UNIVERSITY OF LEEDS

Deep Learning in Cardiac Magnetic Resonance Image Analysis and Cardiovascular Disease Diagnosis

Xiang Chen

University of Leeds

School of Computing

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

November, 2023

Intellectual Property Statement

This candidate confirms that the work submitted is his own, except where work which has formed part of jointly authored publications has been included. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others. The contribution of the candidate and other authors to this work is explicitly indicated in the following.

I am the main author of all the following publications. I led the design of the study and wrote the manuscript, including analysis and discussion of the results. I also wrote the analysis code and carried out the data processing, statistical analysis, and experiments. The contribution of other authors was to help me design the study, provide data annotations, discuss the results, and review the manuscript.

- **Chapter 2:**

  1) Chen X, Diaz-Pinto A, Ravikumar N, Frangi AF. "Deep learning in medical image registration." Progress in Biomedical Engineering, 2021, 3(1): 012003.

- **Chapter 3:**

  1) Chen X, Xia Y, Ravikumar N, Frangi AF. "A Deep Discontinuity-Preserving Image Registration Network." International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2021: 46-55.

  2) Chen X, Xia Y, Ravikumar N, Frangi AF. "Joint Segmentation and Discontinuity-preserving Image Registration using Deep Learning." under review.

- **Chapter 4:**

  1) Chen X, Ravikumar N, Xia Y, Attar R, Diaz-Pinto A, Piechnik SK,

Neubauer S, Petersen SE, Frangi AF. “Shape registration with learned deformations for 3D shape reconstruction from sparse and incomplete point clouds.” Medical Image Analysis, 2021, 74: 102228.

- **Chapter 5:**
  1) Chen X, Xia Y, Dall’Armellina E, Ravikumar N, Frangi AF. “Joint shape/texture representation learning for cardiovascular disease diagnosis from MRI.” under review.

Acknowledgements

I would like to thank my main supervisor, Professor Alejandro Frangi, for giving me the opportunity to pursue a PhD degree and work alongside him at CISTIB. I am thankful for his enthusiastic leadership, continuous constructive criticism, and friendly advice, all of which helped me become a better scientist.

I am very grateful to all my lab mates over the last four years in Leeds for all the happy memories. Special thanks to my supervisors, Dr. Nishant Ravikumar, Dr. Yan Xia, and Dr. Andres Diaz-Pinto, for the numerous friendly and fruitful meetings and guidance. I also thank our software engineers Alejandro Lebrero, Kattia Hoz de Vila, and Christopher Kelly for all the technical support throughout different parts of the project.

The Doctoral Scholarship covering the full tuition fee and living cost stipends from the School of Computing, University of Leeds is appreciated.

Finally, I am deeply grateful to my parents for their unflagging love and ongoing support throughout my life.

Abstract

Cardiovascular diseases (CVDs) are the leading cause of death in the world, accounting for 17.9 million deaths each year, 31% of all global deaths. According to the World Health Organisation (WHO), this number is expected to rise to 23 million by 2030. As a noninvasive technique, medical imaging with corresponding computer vision techniques is becoming more and more popular for detecting, understanding, and analysing CVDs. With the advent of deep learning, there are significant improvements in medical image analysis tasks (e.g. image registration, image segmentation, mesh reconstruction from image), achieving much faster and more accurate registration, segmentation, reconstruction, and disease diagnosis.

This thesis focuses on cardiac magnetic resonance images, systematically studying critical tasks in CVD analysis, including image registration, image segmentation, cardiac mesh reconstruction, and CVD prediction/diagnosis. We first present a thorough review of deep learning-based image registration approaches, and subsequently, propose a novel solution to the problem of discontinuity-preserving intra-subject cardiac image registration, which is generally ignored in previous deep learning-based registration methods. On the basis of this, a joint segmentation and registration framework is further proposed to learn the joint relationship between these two tasks, leading to better registration and segmentation performance. In order to characterise the shape and motion of the heart in 3D, we present a deep learning-based 3D mesh reconstruction network that is able to recover accurate 3D cardiac shapes from 2D slice-wise segmentation masks/contours in a fast and robust manner. Finally, for CVD prediction/diagnosis, we design a multichannel variational autoencoder to learn the joint latent representation of the original cardiac image and mesh, resulting in a shape-aware image representation (SAIR) that serves as an explainable biomarker. SAIR has been shown to outperform traditional biomarkers in the prediction of acute myocardial infarction and the diagnosis of several other CVDs, and can supplement existing biomarkers to improve overall predictive performance.

Contents

List of Figures

List of Tables

Abbreviations

n-D	n-Dimensional, n∈ {2,3,4}	ACDC	Automated Cardiac Diagnosis Challenge
AMI	Acute Myocardial Infarction	ARV	Abnormal Right Ventricle
CBCT	Cone Beam Computed Tomography	CC	Cross Correlation
CNN	Convolutional Neural Network	CT	Computerised Tomography
CVD	Cardiovascular Disease	DCM	Dilated Cardiomyopathy
DL	Deep Learning	DLIR	Deep Learning based Image Registration
DSC	Dice Similarity Coefficient	DVF	Displacement Vector Field
ED	End-Diastole	ES	End-Systole
LA	Left Atrium	LAX	Long-Axis View
LVEDV	Left Ventricle End-Diastole Volume	LVESV	Left Ventricle End-Systole Volume
LVEF	Left Ventricle Ejection Fraction	LVMM	Left Ventricle Myocardium Mass
LV	Left Ventricle	MCVAE	Multi-Channel Variational Autoencoder
MI	Myocardial Infarction	MR	Magnetic Resonance
NMI	Normalised Mutual Information	RV	Right Ventricle
RVEF	Right Ventricle Ejection Fraction	RA	Right Atrium
RVEDV	Right Ventricle End-Diastole Volume	RVESV	Right Ventricle End-Systole Volume
SAX	Short-Axis View	STN	Spatial Transformer Networks
SVF	Stationary Velocity Field	TRUS	Transrectal Ultrasound
UKBB	UK Biobank	US	Ultrasound
VM	Voxelmorph		

