of LV-net in Table 3.1. Yet this kind of variability commonly exists in large datasets so we have to decide to accept and cope with it. Furthermore, inconsistency problems may occur in segmentation (as illustrated and discussed), to which the Dice index might not be sensitive. We believe that on this problem more attention should be paid to the Hausdorff distance, according to which our proposed method performs better. For instance, in the third example shown in Figure 3.11, a small spot of false positive of LVC segmentation is predicted by LVRV-no-propagation-net. This is a very typical case of inconsistency: the false positive part is quite small compared to the ground-truth LVC, and therefore only causes a slight reduction of the Dice index. But it certainly brings about an explosion of the Hausdorff distance.

We did not directly measure the human performance in terms of 3D metrics on UK Biobank to compare with our method. However, the authors of [Bai 2017b] did conduct experiments on UK Biobank to measure human performance in terms of 2D metrics. Taking the inter-observer variability of 3 human experts into account, the reported human expert levels are about 0.93(LVC), 0.88(LVM), and 0.88(RVC) in terms of 2D Dice index, and about 3.1mm(LVC), 3.8mm(LVM) and 7.4mm(RVC) in terms of 2D Hausdorff distance. Though these results are not directly comparable to

ours, they may still give a rough idea of human performance. We roughly estimate that our method, while mainly focusing on consistency, has a performance still a little bit lower than that of human experts in terms of accuracy.

Most of the existing segmentation methods do not explicitly take spatial consistency into account. In particular, they do not accurately segment the “difficult” slices around the apex. Our method, segmenting in a spatially consistent manner, is particularly more robust than them on these slices. The importance of correctly segmenting these slices is often underestimated. In many cutting-edge research projects (e.g. cardiac motion simulation and image synthesis), as a primary step, 3D meshes need to be built based on segmentation. Without spatial consistency and success on the apical slices of the segmentation, the generated meshes would be problematic.

Finally, we wonder whether our method, with better performance on distance measures than many state-of-the-art methods, would be a great tool for cardiac motion analysis. Intuitively, the smaller the Hausdorff distance between the predicted and the ground-truth contours at each instant is, the more precisely the trajectory of the corresponding structure (e.g. LVC, LVM, RVC) can be tracked across time, and hence the better the motion can be characterized. We expect to carry out research on this in the future.

3.6 Appendix

3.6.1 Datasets

3.6.1.1 UK Biobank Dataset

It comprises short-axis cine MRI of 4875 participants from the general population. Details of the magnetic resonance protocol are described in [Petersen 2016]. Each time series is composed of 3D volumes with 10mm slice thickness and in-plane resolution ranging from 1.8mm to 2.3mm. Expert manual segmentation using CVI42⁶ for LVC, LVM, and RVC is provided as ground-truth at both ED and ES. The quality of ground-truth varies highly across the cases. We exclude about one thousand cases that are provided with incomplete (e.g. missing ground-truth on some slice(s)) or unconvincing ground-truth (e.g. visually significant image/mask mismatch). Then we split the remaining 3834 cases into 2 sets of 3078 cases and 756 cases, for training and evaluation respectively.

3.6.1.2 Automated Cardiac Diagnosis Challenge (ACDC) Dataset

The ACDC dataset comprises short-axis cine MRI of 100 subjects, which are divided into 5 groups of equal size: dilated cardiomyopathy, hypertrophic cardiomyopathy, myocardial infarction, abnormal right ventricle and normal subjects. Each time series is composed of 3D volumes with 5mm to 10mm slice thickness and in-plane

⁶<https://www.circlecvi.com/>

resolution ranging from 0.7mm to 1.9mm. Expert manual segmentation for LVC, LVM, and RVC is provided as ground-truth at both ED and ES phases.

3.6.1.3 Sunnybrook Dataset

The validation and the online sub-datasets of the Sunnybrook dataset, made available for the MICCAI 2009 challenge on automated left ventricle (LV) segmentation, contains short-axis cine MRI from 30 subjects with different cardiac conditions: healthy (6 cases), hypertrophy (8 cases), heart failure with infarction (8 cases), and heart failure without infarction (8 cases). Each time series is composed of 6 to 12 2D cine stacks with a slice thickness of 8mm and in-plane resolution ranging from 1.3mm to 1.4mm. Expert-delineated ground-truth contours of the endocardium, or LVC, are provided at both ED and ES phases. Those of epicardium, or LVM, are provided only at ED phase.

3.6.1.4 Right Ventricle Segmentation Challenge (RVSC) Dataset

The RVSC dataset comprises 16 training 2D short-axis cine MRI stacks consisting of slices located across the ventricle. The in-plane resolution ranges from 0.57mm to 0.97mm. Ground-truth delineation of endocardial borders (LVC contours) and epicardial borders are provided at both ED and ES phases for the training cine stacks.

3.6.2 Metrics

3.6.2.1 Dice Index

The Dice index measures the overlap between two areas (2D Dice index) or two volumes (3D Dice index). It is defined as

$$\mathcal{D}(A, B) = 2 \frac{A \cap B}{A + B} \quad (3.7)$$

for two areas or two volumes A and B . The Dice index varies from 0 (complete mismatch) to 1 (perfect match).

3.6.2.2 Hausdorff Distance

The Hausdorff distance measures the distance between two areas (2D Hausdorff distance) or two volumes (3D Hausdorff distance). It is defined as

$$\mathcal{H}(A, B) = \max \left(\max_{p \in A} \left(\min_{q \in B} d(p, q) \right), \max_{q \in B} \left(\min_{p \in A} d(p, q) \right) \right) \quad (3.8)$$

where d denotes Euclidean distance. A smaller Hausdorff distance implies a better match.

3.6.2.3 Average Perpendicular Distance

The average perpendicular distance (APD) [Radau 2009] measures the distance in mm from one contour to another, averaged over all contour points.

3.6.2.4 Percentage of Good Contours

Given a set of ground-truth contours and the corresponding predicted contours, the percentage of good contours (PGC) defined in [Radau 2009] is the fraction of the predicted contours which have APD less than 5mm away from the ground-truth contours.

3.6.2.5 Presence Rate

Segmentation methods may miss a structure totally on some difficult slices. Given the segmentation predictions on a sub-stack, the presence rate (PR) of a structure is defined as the ratio between the number of predicted masks with the structure and the number of slices in the sub-stack.

On the UK Biobank and ACDC datasets, we use the 3D Dice index and 3D Hausdorff distance as metrics, similar to what has been done for the ACDC STACOM MICCAI 2017 challenge. For Sunnybrook, we use the 2D Dice index, APD, and PGC as in the MICCAI 2009 challenge on automated LV segmentation. For RVSC, we use 2D Dice index and 2D Hausdorff distance as done for the MICCAI 2012 challenge on automated RV segmentation.

Explainable Pathology Classification with Motion Characterization

Contents

4.1	Introduction	50
4.2	Data	53
4.2.1	Dataset	53
4.2.2	Notation	53
4.3	Methods	54
4.3.1	Preprocessing: Region of Interest (ROI) Determination	54
4.3.2	Feature Extraction Step 1: Apparent Flow Generation	55
4.3.3	Feature Extraction Step 2: Segmentation	58
4.3.4	Feature Extraction Step 3: Shape-Related Features	58
4.3.5	Feature Extraction Step 4: Motion-Characteristic Features	59
4.3.6	Classification	63
4.4	Experiments and Results	65
4.4.1	Training ApparentFlow-net	65
4.4.2	Finetuning LVRV-net	67
4.4.3	Proposed Classification Model	67
4.4.4	Variants of the Proposed Classification Model	69
4.5	Conclusion and Discussion	72
4.6	Appendix	75
4.6.1	Loss Function for Training ApparentFlow-Net	75
4.6.2	Variants of the Proposed Classification Model with Different Values of Parameter C	76
4.6.3	Variants of the Proposed Classification Model with Different Classifiers and Input Features	76
4.6.4	Examples of Apparent Flow Generated by the ApparentFlow-net	77

Part of this chapter corresponds to the following scientific article:

- [Zheng 2018a] **Explainable Cardiac Pathology Classification on Cine MRI with Motion Characterization by Semi-Supervised Learning of Apparent Flow**

Qiao Zheng, Hervé Delingette and Nicholas Ayache. Submitted to Medical Image Analysis in November 2018, under minor revision in February 2019

4.1 Introduction

Cine magnetic resonance imaging (cine MRI) is widely used in the clinic as an approach to identify cardiac pathology. For both the patients and the clinicians, there is hence a great need for automated accurate cardiac pathology identification and classification based on MRI images as mentioned in [Rueckert 2016] and [Comaniciu 2016], as well as in the myocardial infarct classification challenge run at the STACOM workshop in 2015 ([Suinesiaputra 2018]). Recently, the state-of-the-art cardiac pathology classification methods extract various features from MRI images and perform classification based on these features. Despite the great results achieved so far, there are still some aspects that need to be further explored.

First, most classification models, including the state-of-the-art models, take many feature values together as input to a single or a group of machine learning classifiers (e.g. [Khened 2017], [Khened 2018], [Wolterink 2017], [Cetin 2017], [Isensee 2017]), and output the predicted probability distribution over several classes. Like many other machine learning methods, or more specifically like most deep learning methods, these classification models are not easy to interpret. On the one hand, most of the models contain at least hundreds of parameters and it is impractical to examine and explain the role of each parameter. On the other hand, as many features are used simultaneously, it is hard to tell in a straightforward manner which feature value contributes to the identification of which category. This drawback on explainability causes many problems as pointed out in [Holzinger 2017]. For instance, the lack of explainability is a significant hurdle for their widespread adoption in the clinic despite their performance. Moreover, under the new European General Data Protection Regulation, it may also generate legal issues in business, as companies are required to be able to explain why decisions have been made by their models upon demand. Hence we propose a simple classification model with 9 input features and 14 parameters in total such that the role and contribution of each feature or parameter are clear and explainable.

Second, in terms of data availability in medical image analysis, we usually have access to a large amount of unlabeled data and a small amount of labeled data. How to make good use of the available data to train automatic methods remains an open question ([Weese 2016]). Semi-supervised learning appears to be a powerful approach to tackle this challenge in general ([Bai 2017a], [Gu 2017], [Cheplygina 2018]). In this paper, while cardiac motion is estimated in a flow-based manner like in many other methods ([Gao 2016], [Parajuli 2017]), we extend it as a semi-supervised learning method to train a network for apparent flow gen-

eration, using the dataset of Automatic Cardiac Diagnosis Challenge (ACDC) of MICCAI 2017 ([Bernard 2018]), for which the ground-truth segmentation mask is only available for 2 time frames. Although the percentage of the segmented frames in the dataset is small, making efficient use of their segmentation masks in training is essential for the generated flow to have better consistency. In particular, with the supervision of the masks in training, we show that cardiac structures are better preserved after warping by the generated flow.

Third, the state-of-the-art classification methods most exclusively focus on features extracted at two instants: the instants of end-diastole (ED) and end-systole (ES). The other instants are often ignored in pathology classification. For example, in the ACDC challenge, 3 out of the 4 cardiac pathology classification methods, including [Khened 2017] (as well as its updated version [Khened 2018]), [Wolterink 2017] and [Cetin 2017], use only features based on ED and ES. The authors of [Isensee 2017] propose the only method in the ACDC challenge which explores the instants other than ED and ES by quantifying the volume change and by measuring the LV-RV dissynchrony. Yet much information about cardiac motion (e.g. how individual myocardial segments move) is still excluded from the extracted features. While more and more research efforts are put on cardiac motion estimation (e.g. [Qin 2018a], [Qin 2018b], [Xue 2018], [Yang 2017], [Yan 2018]) and cardiac disease assessment via motion analysis (e.g. [Gilbert 2017], [Dawes 2017], [Lu 2018]), we propose to explore the impact of specific motion features to learn the detection of cardiac pathologies by extracting some useful time series of simple and straightforward features from cine MRI image sequences. Ideally, the resulting time series should be both informative enough to be used for classification and intuitive to be understood by a physician.

In this paper, we propose a novel and explainable method to classify a subset of cardiac pathologies using deep learning of cardiac motion (in the form of apparent flows) and shape. Our main contribution is threefold:

- **Semi-supervised learning of flow:** a novel semi-supervised learning method is applied to train a neural network model, which outputs apparent flows given two MRI images from the same 2D+t cine MRI image sequence. This allows to learn the motion as apparent flows efficiently from both segmented and non-segmented image data.
- **Motion-characteristic features:** combining the apparent flows across time with cardiac segmentation, time series of the radius and thickness of myocardial segments are extracted to describe cardiac motion. As features, they are easy to interpret and allow to characterize different shapes and motions of cardiac pathologies.
- **Explainable classification model:** we train a set of 4 simple classifiers to perform binary classifications. Each classifier performs a logistic regression and takes no more than 3 feature values as input, which makes it very simple and easy to interpret. On the ACDC challenge training set and testing set, our model achieves 95% and 94% as classification accuracy respectively, which is comparable to the state-of-the-art.

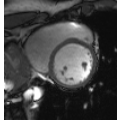
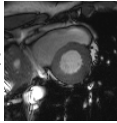
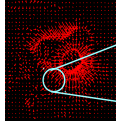
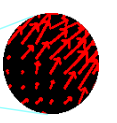
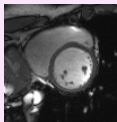




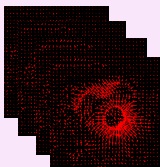

Step	Operation	Input	Output
1	Apparent Flow Generation by ApparentFlow-net	Frame ED  Frame t 	Apparent Flow  
2	Segmentation by LVRV-net	Frame ED or ES 	Mask ED or ES  6-Segment Division of the Myocardial Mask at ED 
3	Extraction of Shape-Related Features	Segmentation Masks of the Stacks at ED and ES  	Volumes at ED and ES, Volume Ratios, Myocardial Thickness
4	Extraction of Motion-Characteristic Features	Apparent Flow along Time  6-Segment Myocardial Mask at ED 	Time Series of Segment Radius and Thickness, Segment Motion Disparity Indices

Figure 4.1: Overview of the feature extraction method: 1. Apparent flow generation given the ED frame and another frame on the same slice; 2. Cardiac segmentation on the ED and ES frames and division of the ED myocardium mask to 6 segments; 3. Extraction of the shape-related features, including the calculation of the volumes, volume ratios and myocardial thickness of a heart given the segmentation masks; 4. Extraction of motion-characteristic features, including the creation of segment radius and thickness time series given a slice with the corresponding apparent flow maps and segmentation mask.

4.2 Data

4.2.1 Dataset

The proposed method is trained and evaluated on the ACDC challenge dataset, which consists of a training set of 100 cases and a testing set of 50 cases. The cine MRIs were acquired with a conventional SSFP sequence ([Bernard 2018]). Most of the cases contain about 10 slices of short-axis MRIs. And the number of frames in the cases varies between 12 and 35. ACDC training set and testing set are respectively divided into 5 pathological groups of equal size (we cite below the properties of each group as provided on the website, though they are only roughly exact according to our measure and observation):

- dilated cardiomyopathy (DCM): left ventricle cavity (LVC) volume at ED larger than 100 mL/m^2 and LVC ejection fraction lower than 40%
- hypertrophic cardiomyopathy (HCM): left ventricle (LV) cardiac mass higher than 110 g/m^2 , several myocardial segments with a thickness higher than 15 mm at ED and a normal ejection fraction
- myocardial infarction (MINF): LVC ejection fraction lower than 40% and several myocardial segments with abnormal contraction
- RV abnormality (RVA): right ventricle cavity (RVC) volume higher than 110 mL/m^2 or RVC ejection fraction lower than 40%
- normal subjects (NOR)

Please note that the abnormal contraction mentioned in the characteristics of MINF is quite vague as a property. In addition, both MINF and DCM cases have low LVC ejection fractions. And sometimes, a myocardial infarction causes a dilated LVC (for which we should classify the case to MINF instead of DCM according to ACDC challenge). As we will present later, it is indeed a challenge to distinguish them.

For the cases of ACDC training set, expert manual segmentation for LVC, RVC and the left ventricular myocardium (LVM) is provided as ground-truth for all slices at ED and ES phases; all other structures in the image are considered as background. For the cases of ACDC testing set, no ground-truth information about classification or segmentation is available. For performance evaluation on the testing set, the predicted results of a model need to be submitted online.

4.2.2 Notation

In this paper, slices in image stacks are indexed in spatial order from the basal part to the apical part of the heart. Given an image stack S , we denote N_S the number of its slices. Given two values a and b between 0 and $N_S - 1$, we note $S[a, b]$ the sub-stack consisting of slices of indexes in the interval $[\text{round}(a), \text{round}(b)[$ ($\text{round}(a)$ is included while $\text{round}(b)$ is excluded).

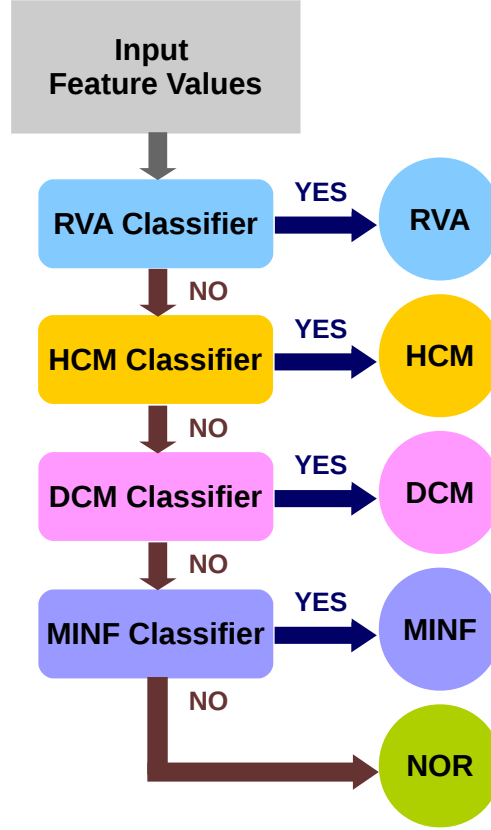


Figure 4.2: Overview of the classification method: the 4 binary classifiers are applied in sequence to classify a case to RVA, HCM, DCM, MINF or NOR.

4.3 Methods

Our method mainly consists of two parts: feature extraction (Figure 4.1) and classification based on features (Figure 4.2). But the region of interest (ROI) needs to be determined first.

4.3.1 Preprocessing: Region of Interest (ROI) Determination

As a preprocessing step, the ROI needs to be determined on the original MRI images. Short-axis MRI images usually cover a zone much larger than that of the heart. To save memory usage and to increase the speed of apparent flow and segmentation methods, it is better to work on an appropriate ROI instead. For this purpose, we directly apply an existing ROI method: we use the trained ROI-net exactly as described in [Zheng 2018b] to define an ROI. Briefly speaking, the ROI-net is a variant of U-net ([Ronneberger 2015]) for heart/background binary segmentation. It is applied on several middle slices on the ED image stack. As shown in [Zheng 2018b], this ROI determination method is very robust and succeeds in all cases of the ACDC dataset. In the remainder of this paper, we only refer to

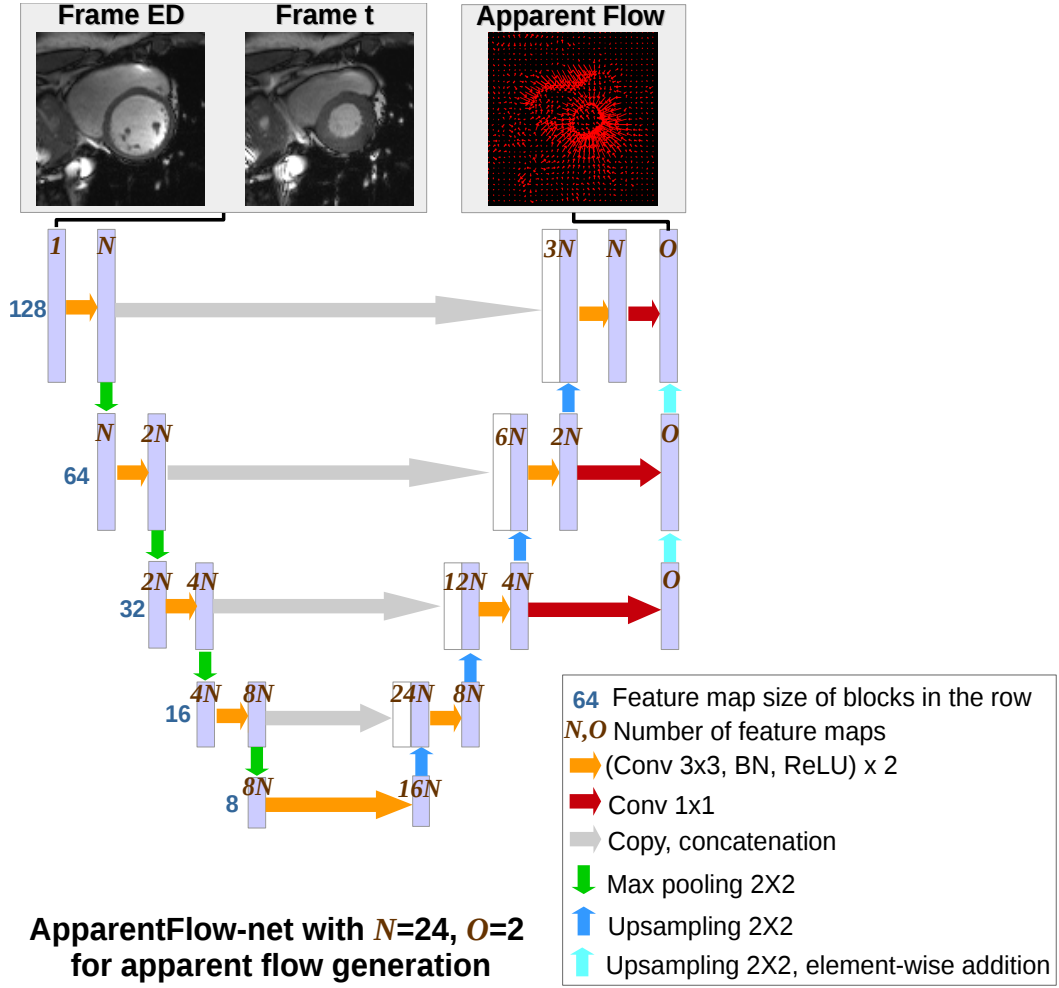


Figure 4.3: ApparentFlow-net: for apparent flow generation. The output is a map of pixel-wise flow F_t .

the automatically cropped ROI of the images.

