DEFORMABLE MODELS AND MACHINE LEARNING FOR LARGE-SCALE CARDIAC MRI IMAGE ANALYTICS

by

DONG YANG

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Computer Science

Written under the direction of

Dimitris N. Metaxas

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

May, 2019

ABSTRACT OF THE DISSERTATION

Deformable Models and Machine Learning for Large-Scale Cardiac MRI Image Analytics

By DONG YANG

Dissertation Director:

Dimitris N. Metaxas

The analysis of left ventricle (LV) wall motion is an important step for understanding cardiac functioning mechanisms, and clinical diagnosis of ventricular diseases. For example, ventricular dyssynchrony is one of the major causes for heart failure; treatment of dyssynchrony, e.g. *Cardiac Resynchronization Therapy* (CRT), can help some patients preventing failure. Conventional diagnosis methods, including *electrocardiogram* (ECG) and ultrasound imaging, provide only coarse characterization of dyssynchrony patterns, such as global function indices or qualitative assessment of motion patterns. To achieve a more comprehensive understanding of ventricular dyssynchrony, we propose a novel approach to study the regional patterns of left ventricle (LV) wall using cardiac magnetic resonance imaging (MRI). Firstly, we extract the myocardial contours from long- and short-axis cine MRI, and compensate for respiration offsets through rigid transformation to reconstruct the 3D shell of the heart wall. Then an unsupervised learning method using deep neural networks is adopted to compute the in-plane deformation field. Next, the 3D volumetric LV wall motion and deformation fields are recovered by using deformable models and spatial interpolation. Finally, in order to characterize the regional motion of the LV wall, a conventional 17-segment model is utilized for dividing the

reconstructed 3D model, so that the local dyssynchrony patterns can be well-determined. Our proposed approach has a great potential to be applied in the analysis of large-scale MRI datasets of various cardiovascular diseases, and used to guide the administration of CRT. Moreover, we include other applications for further demonstration of our approaches.

Acknowledgements

I would like to express my very great gratitude to my supervisor, Professor Dimitris N. Metaxas. The dissertation work would not have been possible without his support and guidance during my Ph.D. study. Professor Metaxas constantly encourages me to conduct research on influential and practical topics, and provides me with opportunities to work with top-tier researchers and doctors. He is an excellent role model in my career and life, given his enthusiasm for research, rich knowledge, and outstanding leadership.

I am also very grateful to the members of my dissertation committee: Professor Konstantinos Michmizos, Professor Yongfeng Zhang, Professor James Duncan (Yale University), for their valuable suggestions on my presentation and dissertation.

I would especially like to thank Dr. Leon Axel (NYU) for the long-term collaboration. I am deeply inspired by his dedication to science and technology.

I am particularly grateful to Dr. Dorin Comanicu (Siemens Healthcare), Dr. Yefeng Zheng (Tencent Inc.), Dr. Kevin Zhou (Chinese Academy of Science), Dr. David Liu (Samsung Inc.), Dr. Daguang Xu (NVIDIA), who provide me with great opportunities of internship.

I would like to thank the graduate students in Computational Biomedicine Imaging and Modeling Center (CBIM) at Rutgers University for the valuable discussion and support. Also, I wish to acknowledge the help provided by the staffs of Computer Science Department: Carol DiFrancesco, Ginger Olszewski, who provide me with countless instructions for the administrative process.

I would like to extend my deepest gratitude to my wife, Dr. Yan Chen, and my parents for their support and love.

Dedication

To my family

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Cardiovascular diseases, such as ventricular dyssynchrony, heart attack, and congestive heart failure, are major causes for human death all over the world. A comprehensive analysis of 3D heart wall motion is fundamental for not only understanding the ventricular functioning mechanism, but also early prevention and accurate treatment of the related diseases. Classical diagnostic tests, including electrocardiogram (ECG), echocardiography (echo), chest X-ray, and cardiac catheterization, are not able to provide sufficient spatial information for 3D motion modeling in detail. 3D echocardiography can be used to study cardiac motion and strains, but its visual appearance may not be clear due to the limited imaging quality.

In the dissertation, we adapt images from cardiac cine magnetic resonance imaging (MRI), which is an a non-invasive imaging technique to visualize the heart conditions both in time and space. MRI is widely used nowadays to study the regional motion of heart chambers [1, 2, 3]. Cardiac MRI is able to provide scans with higher temporal and spatial resolution for the heart wall motion, compared with other imaging techniques. Conventional 2D cine MRI is acquired along both the short-axis and long-axis planes for LV imaging. Short-axis scanning planes are normally parallel to each other, while the long-axis planes are rotated around the main axis of LV shell from base to apex. Typically, 20$\sim$30 cardiac phases are reconstructed within a full cardiac cycle. The MRI scans of individually imaged planes are acquired from different suspended respiration, which means there can be a spatial offset of the heart between the individual long-axis and short-axis planes. Although full 3D MRI scanning can be achieved with free-breathing or breath-holding, it often has a relatively poor imaging quality (e.g. clear artifacts, blurry muscle boundaries). Because the 3D MRI takes much longer to acquire than 2D cine MRI, covering multiple cardiac cycles, subjects typically cannot stop breathing completely for such a long time.

The sequence of 2D MRI acquired along the long-axis (LAX) and the full cardiac cycle provides a complementary view of the shape and function of the left ventricle (LV) for the sequence of 2D MRI acquired across the short-axis (SAX) and time. Thus, analyzing a set of 2D cine MRI sequence can provide a feasible way to fully recover 3D LV wall motion. In order to accomplish the analysis of 2D cine MRI, reconstruct the 3D LV wall motion, estimate the wall motion in 3D and use it to characterize disease such as cardiac dyssyncrhrony, we will provide new methods and solutions to the following open problems. Using those solutions we have developed a system for end-to-end cardiac analytics from from cine MRI data as input.

1.1 Left Ventricle Segmentation in 2D MRI

In order to analyze cardiac motion, one of the most essential clinical tasks is extracting LV contours for myocardium muscle layers at both end-diastolic volume (EDV) and end-systolic volume (ESV) in cardiac MRI. The contour extraction, used for computing ventricular global functions, is equivalent to the heart wall segmentation. In the conventional wall segmentation, the task is done either manually, interactively placing a contour at the best visual estimate of the boundaries of the solid wall, or more automatically, using mathematical optimization and learning-based methods with some smooth outer hull being placed around the "blood" and "muscle" pixels. There are effectively three concentric "zones" in the ventricle: **1)** solid muscle zone consisting of the outer wall and the endocardium wall, **2)** transitional zone with mixed blood and muscle structures, and **3)** mostly blood zone (with possibly a few muscle bundles running through it). The principal challenge associated with the conventional segmentation methods [4] is the difficulty in reliably distinguishing the trabeculae (and papillary muscles) from the underlying solid muscle wall, especially in cardiac phases near end-systole, when the blood is largely squeezed out from between the trabeculae, making them blend with each other and the wall. This causes conventional approaches to tend to fail for the end-systole determination in many cases, especially when there is hypertrophy of the wall and trabecular structures, with a resulting under-estimation for the end-systolic cavity volume and an associated over-estimation of the ejection fraction (EF).

We propose an automatic heart wall (myocardial muscle) segmentation approach using the

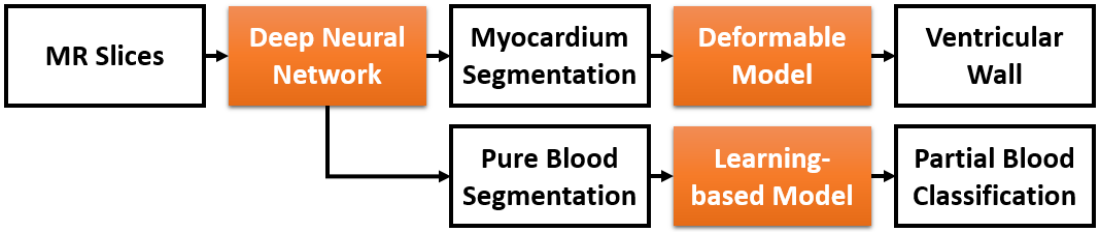

Figure 1.1: The flowchart of the proposed approach.

deep neural networks coupled with a new multi-component deformable model [5]. First, the 2D-3D neural network model provides fine segmentation masks of muscle layers with temporal continuity. Then, the multi-component deformable model is adapted to extract contours dynamically along the cardiac cycle, for both inner and outer heart walls, from the segmentation masks. The neural networks provide external force for the deformable models. The global and local constraints in the deformable model help avoid having the apparent detected boundary move artifactually inward, especially for epicardium/inner wall.

1.2 Blood/Muscle Segmentation in 2D MRI

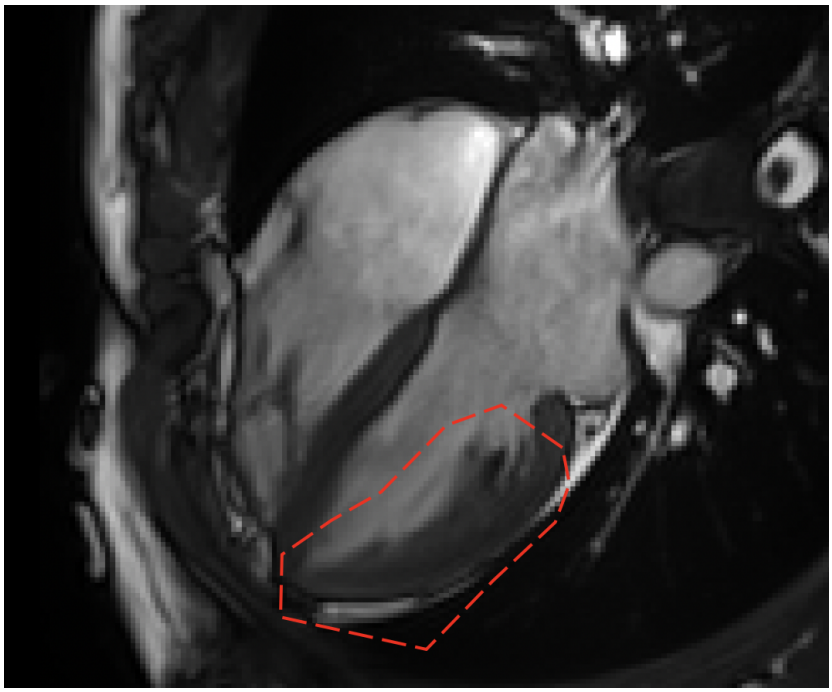

Figure 1.2: The transition zone (marked with a red circle) is the mixed with both blood and muscle. The MRI appearance becomes fuzzy at the transition zone.

The pixels in the transitional zone commonly have mixed contributions from both muscle and blood, shown in Figure 1.2; we can consider them as characterized by pixel/voxel-wise percentages in calculating the segmentation results. It can provide better understanding of cardiac wall motion, comparing with the conventional 2D cine MRI analysis, as it allows for a more realistic modeling of the transition region near the solid wall. Moreover, it can be used for both improved calculation of global function measures, and the dynamic characterization of the transitional zone itself.

Starting at end-diastole (when the ventricular cavity is most open), we can generally produce a good segmentation of the solid portion of the wall from the rest of the cavity with the use of simple conventional thresholding for the muscle intensity, combined with a smoothness constraint to suppress the derived muscle boundary from sticking to smaller structures within the cavity. This can be augmented with the use of machine learning approaches, trained on expertly segmented images. We can then define the transition zone as the region between the solid boundary and the clear cavity. The trabeculae (which can be smaller than or on the order of the size of the pixels/voxels) and blood are mixed inside the transitional zone, which can be seen as a blurry appearance in 2D cine MRI.

The challenging task is to maintain a consistent definition of the boundary of the solid wall, when heart moves into the later phases of the cardiac cycle. Because blood is ejected from the spaces between the muscle structures in the transition zone as we go into systole. It has the effect of causing these structures to appear to merge with each other. And the solid portion of the overlying wall likely to over-estimate the degree of inward motion (squeeze) of the solid wall boundary, even if using some variants of deformable contours for the boundary.

Therefore, a robust blood/muscle segmentation approach is necessary for estimation of the transitional zone. Such segmentation approach would assign probability values to each pixel about how likely it belongs to solid wall (myocardium). Adding the partial label (probability) to the segmentation process, that the total amount of muscle of the transition zone should remain about the same during cardiac cycle (neglecting through-plane motion effect). The total amount of muscle can be obtained by summing the area of the potential pixels in the zone, and weighted by their probability of being muscle rather than blood. As the heart contracts, the initial inner solid wall contour for a given phase would be moved outward (expansion) until about the right

amount of muscle is included in the transition zone, to match the initial condition. That will provide a more reliable way to segment the solid portion of the wall, using this probabilistic segmentation scheme as part of the process.

The "conservation of transition zone muscle" approach provides a reasonable way to approximately correct the simple initial segmentation of the apparent contour of the solid wall for the changing appearance of the transition zone, we can use the statistics of the transition zone to estimate the corresponding changing blood content of the transition zone. This could potentially be a new regional "ejection fraction" rate calculated, which would provide a novel way to characterize cardiovascular functions. While it would likely correlate overall with the regional wall motion, it provides a different way to assess function (independent of an eternal reference frame). The condition of LV non-compaction is also associated intrinsically with local alterations in the degree of trabeculation. And hypertrophy of the heart wall is usually associated with hypertrophy of the trabeculations as well, so that would need to be accounted for. The approach will give us some novel insights into the local function, even just using images from conventional cine imaging. This would be different from the CT approach, in that we are looking overall and regional statistics on the transition zone, rather than attempting to find point correspondence based on the trabecular structure. Codella et al. proposed a fuzzy segmentation approach purely based on image intensity for the global blood content of the ventricle [6]. It did not provide any regional information about blood content of the transitional zone. Also, their approach could be problematic because it was only applied on the 2D cine MRI data (neglecting through-plane motion effect), not in 3D space.

We explore the "partial blood content" approach to estimate the presence of boundary pixels near the trabeculae (and papillary muscles) as well as the solid wall. Adaptively adjusting the associated threshold according to the cardiac cycle phase, we should be able to improve our sensitivity to the smaller bits of blood between the trabeculae near end-systole.

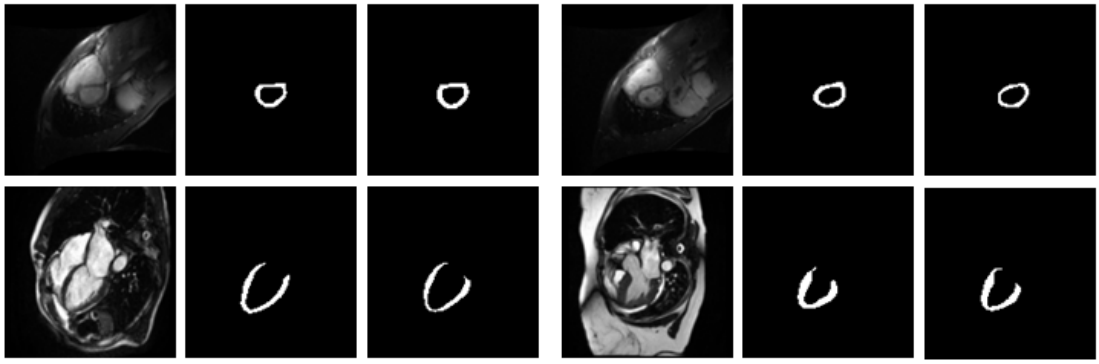

Figure 1.3: For example segmentation: the raw images (left) are segmented by FCN (middle), which are close to the "ground truth" (right).

1.3 3D Left Ventricle Model Reconstruction

3D LV model is critical for heart wall motion analysis. We introduce two model reconstruction approaches for 3D solid LV wall reconstruction and 3D blood/muscle segmentation reconstruction, respectively. They are for understanding LV functions from different perspectives.

1.3.1 3D Left Ventricle Wall Model Reconstruction

In order to analyze the global function and the regional heart wall motion, the contours of the epicardium and endocardium of LV need to be annotated or delineated, either by human experts or machines. However, the annotation procedure is often time-consuming and tedious for doctors and physicians, which becomes a bottleneck for extraction of functional cardiac data in the clinical practice. An automatic method for LV segmentation (or contour extraction), which would reduce both manual labor and annotation time dramatically, has been sought for decades to increase the clinical efficiency of cardiac MRI. Recently, some scholars have proposed several methods to address them. For example, Paragios proposed a level-set method for cardiac MRI segmentation [7] with the gradient vector flow and geodesic active contour model. Jolly also introduced an automatic segmentation method for both CT and MRI images, using multi-stage graph cut optimization in the image plane [8]. In addition, Zhu et al. developed a statistical model, named subject-specific dynamic model (SSDM), to handle the cardiac dynamics and shape variation [9]. Although the ring-shaped structure formed by the paired epicardium and endocardium contours is fairly simple, the cardiac MRI imaging quality can be

inconsistent, because of factors such as different acquisition settings or potential artifacts introduced by respiration during the slow acquisition process. Furthermore, the endocardial contour is intrinsically somewhat ill-defined, due to the presence of the papillary muscles and trabeculations, which tend to be considered as part of the ventricular cavity. Thus, the contours of LV wall segmentation may need to be estimated even when the local image contrast is partially corrupted; conventional intensity-based segmentation methods may fail in such cases. Moreover, the prevailing approaches [10, 11, 4] are mostly concerned with the SAX MRI slices. Without further study of the LAX slices and slice alignment, the calculation of the global functions for LV may not be accurate.

Removal of motion artifacts caused by varying respiration is another important issue to accommodate for analyzing the function of the heart. Although the cine MRI sequences are captured at fixed spatial locations during breath-holding, it is unlikely that the respiration phase would remain the same at different slices of the cine MRI. The MRI slices at different locations are inevitably misaligned with spatial offsets and in-plane deformation. Such misalignment issues can seriously affect the precision and representativeness of a 3D heart model that is built up on the unaligned MRI sequences. Therefore, we need to solve this image registration problem between different MRI slices. In [12], Lotjonen et al. proposed an alignment method maximizing normalized mutual information of image appearances between SAX and LAX slices. However, the optimization procedure is highly non-convex and easily falls into a local minimum. Although Garlapati et al. [13] proposed an effective method to solve the misalignment problems in brain imaging, based on the local boundary detection, it is not applicable in our case because the boundary of LV wall is not always clear in cardiac MRI.

Once the in-plane segmentation and alignment are achieved, 3D LV wall modeling and motion reconstruction of LV wall is the next steps in analysis. The 3D wall shape and motion provide quantitative and visual characteristics to study the normal and abnormal heart functioning mechanisms in a comprehensive way. Park et al. studied the shape and motion of LV using a volumetric deformable model based on tagging MRI [14]. The dynamic deformation of the ventricular wall is computed with Lagrangian dynamics and finite element method.

We present a novel approach to reconstruct 3D shape and motion of LV wall for understanding ventricular functioning mechanisms [15]. First, we adopt a fully convolutional network (FCN) to extract epicardium and endocardium contours from the MRI slices. Second, we develop a new algorithm to align MRI slices in space, compensating the respiration effect. Finally, a deformable model is utilized to recover the 3D shape and motion of LV wall.

1.3.2 3D Blood/Muscle Segmentation

It is also desirable to use the blood/muscle segmentation approach to provide 3D results for the transitional zone at each cardiac phase, because the through-plane motion of the tapered LV wall introduces systematic artifacts into the simple 2D segmentation. In order to recover the full 3D information from multiple separately acquired 2D cine slices, it is necessary to understand the motion in space. Recently, generative adversarial networks (GAN) have been proposed to model the appearance distribution of the image domain for many applications. Fully-connected or fully-convolutional neural networks have been utilized successfully for image generation from 1D noise vectors [16, 17]. The idea of GAN can be further extended for image understanding, completion, reconstruction, segmentation [18], and other similar applications.

We propose a novel 3D blood/muscle cardiac segmentation approach given only 2D cardiac cine MRI acquisitions [19], as shown in Fig. 5.1. In the proposed approach, 2D epi- and endocardial contours are first extracted to provide the 2D in-plane blood/muscle segmentation, and to compensate spatial offset artifacts caused by inconsistent respiration. Each pixel is assigned a probability value characterizing how likely it is to belong to the myocardium. Then, we adopt a generative adversarial network (GAN) to transform multiple 2D blood/muscle segmentation maps in space into a fully 3D blood/muscle segmentation. With the results of the 3D segmentation over the cardiac cycle, we would have a better understanding of the cardiac wall motion. Our work is the first attempt to reconstruct such a 3D blood/muscle segmentation from 2D cine MRI, to the best of our knowledge.

The motivation of using GAN to reconstruct 3D blood/muscle segmentation in our method is as follows. GAN is capable of describing the distribution of contextual appearances, and it has achieved state-of-the-art results on several image-related tasks, for instance, image super-resolution, image denoising, MRI reconstruction, etc. We treat the blood/muscle segmentation

as one kind of contextual appearance, and our goal is to process the partial signals for full-information reconstruction based on the appearance population of training samples, which is equivalent to the aforementioned tasks. Although GAN is not the only way to process such tasks, it is a more efficient way than many others, in terms of accuracy and speed.

1.4 Assessment of Ventricular Dyssynchrony

Heart diseases, including heart rhythm problems, heart defects, etc., are the leading cause of human death around all over the world. Among them, cardiac dyssynchrony is quite common, and one of the major reasons for heart failure. In clinical practice, the electrocardiogram (ECG) signal of LV motion is analyzed to detect dyssynchrony patterns, such as a prolonged QRS complex (for three of the graphical deflections on a standard ECG), and determine the corresponding treatment plans, including cardiac resynchronization therapy (CRT). Although the ECG based QRS provides global quantitative measurements related to the heart motion, it does not have a strong correlation with dyssynchrony symptoms. For example, it is possible for dyssynchrony patients to have relatively normal ECG curves and QRS, as well as for healthy people without dyssynchrony symptoms to have an abnormal ECG. As a result, it is important to study in addition to the electrical ECG-based signal, the regional motion of the cardiac chambers to determine cardiac dyssynchrony. Moreover, studying the regional motion of the cardiac chambers is critical to evaluate the effectiveness of cardiac CRT in improving cardiac pump function. For patients with ventricular dyssynchrony and heart failure, CRT is a popular treatment to reduce heart failure. However, the outcomes of CRT cannot be guaranteed, even when the ECG signal/QRS becomes normal after CRT. The regional analysis also provides useful information for the related surgical planning, for example, how to place pacemaker leads most effectively into the cardiac chambers. This in turn can impact a clinicians assessment on whether CRT will be helpful or not for a particular patient.

We propose an efficient approach to analyze ventricular dyssynchrony, using 2D cardiac MRI and deep neural networks, as shown in Fig. 1.4. Based on the short- and long-axis MRI scanning planes, the boundaries of myocardium are segmented using full convolutional neural networks [5]. Then, any respiration offset is compensated using an iterative method based on

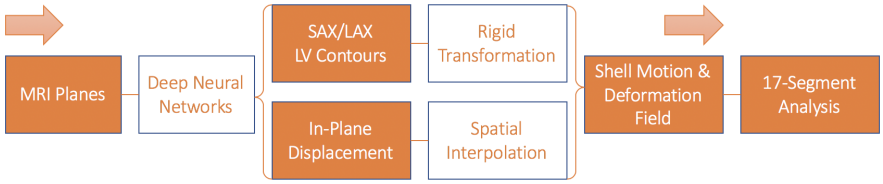

Figure 1.4: The flowchart of the proposed approach.

the location of the boundary contours in 3D [15]. Furthermore, the 2D displacement field within LV contours can be approximately computed using an unsupervised learning method, and the 3D displacement field can be computed through spatial interpolation. Then, the full LV wall and displacement field are reconstructed in 3D at multiple phases of the cardiac cycle. Finally, the regional motion of LV wall is analyzed following the way of 17-segment LV model [20].

1.5 Other Applications in Medical Imaging

In Chapter 7-8, several other applications in medical image analysis using deformable models and deep neural networks will be further discussed. The core of the presented applications is relevant to 3D anatomy understanding. Combining deformable models and deep neural networks is an efficient and robust to enforce local and global constraints for the reconstructed 3D models. Those applications indicate that our proposed approach can be generalized and applied among different scenarios with moderate changes. Therefore, the potential of the proposed approach for large-scale medical image analysis is further validated.

1.6 Dissertation Structure

The structure of the dissertation is as follows. In Chapter 2, we conduct literature reviews for cardiac MRI segmentation, 3D LV modeling, and assessment of ventricular dyssynchrony. In Chapter 3, we consider the novel approaches for LV segmentation and blood/muscle segmentation in 2D cine MRI. In Chapter 4, we discuss how to reconstruct 3D LV wall model with 2D LV segmentation, and 3D blood/muscle segmentation is further explored in Chapter 5. In Chapter 6, the proposed framework is applied for ventricular dyssynchrony assessment. From Chapter 7-8, other applications with deformable models and deep neural networks are further investigated. We reach the conclusions of our proposed framework in Chapter 9.

Chapter 2

Related Work

2.1 Cardiac MRI Segmentation

The LV myocardium segmentation has been addressed by many researchers in the past decades, in order to alleviate the human effort for the time compsuming annotation procedure. For example, Paragios developed a segmentation pipeline using level-set optimization and gradient vector flow (geodesic active contour) [7]. Jolly proposed a multi-stage graph-cut method for cardiac segmentation in both MRI and CT [8]. These methods rely on the image appearance, and they would probably fail when the image contrast is changed. Zhu et al. [9] introduced a subject-specific dynamic model to delineate the ventricular shape variance. However, their results tend to move a bit inward to the blood pool, which may introduce errors for the global function estimation. Recently, the cardiac segmentation has been addressed using deep neural networks [21, 22, 15, 23], benefiting from its advanced feature learning capacities. However, most of these learning models are trained and used to infer the boundaries for individual images, which tend to lead to lack of temporal continuity in the segmentation.

Huang et al. proposed a tracking approach for contours/meshes in echocardiography using sparse representation and dictionary learning [24]. They used the assumptions that high-dimensional local image appearance can be sparsified into a multi-scale apprearance dictionary. Such online dynamic dictionary was utilized in a level set algorithm together with intensity and shape for contour tracking. Their results on $3D + t$ echocardiography of human subjects and animal are promising. And there are few points which can be further improved. Firstly, the initialization of contours/meshes at the first cardiac phase with good quality can be generated with the contemporary state-of-the-art approaches (e.g. neural network based approaches) to reduce human interaction. Second, the computing efficiency can be further improved. In their approach, computing the endo- and epi-cardium contours using the level set algorithm consumes

approximately one minute per cardiac phase.

As an alternative, Codella et al. tried to get the best estimate of the total "true" blood volume in the chamber, by weighting each candidate voxel by its fractional blood content and then summing them [6, 25]. This would then be used for calculation of the conventional global function measures (based on differences in the blood in the ventricle between end-diastole and end-systole), including stroke volume and ejection fraction (EF). It was presented as an alternative to the conventional method of defining the "cavity," the trabeculae, and the papillary muscles, in order to segment out the "solid" wall only. In their approach, the voxel probability distribution was computed directly from the intensity scale, which may cause errors when the imaging quality is compromised. Our approach addresses previous limitations and correctly segments the 3 zones based on the coupled U-Net and the multi-component deformable model.

2.2 Assessment of Ventricular Dyssynchrony

While the diagnosis of ventricular dyssynchrony is fairly accurate in case of dyssynchrony, its treatment usually with CRT is not always successful. The shape and duration of the QRS complex, measured from the ECG, is commonly adopted as an indicator to detect left/right bundle branch block or cardiac hypertrophy in the regular clinical work flow [26]. However, the measurement of QRS is not always reliable to characterize dyssynchrony, because the dyssynchrony motion patterns could exist even with relatively normal QRS values, while QRS prolongation can be associated with relatively normal motion patterns. Without further analyzing interior motion, thus, cardiac dyssynchrony cannot be well characterized purely based on the global electrical conductivity-based metrics and the ECG. Currently, cardiac ultrasound imaging is widely used for clinical diagnosis and surgical planning. Although its imaging process is efficient (effectively real-time), the imaging quality for wall motion assessment is limited because many imaging artifacts are introduced, and the boundaries of the muscle/blood regions can be blurry in the reconstructed images. Thus, ultrasound only provides relatively coarse visual guidance, and it is limited for studying regional cardiac wall motion during dyssynchrony. Cardiac CT is an alternative imaging technique to analyze the cardiac diseases, with high imaging

resolution in 3D space. The cardiac vessels and veins are visualized clearly with contrast-enhanced CT, and the sites to lace pacemaker leads for CRT can be accurately determined [27]. However, cardiac CT has a relatively low temporal resolution for capturing dyssynchrony, and it relies on ionizing radiation, which is harmful for the human body.

Cardiac MRI has sufficient spatial and temporal resolutions to study cardiac functioning. [1] analyzed the ventricular dyssynchrony using cardiac tagging MRI. The MRI-based strain analysis was conducted naturally using the in-plane tagging line movement. Later, [2] used delayed enhancement cardiac MRI as the scar imaging, to study the relationship to circumferential mechanical dyssynchrony before and after CRT. Recently, [28] proposed a study combining evaluation of motion patterns, scar, and electrical timing of CRT using 2D+t cardiac MRI cine displacement encoding with stimulated echoes (DENSE). The imaging technique naturally provides the in-place strain and displacement field. However, it requires a long acquisition period, and 2D-based study can be limited for monitoring the through-plane motion of the myocardium.

Deep neural works (DNN) have been successfully deployed on image processing many applications . The variants of DNN provide efficient solutions for various tasks, such as image classification, image segmentation, and image super-resolution. They can be fitted into different learning settings, including supervised learning, unsupervised learning, semi-supervised learning. Recently, researchers proposed to use DNN to compute optical flow from video clips [29]. The differentiable bi-linear interpolation is embedded in the DNN to transform one frame towards next frame. The loss for training DNN is the difference between the next frame and transformed current frame. After optimization, the network provide dense displacement field as optical flow for videos. Similarly, [30] adopted a similar strategy for 2D/3D medical image registration. The deformation field is achieved when the distance between deformed initial images and target images is minimized. Such an unsupervised learning setting is helpful to estimate displacement field, especial when it is hard to collect the ground truth field. The experiments indicated that such a learning strategy works better and much more efficiently than classic optical flow or image registration methods, given large-scale datasets.

Chapter 3

Myocardium Segmentation in 2D Dynamic Cardiac Magnetic Resonance Imaging

Figure 1.4 shows the flowchart of our proposed approach for myocardium segmentation in cardiac MRI. Initially, coarse segmentation results are generated from the 2D U-Net for individual images. Then, the previous results are stacked into 3D volumes according to their order of cardiac phases and cropped into a centralized region-of-interest (ROI) according to the coarse segmentation for segmentation refinement with a 3D U-Net. Finally, we utilize a multi-component deformable model to determine the myocardium boundaries with global and local constraints. We also compute the probability of pixels belonging to blood, based on the features learned from deep neural networks.

3.1 2D-3D U-Net Model

The concept of U-Net, first proposed in [31] has been successfully applied in many applications of medical image analysis. It has been validated to possess good generalization capacity with few annotated samples. The network consists of a convolutional encoder and decoder, and its U-shape generates multi-scale features and computes them with multi-step convolution and up-sampling. The output of the U-Net shares the same size as input. There are skip connections in between the encoder and decoder to concatenate multi-level feature maps, allowing the decoder to store back the relative features that are lost in the prior stage.

Our 2D-3D U-Net model is described in the Fig. 3.1. For individual phases of MR sequences, we adopt two 2D U-Nets [31] to segment epicardium and endocaridium masks, respectively. Direct predicting myocadium muscle using only one network is also possible. However, in that scenario, the positive samples in the gold standard are generally much less than the negative samples, which makes the learning procedure biased and affects the performance later.