Chapter 1

Introduction: Background, Motivation and Contribution

Cardiovascular diseases (CVDs) are the leading cause of death worldwide. In 2020, it is estimated that there were approximately 19 million fatalities attributed to CVDs [3]. To assist medical professionals in understanding, detecting and analysing CVDs, medical imaging technologies and related processing and analysis algorithms have become increasingly important. With the emergence of deep learning, there has been considerable advancement in the automated analysis of CVDs. In this chapter, we first provide an overview of the fundamental knowledge of cardiac image analysis and CVDs, followed by a summary of our motivations and contributions in this thesis.

1.1 Cardiac Anatomical Structure, Imaging and Cardiovascular Disease

1.1.1 Cardiac Anatomy and Structure

The heart is an essential organ of the human body, which pumps blood through the circulatory system to provide oxygen and nutrients. Generally, the heart is made up of four chambers, including the left ventricle (LV), right ventricle (RV), left atrium (LA) and right atrium (RA), as illustrated in Figure 1.1. The RA receives deoxygenated blood from the body and passes it through the RV, where it is pumped to the lungs for oxygen-carbon dioxide exchange. At the same time, LA receives oxygen-rich blood from the lungs and passes it to LV, then LV pumps it out to the rest of the body. Among the four chambers, the left ventricle and right ventricle (bi-ventricle) are the most studied in image-based cardiac analysis works, since they are available in most cardiac image datasets. To better describe the structure of the LV, it can be divided into three parts, the epicardium (the external layer), the myocardium (the central layer) and the endocardium (the internal layer). The myocardium is also known as the cardiac muscle and is responsible for the contractility of the heart and the pumping action.

The heart of living people keeps a cycle of motion, from diastole to systole, which continues to repeat continuously. In existing publicly available cardiac image datasets (e.g. UK Biobank (UKBB)), the cardiac cycle can be split into 50/30/20 time frames (depending on the imaging operation). In cardiac image analysis, there are two most important time frames in the cardiac cycle, the frames at end-diastole (ED) and end-systole (ES), respectively. The LV volume in the ED frame is the largest in the cardiac cycle, generally in the first frame. In contrast, the ES frame is the time point at which

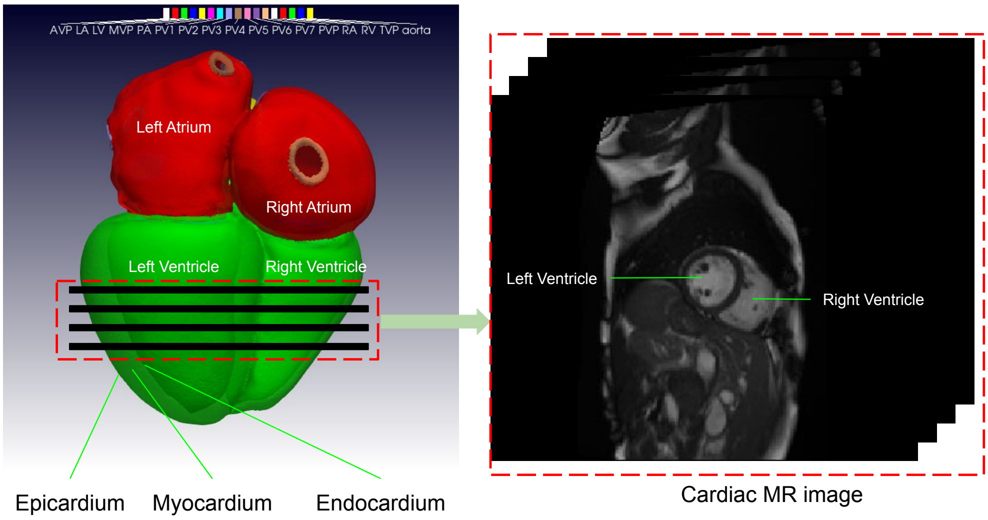

Figure 1.1: The structure of the human heart (the left is a four-chamber mesh, and the right is a short-axis view MR image).

the cardiac volume becomes minimal in the cardiac cycle. Images in these two frames are generally used to analyse cardiac functions and compute clinical indices.

1.1.2 Clinical Cardiac Imaging Techniques

As a noninvasive method, medical imaging has become one of the most essential techniques for understanding the structures and functions of the heart. Similar to other anatomic structures, popular imaging techniques such as magnetic resonance imaging (MR) [4, 5], computed tomography (CT) [6], and ultrasound (US) [7] (often referred to as echocardiography in cardiac imaging [8]) are widely used in cardiac imaging. In addition, certain other imaging modalities, such as single-photon emission computed tomography (SPECT) [9, 10], are also used to diagnose and investigate cardiac pathology.