4.3.2 Feature Extraction Step 1: Apparent Flow Generation

As shown in Figure 4.1, there are four steps for feature extraction. In this first step, the ApparentFlow-net, which is a variant of U-net ([Ronneberger 2015]) as shown in Figure 4.3, is proposed. U-net, with the encoder-decoder structure consisting of layers of various sizes of receptive fields, can effectively integrate local and global information, which is necessary for the analysis of the shape and motion of the heart on MRIs. Previously, we successfully used some variants of U-net for cardiac segmentation ([Zheng 2018b]). So we expect a similar structure would also work for the estimation of cardiac motion. The ApparentFlow-net is applied to generate

pixel-wise apparent flow given a pair of image frames on the same slice as input: the ED frame and another frame of index t on the same slice. In other words, the generated apparent flow map is a displacement field of the slice between ED and instant t . In a later step, combined with the segmentation mask, we will extract cardiac motion features from the sequences of apparent flow maps on a slice. The details of this extraction are available in the sub-section 3.5. While there exists some researches that explore image registration (or equivalently, apparent flow) using unsupervised learning (e.g. [Balakrishnan 2018], [Krebs 2018], [de Vos 2017], [Li 2017]), we propose a semi-supervised learning approach to make efficient use of a large amount of non-segmented images and a small amount of images segmented manually by experts.

In general, the idea of representing motion by apparent flow is based on two assumptions. First, we assume that the pixel intensities of an object do not change much between the two frames. Second, it is assumed that neighboring pixels have similar motion. By observation, we find that these assumptions usually hold on the slices located below the base and above the apex with some margin. This is due to the limited out-of-plane motion on these slices (this is less the case for the slices around the base and the apex). Hence ApparentFlow-net is trained and applied on the middle slices only.

If we note $I_{ED}(\mathbf{P})$ and $I_t(\mathbf{P})$ the pixel intensity of the two input frames of ApparentFlow-net at position $\mathbf{P} = (x, y)$, according to the first assumption above, ApparentFlow-net should generate an apparent flow map \mathbf{F}_t with $\mathbf{F}_t(\mathbf{P}) = (F_t^x(\mathbf{P}), F_t^y(\mathbf{P}))$ between ED and t enabling image reconstruction such that the following intensity discrepancy is minimized:

$$L_{IMG}(\mathbf{F}_t) = \sum_{\mathbf{P}} \left(I_{ED}(\mathbf{P}) - I_t(\mathbf{P} + \mathbf{F}_t(\mathbf{P})) \right)^2 \quad (4.1)$$

Meanwhile, the flow should also preserve the regularity of the motion of neighboring pixels according to the second assumption above. While there are already some methods in the community to impose diffeomorphisms (e.g. demon's algorithm as in [Pennec 1999], LDDMM as in [Hernandez 2008]), we propose a simple one to only discourage the occurrence of the extreme situations such as the crossing between two adjacent pixels or rotations greater than 90° (Figure 4.4). As long as these unrealistic motion patterns do not appear, there is no penalty on the regularity at all and the network is free to generate whatever flow without being influenced by the regularity constraint. More precisely, let us note $W_{\mathbf{F}_t}$ as the warping function such that $W_{\mathbf{F}_t}(\mathbf{P}) = \mathbf{P} + \mathbf{F}_t(\mathbf{P})$. For two adjacent pixels $\mathbf{P} = (x, y)$ and $\mathbf{P}^{x+} = (x + 1, y)$ in a row, we want the warped pixel $W_{\mathbf{F}_t}(\mathbf{P}^{x+})$ to stay on the right of the warped pixel $W_{\mathbf{F}_t}(\mathbf{P})$ (similarly for the adjacent pixels \mathbf{P} and $\mathbf{P}^{y+} = (x, y + 1)$ in a column) (see Figure 4.4). Otherwise, we say that a crossing on the x-components (y-components) of the flow pairs occurs and a penalty should apply. This translates as the following criterion to be minimized (more details about the derivation are available in the

Adjacent Pixel Pairs and Their Transformed Positions by Apparent Flow								
Crossing (and Hence Penalty) on the X-Components	NO	YES	NO	YES	-	-	-	-
Crossing (and Hence Penalty) on the Y-Components	-	-	-	-	NO	YES	NO	YES

Figure 4.4: Examples of adjacent pixel pairs transformed by apparent flow for which the crossing penalty applies or not.

appendix section):

$$L_{CROSS}(\mathbf{F}_t) = \sum_{\mathbf{P}} \min(1 + \frac{\partial F_t^x(\mathbf{P})}{\partial x}, 0)^2 + \min(1 + \frac{\partial F_t^y(\mathbf{P})}{\partial y}, 0)^2 \quad (4.2)$$

Moreover, we further encourage the flow to preserve the segmentation masks of cardiac structures $S \in \{LVC, LVM, RVC\}$. The warped segmentation masks of these structures should approximately match the ground-truth masks on the corresponding frame. Let us note M_{ED}^S and M_{ES}^S the binary ground-truth segmentation mask (of pixel intensity value 0 or 1) of S at the instants of ED and ES (the only instants for which the ground-truth is available in the ACDC training set). This constraint on the flow between ED and ES is based on the Dice coefficient

$$L_{GT}(\mathbf{F}_{ES}) = \sum_{S \in \{LVC, LVM, RVC\}} Dice(M_{ED}^S, M_{ES}^S \circ W_{\mathbf{F}_{ES}}) \quad (4.3)$$

The formula of the *Dice* function is provided in the appendix section.

Finally, the overall loss function for training the ApparentFlow-net is a linear combination of the terms L_{IMG} , L_{CROSS} and potentially L_{GT} . We adopt a semi-supervised approach for which L_{GT} is applied when ground-truth segmentation is available:

$$L_{flow}(\mathbf{F}_t) = L_{IMG}(\mathbf{F}_t) + p_1 L_{CROSS}(\mathbf{F}_t) + p_2 \mathbf{1}_{t=ES} L_{GT}(\mathbf{F}_t) \quad (4.4)$$

where $\mathbf{1}_{t=ES}$ is the indicator function for the event $t = ES$. $\mathbf{1}_{t=ES}$ is necessary as for the instants t other than ED and ES, the ground-truth segmentation is not provided in ACDC. Please note that this is a typical method of semi-supervised learning. It makes use of a small amount of labeled data (the images with ground-truth segmentation) and a large amount of unlabeled data (the images without ground-truth).

4.3.3 Feature Extraction Step 2: Segmentation

In this step, an existing model for segmentation proposed in [Zheng 2018b], the LVRV-net, is applied to segment MRI image stacks as presented in [Zheng 2018b]. With the concept of propagation along the long axis, this method was proven to be robust, as the results achieved on several different datasets are all comparable or even better than the state-of-the-art. For more details about the structure, training and application of the LVRV-net, please refer to [Zheng 2018b]. When we train and evaluate our method on the ACDC training set (100 cases), in each fold of a 5-fold cross-validation, the trained LVRV-net as given by [Zheng 2018b] is finetuned with the 80 cases used for training before being applied on the remaining 20 cases; for the evaluation of our method on ACDC testing set (50 cases), the trained LVRV-net is first finetuned with the 100 cases of ACDC training set.

In fact, in [Zheng 2018b], LVRV-net was trained to start the segmentation propagation from a given slice on which the ventricle cavities are supposed to be present. In other words, it was only trained to identify LV and RV labels on the slices below the base. So it might not work well if the basal slice is not determined in a stack and if the top slice in the volumetric image is located above the base. In this case, if we apply the original LVRV-net starting from the top slice, it might make a false positive prediction. With finetuning on ACDC, we find that this issue is solved. In general, the finetuned LVRV-net successfully learns from the ground-truth segmentation masks of ACDC that no foreground pixel is present (i.e. predict everything to be background) on the slices above the base and start segmentation propagation only when the base is reached. So it is no longer necessary to determine the basal slice manually. On the resulting sets of segmentation masks, we can hence also determine the location of the base, which is necessary for the calculation of volumes and the determination of sub-stacks for motion extraction as we will present later.

With the segmentation mask, we determine \mathbf{B}_L and \mathbf{B}_R , the barycenters of LVC and RVC respectively. Then all the pixels \mathbf{P} labeled to LVM on the segmentation mask are divided into 6 segments, depending on in which interval $[k\pi/3, (k+1)\pi/3[$ for k in $[0, 5]$ the angle between the vectors $\mathbf{B}_L\mathbf{P}$ and $\mathbf{B}_L\mathbf{B}_R$ is. An example of the resulting 6 segments are shown in Figure 4.1. This division of segments is inspired by the 17-segment system recommended by the American Heart Association (AHA) in [Cerqueira 2002]. Indeed, in the AHA system, on all short-axis slices around the base and at the level of mid-cavity, the myocardium is divided into 6 segments.

4.3.4 Feature Extraction Step 3: Shape-Related Features

Based on the segmentation masks generated in the previous step, we estimate the volumes of LVC, LVM and RVC of a case at ED and ES. For each of the two phases, the volume of LVC is calculated by approximating the LVC between two adjacent slices as a truncated cone and summing up all the truncated cone volumes:

$$V_{LVC} = \sum_i (S_i + S_{i+1} + \sqrt{S_i S_{i+1}})(L_{i+1} - L_i)/3 \quad (4.5)$$

Table 4.1: The extracted features used by our classification model

Feature	Notion (and Definition)
RVC volume at ED	$V_{RVC,ED}$
LVC volume at ES	$V_{LVC,ES}$
RVC ejection fraction	$EF_{RVC} (= 1 - V_{RVC,ES}/V_{RVC,ED})$
LVC ejection fraction	$EF_{LVC} (= 1 - V_{LVC,ES}/V_{LVC,ED})$
Ratio between RVC and LV volumes at ED	$R_{RVCLV,ED}$ ($= V_{RVC,ED}/(V_{LVC,ED} + V_{LVM,ED})$)
Ratio between LVM and LVC volumes at ED	$R_{LVMLVC,ED}$ ($= V_{LVM,ED}/V_{LVC,ED}$)
Maximal LVM thickness in all the slices at ED	$MT_{LVM,ED}$
Radius motion disparity	RMD
Thickness motion disparity	TMD

where S_i is the area of LVC on the slice i and L_i is the slice position along the long axis. The volume of LVM and RVC is calculated in a similar way. Then we normalize all the volumes by the corresponding body surface area (BSA) of the subject, which is a traditional practice based on the assumption that BSA is related to the metabolic rate. BSA can be computed from the height and the weight provided in ACDC (using the Mosteller formula $BSA = \sqrt{height * weight/60}$).

With the segmentation masks and volumes at ED and ES, we then compute the 7 shape-related features as listed in the first 7 rows of Table 4.1.

4.3.5 Feature Extraction Step 4: Motion-Characteristic Features

4.3.5.1 Slice Selection

For each case, let S be the image stack at ED (following the Notation part in the previous section). Given the segmentation masks of each slice generated in Step 2, we note i_1 the index of the first slice on which RVC mask is present (roughly the first slice below the base), and i_2 the index of the last slice on which LVC mask is present (roughly the last slice above the apex), and $h = i_2 - i_1 + 1$. Then we focus on extracting motion information from the sub-stack $S_{mid} = S[i_1 + 0.1h, i_2 + 1 - 0.2h]$. Please note that among the slices between the base and the apex, we exclude the top 10% and the bottom 20% and consider the remaining 70% in the middle, since the out-of-plane motion is particularly large in the slices close to the base or the apex.

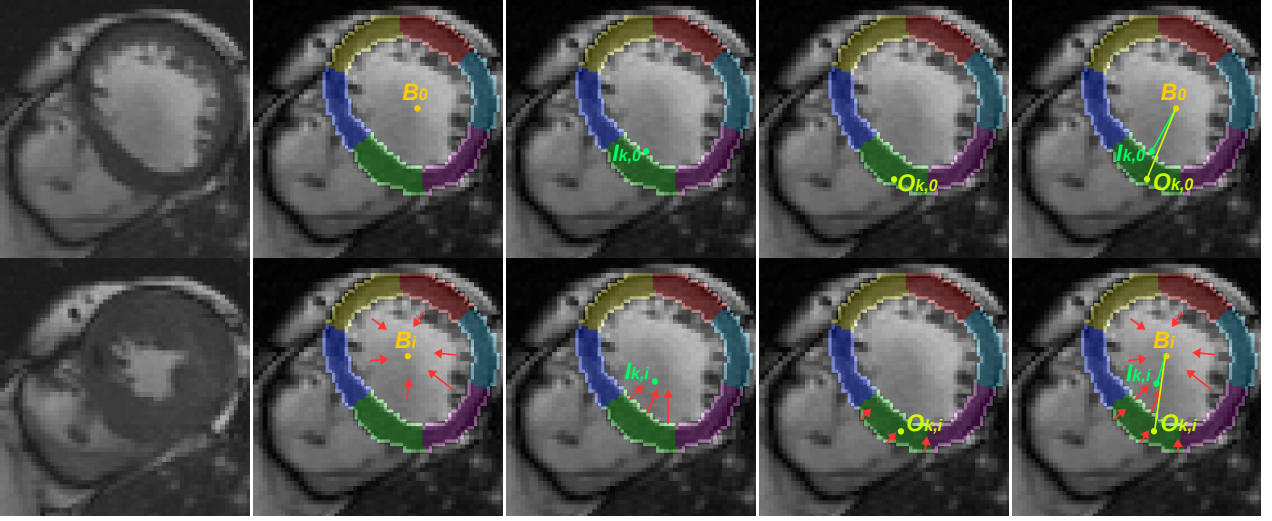


Figure 4.5: Definitions of B_i , $I_{k,i}$, $O_{k,i}$, $RA_{k,i}$ and $T_{k,i}$, for the extraction of motion-characteristic time series. The first row shows the definitions at t_0 ; the second row presents the definitions at t_i for $i \in [1, 9]$. 1st column: Frames at t_0 and t_i , based on which the apparent flow is generated. 2nd column: B_i is the barycenter of warped LVC (segmented at t_0) at t_i . 3rd column: $I_{k,i}$ is the barycenter of the warped inner boundary of segment S_k at t_i . 4th column: $O_{k,i}$ is the barycenter of the warped outer boundary of segment S_k at t_i . 5th column: $RA_{k,i} = |B_i I_{k,i}| / BSA$, $T_{k,i} = |B_i O_{k,i}| / BSA - RA_{k,i}$.

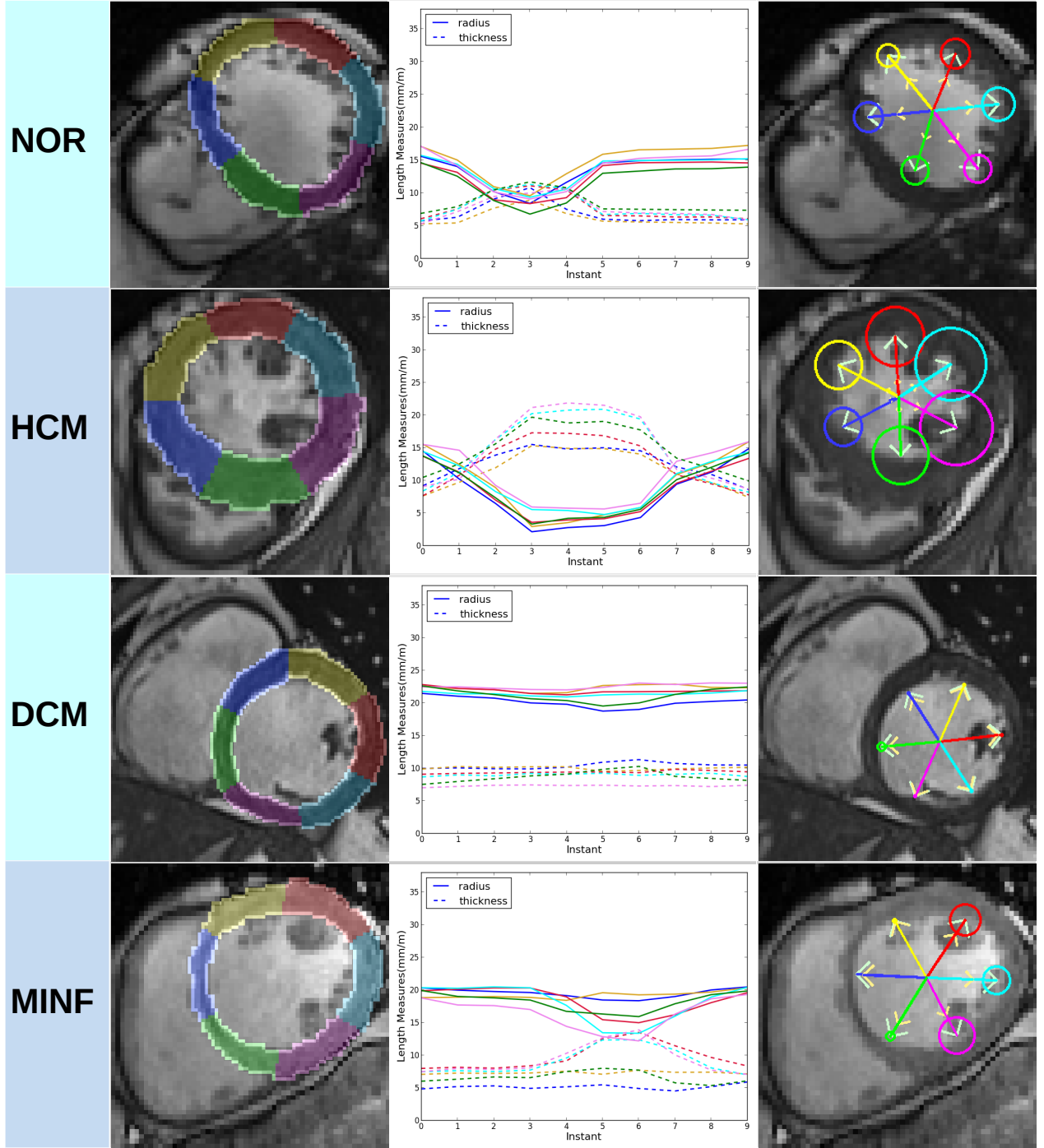


Figure 4.6: Examples of typical slice from 4 of the 5 pathological categories in ACDC. 1st column: the segmentation of the 6 myocardial segments (the boundaries of the segmentation masks are marked by lighter colors). 2nd column: time series of the radius (solid lines) and the thickness (dotted lines) of the 6 segments. 3rd column: a visualization of the motion information. For each segment, the radius connecting the LVC barycenter B_0 and the segment inner boundary barycenter $I_{k,0}$ (marked by the light green arrow) at ED is plotted. The segment inner boundary barycenter at ES is marked by the light orange arrow. The radius of the circle is proportional to the difference of segment thicknesses at ES and ED ($\Delta T_{k,i}$).

4.3.5.2 Frame Sampling

As presented in Figure 4.1, for each slice in S_{mid} , let us note f the number of frames available (all the frames together form a cardiac cycle). We sample 10 frames of instant t_i for i in $[0,9]$, such that t_0 is the instant of ED and $t_i = \text{round}(t_0 + i * f / 10) \bmod f$ for i in $[1,9]$. The 10 sampled frames hence cover the whole cardiac cycle. Applying the ApparentFlow-net of Step 1 in all the 9 pairs of frame (t_0, t_i) , we obtain 9 apparent flow maps \mathbf{F}_{t_i} . Hence for each pixel \mathbf{P} , we get its warped position $W_{\mathbf{F}_{t_i}}(\mathbf{P})$ for i in $[1,9]$.

4.3.5.3 Time Series Extraction

Then, with the convention that \mathbf{F}_{t_0} is the null apparent flow (and hence $W_{\mathbf{F}_{t_0}}$ is the identity function), the barycenter of LVC at t_i for $i \in [0, 9]$, \mathbf{B}_i , is defined as the average of $W_{\mathbf{F}_{t_i}}(\mathbf{P})$ for all the pixels \mathbf{P} labeled as LVC on the segmentation mask M_{ED} at t_0 (the 2nd column of Figure 4.5):

$$\mathbf{B}_i = \text{average}(\{W_{\mathbf{F}_{t_i}}(\mathbf{P}) \mid \mathbf{P} \in LVC \text{ on } M_{ED}\}) \quad (4.6)$$

In a similar way, for each myocardial segment S_k ($k \in [0, 5]$) and each instant t_i ($i \in [0, 9]$), we define $\mathbf{I}_{k,i}$, the barycenter of the inner boundary of the myocardial segment S_k at t_i (the 3rd column of Figure 4.5):

$$\begin{aligned} \mathbf{I}_{k,i} = & \text{average}(\{W_{\mathbf{F}_{t_i}}(\mathbf{P}) \mid \mathbf{P} \in LVC \text{ on } M_{ED} \\ & \& \mathbf{P} \text{ has neighboring pixel}(s) \in S_k\}) \end{aligned} \quad (4.7)$$

and $\mathbf{O}_{k,i}$, the barycenter of the outer boundary of the myocardial segment S_k at t_i (the 4th column of Figure 4.5):

$$\begin{aligned} \mathbf{O}_{k,i} = & \text{average}(\{W_{\mathbf{F}_{t_i}}(\mathbf{P}) \mid \mathbf{P} \in S_k \\ & \& \mathbf{P} \text{ has neighboring pixel}(s) \in \text{background on } M_{ED}\}) \end{aligned} \quad (4.8)$$

Finally, as shown in the 4th column of Figure 4.5, we define the radius of S_k at t_i normalized by BSA:

$$RA_{k,i} = |\mathbf{B}_i \mathbf{I}_{k,i}| / BSA \quad (4.9)$$

and the thickness of S_k at t_i normalized by BSA:

$$T_{k,i} = |\mathbf{B}_i \mathbf{O}_{k,i}| / BSA - RA_{k,i} \quad (4.10)$$

We hence generate two time series $\{RA_{k,i} : i \in [0, 9]\}$ and $\{T_{k,i} : i \in [0, 9]\}$ to represent the contraction and the thickening of S_k .

4.3.5.4 Visual Correspondence between Time Series and Pathologies

We compute the two time series introduced above for all the slices in S_{mid} of all the cases in ACDC. From the majority of the cases, we manage to visually identify

several typical slices with the time series characterizing the motion of the corresponding category. Examples of such typical slices are presented in Figure 4.6. To sum up our observation on the typical slices of each category as shown in Figure 4.6:

- NOR: all segments have similar radius and thickness at all instants; their contraction and thickening are synchronous and with comparable scales.
- HCM: the segments not only look proportionally thicker at ED, but also thicken more and contract stronger in the radial direction.
- DCM: the radiuses are quite large; the segments are moving so little that neither contraction nor thickening is obvious.
- MINF: the radiuses are quite large; some segments are clearly much more active than others.

4.3.5.5 Motion-Characteristic Feature Values

To better distinguish DCM and MINF cases, we define two additional feature values which often indicate the abnormal contraction described in the definition of MINF.