After achieving preliminary segmentation using 2D U-Nets, we crop the region-of-interest (ROI) according to the center of segmentation for both image and segmentation. The same cropping region is applied for all phases in one MR sequence. Then, we stack all the cropped images at the same location, and segment them along the temporal dimension into 3D volumes as input, and adopt 3D U-Nets [32] to refine the previous segmentation. The 3D U-Nets would enforce the smooth prediction in-between consecutive cardiac phases. It may not be easy to directly apply 3D U-Net to the entire image and segmentation regions, since the input with the original size would have large memory consumption and become a bottleneck during computation with GPUs.

Our key point here is to use a convolutional model instead of a recurrent model to handle the temporal data. Although the recurrent model is well established for time-series problems, recent research shows that a fully convolutional model could outperform a recurrent model in some sequential problems, for instance, language translation [33],and video segmentation [22]. Furthermore, similar 2D-3D network models have been adopted for a few medical image segmentation applications and achieved excellent results [34].

3.2 Multi-Component Deformable Model

In clinical practice, doctors and physicians often manually correct overestimated regions of the inner wall contour, relying on playing the serial frames of the cardiac cycle as a movie of the imaged slice to locate trabeculae pretty reliably at end-systole. Then their associated motion can be estimated over the cardiac cycle from the moving animated display, including when they are too close together to reliably detect in an isolated frame. Similarly, we propose a multi-component deformable model to finalize the contours of endo- and epicardium, to simulate the manual correction. At each cardiac phase, the energy function of the deformable model for epicardium can be written as follows.

$$E_{\text{epi}} = \alpha E_{\text{external}} + \beta E_{\text{continuity}} + \gamma E_{\text{smooth}} \tag{3.1}$$

The energy function of the deformable model for endocardium is different from Eq. 3.1, shown as follows.

$$E_{\text{endo}} = \alpha E_{\text{external}} + \beta E_{\text{continuity}} + \gamma E_{\text{smooth}} + \phi E_{\Delta\text{area}} \tag{3.2}$$

The external energy is $E_{\text{external}} = \int \|v_s - v'_s\|_2^2 \, ds$. v_s and v'_s are the corresponding points of the current deformable model and targeting contour from previous deep neural networks. The continuity term $E_{\text{continuity}} = \int \left\| \frac{dv_s}{ds} \right\|_2^2 ds$ ensures that the neighboring points are close. The smoothness term $E_{\text{smooth}} = \int \left\| \frac{d^2 v_s}{ds^2} \right\|_2^2 ds$ guarantees that the model is always a convex smooth shape. Here, $\alpha, \beta, \gamma, \phi$ are all positive constants. In practice $\alpha = 1.0$, $\beta = \gamma = \phi = 0.2$. The epicardium deformable model has an extra energy term

$$E_{\Delta\text{area}} = \left| \int \|v_s - w_s\|_2^2 \, ds - \int \left\|v''_s - w''_s\right\|_2^2 ds \right|. \tag{3.3}$$

w_s are the points of the fixed epicardium contour in the current phase. v''_s and w''_s are the corresponding points from the previous endo- and epicardium deformable models. $\int \|v''_s - w''_s\|_2^2 \, ds$ is equivalent to the myocardium area in the previous phase. The assumption is that the area of myocardium muscle can only change within a limit range among neighboring cardiac phases because the muscle volume is almost unchanged during the cardiac cycle. At inference, we start from the contours of the first phase (normally EDV). The epicardium contour of the next phase is computed by solving Eq. 3.1 and moving contour points along their normal directions, till reaching a minimum status. Then the endocardium contour is computed by minimizing Eq. 3.2 together with the updated epicardium contour. The myocardium contours are derived phase-by-phase in sequence; and the sample results are shown in Fig. 3.2.

In order to calculate the probability of being blood for pixels inside myocardium wall, we apply extra 2D U-Nets to segment pure muscle and blood regions, respectively, and extract the features (length 64) from the second last layers of networks. A logistic regression model is learned with pixels from regions of pure muscle/blood and their extracted features. The pixels of pure blood are labeled as 0, and those of pure muscle are labeled as 1. The probabilities of the other pixels inside the ventricle would be computed using the learned model. For cases where the apparent blood spaces in the transitional zone seem to entirely disappear, we are able to estimate where the corresponding transitional zone was moving to at the times (in earlier systole and later diastole) when we could more reliably see and track it. While this is inherently uncertain, it should still be better than purely relying on image intensity, which can result in significant errors in the estimation of the end-systolic volumes, due to the effective initial assignment of the "wall" contour to the solid wall-transitional zone boundary at end-diastole,

but then moving it inward toward the transitional zone-blood boundary at end-systole.

3.3 Experiments

3.3.1 Dataset and Myocardium Segmentation

We adopted a cardiac MRI dataset consisting of 22 normal volunteers and 3 patients with cardiac dyssynchrony disease. All LV contours of these SAX images over different spatial locations and different cardiac phases are manually annotated by experts. In-plane resolution of images ranged from 1.17 mm to 1.43 mm, and size varied from 224×204 pixels to 240×198 pixels. Each cardiac cycle contains 25 phases. As for [11, 35], we conducted our evaluation procedure in the following way. We run the 5-fold cross validation, and make sure that each of the 25 subjects (containing both normal subjects and patient, around 4000 2D slices with manual annotation) in the test set exactly once.

To boost the robustness of our model, we used data augmentation by 90-degree rotation and mirroring. All images were scaled to resolution of 1.25 mm and padded with zero to gain the same image size. The filter numbers of 3D U-Nets was reduced by half to fit the data and reduce GPU memory consumption. We adopted the soft Dice loss [36] during training. For both tasks, we used ADAM optimizer with a fixed learning rate 0.001 and weight decay of 2.0^{-5}. The results reported below were obtained after training for 30 epochs with batch size 32 for 2D U-Net and 15 epochs with batch size 16 for 3D U-Net. Training one model takes 12 hours on one NVIDIA K80 GPU, and inference takes about 1 sec. for a full cardiac cycle.

In order to make the fair comparison, we re-implemented the state-of-the-art methods[22, 23], and ran them on the same dataset with our proposed method. For endocardium, the average Dice' score of the original 2D U-Net is 0.864 better than the FCN8 (0.855), FCN16 (0.848) and FCN32 (0.639). Our proposed method gains much better results, i.e., 2D-3D U-Net is 0.886 and 2D-3D U-Net + Deformable Model is 0.902 which is the highest one to date. In addition, our proposed methods outperform others in terms of the Jaccard index and APD. To further evaluate our method, we also calculated the percentage of good contours (a perentage of the predicted contours, out of all contours, that have APD less than 5 mm from the gold standard [35]). Among all the segmentation results, Our best model 2D-3D U-Net + Deformable had 97.5%

Table 3.1: Evaluation of endo- and epicardium segmentation, A, B, C represents 2D U-Net, 2D-3D U-Net (Ours), and 2D-3D U-Net + Deformable Model (Ours), respectively.

	Method	Dice	Jaccard	APD (mm)
Endo.	FCN8	0.855 ± 0.218	0.759 ± 0.195	3.845 ± 4.950
	FCN16	0.848 ± 0.212	0.738 ± 0.214	3.134 ± 5.595
	FCN32	0.639 ± 0.274	0.475 ± 0.238	8.094 ± 7.534
	[22]	0.850 ± 0.204	0.742 ± 0.219	6.278 ± 17.801
	[23]	0.859 ± 0.203	0.758 ± 0.213	2.470 ± 3.967
	A	0.864 ± 0.180	0.764 ± 0.196	3.799 ± 8.930
	B	0.886 ± 0.035	0.821 ± 0.038	2.768 ± 1.946
	C	$\mathbf{0.902} \pm 0.035$	$\mathbf{0.847} \pm 0.037$	$\mathbf{1.647} \pm 0.609$
Epi.	FCN8	0.877 ± 0.177	0.731 ± 0.276	1.964 ± 4.986
	FCN16	0.883 ± 0.244	0.763 ± 0.2553	2.994 ± 4.213
	FCN32	0.833 ± 0.165	0.716 ± 0.172	4.318 ± 3.619
	[22]	0.857 ± 0.194	0.746 ± 0.204	3.050 ± 7.291
	[23]	0.821 ± 0.231	0.712 ± 0.243	4.071 ± 4.976
	A	0.886 ± 0.168	0.797 ± 0.187	2.821 ± 4.922
	B	0.895 ± 0.327	0.839 ± 0.035	2.387 ± 0.839
	C	$\mathbf{0.905} \pm 0.037$	$\mathbf{0.855} \pm 0.039$	$\mathbf{2.094} \pm 0.535$

good contours. In the epicardium experiment, both our proposed model 2D-3D U-Net (0.895) and 2D-3D U-Net + Deformable Model (0.905) achieved very high Dice's score as well. Also, it's obvious that the average Jaccard index of the 2D-3D U-Net + Deformable Model (0.855) is higher than that of both 2D U-Net (0.797) and 2D-3D U-Net (0.839), and the good contour percentage of the 2D-3D U-Net + Deformable Model is 93.3%. Some good segmentation examples of our 2D-3D U-Net + Deformable Model are shown in Fig. 3.3. As the 2D-3D U-Net utilizes three dimensional information, it outperforms the traditional 2D U-Net in all three evaluation metrics. Moreover, the 2D-3D U-Net + Deformable Model uses the temporal information to further refine the result. Overall, our methods generate results very close to the gold standard compared with other methods for both endo- and epicardium.

3.3.2 2D Blood/Muscle Estimation

Since it is hard to define a gold standard for the probability maps of partial blood estimation, we only evaluated our results visually, as shown in Fig. 3.4. We can see that the region where muscle and blood mix are assigned with a probability value between 0 and 1, not purely affected by the local appearance.

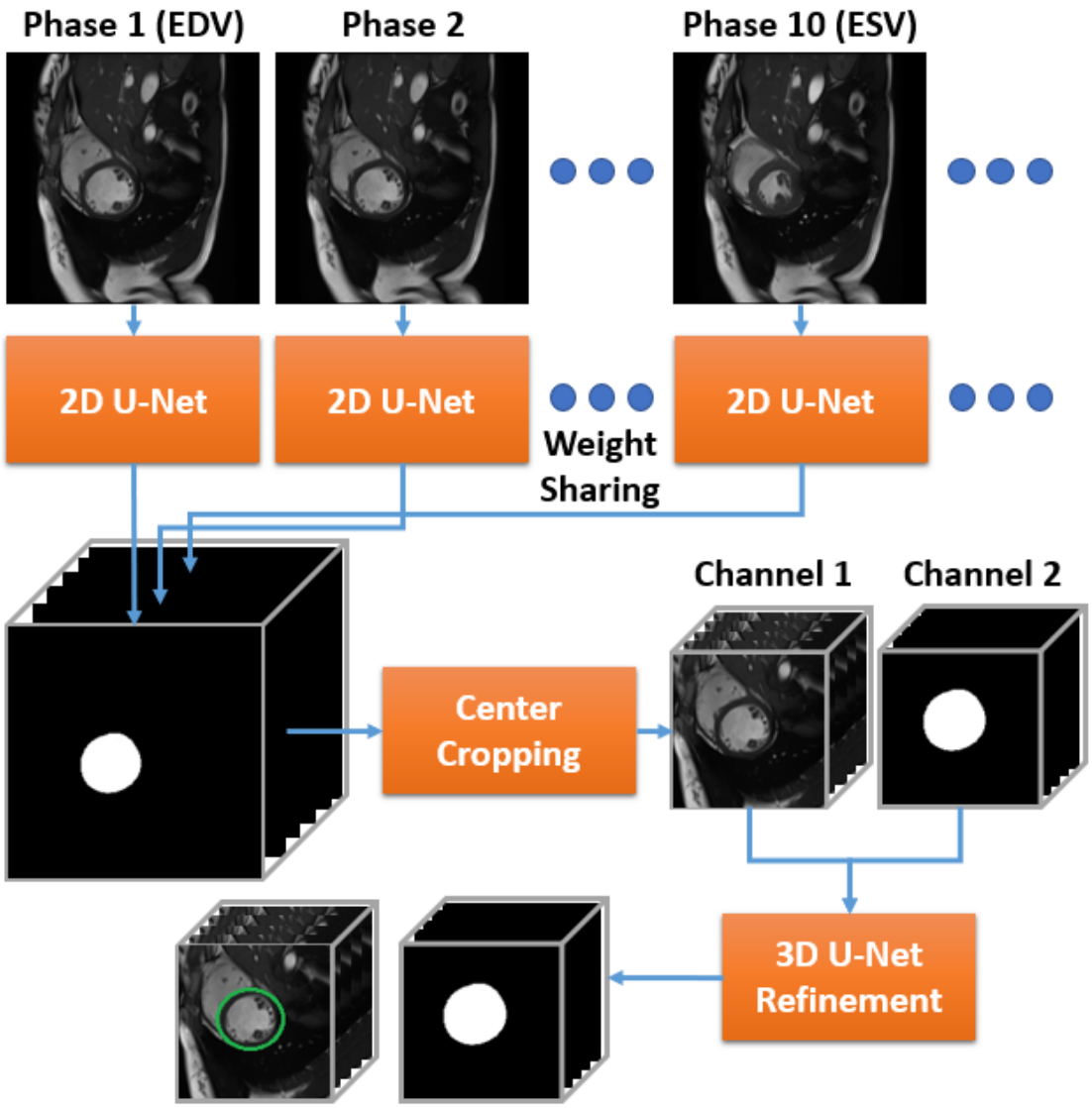

Figure 3.1: The flowchart of the proposed 2D-3D U-Net method. The 2D U-Net is used for generating segmentation priors at each individual cardiac phase, and the 3D one further refines the segmentation results along the temporal dimension with a small cropping region.

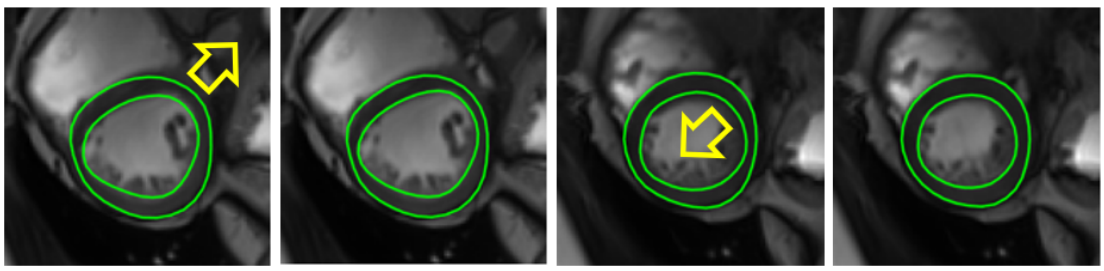

Figure 3.2: The contours before and after applying deformable models. Left: the contours from previous frame, right: the updated contours. The yellow arrows indicate the updating direction of contours.

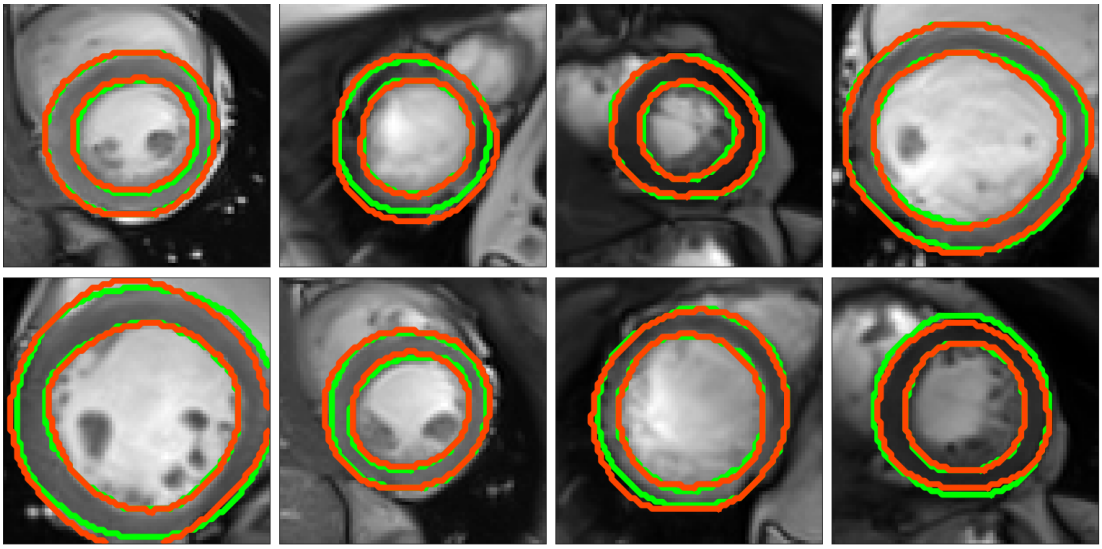

Figure 3.3: Sample results of proposed methods. Green contours are the gold standard, and red contours are the prediction.

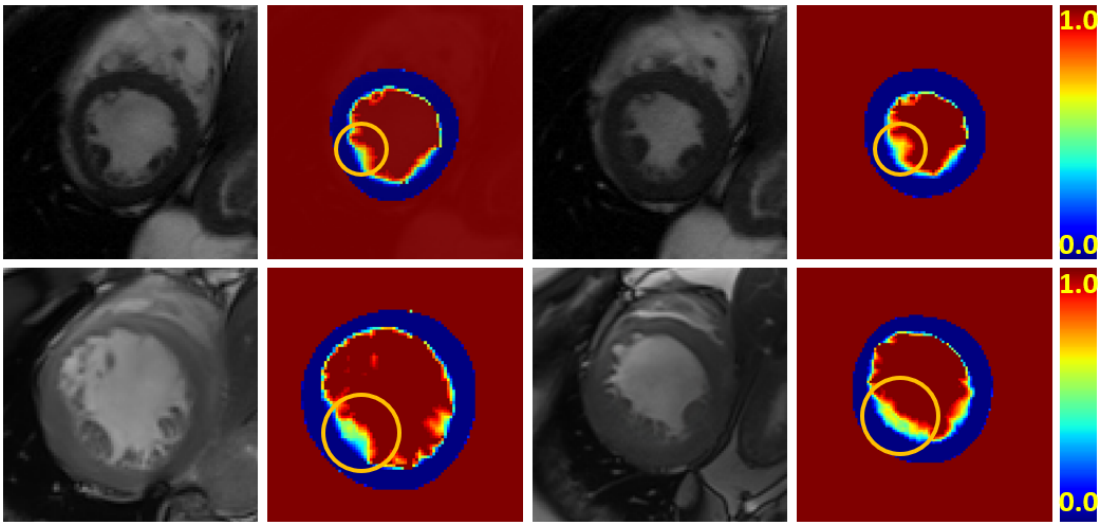

Figure 3.4: Sample results of partial blood segmentation. Left: original image, right: probability map of blood. The yellow circles denote the transition zone.

Chapter 4

3D Modeling and Reconstruction of LV Wall

4.1 Myocardium Contour Extraction

In this section, we introduce an alternative approach for endo- and epi-cardium contour extraction from 2D short-axis cine MRI. The approach utilizes the deep neural networks and group sparsity.

In our framework, LV segmentation is defined as a pixel-wise semantic classification problem, that is, segmentation with class labels. The pixels of myocardium muscle within a semi-ring shape (formed by the epicardium and endocardium) are labelled as one class; pixels of blood pool and other contents are labelled as another class. We adopt the fully convolutional network (FCN), U-net [31], as the learning model following the end-to-end convention during the training and testing. The initial segmentation results are shown in Figure 1.3, which don't all resemble the golden standard.

We enforce strong shape constraints for the segmented contours resulting from the previous step, since the ring-shaped structure of the LV contours is an important prerequisite and the smoothness of contours needs further refinement. However, the raw prediction from the FCN sometimes forms ring-shapes with unreasonable patterns, e.g., zig-zag curves or the intersection of two contours, as shown in the fourth example of Figure 4.2. The initial shape is generated from the shape pool and can be reliably placed in the image plane even when the appearance cue is misleading. The shapes of the LV wall vary from the phase of end-diastole (ED) to that of end-systole (ES), and from the slices near the aorta to those near LV apex. For instance, the contours close to the aorta may be partially merged together, particularly in the membranous portion of the interventricular septum, and the myocardium muscle close to the LV apex is thinner compared to the muscle at other locations (although the typically oblique intersection of the image plane with the apical LV wall in SAX images can result in apparent increased wall

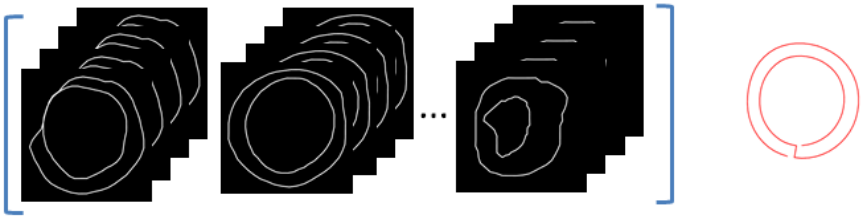

Figure 4.1: Left: clusters in the shape pool; right: mean shape.

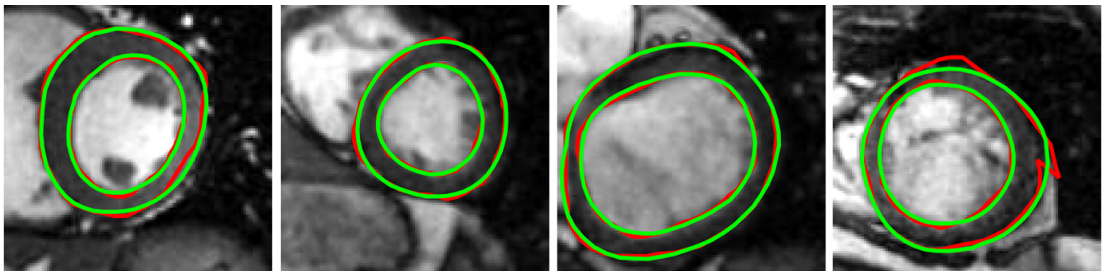

Figure 4.2: Four sample results before and after applying the group sparsity constraints: red contours are the results from the proposed fully convolutional network (FCN), green ones are the refined results after applying group sparsity constraints.

thickness, due to volume averaging). We cluster training shapes into different groups by the geometry and muscle thickness, and compute the refined contours of testing data by optimizing the dictionary learning formulation with the group sparsity constraints shown in Equation 4.1:

$$\underset{x,e,\beta}{\text{minimize}} \left\{ \|T(y,\beta) - Dx - e\|_2^2 + \lambda_1 \sum_{s \subseteq S} \|x_s\|_2 + \lambda_2 \|e\|_1 \right\} \tag{4.1}$$

where $T(y,\beta)$ is the similarity transformation with parameter β for aligning the initial shape y, generated by FCN, to the mean shape of the shape pool. Matrix $D = [d_1, d_2, \cdots, d_k]$ represents the training shape pool, column vector $d_i \in \mathbb{R}^{3n}$ contains the coordinates of n vertices on the contours. $S = \{1, 2, \cdots, k\}$ is the set of indices of x. The clustering process divides S into several non-overlap subsets, $S = \bigcup_i s_i$, $s_i \bigcap s_j = \emptyset, \forall i \neq j$. Vector $x \in \mathbb{R}^k$ contains the weights for the linear combination of shapes in the pool. x_s is the sub-vector for the group $s \in S$, and the term $\sum_{s \subseteq S} \|x_s\|_2$ is a standard group-sparsity regularization ($l_{2,1}$ norm). Vector e models the non-Gaussian error in the case that partial contour information is missing. λ_1 and λ_2 control the weights of two sparsity terms. After solving the optimization, the myocardium contours are refined with the most correlated shapes from a small group of shapes from the training pool. The sample results are shown in Figure 4.2. The similar process is conducted for the LAX slices as well.

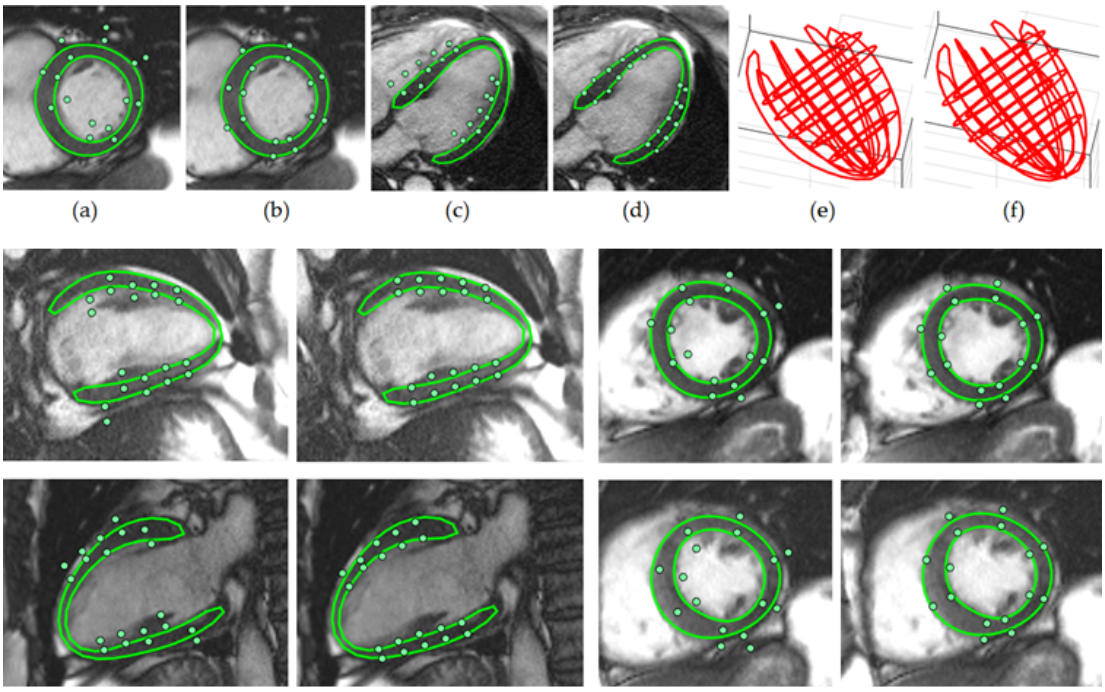

Figure 4.3: Results (before and after) MR slice alignment. (a,b): SAX myocardium contours and intersection points with LAX contours; (c,d): LAX myocardium contours and intersection points with SAX contours; (e,f): all contour points in 3D space; bottom: four sample slices with intersection points before and after alignment.

4.2 Rigid Image Registration for Spatial Alignment

The heart motion under respiration is mainly a rigid-body translation in the craniocaudal (CC) direction, with minimum deformation[37]. Therefore, we assume that in-plane rigid translation is sufficient to compensate the respiration effect for SAX. We also assume the offset of one cardiac phase in a slice can be applied for all cardiac phases at the location, because the respiration phase is almost identical in one-slice acquisition with breath-holding. For simplicity, the registration is carried out only at the state of end-diastolic (ED) for all slices simultaneously.

We propose a novel slice alignment algorithm, described in Algorithm 1, to adjust both SAX and LAX slices, using the contours from the previous step and slice intersection relations. Since SAX slices are almost parallel to each other, we take intersections between SAX and LAX slices, or different LAX slices, into consideration. At SAX slice s, the corresponding image plane is T_s and 2D contours (epicardium and endocardium) are v_s. Contours v_l in LAX slice l have intersection points p_l with slice s. Then, the closest points $p_s \in v_s$ are computed corresponding to all points in p_l. $\|p_s - p_l\|_2 = 0$ ideally if no respiration effect exists during

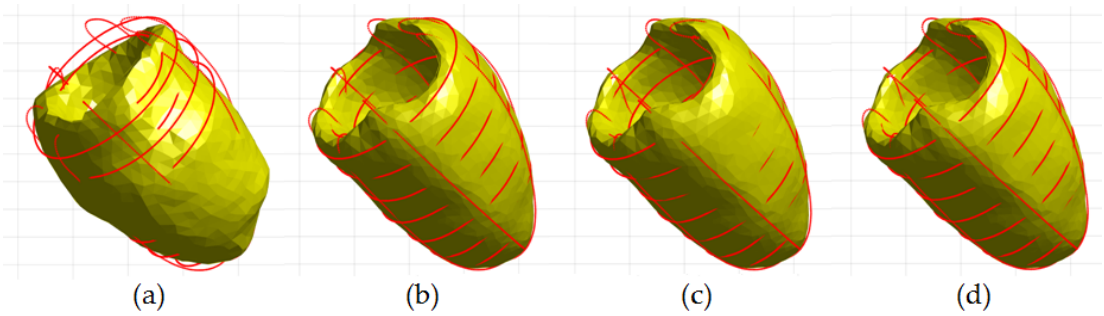

Figure 4.4: 3D yellow models are the LV models, red curves are the 2D aligned contours from SAX and LAX slices in space. (a) Initial model from the referenced LV model at the phase of ED using CPD; (b) fitted model for the phase of ED using deformable model based on the contours; (c) LV model at the phase $k-1$ and contours at the phase k; (d) final fitted model at phase k.

the acquisition. However, as shown in Figure 4.3, p_s and p_l may not intersect with each other. The difference $p_s - p_l$ provides the direction to shift the image plane (or shift the contours equivalently). Computing all the intersection points from LAX contours, the final translation displacement can be determined by taking the average on $p_s - p_l$. The procedure is analogous for LAX slices. The whole procedure is repeated if the marginal update of alignment is greater than a fixed threshold. The complete algorithm, shown in Algorithm 1, is guaranteed to converge to a stable condition where most intersection points are on the in-plane contours among all slices and frames.

4.3 3D Shape Modeling and Motion Reconstruction

Deriving 3D shape and motion of LV wall from the well-aligned contours of different slices is essential for understanding heart functioning mechanism. Analyzing motion of a sequence of 2D contours along an axis and time is able to show some characteristics of heart motion. However, 2D image slices, at the same location but at different phases of the cardiac cycle, actually may present different parts of heart, due to the 3D ventricular motion. Thus, the sequence of 2D MRI slices does not show the true pattern of heart dynamics (shape, strain, etc.). In order to achieve better analysis, we recover the 3D LV wall shapes over the whole cardiac cycle from the sparse in-plane contours. We propose a new method, shown in Algorithm 2, to reconstruct 3D LV shapes and motion, adopting the deformable model. We use the rigid point-wise registration method, coherent point drifting (CPD) [38], to initialize the 3D shape

for the cardiac phase of end-diastolic (ED) from a reference shape towards the aligned contours in space. The shapes for the whole cardiac cycle are computed along the direction from ED to end-systolic (ES). Next we construct the deformable model directly on the triangular mesh from results of CPD registration. The point locations of the deformable model [39] are a function of time t and vector q:

$$x\left(q,t\right)=c+R\left(s+d\right) \tag{4.2}$$

where c is the origin of local coordinates, R is a rotation matrix, s and d are global and local deformation, respectively. q is defined as a vector of parameters in kinematics and dynamics and $\dot{x}=L\dot{q}$, where matrix L is derived from Equation 4.2. According to Lagrangian dynamics, we have the following equation:

$$D\dot{q}+Kq=f_t \tag{4.3}$$

where D is the damping matrix, and K is the stiffness matrix. The external force f_t at phase t is proportional to the Euclidean distance between contour points and initial shape S within a local neighborhood. Once we have the initial shape, we can update the deformable model and the corresponding mesh by solving Equation 4.3. Therefore, the shape at each phase can be computed using the computed shape of the previous phase as initialization for deformation (Figure 4.4). Then, we can recover the whole motion of the LV wall phase-by-phase with proper smoothness (guaranteed by the deformable model).

