Deep learning for automated analysis of cardiac imaging: applications in Cine and 4D flow MRI

Sun, X.

Deep Learning for Automated Analysis of Cardiac Imaging: Applications in Cine and 4D Flow MRI

Xiaowu Sun

2023

Colophon

About the cover: A word cloud and one heart appear on the cover. It denotes that in this thesis, approaches and metrics depicted in the word cloud were used to analyze the function of the heart.

Deep Learning for Automated Analysis of Cardiac Imaging: Applications in Cine and 4D Flow MRI
Xiaowu Sun

ISBN: 978-94-6483-185-6
Thesis layout: Xiaowu Sun
Cover design: Xiaowu Sun
Printed by Ridderprint BV

The research in this thesis was performed at the Division of Image Processing (LKEB), Department of Radiology of Leiden University Medical Center, The Netherlands.

This work was carried out in the ASCI graduate school. ASCI dissertation series number: 445

Financial support for the publication of this thesis was kindly provided by:
ASCI research school,
Hart Onderzoek Nederland,
Dutch Heart Foundation,
Library of Leiden University,
Bonitus Stichting

Financial support by the Dutch Heart Foundation for the publication of this thesis is gratefully acknowledged.

Deep Learning for Automated Analysis of Cardiac Imaging: Applications in Cine and 4D Flow MRI

Proefschrift

ter verkrijging van

de graad van doctor aan de Universiteit Leiden,

op gezag van rector magnificus prof.dr.ir. H. Bijl,

volgens besluit van het college voor promoties

te verdedigen op Woensdag 5 Juli 2023

klokke 15:00 uur

door

Xiaowu Sun
geboren te Yantai, Shandong Province, China
in 1992

Promotor:	Prof. dr. ir. B. P. F. Lelieveldt
Co-promotor:	Dr. ir. R. J. van der Geest
Leden promotiecommissie:	Prof. dr. P.H.A. Quax Prof. dr.ir. J.H.C. Reiber Prof. A.J. Nederveen *University of Amsterdam* Prof. dr. I. Isgum *University of Amsterdam*

Contents

Chapter 1 General Introduction

Cardiovascular disease (CVD) is the leading cause of death globally, taking an estimation of 17.9 million lives each year, representing 32% of all deaths worldwide [1]. Echocardiography, computer tomography (CT) and magnetic resonance imaging (MRI) are the prevailing non-invasive imaging techniques for CVD diagnosis in clinical practice. Compared to the other two modalities, due to its excellent image quality and good soft tissue contrast, cardiac magnetic resonance (CMR) established itself as the reference standard for quantification of cardiac dimensions and function, including assessment of left ventricular volume, ejection fraction (EF) and myocardial mass. These clinical parameters can be derived with high precision from cine MRI. Other hemodynamic parameters, including trans-valvular blood flow, peak velocities, kinetic energy and wall shear stress, which also greatly aid in the diagnosis and prognostication of CVD, can be derived from the four-dimensional (4D) flow MRI.

In recent years, deep learning (DL), especially convolution neural network (CNN), has been successfully applied to automatically analyze medical images and derive clinical measures. Therefore, this thesis investigated deep learning techniques and its applications in both cardiac cine and 4D flow MRI.

1.1. Cine cardiac MRI

Cine cardiac MRI is typically obtained by repeatedly imaging the heart at a single slice location at multiple time points throughout one cardiac cycle. To fully image the whole heart, multiple slices at various locations must be obtained. Therefore, cine cardiac MRI provides a complete 3D visualization of the heart supporting detailed analysis of cardiac function. The short-axis (SAX) view and long-axis (LAX) four-chamber (4-CH) and two-chamber (2-CH) views, as shown in Figure.1.1, are routinely obtained anatomical views in cine cardiac MRI [2]. The images in the long-axis view are extracted as imaging planes parallel to a line extending from the cardiac apex to the center of the mitral valve. The SAX sequences are acquired as a stack of multiple 2D slices from the apex to the base of the heart perpendicular to the LAX view. The SAX view provides an excellent cross-sectional view of left ventricle (LV) and right ventricle (RV). Therefore, the images in SAX view are routinely considered as the standard approach to derive volumetric measurements for LV function assessment [3]. The end-diastolic (ED) and end-systolic (ES) phases, i.e. the phases with the largest and smallest LV blood volume, are two crucial phases in the cardiac cycle. ED volume (EDV) and ES volume (ESV) can be used to measure the stroke volume (SV) and ejection fraction (EF) which are important parameters quantifying global cardiac function [4].

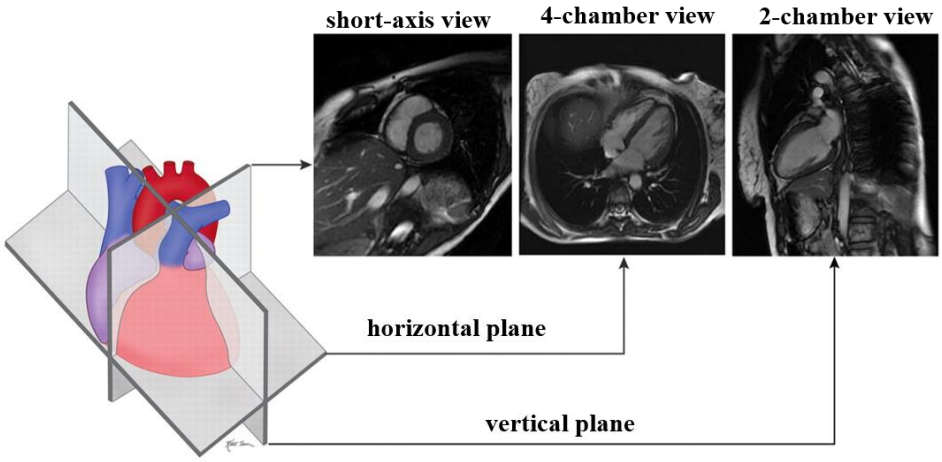


Figure.1.1 Major cardiac imaging planes and their corresponding views in cine cardiac MRI.

There are still some significant barriers to widespread use of cardiac MRI [5,6]. Poor breath-holding related respiratory motion, which is common in patients with heart failure, may introduce artifacts, resulting in low quality images. Additionally, numerous applications of cardiac MRI have been relying on segmentation of the cardiac structures. Manual image segmentation is tedious and time-consuming work and also prone to inter-observer variability. Therefore, in this thesis we address two aspects including the image quality and data analysis in cine MRI.

1.2. 4D flow cardiac MRI

2D cine cardiac MRI allows quantification of volumetric clinical parameters, but it cannot be used to identify hemodynamic markers in the heart or great vessels. 4D flow MRI is a state-of-the-art MR imaging technique encoding time-resolved three-directional velocities, allowing to visualize and quantity the flow direction, peak velocity and flow volume. 4D flow MRI provides sets of 3D volumes over time. Each 4D flow volume contains one magnitude volume and three velocity volumes. 4D flow MRI is particularly used to derive relevant flow parameters, such as flow components, kinetic energy, pulse wave velocity and pressure gradient, to evaluate various cardiovascular conditions and help guide diagnosis treatment and follow-up care for patients.

4D flow MRI shows promising applications in clinical practice, however, it has not been widely used yet. One of the main limitations is that the post-processing takes time and labor. Quantitative analysis relies on segmentation of anatomical regions in the images. But the extremely poor contrast between the heart chamber and surrounding tissues aggravates the difficulty of manual segmentation. As shown

in Figure.1.2, the prevailing segmentation approach in 4D flow MRI depends on the registration between cine MRI and 4D flow data. However, the registration is computationally expensive, resulting in long runtimes. Additionally, differences in heart rate and spatial resolution between those two MRI acquisitions will introduce some misalignment during the registration. Therefore, fully automatic segmentation methods for 4D flow MRI segmentation are needed. The relatively long scan time, ranging from 5 to 20 minutes and limited spatial resolution, are other barriers of 4D flow MRI, which also restrict its analysis.

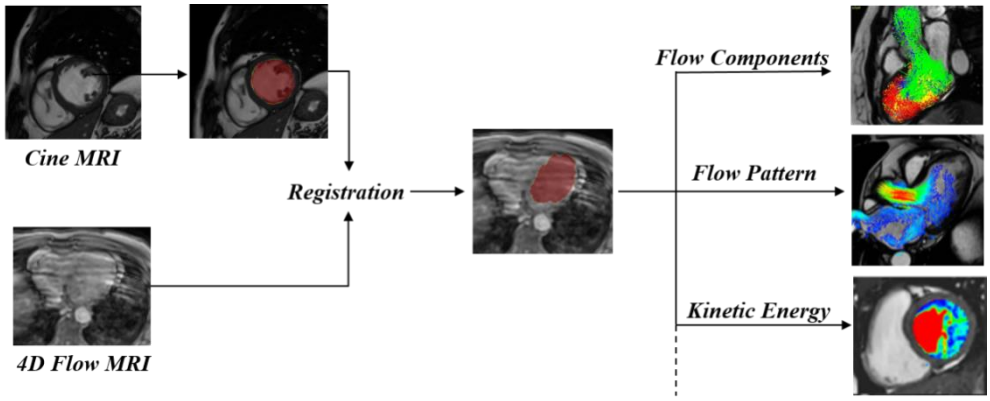


Figure.1.2. Existing workflow of quantitative analysis of 4D flow data. It requires two MRI sequences, firstly generates the mask on cine MRI, then registration is introduced to propagate the segmentation mask from cine MRI to 4D flow MRI.

1.3. Deep learning in cardiac MRI analysis

Based on the availability of the labels in the given data, DL can be divided into unsupervised learning and supervised learning. Unsupervised learning, where the labels are not available, tries to reveal the structure within the data on its own. In supervised learning, the model aims to mimic human performance by learning a mapping from the input data to the annotated labels. Supervised learning is the most commonly used approach in the field of CMR.

CMR has been a crucial technique in the evaluation of cardiac function and disease diagnosis. However, the analysis of CMR is complicated and time-consuming, requiring expert knowledge. Recently with the advance of DL, a variety of DL-based methods have been proposed enabling automated analysis of medical images, including cardiac MRI. For instance, given the manual segmentation, many DL-based frameworks were proposed for automated cardiac MR image segmentation [7-11] enabling quantification of volumetric parameters. As summarized in Figure1.3, the developments of DL relevant to cardiac MRI provide an efficient and effective way in the areas of image acquisition, reconstruction, image quality assessment,

segmentation and diagnosis evaluation. The review [12] provides more details about the applications of DL in cardiac MRI analysis.

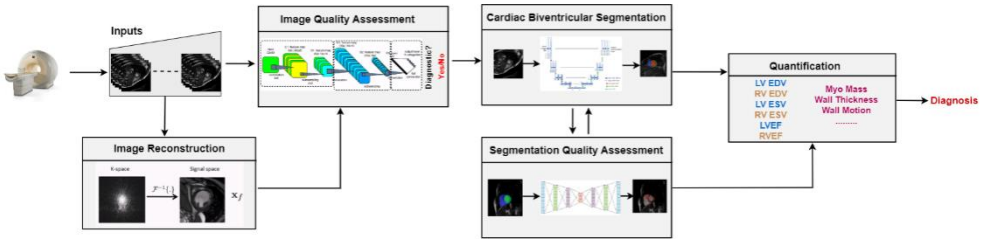


Figure.1.3. Deep learning applications for cardiac function diagnosis including image reconstruction, image quality assessment, segmentation and segmentation quality control [6].

Although deep learning has achieved immense success in the field of medical images, there is still a long way ahead to deploy them to real-world applications. Compared to natural images, medical images are less widely available, especially for cases with rare diseases and expert annotations are expensive. Additionally, model's generalization is another limitation for the deployment in real world, due to the data distribution heterogeneity across multiple modalities, scanners and centers.

1.4. Thesis outline

The work described in this thesis aims to develop deep learning based techniques to achieve fully automatic analysis of cine and 4D flow cardiac MRI. The research topics and connections between each chapter are summarized in the Figure.1.4.

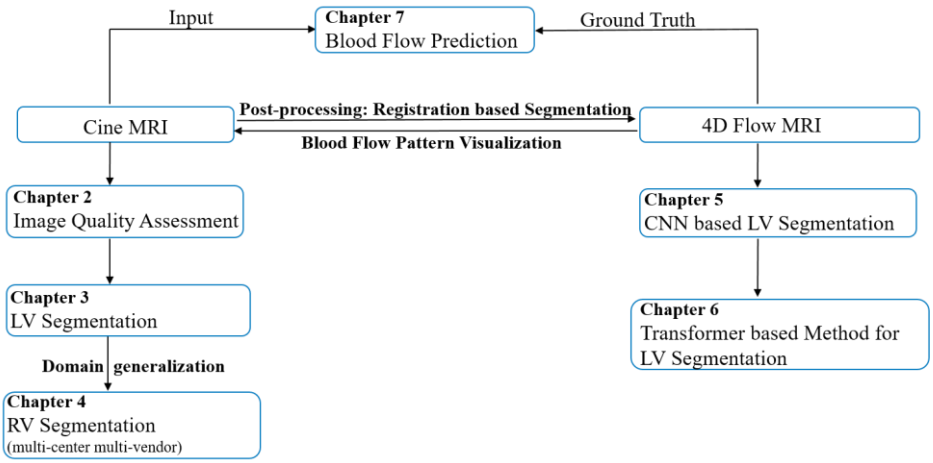


Figure.1.4 Overview of the research topics in the this thesis.

Chapter 2-4 present our work on cine MRI. Motion-related artifacts may result in non-diagnostic image quality. We first propose a method to automatically classify

the image quality. Then, we investigate methods for left ventricular segmentation in short-axis cine images to derive LV volumetric parameters, which can be used for disease classification. A common issue with deep learning based models is that a model trained on one dataset does not generalize well to other unseen datasets due to the distribution heterogeneity between the data sets from various centers or vendors. Therefore, chapter 4 focuses on domain generalization.

Chapter 2 presents a method for cardiac MR image quality assessment combining data augmentation and deep learning network. Given the limited dataset, three specially designed data augmentation techniques are proposed. We also introduce a CNN model to mimic the sample inspection. The method has been evaluated on a public data set and achieved a promising results, ranking the 4th in an international challenge.

Chapter 3 proposes two stack modules to integrate the temporal or spatial information from neighboring slices for left ventricle segmentation in short-axis view. A stack attention module is presented to weigh the features in the channel dimension. The stack attention module can be inserted into the U-Net to improve the segmentation performance. The approach was evaluated on two data sets, one in-house data set and one public data set.

Chapter 4 studies domain generalization in cardiac MR segmentation. **Chapter 3** solves the segmentation task given a specific single-center, single-vendor data set. Instead the use of fine-tuning or adaption to train a new model for a new data set, **Chapter 4** introduces a registration-based method to generate more pseudo data to enlarge the dataset. Additionally, the stack model, introduced in **Chapter 3,** is also applied to explore more features for the segmentation. The trained model is directly validated on an unseen data set.

Chapters 5 and 6 describe methods for LV segmentation in cardiac 4D flow MRI. Due to the poor contrast in 4D flow MRI, conventional segmentation in 4D flow MRI relies on the registration between cine and 4D flow MRI. **Chapter 5** investigates the feasibility of LV segmentation directly from 4D flow data without the use of any additional cine MRI. In contrast to previous studies, this is the first work to fuse the features from two modalities in 4D flow MRI to automate LV segmentation via deep learning. In this chapter we also compared the impact of different network structures and data pre-processing methods on the performance of LV segmentation from 4D flow MRI.

Chapter 6 extends the work of **Chapter 5** into an efficient feature fusion module to aggregate the information from magnitude and velocity images. The proposed module contains a Transformer based cross- and self-fusion layer to explore the inter-relationship between two modalities and the intra-relationship within the same

modality. The clinical parameters derived from the proposed segmentation method are in good agreement with the ground truth.

Chapter 7 aims to build a bridge between cine and 4D flow MRI. 4D flow MRI provides quantitative information on intra-cardiac blood flow, but quantification requires complicated post-processing. A novel deep learning-based approach is presented to predict intra-cardiac blood flow directly from long-axis cine MRI. The intensity fluctuations within the cardiac cavities provide a visual clue about the global blood flow pattern. The model takes temporal neighboring frames as the input and the velocity field derived from 4D flow MRI as the ground truth. The prediction is validated against 4D flow data.

Chapter 8 summarizes the achievements of this thesis and provides a future outlook.

References

1. World Health Organization. Cardiovascular diseases (CVDs) fact sheet. Accessed 19 December 2022. http://www.who.int/mediacentre/factsheets/fs317/en/
2. Ginat, Daniel T., et al. "Cardiac imaging: Part 1, MR pulse sequences, imaging planes, and basic anatomy." American Journal of Roentgenology 197.4 (2011): 808-815.
3. Childs, Helene, et al. "Comparison of long and short axis quantification of left ventricular volume parameters by cardiovascular magnetic resonance, with ex-vivo validation." Journal of Cardiovascular Magnetic Resonance 13.1 (2011): 1-9.
4. van der Geest, Rob J., and Johan HC Reiber. "Quantification in cardiac MRI." Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine 10.5 (1999): 602-608.
5. Ismail, Tevfik F., et al. "Cardiac MR: from theory to practice." Frontiers in cardiovascular medicine 9 (2022).
6. Galati, Francesco, and Maria A. Zuluaga. "Efficient model monitoring for quality control in cardiac image segmentation." International Conference on Functional Imaging and Modeling of the Heart. Springer, Cham, 2021.
7. Tran PV. A fully convolutional neural network for cardiac segmentation in short-axis MRI. arxiv (2016) abs/1604.00494.
8. Baumgartner CF, Koch LM, Pollefeys M, Konukoglu E. An exploration of 2D and 3D deep learning techniques for cardiac MR image segmentation. International Workshop on Statistical Atlases and Computational Models of the Heart, Vo. 10663. Cham: Springer (2017). p. 1–8.
9. Isensee F, Jaeger PF, Full PM, Wolf I, Engelhardt S, Maier-Hein KH. Automatic cardiac disease assessment on cine-MRI via time-series segmentation and domain specific features. Proceedings of the 8th International Workshop, STACOM 2017. Springer International Publishing (2017). p. 120–9.
10. Bai W, Sinclair M, Tarroni G, Oktay O, Rajchl M, Vaillant G, et al. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. J Cardiovasc Magn Reson. (2018) 20:65.
11. Fahmy AS, El-Rewaidy H, Nezafat M, Nakamori S, Nezafat R. Automated analysis of cardiovascular magnetic resonance myocardial native T1 mapping images using fully convolutional neural networks. J Cardiovasc Magn Reson. (2019) 21:1–12.
12. Leiner, Tim, et al. "Machine learning in cardiovascular magnetic resonance: basic concepts and applications." Journal of Cardiovascular Magnetic Resonance 21.1 (2019): 1-14.

Chapter 2 Combination special data augmentation and sampling inspection network for cardiac magnetic resonance imaging quality classification

This chapter was adapted from:

Xiaowu Sun, Li-Hsin Cheng, Rob J. van der Geest. **Combination Special Data Augmentation and Sampling Inspection Network for Cardiac Magnetic Resonance Imaging Quality Classification.** International Workshop on Statistical Atlases and Computational Models of the Heart (STACOM). Springer, Cham, 2022.

Abstract

Cardiac magnetic resonance imaging (MRI) may suffer from motion-related artifacts resulting in non-diagnostic quality images. Therefore, image quality assessment (IQA) is essential for the cardiac MRI analysis. The CMRxMotion challenge aims to develop automatic methods for IQA. In this paper, given the limited amount of training data, we designed three special data augmentation techniques to enlarge the dataset and to balance the class ratio. The generated dataset was used to pre-train the model. We then randomly selected two multi-channel 2D images from one 3D volume to mimic sample inspection and introduced ResNet as the backbone to extract features from those two 2D images. Meanwhile, a channel-based attention module was used to fuse the features for the classification. Our method achieved a mean accuracy of 0.75 and 0.725 in 4-fold cross validation and the held-out validation dataset, respectively. The code can be found here (https://github.com/xsunn/CMRxMotion).

2.1. Introduction

Cardiac magnetic resonance imaging (MRI) is considered as the standard reference for the evaluation of cardiac function due to its excellent image resolution and soft-tissue contrast. However, the MR scanner's hardware itself or the interaction of patient with hardware can result in artifacts in MRI, yielding a low quality imaging, which is often detrimental to the analysis of cardiac function especially in the large-scale imaging studies [1]. Although the artifacts can be minimized by carefully designed image protocols, they still cannot be fully eliminated [2]. Visual inspection of imaging quality is time-consuming and high-cost labor, and also relies on experienced radiologists. Therefore, an automatic method is needed to classify the MR image quality.

In the field of natural images, the approaches to image quality assessment (IQA) can be divided into two categories: full-reference and no-reference, depending on the availability of the original reference image. Meanwhile, recent Convolutional Neural Network (CNN) based methods, such as ResNet [3] and VGG [4], demonstrate promising performance in the automatic image classification task. Bosse et al. [5] employed a Siamese network to extract the features from the distorted and reference patch respectively and fused the difference of those features for IQA. Su et al. [6] proposed a self-adaptive hyper network to blindly assess image quality in the wild without any reference.

However, unlike IQA in natural images, in medical imaging it is particularly challenging for several reasons. There is no large-scale publicly available medical image dataset for IQA. In addition, the distinction between the diagnostic and non-diagnostic imaging is not always evident. Therefore, the labels annotated by radiologists are often subjective [7]. Previously, Fu tried to integrate the information from different color-spaces at feature-level and prediction-level to assess retinal image quality [8]. Oksuz et al. proposed a CNN model to automatically detect and correct motion-related artifacts in cardiac MRI using the K-space lines [9]. Lyu et al. used a recurrent generative adversarial network to reduce motion artifacts in cardiac MRI [10].

The CMRxMotion challenge aims to encourage the participants to develop an IQA model and a segmentation method for the extreme cardiac MRI dataset. In this paper we focuse only on the task of image quality assessment. Our contributions are as follows: (1) We designed specific data augmentation methods to enlarge the given limited data. (2) We proposed a two-branch network and combined a channel-based attention mechanism to fuse features from two random samples of the 3D volume, improving the IQA performance.

2.2. Dataset

The challenge provides short-axis cardiac MR images of 45 healthy volunteers (20 for training, 5 for validation and 20 for testing), obtained through the same 3T MR system (Siemens MAGNETOM Vida) under four different levels of respiratory motion, including full breath-hold, half breath-hold, free breath and intensive breath. Only the images of the end-diastolic (ED) and end-systolic (ES) phase are available. Therefore, there are 160 (20 volunteers × 4 scans × 2 phases), 40 and 160 3D volumes for training, validation and testing. The number of slices in one phase ranges from 9 to 13. The image resolution varies from 0.66×0.66×9.6 mm^3 to 0.76×0.76×10 mm^3, and the range of field of view (FOV) varies from 400×512 mm^2 to 512×512 mm^2. Independent from motion levels, all images were reviewed and scored by multiple radiologists using a standard 5-point Likert scale which can be found in https://www.synapse.org/#!Synapse:syn32407769/wiki/618241. For better reproducibility, the organizer divided those images into three classes based on the 5-point scores: mild motion, intermediate motion, and severe motion.

During data preprocessing, we excluded slices outside of the heart region and selected the 9 slices in the center to make each processed case having the same number of slices. Afterwards, all the cases were cropped or zero-padded into a uniform matrix size of 192×192×9 and the image intensity was normalized to [0,1] using the min-max method.

2.3. Methods

2.3.1. Data augmentation

In this section, we describe the specially designed data augmentation method for IAQ in detail. The first two strategies are based on weighted interpolation of images from the same subject, while the third strategy employs histogram matching plus interpolation to generate new images. All of the data augmentation methods are based on the 3D volume.

Generating transition phases between ED and ES. The ED and ES phases capture the two extreme scenarios in a cardiac cycle. The transition phases between ED and ES in the same cardiac cycle have almost identical intensity distribution [12]. Therefore, given the available ED and ES phases, we first generate new transition phases between ED and ES using weighted interpolation defined as following:

$$wp = wI_1 + (1-w)I_2 \ , \ n\_label \approx wL_1 + (1-w)L_2 \tag{2.1}$$

where *wp* is the generated volume and *n_label* is its corresponding label, I_1, I_2, and L_1, L_2 are the 3D volume and labels of ED and ES phases, and w is the weight. In

this work, we used three values for w, namely 0.2, 0.5 and 0.8, to generate transition phases. Figure.2.1 shows an example of the generated images using this approach.

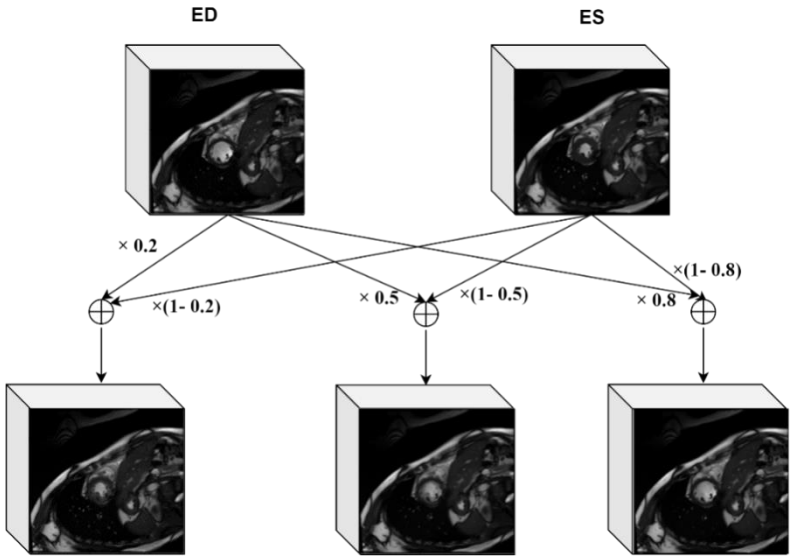


Figure.2.1. An example of generated new phases by weighting ED and ES phases from the same case. The first row implies the 3D volume selected from ED and ES phase. The second row presents 3D volume selected from the generated phases using the weights of 0.2, 0.5 and 0.8, respectively.

Generating intermediate images from different levels of respiratory motion. Interestingly, each volunteer was scanned four times under different levels of respiratory motion. Those paired cases from the same volunteer have the same anatomy structures but with different image qualities. Therefore, we used the paired images at the same phase but from different respiratory motion to generate new images. As illustrated in Figure.2.2, two paired images (P004-1-ED, P004-4-ED for example) both from the ED phase of the same volunteer, but possibly with different image qualities, are selected randomly as the source images. After an intensity-based registration, the method described in formula (2.1) is used to generate the new image and its corresponding label. Similar as the previous augmentation strategy, the new images are generated using weights of 0.2, 0.5 and 0.8.

Generating degraded images with histogram matching and linear interpolation. Within the 160 training cases, the numbers of cases with mild, intermediate, and severe motion artifacts are 70, 69, 21, respectively. To enlarge the subset with severe motion artifacts, the cases with intermediate artifacts were degraded into a lower-quality ones using the linear interpolation approach. As shown in Figure.2.3 a 3D volume with severe artifacts is randomly selected as the reference and another one with intermediate artifacts is considered as the source image. The pixel intensity

distribution of the source image is matched to that of the reference image. We then randomly choose 5% of the pixels from the matched image, and apply the linear interpolation approach on those selected pixels to expand into a new image. The label of the generated image is assigned as severe.

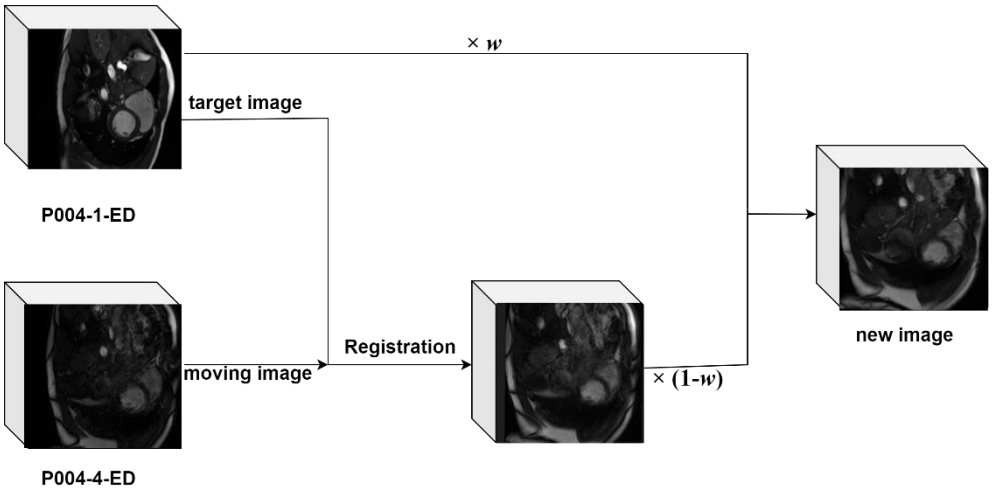


Figure.2.2. An illustration of using two ED phases under different respiratory motion levels of the same volunteer to generate a new image.

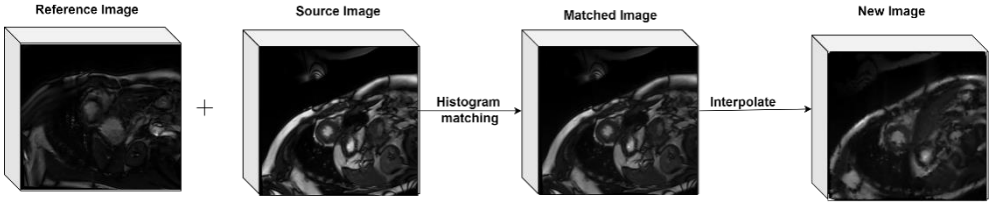


Figure.2.3. The procedure of image degradation includes histogram matching and linear interpolation. The generated result is considered as a new image with lower quality.

2.3.2. Sampling Inspection Network Architecture

To mimic the sampling inspection, the quality of a 3D volume is determined by estimating the quality of random samples drawn from the volume. The advantage of the random selection strategy is that it can generate more data from a single volume to train a model. However, because the selected sample occasionally missed certain critical slices, we introduced another sample with different combinations of 2D slices, mimicking ensembling two times of sampling inspection. The model architecture is shown in Figure.2.4.

The two samples were regarded as 2D multi-channel images, and were each processed by a convolution layer. In addition, according to our intuition, the slices from the apical, middle and basal regions contribute differently for IQA. The slices

in the middle section, with a relatively larger size of the left ventricle than those in apex and base, have a significant impact on the quality assessment. Therefore, the channel attention module (CAM) proposed in [11] was introduced to explore the intra-channel relationship of the input. After that, ResNet was introduced as a backbone to extract the features for each input. The features from those two branches were concatenated along the channel dimension, and a Feature Fusion Module (FFM) was introduced to fuse those features. The FFM block contains one channel attention module to explore the inter-channel relationship and one averaged pooling layer to extract the global information. Lastly, a fully connected layer was used to predict the result.

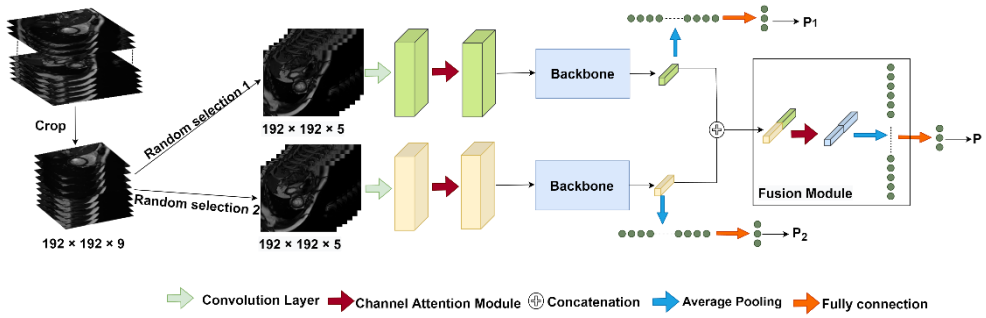


Figure.2.4. Sampling inspection network architecture. The input of each branch is a multi-channel 2D image.

Inspired by the idea of deep supervision [13], besides the final prediction P, each branch also has one prediction denoted as P_1 and P_2. Therefore, the total loss function can be expressed as:

$$Loss = CE(P, L) + \sum_{i=1}^{2} CE(P_i, L) \tag{2.2}$$

where CE is cross-entropy loss. Only the prediction P was used for the validation and testing. During the validation and testing, the sampling inspection was repeated 50 times for one 3D volume, the averaged result was regarded as the final prediction.

2.4. Experiments and results

The training data was divided into 4-fold for cross validation. The metrics, including accuracy, precision, recall, F1-Score and Cohen's Kappa were used to evaluate the performance. All the results were reported as the mean value of four folds. All the experiments were implemented with Pytorch trained on a machine with a NVIDIA Quadro RTX 6000 GPU with 24 GB memory. Adam was employed as the optimizer with 0.00001 as the learning rate.

The ResNet was employed as the baseline, and it took a multi-channel 2D image with a size of 192×192×9 as the input. We first compared our method with the baseline. Due to the small and imbalanced dataset, the baseline failed to predict the severe class, yielding a relatively poor result with accuracy being 0.41, as reported in Table.2.1. The performance indicated that a larger and balanced dataset is needed.

Table.2.1. The 4-fold cross validation performance. Over-Acc: the overall accuracy based on all classes. DA: data augmentation. P: Precision. R: Recall. F: F1-Score.

Model	DA	Over-Acc	Cohen's Kappa	Severity Level	Acc	P	R	F
ResNet	No	0.41	-0.04	Mild	0.76	0.42	0.76	0.54
				Intermediate	0.19	0.38	0.19	0.25
				Severe	0.00	0.00	0.00	0.00
Ours	Yes	**0.75**	**0.58**	Mild	0.77	**0.83**	0.77	**0.80**
				Intermediate	0.81	0.69	0.81	0.75
				Severe	0.48	0.71	0.48	0.57
	No	0.68	0.45	Mild	**0.86**	0.74	**0.86**	0.79
				Intermediate	0.68	0.64	0.68	0.66
				Severe	0.10	0.40	0.10	0.15

We also evaluated the performance of the proposed network and data augmentation (DA) techniques. The new data generated from the offline data augmentation approach was used to pre-train the model and the pre-trained model was fine-tuned using the original training data. Table.2.1 also reports the classification results derived from the proposed methods. It shows that the overall accuracy increased from 0.68 to 0.75 after using DA. Although on the class of mild, the accuracy using DA is a little lower than that without DA, the accuracy for the other two classes is better. The method using DA achieved the best performance on the metric of F1-Score in all classes. The confusion matrix in Figure.2.5 further reveals that the number of false negatives for the severe class reduced after introducing DA. Therefore, the performance confirmed that the carefully designed DA works well for the IQA task.

For the validation part, the labels were hidden by the organizer, we submitted our predicted results and evaluated the performance online. Our method achieved a competitive results, yielding accuracy of 0.725 and Cohen's Kappa 0.645. The best model in the validation data was chosen as the final model, and we submitted it to the organizer and evaluated the performance in the testing dataset with 120 image volumes [14], achieving accuracy of 0.6417 and Cohen's Kappa 0.456.

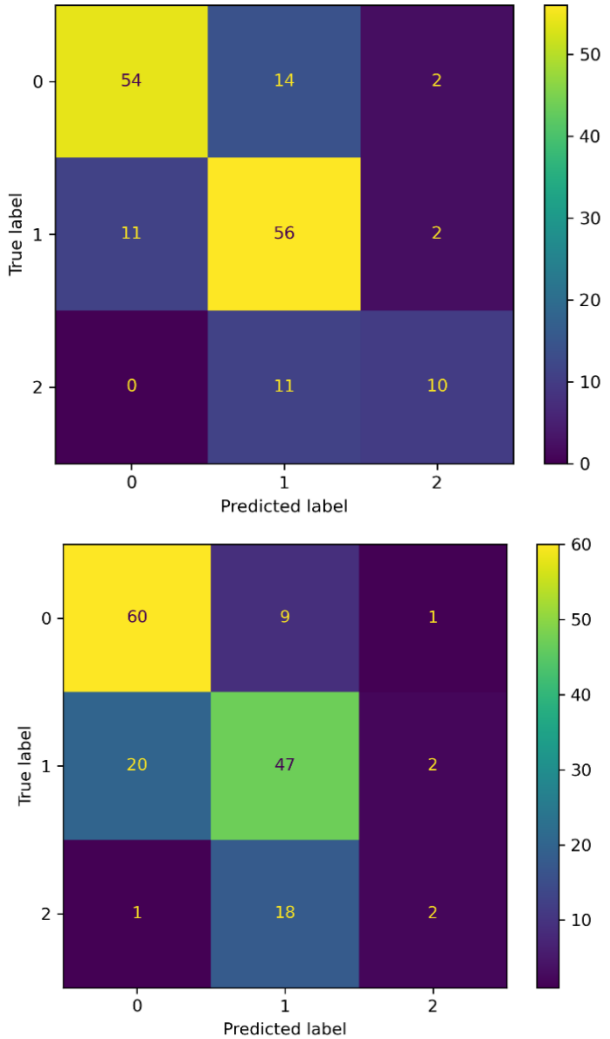


Figure.2.5. Confusion matrix derived from the proposed network. 0, 1, 2 represent the classes of mild, intermediate and severe. The upper one is the result using data augmentation and the bottom one is the result without data augmentation.

Ablation. In the proposed classification network, a module named FFM was used to fuse the features from two branches. To reveal the effectiveness of FFM, we evaluated the accuracy and confusion matrix derived from the three predictions P, P_1, P_2 as reported in Table.2.2, Figure.2.5 and Figure.2.6. P_1, P_2 were derived from two individual braches, while P was generated using the FFM. Compared with the other two predictions, P achieved the best performance on all those classes and the overall.

Table.2.2. Comparison of the accuracy for each class derived from different branches.