The first one is “radius motion disparity” (RMD). Given a case, we consider the set of radius series $\{RA_{k,i} : i \in [0, 9]\}$ of all the segments S_k on all the slices in the sub-stack S_{mid} (e.g. if there are 4 slices in S_{mid} , we consider a set of $6 \times 4 = 24$ time series). We first define the disparity of motion over all the segments in S_{mid} at the instant t_i as the difference between the maximum and minimum contraction at t_i :

$$RD_i = \max_{S_k \in S_{mid}} RA_{k,i}/RA_{k,0} - \min_{S_k \in S_{mid}} RA_{k,i}/RA_{k,0} \quad (4.11)$$

Then RMD is defined as the maximum disparity along the cardiac cycle:

$$RMD = \max_{i \in [0, 9]} RD_i \quad (4.12)$$

The second motion-characteristic feature value is named “thickness motion disparity” (TMD). For each slice s in S_{mid} and each t_i , we define the thickness motion disparity of the slice s at t_i as

$$TD_{s,i} = (\max_{k \in [0, 5]} T_{k,i} - \min_{k \in [0, 5]} T_{k,i}) / \min_{k \in [0, 5]} T_{k,0} \quad (4.13)$$

where we normalize the thicknesses by the minimum segment thickness at t_0 on slice s taking into account that myocardial thickness may vary across slice.

Finally, TMD is defined as

$$TMD = \max_{s \in S_{mid}, i \in [0, 9]} TD_{s,i} \quad (4.14)$$

4.3.6 Classification

4.3.6.1 4-Classifier Classification Model

Using the 7 shape-related features and the 2 motion-characteristic features as input, a classification model is proposed (Figure 4.2) to classify the 5 pathological

categories of ACDC. It consists of 4 binary classifiers:

- RVA classifier: RVA cases v.s. all the other cases.
- HCM classifier: HCM cases v.s. MINF, DCM and NOR cases.
- DCM classifier: DCM cases v.s. MINF and NOR cases.
- MINF classifier: MINF cases v.s. NOR cases.

The 4 binary classifications are arranged in increasing order of difficulty of the binary classification tasks. RVA and HCM cases can be identified based on the commonly used shape-related features. So they are classified first. DCM and MINF cases are somewhat similar in terms of sizes and ejection fractions. We use the novel motion-characteristic features to better distinguish them. Hence this more difficult classification is performed at the end.

4.3.6.2 Explainable Manual Feature Selection

To keep the classifiers simple, limit their risk of overfitting and increase their explainability, we chose no more than 3 features for each classifier as shown in Table 4.2:

- For RVA classifier, according to the definition provided by ACDC, the RVC volume at ED and the RVC ejection fraction are the most relevant features. We add one more feature, the ratio between RVC and LV volumes at ED, as we find that most RVA cases have disproportionately large RVC.
- For HCM classifier, LVC ejection fraction and maximal LVM thickness are selected according to the definition of HCM. The ratio between LVM and LVC volumes at ED is added because with most HCM cases this ratio is exceptionally high due to the exceptional myocardial thickness .
- For DCM classifier, as DCM cases are usually dilated at ED and inactive from ED to ES, their volumes of LVC at ES can be exceptionally large. So this feature is used. In addition, we also use radius motion disparity and thickness motion disparity.
- For MINF classifier, by definition, LVC ejection fraction is enough to distinguish MINF cases from NOR cases

4.3.6.3 Model of Classifiers

Each of the 4 classifiers is just a ridge logistic regression model. For a training case of index m , if we note $f_{m,i}$ the i -th feature values used as input of the classifier and y_m (-1 or 1, corresponding to no or yes) the binary ground-truth of the case, then the classifier is trained by minimizing

$$L_{classifier}(\{p_i\}, b) = \frac{1}{2} \sum_i p_i^2 + C \sum_m \log \left(\exp \left(-y_m \left(\sum_i p_i f_{m,i} + b \right) \right) + 1 \right) \quad (4.15)$$

with respect to the parameters $\{p_i\}$ and b . C is the inverse of regularization strength. After the training is done, given a case of index l and feature values $f_{l,i}$, the pre-

Table 4.2: The input features of the 4 binary classifiers

	Input Feature(s)
RVA Classifier	$V_{RVC,ED}, EF_{RVC}, R_{RVCLV,ED}$
HCM Classifier	$EF_{LVC}, R_{LVMLVC,ED}, MT_{LVM,ED}$
DCM Classifier	$V_{LVC,ES}, RMD, TMD$
MINF Classifier	EF_{LVC}

diction the sign of $\sum_i p_i f_{l,i} + b$. If it is non-negative, the prediction of the trained classifier is yes; otherwise it is no.

4.3.6.4 Flexibility and Versatility of the Model

Finally, we would also like to point out that the 4 classifiers function independently. While they are grouped together to form the proposed classification model in this paper, they can certainly be applied separately or grouped in a different manner in other situations if appropriate. This proposed classification model is hence very flexible and versatile.

4.4 Experiments and Results

We evaluate our method in two different ways. On the one hand, the model is trained with ACDC training set and then tested on ACDC testing set. On the other hand, a 5-fold cross-validation is performed on ACDC training set. For the latter, the 100 cases of ACDC training set are partitioned into 5 subsets of 20 cases, such that in each subset there are exactly 4 cases of each of the 5 categories.

In addition, we also analyze the proposed model by comparing it with various other models. Since the ground-truth category is only available for the cases in the training set (and not for those in the testing set), this analysis is based on the results on the training set.

4.4.1 Training ApparentFlow-net

4.4.1.1 Parameters and Data

In the training process with the whole ACDC training set, as well as in each of the 5 training processes of the 5-fold cross validation, the ApparentFlow-net is trained using the loss function $L_{flow}(\mathbf{F}_t)$ introduced in the Method section for 50 epochs with batch size 16. In terms of loss function parameter, we empirically find that $p_1 = 10^3$ and $p_2 = 10^5$ work well. These values are hence used for training. In terms of training data, for each case in the corresponding training set, we use the slices in the sub-stack $S[i_1 + 0.2h, i_2 + 1 - 0.2h]$ (with the notation introduced in the sub-section 3.5.1). In other words, we approximately exclude the top 20% and the bottom 20% of all the slices covering the LV cavity, and select the remaining 60% in the middle. This slice selection for training (middle 60%) is slightly more

Table 4.3: The mean(standard deviation) of Dice coefficients achieved by comparing $M_{ES} \circ W_{F_{ES}}$ and M_{ED} for 3 cardiac structures in the 5-fold cross-validation on ACDC training set.

Training Method	Dice		
	LVM	LVC	RVC
semi-supervised (proposed)	0.84(0.07)	0.94(0.07)	0.87(0.19)
unsupervised	0.76(0.08)	0.93(0.06)	0.83(0.22)

conservative than that for the application of the method (middle 70%). This design is aimed to further reduce the impact of the out-of-plane motion in training. For each selected slice, the frame pairs of indices (ED, t) for all frame index t are used to train the ApparentFlow-net. Only when $t = ES$, the term $L_{GT}(\mathbf{F}_t)$ in $L_{flow}(\mathbf{F}_t)$ using the segmentation ground truth is applied. With our automatic slice selection approach, in total, there are 13672 frame pairs used for training in the ACDC training set. Among the 13672 frame pairs, only 515 pairs (3.77%) come with segmentation ground-truth such that the term $L_{GT}(\mathbf{F}_t)$ applies.

4.4.1.2 Performance

To measure its performance, in each evaluation of the 5-fold cross-validation, for all the slices in the sub-stack $S[i_1 + 0.2h, i_2 + 1 - 0.2h]$ of all the 20 cases for evaluation, we apply the trained ApparentFlow-net to generate \mathbf{F}_{ES} . Then we use it to warp the ground-truth segmentation mask at ES, noted as M_{ES} , to obtain $M_{ES} \circ W_{F_{ES}}$. $M_{ES} \circ W_{F_{ES}}$ is then compared with M_{ED} , the corresponding ground-truth masks at ED, using Dice coefficient (2D version) on LVM, LVC and RVC. Overall, the means(standard deviations) of Dice coefficients achieved on LVM, LVC and RVC in the 5-fold cross-validation are reported in Table 4.3.

Additionally, we also visually evaluate the apparent flow generated by the ApparentFlow-net. We find that the apparent flow is indeed good enough to characterize the cardiac motion of the typical cases in the pathological categories. Several examples are given in the appendix section.

4.4.1.3 Importance of Supervision in Training

In order to understand the importance of the small amount of segmentation ground-truth used in the proposed semi-supervised learning method, we also train a variant of ApparentFlow-net using only unsupervised learning. The only modification is the removal of the term $L_{GT}(\mathbf{F}_t)$ from $L_{flow}(\mathbf{F}_t)$ such that the variant is trained without any ground-truth for supervision. As reported in Table 4.3, the means of Dice coefficients on LVM, LVC and RVC are all lower than the corresponding results achieved by the semi-supervised learning method. In particular, there is a large drop from 0.84 to 0.76 on the mean of Dice coefficient on LVM. So the proposed semi-

Table 4.4: The classification performance on the testing set (50 cases) and training set (100 cases) of ACDC by different models

Model	Testing Set Accuracy	Training Set Accuracy	Evaluation Method on Training Set
proposed model	94%	95%	5-fold cross-validation
[Isensee 2017]	92%	94%	5-fold cross-validation
[Wolterink 2017]	86%	91%	4-fold cross-validation
[Cetin 2017]	92%	100%	forward feature selection and leave-one-out cross-validation
[Khened 2017]	96%	90%	70 cases for training, 20 for validation, 10 for evaluation
[Khened 2018]	100%	N.A.	N.A.

supervised learning method is indeed better than its unsupervised learning variant by making efficient use of the small amount of segmented images.

4.4.2 Finetuning LVRV-net

LVRV-net is already trained in [Zheng 2018b] for 80 epochs on a subset of about 3000 cases of UK Biobank ([Petersen 2016]). In the training process with the whole ACDC training set, as well as in each of the 5 training processes of the 5-fold cross validation, LVRV-net is finetuned for 920 epochs on the corresponding training data, with exactly the same loss function and training parameters as given in [Zheng 2018b]. With the finetuning, the means (standard deviations) of 3D Dice coefficients achieved on LVC, LVM and RVC segmentation volumes in the 5-fold cross-validation are 0.94(0.06), 0.90(0.03) and 0.89(0.12).

4.4.3 Proposed Classification Model

Apparent flows and segmentation masks are generated by the ApparentFlow-net and the finetuned LVRV-net, from which the 7 shape-related features and the 2 motion-characteristic features are extracted. Then the 4 ridge logistic regression binary classifiers are implemented using Scikit-learn [Pedregosa 2011] and trained on the cases of the categories they are supposed to classify. For example, DCM classifier is trained on the cases of NOR, MINF and DCM; the cases of RVA or HCM are not used to train it. In terms of classifier parameter, we empirically find that $C = 50$ works well and use it in this paper. The performances of some variants with different values of C are provided in the appendix section.

4.4.3.1 Classification Performance

As presented in Table 4.4, on the testing set, the accuracy of our model is 94%. In the 5-fold cross-validation on the training set, our method achieves an accuracy of 95%. Hence our model achieves performances that are comparable to those of the

		Predicted				
		NOR	RVA	HCM	DCM	MINF
Ground-Truth	NOR	20	0	0	0	0
	RVA	2	18	0	0	0
	HCM	1	0	19	0	0
	DCM	0	0	0	20	0
	MINF	0	0	0	2	18

Figure 4.7: The confusion matrix of the predictions by the proposed classification model in the 5-fold cross-validation on the training set of ACDC.

state-of-the-art on both the training set and the testing set. This is quite remarkable because, in contrast to the state-of-the-art, each classifier in our model uses only up to three features and has only up to 4 parameters. In total, our model uses 9 features and has 14 parameters. And each feature is selected in a clearly explainable manner. On the testing set, among the two methods with performances better than ours, [Khened 2017] uses a random forest of 100 trees and [Khened 2018] applies a more sophisticated ensemble system. Therefore, those classification models are less straightforward to interpret than ours. Furthermore, since our model has very similar performances on the training and testing sets, there seems to be little overfit.

Based on the confusion matrix of the prediction in the 5-fold cross-validation on the ACDC training set (Figure 4.7), for the binary classification of NOR, RVA, HCM, DCM and MINF, we calculate and find that the precision values are 0.87, 1.00, 1.00, 0.91 and 1.00; the recall values are 1.00, 0.90, 0.95, 1.00 and 0.90.

4.4.3.2 Interpretation of Mis-Classification

As our classifier can be interpreted easily, we figure out for each of the 5 misclassified cases (Figure 4.7) why the prediction is different from the ground-truth. In fact, they all seem to be somewhat ambiguous in terms of pathological category:

- Patients 082 and 088 are of ground-truth RVA but are classified as NOR. According to our segmentation, they have $V_{RVC,ED}$, EF_{RVC} and $R_{RVCLV,ED}$ values all very similar to that of the NOR cases. For instance, they have the third and the first lowest $R_{RVCLV,ED}$ values (0.755 and 0.691) among all the RVA cases, which are well in the range of that of the NOR cases.
- Patient 022 is of ground-truth HCM but is predicted as NOR. Unlike all the other HCM cases, patient 022 has both EF_{LVC} (0.622) and $MT_{LVM,ED}$ (14.7mm) in the normal ranges, which makes it look like a NOR case.
- Patients 050 and 060 are of ground-truth MINF but are predicted as DCM. Their values of $V_{LVC,ES}$ (118.0mL/m² and 83.5mL/m²) are the two highest among all

Table 4.5: The parameters of the 4 logistic regression binary classifiers trained on the training set of ACDC

	Parameters of the Trained Classifier
RVA Classifier	$0.010 V_{RVC,ED} - 4.695 EF_{RVC} + 14.012 R_{RVCLV,ED} - 9.906$
HCM Classifier	$8.434 EF_{LVC} + 4.614 R_{LVMLVC,ED} + 0.420 MT_{LVM,ED} - 16.580$
DCM Classifier	$0.104 V_{LVC,ES} - 0.918 RMD - 7.758 TMD - 0.321$
MINF Classifier	$-17.122 EF_{LVC} + 7.994$

the non-DCM cases and well in the range of that of the DCM cases. In terms of motion disparity, on RMD and TMD , unlike the majority of the MINF cases, their values (0.245 and 1.173 for patient 050, 0.316 and 1.246 for patient 060) are also in the ranges of that of the DCM cases. For these reasons, the DCM classifier predicts them to be DCM cases.

4.4.3.3 Explaining the Classifiers

The 4 binary classifiers are just logistic regression models. As presented in the previous section, their prediction depends on the sign of the sum $\sum_i p_i f_{l,i} + b$. To understand what is learned from the data by the trained classifiers, in Table 4.5 we show the coefficients of the classifiers trained with all the relevant cases in ACDC. We find that the signs of the parameters p_i all correspond to the positive or negative correlation between the feature and the binary classification task. For instance, in the trained RVA classifier, the signs of the coefficients of $V_{RVC,ED}$ and $R_{RVCLV,ED}$ are both positive, as a large RVC volume and a high ratio between the RVC and LV volumes are both indicators of RV abnormality; on the other hand, since low RVC ejection fraction usually signifies RV abnormality, the coefficient of EF_{RVC} is negative. Similarly, such a correspondence applies to all the coefficients of the 3 other trained classifiers. In particular, for MINF classifier, the learned threshold on EF_{LVC} to distinguish MINF cases from NOR cases is $7.994/17.122 = 0.467$, which can well separate them according to their definitions.

4.4.4 Variants of the Proposed Classification Model

We compare the proposed classification model with its variants for a justification of our design and a more comprehensive understanding of the model.

4.4.4.1 Importance of Motion-Characteristic Features

To better understand the value of the two proposed motion-characteristic features, we further train three variants of DCM classifier which use zero or one motion-characteristic feature as input. And the set of input features is the only difference

Table 4.6: The performance of the variants of DCM classifier on the training set of ACDC

DCM Classifier Input	# of Mis-Classification on the 60 DCM, MINF and NOR cases
$V_{LVC,ES}, RMD, TMD$ (proposed)	2
$V_{LVC,ES}, TMD$	2
$V_{LVC,ES}, RMD$	3
$V_{LVC,ES}$	4

between these models. As shown in Table 4.6, on the 60 cases of NOR, MINF and DCM, while DCM classifier makes only two errors, the variant using only shape-related feature $V_{LVC,ES}$ misclassifies 4 cases. But improvements can be made by using at least one motion-characteristic feature. As can be visualized in Figure 4.8, the motion characteristic features RMD and TMD allow the separation of the majority of the cases of DCM and MINF. Combining them with the shape-related feature $V_{LVC,ES}$ together as the input, the DCM classifier can make more accurate classification.

4.4.4.2 Proposed Model on Non-Normalized Features

We test whether BSA normalization is required for our model to achieve high performance. Among the 9 proposed features, only the values of $V_{RVC,ED}$ and $V_{LVC,ES}$ would be different without BSA-normalization. And only RVA and MINF classifiers which use these two features as input would be affected. As presented in Table 4.7, without BSA-normalization on the features, the 5-fold cross validation accuracy on ACDC training set only drops a little bit to 94%. The proposed model still remains accurate.

4.4.4.3 Proposed Model with Inversed Classifier Order

As presented previously, the 4 classifiers in the proposed model are arranged according to the estimated difficulties of the corresponding classification task. To confirm the importance of this order, we create another model by inverting the order of the classifiers. So, unlike the proposed model shown in Figure 4.2, in this variant, a case goes through successively MINF, DCM, HCM and RVA classifiers instead. As shown in Table 4.7, the accuracy of this variant is quite low. Hence the proposed order of the classifiers is indeed important.

4.4.4.4 Variants with Other Classifier Models

We replace the proposed logistic regression classifiers with 3 other types of classifiers on the same sets of input features, including Lasso, LassoCV (Lasso with model selection by cross-validation) and random forest. Details of these models are available

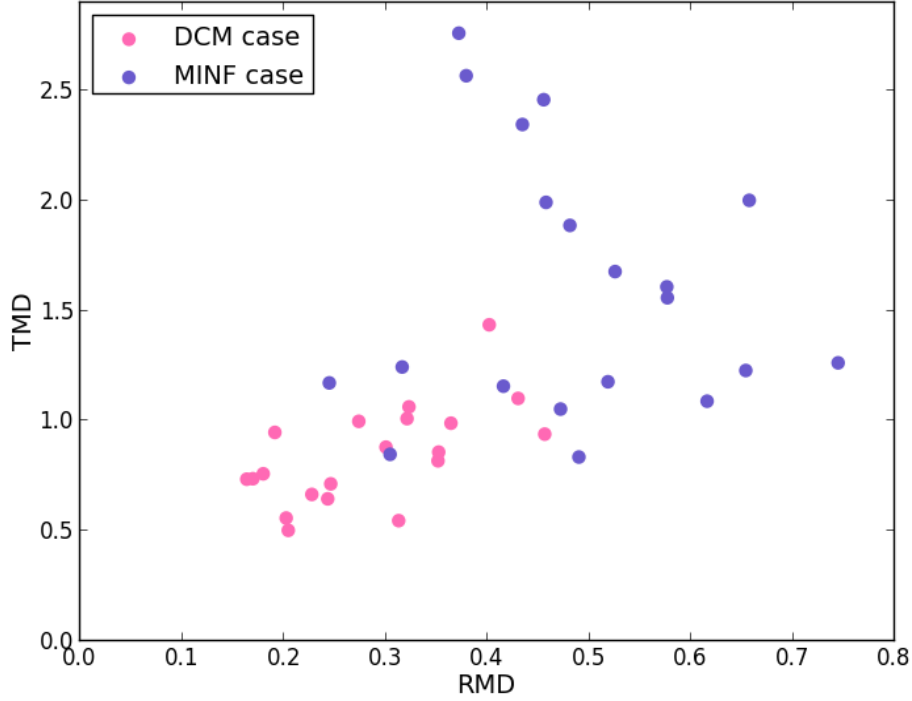


Figure 4.8: The motion-characteristic features (RMD and TMD) of the DCM and MINF cases in the training set of ACDC. The majority of the cases are well separable with these two features.

in the appendix section. As reported in Table 4.7, their performances are clearly below that of the proposed model. Our choice of logistic regression as the classifier model is hence justified.

4.4.4.5 Variants without Manual Input Feature Selection

To evaluate if the manual feature selection is useful for the model to be accurate, we train several modified versions of the proposed model without manual feature selection. They all consist of 4 classifiers to perform the same binary classification tasks as in the proposed model. But each of the 4 classifiers of these variants takes all 9 features together as input. In total, we implement 6 models with the following models as their classifiers respectively (details of these models are available in the appendix section): support vector machine (SVM), relevance vector machine (RVM), Lasso, LassoCV, random forest and high dimensional discriminant analysis (HDDA) model.

As reported in Table 4.7, on the BSA-normalized features as well as on the non-normalized features, they all have accuracy lower than that of the proposed model

by at least 6%. This justifies the necessity of manual feature selection, at least on a relatively small dataset like ACDC. We are not yet able to examine the importance of manual feature selection on large datasets.

To better understand the roles of the features, we further examine the variant with random forest classifiers without manual feature selection trained on the 100 cases of ACDC training set. For each of the 4 classifiers, we compute the feature importance for each of the 9 features to determine the most important one. The importance of a feature is defined as the total reduction of the entropy brought by that feature in all the trees in the random forest. We find that for RVA, HCM and DCM classifiers, the most important features are $R_{RVCLV,ED}$, $R_{LVMLVC,ED}$ and $V_{LVC,ES}$ respectively, which are among the features manually selected for the corresponding proposed classifiers. For MINF classifier, the two most important features are $R_{RVCLV,ED}$ and EF_{LVC} , which have roughly the same importance (0.35 and 0.32). Only EF_{LVC} is used in the proposed model according to its direct relevance. These observations provide further support for our manual feature selection.

4.4.4.6 Variants without Binary Classification

The proposed model divides the 5-category classification task into 4 binary classification sub-tasks. In order to understand whether this special design contributes to the achieved high accuracy, we train and evaluate 2 variants on the same set of 9 features. A random forest and a multi-layer perceptron (MLP) are respectively trained to predict a case to be one of the five categories directly without binary classification (details of these models are available in the appendix section). As reported in Table 4.7, their performances are not as good as that of the proposed model. Hence the strategy of performing a series of binary classification makes sense.