Different imaging modalities are capable of capturing diverse aspects of the structures and functions of the heart, thus accommodating varying clinical settings. Echocardiography (US imaging) is one of the first-line imaging modalities for cardiac assessment, which works by utilising sound waves that are described regarding frequency, to view images. It is safe, noninvasive, portable, cost-effective, and does not require patients to maintain restricted positions, but the resultant images provide limited as-

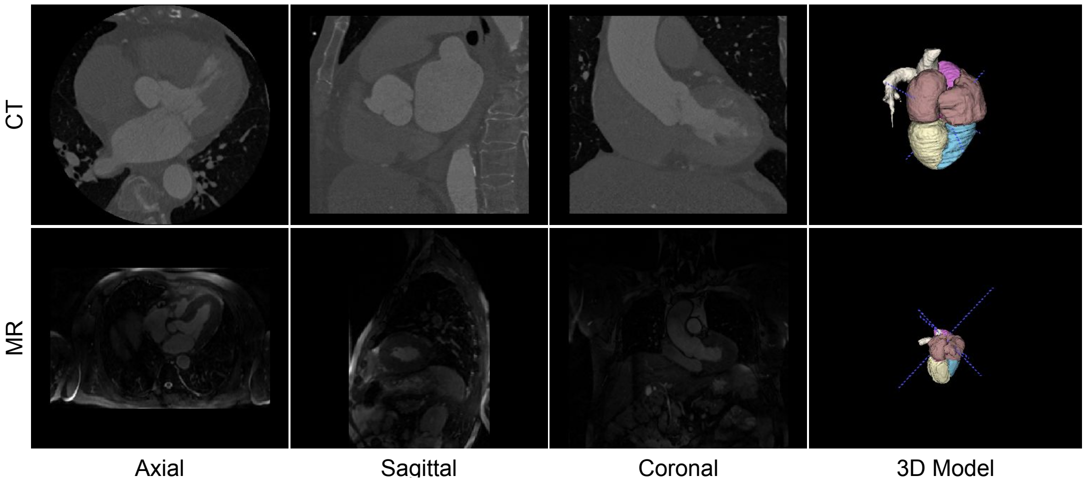

Figure 1.2: Cardiac CT and MR images (from MM-WHS). The first and second rows are CT and MR images, respectively, presenting the images from axial, sagittal, and coronal views. The corresponding meshes are also shown in the right column.

sessment of soft-tissue characteristics and extracardiac structures, and are limited by the acoustic window [8]. Compared with echocardiography, CT and MR generally require patients to be immobile, and to hold their breath. Correspondingly, the obtained images by CT and MR are of higher quality than echocardiography. Figure 1.2 gives an example of MR and CT images of the same person in the multimodal Whole Heart Segmentation Data Set (MM-WHS) [11, 12]. Cardiac CT is a fast imaging technique with X-ray that provides high-quality images with superior spatial resolution [13]. The advantages of CT images are high isotropic spatial and temporal resolution, fast acquisition times, multiplanar image reconstruction capabilities, which make it serve as an alternative to MR imaging in certain scenarios, while the potential adverse effects of radiation exposure cannot be overlooked. Cardiac MR imaging creates images from atomic nuclei with uneven spin using radiowaves in the presence of a magnetic field [14], which has the unique ability to provide quantitative information on cardiac function, perfusion and viability. Compared to CT, MR imaging is safer (without radiation damage) and can provide higher contrast on tissues and anatomical structures. However, despite these benefits, the use of MR imaging is generally associated with increased costs and a longer acquisition time.

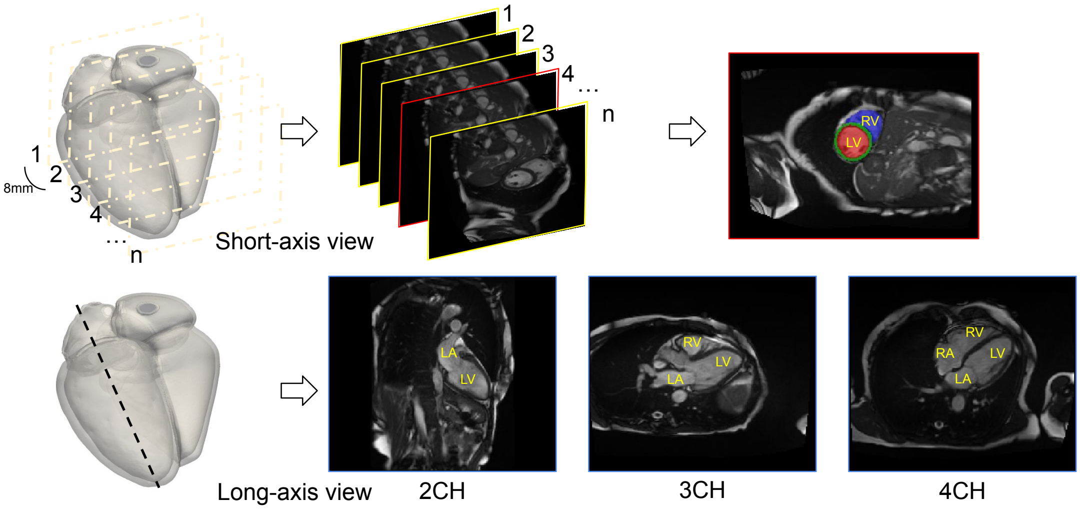

Figure 1.3: An example of cardiac cine MR images. The first row is the SAX images, which are a stack of 2D images with a large slice thickness (for example, in UKBB, the slice thickness is 8 mm). The second row is the corresponding 2CH, 3CH and 4CH LAX images, respectively, which are all 2D images.