4.4 Experiments

We used a cardiac MRI dataset containing MR image sets of 22 normal volunteers and 3 patients for the initial study. The patients all had heart failure with dyssynchrony, and were scheduled for cardiac resynchronization therapy (CRT). We manually annotated LV contours for all LAX and SAX images at each location over different cardiac phases, except the slice planes that did not cut through the LV. Image size varied between 224×204 pixels and 240×198 pixels, and its resolution varied from 1.17 mm to 1.43 mm. In total, 25 subjects (approximately 5625 images, both SAX and LAX) were used from our dataset, randomly divided into training set (20 subjects) and testing set (5 subjects). The SAX and LAX network models were trained separately. A 5-fold cross-validation was used in the training set. We compared the results for

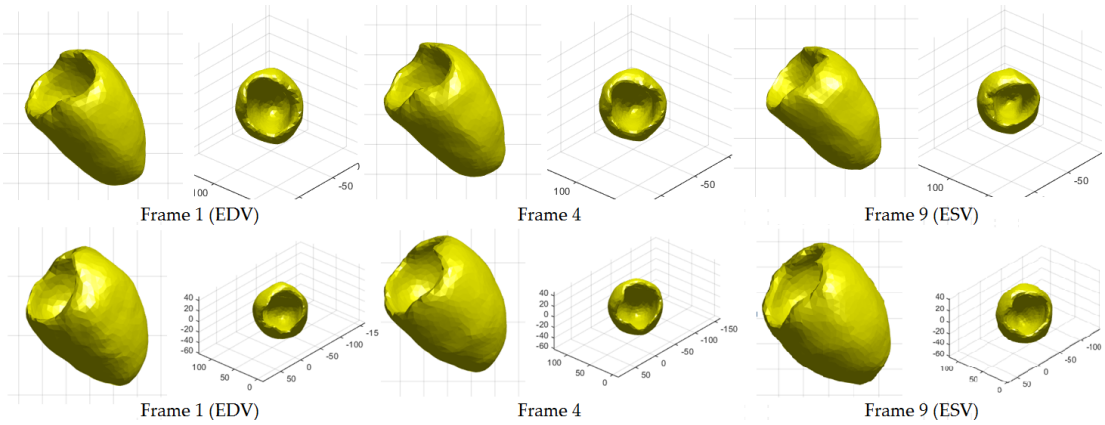

Figure 4.5: Two views of LV model at three frames: first row for a volunteer, second row for a patient.

FCN and the proposed methods, using Dice's coefficient as the evaluation metric for segmentation. The result in Table 1 shows that the proposed method has better performance than FCN, because the shapes of output contours are regularized. For the motion reconstruction, some manual adjustment of segmented contours is necessary in terms of accuracy, which takes a few minutes for each case, on average. Once the adjustment of contours is finished, we conduct the processing steps without any further update for the contours.

We also evaluated our methods with the public dataset from the cardiac MRI segmentation challenge of MICCAI 2009 [35], as well. The dataset, from the Sunnybrook Health Sciences Center, contains 45 cine SAX slices, covering both normal and abnormal cardiac conditions. Image size is 256×256 pixels, and its resolution varies from 1.2500 mm to 1.3672 mm. Expert annotations of endocardium and epicardium contours are provided for some slices at EDV and ESV phases. We only evaluated the cases where both endocardium and epicardium annotations are given for the same image. The dataset is divided into three subsets: training, validation, and online, following the standard nested cross-validation. We trained our model with the training set (135 images), evaluate the model with the evaluation set (138 images) and tested with the online set (147 images). Accuracy was measured with the Dice's coefficient, as well shown in Table 4.1. The accuracy is slightly less than previous experiments since the training set is fairly small.

The average distance between the contour points and the reconstructed model is utilized as the metric to evaluate the performance of the rigid alignment. The result for the whole dataset

Table 4.1: Evaluation results

Dice's coefficient	our dataset	challenge dataset
U-Net (mean)	0.70	0.53
U-Net (std)	0.07	0.15
proposed method (mean)	0.86	0.70
proposed method (std)	0.04	0.12

along the cardiac phase is shown in the Figure 4.6. We find that the distance at each time point is much smaller when applying alignment than that without any alignment. This means our alignment strategy well improves the consistency of contours in 3D space well. The model from the aligned contours is also improved, as shown in Figure 4.7.

Figure 4.5 shows the reconstructed shapes at different frames of the cardiac cycle. There is a clear difference between LV motions of normal volunteers and those of patients with heart dyssynchrony. In the ES phase, the LV contracts well to pump the blood out for normal people; whereas, it does not deform as much for patients, which means the patients' hearts are unable to function properly. Based on the reconstructed model, we can study the LV volumes along time for normal volunteers and patients shown in Figure 4.8. Comparing with normal people, the patient's LV contains more blood and it does not contract much during the cardiac motion (which also can be proved by the ejection fraction rate: 55% for a normal volunteer and 28% for a patient). 2D myocardium contours in tagged MR slices, which are useful for further studying the interior dynamics of the LV wall, can also be located and mutually registered, based on the reconstructed 3D LV model and its intersection with the MR planes (as shown in Figure 4.9).

Algorithm 1 Joint alignment of 2D MR short- and long-axis slices

Data: all 2D contours $\mathbf{v}$ on different image planes $\mathbf{T}$

Result: in-plane translation $(\delta x, \delta y)_s$ for each MR slice s

initial step coefficient $\gamma = 0.5$, initial gap threshold $\theta = 0.1$ initial $(\delta x, \delta y)_s = (0, 0)$ compute intersection points $\mathbf{p}$ of $\mathbf{v}$ and $\mathbf{T}$ compute the closest in-plane points $\mathbf{p}' \in \mathbf{v}$ to $\mathbf{p}$ iteration index $i = 1$, maximum iteration number $i_{max} = 100$ **while** $i \leq i_{max}$ *and* $\|\mathbf{p} - \mathbf{p}'\| \leq \theta$ **do**

  $i \leftarrow i + 1$ **for** *each slice* s **do**

    $\delta x_s \leftarrow \delta x_s + \gamma \cdot \sum (\mathbf{p}'_s - \mathbf{p}_s)_x$ $\delta y_s \leftarrow \delta y_s + \gamma \cdot \sum (\mathbf{p}'_s - \mathbf{p}_s)_y$ $\mathbf{v}_s \leftarrow \mathbf{v}_s + (\delta x, \delta y)_s$

  end

  compute intersection points $\mathbf{p}$ of $\mathbf{v}$ and $\mathbf{T}$ compute the closest in-plane points $\mathbf{p}' \in \mathbf{v}$ to $\mathbf{p}$

end

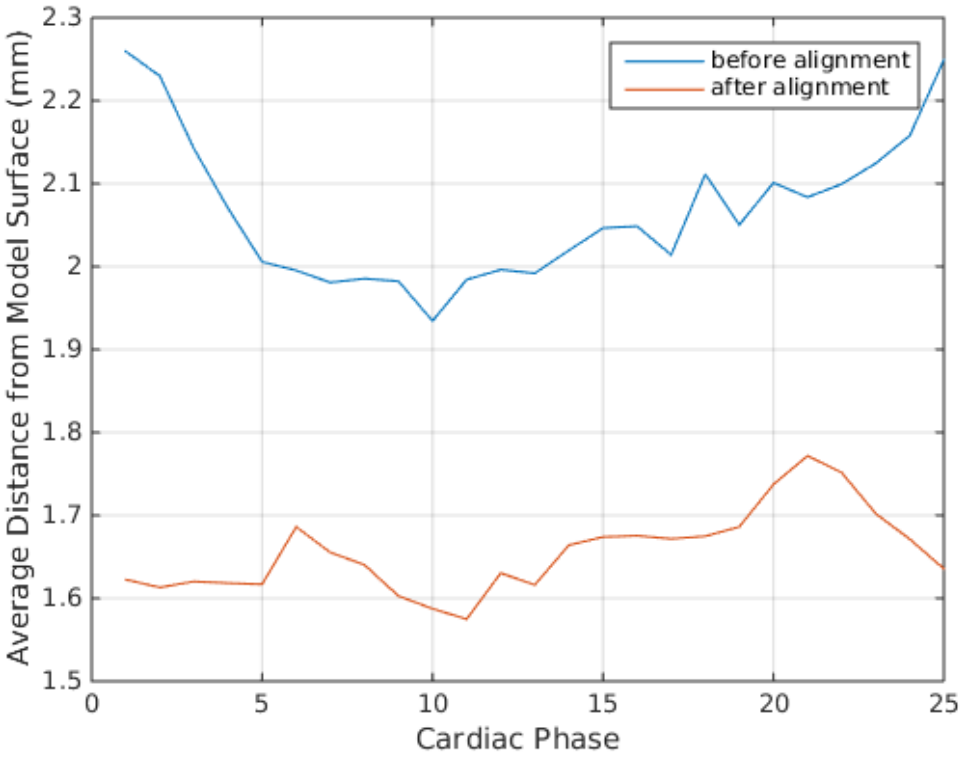

Figure 4.6: The average distance (in mm) between contour points and reconstructed model along the full cardiac cycle for the whole dataset.

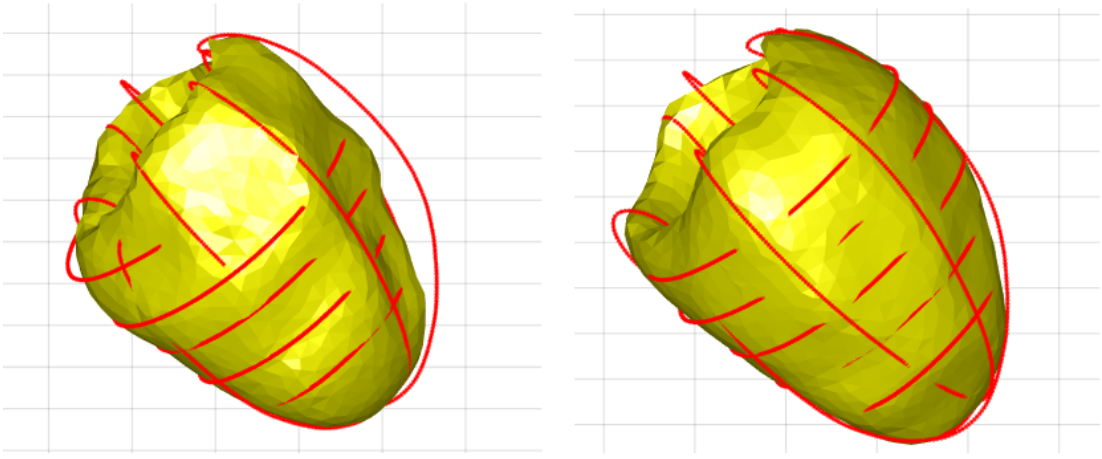

Figure 4.7: Left: the model reconstructed from the contours without aligned; right: the model from the contours with alignment. The model shape with alignment becomes more proper and smooth comparing result without alignment.

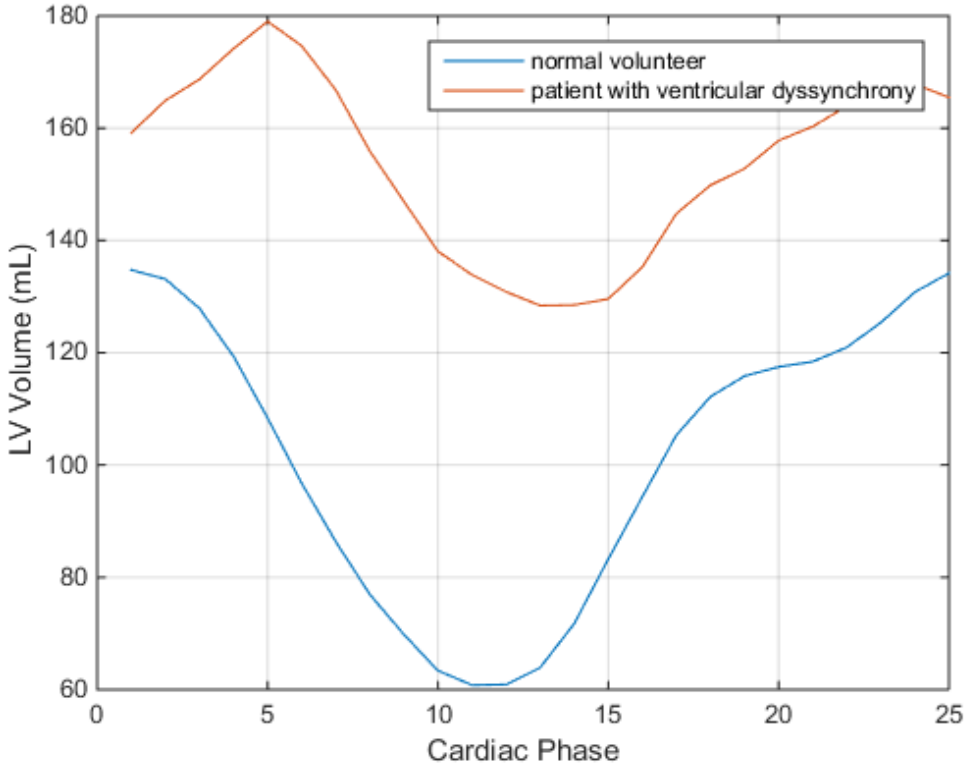

Figure 4.8: LV volume change along time within a full cardiac cycle for a normal volunteer and a patient with heart dyssynchrony.

Algorithm 2 LV wall motion computation over the whole cardiac cycle.

Data: all 2D contours $\mathbf{v}$ in space and time, a 3D reference shell shape S_0

Result: shapes of LV wall at all cardiac phases

compute initial ED shape S from S_0 to $\mathbf{v}_{ED}$ using non-rigid CPD initial threshold $\theta > 0$, initial overall displacement update $\Delta > \theta$ **while** $\Delta > \theta$ **do**

  compute the closest point sets $\mathbf{u} \in S$ corresponding to points in $\mathbf{v}_{ED}$ calculate forces based on difference $\mathbf{v}_{ED} - \mathbf{u}$ interpolate forces f for vertices $V \in S$ calculate $\dot{q}$ and update q $S' \leftarrow$ update surface mesh S using q $\Delta \leftarrow \|S' - S\|$, and $S \leftarrow S'$

end

$S_{initial} \leftarrow S$ **for** *cardiac phase t from ED to ES* **do**

  initial overall displacement update $\Delta > \theta$ **while** $\Delta > \theta$ **do**

    compute the closest point set $\mathbf{u} \in S_{initial}$ of points in $\mathbf{v}_t$ calculate forces based on difference $\mathbf{v}_t - \mathbf{u}$ interpolate force f_t for vertices in $S_{initial}$, calculate $\dot{q}$ and update q $S' \leftarrow$ update surface mesh $S_{initial}$ using q $\Delta \leftarrow \|S' - S_{initial}\|$ $S_t \leftarrow S'$ $S_{initial} \leftarrow S'$

  end

end

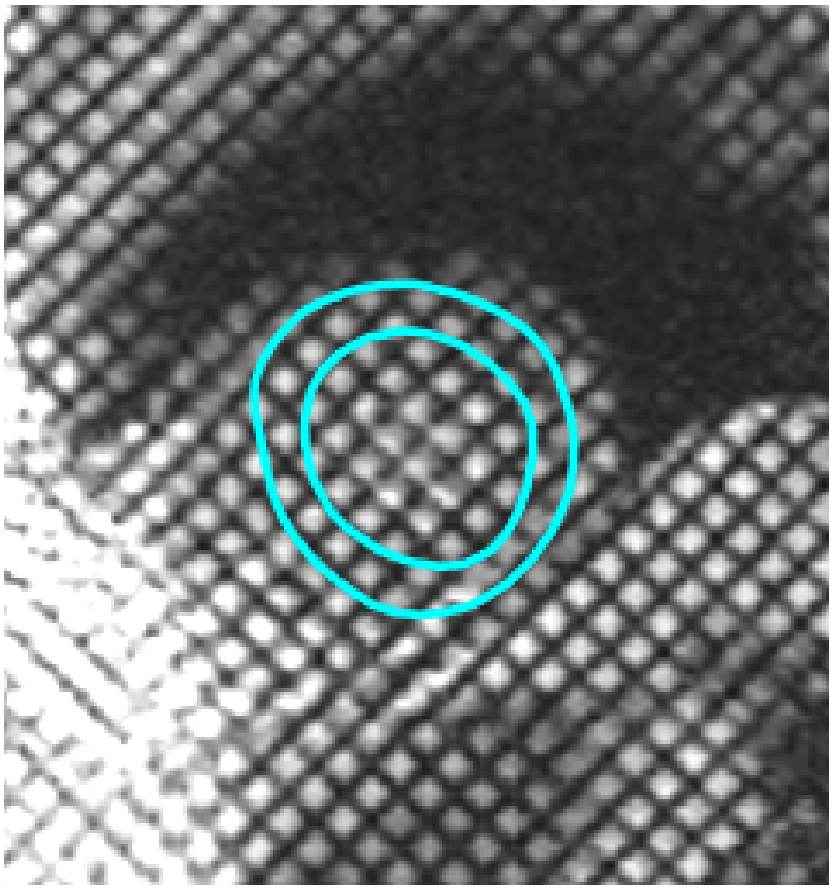

Figure 4.9: Intersected contours of fitted model on a tagged MRI slice.

Chapter 5

3D Blood/Muscle Segmentation using Generative Adversarial Network

5.1 Blood/Muscle Segmentation on 2D Cine MRI and Respiration Compensation

Initially, we adopt a LV segmentation framework using 2D/3D U-Net and multi-component deformable model enforcing the spatial and temporal smoothness on short-axis cardiac cine MRI [5]. The output from the framework is the epi- and endo-cardium contours during cardiac cycle. Similarly, we apply the framework to the long-axis cine MRI to extract the LV wall with the single-component deformable model. Once the LV contours are finalized from the segmentation framework and verified by doctors, they are adopted for generating 2D probabilistic segmentation and compensating artifacts caused by respiration.

Based on the epi- and endo-cardial contours, we adopted a novel approach to estimate the presence of boundary pixels close to trabeculae, papillary muscles, and solid wall, and to estimate the corresponding classification probability [5]. Initially, to compute the probability of belonging to blood or myocardium for the boundary pixels in the LV cavity, a 2D U-Net [31] is trained with pure LV cavity blood and pure myocardium regions (determined by intensity). Next, the features from the second-to-last layer of the U-Net are extracted from each pixel accordingly, and a logistic regression classifier is further trained using those features. The clear myocardium pixels are labeled as 1, and the clear blood pixels are labeled as 0. Finally, the trained classifier assigns the probability values to the pixels inside the transitional zones with fuzzy appearance between myocardium and cavity shown in Fig. 5.1. Normally, in the transitional zones, with mixed blood and muscle, we cannot reliably determine the regional proportion of blood or muscle from their appearance alone. Using the proposed approach, we are able to track the mixed spaces and estimate the regional percentage of blood/muscle

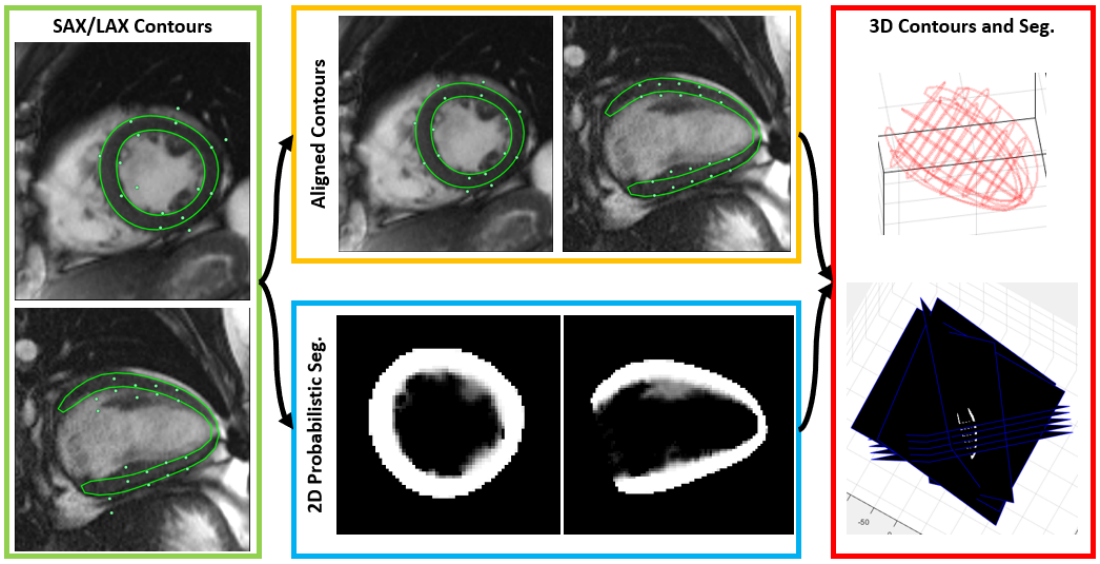

Figure 5.1: 2D probabilistic segmentation and respiration offset artifact removal, using in-plane contours for both short-axis (SAX) and long-axis (LAX) cine MRI that may initially not be well aligned. The output is the aligned contours and the aligned 2D probabilistic segmentation in space.

using the fuzzy classifier. By adjusting the threshold of the classifier within the range [0,1] at different cardiac phases (especially near end-systole), the sensitivity to the relatively small amount of blood near boundary pixels is increased. The adjustment of the threshold can be made based on the the criterion of maintaining a consistent amount of muscle at each cardiac phase.

The epi- and endo-contours are further utilized to remove imaging artifacts caused by inconsistent respiratory state between image acquisitions [15, 23]. The artifacts are seen as apparent mis-alignment between different cine MRI slices in space, which increases the difficulty of recovering the full 3D heart motion. Different cine MRI image slices are typically acquired at different suspended respiration phases, with associated different spatial offsets, even though they are all synchronized by ECG according to cardiac phases; Conventional breath-holding MRI cannot completely diminish this effect, even with cooperative subjects. We assume that the mis-alignment primarily causes an in-plane translation for 2D MR images. Thus, we iteratively minimize the overall distance between the intersectional points of contours recovered from slices perpendicular to a slice and contours within the slices, by serially translating all contours within the planes of their slices. After several iterations, the intersecting contour

points would approximately meet the in-plane contours. When we have thus derived estimates of the translation vectors of the MRI slices that best align the contours,, we then apply them to shift the aforementioned 2D probabilistic segmentations for better spatial consistency between slices.

5.2 3D Label Propagation using Generative Adversarial Network

In order to propagate the label information (probabilistic segmentation) of 2D slices to the whole 3D volume, we propose a generative adversarial network (GAN) [16] based model for the propagation. The proposed GAN based model is shown in Fig. 5.2. The generator (G) is fed with the 3D volume containing multiple 2D slices of probabilistic segmentation, denoted by x. The in-plane label is probability of belonging to myocardium region of each pixel, obtained from Section 2, the voxels outside slices are valued to be 0 by default. With the 3D U-Net, we predict the label for the whole 3D volume. The discriminator (D) is fed by ground truth of 3D volume label as real sample, and the output of the G as fake sample. It aims to discriminate the ground truth from the prediction results. The learning objective is designed to be the weighted combination of 1) the prediction error of G, evaluated by the MSE error with ground truth; 2) the minimax loss of the GAN model. Those two parts are weighted by λ, which is

$$\begin{aligned}\min_G \max_D V(G,D) =& \mathbb{E}_{x\sim p_{data}(x)}[\log(D(x))] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))] \\ &+ \lambda L_{MSE}(Y, G(z))\end{aligned} \tag{5.1}$$

where $\mathbb{E}$ is the empirical estimate of expected value of the probability; Y denotes the ground truth of the 3D volume (probabilistic segmentation). The model parameters in G and D are learned by iteratively maximizing $V(G,D)$ with respect to D and minimizing $V(G,D)$ with respect to G. After the model learning, the 3D U-Net in G is used for label propagation. The probabilistic estimation of myocardium region is obtained from label propagation result.

For the 3D U-Net in generator G, we adopt the structure defined in [32], with the same layer number and convolutional filter number in each layer.The 3D CNN in discriminator D is designed to be a 7-layer convolutional neural network including five convolutional layers, single fully connected layer, and a sigmoid layer. For both networks, batch normalization is used between each two neighboring layers and leaky ReLU are used as the activation functions.

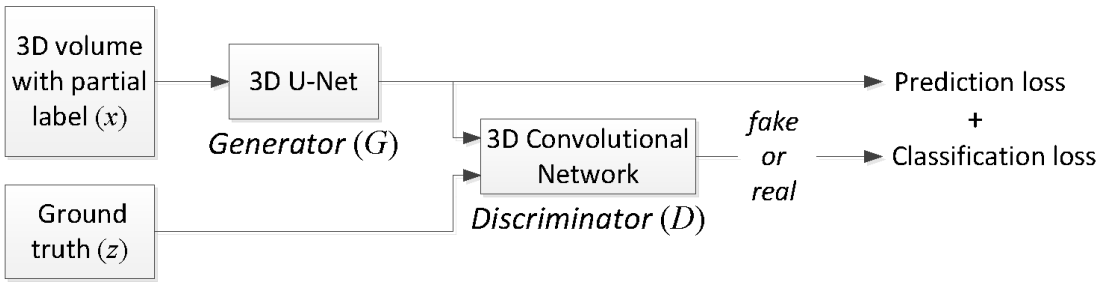

Figure 5.2: The proposed GAN model for label propagation.

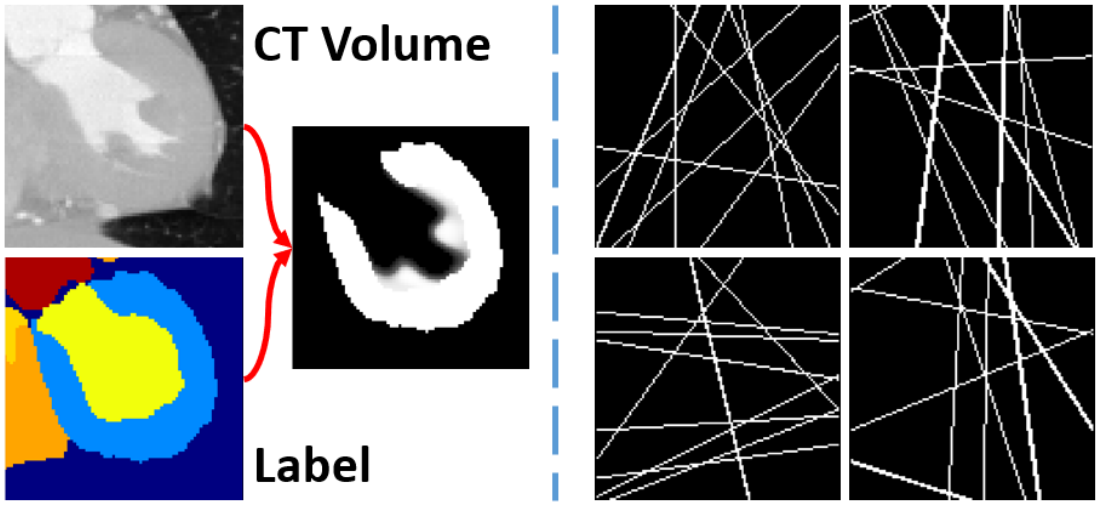

Figure 5.3: Left: synthetic data of 3D probabilistic segmentation using CT volume and its label; right: four samples of pattern masks used for synthetic data (probabilistic segmentation of CT volumes).

5.3 Experiments

There is no ground truth for 3D probabilistic segmentation for 2D MRI acquisitions. Because the 2D MRI is sparsely sampled around the LV area. So we adopt a synthetic dataset from CT to train our neural networks. The CT dataset contains both appearance volumes and the corresponding labels, including both blood pool and myocardium [40]. All volumes and labels are re-sampled to 1.0 mm isotropically for simplicity. We follow the similar strategy in Section 2 to generate ground truth of 3D probabilistic segmentation, and use a 3D U-Net instead for feature extraction. Then, we randomly sample 10 planes (with arbitrary angles and locations) crossing the volume as a pattern, and make them the binary masks with the same size shown in Fig. 5.3. The input to the neural network is the element-wise multiplication of 3D probabilistic segmentation and a pattern, and the output is the probabilistic segmentation itself. The

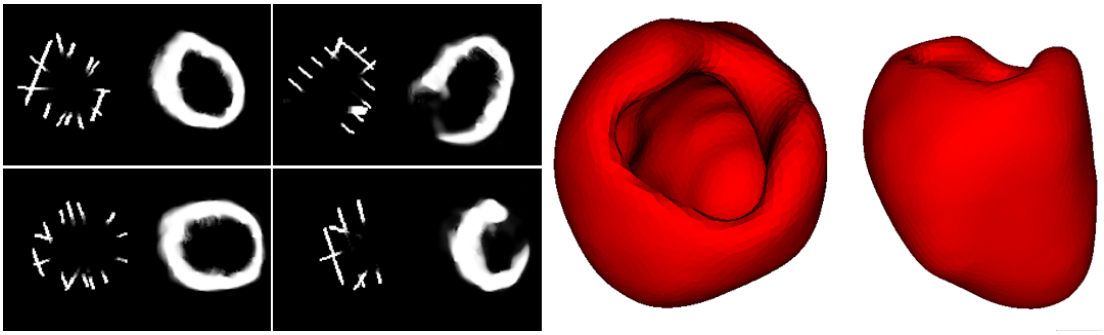

Figure 5.4: Left: the cross sections of 2D probabilistic segmentation maps in space before and after 3D reconstruction; right: 3D myocardium model from the 3D probabilistic segmentation with threshold 0.5.

problem formulation is similar to the task of image completion. During training, we have 15 segmentation maps and 1000 different patterns, which gives us 15000 training samples. We collect another 5 segmentation maps for testing, and in total we have 5000 testing samples. The training details follows the setting in [35]. λ is set as 1000 in our experiments.

In Table 5.1, we can see that our proposed method using GAN performs much better than the baseline 3D U-Net in terms of peak signal-to-noise ratio (PSNR) and mean squared error (MSE). If we provide more planes of probabilistic segmentation (e.g. 15 planes) as input to the neural network, and the output would be closer to the ground truth. It fits the intuition that more information introduces better local and global constraints. After the model is finalized, we deploy it to the real 2D cine MRI dataset. The multiple 2D probabilistic maps in space as a volume are directly treated as input of the generator. The output is full 3D probabilistic segmentation shown in Fig. 5.4. We can achieve myocardium segmentation simply by placing a threshold to the 3D probabilistic segmentation.

We applied our generative adversarial network in the domain of 3D probabilistic segmentation, instead of CT/MRI volume domains, because the synthetic training data is difficult to generate in terms of appearance. Both 2D cine MRI and high-dose CT have excellent image quality, from which the details of trabeculae can be well-observed. Directly acquired 3D cardiac MRI commonly suffers from respiratory motion, and the appearance is generally more blurry and noisy with artifacts, compared to 2D cine MRI. However, we are able to obtain 3D probability segmentation from CT volumes, and can learn the generative model using it. The generative model perfectly fits the domain of 2D probability segmentation from 2D cine MRI

PSNR (dB)			
	5 Planes	10 Planes	15 Planes
U-Net	13.57	14.13	14.61
GAN	**15.01**	**15.22**	**15.34**
MSE			
	5 Planes	10 Planes	15 Planes
U-Net	0.0442	0.0388	0.0348
GAN	**0.0331**	**0.0316**	**0.0301**

Table 5.1: Evaluation on the test dataset.

in space. Therefore, it would be ideal to study the 3D cardiac motion using 2D cine MRI in this way (normally 3D CT acquisitions have low temporal resolution).