4.5 Conclusion and Discussion

We propose a method of cardiac pathology classification based on originally designed and trained neural networks and classifiers¹. A novel semi-supervised training method is applied to train ApparentFlow-net which provides pixel-wise motion information. Combining the apparent flow generated by ApparentFlow-net and the segmentation masks predicted by LVRV-net, we introduce two novel features that characterize the motion of myocardial segments. These motion-characteristic features are not only intuitive for visualization but also very valuable in classification. The proposed classification model consists of 4 small binary classifiers. Each classifier works independently and takes up to 3 features with clearly explainable relevance as input. On ACDC training set and testing set, the proposed model achieves 95% and 94% respectively as classification accuracy. Its performance is hence comparable to that of the state-of-the-art. To justify our design of the proposed classification model, we also quantitatively compare it with other models.

¹The code and the model will be available in this repository: <https://github.com/julien-zheng/CardiacMotionFlow>

The apparent flow generated by ApparentFlow-net and the originally designed time series of myocardial segment motion are not only straightforward to understand but also useful for classification. We believe that making the automatic methods more understandable and explainable is important, as it is not only helpful to facilitate the implementation and application of the research of medical image analysis in clinics but also useful to improve transparency and to gain trust in medical practice ([Holzinger 2017], [Rueckert 2016]).

Furthermore, the motion information we extract from the apparent flow is fairly rich. We believe that ApparentFlow-net may be a powerful tool of motion extraction for the community. The way we extract the time series and the motion-characteristic features from the flow maps is just one of the so many possible applications. Also, ApparentFlow-net is trained in a semi-supervised manner. This training approach is highly relevant to the current situation of data availability in medical image analysis, as we usually have access to a relatively large amount of unlabeled data and a relatively small amount of labeled data. In a word, much more potential applications in various circumstances of apparent flow are yet to be explored.

Regarding the extraction of the motion-characteristic features, one could use the segmentation network to segment all frames and then derive the motion-characteristic features from the segmentation masks. However, we find that the resulting time series characterizing the cardiac motion (e.g. the time series of the radius and thickness as shown in the second column of Figure 4.6) by this approach are not as temporally consistent as we would expect. In fact, the segmentation network was trained to segment the frames at ED and ES only. And no constraint has ever been imposed to make the segmentation masks temporally consistent. The problem would be clearer if we look at the two frames in the first column in Figure 4.5. While the ED frame (upper image) is easy to segment, the other frame (lower image) appears to be more challenging due to the presence of massive trabeculations. Moreover, as the segmentation network segments the two frames independently, it is not obvious how to ensure the consistency of the segmentation masks. This problem can be solved using the ApparentFlow-net instead to extract motion. As shown in the second column of Figure 4.6, with the ApparentFlow-net, the extracted time series of the radius and thickness of the segments are reasonably smooth, which reflects the enforced temporal consistency.

We could have used existing traditional registration models to supervise the training of ApparentFlow-net or even replace ApparentFlow-net by a deformable registration algorithm (e.g. LDDMM, LCC-Demons, etc.). However, we notice that in order to make the traditional registration models work reasonably well on an unseen dataset like ACDC, the estimation and finetuning of key parameters in these models are necessary. For instance, the authors of [Krebs 2019] empirically estimate the key parameters of LCC-Demons ([Lorenzi 2013]) and SyN ([Avants 2008]) before applying them on ACDC. Our method is simpler in the sense that it learns everything from data and requires no manual model/parameter estimation/adjustment. Hence, on the one hand, our method is easier and more convenient to be applied to various datasets that are reasonably similar to the training set. On the other

hand, it allows us to take advantage of the increasing number of data available in the community. We believe a method with these advantages is very interesting and worth trying. Moreover, as far as we observe, our registration method is accurate enough to characterize the cardiac motion. Some examples are provided in Figure 4.9 to show that the generated apparent flow characterizes the motion patterns of typical cases in several pathological categories. And as shown in the second column of Figure 4.6, with the apparent flow generated by the ApparentFlow-net, the extracted time series of the radius and thickness of the segments enable us to easily distinguish the typical cases of different cardiac pathologies.

A straightforward comparison with prior works on 2D registration methods on the ACDC dataset shows that the ApparentFlow-net performs rather well. Indeed when looking at the Dice coefficients achieved on LVC and RVC, our approach leads to 0.94 and 0.87 respectively (see Table 4.3). In [Hering 2019], the authors describe a learning-based method leading to Dice coefficients at best equal to 0.90 on the same structures (based on Fig.3 of [Hering 2019]) and also performances of a non-learning-based method similar to [Rühaak 2013] with at most 0.80 of Dice. Note however that in this comparison, differences on cross-validation (5-fold v.s. 10-fold), slice selection and ROI determination may hinder the analysis.

While analyzing and extracting the cardiac motion, we adopt a 2D slice-by-slice processing method, without taking the motion on neighboring slices into account. The reason behind this choice is the fact that the inter-slice distance in the short-axis MRIs in ACDC is quite large. Usually, the inter-slice distance between two adjacent slices in MRI stacks is 5 to 10mm. The heart may hence have obviously different shape and motion even on two adjacent slices. In this case, ignoring the neighboring slices for motion estimation might be reasonable. However, if our method is to be applied on some volumetric images with small inter-slice distance, a modification of the approach by taking neighboring slices into account might be beneficial.

An issue that would hinder the generalization of pathology classification models like ours is the lack of a standard and universal definition of pathological category [Suinesiaputra 2016]. For instance, there is another public dataset made available for the MICCAI 2009 challenge on automated LV segmentation [Radau 2009] (the dataset is also known as the Sunnybrook dataset (SD)) containing pathological cases. The 4 categories of SD are heart failure with infarction, heart failure without infarction, LV hypertrophy and healthy. While a hypertrophic case in ACDC has a LV cardiac mass over $110g/m^2$ and several myocardial segments of thickness over 15mm at ED by definition, the hypertrophic cases according to SD definition only need to have a LV cardiac mass over $83g/m^2$. And no threshold is proposed for the myocardial segment thickness by the SD definition. In fact, we find multiple cases in SD which are of LV cardiac mass between $83g/m^2$ and $110g/m^2$ and maximal segment thicknesses well below 15mm. They are identified as hypertrophic cases in SD. But they would not be considered as hypertrophic at all according to ACDC. Similarly, the category of infarction is defined differently in SD and ACDC. In SD, the infarction is determined by the evidence of late gadolinium enhancement; abnormal cardiac motion might not be observable. Yet in ACDC, the infarction

category is defined by the presence of abnormal motion. With such discrepancies between the definitions in different datasets, it is difficult for the community to train a classification model on a dataset such that it generalizes well to the others. We hence appeal for more attention on this issue.

Another issue that may limit the generalization of our classification model is the small size of the ACDC dataset used for training. ACDC training set has only 100 cases of 5 pathological categories. Moreover, in each pathological category, there are only 20 cases. Consequently, on the one hand, many pathological categories are not included in ACDC. On the other hand, for each of the 5 pathological categories in ACDC, we would expect that the 20 cases might not be enough to represent all cases in the category. In order to achieve good generalization, we may need larger datasets with more pathological categories to train the model.

Also, notice that the proposed simple classification model of only 14 parameters is somewhat specific to the ACDC dataset. If we need to adapt our model to perform classification on a larger dataset with more pathological categories, it may be necessary to increase the size and hence the number of parameters of the model.

Finally, we would like to point out that although some single-value hand-crafted motion-characteristic features (e.g. *RMD* and *TMD*) are used in this paper, we believe that for some pathology it would be better to use the whole time series of segment radius or thickness as input to a classification model. For instance, if we aim to discover subtler characteristics related to the motion (e.g. dyssynchrony, septal flash) from a larger dataset, doing so might become appropriate and necessary. We expect to carry out research on this topic in the future.

4.6 Appendix

4.6.1 Loss Function for Training ApparentFlow-Net

To penalize the crossing or large rotations of flows, we compute the difference between of the warped x-components (resp. y-components) of each pair of horizontally (resp. vertically) adjacent pixels \mathbf{P}^{x+} and \mathbf{P} (resp. \mathbf{P}^{y+} and \mathbf{P}). There is a crossing if and only if this difference is smaller than 0, for which a penalty which is equal to the square of this difference applies. Otherwise no penalty applies. Hence we come up with the term $L_{CROSS}(\mathbf{F}_t)$ to penalize the crossing of flows:

$$\begin{aligned}
 & L_{CROSS}(\mathbf{F}_t) \\
 &= \sum_{\mathbf{P}} \min\left((x + 1 + F_t^x(\mathbf{P}^{x+})) - (x + F_t^x(\mathbf{P})), 0\right)^2 \\
 &\quad + \sum_{\mathbf{P}} \min\left((y + 1 + F_t^y(\mathbf{P}^{y+})) - (y + F_t^y(\mathbf{P})), 0\right)^2 \\
 &= \sum_{\mathbf{P}} \min\left(1 + \frac{\partial F_t^x(\mathbf{P})}{\partial x}, 0\right)^2 + \min\left(1 + \frac{\partial F_t^y(\mathbf{P})}{\partial y}, 0\right)^2
 \end{aligned} \tag{4.16}$$

in which $\partial F_t^x(\mathbf{P})/\partial x$ is computed with finite difference as $F_t^x(\mathbf{P}^{x+}) - F_t^x(\mathbf{P})$ (and similarly for $\partial F_t^y(\mathbf{P})/\partial y$).

The Dice function in the term $L_{GT}(\mathbf{F}_{ES})$ is defined on two images U and V as described in [Zheng 2018b]

$$Dice(U, V) = -\frac{2 \sum_{\mathbf{P}} U(\mathbf{P})V(\mathbf{P}) + \varepsilon}{\sum_{\mathbf{P}} U(\mathbf{P}) + \sum_{\mathbf{P}} V(\mathbf{P}) + \varepsilon} \quad (4.17)$$

with $\varepsilon = 1$ a term for better numerical stability in training.

4.6.2 Variants of the Proposed Classification Model with Different Values of Parameter C

We also perform 5-fold cross-validation on the ACDC training set for the variants of the proposed classification model by varying the parameter C in the 4 logistic regression classifiers. Their performances are reported in Table 4.8.

4.6.3 Variants of the Proposed Classification Model with Different Classifiers and Input Features

4.6.3.1 Variants with Other Classifier Models

We replace the proposed ridge logistic regression classifiers by other types of classifiers on the same sets of input features:

- Lasso classifiers: in this variant, each of the 4 classifiers is a least absolute shrinkage and selection operator (Lasso). The constant *alpha* that multiplies the L_1 term is empirically chosen to be 10^{-4} .
- LassoCV classifiers: each of the 4 classifiers is a Lasso model with model selection by cross-validation (LassoCV). The optimal constant *alpha* is searched in the range $[10^{-4}, 10^{-0.5}]$ in a 4-fold cross-validation on the training data.
- random forest classifiers: each of the 4 classifiers is a random forest of 1000 trees which expand their nodes in training until all leaves are pure or all leaves contain less than 2 samples. Entropy is used to measure the quality of a split in training.

4.6.3.2 Variants without Manual Input Feature Selection

We train several variants of the proposed model without manual feature selection. They all consist of 4 classifiers arranged in the same order to perform the same binary classification tasks as in the proposed model. But each of the 4 classifiers in these variants takes all the 9 features together as input. In total, we implement and examine 6 variants with the following models as their classifiers respectively:

- Variant with SVM classifiers: each of the 4 binary classifiers is a support vector machine (SVM) with linear kernel and penalty parameter $C=50$.
- Variant with RVM classifiers: each of the 4 binary classifiers is a relevance vector machine (RVM) as introduced in [Tipping 2003] with linear kernel.
- Variant with Lasso classifiers: each of the 4 binary classifiers is a Lasso. Lasso is

known as a model capable of performing both variable selection and regularization. α , the constant that multiplies the L_1 term, is empirically set to 10^{-4} .

- Variant with LassoCV classifiers: each of the 4 binary classifiers is a Lasso with model selection in a 4-fold cross-validation on the training data. The optimal constant α is searched in the range $[10^{-4}, 10^{-0.5}]$.
- Variant with random forest classifiers: in this variant, each of the 4 binary classifiers is a random forest of 1000 trees which expand their nodes in training until all leaves are pure or all leaves contain less than 2 samples. Entropy is used to measure the quality of a split in training.
- Variant with HDDA classifiers: each of the 4 binary classifiers is a high dimensional discriminant analysis (HDDA) model, which is an expectation-maximization algorithm designed for high-dimensional data clustering based on the ideas of dimension reduction and parsimonious modeling ([Bouveyron 2007], [Orlhac 2018]). Though the 9-feature space in this paper is not high dimensional, we show the performance of such a sophisticated method for comparison.

4.6.3.3 Variants without Binary Classification

We train and evaluate the following 2 variants on all the 9 input features. These variants are obtained by replacing the 4 binary classifiers with a single multi-class one:

- Variant using random forest: it is a random forest of 1000 trees which expand their nodes in training until all leaves are pure or all leaves contain less than 2 samples. Entropy is used to measure the quality of a split in training.
- Variant using MLP: it is a multi-layer perceptron (MLP). It has 2 hidden layers of 32 neurons with tanh activation function. Adam optimizer is used to train it for 10^5 epochs with learning rate 0.001.

4.6.3.4 Implementation of the Variants

Among the above variants of the proposed classification model with different classifiers and input features, the HDDA classifiers are implemented using the HDDA python toolbox downloaded from the GitHub page <https://github.com/mfauvel/HDDA>, the RVM classifiers are implemented in Python according to the method described in [Tipping 2003], and all the other variants are implemented with Scikit-learn.

4.6.4 Examples of Apparent Flow Generated by the ApparentFlow-net

We provide 4 examples of the apparent flow generated by the ApparentFlow-net of 4 ACDC training set cases in different pathological categories. In Figure 4.9, given the frames at ED (first column) and the frames around ES (second column),

we apply the trained ApparentFlow-net to generate the apparent flow maps (third column).

Visually, we find that the apparent flow can indeed characterize the cardiac motion of the typical cases in the pathological categories. As expected, the apparent flow on the LVM of a NOR case is oriented along the gradient of the image intensity and has roughly the same amplitude throughout the left ventricle, signifying the synchronous and uniform contraction and thickening of the LVM of the NOR case. For a HCM cases, we can see that the flow on LVM is excessively large, which means that the contraction and thickening is excessive, a typical phenomenon we find on HCM cases. Conversely, the flow on the LVM of a DCM case is small since the hearts of DCM cases usually do not contract or thicken enough. Finally, the flow on the LVM of a MINF case is not uniform: some myocardial segments contract and thicken much less than the others. This is a typical symptom that we can find on MINF cases.

Table 4.7: The 5-fold cross validation accuracy on the training set of ACDC of the variants of the proposed classification model

Method	BSA-Normalized Features	Non-Normalized Features
logistic regression classifiers (proposed model)	95%	94%
logistic regression classifiers in inversed order	63%	64%
Lasso classifiers	89%	91%
LassoCV classifiers	80%	81%
random forest classifiers	85%	87%
logistic regression classifiers w/o manual feature selection	88%	88%
SVM classifiers w/o manual feature selection	87%	84%
RVM classifiers w/o manual feature selection	88%	72%
Lasso classifiers w/o manual feature selection	85%	86%
LassoCV classifiers w/o manual feature selection	84%	87%
random forest classifiers w/o manual feature selection	86%	88%
HDDA classifiers w/o manual feature selection	49%	46%
one single random forest w/o binary classification	87%	88%
one single MLP w/o binary classification	84%	84%

Table 4.8: The 5-fold cross-validation performance on the ACDC training set of some variants of the proposed classification model with various values of parameter C

C	Training Set Accuracy
1	76%
5	88%
10	92%
50	95%
100	95%
500	93%
1000	93%
5000	93%

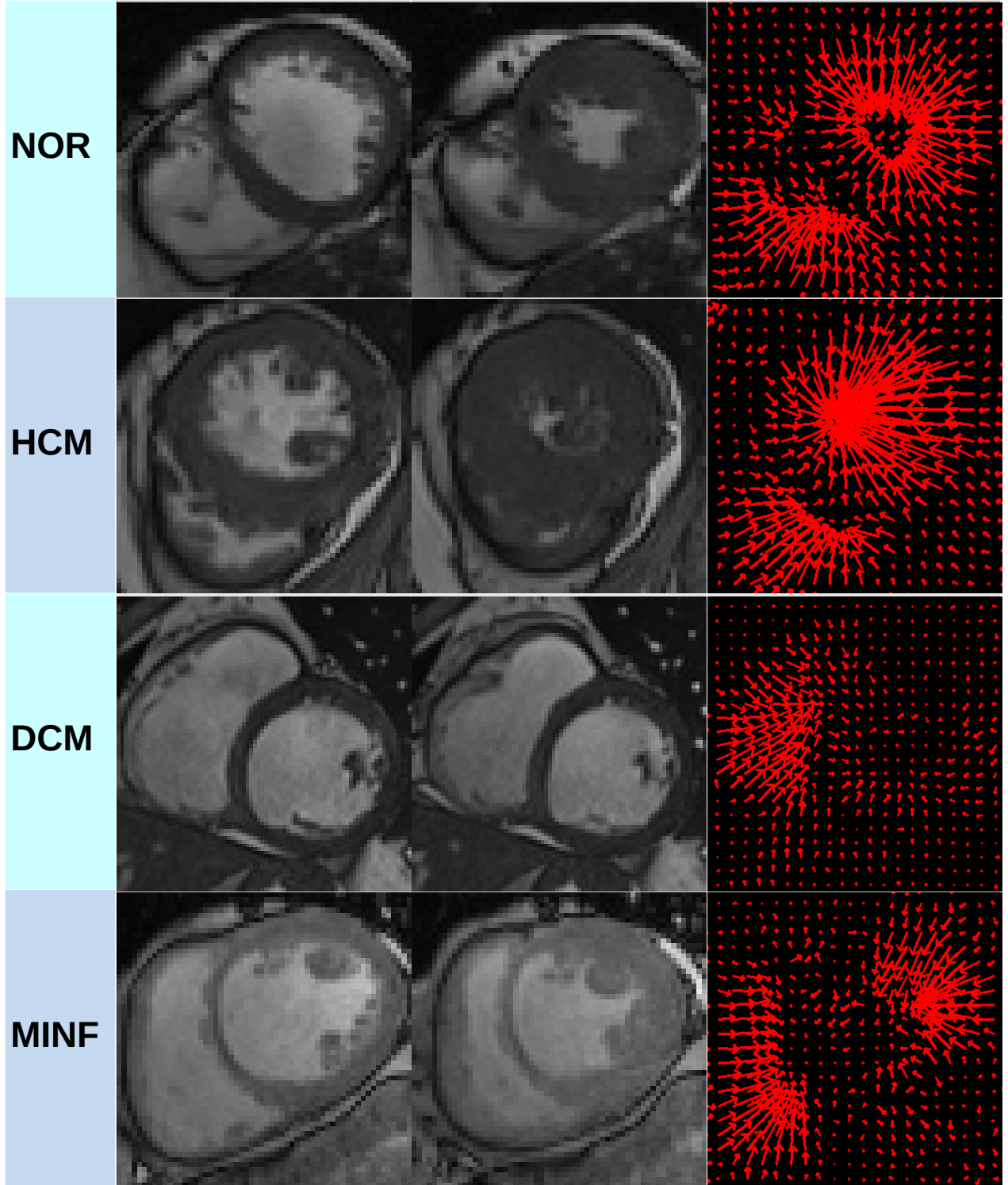


Figure 4.9: Four examples of the apparent flow generated by the ApparentFlow-net of 4 ACDC training set cases in different pathological categories. The images in the first column are the frames at ED. The images in the second column are the frames around ES. The apparent flow maps corresponding to the pairs of frames in the first and second columns are shown in the third column. The apparent flow can indeed characterize the cardiac motion of the typical cases in the pathological categories. NOR: synchronous and uniform flow on LVM; HCM: excessively large flow on LVM; DCM: very small flow on LVM; MINF: asynchronous and ununiform flow on LVM.

Cluster Analysis of Image-Derived Features

Contents

5.1	Introduction	83
5.2	Data	85
5.2.1	UK Biobank	85
5.2.2	ACDC	85
5.3	Methods	86
5.3.1	Feature Extraction	86
5.3.2	Feature Selection	87
5.3.3	Cluster Analysis	87
5.4	Experiments and Results	88
5.4.1	Feature Extraction	88
5.4.2	Feature Selection	88
5.4.3	Cluster Analysis	89
5.4.4	Further Analysis for Confirmation	92
5.5	Conclusion and Discussion	96

Part of this chapter corresponds to the following scientific article:

- [Zheng 2019] **Unsupervised Shape and Motion Analysis of 3822 Cardiac 4D MRIs of UK Biobank**
Qiao Zheng, Hervé Delingette, Kenneth Fung, Steffen E. Petersen and Nicholas Ayache. Submitted to Computerized Medical Imaging and Graphics in February 2019

5.1 Introduction

In recent years, more and more data are made accessible for research in medical image analysis. For instance, the UK Biobank study of [Petersen 2017] has released a dataset containing the cardiac cine MRI images of thousands of volunteers, from which various key cardiovascular functional indexes can be extracted for analysis ([Attar 2019]). The Alzheimer’s Diseases Neuroimaging Initiative (ADNI, [Toga 2015]) has accumulated brain scan images of about two thousand participants.

The abundant data available in the community is certainly a highly valuable resource ([Rueckert 2016], [Suinesiaputra 2016]). Researchers are hence less constrained by the scarcity of data which has been a prevailing challenge for a long time. Further research is necessary ([Zhang 2016], [Barillot 2016]) on new topics associated with big data. For example, one major challenge is how to make good use of unlabeled data ([de Bruijne 2016], [Weese 2016]). In fact, while there are more and more labeled data available, an important part of medical images are still unlabeled. This is understandable as it is in general expensive and tedious to diagnose and label cases by human experts. Methods that can extract useful information from unlabeled data are hence interesting and might potentially save a lot of time and effort.

Many research projects have been developed to perform pathology-related analysis using features extracted from medical images. Many of these works focus on brain scan images. For example, in [Parisot 2018], feature vectors extracted from brain images are used for the prediction of autism spectrum disorder and Alzheimer’s disease. An anatomical landmark based deep feature representation for MRI is proposed in [Liu 2018] for diagnosis of brain disease. Some other studies are based on digital histopathological images. For instance, [Madabhushi 2017] discusses the predictive modeling of digital histopathological images from a detection, segmentation, feature extraction, and tissue classification perspective. [Komura 2018] reviews the machine learning methods for histopathological image analysis. But there are less pathology-related and feature-based researches on cardiac images than on brain scan images and digital histopathological images. And currently, this research ([Zheng 2018a], [Khened 2018], [Khened 2017], [Isensee 2017], [Wolterink 2017], [Cetin 2017]) is mostly about pathology classification in the dataset of Automatic Cardiac Diagnosis Challenge (ACDC) of MICCAI 2017, which contains 100 cases with labels. The work of [Attar 2019] is one of the very first projects to propose a fully automatic, high throughput image parsing workflow for the analysis of cardiac MRI in UK Biobank with systematic tests of the performance. As an extension of the previous works and a challenge to ourselves, we wish to conduct unsupervised analysis on large unlabeled cardiac image datasets.