In the analysis/diagnosis of CVDs, MR images are frequently considered the gold standard, due to the high contrast on tissues. Accordingly, the present research described in this thesis focusses mainly on the analysis of cardiac cine MR images, which comprise short-axis images (SAX) and long-axis images (LAX). SAX images are mainly used to visualise the structure of LV and RV. Constrained by the inherent limitations of cardiac MR imaging, SAX images generally have a high in-plane resolution and a huge slice thickness. Unlike other 3D images, SAX images are similar to a sequence of 2D images (usually containing less than 15 slices). For LAX, it serves as a supplement of SAX images and can further provide the structure of RA and LA, containing three slices of 2D images from different LAX views (i.e., 2-chamber (2CH), 3-chamber (3CH) and 4-chamber views (4CH)). In certain cases, there may be only 1 or 2 slices of LAX images available. An example of cine MR images from UKBB is shown in Figure 1.3. The SAX images and LAX images are shown in the first and second rows, respectively, where the SAX image is a stack of 2D images with a large slice thickness, and the LAX images in 2CH, 3CH, and 4CH views are all single 2D images.

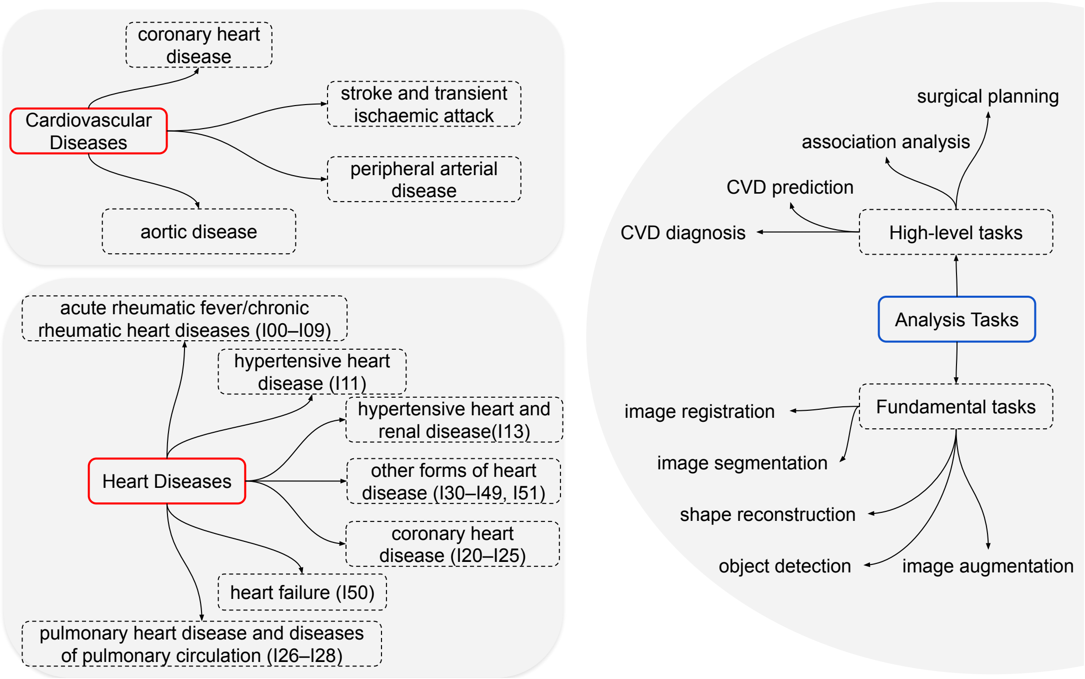
coronary heart disease
Cardiovascular Diseases
stroke and transient ischaemic attack
peripheral arterial disease
aortic disease
acute rheumatic fever/chronic rheumatic heart diseases (I00–I09)
hypertensive heart disease (I11)
hypertensive heart and renal disease(I13)
Heart Diseases
other forms of heart disease (I30–I49, I51)
coronary heart disease (I20–I25)
heart failure (I50)
pulmonary heart disease and diseases of pulmonary circulation (I26–I28)
surgical planning
association analysis
CVD prediction
CVD diagnosis
High-level tasks
Analysis Tasks
Fundamental tasks
image registration
image segmentation
shape reconstruction
object detection
image augmentation

Figure 1.4: List of the main cardiovascular diseases and corresponding analysis tasks. For the categories of heart disease, the corresponding International Classification of Diseases (ICD) 10th Revision codes are also listed.