Fig. 5.5 shows the calculated myocardial volumes at different cardiac phases, using two different methods, from 2D MRI sequences of a patient with cardiac dyssynchrony. The orange curve is from the volumes of shells built using 3D deformable models recovered with simple thresholding [15], and the blue curve is computed by summing up myocardium probabilities over all voxels. We can observe that our 3D probabilistic segmentation clearly has larger volumes comparing to the deformable shell, because our probabilistic model is able to capture the fractional volume components of trabeculae and papillary muscles in the transition zone inside LV cavity. Moreover, the volume quantity of the probabilistic segmentation is more stable temporally, which meets the assumption that myocardium volume should be close to a constant value during cardiac cycle. The appearance of our 3D probabilistic segmentation is qualitatively verified by clinicians, and it has a potential for improved basic understanding and clinical study of cardiac function (e.g., improving the estimation of ejection fraction and related global measures).

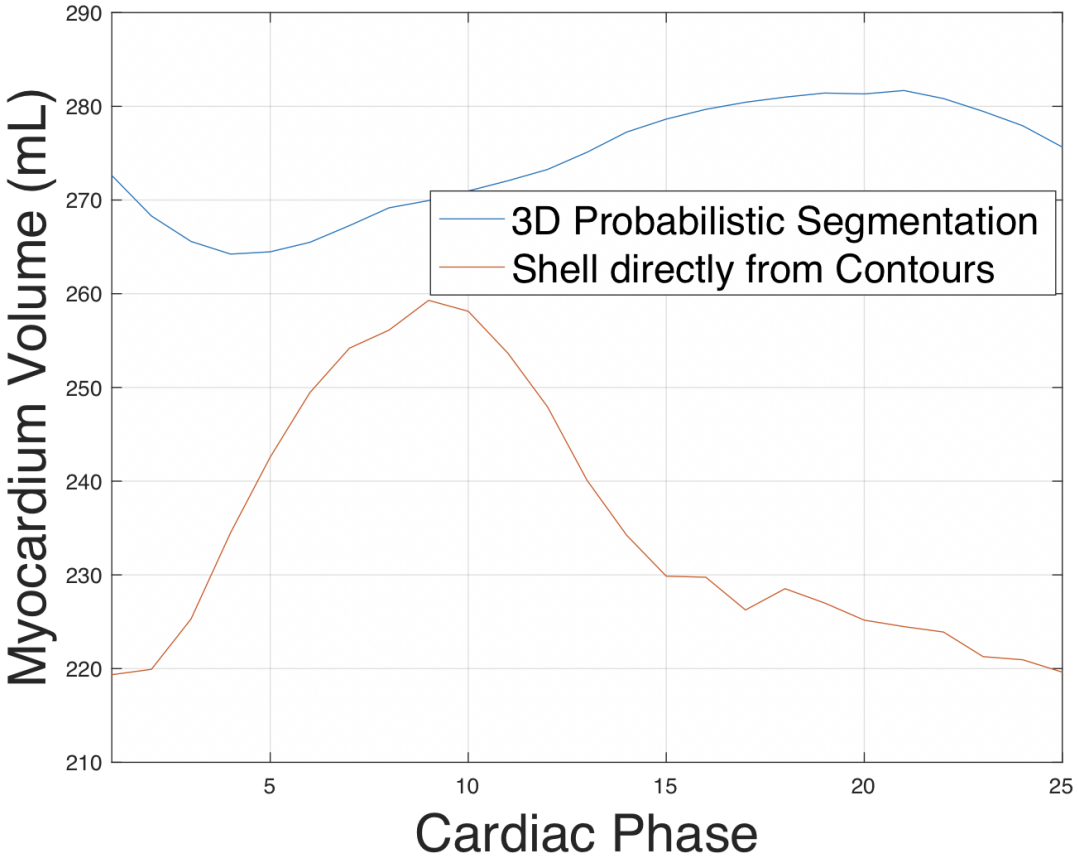

Figure 5.5: LV myocardium volumes at different cardiac phases, for conventional and probabilistic segmentations.

Chapter 6

3D Motion Field Reconstruction and Assessment of Ventricular Dyssynchrony

6.1 3D Motion Reconstruction

For the motion reconstruction, we start from the 2D LV myocardium segmentation using 2D-3D neural networks [5]. Slight manual correction for the extracted contours is required based on the guidance of radiologists and cardiologists. Because none of the segmentation methods could achieve perfect results without downgrading the performance of latter motion reconstruction. Then, the respiration offset, caused during acquisition, is compensated using rigid transformation [15]. The 3D shell model is created using a deformable shape model at each cardiac phase [15].

Meanwhile, we train an individual U-Net [31] to compute dense displacement field between neighbouring phases for each subject. The problem is defined as an unsupervised learning problem, and we use the U-Net model to "overfit" the MRI image pairs. When optimization is finished, the network is able to generate 2D displacement field (no prior assumption in the displacement except necessary smoothness) for current subject. Our proposed network model h follows the design of U-Net shown in 6.2. It takes the pair of images I_t, I_{t+1} at two neighboring phases as input, and produces 2-channel output corresponding to X, Y values of the dense displacement field. We add a sampler at the end of the network after the displacement field is generated. The sampler warps I_t to I'_t towards I_{t+1} through differentiable bilinear interpolation. Then the difference between I'_t and I_{t+1} is treated as optimization target. The training loss consists three components listed as follows.

$$l = \lambda_1 \left\| I'_t - I_{t+1} \right\|_1 + \lambda_2 \left\| h\left(I_t\right) \right\|_2 + \lambda_3 \left\| h\left(I_t\right) \right\|_{TV} \tag{6.1}$$

The first component is the standard image reconstruction loss, and minimizing l_1 loss keeps

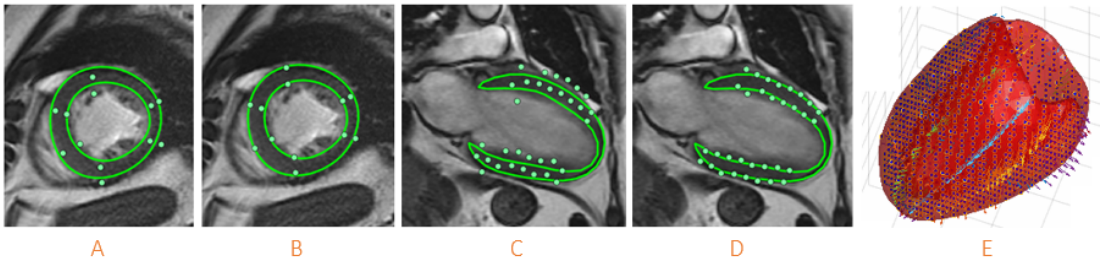

Figure 6.1: **A/C**. Image segmentation; **B/D**. segmentation after alignment, the green dots are the intersection points from other contours; **E**. interpolated 3D displacement field.

the majority of the high-frequency parts of images. The second component is the Euclidean loss on the magnitude of displacement to suppress unreal large displacement. And the third component is the total-variation loss on displacement $h\left(I_t\right)$ to make sure the displacement is locally smooth.

Based on the aligned contours in 3D space, a deformable shape model is adopted for 3D shell model reconstruction from sparse contours [15]. Apparently, the reconstructed shell at each cardiac phase does not have actual vertex correspondence with each other. Based on the shell information, we use the 2D in-plane displacement and conduct interpolation for 3D displacement field. Each sampled grid p inside shell gathers the displacement from all other grids within the neighborhood δp. The deformation d_p is the weighted sum of all displacement vectors from neighbors. The weights are the reciprocal of distances between the grid itself and p with normalization. The whole process is repeated for several times until the displacement field is converged. Therefore, the dense point correspondence (3D displacement field) is built approximately for the shell models of neighboring frames.

6.2 17-Segment Shell Model Analysis

The American Heart Association (AHA) writing group on myocardial segmentation and registration for cardiac imaging introduced a 17-segment shell model of the left ventricle to study the regional activities for various cardiovascular diseases, given several imaging techniques [20]. We follow the same way to divide our LV shell model into sub-regions. Additional annotation is further required to distinguish septum and free-wall regions given an LV shell model. Here we use the contours of LV and RV muscle boundaries within a mid-level short-axis plane, and the

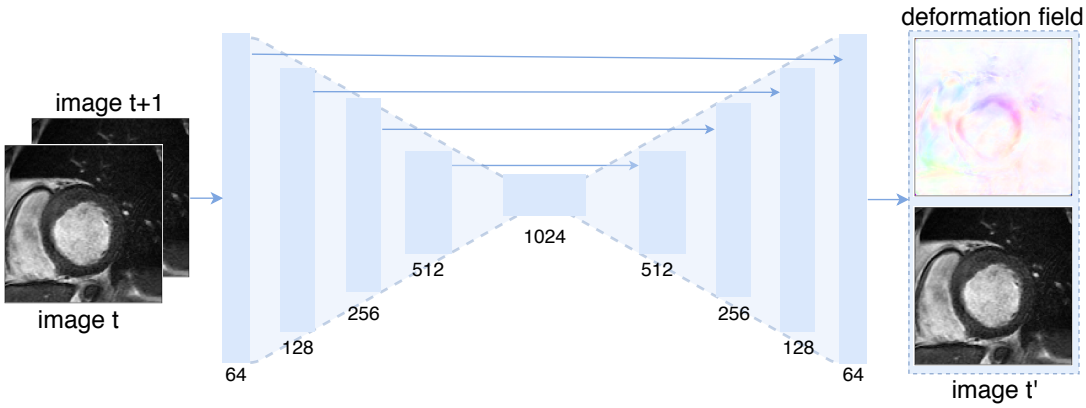

Figure 6.2: The convolutional encoder-decoder we used for displacement field estimation. The numbers next to convolutional layers indicate the quantity of convolution kernel. Image t' is the warped image t using dense displacement field for the training loss computation. The images on the right are the color-coded displacement field, and warped image I_t. We notice that the majority of motion is around the myocardium muscle, which meets our expectation.

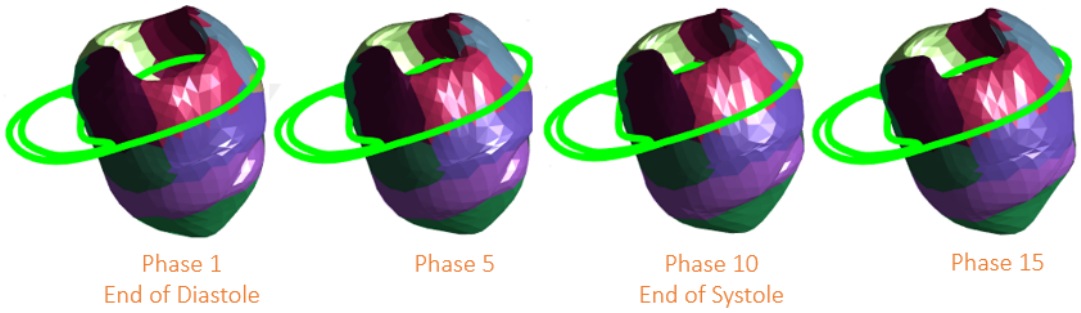

Figure 6.3: The 17-segment shell model of LV at different cardiac phases from a patient data. The green contours indicate boundaries of LV and right ventricle (RV) muscle in the mid-level short-axis plane. Regional colors represents non-overlapping segments of LV wall.

region between LV and RV blood pools is clearly the septum region. Based on the location of septum, we can easily separate the shell model into 17 segments. The motion of each segment is tracked using the computed 3D displacement field. Then we study the regional information based on the segment-wise displacement.

6.3 Experiments

Dataset We utilize a cardiac cine MRI dataset, consisting of 50 patients with potential ventricular dyssynchrony (QRS value greater than 120) with CRT outcome, from GE medical systems. For each subject, the dicom dataset contains 10$\sim$12 short-axis planes during cardiac cycle, and 3 long-axis planes (2-, 3-, and 4-chamber view). In-plane resolution of images

Table 6.1: The Dice's score and Hausdorff Distance of the proposed U-Net displacement model and other methods

Methods	Dice	Hausdorff Distance
Neighbouring	0.9129 ± 0.0631	2.5841 ± 0.5209
Lucas-Kanade	0.9079 ± 0.0551	2.8530 ± 0.4621
U-Net	$\mathbf{0.9250 \pm 0.0427}$	$\mathbf{2.4392 \pm 0.4649}$

ranged from 1.25 mm to 1.33 mm, and size varied from 240×180 pixels to 256×256 pixels. The full cardiac cycle consists of 20$\sim$30 phases.

Methods Because it is almost impossible to obtain the golden standard of dense displacement from cardiac MRI, the evaluation of displacement field computation is conducted using segmentation accuracy between neighboring images of one subject's full MRI scans. We compared our U-Net displacement model with two baseline methods in terms of two common segmentation metrics: Dice's score and Hausdorff Distance. Specifically, Neighbouring is a naive method that simply calculating the metrics using original segmentation of neighbouring phases. Lucas-Kanade [41] is a differential method to compute optical flow. And we first estimate optical flow between image pairs, and then compute the metrics based on warped segmentation mask for the target image. Our method calculates the metrics between the ground truth segmentation mask and the estimated warped segmentation mask. Particularly, to achieve the best accuracy, we set $\lambda_1 = 1.0$, $\lambda_2 = 50.0$ and $\lambda_3 = 0.01$ of Eq. (6.1). We adopt Adam optimizer with initial learning rate of 10^{-4} with decreasing rate of 0.5 at the 20-th,40-th and 60-th epochs. The batch size is 16 and the maximum number of epochs is 150.

Validation with segmentation The results are reported in Table 6.1 and we mainly observe two following aspects. First, our method (U-Net) achieves the highest Dice's score of 0.9250 and the smallest Haudorff Distance of 2.4392 among all tested approaches. For example, the Dice's score of our method is 0.0121 higher than that of the second highest Neighboring method. The Hausdorff Distance of our method is 0.1449 smaller than that of the second best Lucas-Kanade method. Second, the standard deviation of our method is the smallest in Dice's score and comparable to others in Hausdorff Distance. This indicates that our method is more robust compared with other approaches, which is also an important factor in practical medical applications.

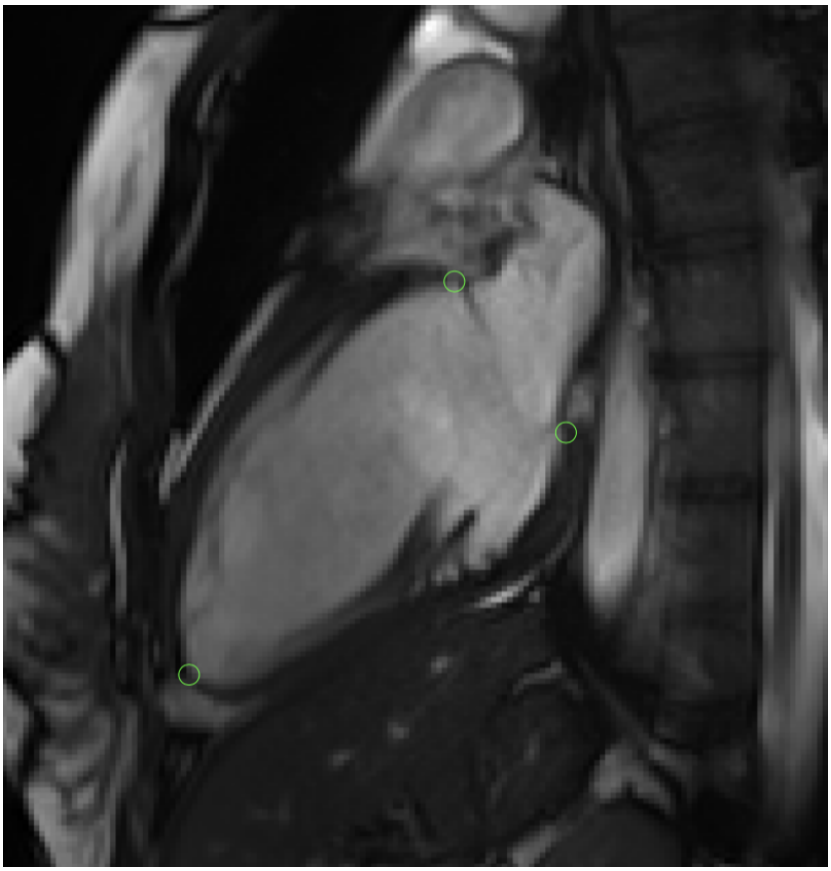

Figure 6.4: Anatomical landmarks (circled) on LV wall: two mitral valve points, and one LV apical point.

Table 6.2: The Euclidean distance errors (mm) of the proposed U-Net displacement model and other methods

Methods	Mean	Median	Max.	Min.	90%
Neighbouring	1.6300	1.4577	9.4240	**0.0000**	2.9546
Lucas-Kanade	1.1945	0.6422	18.3515	0.0673	2.3761
U-Net	**1.0194**	**1.0272**	**1.4232**	0.6718	**1.3803**

Validation with landmark localization To further validate our proposed approach, we adopt the task of landmark localization in 2D cine MRI images to track the landmark between neighboring cardiac phases. Since segmentation task is an indirect way to validate the displacement output without necessary point-to-point correspondence, landmark location tracking is preferred to validate the point-to-point correspondence. As shown in Fig. 6.4, three pre-defined anatomical landmarks are manually annotated by experts: two mitral valve points, and one LV apical point. Here we assume the landmarks remain the same physical points at the entire cardiac cycle, which means the through-plane motion of LV is neglected. We firstly initialize the landmark locations (ground truth locations) for 2D long-axis (LAX) MRI at one phase, and track the movement of landmarks for the next cardiac phase. The validation metric is the Euclidean distance between the ground truth landmark locations and moved landmark locations from the previous phase given the displacement field (output of 2D encoder-decoder). The final results are shown in 6.2. From the table, our proposed approach outperforms the baseline approaches, with the minimum average distance error, the minimum median distance error, and the minimum of the largest distance error. Combining the validation of both segmentation and landmark localization, our proposed approach is validated to be effective and accurate.

Dyssynchrony analysis Based on the 17-segment model, we track the radial movement of each segment at cardiac cycle according to LV axis. In our dataset, the acquisition starts from end-diastolic volume (EDV), and ends at the same phase. According to the normal ventricular functioning mechanism, we expect that each segment would move inward together, and outward later after the end-systolic volume (ESV) for the regular cases. Then the regional dyssynchrony pattern can be easily determined. In general, there are mainly two categories for the dyssynchrony patterns detected in our dataset. First, one or two segments move irregularly. For instance, we study one patient's data shown in Fig. 6.5. So clearly the 14-th segment works differently than others, and that segment is corresponding to apical septal. Second, the majority of segments moves much less than the regular patterns. We can see it from another example shown in Fig. 6.6. The 12-th segment moves unexpectedly, and that segment is corresponding to mid anterolateral. There are also some segments not moving much at the cardiac cycle, which is also a reason causing irregular QRS values.

Dataset We evaluate our approach on a cardiac cine MRI dataset that contains 22 normal

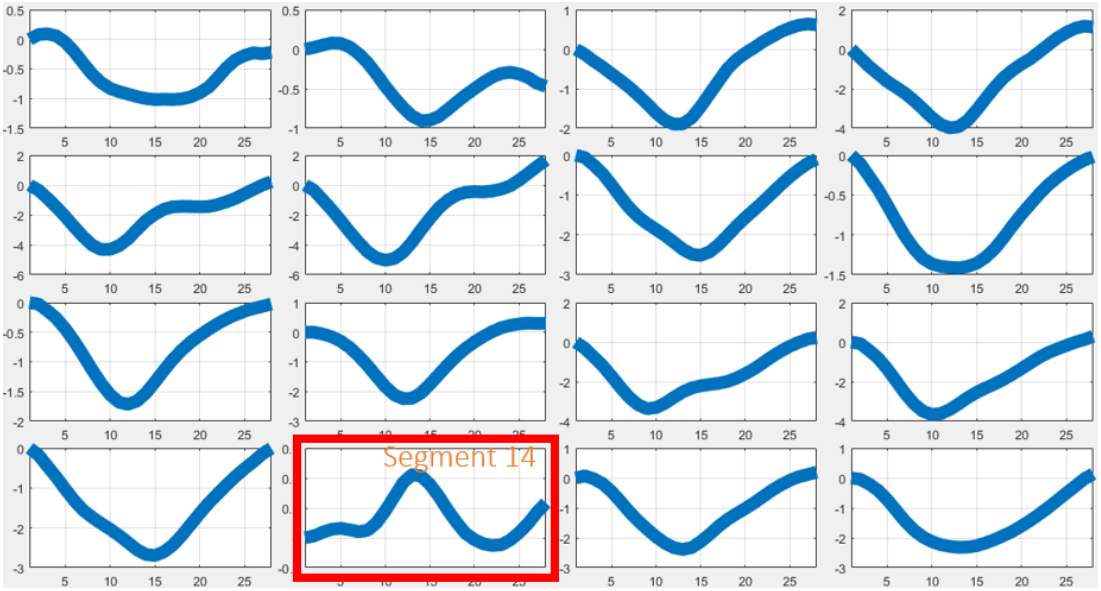

Figure 6.5: Radial movement for segment 1 to 16 from one patient. X-axis represents cardiac phase. Clearly the abnormality is from the 14th segment.

people and 34 ventricular (LBBB type) dyssynchrony patients, whose QRS value are greater than 120 and treated with CRT. In every cardiac cycle, there are around 10$\sim$12 short-axis planes and 3 long-axis planes (2-, 3-, and 4-chamber view). Image is of size varied from 240×180 pixels to 256×256 pixels and with in-plane resolution from 1.25 mm to 1.33 mm. Each cardiac cycle contains 20 to 30 phases. All the analysis results are based on re-sampling cardiac cycle to 20 phases for the purpose of fair comparison. We start to analyze the normal and those with LBBB type dyssynchrony. Meanwhile we have the outcomes of CRT in terms of categories 1, 2, 3. Category 1 means ”not improved”, category 2 means ”remain the same”, category 3 means ”improved”. The outcome of CRT is mostly conducted by asking feeling of patients after treatment.

We set things up so as to be able to analyze the cine data in these cases, so as to be able to objectively show which areas moved first in the images of the ectopic beats. The initially outward motion of the remote areas followed by inward motion there and outward motion of the initially activated areas, as novel sort of generalized septal flash. It is fairly common to qualitatively observe such feature in some cases. Our hypothesis is that the patients that didnt get improved (category 1-2) from CRT treatment exhibit pumping pattern (or patterns) that differs from the patients that experienced improvement (category 3). When it comes to

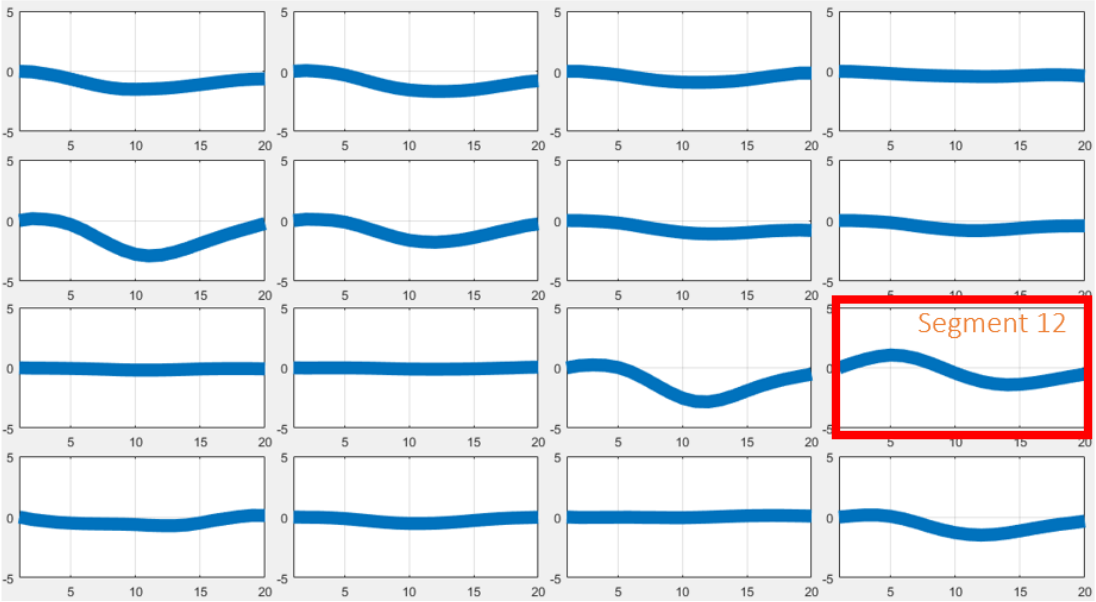

Figure 6.6: Radial movement for segment 1 to 16 from another patient. The abnormality is from the 12th segment.

insertion of the leads, the operators have done what they thought was best at the time.

From the Fig. 6.7, we can see that there are intrinsic difference in various regions. Three curves representing average values of normal subjects and patients in three categories. For example, at mid anteroseptal segment, the patient benefiting from the CRT treatment would be the one with larger regional contraction. Similarly we can find other types of difference in various regions. Fig. 6.8 shows the cavity volume change at the cardiac cycle. Then, the patients heart has more contraction capability would benefit from the CRT treatment. Fig. 6.9 indicates the apex displacement according to its location at the first phase (EDV). The red curve (category 1) has the smallest temporal displacement, which meets our expectation.

There are some interesting things to see, even with this small initial data set: (a) Although the stroke volumes are similar, the initial volume of the patient is higher, and the associated EF is lower, as expected for conventional systolic heart failure patients. (b) The delay in the drop in the volume of the patient LV may reflect the delayed opening of the aortic valve that is often part of LBBB dyssynchrony, if this is a LBBB patient. (c) The behavior of the diastolic filling phase will also be very interesting to characterize; that looks to be different between the normal and patient curves that presented here.

In the patient case we can see that the LV cavity starts to get smaller later compared to the

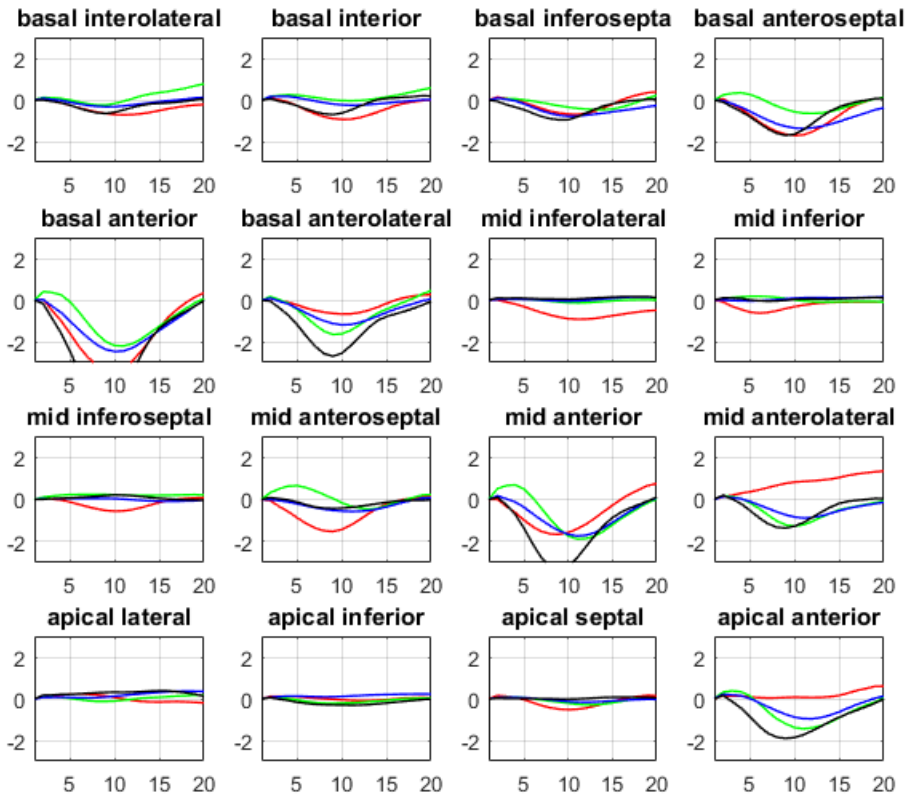

Figure 6.7: Regional radial distance towards LV axis for each segment. Different colors indicates average values for specific categories. Red curve represents category 1, green curve represents category 2, blue curve represents category 3, and black curve represents normal subjects.

normal subject, and this is because the build-up of pressure inside the cavity is delayed. This is very difficult to discern other than looking at when the aortic valve opens ie this is a measure of delayed opening of the aortic valve. Also, we see that the patient almost does not have any plateau during diastole. This can either be due to a relatively higher heart rate, or, as we talked about, diastolic dysfunction.

The pattern of the dyssynchronous motion is likely to be different for both the different cycle lengths and for different positions within the ventricle. That is, the initial, normally excited beat (which may actually be shorter if it is interrupted by a PVC) would be expected to have a uniform and prompt synchronous contraction pattern (assuming a normal conduction system), while the subsequent PVC beat in a bigeminy cardiac cycle pattern would have a more

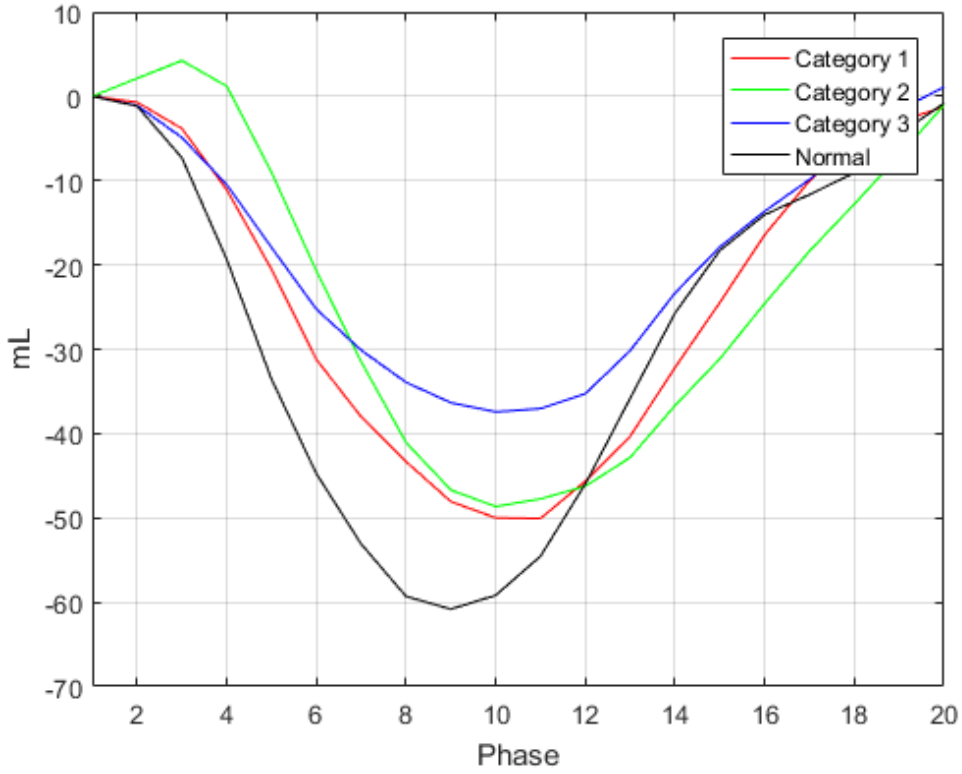

Figure 6.8: The volumes of LV cavity at the cardiac cycle. Different colors indicates average values for specific categories.

prolonged excitation and dyssynchronous contraction pattern. Furthermore, there are likely to be regional differences in the evolution of the contraction in different regions (although not in different image orientations at a given region). Thus, the site of the earliest inward motion may actually more outward when the rest of the ventricle contacts, similar to the septal flash phenomenon we can see in LBBB conduction system abnormalities. The specific temporal/spatial contraction pattern seen may depend on the site (and likely timing) of the initial excitation. However, in this case, it is due to the ectopic excitation spreading around the ventricle through the muscle system rather than the specialized conduction system, even though there is no intrinsic abnormality of the conduction system.