Clustering, an unsupervised machine learning technique that groups similar entities together, might be suitable for analyzing large unlabeled datasets. Up to now, clustering has been widely used on image segmentation in medical image analysis. For example, the authors of [Kinani 2017] develop a tool based on clustering to outline brain lesion contours. Unsupervised segmentation of 3D lung CT images is proposed in [Moriya 2018] based on clustering and deep representation learning. Some studies show that clustering is also a powerful tool for classification. For instance, a clustering method is applied to classify the analyzed brain images into healthy and multiple sclerosis disease in [Moldovanu 2015]. The authors of [Kawadiwale 2014] introduce various clustering techniques to classify brain MR images into normal and malformed. While most of the application of clustering in the domain is on brain images, we aim to extend its application to cardiac images.

In this paper, we perform a cluster analysis of a group of features extracted from the cardiac MR images of the UK Biobank dataset. Our main contributions

are threefold:

- We conduct a cardiac-pathology-related analysis on a large unlabeled dataset.
- As a novel application of a classic method in medical image analysis, clustering is used in our analysis to group cases without supervision.
- Among the resulting clusters, two can indeed be identified as leaning toward pathological categories.

5.2 Data

5.2.1 UK Biobank

The proposed method was applied to the very large UK Biobank cardiac MRI dataset, see [Petersen 2016]¹. It comprises short-axis cine MRI of about five thousand participants from the general population. More details of the magnetic resonance protocol are available in [Petersen 2016]. Each time series consists of 3D volumes with slice thickness of 8mm for short-axis images and 6mm for long-axis images. The in-plane resolution is $1.8\text{mm} \times 1.8\text{mm}$. Volumes at end-diastole (ED) and end-systole (ES) and ejection fraction for left ventricle cavity (LVC) were derived from InlineVF analysis algorithm ([Jolly 2013], [Lu 2010]) performed by UK Biobank (Field 22421-22422). Those values are considered in this paper as ground-truth (or reference) values. To be consistent with our previous research such as [Zheng 2018b] and [Zheng 2018a], we exclude roughly one thousand cases that are provided with incomplete or unconvincing ground-truth. The remaining 3822 cases are then used for cluster analysis. For part of these cases, the measures of LVC volumes at ED and ES and LVC ejection fraction are provided as ground-truth by UK Biobank.

As pointed out on the website of UK Biobank² and in [Fry 2017], while UK Biobank participants are not representative of the general population with evidence of a ‘healthy volunteer’ selection bias (and hence cannot be used to provide representative disease prevalence and incidence rates), valid assessment of exposure-disease relationships are nonetheless widely generalizable and does not require participants to be representative of the population at large.

5.2.2 ACDC

In the experiment part, we will show the correspondence between some resulting clusters and the definition of some pathology categories defined in the ACDC challenge. Furthermore, a classification model trained on ACDC by [Zheng 2018a] will be applied on UK Biobank for comparison with the clustering method proposed in this paper. The ACDC challenge dataset consists of 100 cases, which are divided into 5 pathological groups of equal size according to their pathology on either the left ventricle (LV) or the right ventricle (RV):

¹Application Number 2964.

²<https://www.ukbiobank.ac.uk/scientists-3/>

Table 5.1: The 9 features generated by our feature extraction method. Among them 8 are selected for cluster analysis.

Feature	Notion	Selected
RVC volume at ED	$V_{RVC,ED}$	yes
LVC volume at ES	$V_{LVC,ES}$	yes
RVC ejection fraction	EF_{RVC}	yes
LVC ejection fraction	EF_{LVC}	no
Ratio between RVC and LV volumes at ED	$R_{RVCLV,ED}$	yes
Ratio between LVM and LVC volumes at ED	$R_{LVMLVC,ED}$	yes
Maximal LVM thickness in all the slices at ED	$MT_{LVM,ED}$	yes
Radius motion disparity	RMD	yes
Thickness motion disparity	TMD	yes

- dilated cardiomyopathy (DCM): left ventricle cavity (LVC) volume at ED larger than 100 mL/m^2 and LVC ejection fraction lower than 40%
- hypertrophic cardiomyopathy (HCM): left ventricle (LV) cardiac mass higher than 110 g/m^2 , several myocardial segments with a thickness higher than 15 mm at ED and a normal ejection fraction
- myocardial infarction (MINF): LVC ejection fraction lower than 40% and several myocardial segments with abnormal contraction
- RV abnormality (RVA): right ventricle cavity (RVC) volume higher than 110 mL/m^2 or RVC ejection fraction lower than 40%
- normal subjects (NOR)

5.3 Methods

There are mainly three steps in the proposed method: feature extraction, feature selection and cluster analysis.

5.3.1 Feature Extraction

The feature extraction method used in this paper is the same as the one proposed in our previous work published by [Zheng 2018a]. We briefly describe its principal steps again below.

The first part of the feature extraction method generates 7 shape-related features. Segmentation with spatial propagation has been proven to be consistent and

robust ([Zheng 2018b], [Zheng 2018c]). With the cardiac segmentation method proposed in [Zheng 2018b], the cardiac images are segmented such that we obtain the masks of LVC, left ventricle myocardium (LVM) and RVC on both ED and ES frames. Then the volumes of LVC, LVM and RVC at both ED and ES can be computed directly, as can the thickness of LVM. Finally, 7 shape-related features are generated (the first 7 terms in Table 5.1).

The second part of the method extracts 2 motion-characteristic features. Using a neural network which outputs apparent flow maps given image pairs, we get a series of apparent flow maps characterizing the in-plane motion for each MRI slice of each case. Combined with the LVM segmentation mask obtained as described above, the motion of each myocardium pixel is hence available. Eventually, 2 features are computed to present the disparity of the radial myocardial motion and the myocardial thickening respectively (the last 2 rows in Table 5.1).

In total, from the images of each case, 9 features characterizing the shape and the motion of the heart are extracted.

5.3.2 Feature Selection

As shown in [Zheng 2018a], these extracted features can be used for cardiac pathology classification in the ACDC dataset with performances comparable to the state-of-the-art. However, these features are not necessarily independent. Some might be redundant if there are highly correlated feature pairs. In cluster analysis, if too many variables are used simultaneously, the redundant ones serve only to create noise that harms the clustering. So it is helpful to select a sub-group of features by removing highly correlated feature pairs.

For each pair among the 9 extracted features, we compute the Pearson correlation coefficient and the maximal information coefficient (MIC) ([Reshed 2011]). The former measures the linear correlation between two features, while the latter measures the mutual information between features. If there is any highly correlated pair according to these measures (i.e. Pearson correlation coefficient of absolute value above 0.8, or MIC above 0.5), we will exclude one feature in this pair. The remaining features are then considered as selected.

5.3.3 Cluster Analysis

We perform a model selection of Gaussian mixture model using the Bayesian information criterion (BIC). Then the selected Gaussian mixture model is applied to cluster the 8 selected features.

5.3.3.1 Gaussian Mixture Model Selection

A Gaussian mixture model ([Reynolds 2009]) is a probabilistic model which assumes that the data points are generated from a mixture of a certain number of Gaussian distributions with unknown parameters. An expectation-maximization algorithm is

used to iteratively estimate its parameters from data. Then the fitted model can assign to each sample the Gaussian component it most likely belongs to.

We use the Gaussian mixture model as implemented in scikit-learn ([Pedregosa 2011]). It has two major parameters, the type of covariance matrix and the number of components, upon which a selection is necessary. For this purpose, we calculate the Bayesian information criterion (BIC, [Wit 2012]) for Gaussian mixture models with different types of covariance matrix and numbers of components. In theory, BIC recovers the true number of components approximately. We fit the Gaussian mixture models with the following types of covariance matrix:

- ‘tied’: all components share the same covariance matrix;
- ‘diag’: each component has its own diagonal covariance matrix;
- ‘full’: each component has its own covariance matrix.

The number of components is also varied. By looking for models with the smallest BIC scores, we wish to select the most simple model that can fit the data thereby identifying the most suitable type of covariance matrix and a range of reasonable numbers of components.

The number of components will finally be determined by examining the sizes of resulting clusters of the Gaussian mixture models. More details will be provided in the Experiments and Results section.

5.3.3.2 Analysis of the Resulting Clusters

The clusters generated by the selected model will be examined. In particular, we verify if the cases in any of the clusters correspond to a pathological category according to the definitions of pathologies given by the ACDC challenge.

5.4 Experiments and Results

5.4.1 Feature Extraction

With the feature extraction method introduced in the Methods section, for each of the 3822 UK Biobank cases, 9 feature values are extracted.

5.4.2 Feature Selection

We calculate the Pearson correlation coefficient and MIC for each pair of features among the 9 extracted features. In Figure 5.1, the plot of Pearson correlation coefficient versus MIC, it is clear that the absolute value of the Pearson correlation coefficient and MIC are positively correlated. There is only one point on the upper left corner of the plot representing a highly correlated pair. It corresponds to $V_{LVC,ES}$ and EF_{LVC} , which are of Pearson correlation coefficient -0.80 and MIC 0.51. The strong negative correlation between these two features is reasonable, since by definition $EF_{LVC} = 1 - V_{LVC,ES}/V_{LVC,ED}$, in which $V_{LVC,ED}$ is the LVC volume at ED. Therefore, $V_{LVC,ES}$ and EF_{LVC} appear to be redundant. Hence we exclude EF_{LVC} and select the remaining 8 features for cluster analysis (Table 5.1).

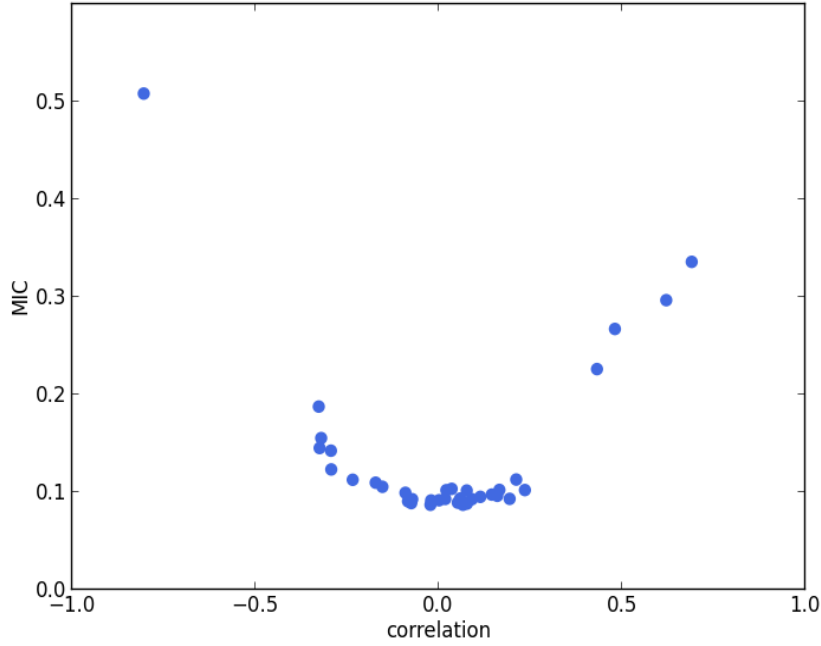


Figure 5.1: Pearson correlation coefficient versus MIC. Each point corresponds to a pair of features. The point in the upper-left corner corresponds to $V_{LVC,ES}$ and EF_{LVC} . The strong negative correlation between these two features is reasonable, since by definition $EF_{LVC} = 1 - V_{LVC,ES}/V_{LVC,ED}$, in which $V_{LVC,ED}$ is the LVC volume at ED.

5.4.3 Cluster Analysis

5.4.3.1 Gaussian Mixture Model Selection

The BIC scores of the Gaussian mixture models with various types of covariance matrix and numbers of components are plotted in Figure 5.2. It is clear that the ‘full’ covariance matrix type is the best among the three. The ‘full’ covariance matrix type is hence selected.

And in terms of the number of components, the Gaussian mixture models with the ‘full’ covariance matrix type of 3 to 10 components have the smallest BIC scores. Among them, we find that:

- The models of 3 to 6 components only generate large clusters, each of which contains at least about one hundred cases.
- The models of 7 and 8 components bring about only one small cluster (less than a dozen cases).
- The models of 9 and 10 components give rise to two small clusters (less than a dozen cases).

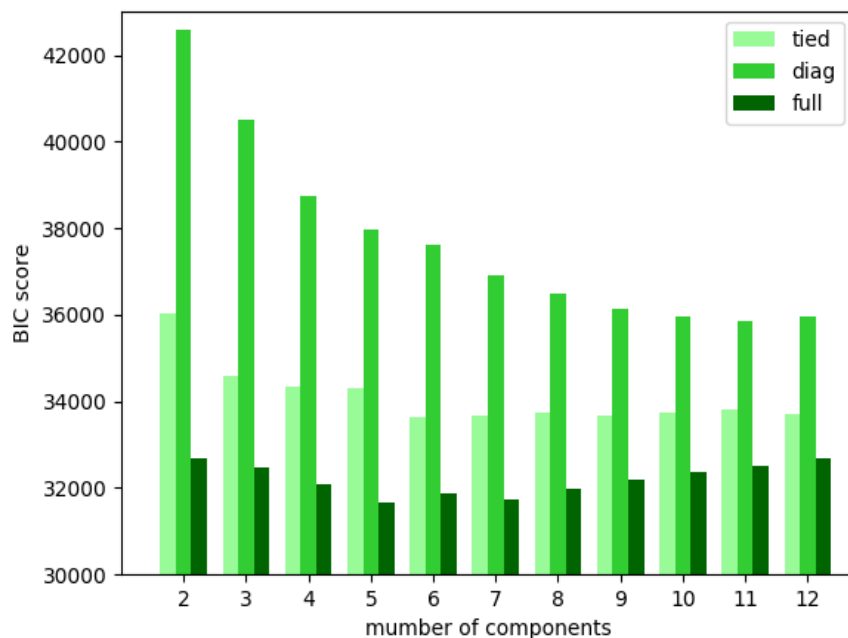


Figure 5.2: BIC scores of Gaussian mixture models with various types of covariance matrix and numbers of components

According to the statistics³ provided by the British Heart Foundation, about 7 million people in the UK are living with cardiovascular diseases, which is about 10.6% of the total population. More specifically, if we look at the most common cardiovascular disease categories, the percentages of UK population living with myocardial infarction, atrial fibrillation and heart failure are about 1.5%, 2.0% and 1.4%, respectively. This means that most of the cases in the general population do not have a cardiac pathology. Taking the ‘healthy volunteer’ selection bias of UK Biobank mentioned in Section 2.1 into account, the cases of cardiovascular diseases are hence probably exceedingly rare in UK Biobank. Thus, if there is any cluster that is related to a specific pathological category in an interpretable manner, its size should be small, say, no more than 76 (2% of the 3822 UK Biobank cases).

So we can now suggest that a component number of 9 or 10 is probably most suitable. We choose the model of 9 components for further analysis. But we would like to point out that the two resulting small clusters of the models of 9 and 10 components are very similar in terms of size and cases. So the results and the conclusions shown below will be roughly the same if we use the model of 10 components.

To summarize, the Gaussian mixture model with the ‘full’ covariance matrix type and 9 components is selected.

³<https://www.bhf.org.uk/what-we-do/our-research/heart-statistics> (accessed January 20, 2019)

Table 5.2: RVC volumes and ejection fraction at ED of the cases of cluster #5 based on our feature extraction method.

ID	RVC volume at ED (mL/m^2)	RVC ejection fraction
2512949	133.13	63.61%
2628396	175.77	43.91%
3423847	140.50	65.24%
3713328	169.65	71.59%
3874816	183.96	56.22%
4366978	134.68	52.53%
4681487	139.82	54.39%
4710306	144.86	29.69%
5101726	145.93	43.82%
5319688	151.30	51.93%
5561149	180.48	41.88%

Table 5.3: LVC volumes at ED and ejection fraction of the cases of cluster #8 based on our feature extraction method (the 2nd and 3rd columns). The same measures provided by the UK Biobank dataset are also shown (the 4th and 5th columns). The two sets of measures are quite close to each other.

ID	LVC volume at ED (mL/m^2)	LVC ejection fraction	Ground-truth LVC volume at ED (mL/m^2)	Ground-truth LVC ejection fraction
2432774	189.28	19.74%	208.24	20%
3378112	213.28	18.75%	213.03	15%
4879002	133.09	27.03%	144.59	29%
5618713	192.87	26.74%	192.43	27%

5.4.3.2 Analysis of the Resulting Clusters

Among the 9 resulting clusters (termed cluster #1 to #9) of the selected model, two are of small sizes (clusters #5 and #8). We find that they actually correspond to two pathological categories according to the definition given by the ACDC challenge (RVA and DCM respectively).

Cluster #5 has 11 cases (examples are given in Figure 5.3). As listed in Table 5.2, these cases have exceptionally large right ventricles, which are above $130 mL/m^2$. In the ACDC challenge, the RVA cases are described as of RVC volumes higher than $110 mL/m^2$ or RVC ejection fraction lower than 40%. Hence according to the definition of ACDC, cluster #5 is a group of cases belonging to RVA.

Cluster #8 has 4 cases (examples are given in Figure 5.3). As shown in Table 5.3, these cases have large LVC volumes at ED (above $130 mL/m^2$) and low LVC ejection fractions (below 30%). In the ACDC challenge, DCM cases are those with

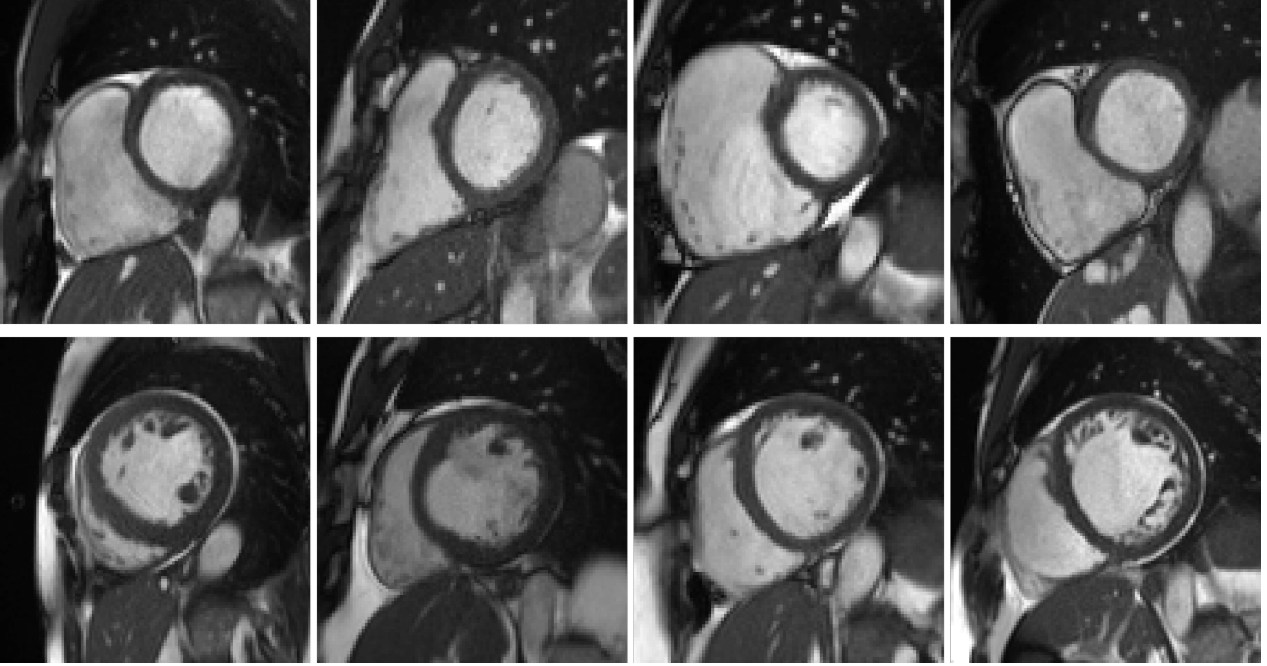


Figure 5.3: Examples of the cases in clusters #5 and #8. First row: example cases in cluster #5, of which the RVs appear to be exceptionally large. Second row: cases in cluster #8, of which the LVs seem to be dilated.

LVC volumes larger than 100 mL/m^2 and LVC ejection fraction lower than 40%. So cluster #8 is a group of DCM cases according to ACDC. In addition, we find that the ground-truth measures of LVC volume at ED and LVC ejection fraction are available for all 4 cases in UK Biobank (last two columns in Table 5.3). It is straightforward to see in Table 5.3 that the measures generated by our feature extraction method are quite close to the ground-truth.

For the other 7 clusters, which are of much larger sizes (above 70), we do not identify any clear correspondence between them and the pathological categories defined in the ACDC challenge.

5.4.4 Further Analysis for Confirmation

To further confirm the discovered correspondence between the two small clusters and the two pathological categories, as well as to verify whether the large clusters represent normal cases, in addition to manual verification of the segmentation masks and apparent flow maps to ensure the exactness of the features, we also conduct the following analysis.

5.4.4.1 Interpretation of the Results of an ACDC Classification Model

We apply a pathology classification model ([Zheng 2018a]) trained using the ACDC dataset on the cases of clusters #5 and #8.

Seven of the eleven cases of cluster #5 are predicted to be RVA, which is as expected. However, the other 4 cases (2512949, 3423847, 4681487 and 5319688) are predicted to be NOR (i.e. normal). We suggest that this is partially due to the difference in the distributions of RVC ejection fraction. In ACDC, a great majority of the RVA cases are of RVC ejection fraction well below 50%. So the trained model has learned to rely on this feature to determine RVA cases. Yet in UK Biobank, some RVA cases, including the 4 listed above, are of RVC ejection fraction above 50%. They are not as severe cases as in ACDC.

All four cases of cluster #8 are predicted to be DCM by the classification model, which supports the correspondence between cluster #8 and DCM. In addition, by manually checking the motion, we can confirm areas of hypokinesia and akinesia for these cases but also dyskinesia for one case (3378112). For case ID 2432774, we also observe discoordinate movement of the LV myocardium suggestive of bundle branch block, which is a type of electrical conduction disease commonly associated with structural heart disease and heart failure. These observations suggest that these cases might also have some relation to MINF. In fact, as pointed out in the ACDC challenge, the increase of LVC volume can be a consequence of the adaptation of LV due to MINF (also called cardiac remodeling).