1.1.3 Cardiovascular Diseases and Corresponding Analysis/diagnosis

According to the definition from British Heart Foundation, cardiovascular disease (CVD), also called heart and circulatory disease, is an umbrella name for conditions that affect human heart or circulation. Heart disease, cardiomyopathy, congenital heart disease, and valvular heart disease are various forms of CVD, a group of conditions that affect the heart or blood vessels, and have become the leading cause of death worldwide in recent years [15]. They are responsible for around 30% of human mortality, as well as 10% of the disease burden in the world [16, 17]. Different CVD classifications have been proposed due to the intricate pathophysiology and systemic impact on the human body. However, the main CVDs are widely recognised in different category lists, such as ischaemic (coronary) heart disease and hypertensive diseases. According to the National Health Service (NHS), there are four main CVDs, including coronary heart disease (as a result of angina, heart attack, and heart failure), stroke, and transient ischaemic attack, peripheral arterial disease, and aortic disease. CVDs are not the same as heart diseases, as the latter are more specific to the heart. Heart diseases include acute rheumatic fever/chronic rheumatic heart disease, hypertensive heart disease, hypertensive heart and kidney disease, coronary heart disease, heart failure, pulmonary heart disease and diseases of the pulmonary circulation, and other forms of heart disease [3] (more details can be found in Figure 1.4). Among those CVDs, the diseases studied primarily in cardiac image analyses (especially learning-based image analyses) are acute myocardial infarction (AMI), myocardial infarction (MI), dilated cardiomyopathy (DCM), hypertrophic cardiomyopathy (HCM), abnormal right ventricle (ARV), as they are available in publicly available datasets (e.g. UKBB, Automated Cardiac Diagnosis Challenge (ACDC) and Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image Segmentation Challenge (M&M)).

To understand and analyse such types of complex diseases, cardiac imaging with the corresponding computer-aided processing and analysis techniques has been widely used. In some scenarios, clinicians can directly make predictions/diagnoses by looking at the cardiac images obtained. However, as different patients may have numerous images of different modalities, scanners, and times, it is a time-consuming and laborious task for clinicians to process and analyse all of these images manually. With computer vision technologies, it is possible to achieve some fundamental pre-processing (e.g. registration, segmentation, detection, denoising, edge detection, restoration, image superresolution, resampling) and higher-level tasks (e.g. classification and recognition) from the raw images automatically. Those techniques can provide higher-quality images, faster image analysis, and sufficient information for decision-making. For example, the registration results provided by automatic registration methods can be used to calculate the strain curve, which is an important parameter to evaluate cardiac motion [18]. With the segmentation masks predicted by automatic segmentation methods in ED and ES frames, it is easy to compute clinical indices, an important biomarker for CVD prediction/diagnosis. Based on these tools, clinicians can focus on critical things such as prediction/diagnosis and decision-making, leading to more efficient and accurate CVD analysis.

1.2 Learning-based Cardiovascular Disease Analysis

Learning-based CVD analysis employs trainable approaches to address CVD-related tasks, where models are trained on a designated data set and then tested on unseen data. As shown in Figure 1.4, automatic CVD analysis tasks that involve cardiac MR images comprise a group of fundamental tasks (e.g. image registration, image segmentation, mesh reconstruction) and high-level tasks (such as CVD diagnosis, CVD prediction, and CVD association analysis, to name a few). The former is generally used as the preliminary tasks of the latter. In this thesis, we mainly study the registration/segmentation and mesh reconstruction in the fundamental tasks. For high-level tasks, this thesis focusses on CVD prediction and diagnosis, two of the most significant tasks among them. The diagnosis of CVD is to check the presence of CVD by examination of cardiac MR images and other possible biomarkers. Similarly, CVD prediction uses the same information as CVD diagnosis to estimate the likelihood of CVD occurring within a specific period (e.g. 10 years). For both, approaches to learning the corresponding biomarkers from MR images are essentially important.

Unlike other diseases (e.g. cancer), CVDs are a group of complex diseases affected by heart motion and vessels, which are difficult to diagnose/predict using the image from a single time frame. Therefore, temporal/motion information is generally required for CVD prediction/diagnosis. As jointly analysing the complete cardiac cycle is computationally intensive and comprises redundant information, previous studies have chosen images from the ED and ES frames for cardiac analysis. The raw images can be used for diagnosis/prediction directly, while it is easy to be affected by background tissues,