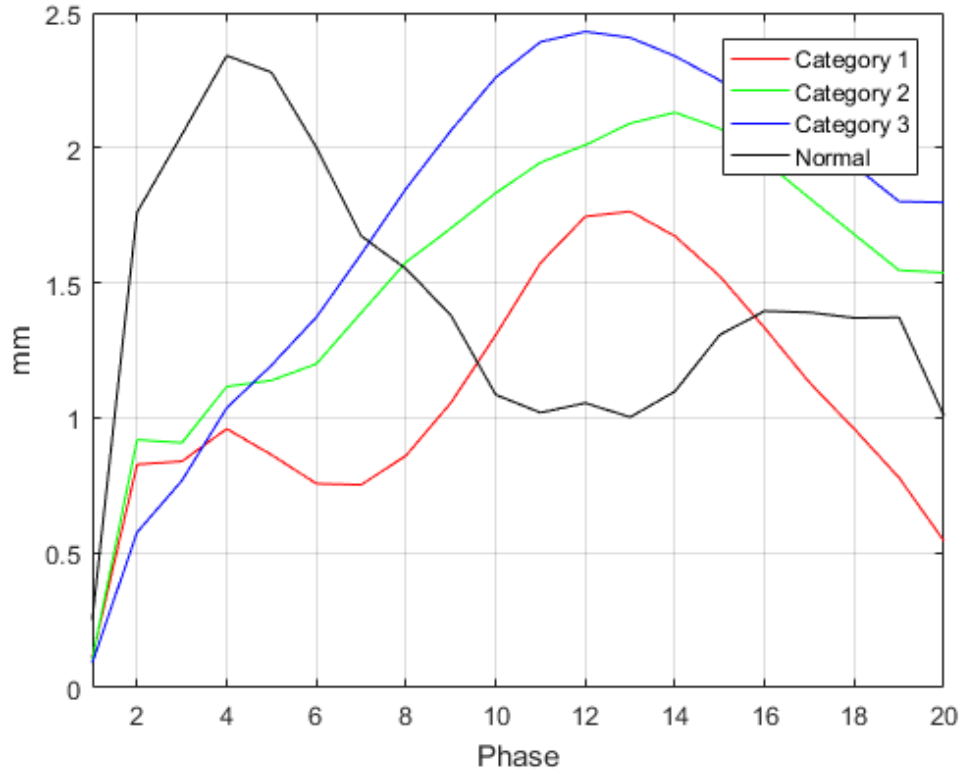

Figure 6.9: The displacement distance of LV apex at the cardiac cycle. Different colors indicates average values for specific categories.

6.4 Case Analysis of Different Categories

Here we describe the detailed analysis for the several subjects with different categories of CRT outcomes. For each individual categories, we describe the common patterns in-between different patient data. With the detailed understanding of the LV motion of patients, the study will help doctors/physicians to make better decision about pace maker placing or CRT treatment. Similar approach can be adopted for the analysis of other cardiovascular disease.

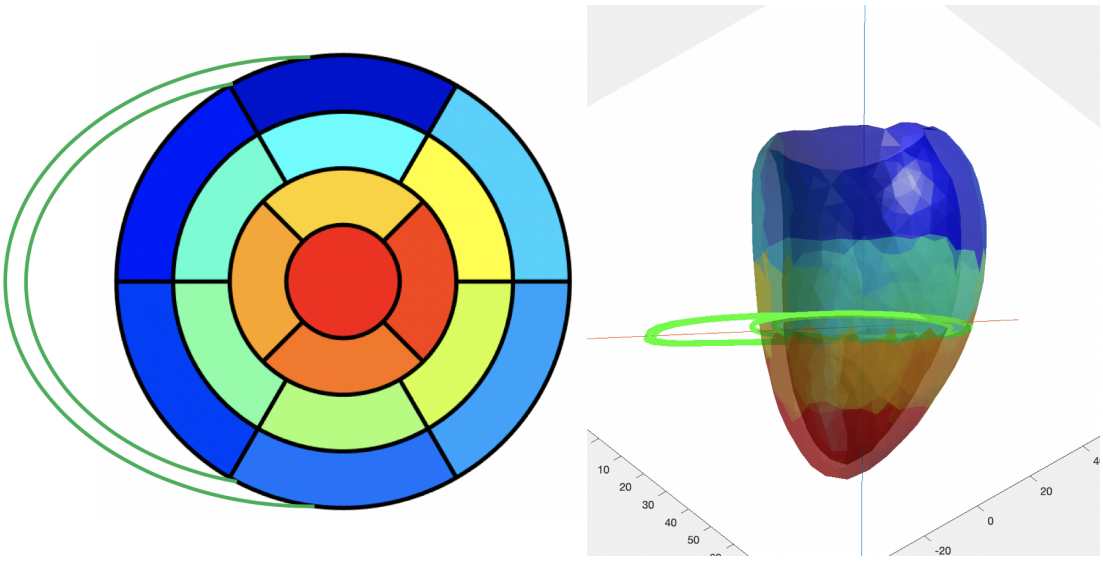

Figure 6.10: Left: the AHA 17-segment LV model, the green contours are the boundary of LV and RV myocardium; right: The transparent visualization of 17-segment model in 3D space. Both inner and outer wall are visible. And the myocardium contours are for dividing 3D models into 17 segments through coarsely defining septum. The blue axis is shown as the LV axis.

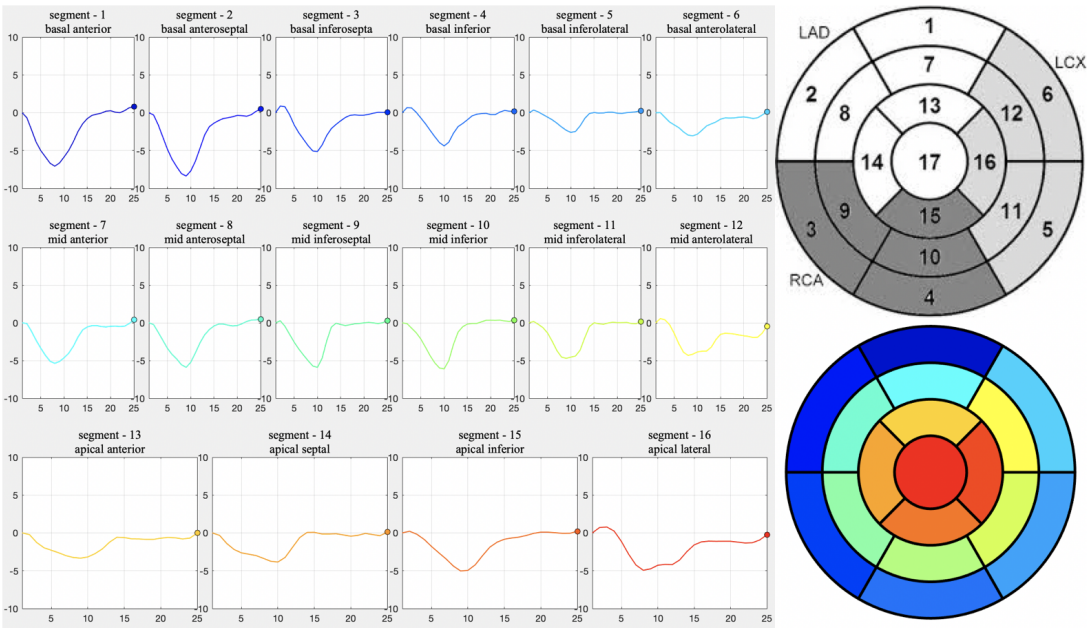

Figure 6.11: Regional motion of normal subjects. Left: the radial distance of 16 segments between itself and LV axis at the entire cardiac cycle. Positive value means moving outward, and negative value means moving inward. Right: 17-segment definition as reference. The colors indicate the correspondence with the results in the left figures.

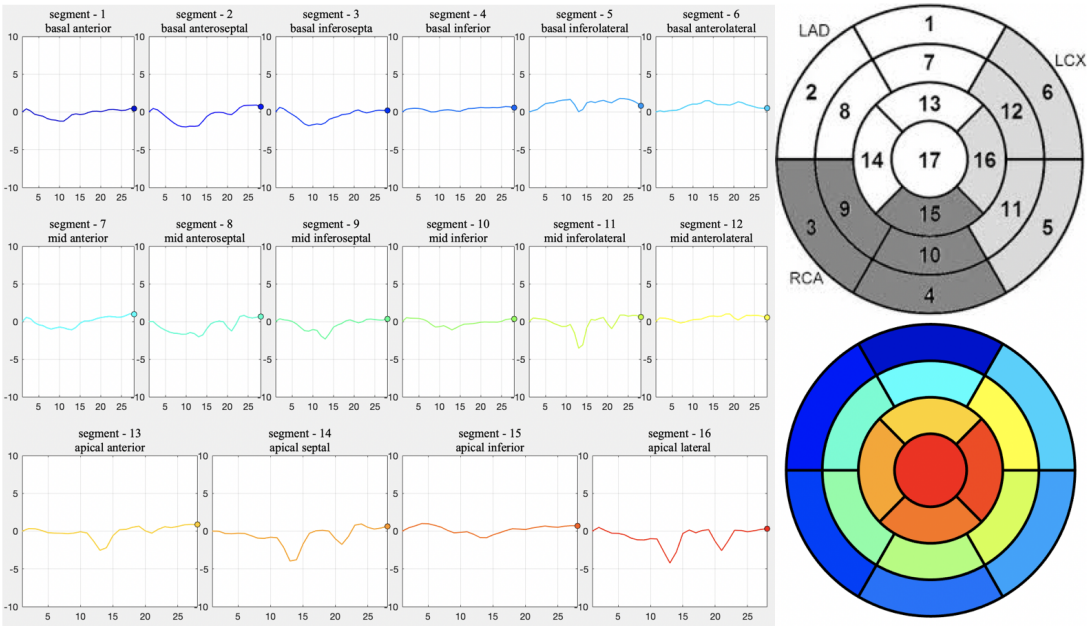

Figure 6.12: Regional motion of one category 1 patient (not improved). Left: the radial distance of 16 segments between itself and LV axis at the entire cardiac cycle. Positive value means moving outward, and negative value means moving inward. From the figure, it is clear that not every segment moves inward simultaneously at the ES phase. From segment 14 and 16, we can notice the rapid apical rocking (twice). It is the common pattern for patient belonging to this category. Right: 17-segment definition as reference. The colors indicate the correspondence with the results in the left figures.

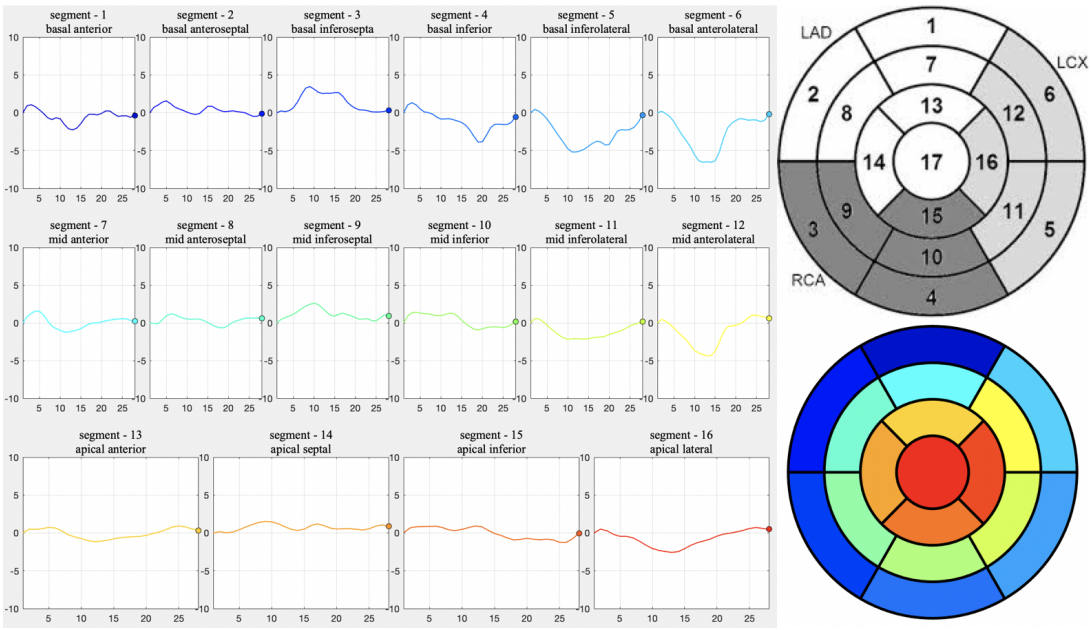

Figure 6.13: Regional motion of one category 2 patient (remain the same). Left: the radial distance of 16 segments between itself and LV axis at the entire cardiac cycle. Positive value means moving outward, and negative value means moving inward. From the figure, it is clear that not every segment moves inward simultaneously at the ES phase. From segment 3 and 6, we can notice the septal flash. It is the common pattern for patients belonging to this category. Right: 17-segment definition as reference. The colors indicate the correspondence with the results in the left figures.

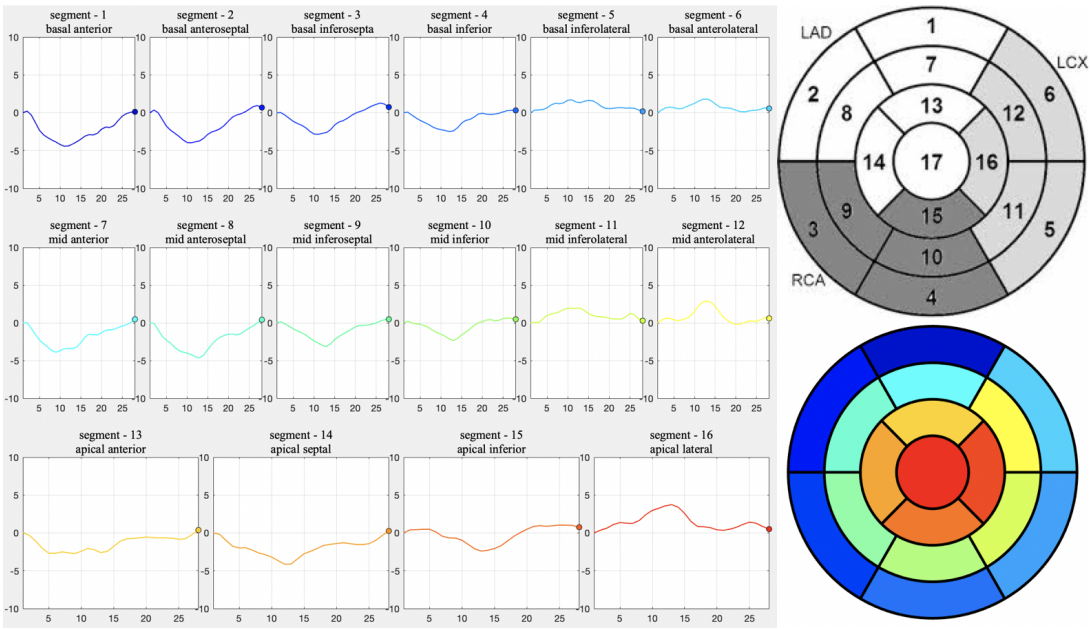

Figure 6.14: Regional motion of one category 3 patient (improved). Left: the radial distance of 16 segments between itself and LV axis at the entire cardiac cycle. Positive value means moving outward, and negative value means moving inward. From the figure, it is clear that not every segment moves inward simultaneously at the ES phase. From segment 14 and 16, we can notice very slow apical rocking. It is the common pattern for patients belonging to this category. Right: 17-segment definition as reference. The colors indicate the correspondence with the results in the left figures.

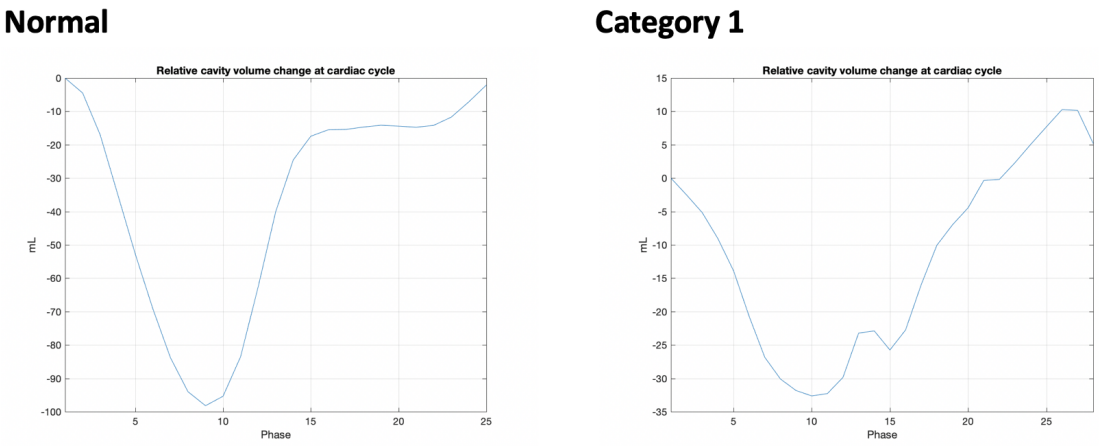

Figure 6.15: LV cavity volumes comparison between normal subjects and one category 1 patient (not improved). Left: the dynamic LV cavity volume of normal subjects. Right: the dynamic LV cavity volume of the patient. We can notice that the quick apical rocking causes the sudden jump of the cavity volume change.

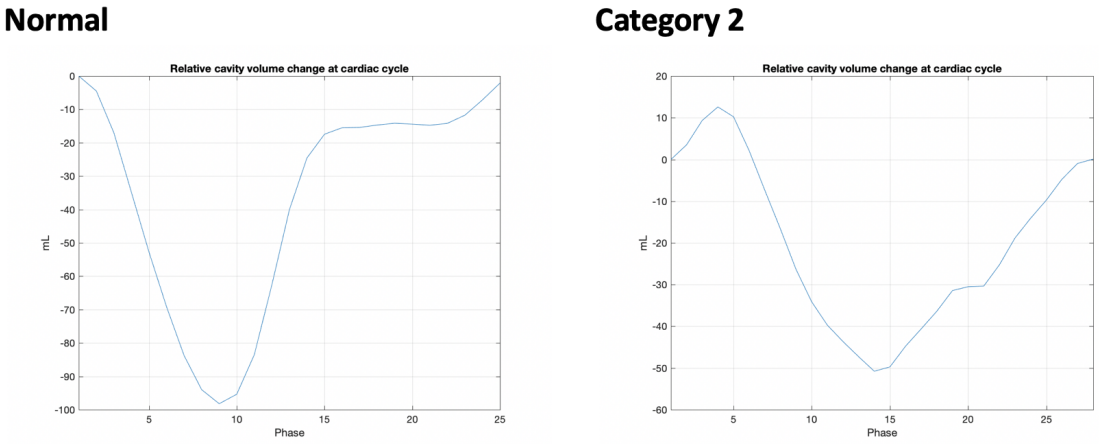

Figure 6.16: LV cavity volumes comparison between normal subjects and one category 2 patient (remain the same). Left: the dynamic LV cavity volume of normal subjects. Right: the dynamic LV cavity volume of the patient. We can notice that the cavity volume goes up first, which is not a healthy pattern and affected by septal flash.

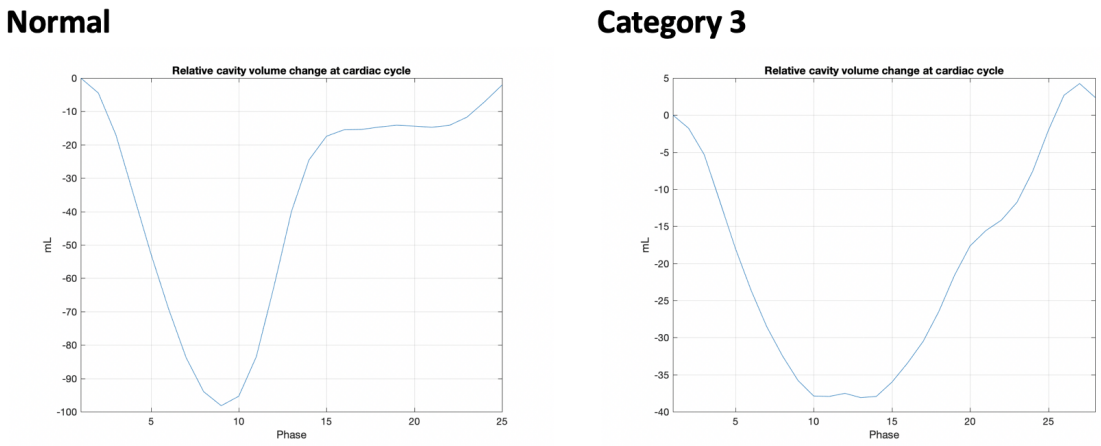

Figure 6.17: LV cavity volumes comparison between normal subjects and one category 3 patient (improved). Left: the dynamic LV cavity volume of normal subjects. Right: the dynamic LV cavity volume of the patient. We can notice that the LV is activated is delayed, which is the clear indicator for heart diseases.

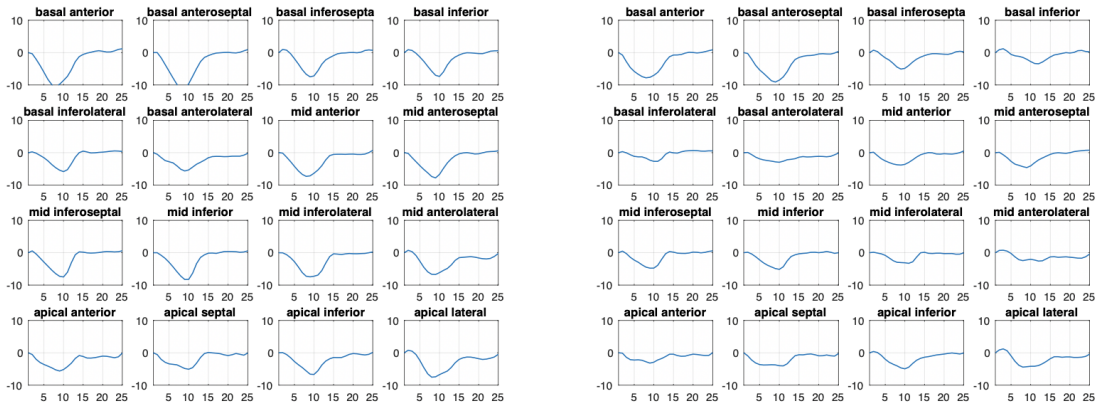

Figure 6.18: LV radial motion of surface centroids of 16 segments for a normal subject. Left: inner wall; right: outer wall. We are able to validate with the plot. Our results fit the fact that inner wall has a larger motion compared with outer wall.

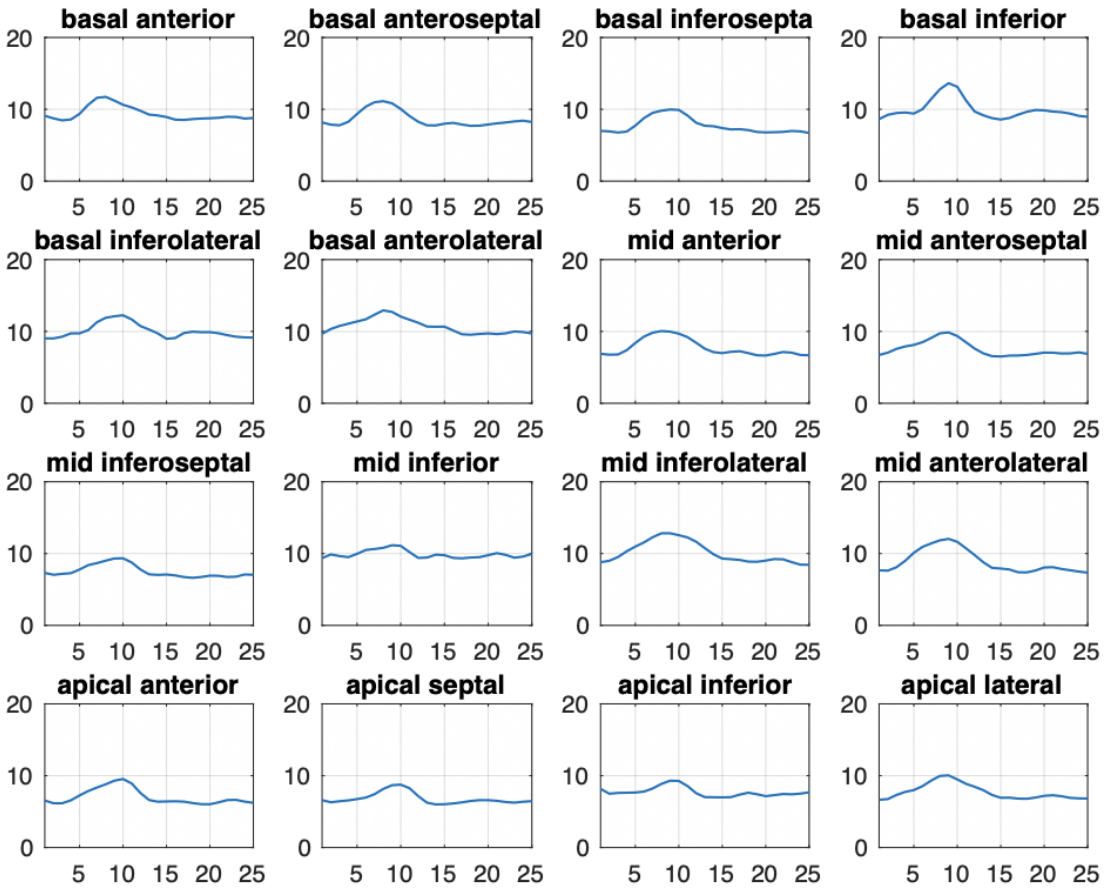

Figure 6.19: Dynamic thickness of 16 segments for a normal subject. We are able to validate with the plot. Our results fit the fact that wall becomes thicker at ESV, and turns thinner after ESV.

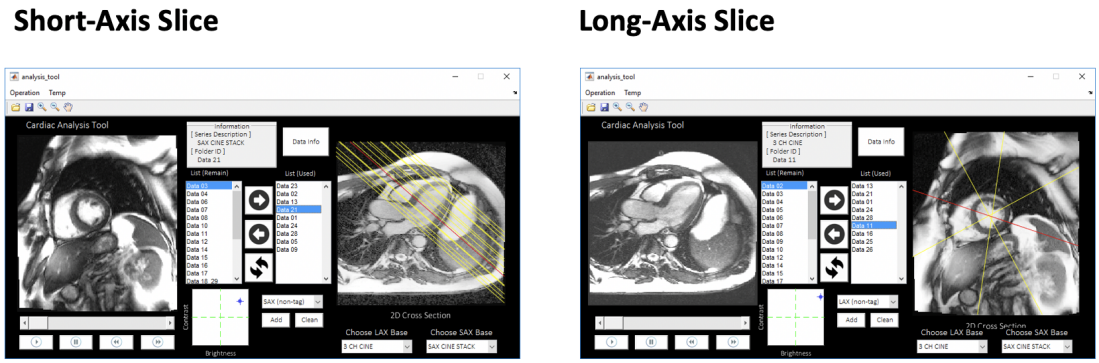

Figure 6.20: Graphical user interface of our proposed framework for MRI visualization.

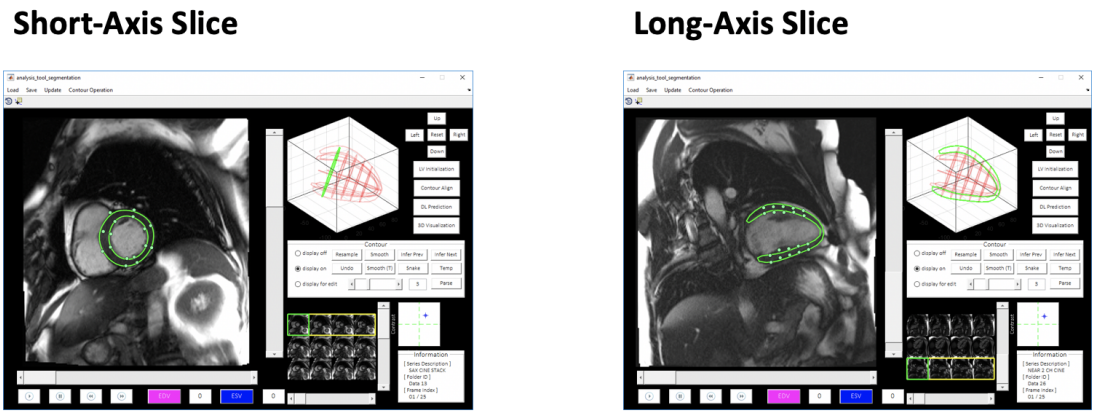

Figure 6.21: Graphical user interface of our proposed framework for MRI segmentation, 3D motion reconstruction, and 17-segment model analysis.

6.5 Conclusions

We have proposed an automated framework to conduct both 3D LV wall motion analysis and 3D blood/muscle segmentation from 2D cardiac MRI. The regional motion and global function of LV are properly studied for MRI acquisitions of both normal subjects and patients, given the reconstructed 3D models from our framework. Our experimental results demonstrate the effectiveness of the proposed approach. From the applications in the thesis, we are able to observe that the CRT outcome is implicitly correlated the LV wall motion. Moreover, our proposed approaches can be applied for the analysis of other heart diseases. We also implement a prototype tool with graphical user interface for concept proof of our entire framework, shown in 6.20 and 6.21.

The proposed framework can be further extended in many medical imaging applications, such as vertebra localization, liver segmentation, and knee joint analysis for which 3D model reconstruction is critical for understanding human anatomy. We will apply our framework to several other important applications in the following chapters.

Chapter 7

Other Applications I: Vertebra Localization

7.1 Background

Automatic and accurate landmark positioning and identification, e.g. for human spine detection and labeling, have been developed as key tools in 2D or 3D medical imaging, such as computed tomography (CT), magnetic resonance imaging (MRI), and X-ray, etc. General clinical tasks such as pathological diagnosis, surgical planning [42] and post-operative assessment can benefit from such locate-and-name tool. Specific applications in human vertebrae detection and labeling include vertebrae segmentation [43, 44], fracture detection [45], tumor detection, registration [46, 47] and statistical shape analysis [48, 49], etc. However, designing such an automatic and accurate vertebrae detection and labeling framework faces multiple challenges such as pathological conditions, image artifacts and limited field-of-view (FOV), as shown in Figure 7.1. Pathological conditions can arise from spinal curvature, fractures, deformity and degeneration, of which spinal shapes are significantly different compared to normal anatomy. Image artifacts such as surgical metal implants change the image intensity distribution and greatly alter the appearance of vertebrae. Furthermore, limited FOVs given by spine-focused scans also add difficulty to the localization and identification of each vertebra due to the repetitive nature of these vertebrae and the lack of global spatial and contextual information. In order to address these challenges, an accurate and efficient spine localization algorithm is required for the potential clinical usage.