	P	P_1	P_2
Mild	**0.77**	0.71	0.73
Intermediate	**0.81**	0.79	0.80
Severe	**0.48**	0.29	0.29
Overall	**0.75**	0.69	0.71

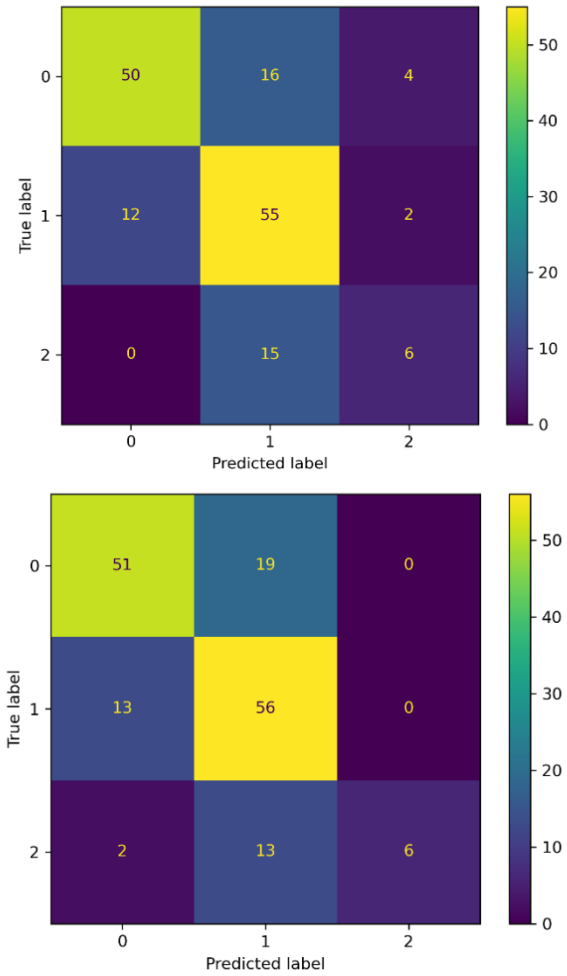


Figure.2.6. Confusion Matrix of two predictions P_1, P_2. The upper one is from the result P_1, the bottom one is derived from P_2.

2.5. Conclusion

In this paper, we designed three data augmentation methods to enlarge the dataset and balance the classes for the cardiac MR image quality assessment task. Inspired by the idea of sample inspection, to enlarge the training data and to extract sufficient features, we randomly selected different combinations of 2D slices as the input of

each branch of the network. The proposed method was trained and evaluated using four-fold cross validation. The results of the classification accuracy, precision, recall and F1-Score demonstrate that our method performed better than the baseline, and the results on the validation dataset shows a competitive performance against the other participants' methods.

Declaration. The authors of this paper declare that they did not use any additional medical image datasets other than those provided by the organizers. They also would like to acknowledge the organizer of the CMRxMotion challenge for collecting and sharing the dataset.

References

1. Krupa, Katarzyna, and Monika Bekiesińska-Figatowska. "Artifacts in magnetic resonance imaging." Polish journal of radiology 80 (2015): 93.
2. Zhang, Le, et al. "Automated quality assessment of cardiac MR images using convolutional neural networks." International Workshop on Simulation and Synthesis in Medical Imaging. Springer, Cham, 2016.
3. He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
4. Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).
5. Bosse, Sebastian, et al. "Deep neural networks for no-reference and full-reference image quality assessment." IEEE Transactions on image processing 27.1 (2017): 206-219.
6. Su, Shaolin, et al. "Blindly assess image quality in the wild guided by a self-adaptive hyper network." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
7. Ma, Jeffrey J., et al. "Diagnostic image quality assessment and classification in medical imaging: Opportunities and challenges." 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). IEEE, 2020.
8. Fu, Huazhu, et al. "Evaluation of retinal image quality assessment networks in different color-spaces." International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2019.
9. Oksuz, Ilkay, et al. "Detection and correction of cardiac MRI motion artefacts during reconstruction from k-space." International conference on medical image computing and computer-assisted intervention. Springer, Cham, 2019.
10. Lyu, Qing, et al. "Cine cardiac MRI motion artifact reduction using a recurrent neural network." IEEE Transactions on Medical Imaging 40.8 (2021): 2170-2181.
11. Wang, Qilong, et al. " ECA-Net: Efficient channel attention for deep convolutional neural networks." Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Seattle, WA, USA. 2020.
12. Zhang, Yao, et al. "Semi-supervised cardiac image segmentation via label propagation and style transfer." International Workshop on Statistical Atlases and Computational Models of the Heart. Springer, Cham, 2020.
13. Wang L, Lee CY, Tu Z, Lazebnik S. Training deeper convolutional networks with deep supervision. arXiv preprint arXiv:1505.02496. 2015.
14. Wang S, Qin C, Wang C, Wang K, Wang H, Chen C, et al. “The Extreme Cardiac MRI Analysis Challenge under Respiratory Motion (CMRxMotion)”. arXiv preprint arXIv: 2210.06385 (2022)

Chapter 3 SAUN: Stack attention U-Net for left ventricle segmentation from cardiac cine magnetic resonance imaging

This chapter was adapted from:

Xiaowu Sun, Pankaj Garg, Sven Plein, Rob J. van der Geest. **SAUN: Stack attention U-Net for left ventricle segmentation from cardiac cine magnetic resonance imaging.** Medical Physics, 48(4), 1750-1763

Abstract

Purpose: Quantification of left ventricular (LV) volume, ejection fraction and myocardial mass from multi-slice multi-phase cine MRI requires accurate segmentation of the LV in many images. We propose a stack attention-based convolutional neural network (CNN) approach for fully automatic segmentation from short-axis cine MR images.

Methods: To extract the relevant spatiotemporal image features, we introduce two kinds of stack methods, spatial stack model and temporal stack model, combining the target image with its neighboring images as the input of a CNN. A stack attention mechanism is proposed to weigh neighboring image slices in order to extract the relevant features using the target image as a guide. Based on stack attention and standard U-Net, a novel Stack Attention U-Net (SAUN) is proposed and trained to perform the semantic segmentation task. A loss function combining cross-entropy and Dice is used to train SAUN. The performance of the proposed method was evaluated on an internal and a public dataset using technical metrics including Dice, Hausdorff distance (HD) and mean contour distance (MCD), as well as clinical parameters, including left ventricular ejection fraction (LVEF) and myocardial mass (LVM). In addition, the results of SAUN were compared to previously presented CNN methods, including U-Net and SegNet.

Results: The spatial stack attention model resulted in better segmentation results than the temporal stack model. On the internal dataset comprising of 167 post-myocardial infarction patients and 57 healthy volunteers, our method achieved a mean Dice of 0.91, HD of 3.37 mm and MCD of 1.08 mm. Evaluation on the publicly available ACDC dataset demonstrated good generalization performance, yielding a Dice of 0.92, HD of 9.4 mm and MCD of 0.74 mm on end-diastolic images, and a Dice of 0.89, HD of 7.1 mm and MCD of 1.03 mm on end-systolic images. The Pearson correlation coefficient of LVEF and LVM between automatically and manually derived results were higher than 0.98 in both datasets.

Conclusion: We developed a CNN with a stack attention mechanism to automatically segment the LV chamber and myocardium from the multi-slice short-axis cine MRI. The experimental results demonstrate that the proposed approach exceeds existing state-of-the-art segmentation methods and verify its potential clinical applicability.

3.1 Introduction

Due to the excellent image resolution and soft-tissue contrast, cardiac cine magnetic resonance imaging (MRI) is considered the reference standard for quantitative assessment of cardiac size and function [1,2]. Typically, imaging is performed in short-axis orientation, and multiple slices and multiple phases are acquired to image the complete left ventricle (LV) over the cardiac cycle. Quantification requires segmentation of many images. Traditional manual segmentation is labor-intensive and relies on experienced experts. In recent years, the convolution neural network (CNN) based approaches have achieved immense success in LV segmentation, and many fully automatic segmentation algorithms based on CNN have been proposed. U-Net [3] and fully convolution network (FCN) [4] are the typical CNN models used in medical image analysis due to their capability of multi-scale feature extraction and fusion. Bai et al. [5] used a training set of 4875 subjects (93500 annotated image slices) to build a basic FCN for segmentation of the LV in short-axis MRI and used a fine-tuning approach to enable segmentation in other datasets. This approach required a massive set of images and also labor-intensive manual annotation effort. Isensee et al. [6] integrated the segmentation and classification task into an ensemble U-Net in which geometrical features extracted from the segmentation results were used for pathology classification. Recently, several unsupervised and self-learning strategies have been proposed, most of these methods use multiple branches to explore additional information and then add these branches to the segmentation backbone [7]. Qin et al. [8] proposed a joint model with two branches: one branch introduced an unsupervised Siamese style spatial transformer network to extract motion features, and the other branch was based on the fully convolutional network for segmentation.

A limitation of previous work is that most of the proposed deep learning methods extract image features from a single 2D image only, which implies that potentially relevant spatiotemporal information that can be derived from neighboring slices and phases is not being exploited [9]. In recent literature, the classical optical flow (OF) method [10, 11, 12] has been introduced to extract temporal coherence among neighboring phases. For example, Zhao et al. [11] coupled the OF from the specified resolution scale to explore the motion features. Yan et al. [12] computed the OF features between two neighboring phases and integrated those features into a U-Net. However, the OF adopts an iterative method, which is time-consuming. Recently some other deep learning methods have been proposed to detect motion features. Zhang et al. [13] applied an LSTM model to incorporate local motion information by regarding several neighboring frames as input. Desai et al. [14] constructed a multi-channel architecture by stacking several neighboring frames to detect the spatiotemporal features. However, which architecture and input depth are optimal for LV segmentation performance in cine MRI is not fully explored. Therefore, we

proposed two image stack models to build a multi-channel architecture. One method is called the spatial stack model, combining the target image which is introduced for the segmentation and its neighboring slices from the same cardiac phase. The other method is called a temporal stack, containing the target image and its neighboring phases at the same slice level. Then a stack attention model is proposed to obtain weighted potential cardiac information from the stack. Traditional local image feature extraction, visual saliency detection, and sliding window methods can all be considered as an attention mechanism. However, in a CNN, the attention module is usually an additional brief neural network that can recognize the important parts from the images or assign different weights to different parts of the input. With the development of deep learning, building a neural network with an attention mechanism has been an active topic of research in computer vision [15, 16, 17]. Because a neural network can learn the attention mechanism autonomously, the inclusion of an attention mechanism can help the network to understand the image better. Due to its excellent performance, attention mechanism is currently widely used in many fields such as machine translation, speech recognition, image caption and computer vision.

To improve the accuracy of LV segmentation, our work mainly focuses on the following aspects:

(1) We introduce two stack models (spatial stack and temporal stack) as a quasi-volumetric architecture to extend the depth of the input.

(2) We propose a stack attention mechanism in which the target image serves as a guide to weigh the features from multiple channels and select the spatiotemporal information.

(3) A novel Stack Attention U-Net (SAUN) based on the stack attention and basic U-Net is proposed for automatic LV segmentation.

3.2 Methods

Different from natural images, MR images only have a single channel (grayscale) and have more complex texture features. Meanwhile, the shape, size and position of the LV only varies slightly between neighboring slices both in the spatial and temporal domain. To address those deformations and contextual information, we will first illustrate how to construct a volumetric architecture using the spatial stack model and temporal stack model respectively, and then integrate the features from the stack model with a novel stack attention mechanism. Finally, we propose the SAUN model based on stack attention and basic U-Net for segmentation.

3.2.1 **Stack model**

Figure.3.1 illustrates the construction of a stack in a case having 30 cardiac phases and 12 slices. Spatial stack $SSM=\{S_{15,4},\cdots,S_{15,7},\cdots S_{15,10}\}$ and temporal stack $TSM=\{S_{12,7},\cdots,S_{15,7},\cdots S_{18,7}\}$ can be used to generate an example image stack of dimension $N=7$ as the input which produces the segmentation result for the central slice $S_{15,7}$.

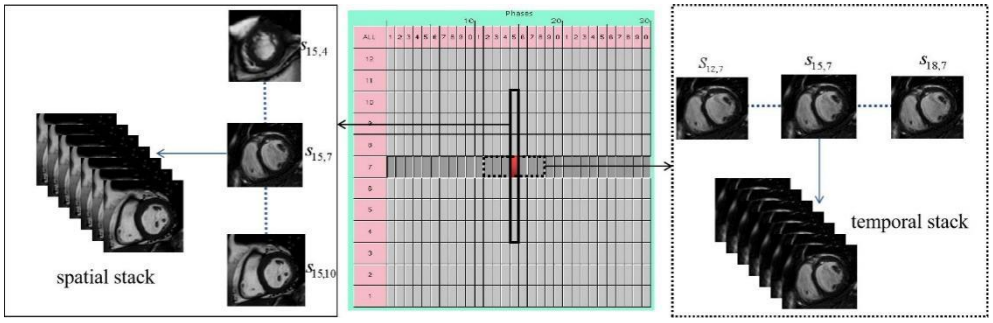


Figure.3.1. Example of the construction of a spatial and temporal stack of dimension 7. Slice $S_{15,7}$ is the target slice; spatial stack model uses slices $\{S_{15,4},S_{15,5},L,S_{15,9},S_{15,10}\}$ from the same phase to build the stack model, while temporal stack model introduces slices $\{S_{12,7},S_{13,7},LS_{17,7},S_{18,7}\}$ from the same slice level to construct another kind of stack model. $S_{i,j}$ is the image from the *i*th phase *j*th slice.

Spatial stack model (SSM) We propose a novel method named spatial stack model to combine the target image with its neighboring spatial slices. The stack model for the central slice $S_{p,t}$ can be described as the following, where $S_{i,j}(i=1,2,\cdots,T;\ j=1,2,\cdots,F)$ represents the image from the *i*th phase *j*th slice, *T* and *F* is the number of phases and slices in the data set respectively, and *N* is the number of the images in the stack.

$$SSM(S_{p,t},N)=\{S_{i,j}\mid i=p,\ j=t-(N-1)/2,\cdots,t+(N-1)/2\}$$
$$and\quad \begin{cases} if & j<1, \quad j=1 \\ if & j>F, \quad j=F \end{cases} \tag{3.1}$$

Temporal stack model (TSM) Similar to the spatial stack, the temporal stack model can be defined as follows

$$TSM(S_{p,t},N)=\{S_{i,j}\mid i=p-(N-1)/2,\cdots,p+(N-1)/2,\ j=t\}$$
$$and\quad \begin{cases} if & i<1, \quad i=i+T \\ if & i>T, \quad i=i-T \end{cases} \tag{3.2}$$

The original MR image is a grayscale image with one channel. The image represented by the stack model can be regarded as a multi-channel image with abundant semantic features. It is important to note that, to our intuition, features derived from images closer (in space or time) to the target image contribute more in segmenting the object in the target slice. Hence, in order to filter out the background noise and extract relevant image information, we further propose the stack attention model.

3.2.2 Stack attention model

In this part, we introduce the target image as a guide to provide the channel information to fuse the neighboring images into the stack.

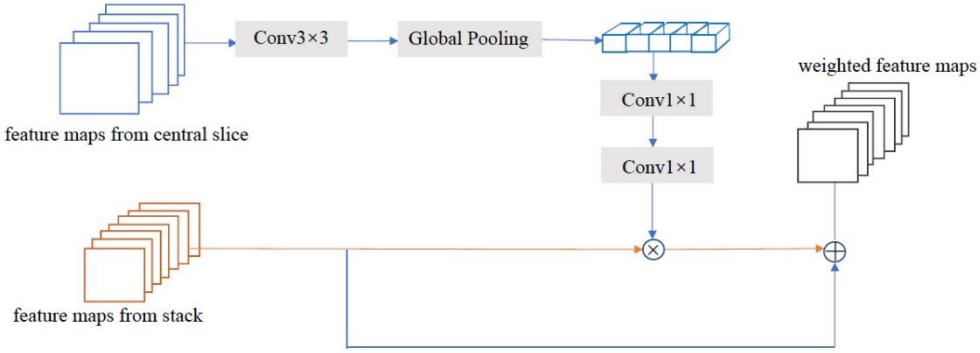


Figure.3.2. Stack attention module structure

In detail, as shown in Figure.3.2, we first perform a 3×3 convolution with ReLu non-linearity function on the feature maps from the central slice to ensure the number of the feature maps generated from central slice and stack is the same. Then the global spatial information is extracted and squeezed to a vector $P=(p_1, p_2, \cdots, p_C)$ through the global average pooling, which can be described as the following equation where $W \times L$ is the size of the feature map, f_c is the feature map of the cth channel and C is the number of channels which is equal to the number of kernels in the convolution layer.

$$p_c = \frac{1}{W \times L} \sum_{i=1}^{W} \sum_{j=1}^{L} f_c(i, j) \quad (c = 1,2,\cdots,C) \tag{3.3}$$

Two different 1×1 convolutions K_1 and K_2 are applied to further compute the weights of each channel as follows:

$$s' = \sigma(\sigma(P * K_1) * K_2) \cdot s \tag{3.4}$$

where $*$ is the convolution operation, σ is ReLu activation function and s is the feature map generated from the stack model. The first convolution K_1 reduces the dimension of vector P from C to $C/2$, and then convolution K_2 resizes the length of vector P into C again. However, the dot production with the weights which range from 0 to 1 repeatedly will degrade the feature values in deep layers, which may lead to negative results. To avoid this problem, finally the weighted stack feature maps are added with the original stack feature maps, which means

$$attS_c(i,j) = (1+P_c) f_c(i,j) \qquad (c = 1,2,\cdots C) \tag{3.5}$$

where $attS_c$ is the *cth* channel of the attention stack. When P_c approaches to 0, $attS_c(i,j)$ will approximate to the original features.

3.2.3 SAUN Network Architecture

Based on the mentioned stack attention and traditional U-Net, we propose the SAUN for the segmentation task. As shown in Figure.3.3, there are two inputs in SAUN, one is the central slice which is the target, and the other one is called the initial stack (either spatial or temporal stack) which is constructed according to $Stack(S,N)$ proposed above. To ensure that the central slice and the stack are at the same feature level, the convolution operation is applied to both of them at the same time.

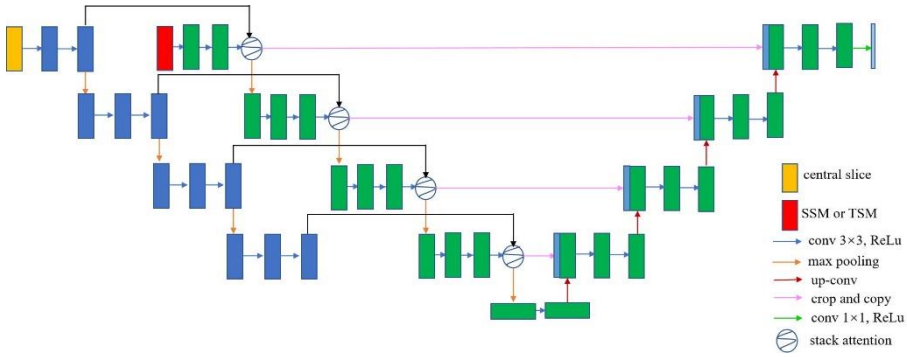


Figure.3.3 Segmentation model structure based on Stack Attention and U-Net (SAUN).

During training SAUN, we aim to optimize the following loss function, which contains the generalized Dice loss and cross-entropy loss. The loss function can be formulated as

$$loss = 1 - 2\frac{\sum_{i=1}^{l} w_i \sum_{j=1}^{n} g_{ij} p_{ij}}{\sum_{i=1}^{l} w_i \sum_{j=1}^{n} g_{ij} + p_{ij}} - \sum_{i=1}^{l}\sum_{j=1}^{n} g_{ij} \log(p_{ij}) \tag{3.6}$$

where the second term is the weighted Dice loss for multiple cardiac structure segmentation, and the third term is cross-entropy loss based on pixel-wise classification. Parameters g, p stand for ground truth and prediction results respectively, l denotes three labels (background, chamber and myocardium), n is the number of the pixels and w_i is the weight of each label, which were set to $w = [0.1, 0.2, 0.7]$.

3.3 Dataset and data preprocessing

3.3.1 Dataset

Leeds University Dataset (LUD). One of the datasets in this work is from the University of Leeds, UK. This dataset contains 168 post-myocardial infarction patients and 57 healthy volunteers. All subjects were scanned on a Philips Ingenia 1.5T MRI system using a slice thickness of 5.0 mm (or sometimes 8.0 mm) and slice gap of 2 mm. The number of slices ranged from 10 to 20, and 30 phases were reconstructed to cover a complete cardiac cycle. The in-plane image resolution varied from 0.78×0.78 mm^2 to 1.18×1.18 mm^2 and the range of field of view (FOV) varied from 280×280 mm^2 to 470×470 mm^2. Expert annotations were derived semi-automatically in all cardiac phases and slices by one observer (RG) with 20 years of experience in cardiac MRI using Mass software (Version V2017-EXP; Leiden University Medical Center, Leiden, the Netherlands), resulting in 6703 annotated images. The subjects' exams were randomly split into three parts with 141, 15 and 69 for training, validation and testing, respectively.

MICCAI 2017 Automated cardiac diagnosis challenge (ACDC 2017). The MICCAI 2017 Automated Cardiac Diagnosis Challenge (ACDC 2017) was organized by the University Hospital of Dijon and the data used in this challenge has become publicly available [18]. The dataset contains short-axis cine MRI exams of 100 subjects of five patient categories (post-myocardial infarction, dilated cardiomyopathy, hypertrophic cardiomyopathy, abnormal right ventricle and healthy subjects). The subjects were scanned on two different scanners (1.5T Siemens Area and 3.0T Siemens Trio Tim) using a typical slice thickness of 5.0 mm (range 5 - 8 mm), an inter-slice gap of 5 mm (range 5 - 10 mm) and pixel spacing ranging from 1.37 to 1.68 mm. For all exams, the manual ground truth annotation was generated by a single clinical expert including contours of the LV cavity and myocardium and the right ventricular cavity in the end-diastolic (ED) and end-systolic (ES) images. In this work, the annotation of the right ventricular cavity was ignored and considered as background in the ground truth. The 100 subjects were randomly divided into five folds, each fold containing five patient categories and each category containing four subjects. We randomly selected three folds to train the network, and the other two folds were chosen for validation and testing, respectively.

3.3.2 Data preprocessing and augmentation

Within the available dataset, the images vary in intensity range, FOV and pixel spacing. The field of view in the LUD data varies from 280 mm to 470 mm, while the heart as the object of interest typically measures 60 mm, occupying only a small proportion of the whole image. For example, in our LUD dataset, the average proportion occupied by the object relative to the full image is around 2.2%. Hence, several image preprocessing methods were performed to standardize those parameters.

We firstly resample the original images into a common pixel spacing of 1.5 mm, and then the image intensities were normalized according to the following formula where $P_{\min}$ and $P_{\max}$ is the minimum and maximum value of 5% and 95% percentile of image P.

$$p = \frac{P_i - P_{\min}}{P_{\max} - P_{\min}} \tag{3.7}$$

To solve the label imbalance problem, the YOLO model [19] is applied to localize the region-of-interest (ROI). As illustrated in Figure.3.4 each 2D original image is considered as an input, and then YOLO extracts the features from the input to generate the bounding boxes. Lastly, the images are cropped or zero-padded to a uniform matrix size of 128×128, centered at each bounding box. Additionally, in order to train a well generalizing network with limited data, data augmentation was employed, including horizontal and vertical clip, image transpose and elastic deformation.

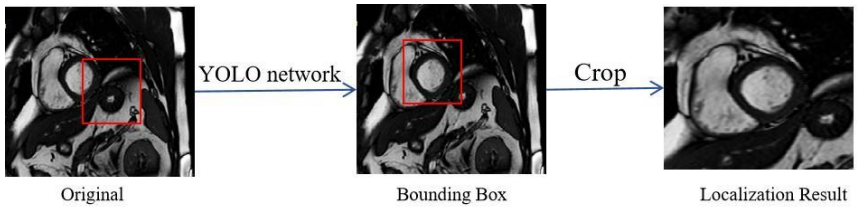


Figure.3.4. An example of localization preprocess. In the image, at the left, the red box is at the center of the image initially, but it didn't detect the heart accurately, but after applying the YOLO model, the position of the object can be extracted precisely. Lastly, it is cropped into a fixed size, centred at the red bounding box.

3.4 Evaluation metrics

For quantitative assessment, two aspects, including segmentation and clinical parameter estimation, are proposed to compare the performance among different segmentation methods. All metrics are evaluated on a per-patient basis.

3.4.1. Segmentation accuracy assessment metrics

Dice is introduced to evaluate the overlap between the automatic and manual segmentation mask. In addition, the distance metrics, including Mean Contour Distance (MCD) and Hausdorff Distance (HD) are employed as the segmentation metrics.

MCD and HD are defined as:

$$MCD = \frac{1}{2|C_A|}\sum_{p\in C_A} d(p,C_B) + \frac{1}{2|C_B|}\sum_{q\in C_B} d(q,C_A) \tag{3.8}$$

$$HD = \max(\max_{p\in C_A} d(p,C_B), \max_{q\in C_B} d(q,C_A)) \tag{3.9}$$

where C_A and C_B are the automatic contour and manual contour respectively, $d(p,C) = \min_{q\in C} d(p,q)$ denotes the minimum distance from point p to contour C.

3.4.2. Clinical metrics

Clinical parameters such as LV volume, LV ejection fraction (LVEF) and myocardial mass (LVM) are another essential aspect of assessing the quality of automatic segmentation. The volume is computed by summation of the number of pixels corresponding to the LV or myocardium binary mask, multiplied by the pixel dimension. Myocardial mass is calculated by the following formula:

$$LVM = Myo - Volume \times 1.05(gram/cm^3) \tag{3.10}$$

and LVEF is defined as:

$$LVEF = \frac{EDV - ESV}{EDV} \times 100\% \tag{3.11}$$

where EDV and ESV are the LV volumes at the end-diastolic and end-systolic phases, respectively.

3.4.3. Statistical analysis

Pearson correlation coefficient (PCC), mean of differences (Bias) and limits of agreement (LOA, 1.96×standard deviation) are assessed to describe the differences and the agreement between automatically and manually derived segmentation. In addition, Bland-Altman is used to further describe the results.

To investigate the statistical significance of the differences between different segmentation models, the Wilcoxon signed-rank test is used to compare the

difference between paired Dice, HD and MCD without assuming the underlying distribution, $P<0.05$ indicates a significant difference.

3.5 Experiments and Results

We trained and evaluated our method on both LUD and ACDC datasets. The network is firstly trained on LUD from scratch, and then we performed transfer learning to train the network on ACDC. All the experiments were executed on a machine equipped with an NVIDIA Quadro RTX 6000 GPU with 24 GB internal memory. The networks were implemented using Keras with the following parameters: Adam optimizer, batch size as 50, learning rate as 10^{-5}, 150 epochs, as well as early stopping, to avoid overtraining the network.

First, we explored and determined the optimal value of parameter *N* in the spatial and temporal stack. Second, we compared the results of three classical segmentation networks, U-Net, SegNet [20] and 3D U-Net, with SAUN based on Dice, MCD, HD, LVEF and LVM on both LUD and ACDC datasets. Meanwhile, to further explore the impact of using YOLO for localization and spatial stack for extracting potential features on the segmentation performance, another two networks named YOLO+U-Net (YUN) and SSM+U-Net (SUN) were employed. The cropped images with a uniform matrix size of 128×128, centered to the original image, were used as the input of U-Net and SegNet. The input of YUN is presented after localization, and input of SUN and SAUN are preprocessed with localization and SSM. For the input of 3D U-Net, for both datasets, all the 2D slices in the ED or ES phase together are stacked to construct a 3D image. Then, all 3D images were resampled into the same resolution of 2.5×2.5×5 mm^3 and the signal intensity normalized to (0,1). Lastly, all 3D images were cropped or padded to a size of 112×112×24 as the input of the 3D U-Net. For the post-processing, the predictions were resampled to their original resolution. All of the networks are assessed using the defined evaluation metrics for different levels of the LV, including apex (25% slices in the apical region and beyond), middle (50% mid slices) and base (25% slices in the basal region and beyond).

3.5.1 Multi-Channel architecture

To analyze the impact of the two multi-channel architectures (SSM and TSM) of different dimensions on the segmentation results, we trained SAUN using SSM and TSM with different dimension parameters N as input. The results presented in Table.3.1 illustrates the segmentation performance for LV chamber and myocardium.

Results of multi-channel architecture showed four TSM versions (*N*=3,5,7,9) achieved stable segmentation performance for LV chamber and myocardium with the best Dice of 0.93 and 0.84, respectively. SSM, however, did work significantly

better than TSM with best performance Dice of 0.95 and 0.86 for chamber and myocardium with *N* set to 3. Hence, SSM with dimension N=3 is regarded as the optimal input of SAUN.

Table.3.1. Dice of segmentation results generated from different multi-channel architectures with various values of parameter N at LUD using SAUN method. N is the dimension parameter.

	Chamber		Myocardium	
Parameters	SSM	TSM	SSM	TSM
N=3	**0.95(0.05)**	**0.93(0.07)**	**0.86(0.07)**	0.84(0.11)
N=5	0.93(0.11)	0.92(0.01)	0.84(0.14)	0.83(0.13)
N=7	0.93(0.12)	0.92(0.10)	0.84(0.13)	0.81(0.14)
N=9	0.92(0.13)	0.92(0.11)	0.82(0.13)	0.82(0.12)

3.5.2 Results in LUD

The performance of the SAUN method was evaluated in the LUD testing data set (69 subjects, 1611 2D images). We compared the segmentation performance for different heart structures between multiple neural networks using the evaluation metrics defined. As the cross-sectional area of the left ventricle at the apical level is very small and the image quality at this level is degraded due to particle voluming, segmentation errors are more likely to occur at this level, although it will have only little effect on the clinical metrics, especially on the LVEF. Hence, we further evaluated the segmentation performance on apex, middle and base level, respectively. Finally, we report the results of the clinical functional parameters.

Table 3.2 and Table 3.3 respectively show the Dice and distance metrics (HD and MCD) comparing manual with automatic segmentation. It can be observed that the networks with localization perform better than those without localization, which confirms that localization can filter out the data noise effectively for the label unbalanced data. Moreover, the SAUN method achieved the best segmentation results compared to the other networks on Dice, HD, and MCD. The results for the individual LV levels further indicate that the SAUN model provides much more precise feature maps, leading to the best evaluation metric scores for both LV chamber and myocardium at all LV levels.

Table.3.2. Comparison of the mean and standard deviation (in parenthesis) of Dice metric on LUD for LV chamber and LV myocardium predicted by different networks. (1) U-Net:basic U-Net without localization, (2)YUN: combine YOLO for localization and basic U-Net, (3) SUN: SSM with N=3 as the input of basic U-Net, (4) SegNet: basic SegNet without localization, (5) SAUN: SSM with N=3 as the input of proposed SAUN network

	Apex		Middle		Base		Average	
Networks	chamber	myocardium	chamber	myocardium	chamber	myocardium	chamber	myocardium
U-Net	0.821 (0.210)	0.692 (0.220)	0.939 (0.040)	0.817 (0.086)	0.924 (0.067)	0.800 (0.105)	0.922 (0.120)	0.799 (0.140)
YUN	0.897 (0.100)	0.794 (0.110)	0.945 (0.036)	0.840 (0.069)	0.909 (0.066)	0.793 (0.110)	0.932 (0.066)	0.825 (0.096)
SUN	0.849 (0.180)	0.752 (0.170)	0.949 (0.035)	0.867 (0.058)	0.938 (0.056)	0.839 (0.096)	0.935 (0.100)	0.848 (0.110)
SegNet	0.794 (0.200)	0.654 (0.200)	0.924 (0.039)	0.812 (0.073)	0.919 (0.062)	0.786 (0.107)	0.908 (0.110)	0.788 (0.130)
SAUN	**0.911 (0.080)**	**0.823 (0.080)**	**0.952 (0.034)**	**0.876 (0.042)**	**0.941 (0.046)**	**0.847 (0.069)**	**0.945 (0.053)**	**0.864 (0.066)**

Table.3.3. Comparison of the mean and standard deviation (in parenthesis) of HD and MCD metrics on LUD dataset for LV chamber and LV myocardium at apex, middle and base regions predicted by different networks. (1) U-Net: basic U-Net without localization, (2) YUN: combine YOLO for localization and basic U-Net, (3) SUN: SSM with N=3 as the input of basic U-Net, (4) SegNet: basic SegNet without localization, (5) SAUN: SSM with N=3 as the input of proposed SAUN

	Apex				Middle				Base			
	Chamber		Myocardium		Chamber		Myocardium		Chamber		Myocardium	
Networks	MCD	HD	MCD	HD	MCD	HD	MCD	HD	MCD	HD	MCD	HD
U-Net	1.584 (1.88)	5.373 (5.99)	1.706 (1.69)	6.969 (7.68)	1.157 (0.52)	4.785 (3.85)	1.311 (0.49)	5.372 (3.97)	1.496 (0.94)	5.467 (4.28)	1.434 (0.72)	6.782 (5.48)
YUN	1.150 (0.51)	3.677 (3.48)	1.228 (0.49)	4.250 (3.57)	1.127 (0.57)	4.072 (3.13)	1.205 (0.45)	4.382 (2.37)	1.528 (0.95)	5.125 (3.89)	1.449 (0.78)	6.556 (5.28)
SUN	1.309 (1.13)	3.491 (3.11)	1.336 (0.91)	4.254 (3.51)	1.028 (0.55)	2.994 (1.73)	1.009 (0.41)	3.302 (1.46)	1.303 (1.01)	4.254 (3.43)	1.187 (0.64)	5.763 (4.85)
SegNet	1.808 (2.98)	4.978 (3.18)	1.785 (0.99)	6.108 (2.98)	1.517 (0.69)	4.260 (1.46)	1.383 (0.36)	4.815 (1.52)	1.639 (0.84)	5.197 (3.22)	1.541 (0.55)	6.864 (4.59)
SAUN	**0.942 (0.46)**	**2.913 (2.52)**	**1.067 (0.41)**	**3.817 (3.58)**	**0.978 (0.48)**	**2.796 (1.41)**	**0.950 (0.36)**	**3.185 (1.29)**	**1.234 (0.66)**	**4.032 (2.85)**	**1.148 (0.47)**	**5.367 (4.11)**

The PCC, bias and LOA of the clinical evaluation metrics comparing automated segmentation results with results from manual segmentation are reported in Table.3.4 and Figure.3.5. For both LVEF and LVM assessment, the proposed SAUN network achieves the highest PCC, the smallest bias and LOA. Table.3.5 summarizes the significance test results between SAUN and the other state-of-the-art methods on LUD, all the P-values are smaller than 0.05, which confirms the significantly better results of SAUN compared to the other methods.

Figure.3.6 illustrates examples of segmentation results obtained by automated SAUN method and conventional manual method from randomly selected cases from the test data. It shows that the automated results are highly similar to the manual reference at both ED and ES phases.

Table.3.4. Results of clinical evaluation metrics from all networks against the reference. (1) U-Net: basic U-Net without localization, (2) YUN: combine YOLO for localization and basic U-Net, (3) SUN: SSM with N=3 as the input of basic U-Net, (4) SegNet: basic SegNet without localization, (5) SAUN: SSM with N=3 as the input of proposed SAUN network.

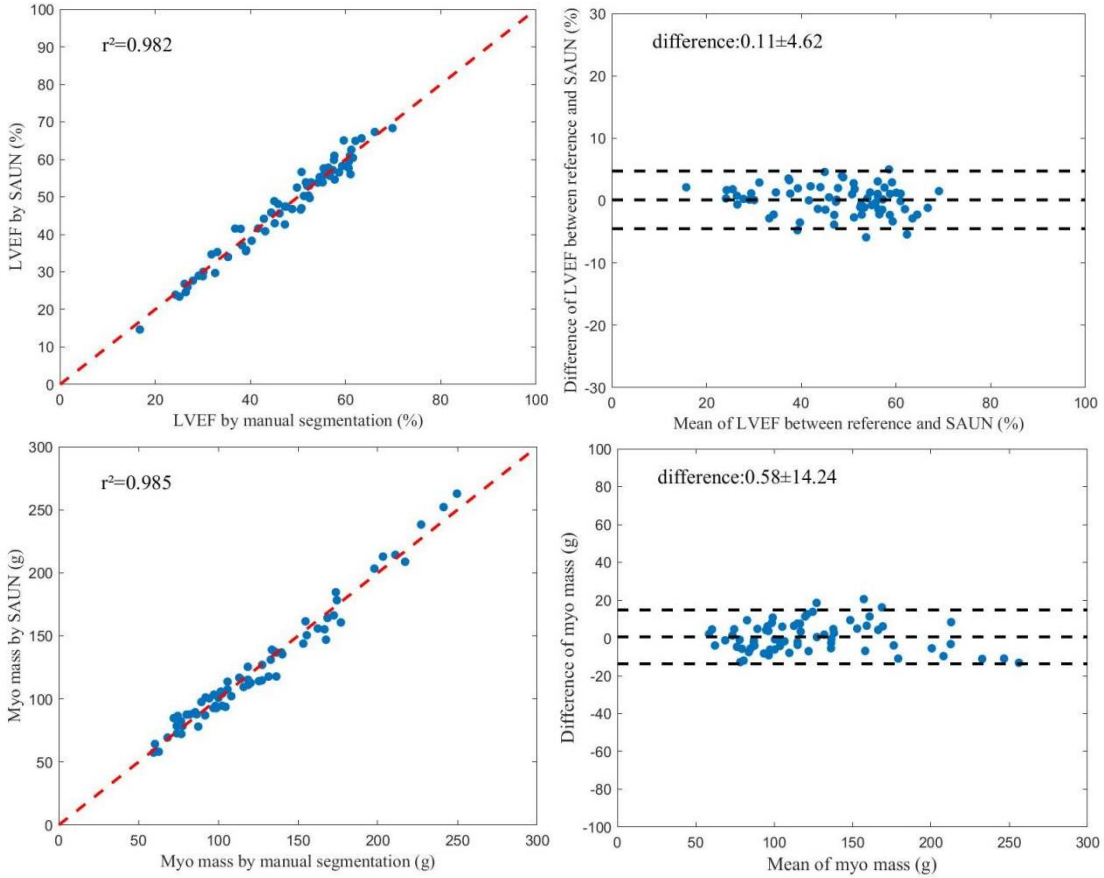
	LVEF		LVM	
Networks	PCC (%)	Bias ± LOA (%)	PCC (%)	Bias ± LOA (g)
U-Net	0.969	2.22±5.89	0.957	3.11±28.60
YUN	0.972	0.52±5.76	0.971	1.51±20.29
SUN	0.974	0.22±5.36	0.976	1.47±19.21
SegNet	0.967	3.34±6.01	0.954	4.51±31.57
SAUN	**0.982**	**0.11±4.62**	**0.985**	**0.58±14.24**

Table.3.5. Wilcoxon signed-rank test based significance test results on LUD dataset. (1) W(SAUN,U-Net): Wilcoxon signed-rank test's P-value between SAUN and U-Net, (2) W(SAUN,YUN): Wilcoxon signed-rank test's P-value between SAUN and YUN(YOLO+U-Net), (3) W(SAUN,SUN): Wilcoxon signed-rank test's P-value between SAUN and SUN(SSM stack + U-Net), (4) W(SAUN, SegNet):Wilcoxon signed-rank test's P-value between SAUN and SegNet.