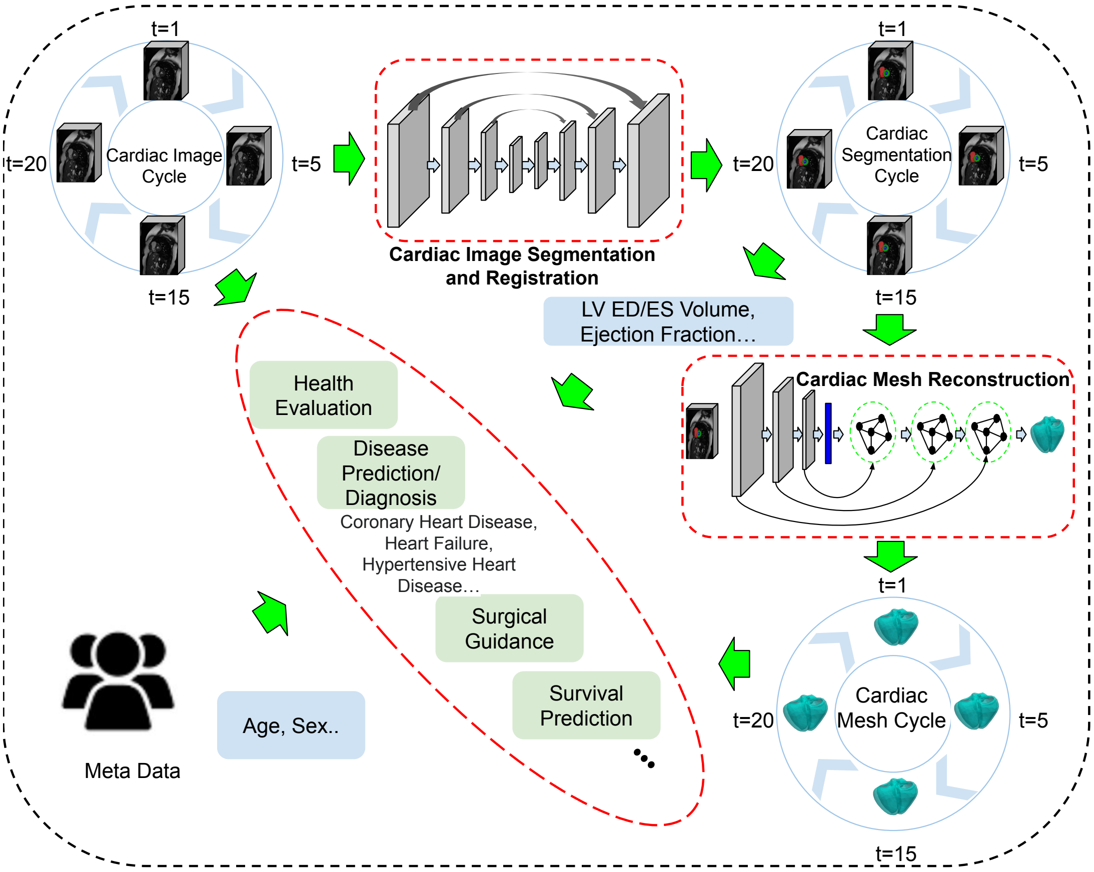
t=1
t=20
Cardiac Image Cycle
t=5
t=15
Cardiac Image Segmentation and Registration
t=1
t=20
Cardiac Segmentation Cycle
t=5
t=15
LV ED/ES Volume, Ejection Fraction…
Cardiac Mesh Reconstruction
Health Evaluation
Disease Prediction/ Diagnosis
Coronary Heart Disease, Heart Failure, Hypertensive Heart Disease…
Surgical Guidance
Survival Prediction
Age, Sex..
Meta Data
t=1
t=20
Cardiac Mesh Cycle
t=5
t=15

Figure 1.5: An overview figure of deep learning-based CVD analysis.

and brings a high computation burden for traditional machine learning approaches. Moreover, while deep learning networks can directly take raw images as input and predict the corresponding diagnosis/prediction results, they lack interpretability and are unable to provide efficient feedback in a clinical sense to clinicians. To obtain more efficient and explainable biomarkers, some preprocessing steps are necessary to exclude background tissues or generate new representations that focus on the heart.

In Figure 1.5, we present a general CVD analysis pipeline (using MR images as an example). In addition to raw images, other cardiac representations such as segmentation masks and meshes are widely used in CVD analysis. The segmentation masks allow for the identification of various anatomical structures of the heart (e.g. LV blood pool, LVM and RV) by distinguishing them from the background. The cardiac meshes are 3D structures with the coordinates of vertices/surfaces along the boundary of the heart in real-world space, which provides a more intuitive depiction of the cardiac shape in 3D. Both segmentation masks and cardiac meshes contain only heart-related information. However, the latter provides complete 3D spatial structures and enables a better presentation of cardiac motion. The segmentation masks and meshes are derived from the raw images using image segmentation or mesh reconstruction methods but yield enhanced anatomical information beyond raw images by leveraging anatomical priors. Cardiac image registration, another important tool for capturing motion information, predicts point-to-point correspondence between images at different time frames of the cardiac cycle. The deformation fields obtained from this process enable the calculation of strain curves and other motion-related biomarkers critical for the analysis of CVD [19, 20]. Automatic CVD prediction/diagnosis methods can rely on a joint consideration of metadata (the fundamental information of patients, e.g. sex, age) and the three aforementioned representations (using part or all of them) to make decisions.

Given raw MR images, segmentation is applied to segment the region of the heart from the background, either automatically (using machine learning-based segmentation methods) or manually. Based on the segmentation results in the ED and ES frames of the cardiac cycle, nine important clinical indices are calculated for subsequent analysis, including left ventricle end-diastole volume (LVEDV), left ventricle end-systole volume (LVESV), left ventricle ejection fraction (LVEF), left ventricle myocardium mass (LVMM), right ventricle end-diastole volume (RVEDV), right ventricle end-systole volume (RVESV) and right ventricle ejection fraction (RVEF). These clinical indices serve as fundamental features in the prediction/diagnosis of CVDs. Beyond these indices, radiomic features [21] can also be extracted from original cardiac images and the corresponding segmentation masks, as an additional efficient predictor. In traditional methods, metadata, clinical indices, and radiomic features are the most popular features/predictors for CVD prediction/diagnosis.

The 3D cardiac mesh is another important representation of the heart, which plays a critical role in surgical planning, surgical navigation, and many other CVD analysis tasks. Unlike cardiac images, cardiac meshes only display the structure of the heart, without interference from surrounding tissues. It can be obtained from the cardiac segmentation masks [22] or directly from raw images [23]. The cardiac meshes correspond to the spatial coordinates of the heart in real 3D space, thereby providing an accurate and more intuitive way to present the cardiac motion. Therefore, some researchers have proposed to apply cardiac mesh for the prediction of survival in patients with CVDs [24]. Furthermore, reconstructed meshes can also be applied for cardiac image segmentation by overlaying them on the corresponding images [23], thus it can be seen as an advanced representation of segmentation masks.