5.4.4.2 Reduced Dimensionality Visualization Using Principal Component Analysis

To better visualize the two isolated clusters (#5 and #8), we perform a principal component analysis to reduce the dimensionality of the 3822 vectors of size 8 (8 selected features of 3822 cases) of UK Biobank to 2. Furthermore, the centers of the 9 clusters are also projected to the sample space of the 2 principal components. As can be seen in Figure 5.4, the points corresponding to the cases of clusters #5 and #8, as well as the centers of the two clusters, are indeed located far away from most of the other points. This supports the suggestion that the cases in clusters #5 and #8, which are pathological, are quite different from most of the cases in the general population.

5.4.4.3 Visualization using t-SNE

Similarly, another tool to visualize high-dimensional data called t-SNE (t-distributed stochastic neighbor embedding, [van der Maaten 2008]) is applied. Its main advantage is the ability to preserve local structure. So roughly speaking, points which are close to one another in the high-dimensional space will still be close to one another after the dimensionality reduction. t-SNE is applied to the set of the 3822 vectors of the UK Biobank cases, as well as to the set of 3831 vectors which consists of the 3822 UK Biobank cases and the 9 cluster centers. Before applying t-SNE, a normalization is performed for each feature of the original data. The purpose is to make sure that each feature is on the same scale and hence has the same importance in t-SNE. As shown in Figure 5.5, the points of the cases and the centers of clusters

#5 and #8 are at the edge of the ensemble of points in the embedding space. This phenomenon is again consistent with the suggestion that clusters #5 and #8 correspond to pathological cases which are rather different from the other cases in the general population.

5.4.4.4 Examination of the Two Largest Clusters

As pointed out previously, while the pathological categories of clusters #5 and #8 are identifiable, we do not see how the other seven large clusters correspond to any cardiac pathology. In particular, the largest clusters which are of several hundreds or even more cases probably represent groups of normal cases. To verify this, we further examine the two largest clusters (#1 and #4, 889 and 1075 cases, respectively).

We plot the histograms of their ventricle volumes and ejection fractions, as well as their maximal myocardial thicknesses (Figure 5.6). The distributions of #1 and #4 look pretty similar in terms of LVC volume and LVC ejection fraction. But they are different on RVC volume, RVC ejection fraction and maximal myocardial thickness. On average, the cases of #4 have larger RVCs with higher ejection fractions. And their myocardiums also tend to be thicker than that of the cases of #1. Furthermore, we perform the unpaired unequal variance t-test to prove that the corresponding means of the distributions of #1 and #4 are different. Under the null hypotheses that the corresponding distributions have the same mean, the p-values for LVC volume, LVC ejection fraction, RVC volume, RVC ejection fraction and maximal myocardial thickness are all much below 0.05 (lower than 10^{-7}), which are small enough to reject the null hypotheses. This means that clusters #1 and #4 actually exhibit significant different values of the 5 features (LVC volume at ED, LVC ejection fraction, RVC volume at ED, RVC ejection fraction and maximal myocardial thickness).

For both clusters, at least a great majority of the cases satisfy:

- LVC volumes at ED less than $100 \text{ mL}/\text{m}^2$
- LVC ejection fraction above 40%
- RVC volumes at ED less than $110 \text{ mL}/\text{m}^2$
- RVC ejection fraction above 40%
- Maximal myocardial thickness less than 15 mm

Hence according to the definitions in ACDC, these two clusters do not correspond to any of the 4 pathological categories (DCM, HCM, MINF, RVA).

5.4.4.5 Examination of the Seven Large Clusters

To further understand the seven large clusters, we first systematically perform the unpaired unequal variance t-test. For each pair of clusters in the seven large clusters, and for each of the 8 extracted features, under the null hypothesis that the distributions of the feature has the same mean for both clusters, the p-value is computed. In this way $21 \times 8 = 168$ p-values are obtained. In total, 149 p-values among them are below 0.05, which are small enough to reject the corresponding null hypotheses.

Table 5.4: The large p-values of the unpaired unequal variance t-tests for the 21 pairs of clusters in the seven large clusters, and for the 8 extracted features, under the null hypothesis that the distributions of the feature has the same mean for both clusters. For most of the cluster pairs and features, the p-values are below 0.05

cluster pair	p-values above 0.05 (and the corresponding features)
(#1, #4)	0.07 ($V_{LVC,ES}$)
(#1, #6)	0.56 (RMD), 0.05 (TMD)
(#1, #9)	0.55 ($V_{RVC,ED}$), 0.76 ($R_{RVCLV,ED}$)
(#2, #3)	0.17 ($R_{RVCLV,ED}$), 0.80 ($R_{LVMLVC,ED}$)
(#2, #4)	0.31 (TMD)
(#2, #7)	0.85 ($R_{RVCLV,ED}$), 0.76 (RMD)
(#3, #4)	0.29 ($R_{RVCLV,ED}$)
(#3, #6)	0.12 (EF_{RVC})
(#3, #7)	0.07 (EF_{RVC}), 0.28 ($R_{RVCLV,ED}$), 0.61 ($MT_{LVM,ED}$), 0.25 (TMD)
(#4, #6)	0.70 ($R_{LVMLVC,ED}$), 0.14 (TMD)
(#6, #7)	0.27 (EF_{RVC})

Table 5.5: The means and standard deviations of the measures (in mL/m^2) by the automatic pipeline versus the ground-truth.

	Automatic pipeline	Ground-truth
LVC volume at ED (mL/m^2)	70.56 (13.91)	75.48 (28.62)
LVC volume at ES (mL/m^2)	24.06 (9.02)	33.87 (22.82)
LVC ejection fraction	66.41% (7.33%)	56.04% (6.53%)

This confirms that the clusters have different distributions on the features. Nineteen p-values among them are above 0.05, which signify a kind of similarity between pairs of clusters (Table 5.4). Similarly, we perform the unpaired two-sided Mann-Whitney rank tests, under the null hypotheses that the corresponding distributions of the features are the same for both clusters. And we find again that a great majority (147) of the p-values are below 0.05 such that the corresponding null hypotheses can be rejected.

5.4.4.6 Measures by the Automatic Pipeline versus the Ground-Truth

As mentioned previously, for part of the UK Biobank cases, the ground-truth measures given by the InlineVF analysis algorithm of LVC volumes at ED and ES and LVC ejection fraction are available. In particular, among the 3822 cases used in this paper, we have access to all of the three ground-truth measures for 3212 cases. The comparison between the means and standard deviations of the measures generated by the automatic pipeline used in this paper and the ground-truth measures are shown in Table 5.5. It is clear that the ground-truth measures of the volumes

are higher and of larger standard deviations than those estimated by the automatic pipeline.

To better understand the cause of these differences, we plot the points of the measures in Figure 5.7. We can see that the ground-truth values contain some obvious outliers, which are often of values well above the realistic range of LVC volumes. This explains the fact that the ground-truth volumes have higher means and larger standard deviations than those estimated by the automatic pipeline. Moreover, proportionally, the mean of the ground-truth values of LVC volume at ED is 7.0% ($= 75.48/70.56 - 1$) above that of the estimates by the automatic pipeline, while for LVC volume at ES the ground-truth is on average 40.8% ($= 33.87/24.06 - 1$) higher than the values obtained via the automatic pipeline. This also explains why the ground-truth of LVC ejection fraction is on average lower than that given by the automatic pipeline. The models obtained by the robust linear regression using Huber’s criterion for LVC volume at ED and ES are $ground-truth = 1.002 \times automatic-pipeline + 3.373$ and $ground-truth = 0.923 \times automatic-pipeline + 10.303$, respectively. The lines corresponding to the robust linear regression models (red) and the lines corresponding to $ground-truth = automatic-pipeline$ (black) are plotted in Figure 5.7. On both graphs in Figure 5.7, the red line and the black line almost overlap with each other. This means that our regression lines are near the lines of identity, which signifies a similarity between the measures by our method and those based on the InlineVF algorithm. By comparing the regression lines and identity lines in Fig. 4 of [Suinesiaputra 2018], we can also conclude a similarity between the measures derived from manual segmentation and those based on the InlineVF algorithm. Hence our method actually generates measures which are close to both manual and InlineVF values.

We believe that the differences between the measures by the automatic pipeline used in this paper and the ground-truth are partially due to the lack of quality control on the ground-truth. In fact, as pointed out in [Suinesiaputra 2018], the ground-truth is generated by the InlineVF algorithm, which may fail and hence make unreliable predictions on some cases. Without quality control, these failures causes the outliers in Figure 5.7.

5.5 Conclusion and Discussion

In this paper, we proposed a method of unsupervised cluster analysis on a large unlabeled dataset (UK Biobank) of the general population to identify pathological cases based on shape-related and motion-characteristic features extracted from cardiac cine MRI images. As far as we know, this is a topic that has rarely been studied before. In our cluster analysis, a Gaussian mixture model is applied to cluster similar cases together without supervision. As a result, among the generated clusters, we identify two that probably correspond to two cardiac pathological categories. This idea is further supported by the observations on the results of a trained classification model and of the dimensionality reduction tools including principal component analysis and t-SNE.

As more and more large and unlabeled datasets are available in the community, researchers will be able to extract interesting information by data mining. Identification of cardiac pathology is just one among other topics such as the analysis of motion patterns, the relationship between motion and shape features, etc. In the future, more research may be carried out by including more data and different types of data ([Kohli 2017]), using more features, targeting other abnormalities or phenotype properties, etc. Various unsupervised learning methods ([Raza 2018]) other than a Gaussian mixture model can also be applied.

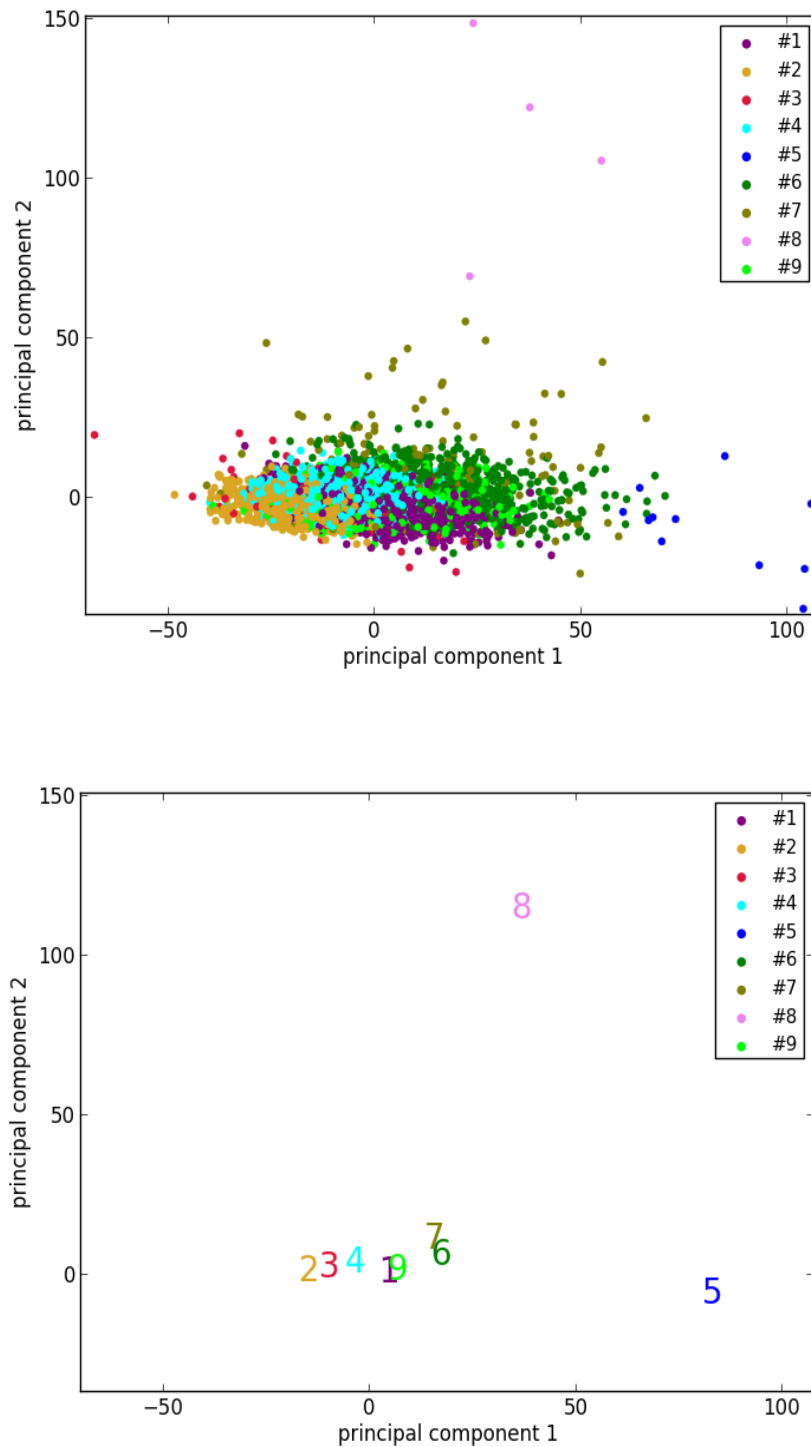


Figure 5.4: The results of dimensionality reduction by principal component analysis. (Upper) The data points of the 3822 UK Biobank cases projected to the space of the 2 principal components. Each data point is colored according to its cluster. (Lower) Projection of the centers (marked by the corresponding indexes and colors) of the 9 clusters to the same space.

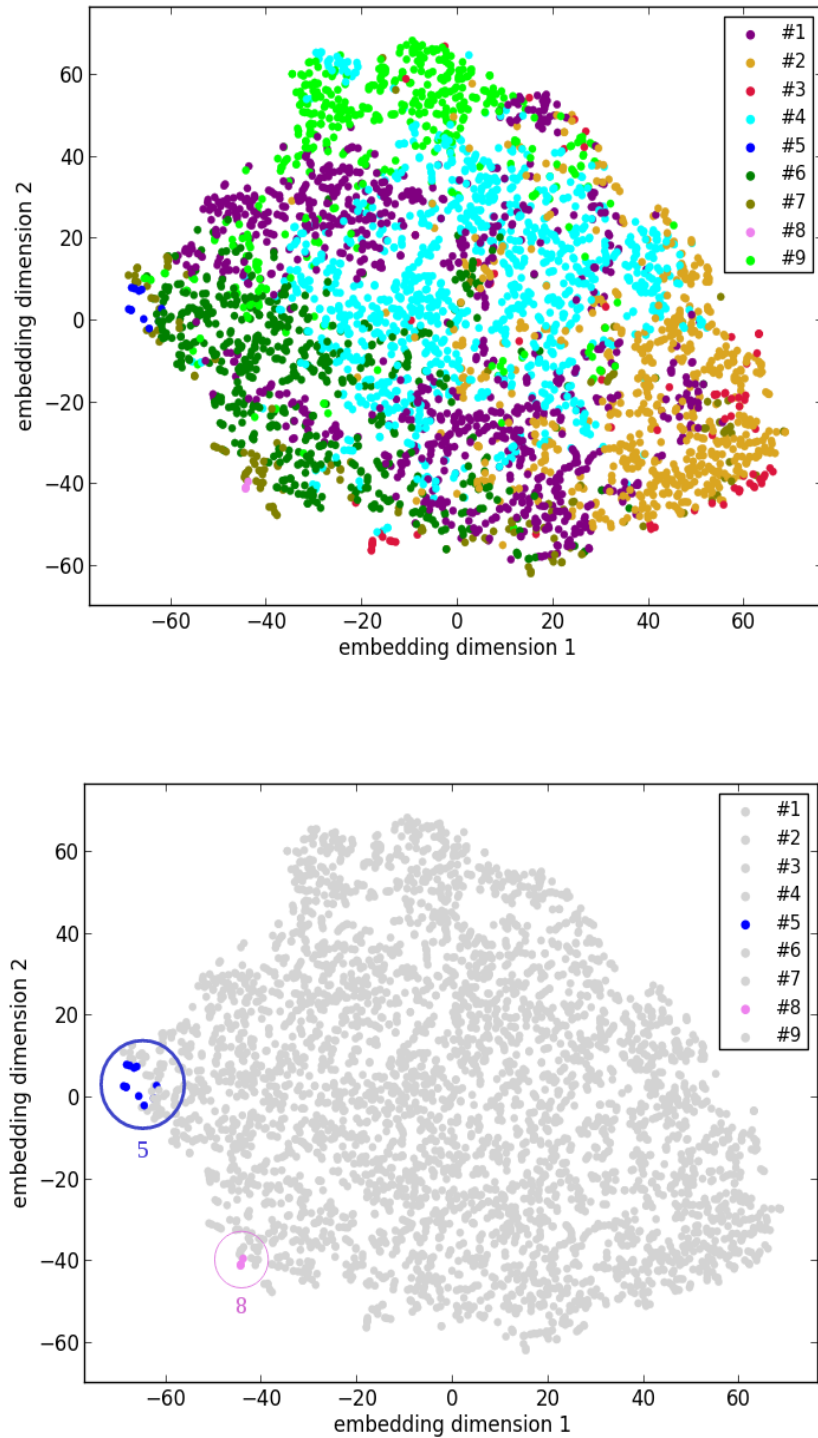


Figure 5.5: The results of dimensionality reduction by t-SNE. (Upper) The data points of the 3822 UK Biobank cases in the space of the 2 embedding dimensions after t-SNE. Each data point is colored according to its cluster. (Lower) A plot similar to the left one with only differences on coloring. Only the points of clusters #5 and #8 are highlighted with colors and circles.

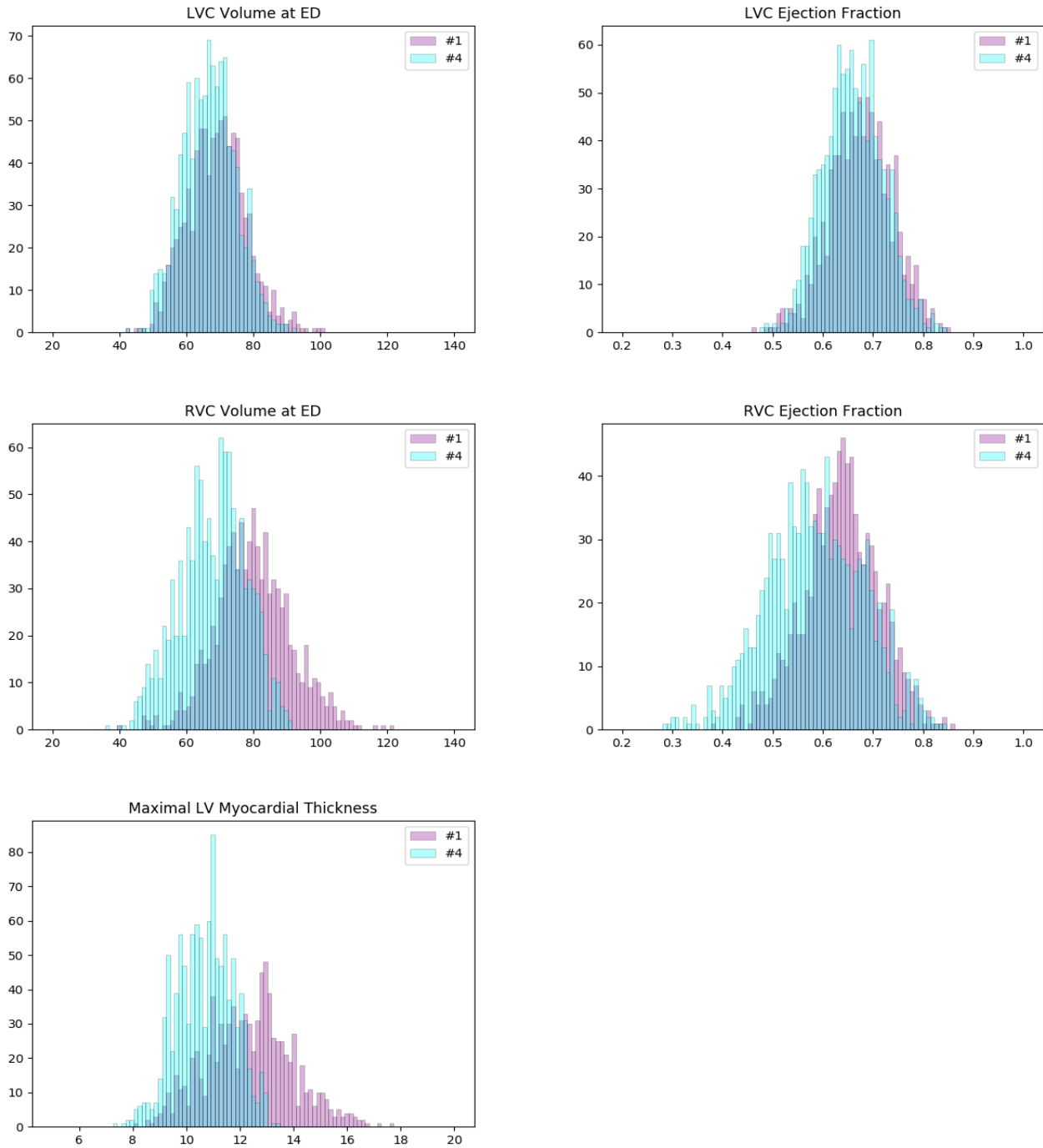


Figure 5.6: Histograms of some important measures of the cases in clusters #1 (pink) and #4 (cyan). The colors of the columns are set to be partially transparent such that their overlaps appear to be of color dark blue. The distributions of #1 and #4 are pretty similar in terms of LVC volume and LVC ejection fraction (1st row). But they are different on RVC volume, RVC ejection fraction and maximal myocardial thickness (2nd and 3rd rows). On average, the cases of #1 have larger RVCs with higher ejection fractions. And their myocardiums also tend to be thicker than that of the cases of #4. For both clusters, the measures are well in normal ranges according to the definitions given by ACDC.