	Chamber			Myocardium		
	Dice	HD	MCD	Dice	HD	MCD
W(SAUN, U-Net)	1.36E-05	7.09E-12	0.0129	2.86E-09	1.55E-12	6.22E-04
W(SAUN, YUN)	5.21E-10	5.27E-09	1.37E-08	2.09E-12	1.45E-11	3.00E-10
W(SAUN, SUN)	1.81E-10	6.70E-03	4.87E-07	4.10E-09	2.56E-04	3.24E-07
W(SAUN, SegNet)	1.07E-08	1.36E-12	6.56E-07	1.29E-11	7.99E-13	1.19E-06

1.36E-05 means 1.36×10^{-5}.

Figure.3.5. Correlation and Bland-Altman plots comparing LV ejection fraction and LV mass derived from either the SAUN method and manual segmentation on LUD.



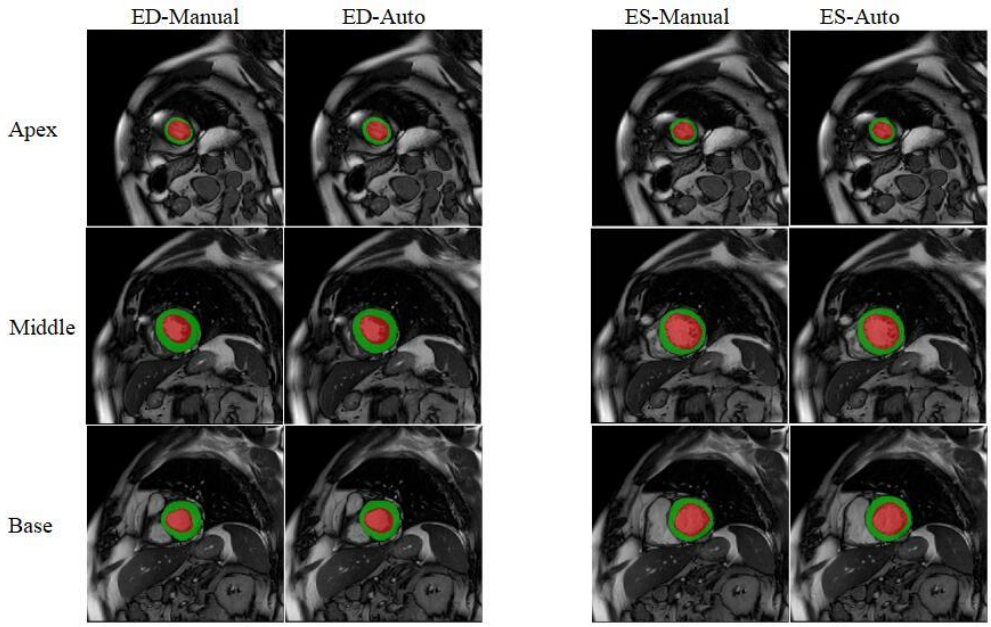


Figure.3.6. Examples of the segmentation results from the SAUN method. The left two columns show ED images, and the right two columns show images of ES phase. For each phase, images at the apex, middle and base levels are shown.

3.5.3 Results in ACDC

We also compared our method with other approaches on the public ACDC 2017 dataset, which includes short-axis Cine MR exams of 100 patients with manual contours. As in this dataset, manual contours are only defined in the ED and ES phases; all results are based on those two phases only.

Table.3.6 summarizes the segmentation results for the ACDC dataset. The best segmentation results on both ED and ES phases are obtained using the SAUN method. In Table.3.7 and Figure.3.7, the PCC, bias and LOA are presented and illustrated for the comparison of the clinical parameters. It shows that the prediction results are highly correlated to the reference with a PCC of 0.985 for LVEF and 0.981 for LVM. The Bland-Altman analysis illustrated in Figure.3.7 reveals a bias for LVEF and LVM, which is close to zero, while the LOA is less than 5% for LVEF and less than 6 g for LVM.

Table.3.6. Comparison of the mean and standard deviation (in parenthesis) of segmentation results on ACDC dataset for LV chamber and LV myocardium segmentation by different networks. (1) U-Net: basic U-Net without localization, (2) YUN: combine YOLO for localization and basic U-Net, (3) SUN: SSM with N=3 as the input of basic U-Net, (4) SegNet: basic SegNet without localization, (5) 3D U-Net: basic 3D U-Net without localization (6) SAUN: SSM with N=3 as the input of proposed SAUN network.

	ED						ES					
	Chamber			Myocardium			Chamber			Myocardium		
Networks	Dice	HD	MCD	Dice	HD	MCD	Dice	HD	MCD	Dice	HD	MCD
U-Net	0.940(0.051)	10.780(7.29)	0.596(0.39)	0.836(0.089)	11.538(6.72)	0.791(0.47)	0.841(0.074)	11.854(7.94)	1.529(0.96)	0.812(0.083)	13.338(11.12)	1.358(1.15)
YUN	0.942(0.067)	10.163(7.06)	0.571(0.45)	0.847(0.057)	11.144(6.78)	0.714(0.31)	0.861(0.076)	11.255(6.47)	1.057(0.76)	0.838(0.061)	12.478(6.18)	1.206(0.78)
SUN	0.941(0.058)	11.347(4.97)	0.614(1.33)	0.824(0.130)	11.975(7.88)	0.746(0.84)	0.841(0.069)	13.259(5.91)	1.618(0.98)	0.808(0.096)	13.750(7.31)	1.724(0.69)
SegNet	0.932(0.056)	10.240(6.11)	0.577(0.58)	0.833(0.081)	11.241(7.74)	0.757(0.64)	0.839(0.084)	11.311(6.32)	1.088(0.86)	0.798(0.092)	12.673(8.68)	1.188(0.81)
3D U-Net	0.935(0.048)	10.634(7.42)	0.645(1.67)	0.811(0.087)	12.217(6.37)	0.849(1.32)	0.847(0.089)	11.281(7.71)	1.106(1.94)	0.792(0.067)	13.681(9.69)	1.922(1.06)
SAUN	**0.956 (0.031)**	**9.759(3.45)**	**0.541(0.28)**	**0.877(0.064)**	**10.192(4.37)**	**0.672(0.189)**	**0.887(0.061)**	**10.132(5.35)**	**1.024(0.52)**	**0.873(0.058)**	**11.711(7.36)**	**0.939(0.62)**

Table.3.7. Results of clinical evaluation metrics from all networks against the reference. (1) U-Net:basic U-Net without localization, (2) YUN: combine YOLO for localization and basic U-Net, (3) SUN: SSM with N=3 as the input of basic U-Net, (4) SegNet: basic SegNet without localization, (5) 3D U-Net: basic 3D U-Net without localization (6) SAUN: SSM with N=3 as the input of proposed SAUN network.

	LVEF		Myo Mass	
Networks	PCC	Bias±LOA	PCC	Bias±LOA
U-Net	0.956	7.40(12.06)	0.965	6.06(9.51)
YUN	0.972	2.04(15.64)	0.974	2.60(7.08)
SUN	0.962	0.29(11.09)	0.953	-2.69(10.56)
SegNet	0.947	8.17(11.59)	0.958	9.26(10.97)
3D U-Net	0.948	9.51(17.61)	0.963	5.37(8.96)
SAUN	**0.985**	2.39(**4.61**)	**0.981**	**-0.42(5.74)**

Table.3.8 reports almost all the P-values between SAUN and U-Net, SegNet and 3D U-Net on ACDC dataset are smaller than 0.05, which confirms there is a significant improvement of SAUN compared to the other state-of-the-art methods. Figure.3.8. shows the example segmentation results of two randomly selected cases from the testing set.

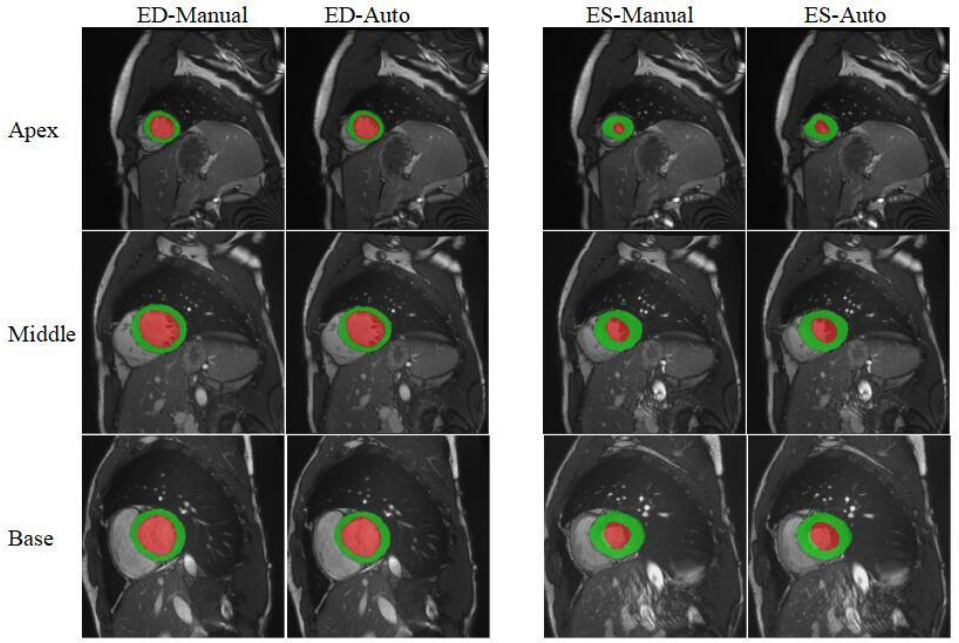


Figure.3.8. Examples of segmentation from the SAUN method from two randomly selected cases from the ACDC dataset. The left two columns show ED images, and the right two columns images of ES phase. For each phase, images at the apex, middle and base levels are shown.

Figure.3.7. Correlation and Bland-Altman plots of clinical parameters comparing results derived from the SAUN method and manual segmentation on ACDC data set.

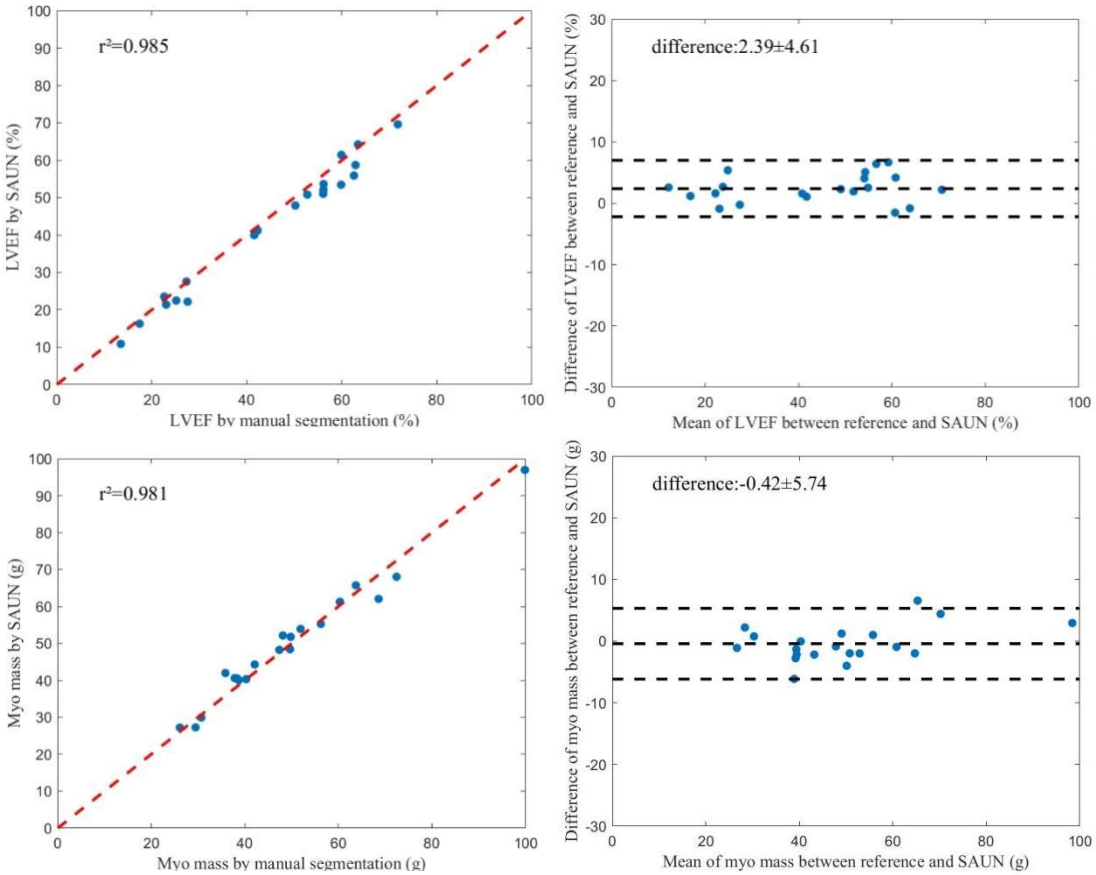


Table.3.8. Wilcoxon signed-rank test based significance test results on ACDC dataset. (1)W(SAUN,U-Net): Wilcoxon signed-rank test's P-value between SAUN and U-Net, (2)W(SAUN,YUN): Wilcoxon signed-rank test's P-value between SAUN and YUN(YOLO+U-Net), (3)W(SAUN,SUN): Wilcoxon signed-rank test's P-value between SAUN and SUN(SSM stack+U-Net), (4)W(SAUN,SegNet):Wilcoxon signed-rank test's P-value between SAUN and SegNet, (5)W(SAUN,3D U-Net):Wilcoxon signed-rank test's P-value between SAUN and 3D U-Net.

	ED						ES					
	Chamber			Myocardium			Chamber			Myocardium		
	Dice	HD	MCD	Dice	HD	MCD	Dice	HD	MCD	Dice	HD	MCD
W(SAUN,U-Net)	3.65E-03	0.475	5.58E-03	3.62E-05	0.189	1.34E-05	2.10E-04	1.02E-03	7.08E-04	3.81E-06	3.22E-04	7.08E-04
W(SAUN,YUN)	2.61E-04	0.0241	3.28E-03	1.34E-04	7.30E-03	3.22E-04	6.48E-03	1.34E-05	8.31E-03	3.12E-04	6.29E-05	3.65E-03
W(SAUN,SUN)	1.49E-06	0.0583	2.10E-04	5.72E-06	0.0441	1.33E-06	0.0215	0.0355	1.43E-03	9.54E-06	0.1893	1.91E-06
W(SAUN,SegNet)	3.97E-05	4.22E-03	1.91E-06	2.91E-06	3.65E-03	1.19E-05	9.54E-06	1.68E-04	1.69E-03	7.01E-05	1.05E-04	3.95E-04
W(SAUN,3D U-Net)	1.89E-04	0.0124	6.81E-04	8.86E-05	2.20E-03	1.40E-04	0.0407	0.232	5.93E-04	1.03E-04	0.0479	1.63E-04

3.65E-03 means 3.65×10^{-3}

3.6 Discussion

To explore more spatiotemporal information for automatic cine MRI segmentation, we proposed two stack models to construct a multi-channel architecture, then introduced a segmentation network based on a stack attention mechanism to weight the feature maps from different channels. The method was evaluated on an internal and a public dataset demonstrating competitive results compared other typical CNN networks.

3.6.1. Multi-Channel Architecture Comparison

Our results demonstrate that, when the spatial stack was used to combine the target slice and its neighboring slices from the same phase together as the input of the network, the performance improved in the testing data. The segmentation results were found to be sensitive to the dimension of the spatial stack model. For both spatial and temporal stack the optimal value for the dimension parameter N was found to be 3. However, the use of temporal stack had a negligible impact on the cardiac segmentation results. It also can be observed that all of the evaluation metrics from the spatial stack and SAUN are much better than those predicted from basic U-Net and SegNet whose input is a single 2D image, which illustrates the spatial stack model can provide more useful information than a single MRI slice and temporal stack. The images in the temporal stack are similar to each other and provided comparable features for the network. Whereas, the images from the spatial stack vary obviously with the heart region, and when combining the target slice and its neighboring spatial slices together as the input of the network, the spatial stack contains more information about position, size and shape of the heart.

However, including more slices in the stack does not necessarily result in better segmentation results. This was clearly demonstrated by the multi-channel architecture comparison experiment, which showed that when the parameter N was set to a value higher than 3, which implies introducing more spatiotemporal information, the performance degraded. In addition, the limited difference between the stack network and 2D network is only at the first convolution layer. The stack network regards a (W×L×N) volume as an input, and the 2D network accepts a single slice (W×L×1) as an input, while the other parts of the network are the same, which leads to the stack networks having more parameters only at the first convolution layer.

Unlike the stack model transferring stack input into multiple 2D feature maps, Çiçek [21] and Perslev [22] proposed a 3D network for the segmentation. A 3D network will introduce more parameters to extracts the depth-wise features through the entire network than 2D and stack networks. During the training process, in order to fit the scan volume in memory, Arjun [14] set the batch size to 1, resulting in less

stable feature regularization. Other researchers set the number of filters at the initial convolution layers into a low value to reduce the number of parameters, but a lower number of filters will likely contribute to inferior feature representation and in turn cause less accurate segmentation. Another disadvantage of repeated pooling and convolution operation is the loss in spatial information in cases with fewer slices. Whereas, the spatial stack network can maintain the spatial information and keep the approximately same number of parameters as a 2D network resulting in improved segmentation performance.

3.6.2. Effect of stack attention

SUN achieved better performance than YUN in LUD; however, in ACDC SUN did not get as good results as YUN. Because the slice gap (5mm or 10mm) in ACDC is larger than in LUD (2mm), the recognizable variance between the spatially neighboring images, such as the shape, size or outline of LV is larger in ACDC, which will confuse the network. Meanwhile, when we combine neighboring slices having imbalanced labels to build the spatial stack, the proportion of the background will increase, compared with a single 2D image, which results in the spatial stack generating more data noise. This issue is overcome by employing the proposed attention mechanism which weighs and fuses the feature maps of different channels from the spatial stack and balances the noise.

During fusion of the features, the features from the target slice should be regarded as the primary components, and the others from the neighboring slices should be considered as the additional information. In the stack attention mechanism, the target slice serves as a guideline to keep the primary features, and the global pooling is used to compute the weights of different channels to select the feature maps generated from the target slice. Therefore, the stack attention can not only reserve the primary feature information but also balance the importance of different channels to pick up the more important maps. Figure.3.9 illustrates the process of SAUN method extracting the feature maps from a random sample taken from the LUD data set. The first row illustrates the features for the LV chamber, and the second row is the features for the myocardium. The first column is one test case, the last column is the ground truth segmentation, and the middle four columns represent feature maps from the low, middle, high level and final layer.

It can be observed from the performance of Dice in LUD that the segmentation predicted by SAUN for the apical level is much better than the other approaches., When comparing the results from SUN and SAUN, it can be found that the LOA from the SAUN is further improved. The clinical evaluation results in ACDC illustrate that the PCC, bias and the limit of agreement computed by SUN is inferior compared to the other networks. The evaluations predicted by SAUN achieve best

with the attention mechanism. The Bland-Altman plots show almost all of the subjects from LUD and ACDC distribute between the upper bound and lower bound, which confirms that in the clinical measures the automated method is almost unbiased to the manual results. The experiments demonstrate that the proposed stack attention mechanism performs well in filtering out data noise during integrating neighboring spatial information, weighting and confusing the feature maps of various levels as well.

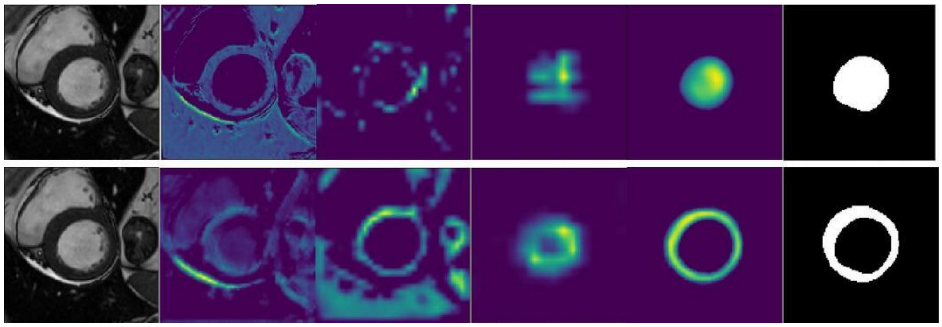


Figure.3.9. Feature map visualization of SAUN. There are 42 convolutional layers in SAUN, we did the visualization for each convolutional layer. The first and last columns are the original image and the ground truth, the other four columns represent the feature maps from low, middle, high levels (from 3rd, 18th, 32nd convolutional layer) and the output of the final layer.

Our proposed method has several limitations. It ignores the right ventricle (RV) and only provided segmentation for the left ventricle and myocardium. If more annotation information about the RV is provided for the network, the segmentation results could become more accurate. In the current implementation, we separately trained the localization and segmentation networks. As for both tasks, the MR image features need to be explored; integration of both tasks into a single network would result in improved efficiency of the segmentation algorithm.

3.7 Conclusion

In this work, we proposed a Stack Attention U-Net based method for automatic LV segmentation in short-axis cine MRI and confirmed its benefits in integrating more information from neighboring spatial images by employing an attention mechanism to weight each channel of the feature maps. The experimental results demonstrate that the proposed approach exceeds existing state-of-the-art segmentation methods and verify its potential clinical applicability.

Acknowledgment

Prof. Sven Plein from the University of Leeds is acknowledged for granting access to the image data used in this work. We also would like to acknowledge the organizer

of ACDC 2017 challenge to collect and public the dataset. X. Sun is supported by the China Scholarship Council No. 201808110201.

DISCLOSURE OF CONFLICTS OF INTEREST

The authors have no conflicts of interest to disclose.

Data Availability Statements

The LUD datasets generated and analysed during the current study are not publicly available, due to the nature of this research, participants of this study did not agree for their data to be shared publicly.

The ACDC data that support this study are openly available at https://acdc.creatis.insa-lyon.fr/description/databases.html.

References

1. Kaus MR, Von BJ, Weese J, Niessen W, Pekar V. Automated segmentation of the left ventricle in cardiac MRI. Med Image Anal. 2004;8:245-254.
2. Khened M, Kollerathu VA, Krishnamurthi G. Fully convolutional multi-scale residual DenseNets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers. Med Image Anal. 2019;51:21-45.
3. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention; 2015.
4. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. in IEEE CVPR, pages 3431-3440; 2015.
5. Bai W, Sinclair M, Tarroni G, et al. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. J Cardiovasc Magn Reson. 2018; 20(1):65-77.
6. Isensee F, Jaeger PF, Full PM, et al. Automatic Cardiac Disease Assessment on cine-MRI via Time-Series Segmentation and Domain Specific Features, in STACOM, pages 120–129, 2018.
7. Khened M, Kollerathu VA, Krishnamurthi G. Fully convolutional multi-scale residual DenseNets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers. Med Image Anal. 2019;51:21-45.
8. Qin C, Bai W, Schlemper J, Petersen SE, Piechnik SK, Neubauer S, et al. Joint learning of motion estimation and segmentation for cardiac MR image sequences. International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 472-480; 2018.
9. Tao, Qian, Wenjun Yan, Yuanyuan Wang, et al. Deep learning–based method for fully automatic quantification of left ventricle function from cine MR images: a multivendor, multicenter study. Radiology. 2019;290(1): 81-88.
10. Cheng J, Tsai YH, Wang S, Yang MH. Segflow: Joint learning for video object segmentation and optical flow. IEEE international conference on computer vision, pages 686-695;2017.
11. Zhao N, O'Connor D, Gu W, Ruan D, Basarab A, Sheng K. Coupling reconstruction and motion estimation for dynamic MRI through optical flow constraint. SPIE:Image Processing; 2018.
12. Yan W, Wang Y, Li Z, Van Der Geest RJ, Tao Q. Left ventricle segmentation via optical-flow-net from short-axis cine MRI: preserving the temporal coherence of cardiac motion. International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 613-621;2018.

13. Zhang N, Yang G, Gao Z, Xu C, Zhang Y, Shi R, et al. Deep learning for diagnosis of chronic myocardial infarction on nonenhanced cardiac cine MRI. Radiology. 2019;291(3):606-17.
14. Desai AD, Gold GE, Hargreaves BA, Chaudhari AS. Technical considerations for semantic segmentation in MRI using convolutional neural networks. arXiv preprint arXiv:1902.01977. 2019.
15. Abraham N, Khan NM. A novel focal tversky loss function with improved attention u-net for lesion segmentation. International Symposium on Biomedical Imaging, pages 683-687; 2019.
16. Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, et al. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999. 2018.
17. Huang Q, Yang D, Wu P, Qu H, Yi J, Metaxas D. MRI reconstruction via cascaded channel-wise attention network. International Symposium on Biomedical Imaging, pages 1622-1626; 2019.
18. Bernard O, Lalande A, Zotti C, Cervenansky F, Yang X, Heng PA, et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?. IEEE Trans Med Imaging. 2018;37:2514-2525.
19. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. IEEE conference on computer vision and pattern recognition, pages 779-788; 2016.
20. Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell. 2017; 39(12):2481-95.
21. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. International conference on medical image computing and computer-assisted intervention, pages 424-432; 2016.
22. Perslev M, Dam EB, Pai A, Igel C. One network to segment them all: A general, lightweight system for accurate 3d medical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 30-38; 2019.

Chapter 4 Right Ventricle Segmentation via Registration and Multi-input Modalities in Cardiac Magnetic Resonance Imaging from Multi-disease, Multi-view and Multi-center

This chapter was adapted from:

Xiaowu Sun, Li-Hsin Cheng, Rob J. van der Geest. **Right Ventricle Segmentation via Registration and Multi-input Modalities in Cardiac Magnetic Resonance Imaging from Multi-disease, Multi-view and Multi-center.** International Workshop on Statistical Atlases and Computational Models of the Heart (STACOM). Springer, Cham, 2021.

Abstract

Quantitative assessment of cardiac function requires accurate segmentation of cardiac structures. Convolutional Neural Networks (CNNs) have achieved immense success in automatic segmentation in cardiac magnetic resonance imaging (cMRI) given sufficient training data. However, the performance of CNN models greatly degrade when the testing data is from different vendors or different centers. In this paper, we introduce the use of image registration to propagate annotation masks from labeled images to unlabeled images as to enlarge the training dataset. Furthermore, we investigated various input modalities including 3D volume, single-channel 2D image, multi-channel 2D image constructed from spatial and temporal stack to extract more features to improve do-main generalization in cMRI segmentation. We evaluated our method in M&Ms-2 challenge testing data (https://www.ub.edu/mnms-2/), achieving averaged Dice scores of 0.925, 0.919 and Hausdorff Distance of 10.587 mm, 6.045 mm in right ventricular segmentation in short-axis view and long-axis view respectively.

4.1 Introduction

In clinical routine, cardiac magnetic resonance imaging (cMRI) is considered a standard reference for the diagnosis of cardiac disease. Accurate segmentation of cardiac structures such as left ventricle (LV), myocardium and right ventricle (RV) is essential to quantitatively assess the cardiac function. Traditional manual segmentation method not only is time-consuming but also prone to inter-rater experience.

In recent years, deep learning based automatic segmentation approaches have been achieved immense success in cardiac segmentation. Tran et al. was the first to employ the fully convolutional neural (FCN) network for LV and RV segmentation in short-axis MRI [1]. Poudel proposed a recurrent FCN network ensembling the spatial information for LV segmentation [2]. However, the performance of most of those deep learning based models degrade dramatically when the trained model is applied directly on other unseen datasets from different centers or vendors. Differences in image protocols, disease characteristics, scanner-specific bias and the other factors remain even after careful pre-processing [3]. In addition, the RV has a more complex shape and border characteristics compared to the LV. Hence, the M&Ms-2 challenge is motivated to build a method to segment the RV using multi-center, multi-disease and multi-view cMRI data.

The most straight forward approach to tackle this problem is to collect and annotate data from multiple centers, vendors and patient pathologies. Tao used a large heterogeneous data with 41,593 images from different centers and different vendors to train a CNN model and achieved a good generalization [4]. Chen demonstrated that applying data augmentation strategies on a single-site single-scanner dataset could improve the performance on an unseen dataset across different sites or scanners [5]. Based on those studies, we hypothesize that a large-scaled pooling data from different domains could improve a model's performance on an unseen dataset. Additionally, in the conventional CNN models, the information derived from the neighboring images is usually ignored. Hence, we introduced two stack model to extract the spatiotemporal features to improve the performance.

In this paper, given limited data, we investigated several methods to generate more training data and extract more features including 1): The use of image registration to propagate annotation masks to unlabeled phases 2): Introducing the spatial and temporal neighboring images to construct a multi-channel 2D image to integrate more spatiotemporal information for the RV segmentation task.

4.2 Data

Table.4.1. Description of training, validation and testing dataset

Pathology	Num. of training	Num. of validation	Num. of testing
Normal subjects	40	5	30
Dilated Left Ventricle	30	5	25
Hypertrophic Cardiomyopathy	30	5	25
Congenital Arrhythmogenesis	20	5	10
Tetralogy of Fallot	20	5	10
Interatrial Communication	20	5	10
Dilated Right Ventricle	0	5	25
Tricuspid Valve Regurgitation	0	5	25
Total	160	40	160

The M&Ms-2 challenge provides 360 cases (160 for training, 40 for validation and 160 for testing) in both short-axis (SA) and long-axis (LA) views from four different centers, acquired with three vendors (General Electric, Philips and Siemens). As shown in Table.4.1, except the normal subjects, there are five pathologies in the training dataset, two pathologies are not present in the training dataset but only in the validation and testing dataset. In addition, only end-diastolic (ED) and end-systolic (ES) phases in the training data are annotated by experienced experts, including LV, RV and left ventricular myocardium (MYO). Although this challenge focused on the RV segmentation, in our experiments, LV and MYO annotations were also used to constrain the RV segmentation.

4.3 Method

4.3.1. Registration

In the available dataset, only the ED and ES phases are labeled, while the other phases are continuous in time consistent with the ED and ES phases. All the phases from the same case have an almost identical intensity distribution, which will alleviate the errors caused by inter-subject variability [6]. Hence, we used intensity-based registration method to propagate the labels, regarding the ED and ES as the template. The progress is described in Figure.4.1. Given three phases (ED, ES and unlabeled), the ED and ES are firstly registered to the unlabeled phase, generating two geometric transformation matrixes named ES-Tf and ED-Tf, then the transformation matrix with smaller norm was used to propagate the mask. Matlab inbuilt functions imregtform and imwarp were used to implement the registration

[11]. Mean square error (MSE) and Affine were set as the similarity metric and transformation type.

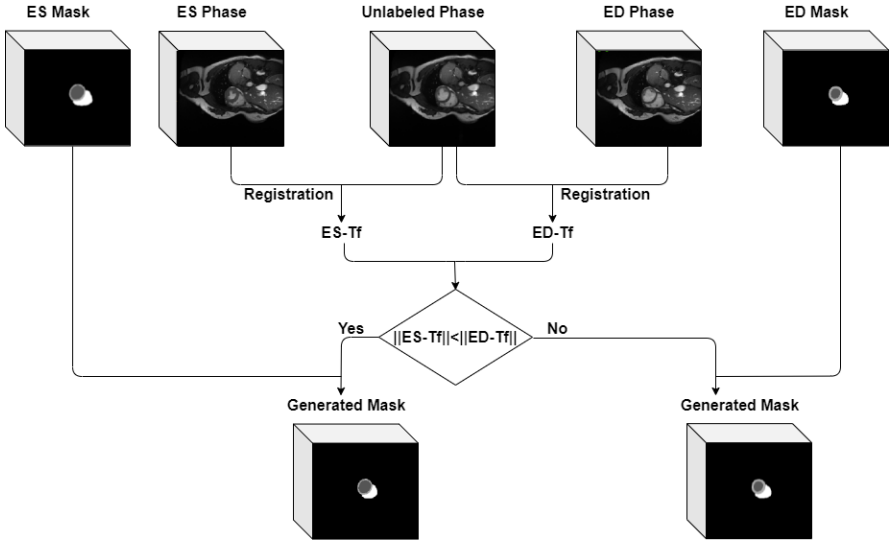


Figure.4.1. Registration method to generate a mask for a unlabeled phase in SA view.

4.3.2. Input modality of network

A illustrated in Figure.4.2 a short-axis cine MRI scan contains multiple slices and multiple phases. Images from the same slice level describe a cardiac cycle, while images from the same phase describe the complete heart structure. In conventional methods [4,5], each single-channel 2D image or a 3D volume with the whole images from the same phase is usually considered as the input of a network. Although using a single-channel 2D image as the input could enlarge the training dataset, a 3D volume can provide more spatial information for the segmentation than using a 2D image. As a compromise, a spatial stack or temporal stack model, as proposed in our previous work [7], can be used to build a multi-channel 2D image which can provide accompanying spatial or temporal information respectively.

Spatial Multi-channel 2D image (SMI). The slices from the same phase are used to construct a SMI. As illustrated in Figure.4.2, three 2D images from Phase 9 Slice 5,6,7 are used to build a 3-channels SMI for the image of Slice 6 Phase 9.

Temporal Multi-channel 2D image (TMI). In a similar way, a TMI consists of several neighboring phases of a particular slice. As shown in Figure.4.2, images from Slice 6 Phase 8,9,10 are used to construct a 3-channels TMI for the image of Slice 6 Phase 9.

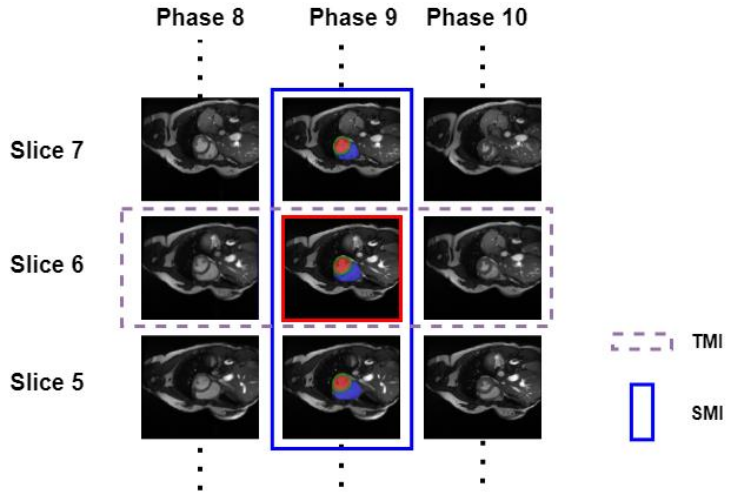


Figure.4.2. An example of constructing a spatial multi-channel 2D image and temporal multi-channel 2D image. The image in the red box is the target image which will be segmented. The three spatial neighboring images in the blue box is called an SMI with three channels, where the top one is the first channel, the middle one is the second channel and so on. The TMI consists of three temporal neighboring images in the dash-line box, the left one is the first channel and the right is the last channel.

Table.4.2 shows a brief summary of the training data size in SA view after combining the registration, SCI (single-channel 2D image), SMI and TMI. The original MnMS-2 dataset contains 320 3D volumes and 2,704 2D images with labeled annotation for training, when applying the registration approach to propagate the annotation masks, the data size of 3D volume and single-channel 2D image increased to 4,152 and 32,330 respectively. The LA-view images were acquired as single slice, resulting in the LA images being multi-phase single-slice. The 3D volumes and SMI cannot be constructed in LA view. Hence, the model achieving the best performance in SA view was used as the pre-trained model for the LA view instead of using different input modalities.

Table.4.2. Training dataset description in SA MRI. **SCI**: single-channel 2D image. The number of channel in SMI and TMI is set to 5.

Data modality	Used registration	Training Data Size
SCI	No	2,704
SCI	Yes	32,330
3D volume	No	320
3D volume	Yes	4,152
SMI	No	2,704
SMI	Yes	32,330
TMI	No	2,704
TMI	Yes	32,330

4.3.3. Network Architecture

nnUNet [8] based on the U-Net architecture is a fully automatic and out-of-the-box medical image segmentation framework. To improve the robustness of domain shift in cardiac MRI, nnUNet_MMS [9] was specially designed by investigating various data augmentation techniques and ranked first at the first edition of M&Ms [10]. Hence, we introduced nnUNet as the baseline, and built our method upon nnUNet_MMS. All the models in this study are based on a 2D network. The data augmentation methods are the same in nnUNet_MMS model.

Since the propagated masks are not as accurate as the manual masks, those pseudo data was used to pre-train the model and the manually labeled data was applied to fine-tune the pre-trained network. The results are reported using Dice and Hausdorff Distance (HD). All experiments were executed on an NVIDIA Quadro RTX 6000 GPU with 24 GB internal memory.

4.4 Experiments and Results

4.4.1. Validation Set Results

We first evaluated the performance of different networks with different input modalities in the SA view from the validation dataset. Then we compared the results in the LA view with or without pre-training from SA view.

Table.4.3. Segmentation results generated from different networks with different input modalities in the validation dataset in SA view.

Network	Input	Used registration data to pre-train network	Dice	HD (mm)
nnUNet(Baseline)	3D volume	No	0.912	10.318
nnUNet_MMS	3D volume	No	0.915	10.475
	3D volume	Yes	**0.922**	**9.472**
	SMI	No	0.919	9.577
	SMI	Yes	0.920	9.539
	TMI	No	0.917	10.343
	TMI	Yes	0.914	12.221
	SCI	No	0.915	11.354
	SCI	Yes	0.914	10.515

Table.4.3 shows that using 3D volume without registration processing as the input, nnUNet_MMS achieved a slightly better dice than nnUNet, but yielded worse HD. However, when the registration method is applied to generate more 3D volume data to pre-train nnUNet_MMS, it achieved the best performance with a Dice of 0.922 and HD of 9.472 mm. It also can be observed that the segmentation results derived from the two stack models (SMI and TMI) are better than that from SCI, which

confirmed that SMI and TMI could provide more spatial and temporal information for the segmentation task.

The results in LA views presented in Table.4.4 illustrates that the performance increased by 0.01 and 0.7 in terms of Dice and HD as a result of transferring the pre-trained model from SA view to LA view. Figure.4.3 and Figure.4.4 show some segmentation examples derived from the best model.

Table.4.4. Segmentation results in the validation dataset in LA view.