The primary challenge in image-based CVD analysis is the development of effective biomarkers for subsequent prediction and diagnosis. Previous research has typically focussed on individual cardiac representations such as raw images, segmentation masks, or 3D meshes. For example, numerous traditional methods only use manually designed biomarkers extracted from raw images or segmentation masks, for CVD prediction/diagnosis [25, 26, 21]. Recent research has observed deep learning-based methods for survival prediction, using cardiac meshes [24]. Compared to raw images, cardiac meshes emphasise cardiac structures and exclude background tissues, providing a better representation of the shape of the heart while losing local details. Raw MR images, on the other hand, provide high-contrast local details and serve as the source of cardiac mesh, but lack a comprehensive overview of the heart. Therefore, it is natural to consider combining the advantages of both representations for CVD analysis.

Learning-based CVD analysis has great potential in various domains of CVD analysis tasks, including prediction, early detection, diagnosis, surgical planning, and surgical navigation, among others. It can provide clinical decision support and serve as an objective reference without human intervention, thus contributing significantly to the prevention of CVDs and extending the human lifespan. Recent advances in deep learning have facilitated faster and more precise analysis of medical images. However, most of the existing research has focussed on individual tasks such as segmentation, registration, and mesh reconstruction, with little attention to a comprehensive perspective. Thus, this thesis aims to identify critical tasks in CVD analysis and to analyse the latent correlations between them, thus establishing a foundation for future research in the domain of automatic CVD analysis.

1.3 Thesis Contributions and Overview

This thesis aims to explore fast, explainable, and precise image-based CVD analysis, with advanced deep learning-based approaches. To do this task, we look at three critical tasks in CVD analysis, cardiac image registration/segmentation, cardiac mesh reconstruction, and CVD prediction/diagnosis. Consequently, there are three main objectives that should be taken into consideration:

- Build more realistic and accurate deep learning-based registration and segmentation methods to address the unique challenges inherent in cardiac imaging scenarios. Current deep learning-based registration methods assume that the deformation fields are entirely smooth, which is not always applicable in medical image registration, particularly in cardiac and abdominal image registration. Moreover, given that image registration and segmentation are fundamental tasks that are closely related to each other, it is essential to consider them jointly, particularly in the context of cardiac images, which can further enhance the performance of both tasks.

- Design fast, accurate and robust shape reconstruction methods to learn the corresponding 3D shapes from raw cardiac images. While raw MR images contain high-contrast details for diagnosis, they are also subject to interference from background tissues. Additionally, the large thickness of the cut in the MR imaging of SAX images leads to a lack of structural information between the slices. As a result, reconstructing 3D cardiac meshes from the original images is necessary to obtain an overall understanding of the cardiac structure and improve the analysis of cardiac motion. Although numerous traditional approaches have been proposed to achieve this task, they typically involve multiple iterations in inference, which is generally time-consuming and limits their performance in realistic scenarios. Therefore, it is important to develop novel methods that can accurately reconstruct 3D cardiac meshes from raw images, in a fast and robust manner.

- Explore an efficient and explainable cardiac biomarker and use it for subsequent CVD prediction/diagnosis. For accurate CVD prediction/diagnosis, it is crucial to utilise all the information available from the raw data. However, previous research typically uses individual representations of the heart (e.g. raw images, segmentation masks, or meshes) and may not achieve optimal prediction/diagnosis performance, due to the inherent limitations on specific representations. To address this limitation, it is necessary to explore efficient and explainable cardiac biomarkers by jointly considering multiple representations of the heart, rather than relying on a single representation.

The following chapters in this thesis are organised as follows:

Chapter 2: This chapter gives a thorough review of the literature on deep learning-based image registration approaches. We review all deep learning-based image registration methods since 2013, make a detailed summary of existing deep learning-based image registration methods and provide an in-depth analysis of the current development trends, as well as limitations that need to be addressed. We also discuss potential future directions for research in this area.

Chapter 3: This chapter is to capture motion information in the cardiac cycle by image registration and segmentation. We first propose a weakly supervised discontinuity-preserving image registration network, DDIR, to predict more realistic deformation fields in intra-subject registration. On the basis of DDIR, a joint registration and segmentation network is further proposed by introducing an additional segmentation subnetwork. We simultaneously accomplish registration that preserves discontinuities and segmentation based on co-attention in our joint registration and segmentation framework. With only moving and fixed images as input, we can achieve more realistic and accurate registration performance than previous registration approaches while obtaining precise segmentation masks.

Chapter 4: This chapter presents a 3D cardiac shape reconstruction approach from MR images/contours. We propose an end-to-end deep graph convolution network, named MR-Net, that enables fast and robust reconstruction of accurate 3D shapes from stacked 2D contours, significantly outperforming traditional approaches. In addition, our proposed method can rapidly and accurately reconstruct 3D cardiac meshes from raw MR images, with the support of pre-trained deep learning-based segmentation approaches.

Chapter 5: This chapter aims to learn efficient and explainable biomarkers for the prediction and diagnosis of CVDs. To do this task, we built a mesh-image variational autoencoder (MIVAE) to learn the joint latent representations of cardiac images and meshes. MIVAE enables the extraction of shape-aware image representations that take into account the local details present in MR images, as well as the global structure of cardiac meshes, leading to improved accuracy in CVD prediction/diagnosis (compared to traditional biomarkers). Additionally, the MIVAE can work as a mesh reconstruction method when MR images are given alone as input.

These four chapters of this thesis are independent and based on articles that are already published or under review in peer-reviewed conferences/journals. The last chapter, Chapter 6, summarises the thesis and examines current restrictions and potential future directions.