To meet the requirements of both accuracy and efficiency, many approaches have been presented in the recent decade. Generally, they can be divided into two categories: conventional machine learning based approaches and deep neural network based approaches. Schmidt *et al.* [50] proposed an efficient method for part-based localization of spine detection which incorporates contextual shape information into a probabilistic graphic model. Features for detecting

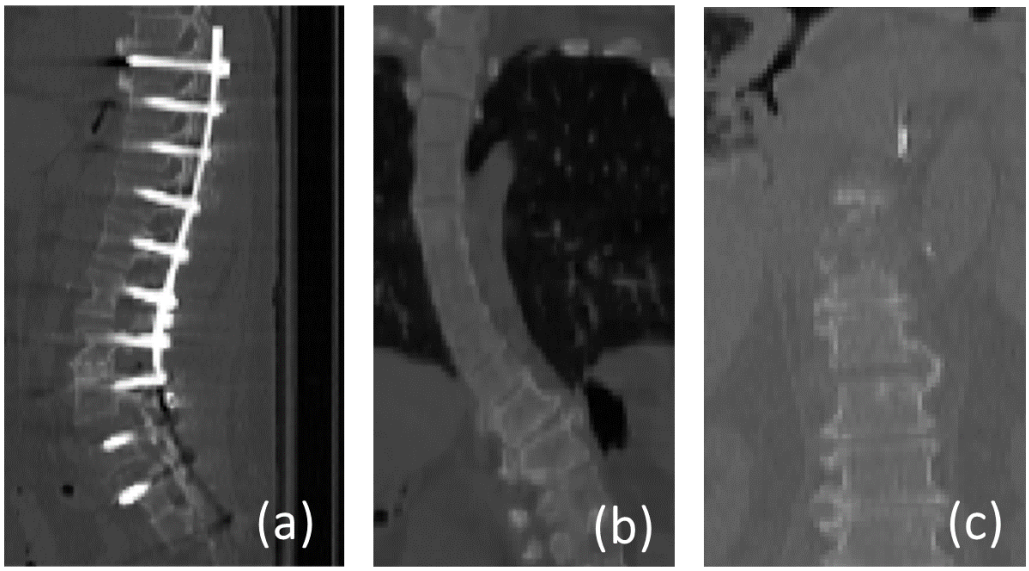

Figure 7.1: Demonstration of uncommon conditions in CT scans. (a) Surgical metal implants (b) Spine curvature (c) Limited FOV

parts are learned from the training database and detected by a multi-class classifier followed by a graphical model. Their method is evaluated on an MRI database and demonstrates robust detection even when some of vertebrae are missing in the image. Glocker *et al.* [51] presents an algorithm based on the regression forests and probabilistic graphical models. This two-stage approach is quantitatively evaluated on 200 CT scans, which achieves an identification rate of 81%. Furthermore, Glocker *et al.* [52] extends this vertebrae localization approach to address the challenge in pathological spine CT scans. Their approach is built on the supervised classification forests and evaluated on a challenging database of 224 pathological spine CT scans. It obtains an overall mean localization error of less than 9 mm with an identification rate of 70%, which outperforms state-of-the-art on pathological cases at that moment. Recently, deep neural networks (DNN) have been achieving great progress in solving low-level computer vision tasks such as image classification, scene segmentation and object detection. DNN has been highlighted in the research of landmark detection in medical imaging and demonstrated its outstanding performance compared to the conventional approaches. Chen *et al.* [53] proposed a joint learning model with convolutional neural networks (J-CNN) to effectively localize and identify the vertebrae. This approach, which is composed of a random forest classifier, a

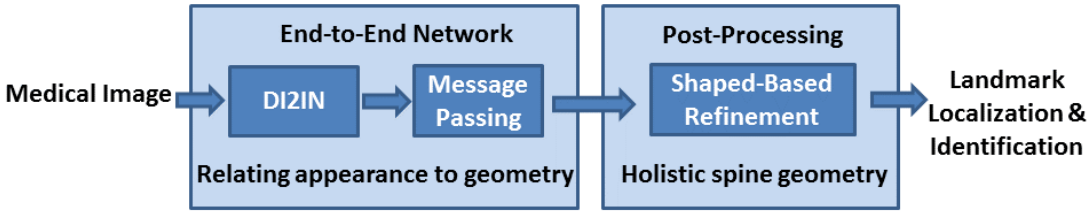

Figure 7.2: Proposed method which consists of three major components: deep Image-to-Image Network (DI2IN), message passing and shape-based refinement.

J-CNN and a shape regression model, improved the identification rate (85%) with a large margin with smaller localization errors in the same challenging database [52]. Suzani *et al.* [54] presented a fast automatic vertebrae detection and localization approach using deep learning. Their approach first extracts intensity-based features from the voxels in the CT scans; then applied a deep neural network on these features to regress the distance between the center of vertebrae and the reference voxels. It achieves a higher detection rate with faster inference but suffers from a larger mean error compared to other approaches [52, 53]. While most approaches are conducted on CT scans, Sun *et al.* [55] proposed the method of structured support vector regression for spinal angle estimation and landmark detection in 2D X-ray images. Their method has strong dependence on the hand-crafted features. The original work is published in an international conference [56].

In order to take the advantage of deep neural networks and overcome the limitations in vertebrae detection, we propose an effective and automatic approach, as shown in Figure 7.2, with the following contributions.

a) Deep Image-to-Image Network for Voxel-Wise Regression

Compared to the approaches that require hand-crafted features from input images, the proposed deep image-to-image network (DI2IN) performs directly on the 2D X-ray images or 3D CT volumes and generates the multi-channel probability maps which are associated with different vertebrae. The probability map itself explicitly indicates the location and type of vertebra. Additionally, the proposed DI2IN does not adopt any classifier to coarsely remove outliers in pre-processing. By building the DI2IN in a fully convolutional manner, it is significantly efficient in terms of computation time, which sets it apart from the sliding window approaches.

b) Response Enhancement with Message Passing

Although the proposed DI2IN usually provides high confident probability maps, sometimes it produces few false positives due to the similar appearance of vertebrae. The anatomical structure of spine provides a strong geometric prior for vertebral centroids. In order to fully explore such prior, we introduce a message-passing scheme which can communicate information of the neighborhood in space. At first, the chain-structured graph is constructed based on the prior on vertebra structure. The graph connection directly defines the neighborhood of each vertebra. Second, for the neighboring centroids, we learn the convolutional kernels between the probability maps. At inference, the probability maps from previous step are further convoluted with the learned kernels to help refine the prediction of neighbors' probability maps. The messages are passed via the convolution operations between neighbors. After a few iterations of message passing, the probability maps converge to a stable state. The probability maps of vertebrae are enhanced, and the issues, such as missing response or false positive response, are well compensated.

c) Joint Refinement using Shape-Based Dictionaries

Given the coordinates of vertebrae, which are the outputs of DI2IN and message passing, we present a joint refinement approach using dictionary learning and sparse representation. In details, we first construct a shape-based dictionary in the refinement, which embeds the holistic structure of the spine. Instead of learning a shape regression model [53] or Hidden Markov Model [51] to fit the spinal shape, the shape-based dictionary is simply built from the coordinates of spines in the training samples. The refinement can be formulated as an $\ell 1$-norm optimization problem and solved by the sparse coding approach in a pre-defined subspace. This optimization aims to find the best sparse representation of the coordinates with respect to the dictionary. By taking the regularity of the spine shape into account, ambiguous predictions and false positives are removed. Finally, the coordinates from all directions are jointly refined, which leads to further improvement in performance.

In the previous published version of this section [56], we validated our proposed method in a large-scale CT database. In this journal version, we extend our work with more analysis, results and implementation details. Several typical failure cases are well studied and solved with sufficient explanation. In addition, we validate our method in another large-scale database, 2D

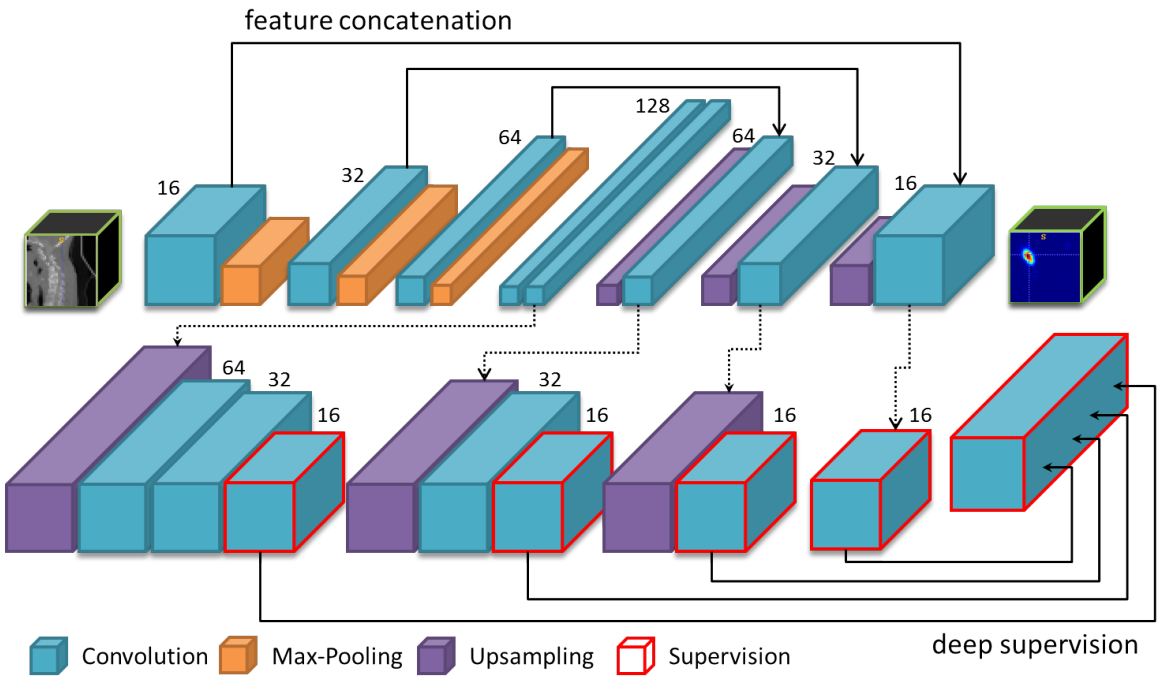

Figure 7.3: Proposed deep image-to-image network (DI2IN) used in 3D CT images experiments. The front part is a convolutional encoder-decoder network with feature concatenation, while the backend is a multi-level deep supervision network. Numbers next to convolutional layers are the channel numbers.

chest X-ray scans, which is also challenging due to similar imaging appearance. The experimental results show that our method has large potentials for any general applications of the anatomical landmark location.

The remainder of this section is organized as follow: In Section II, we present the details of the proposed approach for vertebrae localization and identification, which consists of three subsections. In Section III, we evaluate the proposed approach on both 2D X-ray and 3D CT databases. Our results are compared with other state-of-the-art works. Section IV presents the conclusion.

7.2 Methodology

7.2.1 The Deep Image-to-Image Network (DI2IN) for Spinal Centroid Localization

In this section, we present a deep image-to-image network (DI2IN) model, which is a multi-layer fully convolutional neural network [57, 58] for localization of the vertebral centroids. Figure 7.3 shows the configuration of 3D DI2IN used in the 3D CT images experiments. The 2D DI2IN used in X-Ray experiments has similar structure except all layers are 2D-based. As can be seen, the deployment of DI2IN is symmetric and can be considered as a convolutional encoder-decoder model. DI2IN follows the end-to-end learning fashion, which also guarantees the efficiency at inference. For such purpose, the multi-channel ground truth data is specially designed using the coordinates of vertebral centroids. The 3D Gaussian distribution $I_{\text{gt}} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\|\mathbf{x}-\mu\|^2/2\sigma^2}$ is defined around the positions of the vertebrae in each channel. Vector $\mathbf{x} \in \mathbb{R}^3$ denotes the voxel coordinate inside the volume, and vector μ is the ground truth position of each vertebra. Variance σ^2 is predefined, which controls the size of the Gaussian distribution. The prediction of each channel $I_{\text{prediction}}$ is corresponding to the unique vertebral centroid. It shares the same size with the input image. Thus, the whole learning problem is transformed into a multi-channel voxel-wise regression. In the training process, we use the square loss of $\left\| I_{\text{gt}} - I_{\text{prediction}} \right\|^2$ in the output layer of each voxel. The reason that we define the centroid localization as a regression task instead of classification, is that the highly unbalanced labeling of voxels is unavoidable in the classification approach, which may cause misleading classification accuracy.

The encoder part of the proposed network uses convolution, rectified linear unit (ReLU), and maximum pooling layers. The pooling layer is vital because it helps increase the receptive field of neurons, while reducing the GPU memory consumption at the same time. With larger receptive field, each neuron in different levels considers richer contextual information, therefore the relative spatial positions of the vertebral centroid is better understood. The decoder section consists of convolution, ReLU, and upsampling layers. The upsampling layer is implemented as the bilinear interpolation to amplify and densify the activation. It enables the voxel-wise end-to-end training scheme. In Figure 7.3, the convolution filter size is $1 \times 1 \times 1$ at the final output

layer (1×1 for 2D images), and $3 \times 3 \times 3$ for other convolution layers (3×3 for 2D images). The filter size of the maximum pooling layers is $2 \times 2 \times 2$ (2×2 for 2D images). The stride number in the convolution layer is set to 1, so that each channel remains the same size. The stride number in the pooling layer is set to 2 which down-samples the size of feature maps by 2 in each dimension. The number of channels in each layer is illustrated next to the convolution layers in Figure 7.3. In the up-sampling layers, the input feature maps are up-sampled by 2 in all directions. The network takes a 3D CT image (volume) or 2D X-ray scans as input and directly outputs probability maps associated with vertebral centroids within different channels. Our framework computes the probability maps and the centers of gravity positions, which is more efficient than the methods of classification or regression methods in [53, 54].

Our DI2IN has adopted several popular techniques. We use feature concatenation (skip connection) in the DI2IN, which is similar to the references [59, 36]. The short-cut bridge is built directly from the encoder layers to the decoder layers. It forwards the feature maps of the encoder; then concatenates them to the corresponding layers of the decoder. The outcome of concatenation is used as input of the following convolution layers. Based on the design, the high- and low-level features are clearly combined to gain the benefits of local and global information into the network. In [60], the deep supervision in neural network depth monitoring enables excellent boundary detection and segmentation results. In this work, we introduce a more complex deep supervision method to improve the performance. Multiple branches are separated from the middle layers of the decoder in the master network. They up-sample each input feature map to the same size of the image, followed by several convolution layers to match the channel number of ground truth data. The supervision happens at the end of each branch i and shares the same ground truth data in order to calculate the loss item l_i. The final output is determined by another convolution of the concatenation of all branches' outputs and the decoder output. The total loss l_{total} is a sum of the loss from all branches and that from the final output, as shown in the following equation:

$$l_{\text{total}} = \sum_i l_i + l_{\text{final}} \tag{7.1}$$

7.2.2 Probability Map Enhancement with Message Passing

Given an input image I, the DI2IN usually generates a probability map $P(v_i|I)$ for the centroid coordinate v_i of vertebra i with a high confidence. The location with highest probability shall be marked as the prediction of v_i. However, the probability maps from DI2IN are not always perfect, which may result in errors in the vertebra location prediction. In the worst-case scenario, there are no clear responses in the corresponding probability maps for few vertebrae because the imaging appearance of those vertebrae are very similar. In order to handle the issue of the missing response and reduce the false positive response, we propose a message-passing scheme to enhance the probability maps from the DI2IN utilizing the spatial relationship of vertebrae.

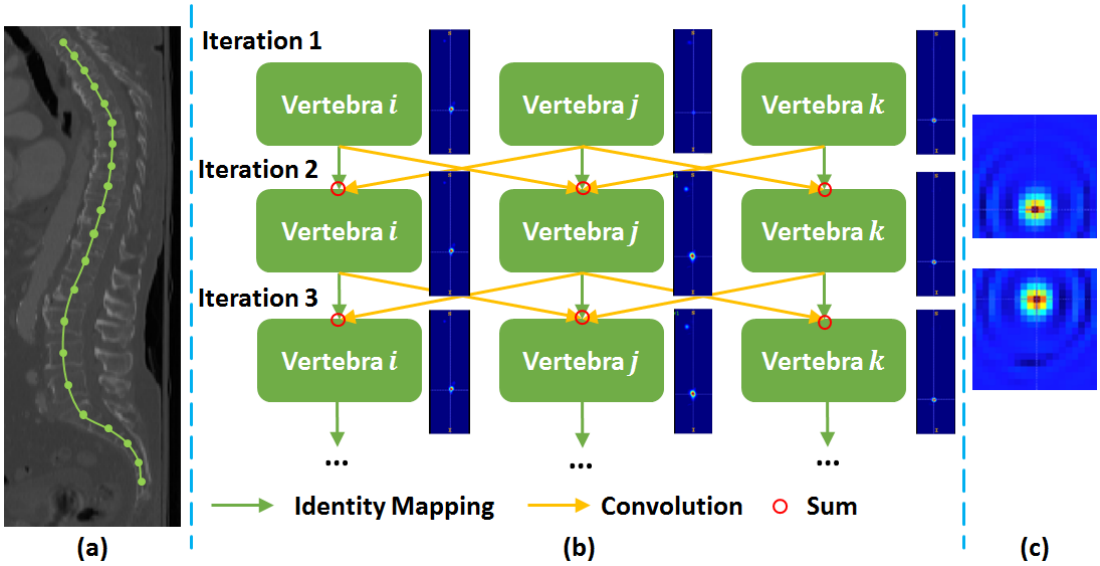

Figure 7.4: (a) The chain-structure model for vertebra centroids shown in CT image; (b) Several iterations of message passing (landmarks represents vertebra centers): the neighbors' centroid probability maps help compensating the missing response of centroids. (c) Sample appearance of the learned kernels.

The concept of message-passing algorithm, also known as belief propagation, has been brought up on the graphical models for decades [61]. It is used to compute marginal distribution of each unobserved nodes (sum-product algorithm) or infer the mode of joint distribution (max-product algorithm). The algorithm has been prevailing in the field of computer vision for many applications [62, 63, 64]. The key idea is to pass mutual information between neighboring nodes for multiple iterations until convergence and enable the model to reach the global

optimization. Similarly, we introduce a chain-structured graph based on the geometry of spine. Each node i represents a vertebral centroid, and has at most two neighboring nodes (vertebrae). We propose the following formulation to update the probability map $P\left(v_i|I\right)$ at the t-th iteration of message passing.

$$P_{t+1}\left(v_i|I\right) = \frac{\alpha \cdot \frac{\sum_{j\in\partial i} m_{j\to i}}{|\partial i|} + P_t\left(v_i|I\right)}{Z} \tag{7.2}$$

$$= \frac{\alpha \cdot \frac{\sum_{j\in\partial i} P_t(v_j|I) * k(v_i|v_j)}{|\partial i|} + P_t\left(v_i|I\right)}{Z} \tag{7.3}$$

where ∂i denotes the neighbors of node i in the graph, which is also corresponding to the adjacent vertebrae. α is a constant to adjust the summation weights between the passed messages and the previous probability map. Z is another constant for normalization. The message $m_{j\to i}$ is passed from node j to its neighboring node i, defined as $P_t\left(v_j|I\right) * k\left(v_i|v_j\right)$. $*$ denotes the convolution operation and the kernel $k\left(v_i|v_j\right)$ is learned from the ground truth Gaussian distributions of i and j. The convolution using the kernels actually shifts the probability map $P\left(v_i|I\right)$ towards $P\left(v_j|I\right)$. If DI2IN provides a confident response at the correct location of vertebra i, its message would be strong as well around the ground truth location of vertebra j after convoluting with the learned kernel. The messages from all neighbors are aggregated to enhance the response. After several iterations of message-passing, the probability maps will converge to a stable state and the issue of the missing response would be compensated. The locations of vertebrae are determined at the peak positions of the enhanced probability maps at the moment. The underlying assumption of message-passing is that DI2IN has given the correct and confident prediction for most vertebrae, which has already been proved in the experiments. Another advantage of the scheme is that it enables the end-to-end training (or fine-tuning) together with DI2IN for better optimization when the iteration number is fixed.

Several recent works have applied the similar message-passing schemes in different applications of the landmark detection. Chu *et al.* [65] introduced a similar message-passing method for human pose estimation (or body joint detection). However, the effectiveness of their implicit passing method may not be clear because it is conducted between feature maps of different landmarks. Our message-passing is directly applied between the probability maps of vertebrae. It is more intuitive to understand how the kernel works and justify the quality of messages. Yang *et al.*[66] also proposed an analogous message-passing method for human pose estimation. They

used the hand-crafted features, which usually have limitation on generalization, to describe the spatial relationship of landmarks. Our method uses the learnable kernels to describe the geometric relationship of vertebrae. The convolution kernels enables the pair-wise communication between vertebrae. Payer *et al.*[67] brought up a one-time message passing method for the anatomical landmark detection. Their passing scheme used dot-product for message aggregation and mainly for outlier removal. But in our framework, the missing response is the major issue instead of noisy probability maps, then the dot-product is not applicable for our passing scheme.

7.2.3 Joint Refinement using Shape-Based Dictionaries

Given the probability maps generated by DI2IN and message-passing enhancement, it may still generate some outliers or false positives. For example, even though the DI2IN followed by message-passing enhancement outputs quite clear and reasonable probability maps, there is still false positive as shown in Figure 7.5. This might arise from the low resolution scans, image artifacts or lack of global contextual information. In order to overcome these limitations, localization refinement has been introduced in many works[51, 53]. In [51], a hidden Markov model (HMM) with hidden states is defined for vertebrae location, appearance likelihoods and inter-vertebra shape priors, which could yield a refined localization based on several thousands of candidate locations from the forest prediction. In [53], a quadratic polynomial curve is proposed to refine the coordinate in the vertical axis. By optimizing an energy function, the parameters for the shape regression model are learned to refine the coordinates of vertebrae. However, this model assumes the shape of the spine could be represented by a quadratic form. In addition, only coordinates in the vertical axis (head to foot direction) are refined.

Inspired by dictionary learning and sparse representation [68, 69], we design a joint refinement using a shape-based dictionary. For illustration purpose, we are using 3D representation in this section which is used in 3D CT experiments. Given a pre-defined shape-based dictionary, the coordinates are refined jointly in all x, y and z axes. The refinement itself can be formulated as an ℓ_1-norm optimization and solved by the sparse coding approach. In details, given the shape-based dictionary $\mathbf{D} \in \mathbb{R}^{M \times N}$ and the coordinate prediction $\mathbf{v} \in \mathbb{R}^N$, we propose a joint refinement algorithm as shown in Algorithm 3 to solve the sparse coefficient vector

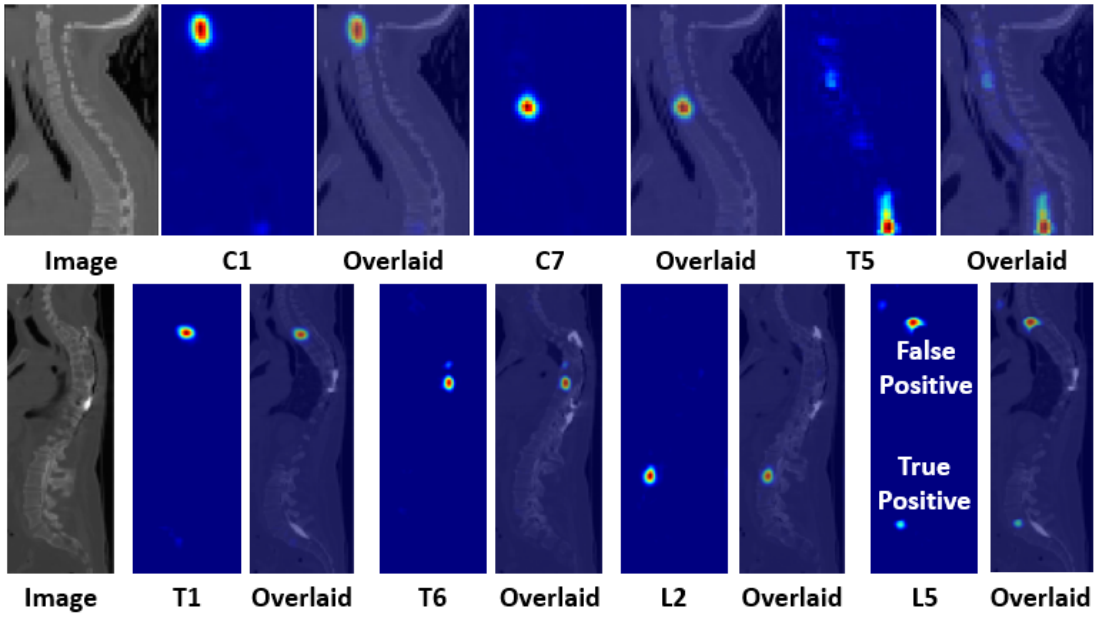

Figure 7.5: Demonstration of two prediction examples in CT images. Only one representative slice is shown for demonstration purpose. Left: CT image. Middle: Output of one channel from the network. Right: Overlaid display. The most predicted responses are close to ground truth location. In the second row, a false positive response exists remotely besides the response at the correct location.

$\mathbf{a} \in \mathbb{R}^M$. Then the refined coordinate vector is defined as $\hat{\mathbf{v}} = \mathbf{D}\mathbf{a}$. Specifically, the shape-based dictionary $\mathbf{D}$ is simply built by the coordinates of vertebrae in training samples. For example, the notation $\mathbf{D}_z$ indicates the shape-based dictionary associated with vertical axis or z direction. $\mathbf{d}_{z,i} \in \mathbb{R}^M$, which is a column of $\mathbf{D}_z$, is defined as $[z_{i,1}\ z_{i,2}\ ...\ z_{i,26}]^T$. For instance, $z_{i,1}$ denotes the vertical ground truth coordinate of ith sample corresponding to vertebrae C_1. The $\mathbf{D}_x$ and $\mathbf{D}_y$ denote the dictionaries associated with x and y directions, respectively. They are both build in the same manner as $\mathbf{D}_z$. Similarly, $\mathbf{v_z}$, defined as $[v_{z,1}\ v_{z,2}\ ...\ v_{z,26}]$, is the vertical coordinate of prediction. $\mathbf{v}_x$ and $\mathbf{v}_y$ are defined in the same manner.

In order to address the challenges such as outliers and limited FOV in spinal scans, we define the original space ϕ_0 and a subspace ϕ_1 in proposed refinement approach. The original space denotes a set which contains all indexes of 26 vertebrae. In our case, ϕ_0 contains the indexes from 1 to 26 which are corresponding to vertebra C_1 to S_2. Compared to the original space ϕ_0, the subspace ϕ_1 denotes a subset which only contains the partial indexes of ϕ_0. Based on the subspace ϕ_1, we define sub-dictionary $\mathbf{D}_{\phi_1}$ and sub-coordinate vector $\mathbf{v}_{\phi_1}$. Intuitively, $\mathbf{D}_{z,\phi_1}$ indicates the sub-dictionary associated with axis z, which is also simply a sub-matrix of

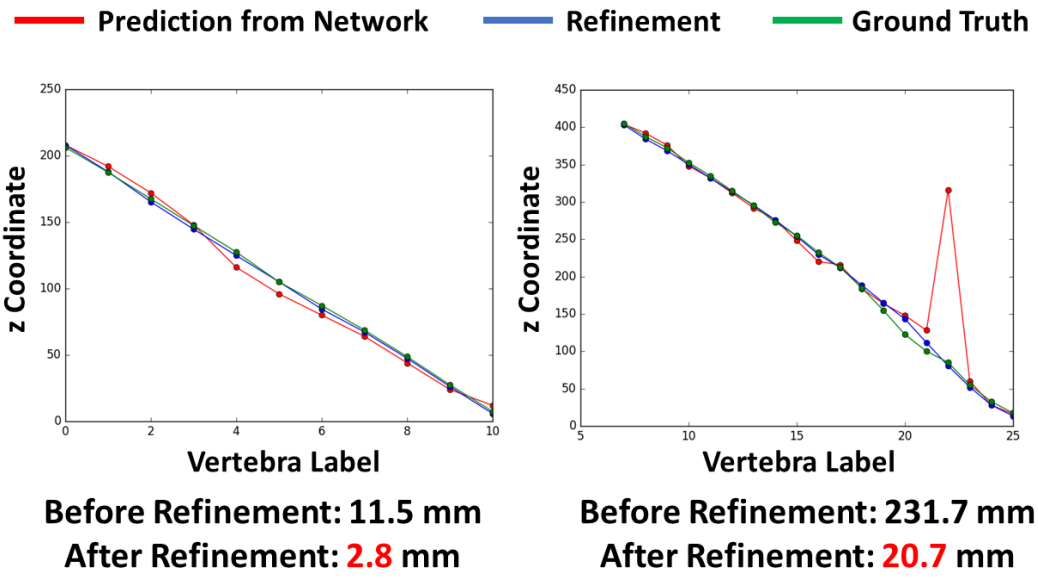

Figure 7.6: Maximum errors of vertebra localization before and after the joint shape-based refinement in 3D CT experiments.

$\mathbf{D}_{z,\phi_0}$. Basically, the optimization problem is solved based on the subspace ϕ_1 instead of the original space ϕ_0.

The details are demonstrated in Algorithm 3. Taking the shape regularity into account, we firstly find the maximum descending subsequence in the coordinate prediction $\mathbf{v}_z$ via dynamic programming. The reason we choose the vertical axis z to determine the maximum subsequence instead of $\mathbf{v}_x$ and $\mathbf{v}_y$ is the vertical axis of the human spine naturally demonstrates the most robust geometric shape compared to x and y axes. Based on the subspace ϕ_1 generated in Step 1, we further remove the indexes of neighboring vertebrae of which distance is too large or too small. Given the subspace ϕ_1, we define the sub-dictionary and sub-coordinate vector for each axis, respectively. Then, the ℓ_1 norm problem in Step 5 is optimized for x, y and z individually based on the same subspace ϕ_1. Finally, all coordinates are refined based on the original space ϕ_0 (i.e. $\mathbf{D}_{z,\phi_0}$ and $\mathbf{v}_{z,\phi_0}$). Intuitively, we remove the ambiguous outliers from the preliminary prediction and then jointly refine the coordinates without these outliers. Based on the subspace, we optimize the refinement problem to find the best sparse combination in the shape-based sub-dictionary. By taking the advantage of the original shape-based dictionary, all coordinates are refined jointly as shown in Figure 7.6.

Algorithm 3 Joint Refinement using Shape-Based Dictionary

Require: The dictionary $\mathbf{D}_{x,\phi_0}$, $\mathbf{D}_{y,\phi_0}$ and $\mathbf{D}_{z,\phi_0} \in \mathbb{R}^{M\times N}$, the predicted coordinates vector $\mathbf{v}_x$, $\mathbf{v}_y$ and $\mathbf{v}_z$, the error threshold ϵ_1 and ϵ_2, and the coefficient λ. M and N indicate the number of landmarks and size of items in dictionary, respectively.