Network	Transfer from SA	Dice	HD
nnUnet(Baseline)	No	0.910	6.004
nnUNet_MMS	No	0.908	6.081
	Yes	**0.920**	**5.343**

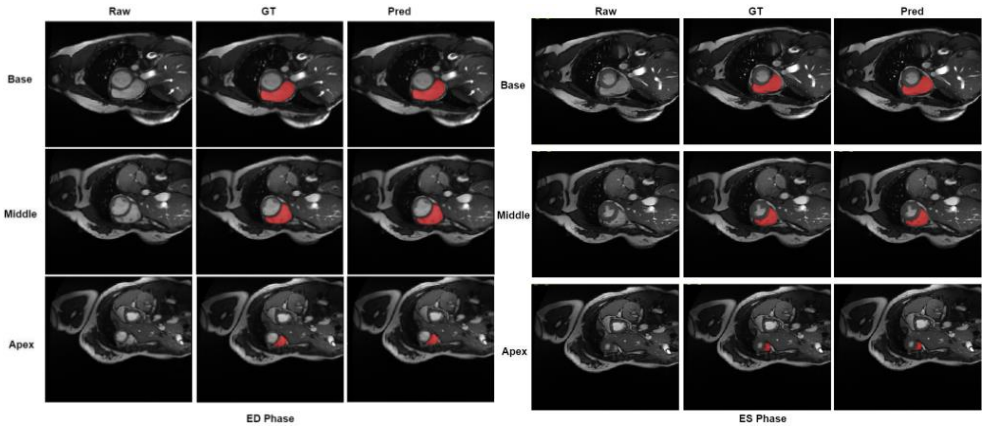


Figure.4.3. A visual example from the apex, middle and base levels at ED (left) and ES (right) phases in SA view.

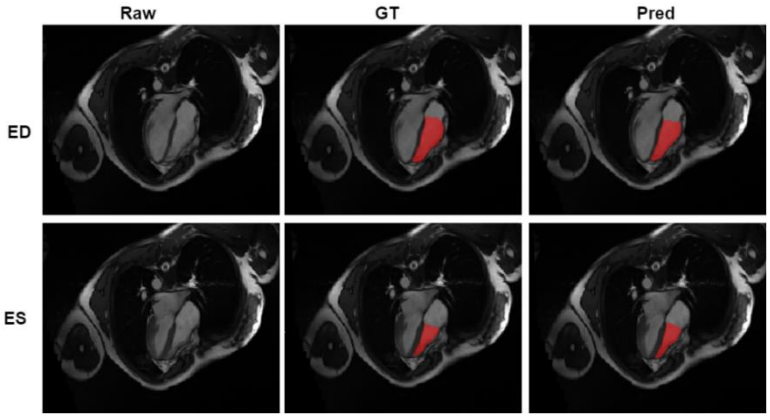


Figure.4.4. A visual example at ED and ES phase in LA view.

4.4.2. Testing Set Results

We chose the model which performs best in the validation data as the final model. As the testing dataset is hidden by the organizer, we submitted our final model to the organizer and evaluated the performance online. Table.4.5 shows the details of our method on the hidden test data. In the SA view our method performed best in congenital arrhythmogenesis yielding 0.949, 8.45 mm for Dice and HD. The best results in LA view are generated from the normal subjects with 0.935 and 5.006 mm for Dice and HD. In addition, two pathologies (dilated right ventricle and tricuspid valve regurgitation) are not present in the training data but only in the testing data. The results on those two pathologies reveal that our approach obtains promising performance on an unseen pathology.

Table.4.5. Segmentation results on 8 pathologies of the hidden test set. The mean and standard deviation are reported.

	Dice		HD (mm)	
Pathology	SA	LA	SA	LA
Normal subjects	0.922±0.050	**0.935± 0.035**	8.999±4.540	**5.006±2.657**
Dilated Left Ventricle	0.922±0.084	0.915± 0.052	13.257±13.134	5.944±3.547
Hypertrophic Cardiomyopathy	0.934±0.057	0.932± 0.033	10.214±5.842	5.343±2.916
Congenital Arrhythmogenesis	**0.949±0.028**	0.934± 0.031	**8.450±4.838**	5.125±1.738
Tetralogy of Fallot	0.920±0.034	0.914± 0.037	14.157±8.232	7.404±3.673
Interatrial Communication	0.910±0.048	0.906±0.066	12.045±4.189	8.021±6.089
Dilated Right Ventricle	0.924±0.045	0.897± 0.121	10.397±5.223	7.064±5.091
Tricuspidal Regurgitation	0.923±0.040	0.914± 0.039	9.236±3.675	6.112±3.349
Overall	0.925±0.055	0.919±0.063	10.587±7.241	6.045±3.824

4.5 Conclusion

In this paper, we investigated label propagation and multiple input modalities to increase the robustness in right ventricle segmentation from multi-disease, multi-view and multi-center cMRI data. To enlarge the training dataset, we explored the use of image registration to propagate annotation masks to unlabeled phases. We further systematically investigated the effect of using different input modalities including 3D volumes, single-channel 2D image, spatial stack and temporal stack. The results illustrate that spatial stack and temporal stack provide more information for the segmentation task, and using 3D volume with label propagation could further improve the generalization ability in a unseen dataset.

Declaration. The authors of this paper declare that the segmentation methods implemented in this challenge has not used any pre-trained models nor additional MRI datasets other than those provided by the organizers.

References

1. Tran, Phi Vu. "A fully convolutional neural network for cardiac segmentation in short-axis MRI." arXiv preprint arXiv:1604.00494 (2016).
2. Poudel, Rudra PK, Pablo Lamata, and Giovanni Montana. "Recurrent fully convolutional neural networks for multi-slice MRI cardiac segmentation." Reconstruction, segmentation, and analysis of medical images. Springer, Cham, 2016. 83-94.
3. Glocker, Ben, et al. "Machine learning with multi-site imaging data: An empirical study on the impact of scanner effects." arXiv preprint arXiv:1910.04597 (2019).
4. Tao, Q., et al.: Deep learning-based method for fully automatic quantification of left ventricle function from cine MR images: a multivendor, multicenter study. Radiology 290(1), 81–88 (2019)
5. Chen, Chen, et al. "Improving the generalizability of convolutional neural network-based segmentation on CMR images." Frontiers in cardiovascular medicine 7 (2020): 105.
6. Zhang, Yao, et al. "Semi-supervised Cardiac Image Segmentation via Label Propagation and Style Transfer." International Workshop on Statistical Atlases and Computational Models of the Heart. Springer, Cham, 2020.
7. Sun, Xiaowu, et al. "SAUN: Stack attention U-Net for left ventricle segmentation from cardiac cine magnetic resonance imaging." Medical Physics 48.4 (2021): 1750-1763.
8. Isensee, Fabian, et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation." Nature methods 18.2 (2021): 203-211.
9. Full, Peter M., et al. "Studying robustness of semantic segmentation under domain shift in cardiac MRI." International Workshop on Statistical Atlases and Computational Models of the Heart. Springer, Cham, 2020.
10. Campello, Víctor M., et al. "Multi-centre, multi-vendor and multi-disease cardiac segmentation: The M&Ms challenge." IEEE Transactions on Medical Imaging (2021).
11. “Intensity-based automatic image registration - MATLAB& Simulink.” [Online]. Available: https://www.mathworks.com/help/images/intensity-based-automatic-image-registration.html.

Chapter 5 Deep learning based automated left ventricle segmentation and flow quantification in 4D flow cardiac MRI

This chapter was adapted from:

Xiaowu Sun, Li-Hsin Cheng, Sven Plein, Pankaj Garg, Rob J. van der Geest. **Deep learning-based automated left ventricle segmentation and flow quantification in 4D flow cardiac MRI.** Journal of Cardiovascular Magnetic Resonance.(under review)

Abstract

Background: 4D flow MRI enables assessment of cardiac function and intra-cardiac blood flow dynamics from a single acquisition. However, due to the poor contrast between the chambers and surrounding tissue, quantitative analysis relies on the segmentation derived from a registered cine MRI acquisition. This requires an additional acquisition and is prone to imperfect spatial and temporal inter-scan alignment. Therefore, in this work we developed and evaluated deep learning-based methods to segment the left ventricle from 4D flow MRI directly.

Methods: We compared five deep learning-based approaches with different network structures, data pre-processing and feature fusion methods. For the data pre-processing, the 4D flow MRI was reformatted into a stack of short-axis view slices. Two feature fusion approaches were proposed to integrate the features from magnitude and velocity images. The networks were trained and evaluated on an in-house dataset of 103 subjects with 69,619 2D images and 3090 3D volumes. The performance was evaluated using various metrics including Dice, average surface distance (ASD), end-diastolic volume (EDV), end-systolic volume (ESV), left ventricular ejection fraction (LVEF), kinetic energy (KE) and flow components. The Monte Carlo dropout method was used to assess the confidence and to describe the uncertainty area in the segmentation results.

Results: Among the five models, the model combining 2D U-Net with late fusion method operating on short-axis reformatted 4D flow volumes achieved the best results with Dice of 84.51% and ASD of 3.13 mm. The averaged absolute error between manual and automated segmentation for EDV, ESV, LVEF and normalized KE was 20.27 ml, 17.21 ml, 7.41% and 0.54 μJ/ml, respectively. Flow component results derived from automated segmentation showed high correlation and small average error compared to results derived from manual segmentation.

Conclusions: Deep learning-based methods can achieve accurate automated LV segmentation and subsequent quantification of volumetric and hemodynamic LV parameters from 4D flow MRI without requiring an additional cine MRI acquisition.

5.1 Background

Four-dimensional flow magnetic resonance imaging (4D flow MRI) provides time-resolved three-dimensional imaging of cardiac geometry and multi-directional intra-cardiac blood flow velocity from a single acquisition [1]. Several quantitative left ventricular (LV) hemodynamic parameters can be derived from the acquired data, including intra-cardiac kinetic energy (KE), vorticity and functional flow components [2, 3]. Quantitative assessment of these parameters relies on accurate segmentation of the LV cavity. However, the contrast between the blood pool and the surrounding tissue is typically extremely poor in the acquired magnitude images of a 4D flow acquisition. For this reason, the segmentation is usually performed using the images of an additionally acquired balanced Steady State Free Precession (b-SSFP) cine MR acquisition [4, 5]. Based on the known spatial relation between the two acquisitions, the obtained segmentation can be transferred to the domain of the 4D flow acquisition. Unfortunately, due to breath-hold inconsistency and differences in heart rate, the cine MR images are prone to a spatial and temporal misalignment resulting in sub-optimal segmentation of the 4D flow acquisition. Therefore, it would be advantageous when the segmentation could be performed directly from the 4D flow acquisition, not requiring any additional acquisition.

Bustamante proposed a multi-atlas registration method to automatically generate a segmentation of the entire thoracic cardiovascular system using eight 3D phase-contrast MR angiogram volumes as atlases [6]. A disadvantage of this approach is the high computational cost of the required image registration. In recent years, deep learning-based segmentation methods have been proposed and achieved immense success in medical image segmentation tasks. U-Net, consisting of a contracting and expanding path, has demonstrated excellent performance in segmentation of MR imaging data of the heart, brain and various other organs [7]. Benefiting from these convolutional neural networks (CNNs), a few studies reported the use of deep learning for the segmentation of 4D flow MRI. Berhane et al. developed a 3D U-Net with DenseNet-based dense blocks to segment the aortic arch from 4D flow MRI [8]. Based on U-Net and attention gate mechanism, Wu demonstrated that incorporating the information from the combination of magnitude and velocity images results in improved performance in LV myocardium segmentation in 4D myocardial velocity mapping MRI [9]. Corrado et al. applied a fine-tuned CNN model trained on cine b-SSFP MRI data and used registration to derive segmentation of the 4D flow MRI [10]. However, this approach relies on the availability of a cine MRI acquisition. Bustamante et al. recently reported a 3D U-Net based method for segmentation of the cardiac chambers and great thoracic vessels directly from 4D flow MRI magnitude images, ignoring the velocity images [11]. An excellent geometric agreement with manual segmentation results was reported (DICE $\geq$ 0.9) and also good agreement of the derived quantitative results, such as end-diastolic (ED) and

and-systolic (ES) volumes and blood flow kinetic energy. However, since the employed 4D flow acquisition was acquired directly after gadolinium contrast administration it remains unknown whether the presented method performs equally well on non-contrast-enhanced imaging data. A CNN-based segmentation method that takes both magnitude and velocity images as input may yield better segmentation performance than one only using magnitude images.

Our main contributions are summarized as follows: (1) We evaluated multiple strategies to take advantage of the magnitude and velocity images of the 4D flow MRI acquisition. (2) We compared the performance of five different U-Net-based networks. (3) We used a Monte Carlo dropout method to evaluate the segmentation uncertainty of the implemented CNN models.

5.2 Methods

5.2.1 Study cohort and imaging protocol

The dataset used in the study included 103 subjects, including 75 post-myocardial infarction (MI) patients (15 females, 60 males; mean age 69±12 years, range 40-94) and 28 healthy volunteers (11 females, 17 males; mean age 48±17, range 23-80). The study was approved by the local medical ethical committee of the University of Leeds, UK, and all participants provided written informed consent. All subjects underwent a comprehensive cardiac MR imaging protocol on a 1.5T MR system (Philips Healthcare), including cine MR imaging in standard cardiac views and 4D flow MR with whole-heart coverage.

A short-axis cine stack was acquired with a slice thickness of 8-10 mm and an inter-slice gap of 2 mm using 10-17 slices to cover the LV from the apex to the base. Imaging was performed during breath-holding in end-expiration. Other imaging parameters were a field of view (FOV) 300×300 mm^2 to 470×470 mm^2, pixel spacing 0.83-1.19 mm, echo time (TE) 1.27-1.62 ms, repetition time (TR) 2.55-3.25 ms. Using retrospective gating 30 phases were reconstructed to cover a full cardiac cycle. 4D flow MRI was acquired using an echo-planar imaging (EPI) accelerated sequence with retrospective electrocardiogram gating during free-breathing without using respiratory motion compensation. The 3D volume of the acquisition was planned in an oblique orientation with a voxel size of 3×3×3 mm^3 , a field of view of 370-400 × 370-400 mm^2 and 33-52 reconstructed slices to cover the whole heart. The orientation of the acquired 3D volume varied from subject to subject and was adjusted such as to encompass the complete heart and proximal aorta using a minimal number of slices. The number of reconstructed cardiac phases was 30. Other scan parameters of the 4D flow MRI acquisition were TE 1.9-3.8 ms, TR 4.8-13.9 ms, flip angle 10˚ and velocity encoding (VENC) 150 cm/s. A more detailed description of the scan parameters can be found in previous work [12]. In patients, the 4D flow

acquisition was added to a regular clinical scan protocol, including late-gadolinium enhancement (LGE) imaging. Typically, the 4D flow acquisition was obtained post-contrast (Magnevist, 0.2 mmol/kg) in the waiting period between contrast administration and LGE imaging.

5.2.2 Ground truth generation

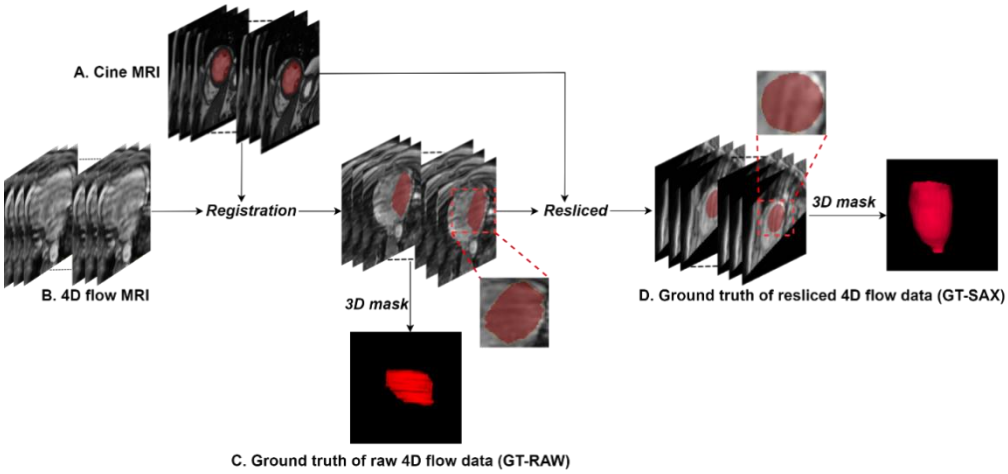


Figure.5.1. The procedure of ground truth generation. A: The mask of left ventricle was first annotated in the short-axis cine MRI. B,C: it was propagated to original 4D flow MRI using rigid registration method. D: Given the orientation of short-axis cine MRI, the raw 4D flow MRI was resliced into short-axis view.

One experienced observer semi-automatically defined the LV endocardial contours in all slices and phases of the short-axis cine stack using in-house developed Mass research software (Version V2017-EXP; Leiden University Medical Center, Leiden, the Netherlands). Following SCMR recommendations, papillary muscles and trabeculations were included within the defined contours in order to derive a consistent and time-continuous segmentation of the LV geometry. Correction for spatial misalignment, resulting from patient movement between the cine MR and 4D flow acquisition, was performed using rigid registration using Elastix software as previously described [13, 14]. Subsequently, we generated two types of LV blood pool masks for the 4D flow MRI acquisition. The first type of mask, further labelled as RAW, was generated by labelling the pixels of the original slices of the 4D flow acquisition as either blood pool or background according to the nearest labelled pixel in the short-axis cine acquisition. Due to the relatively low through-plane resolution of the short-axis stack and the varying orientation of the acquired 4D flow volumes, the resulting RAW masks frequently suffer from jagged boundaries and are less smooth compared to the original contours as defined in the short-axis stack. Therefore a second type of mask, further labelled as SAX, was generated by reformatting the volume of the 4D flow acquisition into a stack of short-axis slices.

Given the known short-axis orientation, the original 4D flow acquisition was resliced into a short-axis view using a slice spacing of 3 mm and a fixed number of 41 slices. The in-plane resolution was chosen to be equal to that of the cine short-axis stack and ranged from 0.83×0.83 mm2 to 1.19×1.19 mm2. Subsequently, the SAX mask was generated by labelling the pixels in the reformatted 4D flow images as either blood pool or background, following the same approach as for the RAW mask. The resulting blood pool regions are more smooth compared to the RAW mask regions and vary less in shape since all masks are defined in short-axis orientation. Accordingly, two ground truths are available for training and testing: GT-RAW for the original 4D flow data and GT-SAX for the resliced 4D flow data. Figure.5.1 describes the procedure of the ground truth generation, illustrating the more irregular GT-RAW masks compared to the GT-SAX masks. After excluding the images without LV, the dataset contained 90,313 SAX 2D image pairs, 69,619 RAW 2D image pairs and 3,090 (103×30) 3D volumes.

5.2.3 Networks

Table.5.1. Different methods with different networks and inputs. SAX indicates that the resliced data in the short-axis view was used as input to the network. RAW indicates that the raw 4D flow data was used as input.

Method	Input orientation	Ground truth	Network	Input Size	Output Size
SAX2D	SAX	GT-SAX	2D U-Net	(256,256,4)	(256,256,1)
RAW2D	RAW	GT-RAW	2D U-Net	(256,256,4)	(256,256,1)
SAX3D	SAX	GT-SAX	3D U-Net	(256,256,40,4)	(256,256,40,1)
RAW3D	RAW	GT-RAW	3D U-Net	(256,256,32,4)	(256,256, 32,1)
SAX2DF	SAX	GT-SAX	2D Fusion Network	(256,256,1), (256,256,3)	(256,256,1)

We compare five deep learning models to investigate the effect of data preprocessing, information fusion strategies and network structures on the segmentation performance. The five proposed methods are summarized in Table.5.1. RAW and SAX represent the two different input orientations. RAW used the original 4D flow data to train the network, either as a 3D volume, or as individual 2D slices and SAX used 4D flow data resliced into the short-axis orientation. Each 2D slice was center-cropped to a fixed size of 256×256. The number of slices in the 3D volume of the original 4D flow data varies from 33 to 52 The middle 32 slices in the original 4D flow data are stacked as the input of RAW3D. In the resliced dataset with fixed number of 41 slices, the last 40 slices are stacked as the input of SAX3D after excluding the first slice, resulting in an even number of spatial input dimension, which is convenient for the repeated down-sampling operations with a factor of 2.

SAX2D and RAW2D models are adapted from 2D U-Net, an encoder-decoder CNN model with long-skip connections. The architecture includes five-scaled resolutions. Each level contains two convolutional blocks composed of a convolution layer with kernel size of 3×3 followed by a instance normalization (IN) layer, a rectified linear unit (ReLU) and one dropout layer. In the encoder feature maps are down-sampled by a max-pooling layer with kernel size of 2×2, while in the decoder transposed convolution layers are used to increase the resolution to its original scale. The long-skip connections are used to concatenate the features from fine to coarse scales at each level. Finally, a convolution layer with kernel size of 1×1, followed by a Sigmoid function, is used to generate the probability map. The final segmentation results are determined by choosing the class with the highest probability at each pixel.

RAW3D and SAX3D models employ a 3D U-Net architecture, which is used to investigate the performance of varying volumetric inputs. The 3D volume generated from each phase is considered as an independent input of 3D U-Net. Compared to 2D U-Net, the kernel size of all convolution layers in 3D U-Net is set to 3×3×3. The 3D U-Net introduced four max-pooling layers for the down-sampling operations. The kernel size of all pooling layers in RAW3D are set to 2×2×2. Whereas in SAX3D the first three pooling layers are set to 2×2×2 and last pooling layer is set to 2×2×1 because the spatial dimension will be reduced to 5 after three down-sampling operations.

Magnitude and velocity images can be considered as different modalities providing different information for the segmentation. To fuse the information from these two modalities, we introduce two approaches named early fusion and late fusion, respectively. SAX2D, SAX3D, RAW2D and RAW3D use the early fusion method where the magnitude and velocity images are concatenated along the channel dimension as the input. Whereas SAX2DF uses the late fusion method. As illustrated in Figure.5.2, separate encoders are used to extract the features from these two modalities. Thereafter, features in the same level from two encoders are added together. The aggregated features in the bottleneck are up-sampled to the original resolution. The other multi-scale aggregated feature maps are then concatenated with the features up-sampled from the lower level. The structure of decoder used in SAX2DF is the same as that in 2D U-Net.

Dice loss and cross-entropy were jointly used as the loss function to train the models. All the experiments were implemented using Pytorch with the following parameters: batch size=50; learning rate=0.0001; optimizer=Adam. Five-fold cross-validation was applied to assess the performance and the averaged values are reported. All the experiments were implemented on a machine equipped with an NVIDIA Quadro RTX 6000 GPU with 24 GB internal memory.

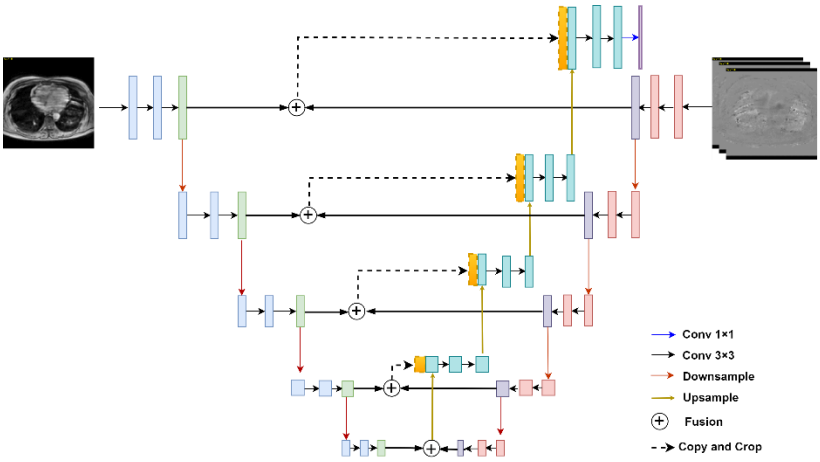


Figure.5.2. The network architecture of SAX2DF. SAX2DF separates magnitude and the three velocity images as two inputs and uses two encoders to extract the features from each input. The late fusion method us used to integrate those features.

5.2.4 Evaluation metrics

The performance of the automated methods was evaluated using segmentation accuracy, uncertainty score and volumetric and flow related clinical metrics.

Segmentation Accuracy. Dice and average surface distance (ASD) were used to assess the segmentation accuracy. Dice measures the overlap between the prediction and the ground truth. ASD is the average of all the distances from all surface points on the boundary of the predicted region to the boundary of the ground truth, which can be described as formula (5.1)

$$ASD = \frac{1}{n_S + n_{S'}} \left(\sum_{p=1}^{n_S} d(p, S') + \sum_{p'=1}^{n_{S'}} d(p', S) \right) \tag{5.1}$$

where $d(p, S') = \min_{p' \in S'} \| p - p' \|_2$ is the minimum of the Euclidean distance between a point p on surface S and the surface S'. Dice and ASD reported in this work are computed based on each 3D volume and averaged over all phases.

Clinical metrics. End-diastolic volume (EDV), end-systolic volume (ESV), LV ejection fraction (LVEF) and kinetic energy (KE) were derived as clinical metrics. The KE was computed as formula 5.2 with ρ_{blood} being the density of the blood (1.06g/cm^3), V_{voxel} the voxel volume and v the velocity magnitude of one voxel. The

total KE is the summation of the KE of each voxel within the LV region. The total KE values were indexed for LV EDV and averaged over the complete cardiac cycle.

$$KE = \frac{1}{2}\rho_{blood} \cdot V_{voxel} \cdot v^2 \tag{5.2}$$

Additionally, three phasic KE parameters were derived: peak systolic, peak E-wave and peak A-wave KE. The result of LV segmentation was also used for LV flow component analysis. Based on previously described methods by Eriksson *et al* the segmented LV blood pool at the ED moment was used to define seeding particles of size 3×3×3 mm^3 and particle pathlines were derived using particle tracing in forward (until the next ES moment) and backward (until the previous ES moment) direction [4]. The particle position at the two ES moments was then used to classify the defined pathlines as either direct flow (DIR), delayed ejection flow (DEL), retained inflow (RET) or residual volume (RES). The relative size of each flow component was expressed as a percentage of the ED volume. The clinical parameters derived from automated LV segmentation were compared to the results derived of manual segmentation.

Uncertainty Score. Segmentation of anatomical structures is inherently ambiguous especially near an object border which is not clearly defined due to the poor contrast or restriction imposed by the image acquisition. The uncertainty score can give some insights into the confidence of a model in its predicted segmentation results [15]. In case of a high uncertainty score, it is more likely that the segmentation result is inaccurate. Usually, a CNN model only produces a single segmentation map without any information to explain its confidence in its prediction. A high probability value in a segmentation map doesn't imply a high confidence score. A model also can be uncertain in pixels with high probability. In order to investigate the segmentation uncertainty of the different models we applied the Monte Carlo (MC) dropout method [16] to quantify the model's confidence in the segmentation result.

Generally during the testing phase, the dropout layers in the network are removed. The uncertainty score can be derived by preserving the dropout layers during testing while executing multiple inference runs. In our experiments the drop rate in the middle-level dropout layers was set to 0.5 and the testing was repeated 20 times resulting in 20 predictions denoted as $P_i(i = 1, 2, \cdots 20)$. The uncertainty score can be derived using equation 5.3 where $P = \frac{1}{20}\sum_{i=1}^{20} P_i$.

$$UQ = -P \times \log_2 P - (1-P) \times \log_2 (1-P) \tag{5.3}$$

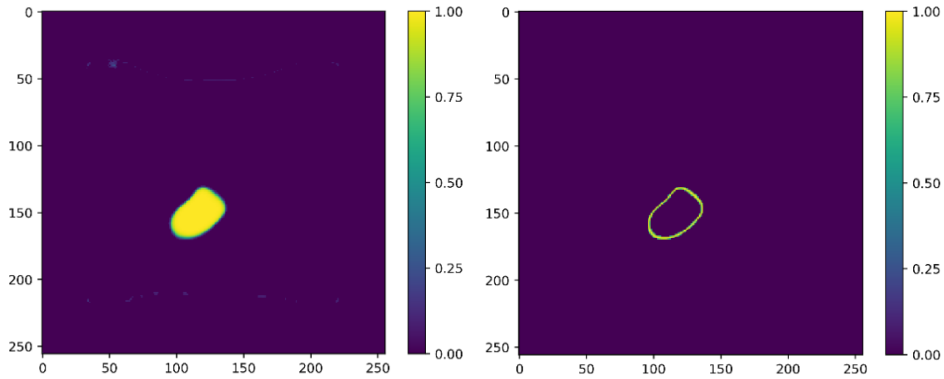


Figure.5.3. An example of segmentation probability and its corresponding uncertainty map. **Left**: Probability map derived from the last layer of RAW2D. **Right**: Corresponding uncertainty map derived from MC method.

Figure.5.3 shows an example of a segmentation probability map and its corresponding uncertainty map. The uncertainty score for the pixels within the LV chamber is low, implying a high confidence of the models' prediction, but due to the poor contrast between the heart chamber and myocardium the uncertainty near the ambiguous LV border with a corresponding probability varying from 0.4 to 0.6 is substantially higher. To compute the mean of uncertainty and to quantify the segmentation quality, we first computed the uncertainty score for the whole LV chamber where each pixel's prediction probability is larger than 0.5. Then to highlight the higher uncertainty in the boundary region, we further computed the score for this area with a prediction probability ranging from 0.4 to 0.6.

Statistical analysis. The correlation of the clinical metrics derived from the manual and predicted segmentation results were assessed using the Pearson correlation coefficient (PCC). Additionally, bias and limits of agreement (LOA, 1.96×standard deviation) were used to describe the agreement of prediction and ground truth.

5.3 Results

First, we compared the results derived by the five models on various evaluation metrics. Second, we explored the impact of the fusion methods on the uncertainty score. Lastly, we investigated the performance of the best model on the KE and flow components.

5.3.1 Segmentation results

Table.5.2 summarizes the segmentation performance derived from different models. SAX2DF achieved the best results with Dice of 84.51%, ASD of 3.13 mm and absolute error of ESV and LVEF of 17.21 ml, 7.41%, respectively. The best results in an absolute error of EDV and KE were obtained using the model of SAX2D, yielding an error of 19.96 ml and 0.41 μJ/ml, respectively.

Table.5.2. Segmentation performance derived from different methods. The reported clinical results are the error between the ground truth and prediction. The best results are shown in bold. KE was normalized to the EDV. SAX2D-M and SAX3D-M represent models that only use the magnitude images as the input. The other methods use both magnitude and velocity images as the input

Method	Dice(%)	ASD(mm)	EDV(ml)	ESV(ml)	LVEF(%)	KE(μJ/ml)
SAX2D	84.33±6.28	3.30±1.89	**19.96±22.08**	19.35±16.04	9.54±7.17	**0.41±0.39**
RAW2D	79.97±7.38	4.01±1.86	29.32±22.91	32.35±24.33	10.84±8.58	1.16±1.18
SAX3D	82.84±6.65	4.41±4.68	22.47±19.19	23.21±22.12	10.07±8.89	0.84±1.02
RAW3D	79.77±7.56	3.67±1.64	24.25±19.56	26.91±24.04	10.58±8.01	0.91±0.85
SAX2DF	**84.51±6.58**	**3.13±2.33**	20.27±20.31	**17.21±16.03**	**7.41±6.07**	0.54±0.51
SAX2D-M	81.46±7.39	4.10±2.91	26.20±23.93	32.65±22.04	21.43±11.03	0.88±2.17
SAX3D-M	80.61±0.09	5.48±5.23	24.53±18.75	38.47±31.27	18.97±16.48	1.01±0.93

Due to the different ground truth masks used, a direct comparison of the performance using Dice and ASD derived from RAW and SAX data is not easily possible. Therefore, also clinical parameters were used to compare the performance of the models. Table.5.2 shows that SAX2D outperformed RAW2D and SAX3D performed better than RAW3D in all clinical metrics including EDV, ESV, LVEF and KE, demonstrating the models using images in short-axis view orientation can generate a better prediction.

We further compared the results derived from only using magnitude images (SAX2D-M and SAX3D-M) and combining magnitude and velocity images. The comparison was restricted to the models using the short-axis view data, since these models provided the best performance. Table.5.2 reveals that the Dice derived from models using the combination of magnitude and velocity data is 3% higher compared to the models using the magnitude images only. Adding velocity images as input to the model is clearly shown to be beneficial. The variation in Dice and ASD over the cardiac phase for each model is illustrated in Figure.5.4. All models achieved the best Dice and ASD in phases 1, 2 and 30 which is around the ED phase. The lowest performance is observed in the phases varying between phase 11-13, which is around the ES phase. These results demonstrate that LV segmentation from 4D flow data is more accurate in the ED phase than in the ES phase.

The PCC, bias and LOA of clinical evaluation metrics comparing manual with automatic segmentation results are reported in Table 5 3. Figure.5.5 and Figure.5.6 show the scatter plots, including PCC and Bland-Altman plots of four clinical metrics. SAX2DF achieved the highest correlation of 90.34%, 92.09% and 75.83% for EDV, ESV and LVEF, respectively. The best PCC for KE was achieved using SAX2D method. Although the PCC in LVEF derived from all five methods are lower than 80%, the results in the other three metrics demonstrate a good linear correlation with the results derived from manual segmentation. Notably, all five models achieved a PCC for KE higher than 90%. Although there is a significant variation in the performance of EDV and ESV estimation derived from the different methods, the biases for those two metrics derived from SAX2D, RAW3D and SAX2DF are smaller than 10 ml. The smallest biases in EDV and ESV are 2.03 ml and 3.35 ml derived from SAX3D and SAX2DF, respectively. RAW2D achieved the worst performance, with a bias of 19.19 ml and 20.52 ml in EDV and ESV, respectively. RAW3D and SAX2DF achieved the smallest bias in LVEF and normalized KE with 3.09% and 0.02 μJ/ml, respectively.

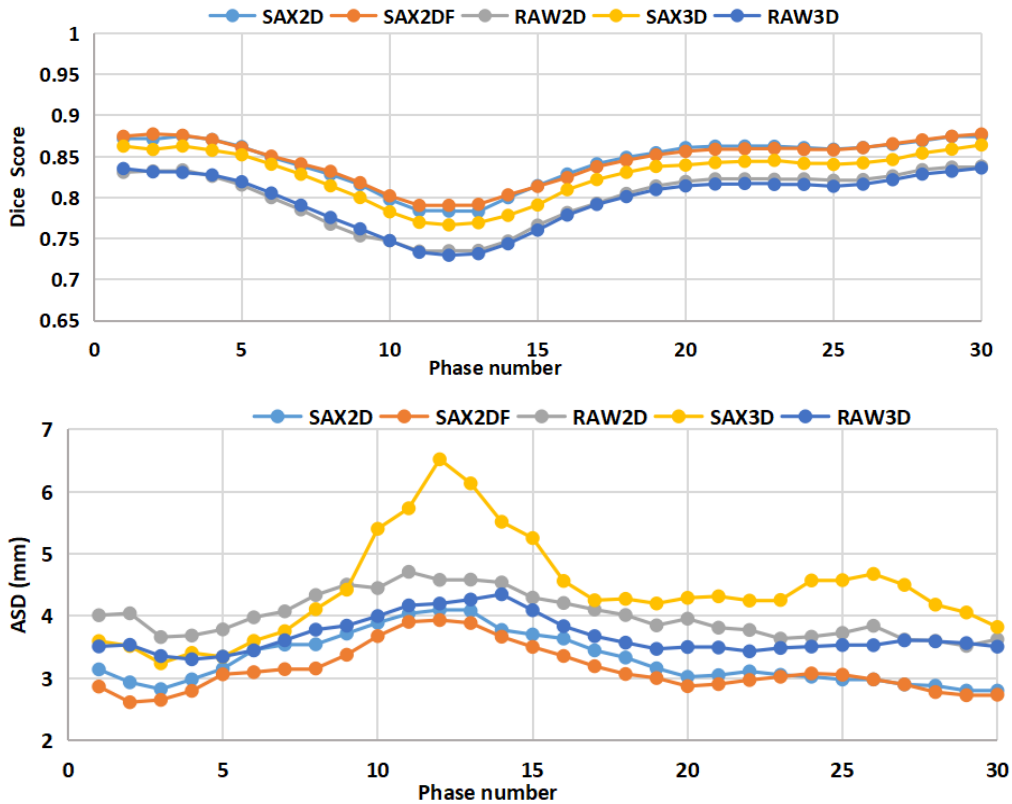


Figure.5.4. Average Dice and ASD results plotted over time (averaged over all subjects). The x-axis is the phase number, y-axis is the averaged Dice (upper) and ASD (bottom) derived from different models.

Table 5-3. PCC and Bias of clinical metrics from the prediction against the reference. The best results are shown in bold.

		SAX2D	RAW2D	SAX3D	RAW3D	SAX2DF
PCC	EDV(%)	89.44	86.51	88.69	87.97	**90.34**
	ESV(%)	91.47	80.61	85.51	79.48	**92.09**
	LVEF(%)	70.29	50.08	50.12	47.77	**75.83**
	KE(%)	**98.63**	92.58	92.82	93.51	94.32
Bias ± LOA	EDV (ml)	7.24±56.71	19.19±62.64	**-2.03±57.94**	-7.24±59.56	7.96±54.15
	ESV (ml)	8.31±46.62	-20.52±68.58	-11.39±58.88	-7.24±69.46	**-3.35±45.75**
	LVEF (%)	7.73±17.86	6.55±23.92	6.89±23.92	**3.09±25.37**	5.11±15.92
	KE(μJ/ml)	0.23±1.02	0.75±2.50	0.44±2.44	0.28±2.37	**0.02±1.47**

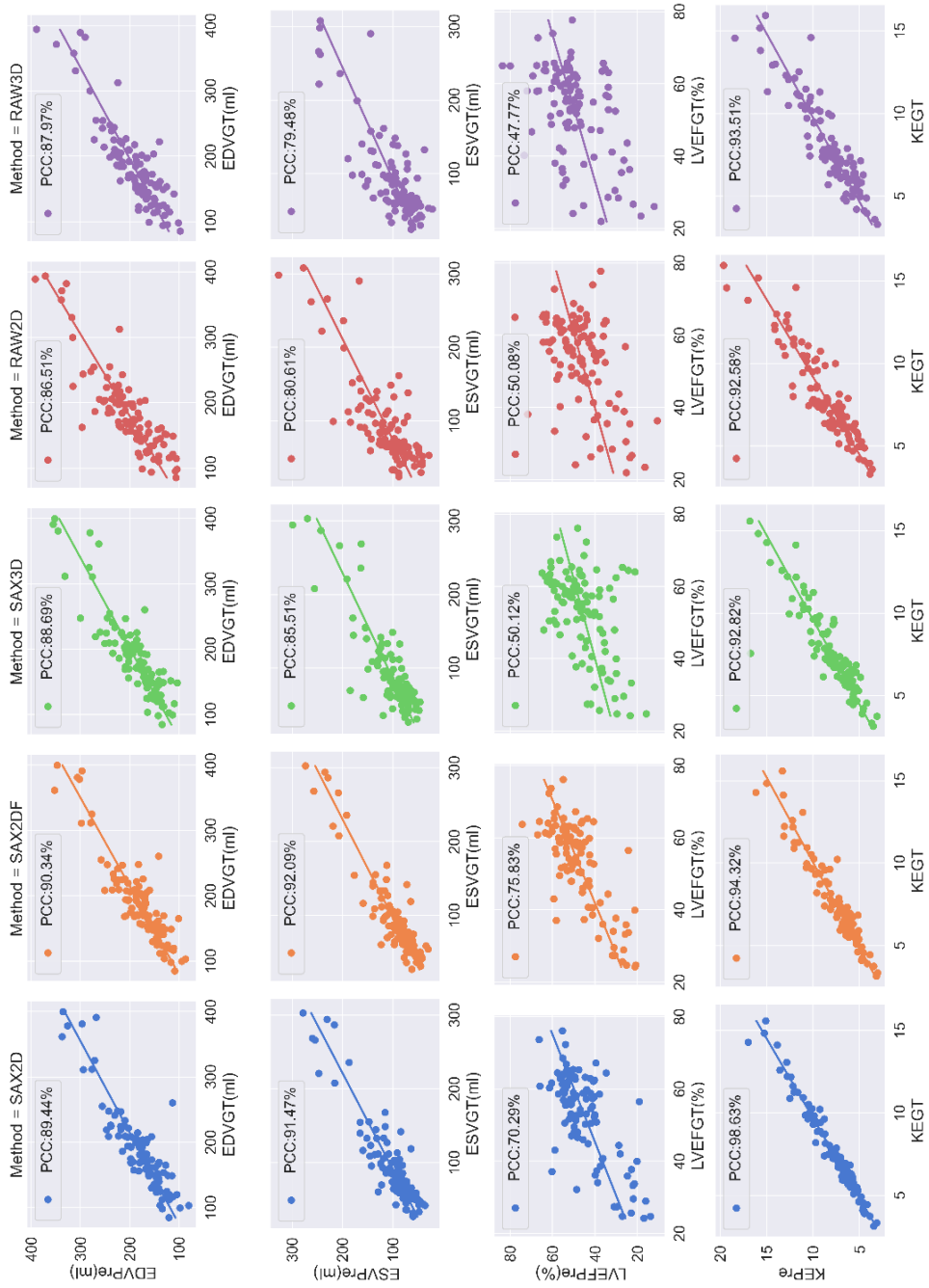


Figure.5.5. Correlation of clinical metrics derived from manual and automated segmentation. Each column represents one CNN model. The four rows denote four clinical metrics including EDV, ESV, LVEF and KE. For each plot, the x-axis is the measure derived from the manual segmentation and y-axis represents the results derived from automated prediction.