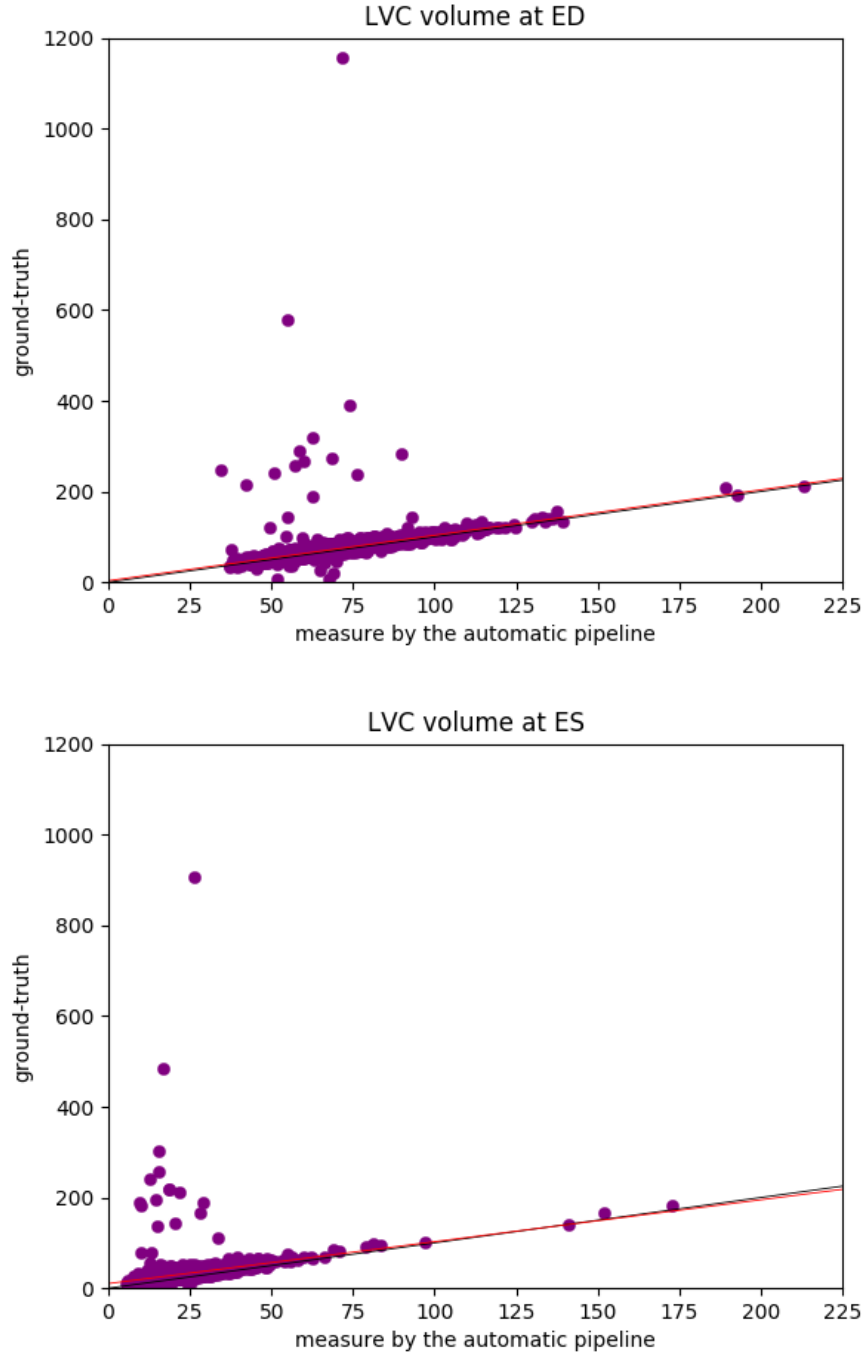


Figure 5.7: The plots of the measures (in mL/m^2) generated by the automatic pipeline against the ground-truth for the LVC volume at ED (upper) and at ES (lower). We can see that the ground-truth values contain some obvious outliers, which are often of values well above the realistic range of LVC volumes. This explains the fact that the ground-truth volumes have higher means and larger standard deviations than those estimated by the automatic pipeline. The lines corresponding to the robust linear regression models (red) and the lines corresponding to $ground-truth=automatic-pipeline$ (black) are also plotted. The red line and the black line almost overlap with each other.

Conclusion and Perspectives

Contents

6.1 Main Contributions	103
6.1.1 Segmentation Using Simulation for Data Augmentation	103
6.1.2 Consistent and Robust Segmentation with Spatial Propagation	104
6.1.3 Explainable Pathology Classification with Motion Characterization	104
6.1.4 Cluster Analysis of Image-Derived Features	104
6.2 Publications	105
6.3 Software	105
6.4 Perspectives	105
6.4.1 Cardiac Mesh Simulation and Image Synthesis for Deep Learning	105
6.4.2 Temporal Consistency of Segmentation	106
6.4.3 Semi-Supervised Learning and Unsupervised Learning	106
6.4.4 More Explainable Models	107

In this thesis, we explored deep learning for robust segmentation and explainable analysis of 3D and dynamic cardiac images. Now we summarize the main contributions and discuss some perspectives.

6.1 Main Contributions

6.1.1 Segmentation Using Simulation for Data Augmentation

In **Chapter 2** and the corresponding study [Zheng 2018c], we show how deep learning can be applied to learn biventricular segmentation from a small dataset, using an existing method of cardiac mesh simulation and image synthesis for large-scale data augmentation. In addition to this data augmentation strategy, the novel spatial segmentation propagation and multi-scale coarse-to-fine networks are proposed. Furthermore, we demonstrate that our method is robust, and combines the assets of 2D (speed) and 3D to provide spatially consistent meshes ready to be used for simulations. Last but not least, a novel loss function is also proposed to overcome class imbalance.

6.1.2 Consistent and Robust Segmentation with Spatial Propagation

In **Chapter 3** and the corresponding publication [Zheng 2018b], we propose a method of segmentation with spatial propagation that is based on originally designed neural networks. By taking the contextual input into account, the spatial consistency of segmentation is enforced. Also, we conduct thorough and unprecedented testing to evaluate the robustness and generalization ability of our model and achieve performances better than or comparable to the state-of-the-art. Furthermore, an exceptionally large dataset (UK Biobank) collected from the general population is used for training and evaluation, which makes the reported results more convincing.

6.1.3 Explainable Pathology Classification with Motion Characterization

In **Chapter 4** and the corresponding publication [Zheng 2018a], we propose a method of cardiac pathology classification based on originally designed and trained neural networks and classifiers. A novel semi-supervised training method is applied to train a network (ApparentFlow-net), which provides pixel-wise motion information. Combining the generated apparent flow and the segmentation masks predicted by another network (LVRV-net), we introduce two novel features that characterize the motion of myocardial segments. These motion-characteristic features are not only intuitive for visualization but also very valuable in classification. The proposed classification model consists of 4 small binary classifiers. Each classifier works independently and takes up to 3 features with clearly explainable relevance as input. On the training and testing datasets (ACDC), the proposed model achieves 95% and 94% respectively as classification accuracy. Its performances are hence comparable to that of the state-of-the-art. To justify our design of the proposed classification model, we also quantitatively compare it with other models.

6.1.4 Cluster Analysis of Image-Derived Features

In **Chapter 5**, we proposed a method of unsupervised cluster analysis on an unlabeled dataset (UK Biobank) of 3822 cases to identify pathological cases based on shape-related and motion-characteristic features extracted from cardiac cine MRI images. As far as we know, this is a topic that has rarely been studied before. In our cluster analysis, a Gaussian mixture model is applied to cluster similar cases together without supervision. As a result, among the generated clusters, we identify two that probably correspond to two cardiac pathological categories respectively. This idea is further supported by the observations on the results of a trained classification model and of the dimensionality reduction tools.

6.2 Publications

This thesis led to several published and submitted publications:

- [Zheng 2018c] **3D Consistent Biventricular Myocardial Segmentation Using Deep Learning for Mesh Generation**
Qiao Zheng, Hervé Delingette, Nicolas Duchateau and Nicholas Ayache. arXiv preprint, 2018
- [Zheng 2018b] **3D Consistent and Robust Segmentation of Cardiac Images by Deep Learning with Spatial Propagation**
Qiao Zheng, Hervé Delingette, Nicolas Duchateau and Nicholas Ayache. IEEE Transactions on Medical Imaging, 2018
- [Zheng 2018a] **Explainable Cardiac Pathology Classification on Cine MRI with Motion Characterization by Semi-Supervised Learning of Apparent Flow**
Qiao Zheng, Hervé Delingette and Nicholas Ayache. Submitted to Medical Image Analysis in November 2018, under minor revision in February 2019
- [Zheng 2019] **Unsupervised Shape and Motion Analysis of 3822 Cardiac 4D MRIs of UK Biobank**
Qiao Zheng, Hervé Delingette, Kenneth Fung, Steffen E. Petersen and Nicholas Ayache. Submitted to Computerized Medical Imaging and Graphics in February 2019

6.3 Software

This thesis led to the development of the following software:

- **CardiacSegmentationPropagation**
CardiacSegmentationPropagation is a Python-based tool for cardiac image segmentation. Using deep learning with the spatial propagation of the segmentation, the method is both consistent and robust. A package containing the code of the method, the pre-trained weights of the model, and the instructions for use can be downloaded from the website of the Epione team (<https://team.inria.fr/epione/en/software/>).

6.4 Perspectives

6.4.1 Cardiac Mesh Simulation and Image Synthesis for Deep Learning

In Chapter 2, we demonstrate how an existing model of cardiac mesh simulation and image synthesis ([Duchateau 2016]) can be used for data augmentation to enable and improve deep learning. However, we notice that simulation and synthesis

models can do much more for deep learning than just this. As pointed out in [Duchateau 2018], based on realistic electromechanical modeling and mesh simulation of the heart, pathological cardiac sequences can be synthesized from real healthy sequences. With such a method as a tool, on the one hand, researchers may generate many pathological cardiac image sequences of various pathological categories, which may be used for learning cardiac pathology classification, detection and localization. In fact, collecting and labeling real cardiac images of pathological cases by human experts is both expensive and tedious. Synthesized pathological images are hence great alternatives of the real ones. In addition, with such a generation process, the synthesized images of a case are associated with a known electromechanical model, which contains many known parameters, which characterize the biophysical factors of the heart (e.g. conductivity, stiffness, contraction, relaxation). Therefore, the electromechanical model associated with the synthesis method can provide a big set of information which might not be available on real images. So combining cardiac mesh simulation and image synthesis with deep learning is research direction with bright prospects.

6.4.2 Temporal Consistency of Segmentation

As presented in Chapter 3, the spatial consistency of segmentation can be reinforced via the spatial propagation of segmentation masks. It is natural to ask whether the temporal consistency of segmentation can also be improved in a similar way. As a matter of fact, the temporal consistency of segmentation is important in the estimation of the cardiac motion, which often characterizes cardiac pathologies as indicated in Chapter 4. Moreover, without the temporal consistency of segmentation, some essential cardiac measures such as the ejection fractions of the left and right ventricles cannot be determined accurately. However, as the ground-truth of segmentation is usually available only at the instants of the end-diastole and end-systole in most of the datasets, it might be practically difficult to train a deep learning model to propagate segmentation across time in a supervised manner as presented in Chapter 3. Fortunately, now we can already see some possible methods to tackle this problem. For instance, incorporation of anatomical properties, such as the shape and location of an organ, has been proved to be helpful in cardiac image enhancement and segmentation ([Oktay 2018]). We thus expect that taking prior knowledge into account in deep learning would make the segmentation more consistent both spatially and temporally.

6.4.3 Semi-Supervised Learning and Unsupervised Learning

In Chapter 4 and Chapter 5, we demonstrate the application of semi-supervised learning and unsupervised learning in medical image analysis respectively. We believe this is a promising research topic. The reality is that there are much more unannotated medical images than annotated ones. Instead of using only the medical images with ground-truth labels, segmentation or annotation, if researchers can

also make good use of unannotated images to train their models in semi-supervised or even unsupervised ways, they might significantly improve the performance the models. The problem of temporal consistency of segmentation discussed above is one of the many possible subjects on which semi-supervised or unsupervised learning may be useful ([Raza 2018], [Cheplygina 2018], [Aganj 2018], [Bai 2017a]). Furthermore, sometimes learning from a large amount of unannotated data may enable a model to discover previously unknown or unnoticed information. The pathological clusters identified by unsupervised learning in Chapter 5 is just one simple example. And this character of semi-supervised learning and unsupervised learning make them especially interesting in the current era of big data.

6.4.4 More Explainable Models

In Chapter 4, we draw attention to the problem of explainability of learning-based models. Machine learning models, in particular, deep learning models, are not easy to interpret. This is because most of the machine models contain at least hundreds of parameters and it is practically infeasible to examine and explain the role of each parameter. Furthermore, as many values or features are used simultaneously as the input of the model, it is hard to tell in a straightforward manner whether and how and to what degree they contribute to the results of the model respectively. This drawback on explainability may cause many problems as suggested in [Holzinger 2017]. For instance, the lack of explainability is an obvious hurdle for the wide adoption of learning-based models in the clinic despite their performance. Moreover, under the new European General Data Protection Regulation¹, they may also generate legal and privacy issues in business. Hence, as a proof of concept, in Chapter 4 we propose a simple classification model with a small number of input features and parameters such that the role and contribution of each feature or parameter are clear and explainable. Fortunately, nowadays, some other researchers are also already interested in building explainable artificial intelligence models for medicine ([Lamy 2019], [Herent 2018]). More research efforts on this topic are necessary and expected.

¹<https://eur-lex.europa.eu/eli/reg/2016/679/oj> (accessed January 30, 2019)

Bibliography

- [Adelman 2014] R Adelman, L Tmanova, D Delgado, S Dion and M Lachs. *Care-giver burden: a clinical review*. JAMA, vol. 311, pages 1052–1060, 2014. (Cited on page 2.)
- [Aganj 2018] I Aganj, M Harisinghani, R Weissleder and B Fischl. *Unsupervised medical image segmentation based on the local center of mass*. Scientific Reports, vol. 8, 2018. (Cited on page 107.)
- [Al-Shamshi 2017] M Al-Shamshi. *Addressing the physicians’ shortage in developing countries by accelerating and reforming the medical education: is it possible?* J Adv Med Educ Prof., vol. 5(4), pages 209–212, 2017. (Cited on page 1.)
- [Attar 2019] R Attar, M Pereanez, A Gooya, X Alba, L Zhang, S Piechnik, S Neubauer, S Petersen and A Frangi. *High throughput computation of reference ranges of biventricular cardiac function on the UK Biobank population cohort*. arXiv preprint arXiv:1901.03326, 2019. (Cited on pages 83 and 84.)
- [Avants 2008] B Avants, C Epstein, M Grossman and J Gee. *Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain*. Med Image Anal, vol. 12(1), pages 26–41, 2008. (Cited on page 73.)
- [Avendi 2016a] M Avendi, A Kheradvar and H Jafarkhani. *A combined deeplearning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI*. Med Image Anal, vol. 30, pages 108–109, 2016. (Cited on pages 5 and 41.)
- [Avendi 2016b] M Avendi, A Kheradvar and H Jafarkhani. *Fully automatic segmentation of heart chambers in cardiac MRI using deep learning*. J Cardiovasc Magn Reson, vol. 18, pages 351–353, 2016. (Cited on page 45.)
- [Avendi 2017] M Avendi, A Kheradvar and H Jafarkhani. *Automatic segmentation of the right ventricle from cardiac MRI using a learning-based approach*. Magn. Reson. Med., vol. 78(6), pages 2439–2448, 2017. (Cited on page 5.)
- [Bai 2017a] W Bai, O Oktay, M Sinclair, M Suzuki, M Rajchl, G Tarroni, B Glocker, A King, P Matthews and D Rueckert. *Semi-supervised learning for network-based cardiac MR image segmentation*. MICCAI, pages 253–260, 2017. (Cited on pages 50 and 107.)
- [Bai 2017b] 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, S Piechnik, S Neubauer, B Glocker and D Rueckert. *Human-level CMR image analysis with deep fully*

- convolutional networks*. arXiv preprint arXiv:1710.09289, 2017. (Cited on pages 35 and 45.)
- [Balakrishnan 2018] G Balakrishnan, A Zhao, M Sabuncu, J Guttag and A Dalca. *VoxelMorph: a Learning framework for deformable medical image registration*. arXiv preprint arXiv:1809.05231, 2018. (Cited on page 56.)
- [Barillot 2016] C Barillot, G Edan and O Commowick. *Imaging biomarkers in multiple sclerosis: from image analysis to population imaging*. Medical Image Analysis, vol. 33, pages 134–139, 2016. (Cited on page 84.)
- [Baumgartner 2017] C Baumgartner, L Koch, M Pollefeys and E Konukoglu. *An exploration of 2D and 3D deep learning techniques for cardiac MR image segmentation*. In Proc. Statistical Atlases and Computational Models of the Heart (STACOM), ACDC challenge, MICCAI’17 Workshop, 2017. (Cited on page 18.)
- [Bernard 2018] O Bernard, A Lalande, C Zotti, F Cervenansky, X Yang, P Heng, I Cetin, K Lekadir, O Camara, M Ballester, G Sanroma, S Napel, S Petersen, G Tziritas, E Grinias, M Khened, V Kollerathu, G Krishnamurthi, M Rohé, X Pennec, M Sermesant, F Isensee, P Jäger, K Maier-Hein, C Baumgartner, L Koch, J Wolterink, I Isgum, Y Jang, Y Hong, J Patravali, S Jain, O Humbert and P Jodoin. *Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?* IEEE Trans Med Imaging, vol. 37(11), pages 2514–2525, 2018. (Cited on pages 51 and 53.)
- [Bouveyron 2007] C Bouveyron, S Girard and C Schmid. *High dimensional data clustering*. Computational Statistics and Data Analysis, vol. 52, pages 502–519, 2007. (Cited on page 77.)
- [Cerqueira 2002] M Cerqueira, N Weissman, V Dilsizian, A Jacobs, S Kaul, W Laskey, D Pennell, J Rumberger, T Ryan and M Verani. *Standardized Myocardial Segmentation and Nomenclature for Tomographic Imaging of the Heart: A Statement for Healthcare Professionals from the Cardiac Imaging Committee of the Council on Clinical Cardiology of the American Heart Association*. Circulation, 2002. (Cited on page 58.)
- [Cetin 2017] I Cetin, G Sanroma, S Petersen, S Napel, O Camara, M Ballester and K Lekadir. *A radiomics approach to computer-aided diagnosis with cardiac cine-MRI*. In Proc. Statistical Atlases and Computational Models of the Heart (STACOM), ACDC challenge, MICCAI’17 Workshop, 2017. (Cited on pages 50, 51, 67 and 84.)
- [Chen 2016] J Chen, L Yang, Y Zhang, M Alber and D Chen. *Combining fully convolutional and recurrent neural networks for 3D biomedical image segmentation*. Advances in Neural Information Processing Systems, pages 3036–3044, 2016. (Cited on page 2.)

- [Cheng 2016] J Cheng, D Ni, Y Chou, J Qin, C Tiu, Y Chang, C Huang, D Shen and C Chen. *Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans*. Nat Sci Rep, vol. 6:24454, 2016. (Cited on page 2.)
- [Cheplygina 2018] V Cheplygina, M de Bruijne and J Pluim. *Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis*. arXiv preprint arXiv:1804.06353, 2018. (Cited on pages 50 and 107.)
- [Comaniciu 2016] D Comaniciu, K Engel, B Georgescu and T Mansi. *Shaping the future through innovations: from medical imaging to precision medicine*. Med Image Anal, vol. 33, pages 19–26, 2016. (Cited on page 50.)
- [Dalmis 2017] M Dalmis, G Litjens, K Holland, A Setio, R Mann, N Karssemeijer and A Gubern-Mérida. *Using deep learning to segment breast and fibroglandular tissue in MRI volumes*. Medical physics, vol. 44(2), pages 533–546, 2017. (Cited on page 2.)
- [Dawes 2017] T Dawes, A de Marvao, W Shi, T Fletcher, G Watson, W Wharton, C Rhodes, L Howard, J Gibbs, D Rueckert, S Cook, M Wilkins and O’Regan D. *Machine learning of three-dimensional right ventricular motion enables outcome prediction in pulmonary hypertension: a cardiac MR imaging study*. Radiology, vol. 283, pages 381–390, 2017. (Cited on page 51.)
- [de Bruijne 2016] M de Bruijne. *Machine learning approaches in medical image analysis: from detection to diagnosis*. Medical Image Analysis, vol. 33, pages 94–97, 2016. (Cited on page 84.)
- [de Vos 2017] B de Vos, F Berendsen, M Viergever, M Staring and I Isgum. *End-to-end unsupervised deformable image registration with a convolutional neural network*. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, vol. 10553, pages 204–212, 2017. (Cited on page 56.)
- [Duchateau 2016] N Duchateau, M De Craene, P Allain, E Saloux and M Sermesant. *Infarct localization from myocardial deformation: Prediction and uncertainty quantification by regression from a low-dimensional space*. IEEE Trans. Med. Imaging, vol. 35(10), pages 2340–2352, 2016. (Cited on pages 12 and 105.)
- [Duchateau 2018] N Duchateau, M Sermesant, H Delingette and N Ayache. *Model-based generation of large databases of cardiac images: synthesis of pathological cine MR sequences from real healthy cases*. IEEE Trans. Med. Imaging, vol. 37, pages 755–766, 2018. (Cited on page 106.)
- [Ecabert 2008] O Ecabert, J Peters, H Schramm, C Lorenz, J von Berg, M Walker, M Vembar, M Olszewski, K Subramanyan, G Lavi and J Weese. *Automatic*

- model-based segmentation of the heart in CT images.* IEEE Trans. Med. Imaging, vol. 27(9), pages 1189–2201, 2008. (Cited on page 6.)
- [Forsberg 2017] D Forsberg, E Sjöblom and J Sunshine. *Detection and labeling of vertebrae in MR images using deep learning with clinical annotations as training data.* J Digit Imaging, vol. 30(4), pages 406–412, 2017. (Cited on page 2.)
- [Fry 2017] A Fry, T Littlejohns, C Sudlow, N Doherty, L Adamska, T Sprosen, R Collins and N Allen. *Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population.* Am J Epidemiol., vol. 186(9), pages 1026–1034, 2017. (Cited on page 85.)
- [Gao 2016] B Gao, W Liu, L Wang, P Liu, P Croisille, P Delachartre and P Clarysse. *Estimation of cardiac motion in cine-MRI sequences by correlation transform optical flow of monogenic features distance.* Phys Med Biol, vol. 61, pages 8640–8663, 2016. (Cited on page 50.)
- [Gilbert 2017] K Gilbert, B Pontre, C Occleshaw, B Cowan, A Suinesiaputra and A Young. *4D modelling for rapid assessment of biventricular function in congenital heart disease.* The International Journal of Cardiovascular Imaging, vol. 34(3), pages 407–417, 2017. (Cited on page 51.)
- [Gu 2017] L Gu, Y Zheng, R Bise, I Sato, N Imanishi and S Aiso. *Semi-supervised learning for biomedical image segmentation via forest oriented super pixels(voxels).* MICCAI, pages 702–710, 2017. (Cited on page 50.)
- [Han 2016] X Han, J Lei and Y Chen. *HEp-2 cell classification using K-support spatial pooling in deep CNNs.* DLMIA, vol. 10008, pages 3–11, 2016. (Cited on page 2.)
- [Herent 2018] P Herent, S Jegou, G Wainrib and T Clozel. *Brain age prediction of healthy subjects on anatomic MRI with deep learning: going beyond with an “explainable AI” mindset.* bioRxiv preprint 413302, 2018. (Cited on page 107.)
- [Hering 2019] A Hering, S Kuckertz, S Heldmann and M Heinrich. *Enhancing label-driven deep deformable image registration with local distance metrics for state-of-the-art cardiac motion tracking.* Bildverarbeitung für die Medizin 2019, pages 309–314, 2019. (Cited on page 74.)
- [Hernandez 2008] M Hernandez, S Olmos and X Pennec. *Comparing algorithms for diffeomorphic registration: stationary LDDMM and diffeomorphic demons.* Proc. MFCA 2008, pages 24–35, 2008. (Cited on page 56.)