1: Given the predicted coordinates $\mathbf{v}_z$ from the DI2IN and message passing, the maximum descending subsequence is found via dynamic programming.
2: Add the indexes associated with the maximum descending subsequence into the set ϕ_1.
3: Remove the pair of neighboring indexes if $|\mathbf{v}_{z,i} - \mathbf{v}_{z,j}| \leq \epsilon_1$ or $|\mathbf{v}_{z,i} - \mathbf{v}_{z,j}| \geq \epsilon_2$, where $i, j \in \phi_1$ and $|i - j| = 1$.
4: Based on the subspace ϕ_1, define the sub-dictionary $\mathbf{D}_{x,\phi_1}$, $\mathbf{D}_{y,\phi_1}$, and $\mathbf{D}_{z,\phi_1}$ and the sub-coordinate predictions $\mathbf{v}_{x,\phi_1}$, $\mathbf{v}_{y,\phi_1}$ and $\mathbf{v}_{z,\phi_1}$.
5: Solve the optimization problem below by ℓ_1 norm recovery for the vertical axis z:

$$\min_{\mathbf{a}_z} \frac{1}{2}||\mathbf{v}_{z,\phi_1} - \mathbf{D}_{z,\phi_1}\mathbf{a}_z||_2^2 + \lambda||\mathbf{a}_z||_1.$$

6: Solve the same optimization problem in Step 3 for $\mathbf{v}_{x,\phi_1}$ and $\mathbf{v}_{y,\phi_1}$, respectively.
7: Return the refined coordinate vectors $\hat{\mathbf{v}_x} = \mathbf{D}_{x,\phi_0}\mathbf{a}_x$, $\hat{\mathbf{v}_y} = \mathbf{D}_{y,\phi_0}\mathbf{a}_y$ and $\hat{\mathbf{v}_z} = \mathbf{D}_{z,\phi_0}\mathbf{a}_z$.

7.3 Experiments

In this section, we evaluate the performance of the proposed approach on two different and large databases. The first one has been introduced in [52] which contains 302 spine-focused 3D CT scans with various pathologies. These unusual appearances include abnormal curvature, fractures and bright visual artifacts such as surgical implants in post-operative cases. In addition, the FOV of each 3D CT scan varies greatly in terms of vertical cropping. The whole spine is visible only in a few samples. Generally, most of the 3D CT scan contain 5-15 vertebrae. In particular, in order to boost the performance of our approach and validate that DNN favors more training data, we further introduce extra 1000+ 3D CT scans in our experiments. The second database consists of 1000+ 2D X-ray scans described in [70, 71, 72]. The ground truth of each database is marked on the center of each vertebra. The location and label of each ground truth is manually annotated by clinical experts. It should be noted that there is no overlap between the training and testing samples.

For 3D CT scans, there are two different settings that have been adopted in previous works[52, 53, 54]. The first setting uses 112 scans as training samples and another 112 scans as testing samples in[52, 54]. The second setting uses overall 242 scans as training samples

and the other 60 scans as testing samples in[52, 53]. In order to fairly compare to other state-of-the-art works[52, 53, 54], we follow the same training and testing configurations, which are denoted as Set 1 and Set 2 in Table 7.2 and 7.3, respectively. For 2D X-Ray scans, we adopt 1170 images as training samples and 50 images as testing samples.

Table 7.2 and 7.3 summarize the quantitative results in terms of localization mean error, identification rate defined by [51] on Set 1 and Set 2 and other metrics. We compare our approach to other results reported in[52, 54, 53] on the 3D CT scans. In details, "DI2IN", "MP" and "S" denote the deep image-to-image network, message passing and shape-based refinement, respectively. "1000" indicates this model is trained with additional 1000 scans and evaluated on the same testing samples. In order to show the improvement of the performance, we list the results after each step for comparison.

Overall, our approach outperforms the state-of-art approaches[52, 53] by 13% and 6% on the same evaluation settings respectively. For Set 1, the DI2IN itself improves the Id. Rates by a margin of 6% compared to the approach in [52]. Message passing and shape-based refinement further increase the Id. Rates to 77% and 80%, respectively. In addition, we have demonstrated that extra 1000 samples boost the performance to 83%. Similarly, the proposed approach also demonstrates better performance in Set 2 compared to [52, 53, 54]. Our approach has achieved a Id. Rates of 85% and a localization mean error of 8.6 mm, which is better than the state-of-art work[53]. Taking advantage of extra 1000 samples, the Id. Rates has achieved 90%. Furthermore, other metrics such as stand deviation (Std), median (Med) and maximum (Max) also intuitively demonstrate the efficiency of our approach. For example, the maximum errors in both sets are significantly reduced to 42.3 mm and 37.9 mm. Figure 7.6 intuitively illustrates the refinement of proposed shape-based refinement in vertical direction. As shown in Figure 7.6, the shape-based refinement takes the shape regularity of spine into account and removes the false positive coordinates. Specifically, the maximum error is significantly reduced.

Additionally, in order to demonstrate the robustness of our approach, we extend our experiments into a 2D X-ray database for training and evaluation. For 2D X-ray scans, the database[70, 71, 72] is randomly divided into two parts: 1170 scans as training samples and 50 scans for testing samples. It is the first time by our knowledge to evaluate such an approach on

2D X-ray scan for human vertebrae localization and identification task. We conducted experiments using the input images with two different resolutions: 0.70 mm and 0.35 mm. They are both re-sampled from the original database. Due to 4 times larger input and output data size, the DI2IN used in 0.35 mm experiment has less number of filters in the convolution layers comparing to the network in 0.70 mm experiment, as well as smaller batch size in training. Table 7.4 and 7.5 demonstrate the performance of each step using our approach in terms of localization error and identification rates on input images with 0.70 mm and 0.35 mm resolution, respectively. Because most of vertebrae in X-ray scans belong to the thoracic region (T_1 - T_{12}), we only present the overall results instead of showing results in individual region. In details, the DI2IN itself achieves a localization error of 8.4 mm and 7.8 mm and an identification rate of 80% and 82% on 0.70 mm and 0.35 mm resolution, respectively. We also introduce message passing scheme and shape-based refinement to evaluate the performance. The quality of performance is further improved compared to the DI2IN itself. The identification rate is also greatly improved after the introduction of message passing and shape-based refinement. Overall, the identification rate has been significantly increased by the message passing and refinement and finally reached 91% on higher resolution settings. Our experiment demonstrates the proposed approach is able to achieve better performance on higher resolution database. Given more memory allocation and model capacity, our approach could further improve the quality of landmark detection.

Although our approach has achieved high identification rates on various pathological cases in both 3D CT scans and 2D X-ray scans, there are still some challenging cases. As shown in Figure 7.7, the proposed approach occasionally fails to refine the coordinates which are jointly offset. This limitation might arise from special pathological cases, limited FOV and low resolution input images. In our approach, the underlying assumption is that majority of the vertebra probability maps are confident and well distributed around the true locations, which is guaranteed by the powerful DI2IN. In order to address this limitation, more sophisticated network will be further studied in the future. From Figure 7.8, we can see that vertebrae in thoracic region are comparatively harder to locate because those vertebrae share similar imaging appearance.

All experiments are conducted on a high-performance cluster equipped with an Intel 3.5

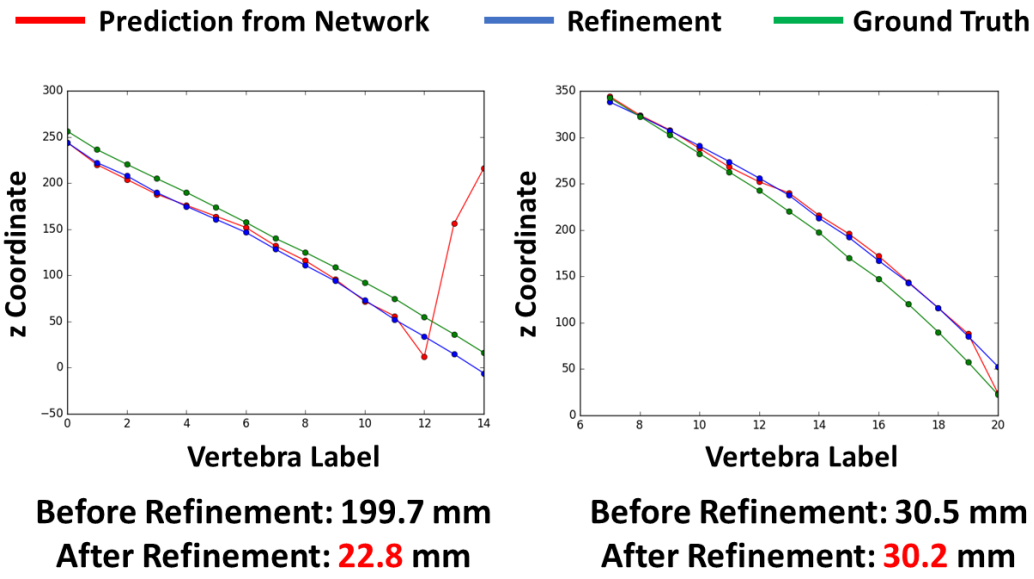

Figure 7.7: Maximum errors of vertebra localization in challenging CT cases before and after the message passing and shape-based network refinement.

GHz CPU as well as a 12 GB Nvidia Titan X GPU. In order to alleviate the pressure of memory, experiments on 3D CT scans and X-rays scans are conducted on a resolution of 4 mm, 0.7 mm and 0.35 mm, respectively. The size of convolutional kernel in message-passing is $23 \times 23 \times 23$ for 3D volume, and 49×49 for 2D images. The evaluation time of our approach is around three seconds per 3D CT case on average using GPU. In order to extract valid information from noisy probability maps, the response maps of DI2IN are compared to a heuristic threshold in an element-wise manner. Only channels with strong response are considered as valid outputs. Then vertebra centroids associated with these channels are identified to be present in the image. The vertebrae associated with other probability maps are identified as non-presented in the image. Therefore, we are able to localize and identify all vertebrae simultaneously in an efficient way.

7.4 Conclusions

We proposed and validated a novel method for vertebral labeling in medical images. The experimental results in both 3D CT volumes and 2D X-ray images show that the proposed method is effective and efficient comparing with the state-of-the-art methods. In addition,

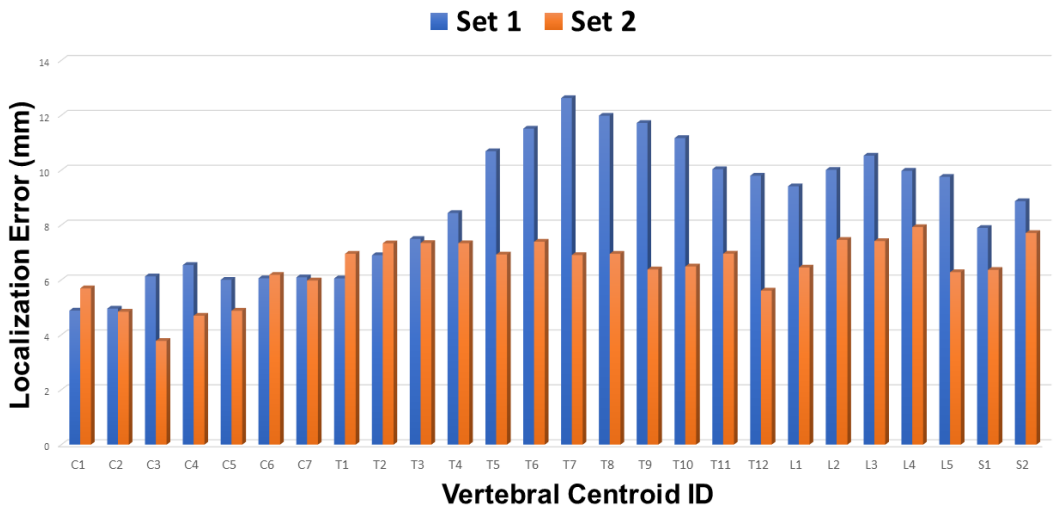

Figure 7.8: Average localization errors (in mm) of the testing database set 1 and set 2 using the proposed methods with extra 1000 training volumes (line "DI2IN+MP+S+1000" in Table 7.2 and 7.3). "C" is for cervical vertebrae, "T" is for thoracic vertebrae, "L" is for thoracic vertebrae, and "S" is for sacral vertebrae.

the extra 1000+ training data in 3D CT experiments evidently boost the performance of the proposed DI2IN, which further acknowledges the importance of large database for deep neural networks.

Table 7.1: Comparison of localization errors in mm and identification rates among different methods for Set 1.

Region	Method	Set 1			Set 1 + 1000			Set 2			Set 2 + 1000		
		Mean	Std	Id.Rates	Mean	Std	Id.Rates	Mean	Std	Id.Rates	Mean	Std	Id.Rates
All	Glocker *et al.*[52]	12.4	11.2	70%	-	-	-	13.2	17.8	74%	-	-	-
	Suzani *et al*[54]	18.2	11.4	-	-	-	-	-	-	-	-	-	-
	Chen *et al.*[53]	-	-	-	-	-	-	8.8	13.0	84%	-	-	-
	DI2IN	17.0	47.3	74%	10.6	21.5	80%	13.6	37.5	76%	7.1	11.8	87%
	DI2IN+MP	11.7	19.7	77%	9.4	16.2	82%	10.2	13.9	78%	6.9	8.3	89%
	DI2IN+MP+Sparsity	9.1	7.2	80%	**8.5**	**7.7**	**83%**	8.6	7.8	85%	**6.4**	**5.9**	**90%**
Cervical	Glocker *et al.*[52]	7.0	4.7	80%	-	-	-	6.8	10.0	89%	-	-	-
	Suzani *et al*[54]	17.1	8.7	-	-	-	-	-	-	-	-	-	-
	Chen *et al.*[53]	-	-	-	-	-	-	**5.1**	8.2	92%	-	-	-
	DI2IN+MP+Sparsity	6.6	3.9	83%	**5.8**	**3.9**	**88%**	5.6	**4.0**	92%	5.2	4.4	**93%**
Thoracic	Glocker *et al.*[52]	13.8	11.8	62%	-	-	-	17.4	22.3	62%	-	-	-
	Suzani *et al*[54]	17.2	11.8	-	-	-	-	-	-	-	-	-	-
	Chen *et al.*[53]	-	-	-	-	-	-	11.4	16.5	76%	-	-	-
	DI2IN+MP+Sparsity	9.9	7.5	74%	**9.5**	**8.5**	**78%**	9.2	7.9	81%	**6.7**	**6.2**	**88%**
Lumbar	Glocker *et al.*[52]	14.3	12.3	75%	-	-	-	13.0	12.5	80%	-	-	-
	Suzani *et al*[54]	20.3	12.2	-	-	-	-	-	-	-	-	-	-
	Chen *et al.*[53]	-	-	-	-	-	-	8.4	8.6	88%	-	-	-
	DI2IN+MP+Sparsity	10.9	9.1	80%	**9.9**	**9.1**	**84%**	11.0	10.8	83%	**7.1**	**7.3**	**90%**

Table 7.2: Comparison of localization errors in mm and identification rates among different methods for Set 1.

Region	Method	Set 1				
		Mean	Std	Id.Rates	Med	Max
All	Glocker *et al.*[52]	12.4	11.2	70%	8.8	-
	Suzani *et al*[54]	18.2	11.4	-	-	-
	Chen *et al.*[53]	-	-	-	-	-
	DI2IN	13.5	32.0	76%	6.7	396.9
	DI2IN+MP	11.7	19.7	77%	6.8	396.9
	DI2IN+MP+S	9.1	7.0	80%	7.1	**42.3**
	DI2IN+1000	10.6	21.5	80%	**5.5**	430.4
	DI2IN+MP+1000	9.4	16.2	82%	6.0	430.4
	DI2IN+MP+S+1000	**8.5**	**7.7**	**83%**	6.2	59.6
Cervical	Glocker *et al.*[52]	7.0	4.7	80%	-	-
	Suzani *et al*[54]	17.1	8.7	-	-	-
	Chen *et al.*[53]	-	-	-	-	-
	DI2IN+MP+S	6.6	3.9	83%	-	-
	DI2IN+MP+S+1000	**5.8**	**3.9**	**88%**	-	-
Thoracic	Glocker *et al.*[52]	13.8	11.8	62%	-	-
	Suzani *et al*[54]	17.2	11.8	-	-	-
	Chen *et al.*[53]	-	-	-	-	-
	DI2IN+MP+S	9.9	7.5	74%	-	-
	DI2IN+MP+S+1000	**9.5**	**8.5**	**78%**	-	-
Lumbar	Glocker *et al.*[52]	14.3	12.3	75%	-	-
	Suzani *et al*[54]	20.3	12.2	-	-	-
	Chen *et al.*[53]	-	-	-	-	-
	DI2IN+MP+S	10.9	9.1	80%	-	-
	DI2IN+MP+S+1000	**9.9**	**9.1**	**84%**	-	-

Table 7.3: Comparison of localization errors in mm and identification rates among different methods for Set 2.

Region	Method	Set 2				
		Mean	Std	Id.Rates	Med	Max
All	Glocker *et al.*[52]	13.2	17.8	74%	-	-
	Suzani *et al*[54]	-	-	-	-	-
	Chen *et al.*[53]	8.8	13.0	84%	-	-
	DI2IN	13.6	37.5	76%	5.9	410.6
	DI2IN+MP	10.2	13.9	78%	5.7	153.1
	DI2IN+MP+S	8.6	7.8	85%	5.2	75.1
	DI2IN+1000	7.1	11.8	87%	4.3	235.9
	DI2IN+MP+1000	6.9	8.3	89%	4.6	108.7
	DI2IN+MP+S+1000	**6.4**	**5.9**	**90%**	**4.5**	**37.9**
Cervical	Glocker *et al.*[52]	6.8	10.0	89%	-	-
	Suzani *et al*[54]	-	-	-	-	-
	Chen *et al.*[53]	**5.1**	8.2	92%	-	-
	DI2IN+MP+S	5.6	**4.0**	92%	-	-
	DI2IN+MP+S+1000	5.2	4.4	**93%**	-	-
Thoracic	Glocker *et al.*[52]	17.4	22.3	62%	-	-
	Suzani *et al*[54]	-	-	-	-	-
	Chen *et al.*[53]	11.4	16.5	76%	-	-
	DI2IN+MP+S	9.2	7.9	81%	-	-
	DI2IN+MP+S+1000	**6.7**	**6.2**	**88%**	-	-
Lumbar	Glocker *et al.*[52]	13.0	12.5	80%	-	-
	Suzani *et al*[54]	-	-	-	-	-
	Chen *et al.*[53]	8.4	8.6	88%	-	-
	DI2IN+MP+S	11.0	10.8	83%	-	-
	DI2IN+MP+S+1000	**7.1**	**7.3**	**90%**	-	-

Table 7.4: Comparison of localization errors in mm and identification rates among different methods for 0.70 mm X-ray Set.

Region	Method	0.7 mm				
		Mean	Std	Id.Rates	Med	Max
All	DI2IN	8.4	14.7	80%	3.7	283.4
	DI2IN+MP	7.7	9.6	82%	3.7	45.9
	DI2IN+MP+S	**7.1**	**9.2**	**88%**	**4.2**	**44.2**

Table 7.5: Comparison of localization errors in mm and identification rates among different methods for 0.35 mm X-ray Set.

Region	Method	0.35 mm				
		Mean	Std	Id.Rates	Med	Max
All	DI2IN	7.8	12.1	82%	3.1	114.0
	DI2IN+MP	7.4	9.8	84%	3.6	57.9
	DI2IN+MP+S	**6.4**	**7.8**	**91%**	**3.0**	**46.2**

Chapter 8

Other Applications II: Liver Segmentation

8.1 Background

Accurate liver segmentation from three dimensional (3D) medical images , e.g. computed tomography (CT) or magnetic resonance imaging (MRI) is essential in many clinical applications, such as pathological diagnosis of hepatic diseases, surgical planning, and postoperative assessment. However, automatic liver segmentation is still a highly challenging task due to the complex background, fuzzy boundary, and various appearance of liver in medical images.

To date, several methods have been proposed for automatic liver segmentation from 3D CT scans. Generally, they can be categorized into non-learning-based and learning-based approaches. Non-learning-based approaches usually rely on the statistical distribution of the intensity, including atlas-based [73], active shape model (ASM)-based [74], levelset-based [75], and graph-cut-based [76] methods, etc. On the other hand, learning-based approaches take the advantage of hand-crafted features to train the classifiers to achieve good segmentation. For example, in [77], the proposed hierarchical framework applies marginal space learning with steerable features to handle the complicated texture pattern near the liver boundary.

Until recently, deep learning has been shown to achieve superior performance in various challenging tasks, such as classification, segmentation, and detection. Several automatic liver segmentation approaches based on convolutional neural network (CNN) have been proposed. Dou, et, al. [78] demonstrated a fully convolutional network (FCN) with deep supervision, which can perform end-to-end learning and inference. The output of FCN is refined with a fully connected conditional random field (CRF) approach. Similarly, Christ, et, al. [79] proposed cascaded FCNs followed by CRF refinement. Lu, et, al. [80] used a FCN with graph-cut based refinement. Although these methods demonstrated good performance, they all used pre-defined refinement approaches. For example, both CRF and graph-cut methods are limited to the use

of pairwise models, and time-consuming as well. They may cause serious leakage at boundary regions with low contrast, which is common in liver segmentation.

Meanwhile, Generative Adversarial Network (GAN) [81] has emerged as a powerful framework in various tasks. It consists of two parts: generator and discriminator. The generator tries to produce the output that is close to the real samples, while the discriminator attempts to distinguish between real and generated samples. Inspired by [82], we propose an automatic liver segmentation approach using an adversarial image-to-image network (DI2IN-AN). A deep image-to-image network (DI2IN) is served as the generator to produce the liver segmentation. It employs a convolutional encoder-decoder architecture combined with multi-level feature concatenation and deep supervision. Our network tries to optimize a conventional multi-class cross-entropy loss together with an adversarial term that aims to distinguish between the output of DI2IN and ground truth. Ideally, the discriminator pushes the generator's output towards the distribution of ground truth, so that it has the potential to enhance generator's performance by refining its output. Since the discriminator is usually a CNN which takes the joint configuration of many input variables, it embeds the higher-order potentials into the network (the geometric difference between prediction and ground truth is represented by the trainable network model instead of heuristic hints). The proposed method also achieves higher computing efficiency since the discriminator does not need to be executed at inference.

All previous liver segmentation approaches were trained using dozens of volumes which did not take the full advantage of CNN. In contrast, our network leverages the knowledge of an annotated dataset of 1000+ CT volumes with various different scanning protocols (e.g., contrast and non-contrast, various resolution and position) and large variations in populations (e.g., ages and pathology). To the best of our knowledge, our experiment is the first time that more than 1000 annotated 3D CT volumes are adopted in liver segmentation tasks. The experimental result shows that training with such a large dataset significantly improves the performance and enhances the robustness of the network. The work originally is published in an international conference [83].

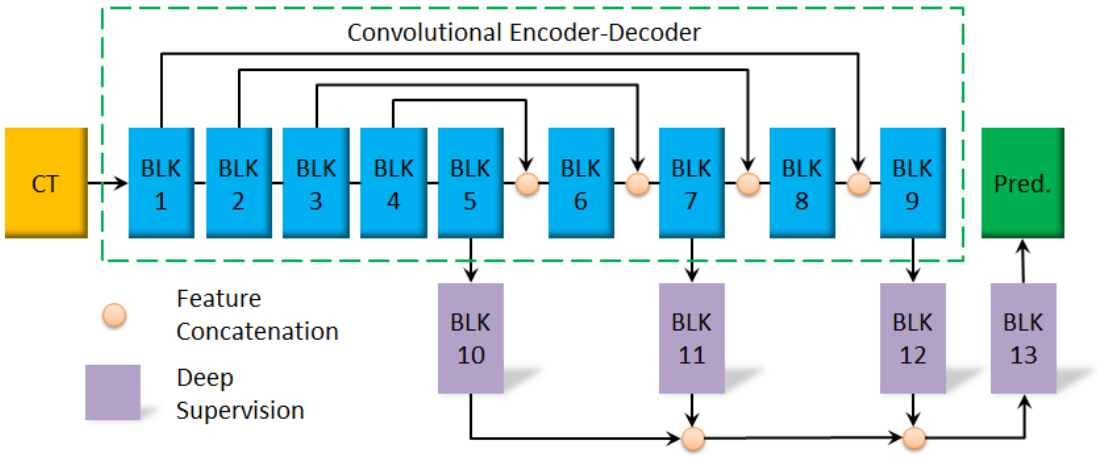

Figure 8.1: Proposed deep image-to-image network (DI2IN). The front part is a convolutional encoder-decoder network with feature concatenation, and the backend is deep supervision network through multi-level. Blocks inside DI2IN consist of convolutional and upscaling layers.

8.2 Methodology

8.2.1 Deep Image-to-Image Network (DI2IN) for Liver Segmentation

In this section, we present a deep image-to-image network (DI2IN), which is a multi-layer convolutional neural network (CNN), for the liver segmentation. The segmentation task is defined as the voxel-wise binary classification. DI2IN takes the entire 3D CT volumes as input, and outputs the probability maps that indicate how likely voxels belongs to the liver region. As shown in Fig. 8.1, the main structure of DI2IN is designed following a symmetric way as a convolutional encoder-decoder. All blocks in DI2IN consist of 3D convolutional and bilinear upscaling layers. The details of the network is described in Fig. 8.3.

In the encoder part of DI2IN, only the convolution layers are used in all blocks. In order to increase the receptive field of neurons and lower the GPU memory consumption, we set stride as 2 at some layers and reduce the size of feature maps. Moreover, larger receptive field covers more contextual information and helps to preserve liver shape information in the prediction. The decoder of DI2IN consists of convolutional and bilinear upscaling layers. To enble end-to-end prediction and training, the upscaling layers are implemented as bilinear interpolation to enlarge the activation maps. All convolutional kernels are $3 \times 3 \times 3$. The upscaling factor in decoder is 2 for x, y, z dimension. The Leaky rectified linear unit (Leaky ReLU) and batch normalization are adopted in all convolutional layers for proper gradient back-propagation.

In order to further improve the performance of DI2IN, we adopt several mainstream technologies with the necessary changes [59, 84, 78]. First, we use the feature layer concatenation in DI2IN. Fast bridges are built directly from the encoder layers to the decoder layers. The bridges pass the information from the encoder forward and then concatenate it with the decoder feature layers. The combined feature is used as the input for the next convolution layer. Following the steps above to explicitly combine advanced and low-level features, DI2IN benefits from local and global contextual information. The deep supervision of the neural network during end-to-end training is shown to achieve good boundary detection and segmentation results. In the network, we introduced a more complex deep supervision scheme to improve performance. Several branches are separated from layers of the decoder section of main DI2IN. With the appropriate upscaling and convolution operations, the output size of each channel for all branches matches the size of the input image (Upscaling factors are 16,4,1 in block 10,11,12 repectively). By calculating the loss item l_i with the same ground truth data, the supervision is enforced at the end of each branch i. In order to further utilize the results of different branches, the final output is determined by the convolution operations of all branches with the leaky ReLU. During training, we apply binary cross entropy loss to each voxel of the output layers. The total loss l_{total} is the weighted combination of loss terms for all output layers, including the final output layer and the output layers for all branches, as follows:

$$l_{total} = \sum_i w_i \cdot l_i + w_{final} \cdot l_{final}$$

8.2.2 Network Improvement with Adversarial Training

We adopt the prevailing idea of the generative adversarial networks to boost the performance of DI2IN. The proposed scheme is shown in Fig.8.2. An adversarial network is adopted to capture the high-order appearance information, which distinguishes between the ground truth and the output from DI2IN. In order to guide the generator to better prediction, the adversarial network provides an extra loss function for updating the parameters of generator during training. The purpose of the extra loss is to make the prediction as close as possible to the ground truth labeling. We adopt the binary cross-entropy loss for training of the adversarial network. D and G represent the discriminator and generator (DI2IN, in the context), respectively. For

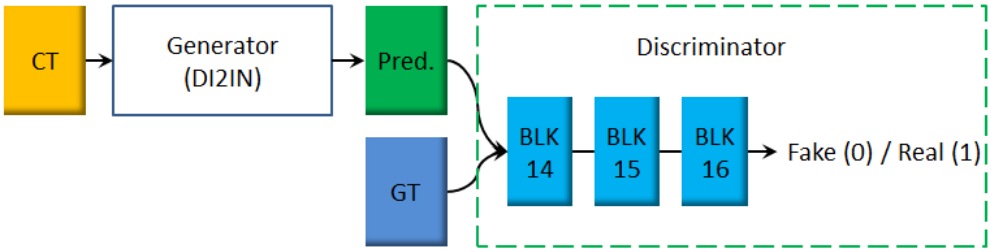

Figure 8.2: Proposed adversarial training scheme. The generator produces the segmentation prediction, and discriminator classifies the prediction and ground truth during training.

the discriminator $D\left(Y;\theta^D\right)$, the ground truth label Y_{gt} is assigned as one, and the prediction $Y_{pred} = G\left(X;\theta^G\right)$ is assigned as zero where X is the input CT volumes. The structure of discriminator network D is shown in Fig. 8.3. The following objective function is used in training the adversarial network:

$$\begin{aligned} l_D &= -\mathbb{E}_{y\sim p_{gt}} \log\left(D\left(y;\theta^D\right)\right) - \mathbb{E}_{y'\sim p_{pred}} \log\left(1 - D\left(y';\theta^D\right)\right) \\ &= -\mathbb{E}_{y\sim p_{gt}} \log\left(D\left(y;\theta^D\right)\right) - \mathbb{E}_{x\sim p_{data}} \log\left(1 - D\left(G\left(x;\theta^G\right);\theta^D\right)\right) \end{aligned} \tag{8.1}$$

During the training of network D, the gradient of loss l_D is propagated back to update the parameters of the generator network (DI2IN). At this stage, the loss for G has two components shown in the Equation 8.2. The first component is the conventional segmentation loss l_b: voxel-wise binary cross entropy between the prediction and ground truth labels. Minimizing the second loss component enables the discriminator D to confuse the ground truth with the prediction from G.

$$\begin{aligned} l_G &= \mathbb{E}_{y\sim p_{pred}, y'\sim p_{gt}}\left[l_{seg}\left(y,y'\right)\right] - \lambda\mathbb{E}_{y\sim p_{pred}} \log\left(1 - D\left(y;\theta^D\right)\right) \\ &= \mathbb{E}_{y\sim p_{pred}, y'\sim p_{gt}}\left[l_{seg}\left(y,y'\right)\right] - \lambda\mathbb{E}_{x\sim p_{data}} \log\left(1 - D\left(G\left(x;\theta^G\right);\theta^D\right)\right) \end{aligned} \tag{8.2}$$

Following suggestions in [81], we replace $-\log\left(1 - D\left(G\left(x\right)\right)\right)$ with $\log\left(D\left(G\left(X\right)\right)\right)$. In another word, we would like to maximize the probability that prediction to be the ground truth in Equation 8.2, instead of minimizing the probability that prediction not to be the generated label map. Such replacement provides strong gradient during training of G and speed up the training process in practice.

$$l_G = \mathbb{E}_{y\sim p_{pred}, y'\sim p_{gt}}\left[l_{seg}\left(y,y'\right)\right] + \lambda\mathbb{E}_{x\sim p_{data}} \log D\left(G\left(x;\theta^G\right);\theta^D\right) \tag{8.3}$$

The generator and discriminator are trained alternatively for several times shown in Algorithm 4, until the discriminator is not able to easily distinguish between ground truth label and the

Algorithm 4 Adversarial training of generator and discriminator.