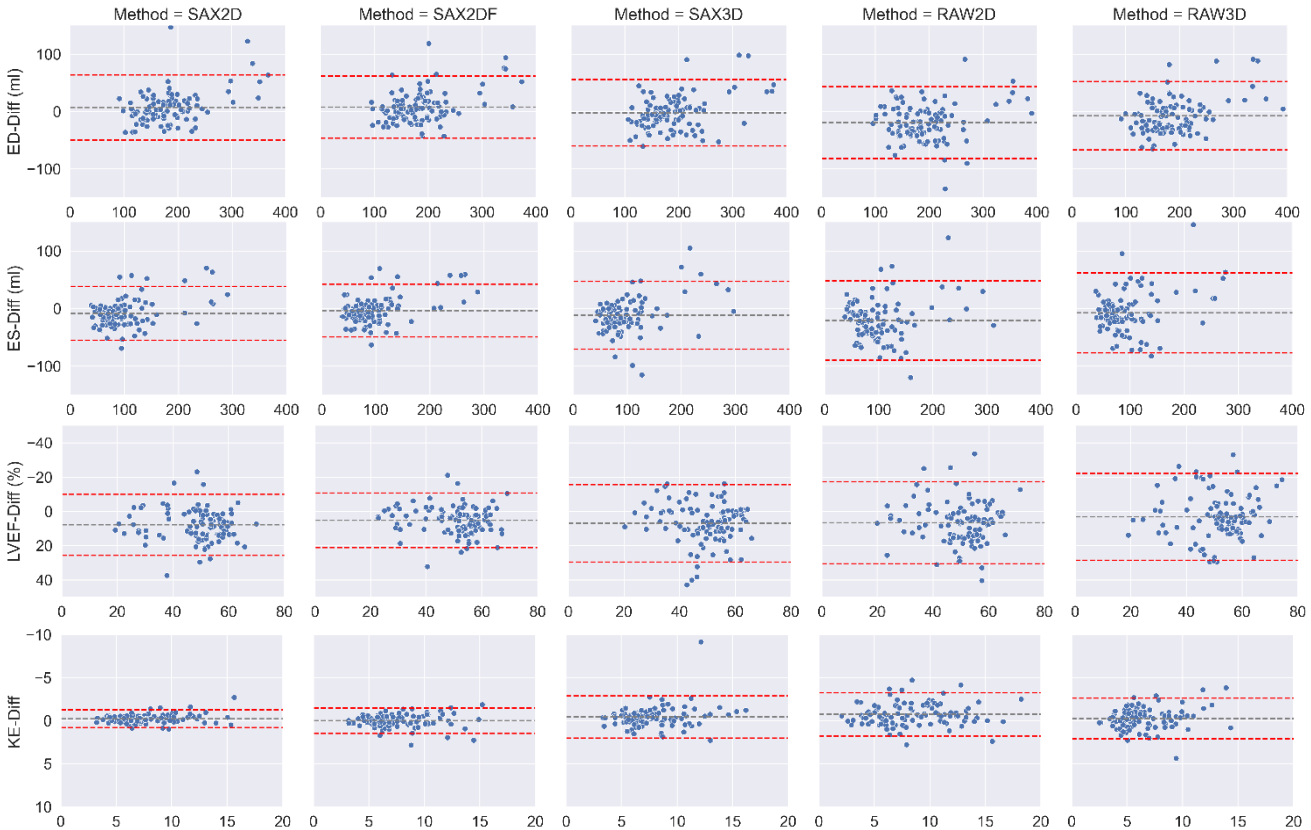


Figure.5.6. Bland-Altman plots of clinical metrics comparing automated and manual segmentation. The columns represent five models and the rows show the results of EDV, ESV, LVEF and KE, respectively. In each Bland-Altman plot, the x-axis denotes the average of two measurements and the y-axis represents the difference between two measurements. The black line represents the bias and the two red lines denote the LOA.

Examples of 2D and 3D segmentation masks derived from the five models are shown in Figure.5.7.

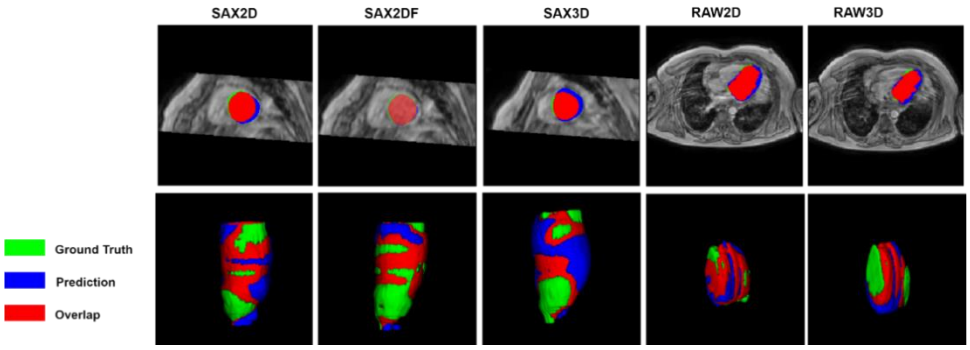


Figure.5.7. Examples of automated LV segmentation results in 2D and 3D. The first two rows are the results of 2D and 3D segmentation results. Green color represents the ground truth, blue color is the prediction, and red parts are the overlap between the prediction and ground truth.

5.3.2 Uncertainty results

Table.5.4 reports the averaged uncertainty scores both in the LV blood pool and the defined boundary area over 3090 phases (30 phases per subject, 103 subjects in total) derived from the five proposed models. SAX2DF achieved the lowest uncertainty scores with 0.12 and 0.75 in the whole LV and boundary area. SAX3D has lower uncertainty than SAX2D (0.13 vs. 0.15, 0.76 vs. 0.83). Similarly, RAW3D has a lower uncertainty than RAW2D (0.13 vs. 0.20, 0.77 vs. 0.82). The 3D models are shown to be more confident in its predictions than the 2D models. When comparing SAX2D and SAX2DF, it can be concluded that the late fusion method resulted in a lower uncertainty score.

Table.5.4. The averaged uncertainty value derived from different defined areas. The LV chamber refers to the area with a probability larger than 0.5. Boundary area refers to the area with probability ranging from 0.4 to 0.6.

Area	SAX2D	SAX3D	RAW2D	RAW3D	SAX2DF
LV chamber	0.15±0.32	0.13±0.36	0.20±0.88	0.13±0.41	**0.12±0.44**
Boundary area	0.83±0.29	0.76±0.54	0.82±0.66	0.77±0.28	**0.75±0.17**

5.3.3 Flow quantitative analysis

SAX2DF is the best segmentation model among the proposed five models, according to the performance on segmentation accuracy, clinical metrics and uncertainty score. Therefore, we further investigated the performance of SAX2DF in quantifying KE and flow components. The low averaged error of indexed KE ranging from -0.03 mJ to 0.04 mJ and flow components varying from -4.58% to 3%, as reported in Table.5.5, shows a good agreement between the prediction and ground

truth. A detailed summary of the PCC of KE and flow components derived from the automatic and manual methods is illustrated in Figure.5.8.

Table.5.5. The averaged difference and PCC between the automatic and manual methods in flow components quantification and KE. RES: Residual volume; RET: retained inflow; DEL: delayed ejection flow; DIR: direct flow. KE parameters were indexed to LV end-diastolic volume.

	flow components (%)				KE (mJ)			
	RES	RET	DEL	DIR	Max RR	Max Systole	Max E-wave	Max A-wave
Mean ± SD	3.00±5.96	0.78±25.9	0.80±2.68	-4.58±4.34	-0.03±0.38	0.04±0.52	0.02±0.45	0.02±0.3
PCC (%)	86.7	91.1	89.1	94.2	97.3	88.2	96.8	95.5

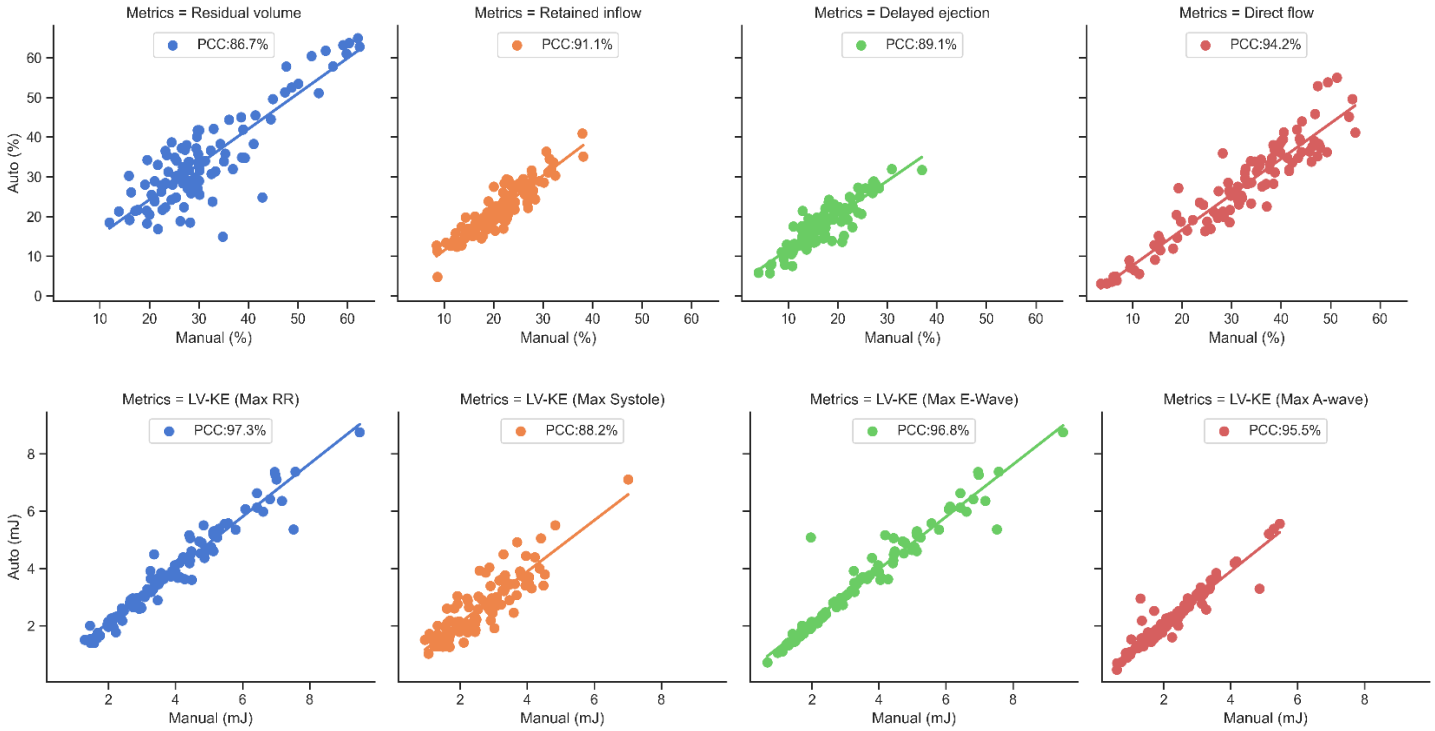


Figure.5.8. Correlation of left ventricle kinetic energy and four flow components derived from the SAX2DF and manual method. **First row:** flow components including residual volume, retained inflow, delayed ejection and direct flow. **Second row:** kinetic energy including max RR, systole, E-wave and A-wave KE

Figure.5.9 visualizes the result of LV flow component analysis derived from manual and CNN based segmentation in five subjects. The results demonstrate a good agreement between those two segmentation methods. More flow components visualization videos and segmentation result videos can be found in https://github.com/xsunn/4DflowLVSegmentation.

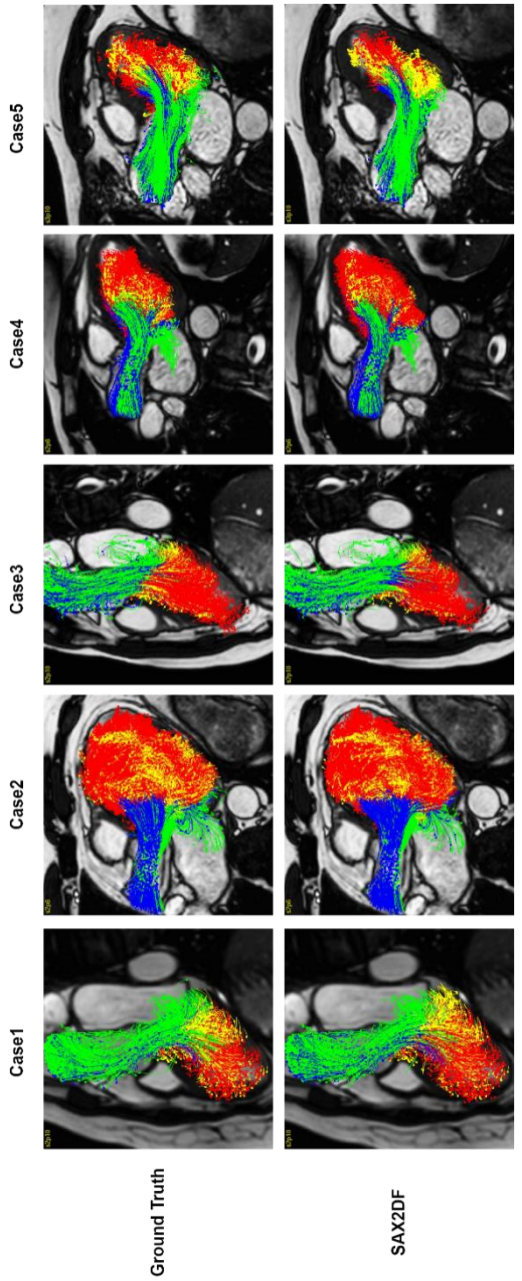


Figure.5.9. Visualization of the different ventricular flow components by track particle derived from the manual segmentation and the method of SAX2DF. Green: direct flow. Yellow: retained inflow. Blue: delayer ejection flow. Red: residual flow.

5.4 Discussion

In this work, we developed and evaluated CNN-based methods for automatic segmentation and LV flow assessment from 4D flow cardiac MRI. The main findings of our study were (1) CNN models showed good performance in LV segmentation with an average Dice of 84.5% across 103 subjects with 90,313 resliced 2D image pairs; (2) Data preprocessing has an impact on the segmentation results; (3) Combining the features from magnitude and velocity images together can benefit the segmentation performance in 4D flow MRI; (4) High correlation and low bias of EDV, ESV, KE and flow components analysis demonstrate CNN-based segmentation can provide reliable quantification of LV flow in 4D flow data.

Segmentation in 4D flow cardiac MRI is challenging due to the poor contrast between the heart chamber and its surrounding tissue. Few approaches have been proposed to overcome this challenge. Atlas-based methods [4] and registration-based methods [10] are two prevailing traditional approaches. The atlas-based method relies on image registration to generate accurate transformation between a labelled atlas and the images. Registration-based segmentation methods rely on the registration between labelled cine MRI data and 4D flow data. Both of these methods require additional data and high computational costs due to the registration. Bustamente *et al.* [11] employed a 3D U-Net architecture for LV segmentation, but in their proposed method, only the magnitude images were used as input and information from velocity images were ignored. In this work, we compared five models named SAX2D, SAX3D, RAW2D, RAW3D and SAX2DF to segment the LV from 4D flow MRI without any additional cine MRI and we also investigated the impact of different data pre-processing approaches, feature fusion methods and model structure on the segmentation results.

The performance derived from our proposed method is not as good as that of Bustamente's. The data cohort used in their work is much larger than ours; in our work, 2472 3D volumes are employed for training, which is significantly smaller than Bustamente's 5760 3D volumes. Meanwhile, our results are averaged over 3090 3D volumes using five-fold cross-validation, whereas their results were directly derived from 1640 3D volumes. Furthermore, simply using the magnitude images as input allows them to introduce various data augmentation techniques to enlarge the training data. However, the conventional data augmentation methods such as rotation, Gaussian noise and transformation cannot directly work on our proposed approach because the velocity images are more complicated than the magnitude images. Therefore, compared to their work, we trained the model with fewer data but evaluated the performance on a larger data set. Since in Bustamente's work all 4D flow acquisitions were obtained post contrast injection in both patients and volunteers and navigator gating breathing motion was applied, it is expected that the image quality of the obtained magnitude images was higher in that study.

For data preparation, given the known orientation, the original 4D flow MRI acquisition volume was resliced into short-axis slices. The raw data and resliced short-axis data served as two independent training data sets to train the networks. Improved segmentation results were derived when using the resliced short-axis data as the training data, demonstrating resliced short-axis data provided more accurate information for the segmentation, which could be explained by the more various shapes and ambiguous borders in the raw data when compared to the more consistent convex left ventricular shape in the short-axis view.

Considering magnitude and velocity images as two different modalities in 4D flow MRI, we proposed two approaches named early fusion and late fusion to fuse the information from these modalities. SAX2D, SAX3D, RAW2D and RAW3D employed early fusion by concatenating two modalities along the channel dimension as the input. While for the late fusion, SAX2DF employed two encoders to extract features from two modalities and then concatenated the features along the spatial dimension. A modestly improved performance was observed in SAX2DF when compared to the other methods, revealing that late fusion works better. We also compared the segmentation performance between 2D and 3D U-Net based methods. The results show that compared to SAX3D, SAX2D achieved better performance in all evaluation metrics. Constrained to the input spatial dimension, in SAX3D the kernel size of the final pooling layer was set to 2×2×1, resulting in the spatial features not being extracted completely. Moreover, a total of 3420 resliced 3D samples (104 subjects, 30 phases in each subject) were used to train and test the 3D U-Net, which is much less than 91,182 2D samples. As a result, the smaller training data size may be the primary reason why SAX3D did not outperform the SAX2D model.

CNN models produce a pixel-level prediction without any knowledge about the confidence of the model in its predictions. In this work, we introduced the Monte Carlo dropout method to estimate the uncertainty of the model in its segmentation results. The uncertainty score assesses segmentation reliability and offers the quantification of error to increase trust into CNN models. The results showed that the most uncertain area in the prediction is near the LV endocardial boundary, which can be explained by the poor contrast in the magnitude images and also because of the low blood flow velocity near the LV wall. Segmenting the myocardium in addition to the LV blood pool may reduce the uncertainty but cannot eliminate the uncertainty. When analyzing the uncertainty scores derived from different models, it reveals that the 3D models (SAX3D and RAW3D) performed better than the 2D models (SAX2D and RAW2D). Because 3D models are able to extract more spatial information from the input than the 2D models. It can be observed that although SAX2DF is a 2D model, benefiting from the late fusion method, SAX2DF achieved the lowest uncertainty score among all five models. A further evaluation of the

results derived from the best model, SAX2DF, was performed by comparing the KE and flow components. The results shows a good agreement between the ground truth and prediction.

There are several limitations in our work. The major limitation is the lack of generalization of the proposed models. The data used in this study was acquired from one vendor and one center. Meanwhile, there is no publicly available 4D flow MRI dataset currently. Therefore the model might not generalize well to the other datasets from different vendors or centers. As Bai [17] pointed out, a CNN model can perform well in other datasets using fine-tuning or transfer learning. Additionally, exploiting advanced data augmentation methods utilizing domain knowledge is also crucial for model generalization and robustness [18]. However, due to the complicated structure of velocity images, commonly used data augmentation methods are not suitable for 4D flow data. A novel efficient late fusion based feature fusion method also needs to be investigated.

5.5 Conclusions

In conclusion, we developed multiple deep learning-based 4D flow MRI LV segmentation models that do not require additional cine MRI. The proposed CNN models were evaluated on a large in-house dataset, achieving good performance on several metrics. The results demonstrate that a model employing late fusion and trained on resliced short-axis view data generates the best performance for left ventricular segmentation in 4D flow MRI.

References

1. Stankovic Z, Allen BD, Garcia J, Jarvis KB, Markl M. 4D flow imaging with MRI. Cardiovascular diagnosis and therapy. 2014;4(2):173.
2. Rizk J. 4D flow MRI applications in congenital heart disease. European Radiology. 2021;31:1160-74.
3. Gupta AN, Avery R, Soulat G, Allen BD, Collins JD, Choudhury L, Bonow RO, Carr J, Markl M, Elbaz MS. Direct mitral regurgitation quantification in hypertrophic cardiomyopathy using 4D flow CMR jet tracking: evaluation in comparison to conventional CMR. Journal of Cardiovascular Magnetic Resonance. 2021;23(1):1-3.
4. Eriksson J, Carlhäll CJ, Dyverfeldt P, Engvall J, Bolger AF, Ebbers T. Semi-automatic quantification of 4D left ventricular blood flow. Journal of Cardiovascular Magnetic Resonance. 2010;12:1-0.
5. Kanski M, Arvidsson PM, Töger J, Borgquist R, Heiberg E, Carlsson M, Arheden H. Left ventricular fluid kinetic energy time curves in heart failure from cardiovascular magnetic resonance 4D flow data. Journal of Cardiovascular Magnetic Resonance. 2015;17:1-0.
6. Bustamante M, Gupta V, Forsberg D, Carlhäll CJ, Engvall J, Ebbers T. Automated multi-atlas segmentation of cardiac 4D flow MRI. Medical image analysis. 2018;49:128-40.
7. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. InMedical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18 2015 (pp. 234-241). Springer International Publishing.
8. Berhane H, Scott M, Elbaz M, Jarvis K, McCarthy P, Carr J, Malaisrie C, Avery R, Barker AJ, Robinson JD, Rigsby CK. Fully automated 3D aortic segmentation of 4D flow MRI for hemodynamic analysis using deep learning. Magnetic resonance in medicine. 2020;84(4):2204-18.
9. Wu Y, Hatipoglu S, Alonso-Álvarez D, Gatehouse P, Firmin D, Keegan J, Yang G. Automated multi-channel segmentation for the 4D myocardial velocity mapping cardiac MR. InMedical Imaging 2021: Computer-Aided Diagnosis 2021 (Vol. 11597, pp. 169-175). SPIE..
10. Corrado PA, Wentland AL, Starekova J, Dhyani A, Goss KN, Wieben O. Fully automated intracardiac 4D flow MRI post-processing using deep learning for biventricular segmentation. European Radiology. 2022 Aug;32(8):5669-78.
11. Bustamante M, Viola F, Engvall J, Carlhäll CJ, Ebbers T. Automatic Time-Resolved Cardiovascular Segmentation of 4D Flow MRI Using Deep Learning. Journal of Magnetic Resonance Imaging. 2023;57(1):191-203.

12. Garg P, Westenberg JJ, van den Boogaard PJ, Swoboda PP, Aziz R, Foley JR, Fent GJ, Tyl FG, Coratella L, ElBaz MS, Van Der Geest RJ. Comparison of fast acquisition strategies in whole-heart four-dimensional flow cardiac MR: Two-center, 1.5 Tesla, phantom and in vivo validation study. Journal of Magnetic Resonance Imaging. 2018;47(1):272-81.
13. Elbaz MS, van der Geest RJ, Calkoen EE, de Roos A, Lelieveldt BP, Roest AA, Westenberg JJ. Assessment of viscous energy loss and the association with three-dimensional vortex ring formation in left ventricular inflow: In vivo evaluation using four-dimensional flow MRI. Magnetic resonance in medicine. 2017;77(2):794-805.
14. Klein S, Staring M, Murphy K, Viergever MA, Pluim JP. Elastix: a toolbox for intensity-based medical image registration. IEEE transactions on medical imaging. 2009;29(1):196-205.
15. Baumgartner CF, Tezcan KC, Chaitanya K, Hötker AM, Muehlematter UJ, Schawkat K, Becker AS, Donati O, Konukoglu E. Phiseg: Capturing uncertainty in medical image segmentation. InMedical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22 2019 (pp. 119-127). Springer International Publishing.
16. Gal Y, Ghahramani Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. Ininternational conference on machine learning 2016 (pp. 1050-1059). PMLR.
17. Bai W, Sinclair M, Tarroni G, Oktay O, Rajchl M, Vaillant G, Lee AM, Aung N, Lukaschuk E, Sanghvi MM, Zemrak F. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. Journal of Cardiovascular Magnetic Resonance. 2018;20(1):1-2.
18. Chen C, Bai W, Davies RH, Bhuva AN, Manisty CH, Augusto JB, Moon JC, Aung N, Lee AM, Sanghvi MM, Fung K. Improving the generalizability of convolutional neural network-based segmentation on CMR images. Frontiers in cardiovascular medicine. 2020;7:105.

Availability of data and materials

The datasets used in this study will not be made publicly available. However, the code for the network is available:https://github.com/xsunn/4DflowLVSegmentation.

Competing interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

XS designed and implemented the method, performed data analysis and wrote the manuscript. RG designed this study, prepared the dataset and revised the manuscript. LC designed the network and revised the manuscript. SP and PG provided support on the clinical aspects and they also provided the data used in the study. All authors read and approved the manuscript.

Acknowledgements

XS is supported by the China Scholarship Council No. 201808110201. LC is supported by the RISE-WELL project under H2020 Marie Skłodowska-Curie Actions. Prof. Sven Plein from the University of Leeds is acknowledged for granting access to the image data used in this work.

Chapter 6 Transformer based feature fusion for left ventricle segmentation in 4D flow MRI

This chapter was adapted from:

Xiaowu Sun, Li-Hsin Cheng, Sven Plein, Pankaj Garg, Rob J. van der Geest. **Transformer based feature fusion for left ventricle segmentation in 4D flow MRI.** International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, Cham, 2022.

Abstract

Four-dimensional flow magnetic resonance imaging (4D Flow MRI) enables visualization of intra-cardiac blood flow and quantification of cardiac function using time-resolved three directional velocity data. Segmentation of cardiac 4D flow data is a big challenge due to the extremely poor contrast between the blood pool and myocardium. The magnitude and velocity images from a 4D flow acquisition provide complementary information, but how to extract and fuse these features efficiently is unknown. Automated cardiac segmentation methods from 4D flow MRI have not been fully investigated yet. In this paper, we take the velocity and magnitude image as the inputs of two branches separately, then pro-pose a Transformer based cross- and self-fusion layer to explore the inter-relationship from two modalities and model the intra-relationship in the same modality. A large in-house dataset of 104 subjects (91,182 2D images) was used to train and evaluate our model using several metrics including the Dice, Average Surface Distance (ASD), end-diastolic volume (EDV), end-systolic volume (ESV), Left Ventricle Ejection Fraction (LVEF) and Kinetic Energy (KE). Our method achieved a mean Dice of 86.52%, and ASD of 2.51 mm. Evaluation on the clinical parameters demonstrated competitive results, yielding a Pearson correlation coefficient of 83.26%, 97.4%, 96.97% and 98.92% for LVEF, EDV, ESV and KE respectively. Code is available at github.com/xsunn/4DFlowLVSeg.

6.1 Introduction

Quantitative assessment of left ventricular (LV) function from magnetic resonance imaging (MRI) is typically based on the use of short-axis multi-slice cine MRI due to its excellent image quality [1,2]. Recently, four-dimensional (4D) Flow MRI has been introduced, encoding blood flow velocity in all three spatial directions and time dimension. 4D Flow MRI can be used for detailed analysis of intra-cardiac blood flow hemodynamics, providing additional information over conventional cine MRI. The segmentation of the cardiac cavities is an important step to derive quantitative blood flow results, such as the total LV kinetic energy (KE) [3]. 4D Flow MRI generates four image volumes including a magnitude image and three velocity images, one for each spatial dimension. Figure.6.1 shows an example of magnitude and velocity images from one slice out of a 4D Flow MRI data set. The example highlights the extremely poor contrast between the heart chambers and the myocardium in the 4D Flow data. Therefore, most authors have used segmentations derived from co-registered short-axis cine MR in order to quantify ventricular blood flow parameters from the 4D Flow data. However, this relies on accurate spatial and temporal registration of the two MR sequences. Inconsistent breath-hold positioning may introduce spatial misalignment while heart rate differences will result in temporal mismatch between the acquisitions. The aim of the current work was therefore to develop an automated method for LV segmentation from 4D Flow MRI data, not requiring additional cine MRI data.

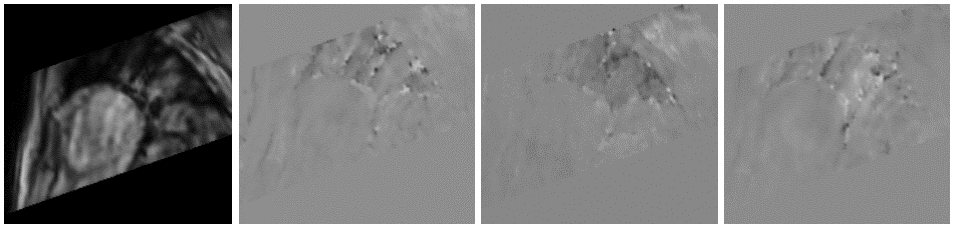


Figure.6.1. A sample of cardiac 4D Flow data in short-axis view. The first image is the magnitude image, and the last three images are the velocities in x, y and z dimensions respectively.

Since U-Net [4] was proposed, convolutional neural networks (CNNs) have been predominant in the task of medical image segmentation. Many variants of U-Net have been proposed further improving the performance. For instance, nnU-Net [5] introducing automated self-configuring outperformed most existing approaches on 23 diverse public datasets. Although those CNN based networks have achieved an excellent performance, restricted by the locality of convolutional kernels, they cannot capture long-distance relations [6,7].

Transformer is considered as an alternative model using its self-attention mechanism to overcome the limitation of CNN. Transformer was designed firstly for

natural language processing (NLP) tasks such as machine translation and document classification. More recently, Transformer-based approaches were introduced in medical image processing. TransUnet [8] applied a CNN-Transformer hybrid encoder and pure CNN decoder for segmentation. However, TransUnet still uses convolutional layers as the main building blocks. Inspired by the Swin Transformer [9], Cao proposed a U-Net-like pure Transformer based segmentation model which uses hierarchical Swin Transformer as the encoder and a symmetric Swin Transformer with patch expanding layer as the decoder [10]. Other Transformer-based networks [11,12,13] also mark the success of Transformer in medical image segmentation and reconstruction.

Although numerous deep learning-based segmentation methods have been proposed in various modalities, the automatic segmentation of the LV directly from 4D Flow data has not been explored yet. A specific challenge is that the magnitude and velocity images of a 4D Flow acquisition have different information content and should be considered as different modalities. Moreover due to velocity noise, a careful fusion method is needed to avoid redundancy or insufficient feature integration [14,15].

In this paper, we present, to the best of our knowledge, the first study to segment the LV directly from 4D Flow MRI data. Our main contributions are: (1) we propose two self- and cross-attention-based methods to fuse the information from different modalities in 4D Flow data; (2) we evaluate our method in a large 4D Flow dataset using multiple segmentation and clinical evaluation metrics.

6.2 Method

6.2.1 Attention mechanism

Attention mechanism, mapping the queries and a set of keys-value pairs to an output, is the fundamental component in Transformer. In this section, we first introduce how the self-attention module models the intra-relationship of features from the same image modality. Then we explain how cross-attention explores the inter-relationship of features from two different modalities. The two attention modules are illustrated in Figure.6.2.

Self-Attention module. In self-attention module [6], the **Q** (queries), **K** (keys) and **V** (values) are generated from the same modality. **Q** and **K** determine a weight matrix after the scaled dot product which is used to compute the weighted sum of **V** as the output. The computing process can be described as in equation (6.1):

$$Atten(Q_a, K_a, V_a) = softmax(\frac{Q_a K_a^{\ T}}{\sqrt{d}})V_a \tag{6.1}$$

where *d* is the key dimensionality, and *a* denotes modality *a*.

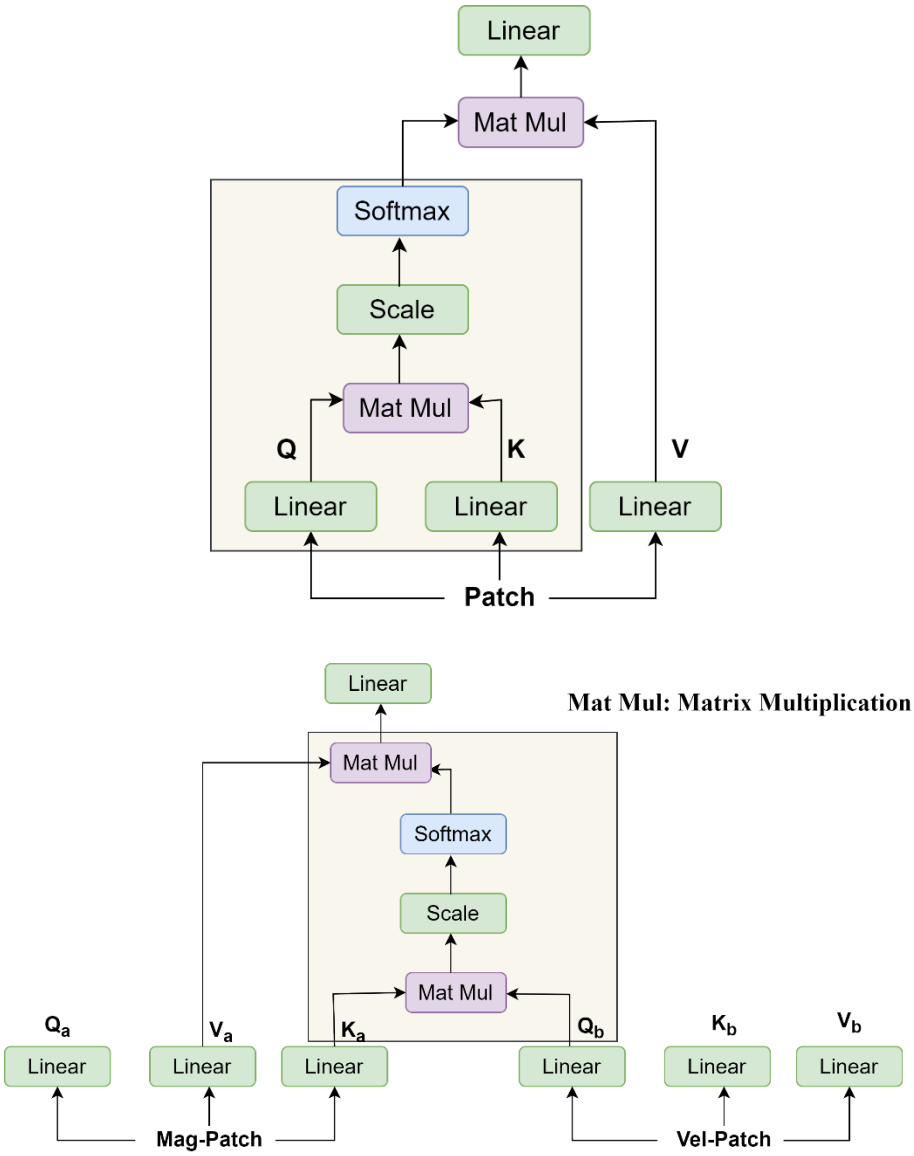


Figure.6.2. The structure of self-attention (upper) and cross-attention (bottom) modules.

Cross-Attention module. Although self-attention explores the intra-modality relationship, the inter-modality relationship, such as the relationship between pixels in the magnitude image and velocity image is not explored. The cross-attention module takes two patches as the input to generate the **Q**, **K**, and **V**. **V** and **K** are generated from the same modality, while **Q** is derived from another modality. The

other operations are kept the same as in self-attention. It can be expressed as equation (6.2). Hence, cross-attention can be adopted to fuse the information from different modalities.

$$Atten(Q_b, K_a, V_a) = softmax(\frac{Q_b K_a^T}{\sqrt{d}})V_a \tag{6.2}$$

Multi-head self(cross)-attention module. To consider various attention distributions and multiple aspects of features, the multi-head attention mechanism [6] is introduced. The multi-head attention is the concatenation of *h* single attentions along the channel dimension followed by a linear projection. Thus, the multi-head attention can be formulated as equation (6.3, 6.4)

$$MultiHead(Q, K, V) = Concat(H_1, H_2, \cdots, H_h)W^0 \tag{6.3}$$

$$H_i = Atten(Q_i, K_{i,} V_i) \tag{6.4}$$

where *Atten* is self-attention or cross-attention, Q_i, K_i, V_i are the *i*-th vector of Q, K ,V . In each single attention head, the channel dimension $d^{'} = d/h$.

6.2.2 Feature Fusion Layer

To fuse the features generated from the magnitude and velocity images, we proposed a feature fusion layer (FFL). The structure of FFL shown in Figure.6.3 contains two branches, each branch has one cross-fusion layer and one self-fusion layer.