- [Holzinger 2017] A Holzinger, C Biemann, C Pattichis and D Kell. *What do we need to build explainable AI systems for the medical domain?* arXiv preprint arXiv:1712.09923, 2017. (Cited on pages 50, 73 and 107.)
- [Isensee 2017] F Isensee, P Jaeger, P Full, I Wolf, S Engelhardt and K Maier-Hein. *Automatic cardiac disease assessment on cine-MRI via time-series segmentation and domain specific features*. In Proc. Statistical Atlases and Computational Models of the Heart (STACOM), ACDC challenge, MICCAI'17 Workshop, 2017. (Cited on pages 18, 36, 40, 50, 51, 67 and 84.)
- [Jang 2017] Y Jang, S Ha, S Kim, Y Hong and H Chang. *Automatic segmentation of LV and RV in cardiac MRI*. In Proc. Statistical Atlases and Computational Models of the Heart (STACOM), ACDC challenge, MICCAI'17 Workshop, 2017. (Cited on pages 36 and 40.)
- [Jolly 2013] M Jolly, C Guetter, X Lu, H Xue and J Guehring. *Automatic segmentation of the myocardium in cine MR images using deformable registration*. Statistical Atlases and Computational Models of the Heart. Imaging and Modelling Challenges, pages 98–108, 2013. (Cited on page 85.)
- [Kawadiwale 2014] R Kawadiwale and M Rane. *Clustering techniques for brain tumor detection*. Proc. of Int. Conf. on Recent Trends in Information, pages 299–305, 2014. (Cited on page 84.)
- [Kayalibay 2017] B Kayalibay, G Jensen and P van der Smagt. *CNN-based segmentation of medical imaging data*. arXiv preprint arXiv:1701.03056, 2017. (Cited on page 26.)
- [Khened 2017] M Khened, V Alex and G Krishnamurthi. *Densely connected fully convolutional network for short-axis cardiac cine MR image segmentation and heart diagnosis using random forest*. In Proc. Statistical Atlases and Computational Models of the Heart (STACOM), ACDC challenge, MICCAI'17 Workshop, 2017. (Cited on pages 50, 51, 67, 68 and 84.)
- [Khened 2018] M Khened, V Alex and G Krishnamurthi. *Fully convolutional multi-scale residual DenseNets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers*. arXiv preprint arXiv:1801.05173, 2018. (Cited on pages 50, 51, 67, 68 and 84.)
- [Kinani 2017] J Kinani, A Silva, F Funes, D Vargas, E Diaz and A Arellano. *Medical imaging lesion detection based on unified gravitational fuzzy clustering*. Journal of Healthcare Engineering, vol. 2017, 2017. (Cited on page 84.)
- [Kohli 2017] M Kohli, R Summers and J Geis. *Medical image data and datasets in the era of machine learning-whitepaper from the 2016 C-MIMI meeting dataset session*. Journal of Digital Imaging, vol. 30, pages 392–399, 2017. (Cited on page 97.)

- [Komura 2018] D Komura and S Ishikawa. *Machine learning methods for histopathological image analysis*. Computational and Structural Biotechnology, vol. 16, pages 34–42, 2018. (Cited on page 84.)
- [Krebs 2018] J Krebs, T Mansi, B Mailhé, N Ayache and H Delingette. *Unsupervised probabilistic deformation modeling for robust diffeomorphic registration*. In Proc. Deep Learning in Medical Image Analysis (DLMIA), MICCAI’18 Workshop, 2018. (Cited on page 56.)
- [Krebs 2019] J Krebs, H Delingette, B Mailhé, N Ayache and T Mansi. *Learning a probabilistic model for diffeomorphic registration*. IEEE Transactions on Medical Imaging, 2019. (Cited on page 73.)
- [Lamy 2019] J Lamy, B Sekar, G Guezennec, J Bouaud and B Séroussi. *Explainable artificial intelligence for breast cancer: a visual case-based reasoning approach*. Artificial Intelligence in Medicine, vol. 94, pages 42–53, 2019. (Cited on page 107.)
- [LeCun 2015] Y LeCun, Y Bengio and G Hinton. *Deep learning*. Nature, vol. 521, pages 436–444, 2015. (Cited on page 2.)
- [Lee 2017] C Lee and H Yoon. *Medical big data: promise and challenges*. Kidney Res Clin Pract., vol. 36(1), pages 3–11, 2017. (Cited on page 2.)
- [Li 2017] H Li and Y Fan. *Non-rigid image registration using fully convolutional networks with deep self-supervision*. arXiv preprint arXiv:1709.00799, 2017. (Cited on page 56.)
- [Litjens 2017] G Litjens, T Kooi, B Bejnordi, A Setio, F Ciompi, M Ghafoorian, J van der Laak, B van Ginneken and C Sánchez. *A survey on deep learning in medical image analysis*. Med Image Anal, vol. 42, pages 60–88, 2017. (Cited on page 2.)
- [Liu 2018] M Liu, J Zhang, D Nie, P Yap and D Shen. *Anatomical landmark based deep feature representation for MR images in brain disease diagnosis*. IEEE Journal of Biomedical and Health Informatics, vol. 22, pages 1476–1485, 2018. (Cited on page 84.)
- [Lorenzi 2013] M Lorenzi, N Ayache, G Frisoni and X Pennec. *LCC-Demons: a robust and accurate symmetric diffeomorphic registration algorithm*. NeuroImage, vol. 81, pages 470–483, 2013. (Cited on page 73.)
- [Lu 2010] X Lu, B Georgescu, M Jolly, J Guehring, A Young, B Cowan, A Littmann and D Comaniciu. *Cardiac anchoring in MRI through context modeling*. Med Image Comput Comput Assist Interv, vol. 13(1), pages 383–390, 2010. (Cited on page 85.)

- [Lu 2018] A Lu, N Parajuli, M Zontak, J Stendahl, K Ta, Z Liu, N Boutagy, G Jeng, I Alkhalil, L Staib, M O'Donnell, A Sinusas and J Duncan. *Learning-based regularization for cardiac strain analysis with ability for domain adaptation*. arXiv preprint arXiv:1807.04807, 2018. (Cited on page 51.)
- [Maas 2013] A Maas, A Hannun and A Ng. *Rectifier nonlinearities improve neural network acoustic models*. Proc. ICML, vol. 30, 2013. (Cited on page 26.)
- [Madabhushi 2017] A Madabhushi and G Lee. *Image analysis and machine learning in digital pathology: Challenges and opportunities*. Med Image Anal, vol. 33, pages 170–175, 2017. (Cited on page 84.)
- [Moldovanu 2015] S Moldovanu, C Obreja and L Moraru. *Threshold selection for classification of MR brain images by clustering method*. AIP Conference Proceedings 2015, vol. 1694, 2015. (Cited on page 84.)
- [Moriya 2018] T Moriya, H Roth, S Nakamura, H Oda, K Nagara, M Oda and K Mori. *Unsupervised segmentation of 3D medical images based on clustering and deep representation learning*. Proceedings of the SPIE, vol. 10578, 2018. (Cited on page 84.)
- [Oktay 2018] O Oktay, E Ferrante, K Kamnitsas, M Heinrich, W Bai, J Caballero, S Cook, A de Marvao, T Dawes, D O'Regan, B Kainz, B Glocker and D Rueckert. *Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation*. IEEE Trans Med Imaging, vol. 37, pages 384–395, 2018. (Cited on pages 2, 19 and 106.)
- [Orlhac 2018] F Orlhac, P Mattei, C Bouveyron and N Ayache. *Class-specific variable selection in high-dimensional discriminant analysis through bayesian sparsity*. preprint HAL 01811514, 2018. (Cited on page 77.)
- [Parajuli 2017] N Parajuli, A Lu, J Stendahl, M Zontak, N Boutagy, I Alkhalil, M Eberle, B Lin, M O'Donnell, A Sinusas and J Duncan. *Flow network based cardiac motion tracking leveraging learned feature matching*. MICCAI, pages 279–286, 2017. (Cited on page 50.)
- [Parisot 2018] S Parisot, S Ktena, E Ferrante, M Lee, R Guerrero, B Glocker and D Rueckert. *Disease prediction using graph convolutional networks: application to autism spectrum disorder and Alzheimer's disease*. Medical Image Analysis, vol. 48, pages 117–130, 2018. (Cited on page 84.)
- [Pedregosa 2011] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay. *Scikit-learn: machine learning in Python*. Journal of Machine Learning Research, vol. 12, pages 2825–2830, 2011. (Cited on pages 67 and 88.)

- [Pennec 1999] X Pennec, P Cachier and N Ayache. *Understanding the “demon’s algorithm”: 3D non-rigid registration by gradient descent*. MICCAI’99, pages 597–605, 1999. (Cited on page 56.)
- [Petersen 2016] S Petersen, P Matthews, J Francis, M Robson, F Zemrak, R Boubertakh, A Young, S Hudson, P Weale, S Garratt, R Collins, S Piechnik and S Neubauer. *UK Biobank’s cardiovascular magnetic resonance protocol*. J Cardiovasc Magn Reson, vol. 18:8, pages 8+, 2016. (Cited on pages 18, 22, 46, 67 and 85.)
- [Petersen 2017] S Petersen, N Aung, M Sanghvi, F Zemrak, K Fung, J Paiva, J Francis, M Khanji, E Lukaschuk, A Lee, V Carapella, Y Kim, P Leeson, S Piechnik and S Neubauer. *Reference ranges for cardiac structure and function using cardiovascular magnetic resonance (CMR) in Caucasians from the UK Biobank population cohort*. J Cardiovasc Magn Reson, vol. 19(1), pages 18+, 2017. (Cited on page 83.)
- [Petitjean 2015] C Petitjean, M Zuluaga, W Bai, J Dacher, D Grosgeorge, J Caudron, S Ruan, I Ayed, M Cardoso, H Chen, D JimenezCarretero, M Ledesma-Carbayo, C Davatzikos, J Doshi, G Erus, O Maier, C Nambakhsh, Y Ou, S Ourselin, C Peng, N Peters, T Peters, M Rajchl, D Rueckert, A Santos, W Shi, C Wang, H Wang and J Yuan. *Right ventricle segmentation from cardiac MRI: A collation study*. Medical Image Analysis, vol. 19(1), pages 187–202, 2015. (Cited on page 22.)
- [Poudel 2016] R Poudel, P Lamata and G Montana. *Recurrent fully convolutional neural networks for multi-slice MRI cardiac segmentation*. arXiv preprint arXiv:1608.03974, 2016. (Cited on pages 18 and 41.)
- [Prakosa 2013] A Prakosa, M Sermesant, H Delingette, S Marchesseau, E Saloux, P Allain, N Villain and N Ayache. *Generation of synthetic but visually realistic time series of cardiac images combining a biophysical model and clinical images*. IEEE Trans. Med. Imaging, vol. 32(1), pages 99–109, 2013. (Cited on page 12.)
- [Qin 2018a] C Qin, W Bai, J Schlemper, S Petersen, S Piechnik, S Neubauer and D Rueckert. *Joint learning of motion estimation and segmentation for cardiac MR image sequences*. MICCAI, pages 472–480, 2018. (Cited on page 51.)
- [Qin 2018b] C Qin, W Bai, J Schlemper, S Petersen, S Piechnik, S Neubauer and D Rueckert. *Joint motion estimation and segmentation from undersampled cardiac MR image*. MLMIR 2018, pages 55–63, 2018. (Cited on page 51.)
- [Queiros 2014] S Queiros, D Barbosa, B Heyde, P Morais, J Vilaca, D Friboulet, O Bernard and J D’hooge. *Fast automatic myocardial segmentation in 4D cine CMR datasets*. Med Image Anal, vol. 18, pages 1115–1131, 2014. (Cited on page 41.)

- [Radau 2009] P Radau, Y Lu, K Connelly, G Paul, A Dick and G Wright. *Evaluation framework for algorithms segmenting short axis cardiac MRI*. The MIDAS Journal - Cardiac MR Left Ventricle Segmentation Challenge <http://hdl.handle.net/10380/3070>, 2009. (Cited on pages 22, 48 and 74.)
- [Ravishankar 2016] H Ravishankar, S Prabhu, V Vaidya and N Singhal. *Hybrid approach for automatic segmentation of fetal abdomen from ultrasound images using deep learning*. IEEE Int Symp Biomedical Imaging, pages 779–782, 2016. (Cited on page 2.)
- [Raza 2018] K Raza and N Singh. *A tour of unsupervised deep learning for medical image analysis*. arXiv preprint arXiv:1812.07715, 2018. (Cited on pages 97 and 107.)
- [Reshed 2011] D Reshed, Y Reshef, H Finucane, S Grossman, G McVean, P Turnbaugh, E Lander, M Mitzenmacher and P Sabeti. *Detecting novel associations in large data sets*. Science, vol. 334, pages 1518–1524, 2011. (Cited on page 87.)
- [Reynolds 2009] D Reynolds. *Gaussian mixture models*. Encyclopedia of Biometrics, pages 659–663, 2009. (Cited on page 87.)
- [Ronneberger 2015] O Ronneberger, P Fischer and T Brox. *U-net: Convolutional networks for biomedical image segmentation*. MICCAI, vol. 9351, pages 234–241, 2015. (Cited on pages 18, 24, 54 and 55.)
- [Rueckert 2016] D Rueckert, B Glocker and B Kainz. *Learning clinically useful information from images: past, present and future*. Medical Image Analysis, vol. 33, pages 13–18, 2016. (Cited on pages 50, 73 and 84.)
- [Rühaak 2013] J Rühaak, S Heldmann, T Kipshagen and B Fischer. *Highly accurate fast lung CT registration*. Medical Imaging 2013: Image Processing, vol. 8669, 2013. (Cited on page 74.)
- [Schulz-Menger 2013] J Schulz-Menger, D Bluemke, J Bremerich, S Flamm, M Fogel, M Friedrich, R Kim, F von Knobelsdorff-Brenkenhoff, C Kramer, D Pennell, S Plein and E Nagel. *Standardized image interpretation and post processing in cardiovascular magnetic resonance: society for cardiovascular magnetic resonance (SCMR) board of trustees task force on standardized post processing*. J Cardiovasc Magn Reson, vol. 15(35), pages 1167–1186, 2013. (Cited on page 23.)
- [Suinesiaputra 2015] A Suinesiaputra, D Bluemke, B Cowan, M Friedrich, C Kramer, R Kwong, S Plein, J Schulz-Menger, J Westenberg, A Young and E Nagel. *Quantification of LV function and mass by cardiovascular magnetic resonance: multi-center variability and consensus contours*. Journal of Cardiovascular Magnetic Resonance, vol. 17(1), 2015. (Cited on page 19.)

- [Suinesiaputra 2016] A Suinesiaputra, A McCulloch, M Nash, B Pontre and A Young. *Cardiac image modelling: breadth and depth in heart disease*. Medical Image Analysis, vol. 33, pages 38–43, 2016. (Cited on pages 74 and 84.)
- [Suinesiaputra 2018] A Suinesiaputra, M Sanghvi, N Aung, J Paiva, F Zemrak, K Fung, E Lukaschuk, A Lee, V Carapella, Y Kim, J Francis, S Piechnik, S Neubauer, A Greiser, M Jolly, C Hayes, A Young and S Petersen. *Fully-automated left ventricular mass and volume MRI analysis in the UK Biobank population cohort: evaluation of initial results*. Int J Cardiovasc Imaging, vol. 34(2), pages 281–291, 2018. (Cited on pages 50 and 96.)
- [Tipping 2003] M Tipping and A Faul. *Fast marginal likelihood maximisation for sparse bayesian models*. AISTATS, 2003. (Cited on pages 76 and 77.)
- [Tobon-Gomez 2013] C Tobon-Gomez, M De Craene, K McLeod, L Tautz, W Shi, A Hennemuth, A Prakosa, H Wang, G Carr-White, S Kapetanakis, A Lutz, V Rasche, T Schaeffter, C Butakoff, O Friman, T Mansi, M Sermesant, X Zhuang, S Ourselin, H Peitgen, X Pennec, R Razavi, D Rueckert, A Frangi and K Rhode. *Benchmarking framework for myocardial tracking and deformation algorithms: an open access database*. Med Image Anal, vol. 17(6), pages 632–648, 2013. (Cited on pages 7 and 12.)
- [Toga 2015] A Toga and K Crawford. *The Alzheimer’s disease neuroimaging initiative informatics core: a decade in review*. Alzheimers Dement., vol. 11, pages 832–839, 2015. (Cited on page 83.)
- [Tran 2016] P Tran. *A fully convolutional neural network for cardiac segmentation in short-axis MRI*. arXiv preprint arXiv:1604.00494, 2016. (Cited on pages 5, 18, 26, 41 and 45.)
- [van der Maaten 2008] L van der Maaten and G Hinton. *Visualizing data using t-sne*. J. Mach. Learn. Research, vol. 9, pages 2579–2605, 2008. (Cited on page 93.)
- [Wang 2013] Y Wang, B Georgescu, T Chen, W Wu, P Wang, X Lu, R Ionasec, Y Zheng and D Comaniciu. *Learning-based detection and tracking in medical imaging: A probabilistic approach*. Deformation Models, vol. 7, pages 209–235, 2013. (Cited on page 7.)
- [Weese 2016] J Weese and C Lorenz. *Four challenges in medical image analysis from an industrial perspective*. Medical Image Analysis, vol. 33, pages 44–49, 2016. (Cited on pages 50 and 84.)
- [Winther 2017] H Winther, C Hundt, B Schmidt, C Czerner, J Bauersachs, F Wacker and J Vogel-Claussen. *ν -net: Deep learning for generalized biventricular cardiac mass and function parameters*. arXiv preprint arXiv:1706.04397, 2017. (Cited on pages 18, 41 and 45.)

- [Wit 2012] E Wit, E van den Heuvel and J Romeijn. *‘All models are wrong ...’: an introduction to model uncertainty*. Statistica Neerlandica, vol. 66, pages 217–236, 2012. (Cited on page 88.)
- [Wolterink 2017] J Wolterink, T Leiner, M Viergever and I Išgum. *Automatic segmentation and disease classification using cardiac cine MR images*. In Proc. Statistical Atlases and Computational Models of the Heart (STACOM), ACDC challenge, MICCAI’17 Workshop, 2017. (Cited on pages 29, 36, 40, 50, 51, 67 and 84.)
- [Xue 2018] W Xue, G Brahm, S Pandey, S Leung and S Li. *Full left ventricle quantification via deep multitask relationships learning*. Med Image Anal, vol. 43, pages 54–65, 2018. (Cited on page 51.)
- [Yan 2018] W Yan, Y Wang, Z Li, R 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*. MICCAI, pages 613–621, 2018. (Cited on page 51.)
- [Yang 2017] D Yang, P Wu, C Tan, K Pohl, L Axel and D Metaxas. *3D motion modeling and reconstruction of left ventricle wall in cardiac MRI*. : International Conference on Functional Imaging and Modeling of the Heart, pages 481–492, 2017. (Cited on page 51.)
- [Zhang 2016] S Zhang and D Metaxas. *Large-scale medical image analytics: Recent methodologies, applications and future directions*. Medical Image Analysis, vol. 33, pages 98–101, 2016. (Cited on page 84.)
- [Zheng 2008] Y Zheng, A Barbu, B Georgescu, M Scheuering and D Comaniciu. *Four-chamber heart modeling and automatic segmentation for 3D cardiac CT volumes using marginal space learning and steerable features*. IEEE Trans Med Imaging, vol. 27, pages 1668–1681, 2008. (Cited on page 6.)
- [Zheng 2018a] Q Zheng, H Delingette and N Ayache. *Explainable cardiac pathology classification on cine MRI with motion characterization by semi-supervised learning of apparent flow*. arXiv preprint arXiv:1811.03433, 2018. (Cited on pages 3, 50, 84, 85, 86, 87, 92, 104 and 105.)
- [Zheng 2018b] Q Zheng, H Delingette, N Duchateau and N Ayache. *3D consistent and robust segmentation of cardiac images by deep learning with spatial propagation*. IEEE Trans Med Imaging, vol. 37(9), pages 2137–2148, 2018. (Cited on pages 3, 17, 54, 55, 58, 67, 76, 85, 87, 104 and 105.)
- [Zheng 2018c] Q Zheng, H Delingette, N Duchateau and N Ayache. *3D consistent biventricular myocardial segmentation using deep learning for mesh generation*. arXiv preprint arXiv:1803.11080, 2018. (Cited on pages 3, 5, 87, 103 and 105.)

- [Zheng 2019] Q Zheng, H Delingette, K Fung, S Petersen and N Ayache. *Unsupervised shape and motion analysis of 3822 cardiac 4D MRIs of UK Biobank*. arXiv preprint arXiv:1902.05811, 2019. (Cited on pages 83 and 105.)
- [Zhou 2017] K Zhou, H Greenspan and D Shen. *Deep learning for medical image analysis*. <https://www.amazon.co.uk/Deep-Learning-Medical-Image-Analysis/dp/0128104082>, 2017. (Cited on pages 2 and 6.)
- [Zilly 2017] J Zilly, J Buhmann and D Mahapatra. *Glaucoma detection using entropy sampling and ensemble learning for automatic optic cup and disc segmentation*. *Comput Med Imaging Graph*, vol. 55, pages 28–41, 2017. (Cited on page 2.)
- [Zuluaga 2013] M Zuluaga, M Cardoso, M Modat and S Ourselin. *Multi-atlas propagation whole heart segmentation from MRI and CTA using a local normalised correlation coefficient criterion*. *Functional Imaging and Modeling of the Heart*, pages 172–180, 2013. (Cited on page 45.)