Input : pre-trained generator (DI2IN) with weights θ_0^G
Output: updated generator weights θ_1^G
for *number of training iterations* **do**
  for k_D *steps* **do**
    sample a mini-batch of training images $x \sim p_{data}$ generate prediction y_{pred} for x with $G\left(x;\theta_0^G\right)$ $\theta^D \leftarrow$ propagate back the stochastic gradient $\triangledown l_D\left(y_{gt}, y_{pred}\right)$
  for k_G *steps* **do**
    sample a mini-batch of training images $x' \sim p_{data}$ generate y'_{pred} for x' with $G\left(x';\theta_0^G\right)$ and compute $D\left(G\left(x'\right)\right)$ $\theta_1^G \leftarrow$ propagate back the stochastic gradient $\triangledown l_G\left(y'_{gt}, y'_{pred}\right)$
  $\theta_0^G \leftarrow \theta_1^G$

output of DI2IN. After the training process, the adversarial network is no longer required at inference. The generator itself can provide high quality segmentation results and its performance is improved.

8.3 Experiments

Most public dataset for liver segmentation only consists of tens of cases. For example, the MICCAI-SLiver07 [85] dataset only contains 20 CT volumes for training and 10 CT volumes for testing. All the data are contrast enhanced. Such a small dataset is not suitable to show the power of CNN: it has been well known that neural network trained with more labelled data can usually achieve much better performance. Thus, in this chapter, we collected more than 1000 CT volumes. The liver of each volume was delineated by human experts. These data covers large variations in populations, contrast phases, scanning ranges, pathologies, and field of view (FOV), etc. The inter-slice distance varies from 0.5mm to 7.0mm. All scans covers the abdominal regions but may extend to head and feet. Tumor can be found in multiple cases. The volumes may also have various other disease. For example, pleural effusion, which brights the lung region and changes the pattern of upper boundary of liver. Then we collected additional 50 volumes from clinical sites for the independent testing. The livers of these data were also annotated by human experts for the purpose of evaluation. We down-sampled the dataset into 3.0mm resolution isotropically to speed up the processing and lower the consumption of computing memory without loss of accuracy. In the adversarial training, we set λ to 0.01, and the

Block	Layer	*s*	*f*	Block	Layer	*s*	*f*	Block	Layer	*s*	*f*	Block	Layer	*s*	*f*
1	Conv.	1	16	6	Up.	2	-	10	Conv.	1	8	13	Conv.	1	1
	Conv.	2	16		Conv.	1	128		Conv.	1	8	14	Conv.	1	16
2	Conv.	1	32	7	Up.	2	-		Conv.	1	8		Conv.	2	16
	Conv.	2	32		Conv.	1	64		Conv.	1	1	15	Conv.	1	32
3	Conv.	1	64	8	Up.	2	-	11	Up.	4	-		Conv.	1	32
	Conv.	2	64		Conv.	1	32		Conv.	1	8		Conv.	2	32
4	Conv.	1	128	9	Up.	2	-		Conv.	1	8	16	Conv.	1	64
	Conv.	2	128		Conv.	1	16		Conv.	1	1		Conv.	1	64
5	Conv.	1	256		Up.	16	-	12	Up.	1	-		Conv.	1	64
	Conv.	1	256		Conv.	1	8		Conv.	1	1		Conv.	1	1

Figure 8.3: Parametric setting of blocks in neural network. s stands for the stride, f is filter number. *Conv.* is convolution, and *Up.* is bilinear upscaling.

number of overall training iterations is 100. For training D, k_D is 10 and the mini-batch size is 8. For training G, k_G is 1 and the mini-batch size is 4. In the segmentation loss, w_i is set as 1.

Table 1 compares the performance of five different methods. The first method, the hierarchical, learning-based algorithm proposed in [77], was trained using 400 CT volumes. More training data did not show performance improvement for this method. For comparison purpose, the DI2IN network, which is similar to deep learning based algorithms proposed in [78, 79, 80] without post-processing steps, and the DI2IN-AN were trained using the same 400 cases. Both the DI2IN network and the DI2IN-AN were also trained using all 1000+ CT volumes. The average symmetric surface distance (ASD) and dice coefficients are computed for all methods on the test data. As shown in Table 1, DI2IN-AN achieves the best performance in both evaluation metrics. All deep learning based algorithms outperform the classic learning based algorithm with the hand-craft features, which shows the power of CNN. The results show that more training data enhances the performance of both DI2IN and DI2IN-AN. Take DI2IN for example, training with 1000+ labelled data improves the mean ASD by 0.23mm and the max ASD by 3.84mm compared to training with 400 labelled data. Table 1 also shows that the adversarial structure can further boost the performance of DI2IN. The maximum ASD error is also reduced. Typical test samples are provided in Fig. 8.4. We also tried CRF and graph cut to refine the output of DI2IN. However, the results became worse, since a large portion of testing data had no contrast and the boundary of liver bottom at many locations was very fuzzy. CRF and graph cut both suffer from serious leakage in these situations. Using an NVIDIA TITAN X GPU and the Theano/Lasagne library, the run time of our algorithm is less than one second,

Table 8.1: Comparison of five methods on 50 unseen CT data.

Method	ASD (mm)				Dice			
	Mean	Std	Max	Median	Mean	Std	Min	Median
Ling *et al.* (400) [77]	2.89	5.10	37.63	2.01	0.92	0.11	0.20	0.95
DI2IN (400)	2.25	1.28	10.06	2.0	0.94	0.03	0.79	0.94
DI2IN-AN (400)	2.00	0.95	7.82	1.80	0.94	0.02	0.85	0.95
DI2IN (1000)	2.15	0.81	6.51	1.95	0.94	0.02	0.87	0.95
DI2IN-AN (1000)	**1.90**	0.74	**6.32**	**1.74**	**0.95**	0.02	**0.88**	0.95

which is significantly faster than most of the current approaches. For example, it requires 1.5 minutes for one case in [78]. More experimental results can be found in the supplementary material.

8.4 Conclusions

In this chapter, we proposed an automatic liver segmentation algorithm based on an adversarial image-to-image network. Our method achieves good segmentation quality as well as faster processing speed. The network is trained on an annotated dataset of 1000+ 3D CT volumes. We demonstrate that training with such a large dataset can improve the performance of CNN by a large margin.

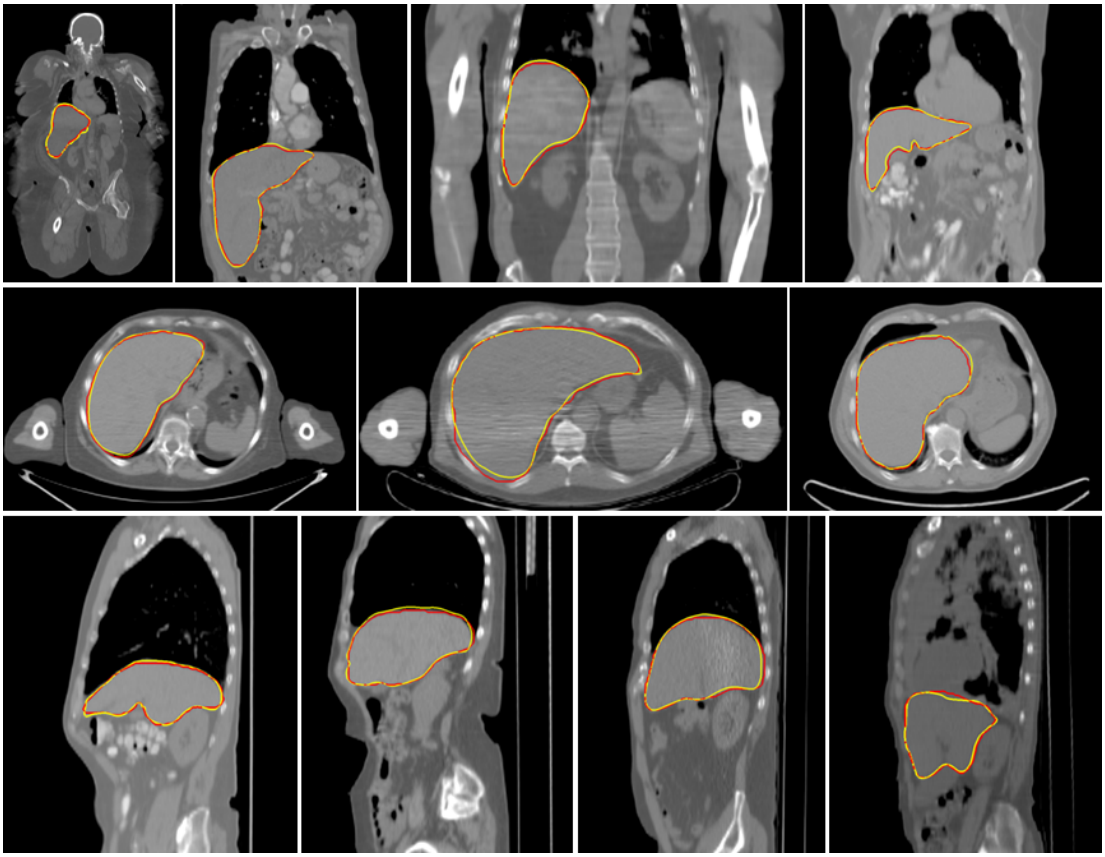

Figure 8.4: Visual Results from different views. Yellow meshes are ground truth. Red ones are the prediction from DI2IN-AN.

Chapter 9

Conclusions and Future Work

In the thesis, we proposed an efficient approach for cardiac MRI segmentation with deep neural networks and multi-component deformable models. And we presented a 2D/3D blood/muscle segmentation to estimate voxel probability distribution, which is an alternative to the conventional approach of defining the "cavity". Furthermore, we proposed a novel way for 3D displacement field computation using unsupervised learning.

9.1 Cardiac MRI Segmentation in 2D Cine MRI

We have proposed a robust and efficient approach for short-axis cardiac MRI segmentation, using both deep neural networks and multi-component deformable models. We first utilize a stack of the segmentation from a 2D U-Net as input of another 3D U-Net. To further improve the segmentation quality, a multi-component deformable model is proposed to integrate the temporal correlation of the cardiac cycle. The evaluation results demonstrated that our approach outperforms other approaches. The similar strategy can be applied on the segmentation of long-axis MRI or the segmentation of other heart chambers.

9.2 3D Left Ventricle Wall Model Reconstruction

In the thesis, we proposed a novel approach to reconstruct 3D shape and motion of LV wall from 2D cardiac cine MRI. The approach is effective and efficient. The further direction is to extend the proposed approach to the tagged MRI, which alignment is still challenging due to the fuzzy imaging quality. It would be also interesting to see the similar analysis conducting on other modalities of cardiac imaging.

9.3 2D/3D Blood/Muscle Segmentation

We have also proposed a probability-based segmentation approach to estimate the fractional blood/muscle classification of boundary pixels near the trabeculae, between the solid wall and the clear blood, in full 3D space, from 2D cardiac MRI acquisitions, through using generative adversarial networks. Our results are quantitatively validated on a synthetic dataset, and also visually validated on a real MRI dataset. This is the first attempt to reconstruct such a 3D probabilistic segmentation from 2D cardiac cine MRI, to the best of our knowledge. The proposed approach has a good potentials for providing improved cardiac motion understanding and clinical applications.

9.4 Assessment of Ventricular Dyssynchrony

Our proposed approach has been applied to study regional motion in 3D space for ventricular dyssynchrony in 2D cardiac MRI. The 3D motion of LV models are created using deep neural networks and deformable models. Then, the 17-segment model is applied for regional motion analysis. Based on the comparison between motion models of normal subjects and patients, we can estimate which type of motion can be treated effectively with CRT. The experimental results demonstrated that our approach is capable to provide a better understanding of dyssynchrony for CRT outcome based on the regional and global measurements from our reconstructed motion models. In addition, our approach has great potentials to be applied for studies of other cardiovascular diseases using cardiac MRI. For the future work, the proposed approach can be applied on the large-scale cardiac MRI dataset for any potential diseases of cardiac functioning mechanisms. As far as our knowledge, it is the first attempt to analyze cardiovascular dyssynchrony using 3D motion models from cardiac MRI.

9.5 Other Medical Imaging Applications

We validate our proposed framework in other applications of medical image analysis. The objectives of the presented applications are highly related with 3D anatomy information. Therefore, together with deep neural networks and deformable model, we can achieve efficient and reliable solutions for those applications.

References

[1] A. C. Lardo, T. P. Abraham, and D. A. Kass, "Magnetic resonance imaging assessment of ventricular dyssynchrony: current and emerging concepts," *Journal of the American College of Cardiology*, vol. 46, no. 12, pp. 2223–2228, 2005.

[2] K. C. Bilchick, V. Dimaano, K. C. Wu, R. H. Helm, R. G. Weiss, J. A. Lima, R. D. Berger, G. F. Tomaselli, D. A. Bluemke, H. R. Halperin, and Others, "Cardiac magnetic resonance assessment of dyssynchrony and myocardial scar predicts function class improvement following cardiac resynchronization therapy," *JACC: Cardiovascular Imaging*, vol. 1, no. 5, pp. 561–568, 2008.

[3] W. Bai, M. Sinclair, G. Tarroni, O. Oktay, M. Rajchl, G. Vaillant, A. M. Lee, N. Aung, E. Lukaschuk, M. M. Sanghvi, and Others, "Automated cardiovascular magnetic resonance image analysis with fully convolutional networks," *Journal of Cardiovascular Magnetic Resonance*, vol. 20, no. 1, p. 65, 2018.

[4] C. Petitjean and J.-N. Dacher, "A review of segmentation methods in short axis cardiac MR images," *Medical image analysis*, vol. 15, no. 2, pp. 169–184, 2011.

[5] D. Yang, Q. Huang, L. Axel, and D. Metaxas, "Multi-component deformable models coupled with 2D-3D U-Net for automated probabilistic segmentation of cardiac walls and blood," in *Proceedings - International Symposium on Biomedical Imaging*, 2018.

[6] N. C. Codella, H. Y. Lee, D. S. Fieno, D. W. Chen, S. Hurtado-Rua, M. Kochar, J. P. Finn, R. Judd, P. Goyal, J. Schenendorf, M. D. Cham, R. B. Devereux, M. Prince, Y. Wang, and J. W. Weinsaft, "Improved left ventricular mass quantification with partial voxel interpolation in vivo and necropsy validation of a novel cardiac MRI segmentation algorithm," *Circulation: Cardiovascular Imaging*, 2012.

[7] N. Paragios, "A variational approach for the segmentation of the left ventricle in cardiac image analysis," *International Journal of Computer Vision*, vol. 50, no. 3, pp. 345–362, 2002.

[8] M. P. Jolly, "Automatic segmentation of the left ventricle in cardiac MR and CT images," *International Journal of Computer Vision*, 2006.

[9] Y. Zhu, X. Papademetris, A. J. Sinusas, and J. S. Duncan, "Segmentation of the left ventricle from cardiac MR images using a subject-specific dynamical model," *IEEE Transactions on Medical Imaging*, vol. 29, no. 3, pp. 669–687, 2010.

[10] S. Queirós, D. Barbosa, J. Engvall, T. Ebbers, E. Nagel, S. I. Sarvari, P. Claus, J. C. Fonseca, J. L. Vilaça, and J. D'Hooge, "Multi-centre validation of an automatic algorithm for fast 4D myocardial segmentation in cine CMR datasets," *European Heart Journal Cardiovascular Imaging*, 2016.

[11] A. Suinesiaputra, B. R. Cowan, A. O. Al-Agamy, M. A. Elattar, N. Ayache, A. S. Fahmy, A. M. Khalifa, P. Medrano-Gracia, M.-P. Jolly, A. H. Kadish, and Others, "A collaborative resource to build consensus for automated left ventricular segmentation of cardiac MR images," *Medical image analysis*, vol. 18, no. 1, pp. 50–62, 2014.

[12] J. Lötjönen, M. Pollari, S. Kivistö, and K. Lauerma, "Correction of Motion Artifacts From Cardiac Cine Magnetic Resonance Images1," *Academic Radiology*, vol. 12, no. 10, pp. 1273–1284, oct 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1076633205005684

[13] R. R. Garlapati, A. Mostayed, G. R. Joldes, A. Wittek, B. Doyle, and K. Miller, "Towards measuring neuroimage misalignment," *Computers in Biology and Medicine*, 2015.

[14] J. Park, D. Metaxas, and L. Axel, "Analysis of left ventricular wall motion based on volumetric deformable models and MRI-SPAMM," *Medical Image Analysis*, vol. 1, no. 1, pp. 53–71, mar 1996. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S1361841501800050

[15] D. Yang, P. Wu, C. Tan, K. M. Pohl, L. Axel, and D. Metaxas, "3D Motion Modeling and Reconstruction of Left Ventricle Wall in Cardiac MRI," in *International Conference on Functional Imaging and Modeling of the Heart*, 2017, pp. 481–492.

[16] Y. B. Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville and Y. Y. Wang, "Generative Adversarial Nets," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2016.

[17] A. Radford, L. Metz, and S. Chintala, "Unsupervised Representation learning with Deep Convolutional GANs," *International Conference on Learning Representations*, 2016.

[18] D. Yang, D. Xu, S. K. Zhou, B. Georgescu, M. Chen, S. Grbic, D. Metaxas, and D. Comaniciu, "Automatic liver segmentation using an adversarial image-to-image network," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017.

[19] D. Yang, B. Liu, L. Axel, and D. Metaxas, "3d lv probabilistic segmentation in cardiac mri using generative adversarial network," in *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer, 2018, pp. 181–190.

[20] M. D. Cerqueira, "American Heart Association Writing Group on Myocardial Segmentation and Registration for Cardiac Imaging: Standardized myocardial segmentation and nomenclature for tomographic imaging of the heart: a statement for healthcare professionals from the Cardiac Imaging Committee of the Council on Clinical Cardiology of the American Heart Association," *Circulation*, vol. 105, pp. 539–542, 2002.

[21] M. R. Avendi, A. Kheradvar, and H. Jafarkhani, "A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI," *Medical image analysis*, vol. 30, pp. 108–119, 2016.

[22] P. V. Tran, "A fully convolutional neural network for cardiac segmentation in short-axis MRI," *arXiv preprint arXiv:1604.00494*, 2016.

[23] M. Sinclair, W. Bai, E. Puyol-Antón, O. Oktay, D. Rueckert, and A. P. King, "Fully Automated Segmentation-Based Respiratory Motion Correction of Multiplanar Cardiac Magnetic Resonance Images for Large-Scale Datasets," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2017, pp. 332–340.

[24] X. Huang, D. P. Dione, C. B. Compas, X. Papademetris, B. A. Lin, A. Bregasi, A. J. Sinusas, L. H. Staib, and J. S. Duncan, "Contour tracking in echocardiographic sequences via sparse representation and dictionary learning," *Medical image analysis*, vol. 18, no. 2, pp. 253–271, 2014.

[25] N. C. F. Codella, J. W. Weinsaft, M. D. Cham, M. Janik, M. R. Prince, and Y. Wang, "Left ventricle: automated segmentation by using myocardial effusion threshold reduction and intravoxel computation at MR imaging," *Radiology*, vol. 248, no. 3, pp. 1004–1012, 2008.

[26] A. Gacek and W. Pedrycz, *ECG Signal Processing, Classification and Interpretation: A Comprehensive Framework of Computational Intelligence*. Springer Publishing Company, Incorporated, 2014.

[27] Y. Zheng, D. Yang, M. John, and D. Comaniciu, "Multi-part modeling and segmentation of left atrium in C-arm CT for image-guided ablation of atrial fibrillation," *IEEE transactions on medical imaging*, vol. 33, no. 2, pp. 318–331, 2014.

[28] K. C. Bilchick, S. Kuruvilla, Y. S. Hamirani, R. Ramachandran, S. A. Clarke, K. M. Parker, G. J. Stukenborg, P. Mason, J. D. Ferguson, J. R. Moorman, and Others, "Impact of mechanical activation, scar, and electrical timing on cardiac resynchronization therapy response and clinical outcomes," *Journal of the American College of Cardiology*, vol. 63, no. 16, pp. 1657–1666, 2014.

[29] Z. Ren, J. Yan, B. Ni, B. Liu, X. Yang, and H. Zha, "Unsupervised Deep Learning for Optical Flow Estimation." in *AAAI*, vol. 3, 2017, p. 7.

[30] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "An Unsupervised Learning Model for Deformable Medical Image Registration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9252–9260.

[31] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.

[32] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-net: Learning dense volumetric segmentation from sparse annotation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016.

[33] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional Sequence to Sequence Learning," *arXiv preprint arXiv:1705.03122*, 2017.

[34] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P. A. Heng, "H-DenseUNet: Hybrid Densely Connected UNet for Liver and Liver Tumor Segmentation from CT Volumes," *arXiv preprint arXiv:1709.07330*, 2017.

[35] P. Radau, Y. Lu, K. Connelly, G. Paul, A. J. Dick, and G. A. Wright, "Evaluation Framework for Algorithms Segmenting Short Axis Cardiac MRI," *The MIDAS Journal - Cardiac MR Left Ventricle Segmentation Challenge*, 2009.

[36] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *3D Vision (3DV), 2016 Fourth International Conference on*. IEEE, 2016, pp. 565–571.

[37] K. McLeish, D. L. Hill, D. Atkinson, J. M. Blackall, and R. Razavi, "A study of the motion and deformation of the heart due to respiration," *IEEE Transactions on Medical Imaging*, 2002.

[38] A. Myronenko and X. Song, "Point set registration: Coherent point drifts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

[39] D. N. Metaxas, *Physics-based deformable models: applications to computer vision, graphics and medical imaging*. Springer Science & Business Media, 2012, vol. 389.

[40] Z. X. and S. J., "Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI," *Medical Image Analysis*, 2016.

[41] B. D. Lucas, T. Kanade, and Others, "An iterative image registration technique with an application to stereo vision," in *Proceedings of the 7th international joint conference on Artificial intelligence*. Vancouver, British Columbia, 1981, pp. 674–679.

[42] O. Schwarzenbach, U. Berlemann, B. Jost, H. Visarius, E. Arm, F. Langlotz, L.-P. Nolte, and C. Ozdoba, "Accuracy of computer-assisted pedicle screw placement: An in vivo computed tomography analysis," *Spine*, vol. 22, no. 4, pp. 452–458, 1997.

[43] J. Yao, J. E. Burns, D. Forsberg, A. Seitel, A. Rasoulian, P. Abolmaesumi, K. Hammernik, M. Urschler, B. Ibragimov, R. Korez *et al.*, "A multi-center milestone study of clinical vertebral ct segmentation," *Computerized Medical Imaging and Graphics*, vol. 49, pp. 16–28, 2016.

[44] T. Klinder, J. Ostermann, M. Ehm, A. Franz, R. Kneser, and C. Lorenz, "Automated model-based vertebra detection, identification, and segmentation in ct images," *Medical image analysis*, vol. 13, no. 3, pp. 471–482, 2009.

[45] H. K. Genant, C. Y. Wu, C. van Kuijk, and M. C. Nevitt, "Vertebral fracture assessment using a semiquantitative technique," *Journal of bone and mineral research*, vol. 8, no. 9, pp. 1137–1148, 1993.

[46] D. Tomazevic, B. Likar, T. Slivnik, and F. Pernus, "3-d/2-d registration of ct and mr to x-ray images," *IEEE transactions on medical imaging*, vol. 22, no. 11, pp. 1407–1416, 2003.

[47] S. Benameur, M. Mignotte, S. Parent, H. Labelle, W. Skalli, and J. de Guise, "3d/2d registration and segmentation of scoliotic vertebrae using statistical models," *Computerized Medical Imaging and Graphics*, vol. 27, no. 5, pp. 321–337, 2003.

[48] J. Boisvert, F. Cheriet, X. Pennec, H. Labelle, and N. Ayache, "Geometric variability of the scoliotic spine using statistics on articulated shape models," *IEEE Transactions on Medical Imaging*, vol. 27, no. 4, pp. 557–568, 2008.

[49] M. Roberts, T. Cootes, and J. Adams, "Vertebral shape: automatic measurement with dynamically sequenced active appearance models," *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2005*, pp. 733–740, 2005.

[50] S. Schmidt, J. Kappes, M. Bergtholdt, V. Pekar, S. Dries, D. Bystrov, and C. Schnörr, "Spine detection and labeling using a parts-based graphical model," in *Information Processing in Medical Imaging*. Springer, 2007, pp. 122–133.

[51] B. Glocker, J. Feulner, A. Criminisi, D. Haynor, and E. Konukoglu, "Automatic localization and identification of vertebrae in arbitrary field-of-view ct scans," *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012*, pp. 590–598, 2012.

[52] B. Glocker, D. Zikic, E. Konukoglu, D. R. Haynor, and A. Criminisi, "Vertebrae localization in pathological spine ct via dense classification from sparse annotations," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2013, pp. 262–270.

[53] H. Chen, C. Shen, J. Qin, D. Ni, L. Shi, J. C. Cheng, and P.-A. Heng, "Automatic localization and identification of vertebrae in spine ct via a joint learning model with deep neural networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer International Publishing, 2015, pp. 515–522.

[54] A. Suzani, A. Seitel, Y. Liu, S. Fels, R. N. Rohling, and P. Abolmaesumi, "Fast automatic vertebrae detection and localization in pathological ct scans-a deep learning approach," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 678–686.

[55] H. Sun, X. Zhen, C. Bailey, P. Rasoulinejad, Y. Yin, and S. Li, "Direct estimation of spinal cobb angles by structured multi-output regression," in *International Conference on Information Processing in Medical Imaging*. Springer, 2017, pp. 529–540.

[56] D. Yang, T. Xiong, D. Xu, Q. Huang, D. Liu, S. K. Zhou, Z. Xu, J. Park, M. Chen, T. D. Tran *et al.*, "Automatic vertebra labeling in large-scale 3d ct using deep image-to-image network with message passing and sparsity regularization," in *International Conference on Information Processing in Medical Imaging*. Springer, 2017, pp. 633–644.

[57] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[58] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *arXiv preprint arXiv:1511.00561*, 2015.

[59] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[60] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1395–1403.

[61] M. J. Wainwright, M. I. Jordan *et al.*, "Graphical models, exponential families, and variational inference," *Foundations and Trends® in Machine Learning*, vol. 1, no. 1–2, pp. 1–305, 2008.

[62] S. Nowozin, C. H. Lampert *et al.*, "Structured learning and prediction in computer vision," *Foundations and Trends® in Computer Graphics and Vision*, vol. 6, no. 3–4, pp. 185–365, 2011.

[63] N. Komodakis, N. Paragios, and G. Tziritas, "Mrf optimization via dual decomposition: Message-passing revisited," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.

[64] S. Ross, D. Munoz, M. Hebert, and J. A. Bagnell, "Learning message-passing inference machines for structured prediction," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 2737–2744.

[65] X. Chu, W. Ouyang, H. Li, and X. Wang, "Structured feature learning for pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4715–4723.

[66] W. Yang, W. Ouyang, H. Li, and X. Wang, "End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3073–3082.

[67] C. Payer, D. Štern, H. Bischof, and M. Urschler, "Regressing heatmaps for multiple landmark localization using cnns," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 230–238.

[68] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1031–1044, 2010.

[69] R. Rubinstein, A. M. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1045–1057, 2010.

[70] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wáng, P.-X. Lu, and G. Thoma, "Two public chest x-ray datasets for computer-aided screening of pulmonary diseases," *Quantitative imaging in medicine and surgery*, vol. 4, no. 6, pp. 475–477, 2014.

[71] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald, "Preparing a collection of radiology examinations for distribution and retrieval," *Journal of the American Medical Informatics Association*, vol. 23, no. 2, pp. 304–310, 2015.

[72] J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K.-i. Komatsu, M. Matsui, H. Fujita, Y. Kodera, and K. Doi, "Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules," *American Journal of Roentgenology*, vol. 174, no. 1, pp. 71–74, 2000.

[73] M. G. Linguraru, J. K. Sandberg, Z. Li, J. A. Pura, and R. M. Summers, "Atlas-based automated segmentation of spleen and liver using adaptive enhancement estimation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2009, pp. 1001–1008.

[74] D. Kainmüller, T. Lange, and H. Lamecker, "Shape constrained automatic segmentation of the liver based on a heuristic intensity model," in *Proc. MICCAI Workshop 3D Segmentation in the Clinic: A Grand Challenge*, 2007, pp. 109–116.

[75] J. Lee, N. Kim, H. Lee, J. B. Seo, H. J. Won, Y. M. Shin, Y. G. Shin, and S.-H. Kim, "Efficient liver segmentation using a level-set method with optimal detection of the initial liver boundary from level-set speed images," *Computer methods and programs in biomedicine*, vol. 88, no. 1, pp. 26–38, 2007.

[76] L. Massoptier and S. Casciaro, "Fully automatic liver segmentation through graph-cut technique," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2007, pp. 5243–5246.

[77] H. Ling, S. K. Zhou, Y. Zheng, B. Georgescu, M. Suehling, and D. Comaniciu, "Hierarchical, learning-based automatic liver segmentation," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[78] Q. Dou, H. Chen, Y. Jin, L. Yu, J. Qin, and P.-A. Heng, "3d deeply supervised network for automatic liver segmentation from ct volumes," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 149–157.

[79] P. F. Christ, M. E. A. Elshaer, F. Ettlinger, S. Tatavarty, M. Bickel, P. Bilic, M. Rempfler, M. Armbruster, F. Hofmann, M. DAnastasi *et al.*, "Automatic liver and lesion segmentation in ct using cascaded fully convolutional neural networks and 3d conditional random fields," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 415–423.

[80] F. Lu, F. Wu, P. Hu, Z. Peng, and D. Kong, "Automatic 3d liver location and segmentation via convolutional neural network and graph cut," *International journal of computer assisted radiology and surgery*, vol. 12, no. 2, pp. 171–182, 2017.

[81] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[82] P. Luc, C. Couprie, S. Chintala, and J. Verbeek, "Semantic segmentation using adversarial networks," *arXiv preprint arXiv:1611.08408*, 2016.

[83] D. Yang, D. Xu, S. K. Zhou, B. Georgescu, M. Chen, S. Grbic, D. Metaxas, and D. Comaniciu, "Automatic liver segmentation using an adversarial image-to-image network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 507–515.

[84] J. Merkow, A. Marsden, D. J. Kriegman, and Z. Tu, "Dense volume-to-volume vascular boundary detection," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016 - 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part III*, 2016, pp. 371–379. [Online]. Available: https://doi.org/10.1007/978-3-319-46726-9_43

[85] T. Heimann, B. van Ginneken, M. Styner, Y. Arzhaeva, V. Aurich, C. Bauer, A. Beck, C. Becker, R. Beichel, G. Bekes, F. Bello, G. K. Binnig, H. Bischof, A. Bornik, P. Cashman, Y. Chi, A. Cordova, B. M. Dawant, M. Fidrich, J. D. Furst, D. Furukawa,

L. Grenacher, J. Hornegger, D. Kainmüller, R. Kitney, H. Kobatake, H. Lamecker, T. Lange, J. Lee, B. Lennon, R. Li, S. Li, H. Meinzer, G. Németh, D. S. Raicu, A. Rau, E. M. van Rikxoort, M. Rousson, L. Ruskó, K. A. Saddi, G. Schmidt, D. Seghers, A. Shimizu, P. Slagmolen, E. Sorantin, G. Soza, R. Susomboon, J. M. Waite, A. Wimmer, and I. Wolf, "Comparison and evaluation of methods for liver segmentation from CT datasets," *IEEE Trans. Med. Imaging*, vol. 28, no. 8, pp. 1251–1265, 2009. [Online]. Available: https://doi.org/10.1109/TMI.2009.2013851