Cross-Fusion Layer (CFL). CFL is proposed to fuse the features from different modalities. The structure of CFL is illustrated in the upper dash box in Figure.6.3. Given **Q**, **K** and **V** generated from two modalities, the Multi-head Cross-Attention (MCA) module followed by a linear projection firstly integrate those information. Then the fused features are added to the original input. Subsequently, another two linear projections and one residual connection followed by a normalization layer are used to enhance the fused information.

Self-Fusion Layer (SFL). The lower dash box in Figure.6.3 shows the structure of SFL. SFL is a simple stack of Multi-head Self-Attention (MSA), linear projection, residual and normalization layer. Different from CFL, the SFL only uses one input to generate the values for the MSA. CFL aims to fuse the features from different image modalities, SFL further enhances the fused features using self-attention.

Having two feature maps from the magnitude and velocity images respectively, we first transform the feature maps into sequence data using the patch embedding. Specifically, the feature $\mathrm{f} \in \mathrm{R}^{H \times W \times C}$ is divided into $N = HW/P^2$ patches, where the

patch size P is set to 16. The patches are flattened and embedded into a latent D-dimension, obtaining a embedding sequence $e \in \mathrm{R}^{N \times D}$. However, dividing feature maps into patches leads to loss of spatial information. Therefore, a learnable positional encoding sequence is added to the embedding sequence to address this issue. Then the sequence data is passed into the FFL. In this work, we used a stack of 4 FFLs as the feature fusion network (FFN).

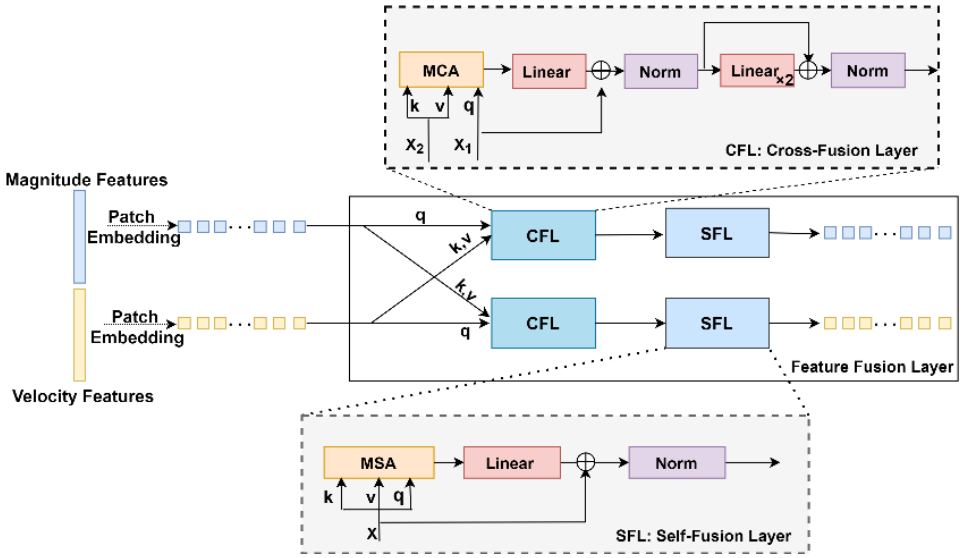


Figure.6.3. Structure of feature fusion layer (FFL). The input of the feature fusion layer is two features derived from magnitude and velocity images respectively. The upper box is the structure of cross-fusion layer (CFL) and the lower one is the structure of self-fusion layer (SFL).

6.2.3 Network Structure

Figure.6.4 illustrates the proposed segmentation network, which takes the U-Net as the backbone. The encoder uses two parallel branches to extract features from magnitude and velocity image separately. The features at the same level are integrated using the feature fusion network. By doing so, the size of integrated features reduces due to the patch embedding. Hence, the fused features are up-sampled first, then added to the original features as the final aggregated features.

The four-level paired aggregated features derived from the encoder are taken as the inputs to the decoder part. The fused features at the same level generated from the magnitude and velocity branch in the encoder are concatenated followed by a convolutional layer to reduce the number of feature maps. The remaining decoder parts including the up-sampling, convolutional and softmax layers are the same as in U-Net.

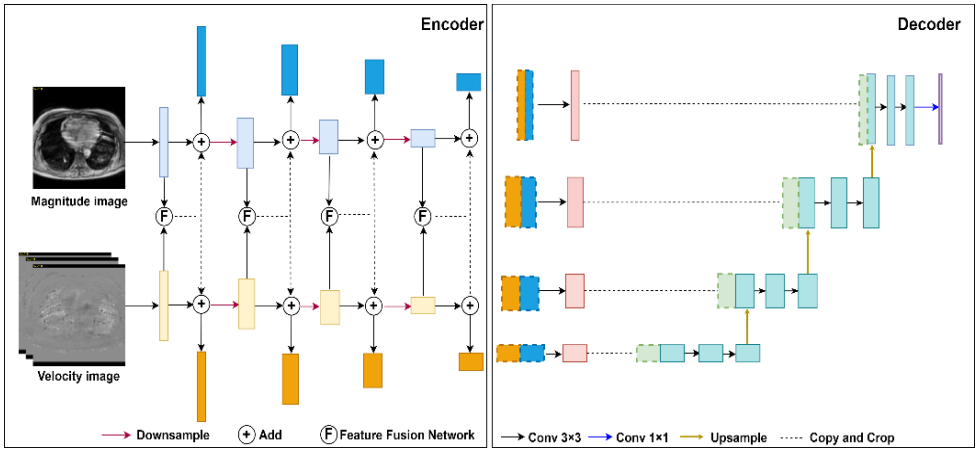


Figure.6.4. The architecture of our proposed segmentation network structure. The feature fusion network is a stack of 4 FFLs.

6.3 Materials

6.3.1 Dataset

4D flow MRI was performed in 28 healthy volunteers and 76 post-myocardial infarction patients on a 1.5T MR system (Philips Healthcare). The 4D flow acquisition covered the complete LV and was acquired in axial orientation with a voxel size of 3×3×3 mm^3 and reconstructed into 30 cardiac phases. The other imaging parameters are as follows: flip angle=10°, velocity encoding (VENC) of 150 cm/s, FOV= 370-400×370-400 mm^2, echo time (TE)=1.88-3.75 ms, repetition time (TR)= 4.78-13.95 ms. In addition, standard cine-MRI was performed in multiple short-axis slices covering the LV from base to apex. More details about the MR acquisition protocol can be found here [16].The short-axis cine acquisition was used to segment the LV endocardial boundaries in all slices and phases. After rigid registration with the 4D flow acquisition, the defined segmentation served as ground truth segmentation of the 4D flow acquisition. Based on the known short-axis orientation, the 4D flow data was resliced into the short-axis orientation using a slice spacing of 3 mm and a fixed number of 41 slices. The spatial in-plane resolution was defined equal to the available short-axis cine acquisition and varied from 0.83×0.83 mm^2 to 1.19×1.19 mm^2.

Excluding the images without any objects this resulted in 91 182 annotated pairs of 2D images, each pair has one 2D magnitude image and three-directional velocity images. The subjects were randomly split into three parts with 64, 20, 20 (total number of images: 55 825, 17 335 and 18 022) for training, validation and testing respectively. We normalized the magnitude image into [0,1] using min-max method. The images were cropped into 256×256.

6.3.2 Evaluation metrics

Segmentation metrics. To quantitatively evaluate the segmentation performance, Dice and Average Surface Distance (ASD) were measured.

Clinical metrics. The clinical metrics, including the end-diastolic volume (EDV), end-systolic volume (ESV), left ventricle ejection fraction (LVEF) and kinetic energy (KE) [3] were measured. The formula of LVEF and KE are defined as:

$$LVEF = \frac{EDV - ESV}{EDV} \times 100\% \qquad KE = \sum_{i=1}^{N} \frac{1}{2} \rho_{blood} \cdot V_i \cdot v_i^2 \qquad (6.5)$$

where N means the number of voxels in the LV, ρ_{blood} represents the density of blood (1.06g/cm^3), V is the voxel volume and v is the velocity magnitude. For each phase, the total KE is the summation of the KE of every voxel within LV. KE was normalized to EDV as recommended by other researchers [3].

Statistical Analysis. The results are expressed as mean ± standard deviation. Pearson correlation coefficient (PCC) was introduced to measure the correlation of the clinical metrics between the manual and automatic segmentation approaches. Paired evaluation metrics were compared using Wilcoxon-signed-rank test with $P < 0.05$ indicating a significant difference.

The Dice, ASD and KE reported in this work are the mean values as computed over 30 phases per subject.

6.4 Experiment and results

All the models were implemented in Pytorch and trained with a NVIDIA Quadro RTX 6000 GPU with 24 GB memory from scratch. We employed Adam as the optimizer with 0.0001 as the learning rate. All of the models were trained for 1000 epochs with a batch size of 15. The sum of Dice loss and cross-entropy loss was used as the loss function. Additionally, due to the complexity of the velocity images, we did not employ any data augmentation methods to enlarge the dataset.

We first evaluated our model against the U-Net, TransUnet [8], and U-NetCon. TransUnet added the self-attention module to the last layer of the encoder. The structure of U-NetCon (shown in the Supplementary) is similar to our proposed network. After removing the feature fusion network, the U-NetCon introduces two U-Net encoders which extract the features from two modalities separately and subsequently, the features from the same level in the encoder are concatenated as the input of the decoder. The input of U-Net and TransUnet is a four-channel stack of one magnitude and three velocity images. Whereas, in our method and U-NetCon, the magnitude and velocity images are taken as two separate input branches.

Table.6.1. Segmentation performance of different methods. Err means the absolute error between the manual and automatic segmentation methods.

Model	Dice (%)	ASD (mm)	EDV-Err (ml)	ESV-Err (ml)	LVEF-Err (%)	KE-Err (μJ/ml)
U-Net	84.62±5.91	2.99±1.66	20.35±31.53	16.01±19.76	7.60±7.10	1.50±1.64
U-NetCon	84.57±6.15	3.19±1.74	22.57±29.46	17.08±24.46	6.11±5.43	0.95±1.94
TransUnet	84.27±5.35	3.09±1.33	18.09±22.91	23.92±16.06	11.79±7.64	0.51±0.48
Ours	**86.52±5.54**	**2.51±1.14**	**9.02±10.03**	**11.86±10.55**	**5.10±4.55**	**0.36±0.34**

Table.6.1 reports the evaluation results of various metrics. It shows our method achieved the best performance for all of the six metrics. In Table.6.2 the PCC of the clinical metrics derived from different models are presented. Our method performs the best on all clinical metrics demonstrating a high correlation. Comparing the results of U-Net and TransUnet, the Dice and ASD only showed marginal improvement, but the performance decreased in LVEF with a low PCC of 48.7%**.** In order to evaluate the effectiveness of feature fusion network, we further compared our method to U-NetCon. As compared to U-NetCon, our method improves the Dice by 2% and the PPC by 3%, 9%, 7% and 16% for LVEF, EDV, ESV and KE, respectively, confirming that the proposed feature fusion network efficiently aggregates the features from magnitude and velocity images. More results about the boxplot and correlation comparing the Dice and four clinical parameters derived from our method and U-NetCon can be found in the supplementary.

Table.6.2. PCC of the clinical metrics derived from manual and automatic segmentation results.

Model	LVEF	EDV	ESV	KE
U-Net	70.65%	84.09%	91.50%	83.76%
U-NetCon	80.61%	88.46%	89.49%	82.46%
TransUnet	48.70%	91.36%	90.33%	97.86%
Ours	**83.26%**	**97.40%**	**96.97%**	**98.92%**

The P-value of Wilcoxon test results between the ground truth and our method in LVEF, EDV, ESV, KE are 0.13, 0.43, 0.35 and 0.43, as shown in Figure.6.5. All of those P-values are larger than 0.05, which confirmed that there is no significant different between the clinical parameters derived from the manual and our automatic segmentation.

6.5 Conclusion

In this paper, we proposed a Transformer based feature fusion network to aggregate the features from different modalities for LV segmentation in 4D flow MRI data. In

the feature fusion network, we introduced a self- and a cross-fusion layer to investigate the inter- and intra- relationship for the features from two different modalities. The proposed method was trained and evaluated in a large in-house dataset and the results of the segmentation accuracy and clinical parameters demonstrate superiority of our method against state-of-arts. We expect that the use of carefully designed data augmentation methods for the velocity images may result in further improvement of the performance of the proposed method.

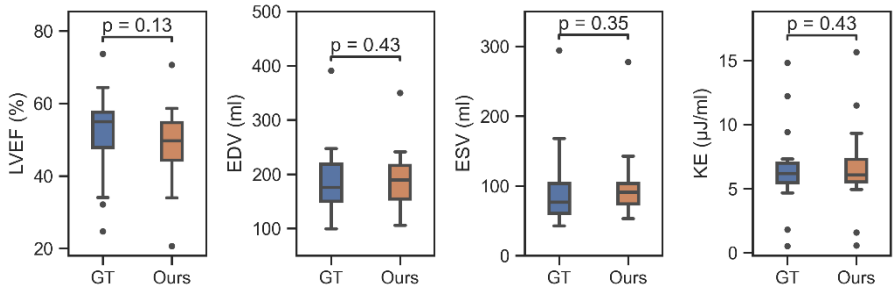


Figure.6.5. Box plots comparing four clinical evaluation metrics including EDV, ESV, LVEF and KE derived from the manual segmentation and our prediction. GT represents the ground truth. P-value was computed using Wilcoxon-signed-rank test. P<0.05 indicate a significant difference between two variables.

References

1. Tao, Q., et al.: Deep learning-based method for fully automatic quantification of left ventricle function from cine MR images: a multivendor, multicenter study. Radiology 290(1), 81–88 (2019)
2. Bai, Wenjia, et al. "Automated cardiovascular magnetic resonance image analysis with fully convolutional networks." Journal of Cardiovascular Magnetic Resonance 20.1 (2018): 1-12.
3. Garg, Pankaj, et al. "Left ventricular blood flow kinetic energy after myocardial infarction-insights from 4D flow cardiovascular magnetic resonance." Journal of Cardiovascular Magnetic Resonance 20.1 (2018): 1-15.
4. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedi- cal image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
5. Isensee, Fabian, et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation." Nature methods 18.2 (2021): 203-211.
6. Vaswani, A., et al. "Attention is all you need. " Adv. Neural. Inf. Process. Syst. 30, 5998–6008 (2017)
7. Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).
8. Chen, J., Lu, Y., Yu, Q., Luo, X., et al. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021).
9. Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." arXiv preprint arXiv:2103.14030 (2021).
10. Cao, Hu, et al. "Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation." arXiv preprint arXiv:2105.05537 (2021).
11. Li, Hang, et al. "DT-MIL: Deformable Transformer for Multi-instance Learning on Histopathological Image." International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2021.
12. Luo, Yanmei, et al. "3D Transformer-GAN for High-Quality PET Reconstruction." International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2021.
13. Ji, Yuanfeng, et al. "Multi-Compound Transformer for Accurate Biomedical Image Segmentation." International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2021.
14. Berhane, Haben, et al. "Fully automated 3D aortic segmentation of 4D flow MRI for hemodynamic analysis using deep learning." Magnetic resonance in medicine 84.4 (2020): 2204-2218.

15. Wu, Yinzhe, et al. "Automated multi-channel segmentation for the 4D myocardial velocity mapping cardiac MR." Medical Imaging 2021: Computer-Aided Diagnosis. Vol. 11597. International Society for Optics and Photonics, 2021.
16. Garg, Pankaj, et al. "Left ventricular thrombus formation in myocardial infarction is associated with altered left ventricular blood flow energetics." European Heart Journal-Cardiovascular Imaging 20.1 (2019): 108-117.

Supplementary

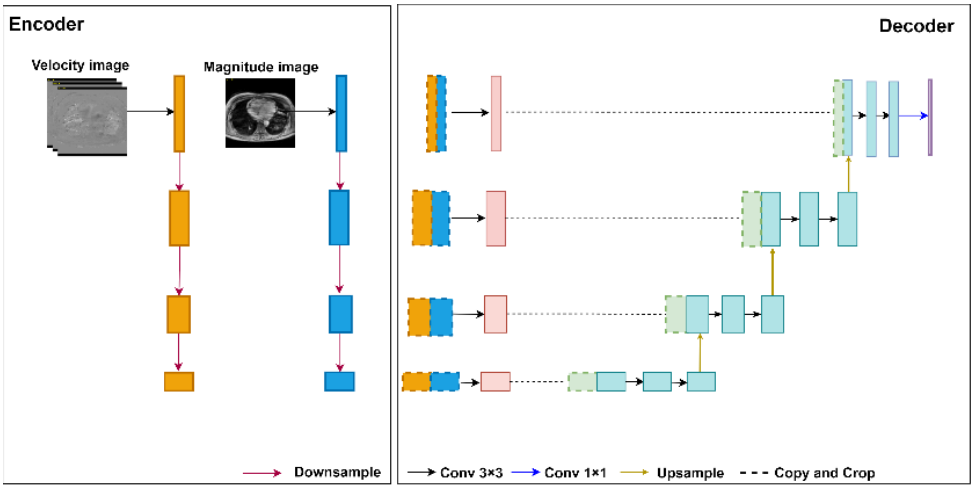


Figure. S6.1. The structure of U-NetCon.

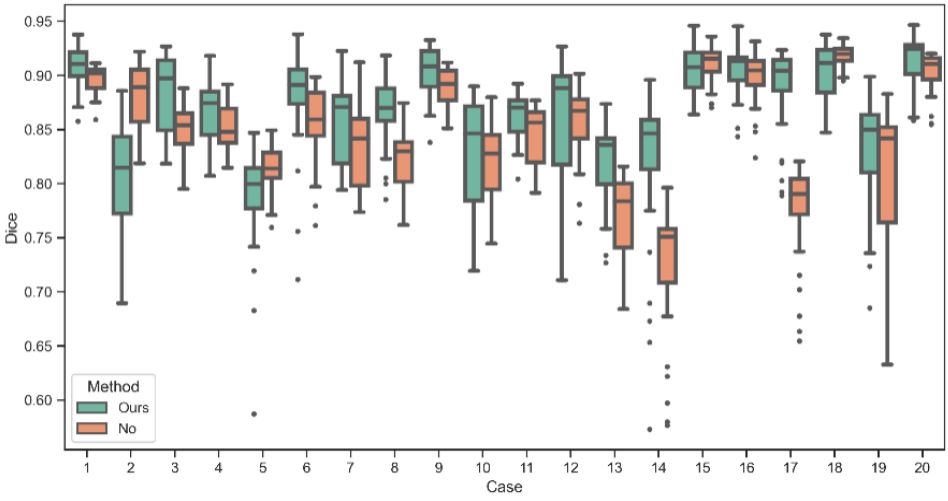


Figure. S6.2. Box plot comparing the Dice derived from our method and U-NetCon on 20 testing cases. The Dice was computed based on each phase, and each box contains 30 phases.

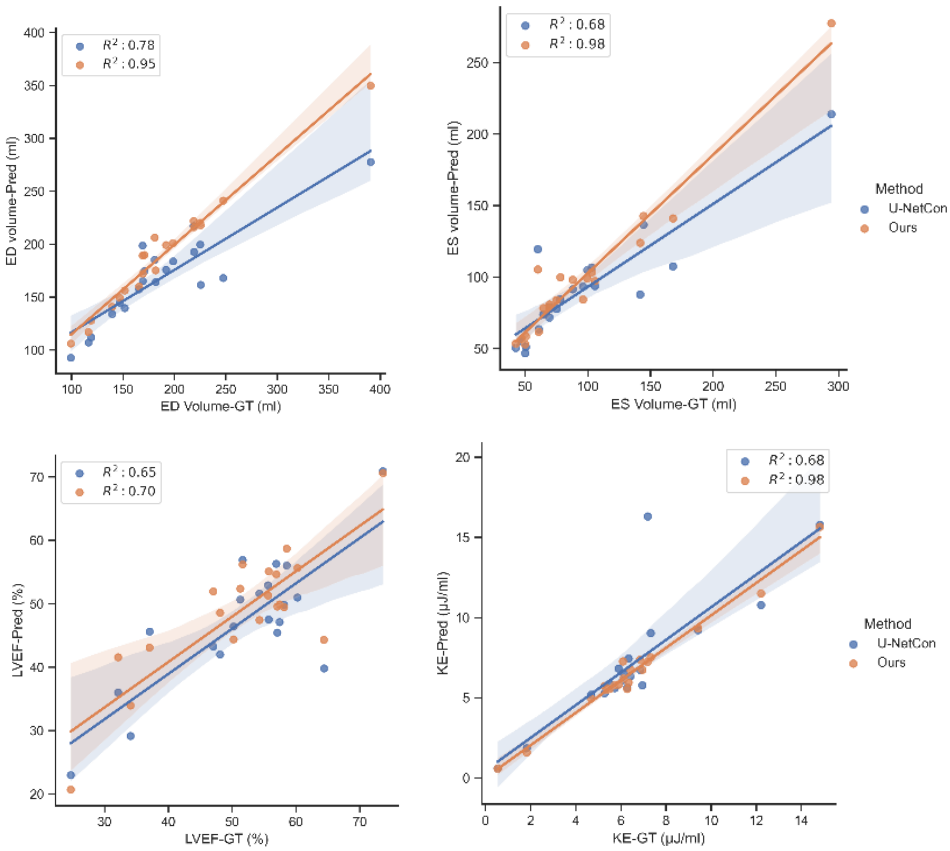


Figure. S6.3. Correlation comparing EDV, ESV, LVEF and KE derived from our method and U-NetCon. GT in the X-axis represents the parameters derived from the ground truth, the Y-axis represents the parameters derived from the prediction of different models.

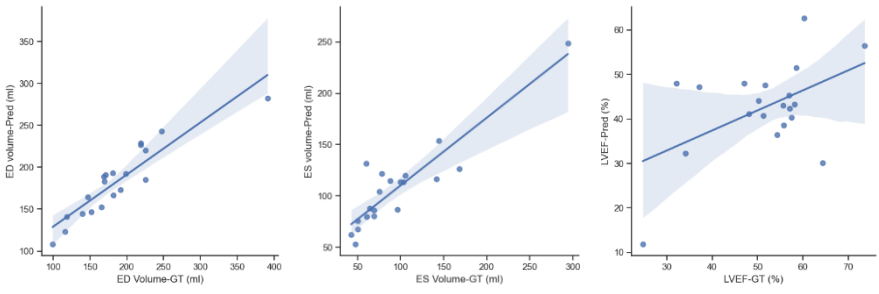


Figure. S6.4. Correlation of EDV, ESV and LVEF derived from TransUnet and ground truth.

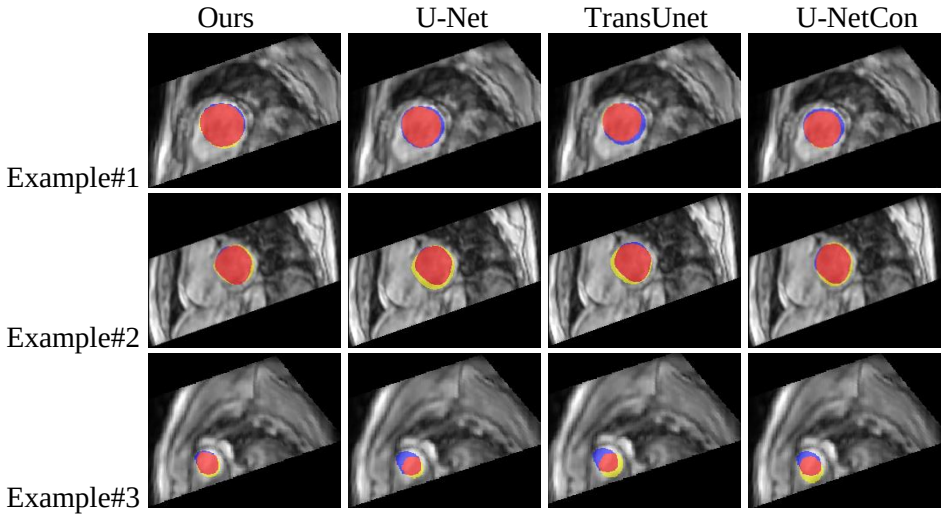


Figure. S6.5. Examples of segmentation results from our method, U-Net, TransUnet and U-NetCon. The blue represents the ground truth, the yellow is the prediction, and the red is the overlap between the prediction and ground truth

.

Chapter 7 Deep Learning-based Prediction of Intra-Cardiac Blood Flow in Long-axis Cine Magnetic Resonance Imaging

This chapter was adapted from:

Xiaowu Sun, Li-Hsin Cheng, Sven Plein, Pankaj Garg, Mehdi H. Moghari, Rob J. van der Geest**. Deep Learning-based Method for Intra-Cardiac Blood Flow Pattern Prediction using 4D Flow Data.** International Journal of Cardiovascular Imaging. (2023): 1-9.

Abstract

Purpose: We aimed to design and evaluate a deep learning-based method to automatically predict the time-varying in-plane blood flow velocity within the cardiac cavities in long-axis cine MRI, validated against 4D flow.

Methods: A convolutional neural network (CNN) was implemented, taking cine MRI as the input and the in-plane velocity derived from the 4D flow acquisition as the ground truth. The method was evaluated using velocity vector end-point error (EPE), angle error and accuracy. Additionally, the E/A ratio and diastolic function classification derived from the predicted velocities were compared to those derived from the 4D flow.

Results: For intra-cardiac pixels with a velocity >5 cm/s, our method achieved an EPE of 8.65 cm/s, angle error of 41.27°. For pixels with a velocity >25 cm/s, the angle error significantly degraded to 19.26°. Although the averaged blood flow velocity prediction was under-estimated by 26.69%, the high correlation (PCC=0.95) of global time-varying velocity and the visual evaluation demonstrate a good agreement between our prediction and 4D flow data. The E/A ratio was derived with minimal bias, but with considerable mean absolute error of 0.39 and wide limits of agreement. The diastolic function classification showed a high accuracy of 86.9%.

Conclusions: Using a deep learning-based algorithm, intra-cardiac blood flow velocities can be predicted from long-axis cine MRI with high correlation with 4D flow derived velocities. Visualization of the derived velocities provides adjunct functional information and may potentially be used to derive the E/A ratio from conventional CMR exams.

7.1 Introduction

Assessment of cardiac function using cardiac magnetic resonance imaging (CMR) is typically based on cine MR imaging. Four-dimensional (4D) flow MRI enables time-resolved three-dimensional visualization of intra-cardiac blood flow to gain a better understanding of the patient's cardiac condition [1, 2]. Cardiac dysfunction is strongly associated with abnormal patterns of blood flow within the cardiac chambers. Therefore, visualization and quantification of intra-cardiac blood flow may provide relevant diagnostic information. However, 4D flow MRI is usually not performed in routine clinical protocols as it requires additional scan time and post-processing. During post-processing typically registration is required of the 4D flow acquisition with the acquired long-axis and short-axis cine views, which may be hampered by variations in respiratory condition and heart rate [3, 4]. Interestingly, in standard long-axis cine MR views, the intensity fluctuations within the cardiac cavities provide a visual clue about the global blood flow pattern. While the signal intensity variations are dependent on various factors such as saturation effects and spin dephasing due to magnetic field inhomogeneity or complex flow [5, 6], we speculate that time-varying flow velocity information can be derived from those intensity variations.

There have been many attempts in using balanced steady-state free precession (SSFP) MR imaging for measuring blood velocity by modifying the SSFP sequence. Markl et al. measured through-plane flow using a SSFP sequence by inverting the slice encode gradient between two consecutive acquisitions [7]. The through-plane velocity was then calculated by subtracting the resulting phase images. Neilson et al. augmented the slice encode gradient in the SSFP sequence for measuring blood velocity in a readout direction [8]. They used the resultant phase information without a reference for measuring the blood velocity in the readout direction. In recent years, convolutional neural networks (CNN) have been introduced to extract cardiac motion information, which could be interpreted as an ensemble of relatively small, periodical variations of the shape and position of heart structures during a cardiac cycle [9, 10, 11]. However, the potential applications for velocity field prediction has not been explored yet.

Accordingly, in this work we proposed a deep learning-based method to track the blood flow displacement within consecutive cardiac frames from long-axis cine MR images. As ground truth, we used the velocity field derived from registered 4D flow MRI. Once the blood flow is tracked and the displacement vectors in X and Y directions are measured, pixel wise blood velocity in each direction can be derived by dividing its displacements to the temporal resolution of each frame. To the best of our knowledge, we are the first to employ deep learning and 4D flow MRI for automated cardiac blood flow prediction. Additionally, in clinical routine, diastolic

function is usually evaluated using Doppler echocardiography. Although, several studies demonstrated the usefulness of CMR in deriving conventional diastolic parameters, those methods rely on additional scan time and extra post-processing, such as the manual localization of regions of interest (ROI), which is time-consuming [12, 13, 14]. In our work the E/A ratio is automatically derived from the predicted blood flow and was used to classify the diastolic function as a potential clinical application.

7.2 Methods

7.2.1 Dataset

The study cohort included 78 post-myocardial infarction (MI) patients and 34 healthy subjects who underwent cardiac MRI on a 1.5T MR system (Philips Healthcare). The study was approved by the local medical ethical committee and all participant in the study provided written information consent. The MR imaging protocol included conventional SSFP cine in 4-chamber (4CH) view and short-axis cine stack. In addition, whole-heart 4D flow MRI was performed for 3D blood flow velocity assessment in the four cardiac chambers. Both cine MRI and 4D flow MRI were reconstructed into 30 phases covering a complete cardiac cycle. MR imaging parameters of the acquisitions are listed in Table.7.1. More details about the MR acquisition protocol have been reported in earlier work [15, 16].

Table.7.1 4D flow and SSFP data acquisition parameters. VENC: velocity encoding; FOV: field of view; TE: echo time; TR: repetition time; bpm: beats per minute.

	4D Flow Data	SSFP
Spatial resolution (mm^3)	3×3×3	0.95-1.25× 0.95-1.25×8
Reconstructed temporal resolution (ms)	20.83-46.73	20.21-48.21
Electrocardiogram gating	retrospective	retrospective
VENC (cm/s)	150	—
FOV (mm^2)	300-440 × 300-440	300-440 × 300-440
TE/TR (ms)	3.10-3.75/7.46-13.95	1.5-1.72/3.0-3.44
Flip angle (°)	10	60
Reconstructed heart phases	30	30
Scan time	7-10 min	6-8 s
Heart rate (bpm)	41-94	42-99
Motion correction	None (free breathing)	Breath hold

Mass software (Version V2017-EXP; Leiden University Medical Center, Leiden, the Netherlands) was used to derive LV volumetric parameters from the short-axis cine stack by semi-automated segmentation of the endocardial and epicardial borders.

The semi-automatically defined ventricular and atrial contours in the 4CH view were used as a mask and for each pixel within the mask the in-plane component of velocity as derived from the aligned 4D flow acquisition was used as the ground truth. To avoid temporal inconsistency, cine acquisitions were excluded if the heart rate deviated from that of the 4D flow acquisition by more than six beats per minute. Based on this exclusion criterion, 92 cases (2760 2D images) remained for training and testing. Table.7.2 summarized the detailed demographics derived from the short-axis cine and 4D flow data.

Table.7.2. Demographics of the study cohort derived from the short-axis cine and 4D flow data. Data is presented as mean ± standard deviation or count. EDV: End-diastolic volume, ESV: End-systolic volume, SV: Stroke volume, EF: Ejection fraction.

Characteristic	Subjects(n=92)
Gender (Male, n)	56
EDV (ml)	179.71±63.93
ESV (ml)	90.14±58.27
SV (ml)	89.58±19.38
EF (%)	53.11±12.27
E/A ratio	1.41±0.54

In-plane spatial alignment was performed between the SSFP cine and reformatted 4D flow images since 4D flow images were acquired during free-breathing while SSFP cine images were acquired during breath-hold. In addition, significant patient motion can occur in between the acquisition of the long-axis cine view and the 4D flow acquisition. Based on the image position information, the in-plane velocity derived from 4D flow was projected on the cine long-axis views. In case a misalignment was observed between the visualized anatomy and the velocity vectors, the cine view images were manually translated in order to optimize the alignment. We further assumed that both 4D flow and SSFP cine images are registered in time since both have the same number of cardiac phases and nearly similar heart rates. Therefore, each cardiac phase of 4D flow is assumed to correspondent to same cardiac phase of SSFP cine. Tri-linear interpolation was used to generate the in-plane velocity components for the 4CH long-axis views.

7.2.2 Data preprocessing

In this work, we aim to predict the blood flow velocity within the cardiac chambers. To filter out irrelevant velocity information, we applied a binary blood pool mask in the long-axis view to exclude the region outside of the cardiac chambers. The signal intensities of the input cine sequence were normalized based on the histogram of the signal intensities within the masked region. The histogram was constructed by aggregating the blood pool pixels of all cardiac phases, which implies that signal loss

information is still preserved and flow-induced artifacts can still be tracked from frame to frame. The normalization can be described as in formula 1, where $P_{\mathrm{norm-i}}$, the normalized value of the pixel-*i* is derived from P_I the signal intensity of pixel-*i*, P_{5th} and P_{95th} represent the 5th and 95th percentile value of the intensity histogram.

$$\boldsymbol{P_{norm-i}} = \frac{\boldsymbol{P_I - P_{5th}}}{\boldsymbol{P_{95th} - P_{5th}}} \tag{7.1}$$

The intensity fluctuations in the cine MR sequence are used to predict the displacement of a pixel, i.e. a blood sample, from frame to frame. However, the 4D flow acquisition provides each pixel's velocity instead of displacement. Therefore, the pixel velocities derived from the 4D flow acquisition are converted into the pixel displacements using formula 7.2,

$$\mathbf{D} = \left(\frac{\Delta t v_x}{ps_x}, \frac{\Delta t v_y}{ps_y}\right) \tag{7.2}$$

in which $\mathbf{V} = (v_x, v_y)$ stands for velocity of each pixel in frame t, v_x, v_y are the velocities projected on the long-axis image, Δt is the time interval between image frame t and t+1, $\mathbf{PS} = (ps_x, ps_y)$ is the pixel spacing. After this preprocessing, the displacement **D** (in pixel units) from frame t to frame t+1 is regarded as the ground truth for model training.

7.2.3 Network structure

The displacement information and moving direction of a pixel, or group of pixels, can only be extracted using the current and its neighboring frames. To predict the in-plane components of blood flow velocity, we consider a sequence of cine MR images containing a central image and its 8 temporal neighboring phases as the input and the displacements in X and Y direction derived from the 4D flow sequence as the ground truth to train an end-to-end network. The proposed CNN architecture is illustrated in Figure.7.1.

The implemented network is a variant of U-Net [17] and ResNet [18] containing a contracting path and an expanding path. In the contracting path, to provide dense per-pixel predictions, one pooling operation and three strided convolutions with a 1×1 kernel size are applied for the down-sampling. The conventional convolution layers in the contracting path of U-Net are replaced with residual convolution modules [18] to extend and deepen the network. In the expanding path, we reserved the concatenation-based skip connections to integrate the local features and the global information.

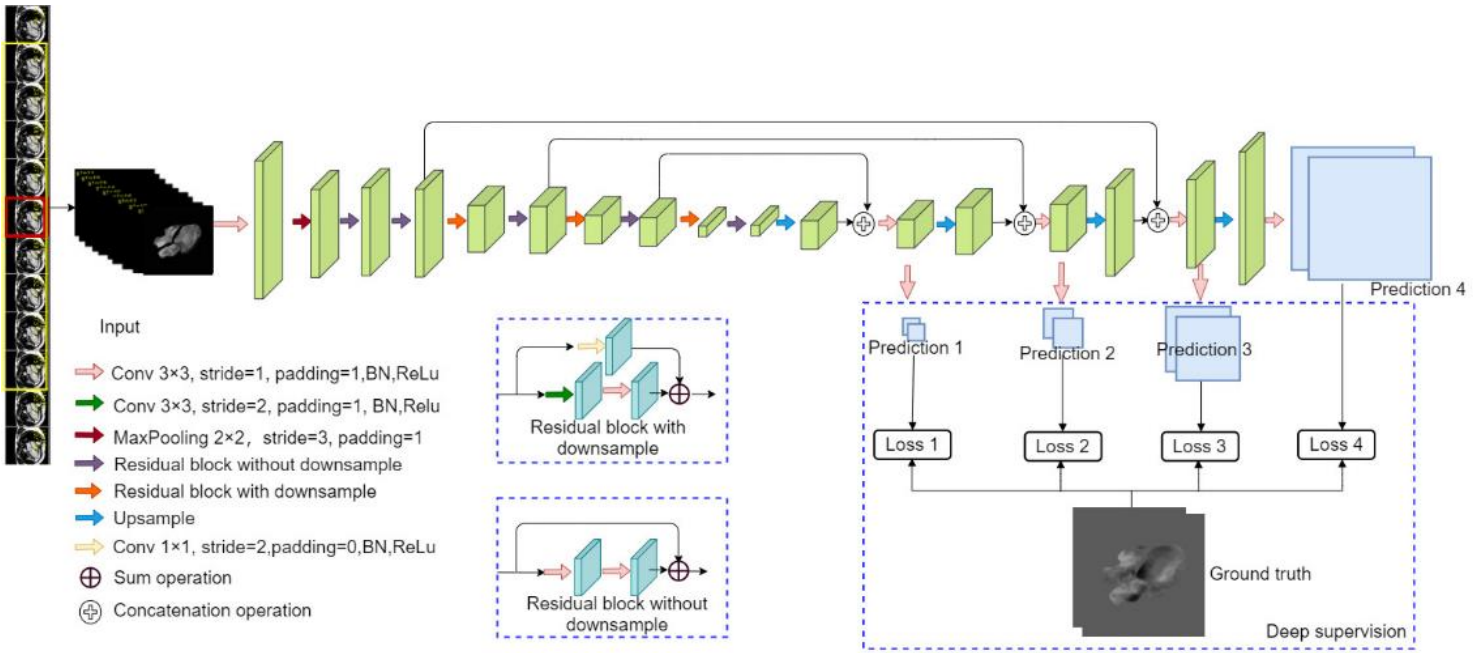


Figure.7.1. Architecture of the proposed network. The input of the network is a sub-sequence of 2D cine MR images including the target image (in red box) and its 8 temporal neighboring frames (in yellow box). Prediction 1,2,3,4 are four outputs with two dimensions on X and Y directions, respectively, which are used to compute the deep supervision loss. Prediction 4 is the final output of the network and is the one evaluated during the test. Only the pixels within the blood pool

Deep supervision [19] is employed to overcome the problem of vanishing gradients in a deep CNN architecture. As shown in Figure.7.1, three auxiliary prediction layers are inserted before the up-sampling operation, each prediction is resampled into the original image size using nearest neighbor interpolation. The end point error (EPE), being the Euclidean distance between two displacement vectors averaged over all pixels within the cardiac cavities, is used as loss function. Given $D_{x,g}, D_{y,g}, D_{x,p}, D_{y,p}$ representing the displacement values of ground truth and prediction in X and Y directions, $\mathbf{D}_{i,g} = (D_{x,g}, D_{y,g})$ and $\mathbf{D}_{i,p} = (D_{x,p}, D_{y,p})$ denoting the displacement vectors for ground truth and prediction of i[th] pixel within the blood pool, then the EPE is defined according to formula 7.3 where *M* indicates the number of pixels within the blood pool.

$$\text{EPE} = \frac{1}{M}\sum_{i=0}^{M}\left\|\mathbf{D}_{i,p} - \mathbf{D}_{i,g}\right\| = \frac{1}{M}\sum_{i=0}^{M}\sqrt{(D_{x,p} - D_{x,g})^2 + (D_{y,p} - D_{y,g})^2} \quad (7.3)$$

The EPE loss is the sum of length of the displacement vector difference to compute the magnitude and angle error between prediction and ground truth for all pixels within the blood pool. The total loss is defined as:

$$\text{Loss} = \text{EPE}\,(\text{G}, \text{O}) + \sum_c w_c EPE_c(G, P_c) \quad (7.4)$$

where G is the displacement generated from the 4D flow data, O is the final output from the network, P_c is the prediction of the c[th] auxiliary prediction layer and w_c is the loss weight of each auxiliary prediction.

To improve the performance and the generalization of the model, five-fold cross-validation was applied. The output of CNN was divided by the temporal resolution to convert to velocity to compute the evaluation metrics.

7.3 Evaluation metrics

7.3.1 Visual evaluation

To visually assess the intra-cardiac blood flow patterns derived from either the CNN prediction and 4D flow, the in-plane velocity was displayed in movie mode as vector overlay projected on the cine MR images. The length and color of the displayed vectors were scaled according to the velocity magnitude.

7.3.2 Quantitative evaluation metrics

The performance of the proposed method was evaluated using EPE and angle error.

To quantitatively assess the performance of predicted blood flow, both the magnitude and angle error are required to be measured. Therefore, EPE described in formula 7.3 and trigonometric function are employed to compute the magnitude error

and angle error, respectively. Here, the EPE was computed using the velocity vectors instead of the displacement vectors. The angle error θ, between the ground truth $\mathbf{V}_{i,g}$ and prediction $\mathbf{V}_{i,p}$ of the i^{th} pixel within the blood pool, is defined as,

$$\theta = \frac{1}{M}\sum_{i=0}^{M} \arccos\left(\frac{\mathbf{V}_{i,p}\cdot\mathbf{V}_{i,g}}{\|\mathbf{V}_{i,p}\|\,\|\mathbf{V}_{i,g}\|}\right) \tag{7.5}$$

where i represents the i^{th} pixel and M indicates the total number of pixels within in the blood pool, $\|\cdot\|$ is the length of a vector and *arccos* means the inverse trigonometric function of cosine. The angle error ranges between 0° and 180°, with 0° denoting two vectors in the same direction and 180° denoting two vectors in the opposite direction.

7.3.3 Clinical parameters

A commonly clinically used flow-related parameter is the E/A ratio. The E/A ratio can be used to classify diastolic function as either normal or abnormal using the cutoff values for E/A ratio as commonly used in cardiac ultrasound. In our work, a region of interest was first defined by three points, being two end points of the defined endocardial contour, which correspond to the valve hinge points, and a third point in the center of LV cavity. A b-spline curve was fitted through the three points, resulting in a region just below the mitral valve plane. The E and A velocities were found by searching for the pixel with maximum (in-plane) velocity within the region to derive the E/A ratio.

7.3.4 Statistical analysis

Results are expressed as mean ± standard deviation (SD). Pearson correlation coefficient (PCC) was used to evaluate the correlation between our prediction and the 4D flow data for the velocity values during a complete cardiac cycle. In addition, Bland-Altman analysis was used to analyze the mean differences (Bias) and limits of agreement (LOA, 1.96×SD) of the E/A ratio derived from either the deep learning method or 4D flow data. Paired t-test was performed to test the statistical significance of the differences between paired E/A ratio measurements, $P<0.05$ indicates a significant difference. PCC was also used to measure the correlation of E/A ratio derived from 4D flow data and our approach.

7.4 Results

We first introduced 9 neighboring cine MR phases in the input (more results using different number of inputs can be found in the Supplementary file), then we reported the predicted results using the defined metrics. At last, the E/A ratio results were reported.

7.4.1 Visual comparison

The predicted and 4D flow derived in-plane blood flow velocity were dynamically visualized as overlay on the original long-axis cine images. The length and colouring of the vectors were used to encode the local blood velocity magnitude. To avoid cluttering of the vectors and to suppress velocity noise the velocity vectors were only generated for image pixels with a velocity >4 cm/s. Figure.7.2 shows an example of selected frames of predicted blood flow velocities compared to 4D flow derived velocities in one of the study subjects. Overall a good agreement is seen in the blood velocity pattern within the cardiac cavities both in systole and diastole. In general it was observed that the visual agreement in flow pattern was better in the ventricles than in the atria. Video examples can be found here (https://github.com/xsunn/BloodFlowPrediction).

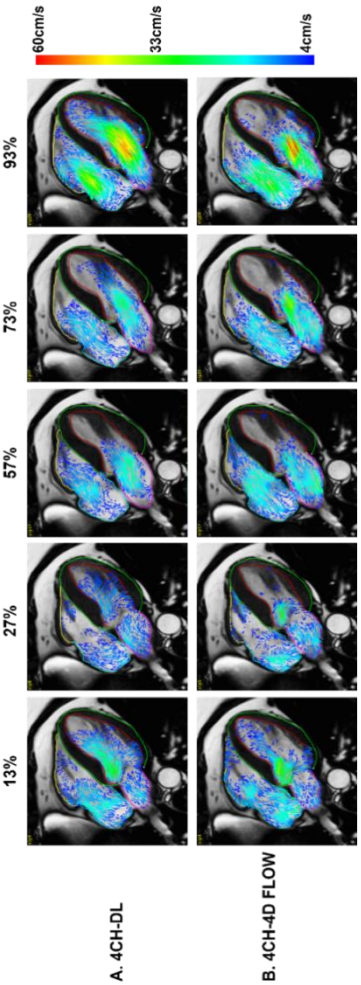


Figure.7.2. Five out of 30 frames of blood flow pattern generated from deep learning-based method and 4D flow. **(A)**: Blood flow pattern in 4CH view using deep learning. **(B):** Corresponding ground truth from 4D flow data in 4CH view. Those five frames are at 13%, 27%, 57%, 73%and 93% of one cardiac cycle.

7.4.2 Quantitative Results

Figure.7.3 shows probability distributions of blood flow velocity in different heart chambers generated from 4D flow data and our prediction. Compared with the ground truth, the predicted velocities were generally lower.

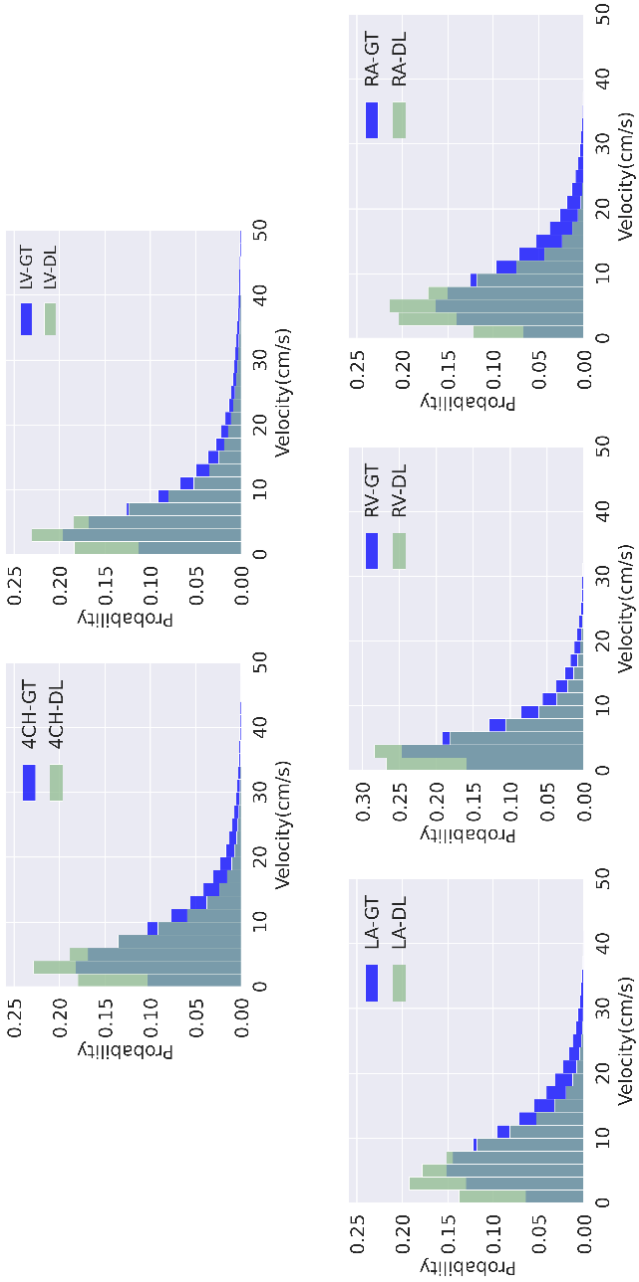


Figure.7.3. Probability distribution of velocity generated from 4D flow data and prediction in 4CH view. The blue color represents the distribution generated from the 4D flow data, and the light green means the distribution generated from the prediction. The light blue represents the overlap between the prediction and 4D flow data.

To quantify the prediction error, those pixels with velocities greater than 5cm/s were involved in computing the EPE and angle error. The accuracy was computed with 30th percentile as a threshold. All pixels were used to compute the relative error (RE) of velocity between the 4D flow and automated velocity prediction. PCC was used to measure the correlation of the time-varying averaged velocity between the 4D flow data and prediction. The results in different heart chambers are reported in Table.7.3. The relative error shows that the velocities were under-estimated by 26.69%. The small standard deviation in the relative velocity difference suggests that potentially a constant correction factor may be applied to the predicted velocity to improve the performance. The PCC of velocity within all four chambers were 0.95, which illustrates a good correlation in the blood flow pattern between the 4D flow and our prediction.Fig.S7.1 in Supplementary shows more details about the performance of our method for different chambers with varying velocity thresholds.

Table.7.3. Prediction results of different chambers. 4CH indicates the results were computed within all 4 chambers; LV, LA, RV and RA mean the results were based on each single chamber separately. RE: relative error. PCC: Pearson correlation coefficient. The mean ± standard deviation are reported.

	EPE (cm/s)	Angle Error (°)	Velocity-RE(%)	Velocity-PCC
4CH	8.65±2.69	41.27±11.39	-26.69± 4.43	0.95
LV	9.10±2.96	37.98±10.94	-24.53±4.29	0.98
LA	8.45±2.20	41.19±12.78	-27.84±6.62	0.94
RV	7.06±1.54	40.99±11.28	-26.18±8.05	0.93
RA	8.64±2.44	47.52±16.90	-29.83±4.53	0.93

7.4.3 E/A ratio results

The average absolute error in E/A ratio estimation were 0.39±0.32. The Bland-Altman analysis as shown in Figure.7.4 reveals a minimal bias with wide limits of agreement (LOA) between our prediction and 4D flow derived E/A ratio and more than 95% of cases are distributed between upper and lower agreement limits.

To investigate the potential clinical applicability of the automated E/A ratio prediction we tested whether the wide LOA effects the classification of diastolic function. Echocardiography is the main imaging modality for assessment of LV diastolic function. It defined 0.75< E/A ratio <1.5 as normal diastolic function and E/A ratio varying in the other ranges as abnormal diastolic function [20]. The confusion matrix of the diastolic function classification experiment are summarized in Figure.7.5. The diastolic function binary classification accuracy was (60+20)/92=86.9%. The other three classification metrics including precision, recall and F1-Score, PCC and P values are reported in Table.7.4. Our method was able to classify 93.75% (60/64) of cases qualified by the 4D flow data as the normal diastolic

function, and 71.43% (20/28) of the abnormal cases were also correctly identified. Due to the wide LOA, the overall PCC of the E/A ratio is 66.71%. The PCC of E/A ratio in the groups with normal and abnormal diastolic function are 39.41% and 75.1%, respectively. But all p values of E/A ratio in both two classes are larger than 0.05, meanwhile, the p value of 0.795 derived from all 92 subjects also confirmed that the E/A ratio generated from our prediction was not significantly different from the 4D flow data.

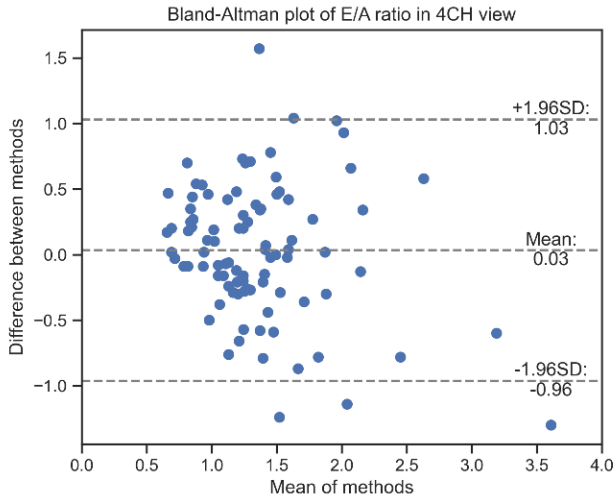


Figure.7.4. Bland-Altman plots of E/A ratio.

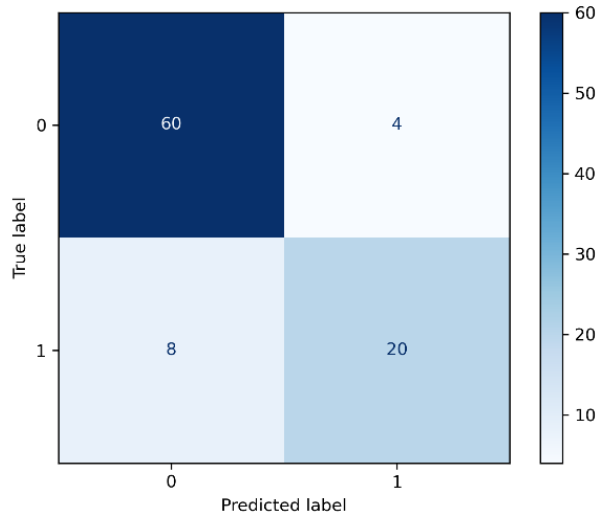


Figure.7.5. Confusion matrix of diastolic function classification derived from the predicted velocities. Label 0 means normal diastolic function, 1 represents abnormal function.

Table.7.4. The results of diastolic function classification, PCC and p value of E/A ratio in each class with normal and abnormal diastolic functions.

	Recall	Precison	F1-Score	PCC	P value
Normal	93.75%	88.24%	90.91%	39.41%	0.052
Abnormal	71.43%	83.33%	76.92%	75.10%	0.088
Overall	-	-	-	66.71%	0.795

7.5 Discussion

We designed and evaluated a deep learning-based method for the prediction of intra-cardiac blood flow velocity from long-axis cine MRI using 4D flow derived velocities as ground truth. The predicted velocities highly correlated with the 4D flow derived velocities with an overall good visual agreement in time-varying flow pattern. Our work shows a potential clinical application to visualize the blood flow pattern without an additional 4D flow data. As the E/A ratio is a well-established clinical parameter used to classify diastolic function, the results demonstrated that the proposed method can be applicable to estimate the E/A ratio without significant bias and to classify the diastolic function with a high accuracy. Although the observed underestimation of the predicted velocities and the variability in the derived measurements indicate that further refinement of the deep learning model using a larger patient cohort is warranted. we believe our results demonstrate the potential of the proposed method.

The variation in blood signal intensity in the cine MR images provides information on the direction and magnitude of the blood flow in the cardiac cavities. The observed displacement of the apparent visible structures in the blood pool in subsequent frames reflects the velocity. Therefore, we performed experiments with different number of neighboring phases as input of the network. Using only three phases as input was shown to result in the worst performance. This may be explained by the fact that the small total displacement like just one pixel in three neighboring temporal phases makes the velocity prediction sensitive to the spatial resolution of the cine images. When using more frames as the input the structures can be followed over a larger time window making it less sensitive to the spatial resolution. It was concluded that more than three neighboring phases are required to predict the blood flow pattern and for the final model 9 neighboring phases was used as input.

The high correlation of the time varying velocity averaged all subjects between our prediction and the 4D flow data, as well as the visual evaluation results, demonstrated a good agreement in the global velocity patterns. However, the velocity values predicted by the proposed model are close to 30% lower than those derived from 4D flow data. In the training data, the low velocities (0-20 cm/s) account for a large proportion which may lead the model to underestimate the

velocities in regions of high velocity. In addition, the evaluation results are sensitive to the selected velocity thresholds, because different levels' velocities are relatively concentrated at certain areas. For example, in the left ventricle, the distribution of the lower velocities are more dispersed and complicated in the apical region, therefore, it is much harder to predict the irregular movement which leads to a relatively large EPE and angle error. The pixels with higher velocities, such as the blood flow from LA to LV in diastole and from the LV towards the aorta in systole, have a relatively fixed direction of motion. Therefore, the angle error decreased when the velocity thresholds increased. However, since the high velocities only account for a small proportion the model is prone to underestimation of high velocities, resulting in a larger EPE for the pixels with higher velocities.

The E/A ratio derived from the velocities could be assessed without bias since both E- and A-velocity were underestimated similarly. Additionally, the statistical test confirmed that there was no significant difference between 4D flow and CNN derived E/A ratio. However, the Bland-Altman analysis revealed a wide limit of agreement. Despite this, the results of diastolic function classification demonstrated that the variability in E/A ratio had minimal effect on the accuracy of diastolic function classification in our study cohort. Echocardiography allows reliable visualization of blood flow pattern. Vector flow mapping (VFM) in echocardiography uses the mass-conservation principle to estimate the azimuthal component of the flow [21]. VFM has been used in many clinical applications including cardiac function evaluation, valvular diseases diagnosis and congenital heart disease. However, VFM is sensitive to out-of-plane flow and boundary conditions [22]. Additionally, the conventional VFM method is applied only to the left ventricle [23]. Our proposed method can be applied to predict the blood flow in the whole heart from any cine long axis view and does not rely on accurate cardiac boundary segmentation. Since cine MRI acquisitions are routinely acquired in standard CMR exams, given the cine MRI, our method can directly predict the in-plane velocities without requiring additional scan time. The combined visualization of blood flow and myocardial motion provides detailed information about cardiac function and hemodynamics. The clinical value of the developed technique should be evaluated in future clinical studies.

There are several limitations in our study. Velocity underestimation is the main limitation since it is patient dependent and varies across the subjects. The use of appropriate data augmentation techniques to artificially enlarge the available set of training data or introducing a weighted loss function by setting larger weights to higher velocities may result in improved performance of the deep learning model. The ground truth generated by projecting the 4D flow data derived in-plane velocities on the long-axis cine MRI is not a perfect reference, due to heart rate difference and patient movement. The heart rate difference cannot be eliminated

completely, even though some cases were excluded to keep the temporal consistency. Registration errors can be corrected for visually by applying in-plane translation of the cine MRI images series. Through-plane misalignment and rotational errors are more difficult to correct for. Additionally, as our method relies on converting predicted pixel displacement to velocity, the limited spatial and temporal resolution of the cine MRI data will have an impact on the velocity magnitude and direction prediction. The 4D flow MRI was acquired during free-breathing while SSFP cine images were acquired during breath-hold, implying a difference in physiological condition of the subject. For regions of low blood flow velocity the noise in the 4D flow data may be non-negligible. Additionally, training and testing the model on a wider range of data from multiple scanner types, centers is also required to gain a further understanding in the potential of the proposed blood flow velocity prediction method. Furthermore it would be valuable to investigate the applicability of our method in patients with valvular regurgitation or stenosis and other patient cohorts with cardiac pathologies associated with abnormal flow patterns, such as patients with dyssynchronous myocardial contraction. Since a full detailed electrocardiographic QRS duration evaluation was not available for the patients in our study, we were unable to perform a patient sub-group analysis.

In conclusion, we proposed a deep learning-based method for automated intra-cardiac blood flow velocity prediction from standard long-axis cine MRI. It was demonstrated that, although the predicted velocity magnitude is underestimated, the global velocity patterns show good correlation with the blood flow patterns derived from 4D flow MRI. The method enables estimation of E/A ratio without significant bias, but with wide limits of agreement. After further improvement of the velocity prediction model the method could potentially be valuable for clinical application.

Acknowledgments

Prof. Sven Plein from the University of Leeds is acknowledged for granting access to the image data used in this work.

References

1. Stankovic, Z., Allen, B.D., Garcia, J., Jarvis, K.B. and Markl, M., 2014. 4D flow imaging with MRI. Cardiovascular diagnosis and therapy, 4(2), p.173.
2. Markl, M., Frydrychowicz, A., Kozerke, S., Hope, M. and Wieben, O., 2012. 4D flow MRI. Journal of Magnetic Resonance Imaging, 36(5), pp.1015-1036.
3. Bock, J., Frydrychowicz, A., Lorenz, R., Hirtler, D., Barker, A.J., Johnson, K.M., Arnold, R., Burkhardt, H., Hennig, J. and Markl, M., 2011. In vivo noninvasive 4D pressure difference mapping in the human aorta: phantom comparison and application in healthy volunteers and patients. Magnetic resonance in medicine, 66(4), pp.1079-1088.
4. Markl, M., Chan, F.P., Alley, M.T., Wedding, K.L., Draney, M.T., Elkins, C.J., Parker, D.W., Wicker, R., Taylor, C.A., Herfkens, R.J. and Pelc, N.J., 2003. Time-resolved three-dimensional phase-contrast MRI. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine, 17(4), pp.499-506.
5. Roes, S.D., Hammer, S., van der Geest, R.J., Marsan, N.A., Bax, J.J., Lamb, H.J., Reiber, J.H., de Roos, A. and Westenberg, J.J., 2009. Flow assessment through four heart valves simultaneously using 3-dimensional 3-directional velocity-encoded magnetic resonance imaging with retrospective valve tracking in healthy volunteers and patients with valvular regurgitation. Investigative radiology, 44(10), pp.669-675.
6. Ridgway, J.P., 2010. Cardiovascular magnetic resonance physics for clinicians: part I. Journal of cardiovascular magnetic resonance, 12(1), pp.1-28.
7. Markl, M., Alley, M.T. and Pelc, N.J., 2003. Balanced phase-contrast steady-state free precession (PC-SSFP): a novel technique for velocity encoding by gradient inversion. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, 49(5), pp.945-952.
8. Nielsen, J.F. and Nayak, K.S., 2009. Referenceless phase velocity mapping using balanced SSFP. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, 61(5), pp.1096-1102.
9. Biffi, B., Bruse, J.L., Zuluaga, M.A., Ntsinjana, H.N., Taylor, A.M. and Schievano, S., 2017. investigating cardiac Motion Patterns Using synthetic high-resolution 3D cardiovascular Magnetic resonance images and statistical shape analysis. Frontiers in pediatrics, 5, p.34.
10. Qin, C., Bai, W., Schlemper, J., Petersen, S.E., Piechnik, S.K., Neubauer, S. and Rueckert, D., 2018, September. Joint learning of motion estimation and segmentation for cardiac MR image sequences. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 472-480). Springer, Cham.

11. Yu, H., Chen, X., Shi, H., Chen, T., Huang, T.S. and Sun, S., 2020, October. Motion pyramid networks for accurate and efficient cardiac motion estimation. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 436-446). Springer, Cham.
12. Rathi, V.K., Doyle, M., Yamrozik, J., Williams, R.B., Caruppannan, K., Truman, C., Vido, D. and Biederman, R.W., 2008. Routine evaluation of left ventricular diastolic function by cardiovascular magnetic resonance: a practical approach. Journal of Cardiovascular Magnetic Resonance, 10(1), pp.1-9.
13. Rubinshtein, R., Glockner, J.F., Feng, D., Araoz, P.A., Kirsch, J., Syed, I.S. and Oh, J.K., 2009. Comparison of magnetic resonance imaging versus Doppler echocardiography for the evaluation of left ventricular diastolic function in patients with cardiac amyloidosis. The American journal of cardiology, 103(5), pp.718-723.
14. Bollache, E., Redheuil, A., Clément-Guinaudeau, S., Defrance, C., Perdrix, L., Ladouceur, M., Lefort, M., De Cesare, A., Herment, A., Diebold, B. and Mousseaux, E., 2010. Automated left ventricular diastolic function evaluation from phase-contrast cardiovascular magnetic resonance and comparison with Doppler echocardiography. Journal of Cardiovascular Magnetic Resonance, 12(1), pp.1-11.
15. Garg, P., Westenberg, J.J., van den Boogaard, P.J., Swoboda, P.P., Aziz, R., Foley, J.R., Fent, G.J., Tyl, F.G.J., Coratella, L., ElBaz, M.S. and Van Der Geest, R.J., 2018. Comparison of fast acquisition strategies in whole-heart four-dimensional flow cardiac MR: Two-center, 1.5 Tesla, phantom and in vivo validation study. Journal of Magnetic Resonance Imaging, 47(1), pp.272-281.
16. Sun, X., Garg, P., Plein, S. and van der Geest, R.J., 2021. SAUN: Stack attention U-Net for left ventricle segmentation from cardiac cine magnetic resonance imaging. Medical Physics, 48(4), pp.1750-1763.
17. Ronneberger, O., Fischer, P. and Brox, T., 2015, October. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.
18. He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
19. Wang, L., Lee, C.Y., Tu, Z. and Lazebnik, S., 2015. Training deeper convolutional networks with deep supervision. arXiv preprint arXiv:1505.02496.
20. Mottram, P.M. and Marwick, T.H., 2005. Assessment of diastolic function: what the general cardiologist needs to know. Heart, 91(5), pp.681-695.
21. Assi, K.C., Gay, E., Chnafa, C., Mendez, S., Nicoud, F., Abascal, J.F., Lantelme, P., Tournoux, F. and Garcia, D., 2017. Intraventricular vector flow mapping—A

Doppler-based regularized problem with automatic model selection. Physics in Medicine & Biology, 62(17), p.7131.

22. Vos, H.J., Voorneveld, J.D., Jebbink, E.G., Leow, C.H., Nie, L., van den Bosch, A.E., Tang, M.X., Freear, S. and Bosch, J.G., 2020. Contrast-enhanced high-frame-rate ultrasound imaging of flow patterns in cardiac chambers and deep vessels. Ultrasound in Medicine & Biology, 46(11), pp.2875-2890.
23. Avesani, M., Degrelle, B., Di Salvo, G., Thambo, J.B. and Iriart, X., 2021. Vector flow mapping: A review from theory to practice. Echocardiography, 38(8), pp.1405-1413.

Supplementary

1. Evaluation metrics

As the third error metric we quantified the "accuracy of the positions" of the pixels with velocities higher than a given threshold. For this, we used the accuracy metric as defined in formula S.1.

$$\text{Accuracy} = \frac{\|G \cap P\|}{\|G\|} \tag{S7.1}$$

where set $\mathrm{G} = \{(i,j) | \|V_g(i,j)\| \geq g_p\}$ and set $\mathrm{P} = \{(i,j) | \|V_p(i,j)\| \geq p_p\}$ contain the pixels whose resultant velocities V are greater than a certain threshold. The threshold g_p and p_p are the pth percentile of the resultant velocity of ground truth and prediction, respectively.

2. Results of input dimension

Table S7.1. Prediction results generated using different input dimensions and different velocity thresholds in 4CH and 2CH view. For EPE and angle error, >5 indicates that only the pixels in the ground truth with a velocity magnitude greater than 5cm/s are included to compute the metrics. ACC >30th indicates that only the pixels with a velocity magnitude greater than 30th percentile of all four chambers (LV,RV,LA, RA) in 4CH view and all two chambers (LV, LA) in 2CH view are included to compute the accuracy. *N* is the dimension of the network input. ACC means the evaluation metric accuracy. The best results within four different dimensions are shown in bold.

View		EPE(cm/s)			Angle Error(°)			ACC(%)		
		>0	>5	>10	>0	>5	>10	>30th	>50th	>70th
4CH	N=3	7.0±1.5	8.7±2.7	10.9±2.3	51.9±9.9	41.9±11.2	33.2±11.0	79.0±4.0	68.0±7.5	55.9±12.1
	N=5	7.0±1.5	8.7±2.6	10.8±2.3	51.7±10.0	41.6±11.3	32.9±11.2	**79.2±3.9**	**68.4±7.5**	**56.4±11.9**
	N=7	**6.9±1.5**	**8.6±2.1**	10.9±2.4	51.7±9.9	41.6±11.1	33.0±10.7	78.9±4.1	68.1±7.6	56.3±12.2
	N=9	6.9±1.5	8.6±2.7	**10.8±2.4**	**51.4±10.1**	**41.3±11.4**	**32.7±11.0**	79.1±4.0	68.2±7.7	56.3±12.1
2CH	N=3	7.2±1.8	9.1±2.0	11.8±2.6	56.9±10.9	47.2±11.9	36.5±12.4	78.8±4.9	66.7±9.5	53.5±16.0
	N=5	7.1±1.80	9.0±2.0	11.7±2.6	56.5±11.1	46.7±12.0	36.4±12.0	78.9±5.0	66.85±9.6	53.7±16.1
	N=7	**7.1±1.8**	9.0±2.0	**11.6±2.5**	**56.4±11.1**	46.6±12.0	36.4±12.4	78.8±5.0	66.9±9.6	53.8±16.2
	N=9	7.1±1.8	**8.9±2.0**	11.7±2.6	56.5±11.6	**46.4±12.3**	**35.9±12.6**	**79.0±4.9**	**67.0±9.4**	**54.1±15.8**

3. Results in four-chamber view

Intra-cardiac flow velocity varies greatly within the cardiac cycle, across regions, cardiac phases and also across patients. Hence, to further analyze the prediction results, various velocity thresholds were used to compute the evaluation metrics for those pixels exceeding a chosen threshold (as shown in Fig.S7.1 and Fig.S7.4). By

excluding the low-velocity pixels, the performance of the model can be more clearly revealed.

Table S7.2. Prediction results of Accuracy in different chambers in 4CH view. Accuracy was computed using the 30th percentile as the threshold. 4CH indicates the results were computed within all 4 chambers; LV, LA, RV and RA mean the results were based on each single chamber separately. The mean ± standard deviation are reported.

	4CH	LV	LA	RV	RA
Accuracy (%)	79.09±4.02	79.40±5.17	78.72±6.38	77.77±5.75	75.62±5.91

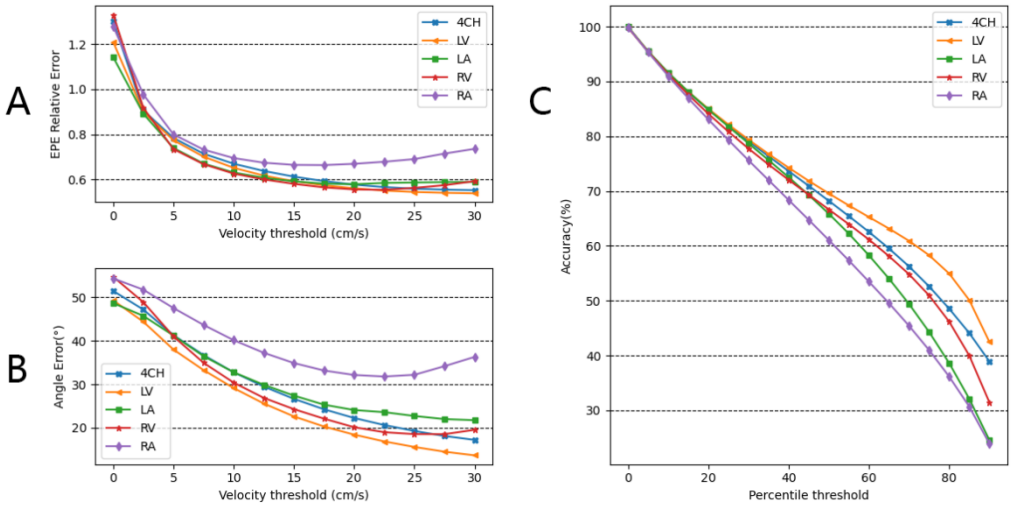


Figure.S7.1. Relative EPE, angle error and accuracy under different threshold values in different chambers in 4CH view. 4CH means all four chambers are included to compute the evaluation metrics. LV, LA, RV and RA means only one chamber was used to compute the metrics. (**A):** The relation between relative EPE, and the velocity threshold. (**B):** The relation between angle error and velocity threshold. (**C):** The relation between the accuracy and the velocity percentile threshold in different chambers.

It defined E/A ratio<0.5 as impaired relaxation pattern, 0.75< E/A ratio <1.5 as normal diastolic function and E/A ratio >2 as restrictive filling. The confusion matrix of the diastolic function classification experiment are summarized in Figure S7.2. The diastolic function classification accuracy was 88.1%.

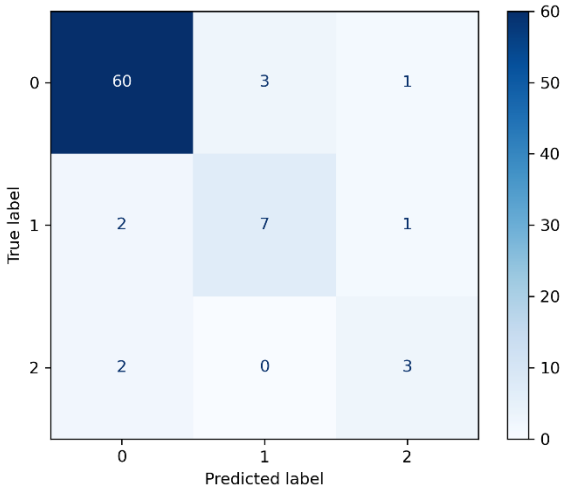


Figure.S7.2. Confusion matrix derived from the predicted velocities in the 4CH views. Label 0 means normal diastolic function, 1 is restrictive filling and 2 is impaired relaxation pattern.

We also test the performance of our model in two-chamber view. 86 cine 2CH views (2580 2D images) were used for training and testing.

4. Results in two-chamber view

Table.S7.3. Prediction results of different chambers in 2CH view. EPE and angle error were computed using a velocity threshold of 5 cm/s. Accuracy was computed using the 30th percentile as the threshold. 2CH indicates the results were computed within all 2 chambers; LV, LA mean the results were based on each single chamber separately. PCC: Pearson correlation coefficient. The mean ± standard deviation are reported.

	2CH	LV	LA
EPE (cm/s)	8.99±2.02	9.18±2.15	8.78±2.40
Angle Error (°)	46.45±12.26	45.19±13.60	48.91±16.12
Accuracy (%)	79.03±4.93	79.90±5.72	76.21±6.76
Velocity-RE (%)	-32.29±4.08	-31.46±4.68	-33.12±8.25
Velocity-PCC	0.971	0.984	0.869

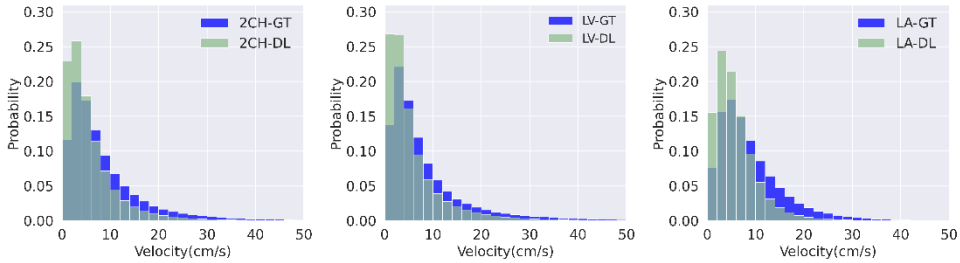


Figure.S7.3. Probability distribution of velocity generated from 4D flow data and prediction in 2CH view. The blue color represents the distribution generated from the 4D flow data, and the light green means the distribution generated from the prediction. The light blue represents the overlap between the prediction and 4D flow data.

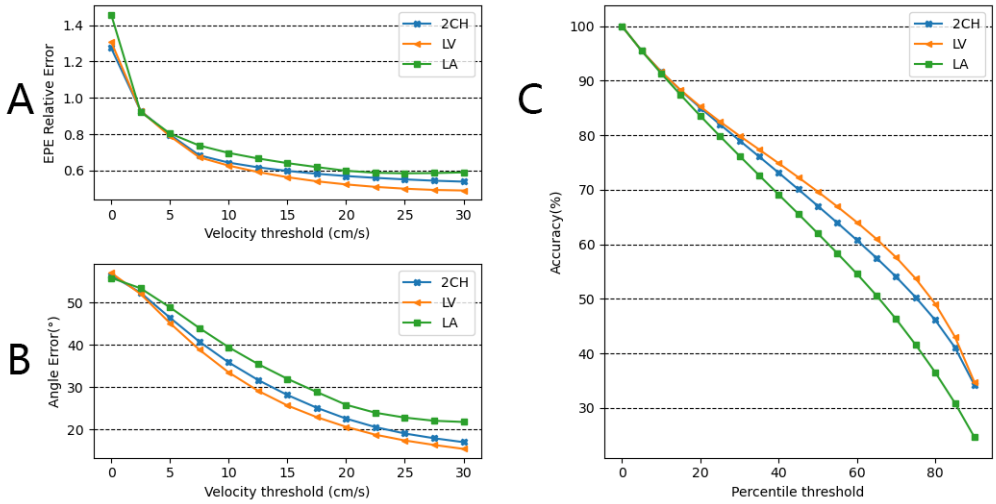


Figure.S7.4. Relative EPE, angle error and accuracy under various thresholds in different chambers in 2CH view. 2CH means LV and LA are included to compute the evaluation metrics. LV, LA means only one chamber was used to calculate the metrics. (**A):** The relation between relative EPE and velocity threshold. (**B):** The relation between angle error and velocity thresholds. (**C):** The relation between the accuracy and the velocity percentile threshold in different chambers.

The average absolute error in E/A ratio estimation in 2CH view was 0.46±0.42. In the 2CH view, there are 47 subjects with normal diastolic function, of those 47 subjects, seven were classified as having restrictive filling and six as having impaired relaxation. Two out of eleven subjects with restrictive filling were classified as normal diastolic function. The confusion matrix of the diastolic function classification experiment are summarized in Figure S7.6. The classification accuracy in 2CH view was 73.3%. The Wilcoxon signed-rank test with P=.67 in 2CH view,

confirmed that the E/A ratio generated from our prediction was not significantly different from the 4D flow data.

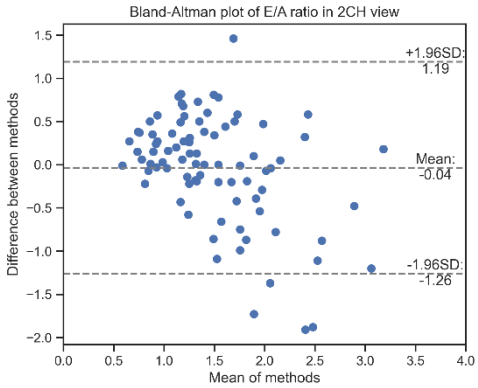


Figure.S7.5. Confusion matrix derived from the predicted velocities in the 2CH (right) views. Label 0 means normal diastolic function, 1 is restrictive filling and 2 is impaired relaxation pattern.

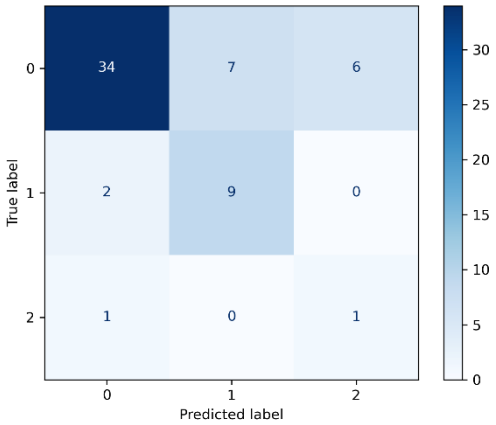


Figure.S7.6. Confusion matrix derived from the predicted velocities in the 2CH views. Label 0 means normal diastolic function, 1 is restrictive filling and 2 is impaired relaxation pattern.

Chapter 8 Summary and future work

Cine and 4D flow cardiac MRI are two important non-invasive MR imaging techniques to assess cardiac function and diagnose cardiovascular diseases. Cine MRI offers great soft tissue detail which allows clinical experts to evaluate structure and function of the heart. 4D flow MRI further has the ability of three-dimensional time-resolved acquisition of blood flow velocity, which can be used to derive intra-cardiac hemodynamic parameters. In this thesis, we developed deep learning-based approaches to analyze cine and 4D flow cardiac MRI. In this chapter, we summarize the previous chapters and discuss potential directions of future work.

8.1 Summary

In **Chapter 1**, we provided a general introduction about cine and 4D flow cardiac MRI and deep learning applications in the field of cardiac MRI. In Chapter 2, we proposed a sampling inspection network combining specially designed data augmentation methods to assess CMR image quality. The proposed method showed a competitive performance against the other methods in the CMRxMotion challenge. In Chapter 3, we proposed temporal and spatial stacks to incorporate temporal or spatial information using stack attention mechanism for left ventricle segmentation in short-axis cine MRI. In Chapter 4, we further studied the concept of domain generalization in the setting of right ventricle segmentation in unseen datasets, such as data with differences in acquisition protocol, across different centers, scanner vendors and diseases. In Chapter 5, we investigated the feasibility of using deep learning-based approaches to segment the LV directly from 4D flow MRI and explored the performance of integrating features extracted from magnitude and velocity images. A transformer based feature fusion model was developed to improve the performance of LV segmentation from 4D flow MRI in Chapter 6. Chapter 7 aimed to train a CNN model to predict blood flow velocity from long-axis cine MRI using the corresponding 4D flow data as ground truth.

Chapter 2 CMR may suffer from motion-related artifacts resulting in non-diagnostic quality images. Visual inspection of image quality is time-consuming and also relies on experienced radiologists. In this chapter, we proposed an automatic method for CMR image quality assessment. Given limited data and an unbalanced class ratio, we proposed three specially designed data augmentation methods to enlarge the dataset including generating transition phases between ED and ES phases, generating images using different levels of respiratory motion and generating images using histogram matching and linear interpolation. To mimic the sampling inspection, we randomly take two subsamples from one 3D volume to estimate the quality of a 3D volume. In the developed model, which was adapted from ResNet,

channel attention is used to explore the intra-channel relationship for the features extracted from each subsample. Subsequently, a feature fusion module is introduced to fuse features from two subsamples to predict the image quality. The proposed method is validated in the 2022 CMRxMotion competition, achieving a mean accuracy of 75% and 72.5% in training and validation dataset, respectively. Additionally, our method ranked at the 4th place in the testing dataset which was hidden by the organizer.

Chapter 3 In this chapter, we leveraged the spatiotemporal information from neighboring slices to improve the segmentation accuracy. The target image is stacked with its spatial or temporal neighboring images as the input. Then a stack attention is developed to extract and weigh the relevant features using the target image as a guide. The stack attention is inserted into U-Net to automatically segment the LV and myocardium from multi-slice short-axis cardiac MRI. An internal data set from one center and one public data set of the 2017 Automated Cardiac Segmentation Challenge (ACDC) were involved in evaluating and validating the proposed method. The model is trained on the internal data set first and then fine-tuned on the public data set. The method achieved a Dice of 0.91 and Hausdorff Distance of 3.37 mm on the in-house data set. The performance on the ACDC data set achieved a Dice of 0.92, 0.89 and Hausdorff Distance of 9.7 mm and 7.1 mm on ED and ED phases, respectively, which confirms a good generalization. Additionally, the results in both data sets show high correlation of LVEF and myocardium mass derived from the model and manual segmentation, demonstrating a potential valuable application in clinical practice.

Chapter 4 This chapter focuses on model generalization, in which the aim is to develop a model that performs well on unseen data sets from different centers, vendors or different diseases. The M&Ms-2 Challenge is motivated to segment the right ventricle based on a multi-disease, multi-view and multi-center samples of 360 cardiac MRI datasets. The most straightforward approach to tackle this problem is to collect more data to train a model. Given limited labeled data, we first introduce an intensity-based registration method to propagate the available labels from the end-diastolic and end-systolic phases to the other unlabeled phases. We subsequently investigate the performance of different input modalities including single 2D image, multi-channel 2D image and 3D volume. The multi-channel 2D image is constructed using the spatial and temporal stack proposed in Chapter 3. On the validation data set, our method achieved a Dice of 0.92 and 0.92, Hausdorff Distance of 9.5 mm and 5.3 mm in short-axis and long-axis view, respectively. Our method also generates a good performance on the hidden testing dataset, yielding a Dice of 0.93, 0.92 and Hausdorff Distance of 10.6 mm, 6.0 mm in short-axis and long-axis view, respectively. The experimental results demonstrate that the multi-channel 2D image

provides more information for the segmentation. Combining volume input and label propagation can further improve the generalization ability.

Previously reported 4D flow segmentation approaches rely on the registration between cine MRI and 4D flow data, which requires high computational cost. **Chapter 5** and **Chapter 6** focus on LV segmentation directly from 4D flow MRI without being dependent on additional cine MRI. In **Chapter 5**, we explored using the combination of magnitude and velocity images together in 4D flow data as input. The poor contrast between the heart chambers and myocardium will result in inherent uncertainty in the segmentation results. Therefore, Monte Carlo dropout method is introduced to assess the segmentation uncertainty. Additionally, five deep learning based models are compared to investigate the effect of using different network architectures, data pre-processing, inputs and feature fusion methods on the segmentation performance. Based on the results, the proposed method was shown to be highly accurate. Additionally, the clinical parameters derived from the best model show a high correlation with results derived from manual annotations, confirming the feasibility of LV segmentation from 4D flow MRI directly.

Chapter 6 presents a transformer based efficient feature fusion method to fuse the information from magnitude and velocity images and to improve the segmentation performance in 4D flow MRI. The network is an encoder-decoder structure based on U-Net. In the encoder, the magnitude and velocity images are considered as the inputs of two branches separately. The features from the same level are integrated using the feature fusion module. The cross- and self-fusion layer in the feature fusion module aim to explore the inter- and intra-relationship between those features. The fused features are added into the original features. The paired multi-level features are concatenated along the channel dimension followed by a convolutional layer as the input of the decoder. The decoder is kept the same as that in U-Net. The proposed methods achieve the best performance compared to the other models and get significant improvement in clinical parameters, yielding a Pearson correlation coefficient of 83.3%, 97.4%, 96.97% and 98.92% for LVEF, EDV, ESV and KE, respectively. The proposed feature fusion method therefore facilitates to aggregate the features from different modalities in an efficient manner.

Chapter 7 In this chapter, we designed and evaluated a deep learning based method to predict the intra-cardiac blood flow pattern from long-axis cine MRI using the velocities derived from 4D flow data as the ground truth. The network, a variant of U-Net and ResNet, takes a subsequence of cine MR images as the input to extract the displacement of blood over the cardiac frames. Although the averaged predicted velocity was shown to be under-estimated by 26.69%, the global time-varying blood flow pattern shows a high correlation with the 4D flow derived velocities. A potential application of the proposed method is to estimate the E/A ratio. The results indicated

that the E/A ratio can be estimated without significant bias and can further classify the diastolic function with a high accuracy. Our study is the first to employ deep learning for blood flow prediction from cine MRI. After further improvement of the model this work could potentially be valuable in clinical applications to visualize the intra-cardiac blood flow without additional 4D flow data.

8.2 Discussion and Future work

The work presented in this thesis aims to develop deep learning based methods for automated analysis of cardiac cine and 4D flow MRI.

The networks developed in chapter 3 and 4 focused primarily on automated segmentation in cine MRI. In chapter 3, we demonstrated that extracting temporal or spatial information from neighboring slices can benefit the segmentation performance in short-axis view cine MRI. The performance derived from introducing spatial features is better than using temporal information. Because the spatial stack can provide more information about the position, size and shape of the heart. While the images in the temporal stack are similar to each other and contain comparable features. In these studies deep learning has shown its promising applications in cardiac MRI segmentation. However, the developed approaches are validated in cases where the testing data is from the same domain as the training data. In a realistic scenario an significant performance drop can be observed when a trained model is applied on data from another domain. For example, when our model trained on the Leeds University dataset (LUD) is applied to the ACDC dataset directly, the segmentation accuracy drops from 90% to 70%. This can be explained by the population bias from different sites, ages, genders, races and diseases, and image appearance differences from various vendors, protocols, and magnetic field strengths resulting in data distribution heterogeneity. The heterogeneity cannot be eliminated completely using data pre-processing. The model needs to be fine-tuned or re-trained on the new data set to achieve a good performance. Therefore, domain generalization is a technical bottleneck for deploying deep learning in real-world clinical environments. Collecting and labeling vast amount of data from various centers and vendors it the most straight-forward solution. However, it is prohibitively expensive to obtain high-quality manual annotations for every domain, as it requires expert knowledge and it is also impossible to cover full spectrum of data. In chapter 4 we introduced registration to propagate the available segmentation labels to unlabeled images in order to enlarge the training data. Additionally, data augmentation techniques are used to increase the variety of training data to improve the model's robustness. The results show that the model has a good generalization on data with unseen pathologies. As a promising direction of future research, it's also worth investigating self-learning and semi-supervised learning to extract more

prior knowledge to improve model's generalization ability when the training data is limited.

In chapter 5, we compared several models for LV segmentation directly from 4D flow MRI without relying on the registration between cine and 4D flow MRI. In chapter 6, we further improve the performance using a novel feature fusion method. However, there are also other considerations that need to be taken into account. Firstly, although most studies focus on introducing novel algorithms, data pre-processing including correction, enhancement, resample and normalization is also significantly important. For example, as shown in chapter 5 the performance derived from resliced data volumes is better than using the original data without data pre-processing. Similarly, the model of nnUnet (no new Unet), a self-configuring method, surpasses most existing deep learning-based segmentation approaches on 23 public datasets. The strong performance is not achieved by introducing a new network architecture for each type of data, but is the result of the carefully designed process of automatic self-configuration. Secondly, it is important to be aware that the final evaluation measurements should be valuable and reliable for the quantitative assessment of model's performance. The proposed models in Chapter 5 achieved similar results in terms of Dice and ASD as reported in Table 5.2, which makes it difficult to select the best model. In general, clinical relevant metrics derived from the segmentation results provide meaningful and actionable information for diagnosis and treatment. As compared to the other state-of-arts, the proposed method in Chapter 6 improves the Dice by only 2%, but the Pearson correlation coefficient in EDV, ESV and KE got improved by 9%, 7% and 16%. Therefore, the clinical parameters should be involved when comparing the performance of different models to ensure that the algorithms are reliable for use in medical applications. It would be possible to develop deep learning based multi-task networks to jointly perform the task of cardiac segmentation task and the regression of volume or ejection fraction prediction. Thirdly, in chapter 6 a transformer based feature fusion module was presented and achieved the best performance. The module can integrate information extracted from two different modalities or views efficiently. It can be adapted to the other applications such as integrating short- and long-axis view cine MRI for disease diagnosis, or combining apical four chamber (A4C) and two chamber (A2C) acquisitions in echocardiography data.

8.3 General conclusions

In conclusion, this thesis proposes deep learning based methods for quantifying cardiac MRI. The described methods can be applied for cine MR image quality classification and ventricle segmentation without any human interactions. Investigating combining and fusing magnitude and velocity images can be helpful for left ventricle segmentation in 4D flow MRI, which is not fully explored yet.

Moreover, we proposed a network to predict the blood flow pattern from the cine MRI. By combining visualization of the blood flow and myocardial motion in the routinely acquired standard CMR exams, the method can be potentially used in clinical studies. All the deep learning methods described in this thesis were evaluated on MRI data, but can potentially also be applied in other imaging modalities such as computed tomography and echocardiography.

Samenvatting en toekomstig werk

Cine en 4D flow cardiale MRI zijn twee belangrijke niet-invasieve MR-beeldvormingstechnieken om de hartfunctie te beoordelen en cardiovasculaire ziekten te diagnosticeren. Cine MRI biedt grote details van het zachte weefsel, waardoor klinische deskundigen de structuur en functie van het hart kunnen beoordelen. 4D flow MRI heeft verder de mogelijkheid tot het maken van driedimensionale tijdsopnamen van de bloedstroomsnelheid, die kan worden gebruikt om intra-cardiale hemodynamische parameters af te leiden. In dit proefschrift hebben we op deep learning gebaseerde benaderingen ontwikkeld voor het analyseren van cine en 4D flow cardiale MRI. In dit hoofdstuk vatten we de voorgaande hoofdstukken samen en bespreken we mogelijke richtingen voor toekomstig werk.

Samenvatting

In **hoofdstuk 1** hebben we een algemene inleiding gegeven over cine en 4D flow cardiale MRI en deep learning toepassingen op het gebied van cardiale MRI. In hoofdstuk 2 hebben we een netwerk voor bemonsteringsinspectie voorgesteld dat speciaal ontworpen methoden voor datavergroting combineert om de CMR-beeldkwaliteit te beoordelen. De voorgestelde methode presteerde concurrerend ten opzichte van de andere methoden in de CMRxMotion challenge. In hoofdstuk 3 hebben we temporele en ruimtelijke stacks voorgesteld om temporele of ruimtelijke informatie op te nemen met behulp van het stack-aandachtsmechanisme voor segmentatie van de linkerventrikel in korte-as cine MRI beelden. In hoofdstuk 4 hebben we het concept van domein generalisatie verder bestudeerd in de setting van rechterventrikel segmentatie in ongeziene datasets, zoals data met verschillen in acquisitie protocol, over verschillende centra, scanner leveranciers en ziekte beelden. In hoofdstuk 5 onderzochten wij de haalbaarheid van op deep learning gebaseerde benaderingen om de LV rechtstreeks te segmenteren op basis van 4D flow-MRI en onderzochten wij de prestaties van de integratie van kenmerken die werden geëxtraheerd uit magnitude- en snelheidsbeelden. In hoofdstuk 6 werd een op *Transformers* gebaseerd feature fusie model ontwikkeld om de prestaties van LV-segmentatie van 4D flow MRI te verbeteren. In hoofdstuk 7 werd een CNN-model getraind om de bloedstroomsnelheid te voorspellen op basis van lange-as cine-MRI beelden, waarbij de corresponderende 4D-flow gegevens als referentie werden gebruikt.

Hoofdstuk 2 CMR kan last hebben van bewegingsgerelateerde artefacten die resulteren in beelden van niet-diagnostische kwaliteit. Visuele inspectie van de beeldkwaliteit is tijdrovend en bovendien afhankelijk van ervaren radiologen. In dit

hoofdstuk stellen we een automatische methode voor om de kwaliteit van CMR-beelden te beoordelen. Met beperkte hoeveelheid data en een onevenwichtige klassenverhouding hebben wij drie speciaal ontworpen datavergrotingsmethoden voorgesteld om de dataset uit te breiden, waaronder het genereren van overgangsfasen tussen de ED- en ES-fasen, het genereren van beelden met verschillende mate van ademhalingsbeweging en het genereren van beelden met behulp van histogrammatching en lineaire interpolatie. Om de bemonsteringsinspectie na te bootsen, nemen we willekeurig twee deelmonsters van één 3D-volume om de kwaliteit van een 3D-volume te schatten. In het ontwikkelde model, dat werd aangepast aan ResNet, wordt kanaalaandacht gebruikt om de intra-kanaalrelatie te onderzoeken voor de kenmerken die uit elk deelmonster worden geëxtraheerd. Vervolgens wordt een feature fusie module geïntroduceerd om features van twee subsamples te fuseren om de beeldkwaliteit te voorspellen. De voorgestelde methode is gevalideerd in de CMRxMotion-wedstrijd van 2022 en behaalde een gemiddelde nauwkeurigheid van 75% en 72,5% in respectievelijk de training- en validatiedataset. Bovendien eindigde onze methode op de vierde plaats in de testdataset die door de organisator verborgen was gehouden.

Hoofdstuk 3 In dit hoofdstuk gebruiken we de spatio-temporele informatie van naburige beelden om de nauwkeurigheid van segmentatie te verbeteren. Het doelbeeld wordt gestapeld met zijn ruimtelijke of temporele naburige beelden als input. Vervolgens wordt een stapel aandacht ontwikkeld om de relevante kenmerken te extraheren en te wegen met het doelbeeld als leidraad. De stackaandacht wordt in U-Net ingevoegd om automatisch de LV en het myocard te segmenteren uit multi-slice korte-as cardiale MRI. Een interne dataset van één centrum en een openbare dataset van de 2017 Automated Cardiac Segmentation Challenge (ACDC) werden betrokken bij het evalueren en valideren van de voorgestelde methode. Het model is eerst getraind op de interne dataset en vervolgens verfijnd op de openbare dataset. De methode behaalde een Dice van 0,91 en een Hausdorff Distance van 3,37 mm op de interne dataset. De prestaties op de ACDC-dataset bereikten een Dice van 0,92 en 0,89 en een Hausdorff Distance van 9,7 mm en 7,1 mm voor respectievelijk de ED- en ES-fasen, wat een goede generalisatie bevestigt. Bovendien laten de resultaten in beide datasets een hoge correlatie zien van LVEF en myocardmassa afgeleid van het model en handmatige segmentatie, wat een potentieel waardevolle toepassing in de klinische praktijk aantoont.

Hoofdstuk 4 Dit hoofdstuk richt zich op modelgeneralisatie, waarbij het doel is een model te ontwikkelen dat goed presteert op ongeziene datasets van verschillende centra, leveranciers of verschillende ziekten. De M&Ms-2-*challenge* is gemotiveerd om de rechterhartkamer te segmenteren op basis van een steekproef van 360 cardiale MRI-datasets met meerdere ziekten, meerdere beeld oriëntaties en meerdere centra. De meest eenvoudige aanpak van dit probleem is het verzamelen van meer data om

een model te trainen. Met beperkte gelabelde data introduceren we eerst een op intensiteit gebaseerde registratiemethode om de beschikbare labels van de eind-diastolische (ED) en eind-systolische (ES) fasen te propageren naar de andere ongelabelde fasen. Vervolgens onderzoeken wij de prestaties van verschillende invoermodaliteiten, waaronder een enkel 2D-beeld, een meerkanaals 2D-beeld en een 3D-volume. Het meerkanaals 2D-beeld wordt opgebouwd met behulp van de in hoofdstuk 3 voorgestelde ruimtelijke en temporele stapeling. Op de validatiedataset behaalde onze methode een Dice van 0,92 en 0,92, Hausdorff Distance van 9,5 mm en 5,3 mm in respectievelijk de korte en lange as beelden. Onze methode levert ook goede prestaties op de verborgen testdataset, met een Dice van 0,93 en 0,92 en een Hausdorff Distance van 10,6 mm en 6,0 mm in respectievelijk korte- en lange-as aanzicht. De experimentele resultaten tonen aan dat het meerkanaals 2D-beeld meer informatie biedt voor de segmentatie. De combinatie van volume-invoer en labelpropagatie kan het generalisatievermogen verder verbeteren.

Eerder gerapporteerde benaderingen voor 4D-flow segmentatie zijn gebaseerd op de registratie tussen cine-MRI en 4D-flow gegevens, wat hoge rekencapaciteit vereist. Hoofdstuk 5 en Hoofdstuk 6 richten zich op LV-segmentatie direct vanuit 4D flow MRI zonder afhankelijk te zijn van aanvullende cine MRI. In **Hoofdstuk 5**, hebben we onderzocht met behulp van de combinatie van magnitude- en snelheidsbeelden samen met 4D-flow gegevens als invoer. Het slechte contrast tussen de hartkamers en het myocard zal resulteren in inherente onzekerheid in de segmentatieresultaten. Daarom wordt de Monte Carlo-uitvalmethode geïntroduceerd om de segmentatie-onzekerheid te beoordelen. Daarnaast worden vijf op deep learning gebaseerde modellen vergeleken om het effect te onderzoeken van het gebruik van verschillende netwerkarchitecturen, datavoorverwerking, invoer en functiefusiemethoden op de segmentatieprestaties. Op basis van de resultaten bleek de voorgestelde methode zeer nauwkeurig te zijn. Bovendien vertonen de klinische parameters die zijn afgeleid van het beste model een hoge correlatie met resultaten die zijn afgeleid van handmatige annotaties, wat de haalbaarheid van LV-segmentatie rechtstreeks uit 4D flow MRI bevestigt.

Hoofdstuk 6 presenteert een op een *Transformer* gebaseerde efficiënte kenmerkfusiemethode om de informatie uit magnitude- en snelheidsbeelden te fuseren en om de segmentatieprestaties in 4D flow MRI te verbeteren. Het netwerk is een encoder-decoderstructuur gebaseerd op U-Net. In de encoder worden de magnitude- en snelheidsbeelden beschouwd als de invoer van twee afzonderlijke takken. De functies van hetzelfde niveau zijn geïntegreerd met behulp van de feature fusie-module. De *cross-* en *self-fusion*-laag in de feature fusie-module is bedoeld om de inter- en intra-relatie tussen die features te verkennen. De gefuseerde kenmerken worden toegevoegd aan de originele kenmerken. De gepaarde kenmerken op meerdere niveaus worden aaneengeschakeld langs de kanaaldimensie, gevolgd door

een convolutionele laag als invoer van de decoder. De decoder wordt hetzelfde gehouden als die in U-Net. De voorgestelde methoden leveren de beste prestaties in vergelijking met de andere modellen en zorgen voor een significante verbetering van de klinische parameters, resulterend in een Pearson-correlatiecoëfficiënt van respectievelijk 83,3%, 97,4%, 96,97% en 98,92% voor LVEF, EDV, ESV en KE. De voorgestelde feature fusie-methode maakt het daarom mogelijk om de features van verschillende modaliteiten op een efficiënte manier samen te voegen.

Hoofdstuk 7 In dit hoofdstuk hebben we een op deep learning gebaseerde methode ontworpen en geëvalueerd om het intra-cardiale bloedstroompatroon te voorspellen op basis van cine-MRI lange as opnamen, waarbij we de snelheden afgeleid van 4D-flow gegevens gebruiken als referentie. Het netwerk, een variant van U-Net en ResNet, neemt een reeks cine-MR-beelden als input om de verplaatsing van bloed over de cardiale frames te extraheren. Hoewel werd aangetoond dat de gemiddelde voorspelde snelheid met 26,69% werd onderschat, vertoont het globale in de tijd variërende bloedstroompatroon een hoge correlatie met de van de 4D-flow afgeleide snelheden. Een mogelijke toepassing van de voorgestelde methode is het schatten van de E/A-ratio. De resultaten gaven aan dat de E/A-ratio zonder significante bias kan worden geschat en daarnaast de diastolische functie kan classificeren met een hoge nauwkeurigheid. Onze studie is de eerste die deep learning gebruikt voor de voorspelling van de bloedstroom op basis van cine-MRI. Na verdere verbetering van het model zou dit werk potentieel waardevol kunnen zijn in klinische toepassingen om de intracardiale bloedstroom te visualiseren zonder aanvullende 4D-flow gegevens.

Discussie en toekomstig werk

Het werk dat in dit proefschrift wordt gepresenteerd, heeft tot doel op deep learning gebaseerde methoden te ontwikkelen voor geautomatiseerde analyse van cardiale cine en 4D flow MRI.

De in hoofdstuk 3 en 4 ontwikkelde netwerken waren voornamelijk gericht op geautomatiseerde segmentatie in cine MRI. In hoofdstuk 3 hebben wij aangetoond dat het extraheren van temporele of ruimtelijke informatie uit naburige coupes de segmentatieprestatie in korte-as cine MRI kan verbeteren. De prestaties van het gebruik van ruimtelijke kenmerken zijn beter dan die van temporele informatie. De ruimtelijke stack kan namelijk meer informatie verschaffen over de positie, grootte en vorm van het hart. Terwijl de beelden in de temporele stack op elkaar lijken en vergelijkbare kenmerken bevatten. In deze studies heeft deep learning zijn veelbelovende toepassingen in cardiale MRI-segmentatie aangetoond. De ontwikkelde benaderingen zijn echter gevalideerd in gevallen waarin de testdata afkomstig zijn uit hetzelfde domein als de trainingsdata. In een realistisch scenario

kan een aanzienlijke prestatiedaling worden waargenomen wanneer een getraind model wordt toegepast op data uit een ander domein. Wanneer ons op de Leeds University dataset (LUD) getrainde model bijvoorbeeld rechtstreeks wordt toegepast op de ACDC dataset, daalt de segmentatienauwkeurigheid van 90% naar 70%. Dit kan worden verklaard door de populatie verschillen van verschillende locaties, leeftijden, geslachten, rassen en pathologieën, en verschillen in beeldvorming door verschillende scanner leveranciers, protocollen en magnetische veldsterktes die resulteren in heterogeniteit van de dataverdeling. De heterogeniteit kan niet volledig worden geëlimineerd door voorbewerking van de data. Het model moet worden verfijnd of opnieuw worden getraind op de nieuwe datareeks om goede prestaties te bereiken. Daarom is domeingeneralisatie een technisch knelpunt voor de toepassing van deep learning in realistische klinische omgeving. Het verzamelen en labelen van een grote hoeveelheid data van verschillende centra en leveranciers is de meest eenvoudige oplossing. Het is echter kostentechnisch onhaalbaar om voor elk domein handmatige annotaties van hoge kwaliteit te verkrijgen, omdat daarvoor expert kennis nodig is, en het is ook onmogelijk om het volledige spectrum van data te bestrijken. In hoofdstuk 4 introduceerden wij registratie om de beschikbare segmentatielabels door te geven aan ongelabelde beelden om de hoeveelheid trainingsdata te vergroten. Bovendien worden technieken voor datauitbreiding gebruikt om de verscheidenheid aan trainingsdata te vergroten om de robuustheid van het model te verbeteren. De resultaten tonen aan dat het model goed generaliseert op data met ongeziene pathologieën. Als veelbelovende richting voor toekomstig onderzoek is het ook de moeite waard *selflearning* en semi-gesuperviseerd leren te onderzoeken om meer voorkennis te extraheren om het generalisatievermogen van het model te verbeteren wanneer de hoeveelheid trainingsdata beperkt is.

In hoofdstuk 5 vergelijken we verschillende modellen voor LV-segmentatie rechtstreeks uit 4D flow-MRI zonder gebruik te maken van de registratie tussen cine en 4D flow-MRI. In hoofdstuk 6 verbeteren we de prestaties verder met behulp van een nieuwe feature fusie methode. Er zijn echter ook andere overwegingen waarmee rekening moet worden gehouden. Ten eerste, hoewel de meeste studies zich richten op de invoering van nieuwe algoritmen, is de voorbewerking van de data, waaronder correctie, verbetering, resampling en normalisatie, ook van groot belang. Zo blijkt uit hoofdstuk 5 dat de prestaties van herschikte datavolumes beter zijn dan die van de oorspronkelijke data zonder datavoorbewerking. Evenzo overtreft het model van nnUnet (*no new Unet*), een zelfconfigurerende methode, de meeste bestaande op deep learning gebaseerde segmentatiebenaderingen op 23 openbare datasets. De goede prestaties worden niet bereikt door de invoering van een nieuwe netwerkarchitectuur voor elk type data, maar zijn het resultaat van het zorgvuldig ontworpen proces van automatische zelfconfiguratie. Ten tweede moeten de uiteindelijke evaluatiemetingen waardevol en betrouwbaar zijn voor de

kwantitatieve beoordeling van de prestaties van het model. De in hoofdstuk 5 voorgestelde modellen behaalden vergelijkbare resultaten voor wat betreft Dice en ASD, zoals gerapporteerd in tabel 5.2, wat het moeilijk maakt het beste model te selecteren. In het algemeen leveren de klinisch relevante metrieken, afgeleid van de segmentatieresultaten, zinvolle en bruikbare informatie op voor diagnose en behandeling. In vergelijking met de andere state-of-arts verbetert de voorgestelde methode in hoofdstuk 6 de Dice met slechts 2%, maar de Pearson correlatiecoëfficiënt in EDV, ESV en KE werd verbeterd met 9%, 7% en 16%. Daarom moeten de klinische parameters worden betrokken bij de vergelijking van de prestaties van verschillende modellen om ervoor te zorgen dat de algoritmen betrouwbaar zijn voor gebruik in medische toepassingen. De mogelijkheid bestaat om op deep learning gebaseerde *multi-task* netwerken te ontwikkelen om gezamenlijk de taak van hartsegmentatie en de regressie van volume of ejectiefractievoorspelling uit te voeren. Ten derde werd in hoofdstuk 6 een op *Transformers* gebaseerde feature fusie module gepresenteerd die de beste prestaties leverde. De module kan informatie uit twee verschillende modaliteiten of beeld oriëntaties efficiënt integreren. De module kan worden aangepast voor andere toepassingen, zoals de integratie van korte-as en lange-as cine-MRI voor automatische diagnosetoepassing, of de combinatie van apicale vier-kamer en twee-kamer opnames in echocardiografische beelddata.

Algemene conclusies

Concluderend stelt dit proefschrift deep learning-gebaseerde methoden voor om cardiale MRI te kwantificeren. De beschreven methoden kunnen worden toegepast voor cine MR-beeldkwaliteitsclassificatie en ventrikelsegmentatie zonder humane interactie. Onderzoek naar het combineren en samenvoegen van magnitude- en snelheidsbeelden kan nuttig zijn voor de segmentatie van de linkerventrikel in 4D flow-MRI, wat nog niet volledig is onderzocht. Bovendien hebben wij een netwerk voorgesteld om het bloedstroompatroon te voorspellen op basis van de cine MRI. Door de combinatie van visualisatie van de bloedstroom en myocardiale beweging in een routinematig verkregen standaard CMR-onderzoek, kan de methode mogelijk worden gebruikt in klinische studies. Alle in dit proefschrift beschreven deep learning-methoden werden geëvalueerd op MRI-beelddata, maar kunnen potentieel ook worden toegepast op andere beeldvormingsmodaliteiten zoals computertomografie en echocardiografie.

Publications

Journal articles

Xiaowu Sun, Pankaj Garg, Sven Plein, Rob J. van der Geest. SAUN: Stack attention U-Net for left ventricle segmentation from cardiac cine magnetic resonance imaging. Medical Physics, 48(4), 1750-1763.

Xiaowu Sun, Li-Hsin Cheng, Sven Plein, Pankaj Garg, Rob J. van der Geest. Deep learning based automated left ventricle segmentation and flow quantification in 4D flow cardiac MRI. Journal of Cardiovascular Magnetic Resonance (under review)

Xiaowu Sun, Li-Hsin Cheng, Sven Plein, Pankaj Garg, Mehdi H. Moghari, Rob J. van der Geest. Deep Learning-based Method for Intra-Cardiac Blood Flow Pattern Prediction using 4D Flow Data. International Journal of Cardiovascular Imaging. (2023): 1-9.

Conference proceedings

Xiaowu Sun, Li-Hsin Cheng, Rob J. van der Geest. Right Ventricle Segmentation via Registration and Multi-input Modalities in Cardiac Magnetic Resonance Imaging from Multi-disease, Multi-view and Multi-center. International Workshop on Statistical Atlases and Computational Models of the Heart (STACOM, **Oral**). Springer, Cham, 2021.

Xiaowu Sun, Li-Hsin Cheng, Sven Plein, Pankaj Garg, Rob J. van der Geest. Transformer based feature fusion for left ventricle segmentation in 4D flow MRI. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI, **Oral**). Springer, Cham, 2022.

Xiaowu Sun, Li-Hsin Cheng, Rob J. van der Geest. Combination Special Data Augmentation and Sampling Inspection Network for Cardiac Magnetic Resonance Imaging Quality Classification. International Workshop on Statistical Atlases and Computational Models of the Heart (STACOM). Springer, Cham, 2022.

Xiaowu Sun, Li-Hsin Cheng, Rob J. van der Geest. Self-and Cross-attention based Transformer for left ventricle segmentation in 4D flow MRI. Medical Imaging with Deep Learning (MIDL). 2022.

Li-Hsin Cheng, **Xiaowu Sun**, Rob J. van der Geest. Contrastive Learning for Echocardiographic View Integration. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI, **Oral**). Springer, Cham, 2022.

Acknowledgements

"*I am not a rich, smart or talented person in the world, but I'm simply an ordinary man who keeps going and going and going*".

I've been using this quotation to encourage myself during the past four and a half years. Being a PhD student outside of China is not easy. I could not have done it without the supporting and help from my supervisors, colleagues, friends and families along this journey. I am deeply grateful to each of you.

First and foremost, I would like to express my gratitude to my promotor Prof. Boudewijn Lelieveldt and my esteemed supervisor Rob van der Geest. Boudewijn, I sincerely appreciate your selfless assistance, unwavering support and the friendly and inclusive work environment that you have fostered.

Rob, throughout our time working together, your friendly and approachable demeanor, open-door policy and willingness to engage in meaningful conversations have been invaluable in helping me navigate complex tasks and projects. Also thank you for organizing the memorable boat trip in Leiden canals and the delightful BBQ in your house. Dankuwel!

I appreciate that I met those friendly and amazing colleagues in LKEB. Niels, it's my honor to have you as a colleague and friend. I enjoyed our funny daily talks and thank you for your concerns during the Covid pandemic. Li-Hsin, I am grateful for our discussions throughout our weekly group meetings, and it is unforgettable how we worked all night to accomplish the tight deadlines for MIDL and MICCAI. Baldur, your knowledge expands my horizons and I learned a lot from you during our in-depth talks. Marius, I appreciate your encouragement for my oral presentation at MICCAI in Singapore. Berend and Els, thank you for your wonderful concert performances. Oleh, thank you for sharing your trip experiences during LKEB beach outgoing. Michèle, you are always there when I need any IT supporting, also thank you for your "Daddy care" when I worked overtime in office. I would like to thank Patrick, Jeoren, Denis, Jouke and Alexander for your invitation for lunch every working day. Thanks to Jingnan, Li-Hsin, Viktor, Patrick, Yunjie, Yanli, Xiaotong, Chang, Ruochen, Chinmay, Mody, Mohamed, Laurens, Simon, Vincent, Konstantinos, Silvia, Bo, Zhiwei, Qing, Hessam, Sahar, Qian, Kilany and Tahereh, all the members of AI meeting for sharing your cutting-edge technologies. I also had a lot of fun with you guys, playing board games, travelling, hiking and bouldering. Of course, I'll never forget how awkwardly we run through Schiphol airport to catch the flight, only to miss it.

Special thankfulness goes to my teachers in China. Prof. Shengde Li, thank you for supervising my scientific contests and bachelor thesis, which is where I first experienced academic research. Your rigorous scientific research attitude also shows me how to be an excellent researcher. Prof. Linghua Kong, you were the one who enlightened and encouraged me when I wanted to drop out. Thank you for your selfless assistance tutoring in my mathematics professional competition. Prof. Lizhen Liu, your valuable advice pointed me in the right direction when I was standing at the crossroads.

I extend my sincere thanks to my dear friends. Qing, Lingling, Ling Lin, Zhiwei, Ningning, Kaixuan, Chenhong, Lu, Qian, Zexu, Wensen, Jiemiao and all Chinese PhDs in LKEB, those moments when we cooked, traveled and played games together are the most precious memories in my study abroad. 谢谢. A special gratitude to Ruizhe in Nottingham, it's a great treasure to meet you in UCL summer school. I appreciate your collaboration during the CMRxMotion Challenge. My volleyball teammates in SKC, thank you for all the hilarious moments on and off the field, you made my life in Leiden joyful.

Things are never quite as scary when you've got a best friend. Xiaofan, we encourage each other from undergrad to master, and then to our PhD studies. What a blessing it is to have a buddy like you for more than 10 years.

Last but most certainly not least, I would like to thank my families. 感谢姥姥、姥爷年逾八十，护我周全；感谢爸妈育我成人，焉得谖草，言树之背；姐姐、姐夫以及我亲爱的外甥昊昊，谢谢你们无微不至的关心，一如既往的支持，让我心无旁骛前行。爱你们, I love you all!

Curriculum Vitae

Xiaowu Sun was born in Yantai, Shandong Province, China in November,1992. In 2011, he started his bachelor in the major of Applied Mathematics at Dalian Ocean University, Liaoning Province, China. He won national scholarship awards (highest honor) in 2013 and 2014, respectively. In 2015, he received "Best Undergraduate Dissertation Award" and graduated as an "outstanding student" of Liaoning Province. At the same year, he began his master study in the major of software engineering at Capital Normal University, Beijing. In his master project, he developed machine learning based approaches to predict the protein complexes from protein-protein interaction network. In 2018, he got his master degree with "outstanding student".

From September 2018, he started his PhD study in the Division of Image Processing (LKEB) under the Department of Radiology at Leiden University Medical Center in the Netherlands. His PhD project mainly focuses on automated analysis for cardiac MRI using deep learning based methods. The results of his research are included in this thesis.

From May 2023, he worked as a post-doctoral researcher in EPFL, Lausanne, Switzerland, with the project of automated analysis of coronary angiography and cardiac ultrasound images using deep learning, under the supervision of Prof. Emmanuel Abbe and Prof. Pascal Frossard.