Chapter 2

Literature Review on Deep Learning in Image Registration

Image registration is a fundamental task in multiple medical image analysis applications. With the advent of deep learning, there have been significant advances in algorithmic performance for various computer vision tasks in recent years, including medical image registration. The last couple of years have seen a dramatic increase in the development of deep learning-based medical image registration algorithms. Consequently, a comprehensive review of the current state-of-the-art algorithms in the field is timely and necessary. This chapter aims to understand the clinical applications and challenges that drove this innovation, analyse the functionality and limitations of existing approaches, and provide insight into open challenges and unmet clinical needs that could shape future research directions. To this end, the main contributions of this chapter are as follows,

- Discussion of all papers on deep learning-based medical image registration published since 2013 with significant methodological and/or functional contributions to the field.
- Analysis of the development and evolution of deep learning-based image registration methods, summarising the current trends and challenges in the domain.
- Overview of unmet clinical needs and potential directions for future research in deep learning-based medical image registration.

2.1 Introduction

Medical image registration has been a central component of various applications in medical image analysis over the last three decades. The field has evolved immensely with growth in computational resources, algorithmic capabilities, and complexities. Various clinical applications that involve disease diagnosis and monitoring, image-guided treatment delivery, and postoperative evaluation use image registration. It is also widely used as a tool to preprocess data for subsequent tasks such as object detection, segmentation, or classification, as variation in the spatial resolution of medical images is very common. Consequently, the performance of the latter is heavily influenced by the quality of the image registration algorithm used to bring the images to a common coordinate frame and fixed size and resolution.

2.1.1 Framework of Registration

Image registration is the process of identifying a spatial transformation that maps two (pair-wise registration) or more (group-wise registration) images to a common coordinate frame so that the corresponding anatomical structures are optimally aligned, or, in other words, a voxel-wise "correspondence" is established between the images. Depending on the degrees of freedom associated with the desired spatial transformation, image registration algorithms may be broadly grouped into rigid, affine, or nonrigid/deformable. In the case of pairwise image registration, this can be formally defined as follows: Let F and M denote fixed and moving images, respectively, and let T be the desired spatial transformation that maps the voxels of M to those of F. Registering the two images can be posed as an optimisation problem expressed as:

$$\widehat{T} = \arg\min_{T} \mathcal{S}(F, T(M)), \tag{2.1}$$

where $\mathcal{S}()$ represents a measure of dissimilarity (or similarity depending on the formulation of the objective function) between the fixed image and the warped moving image. Images are recorded by iteratively improving estimates for the desired T, such that the defined $S()$ in the cost function is maximised or minimised.

Intuitively, nonrigid or deformable image registration is an ill-posed problem, which makes it fundamentally different from other computer vision tasks such as object localisation, segmentation, or classification. For example, given two images as input, deformable image registration aims to find a spatial transformation that warps the moving image to match the fixed image as closely as possible. However, there is no ground truth available for the desired deformation field and, without enforcing any constraints on the properties of the spatial transformation, the resulting cost function is ill-conditioned and highly nonconvex. To address the latter and ensure tractability, all image registration algorithms regularise the estimated deformation field, based on some prior assumptions on the properties of the underlying unknown deformation.

Conventionally, medical image registration algorithms comprise three distinct components: a transformation model, a similarity metric, and an optimisation algorithm, as illustrated in Figure 2.1. The overall process of image registration involves: (1) design/choice of a suitable transformation model (rigid, affine, or nonrigid) and initialisation of its associated parameters, (2) use of the transformation model to warp the moving image, (3) evaluation of the dissimilarity between the warped moving image

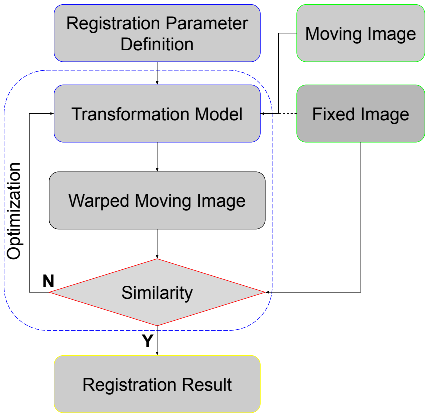

Figure 2.1: A flowchart of medical image registration framework.

and the fixed image, and (4) update of the parameters in the transformation model by optimising the cost function formulated using the dissimilarity metric, using a suitable optimisation algorithm. The registration algorithm iterates between step(2) - step(4) until a suitable convergence criterion is satisfied (usually based on the change in the dissimilarity metric or the transformation parameters between iterations). As image registration using conventional algorithms is an iterative process, they are typically computationally intensive and time-consuming. The overall framework is generic and can be formulated within a Deep Learning (DL) setting, enabling significant acceleration, for registering a pair or group of unseen images using a trained registration network.

2.1.2 Basic Deep Learning Networks

Although the theoretical concepts that underpin neural networks have existed for decades, early attempts to train such algorithms [27], [28] were constrained by the limited computational power available at the time. Recent years have witnessed an almost exponential growth in the development and use of DL algorithms, sustained thus far by rapid improvements in computational hardware (e.g. GPUs). Consequently, clinical applications requiring image classification, segmentation, registration, or object detection/localisation, have witnessed significant improvements in algorithmic performance, in terms of accuracy and/or efficiency. Although DL-based medical image registration algorithms have yet to achieve significant breakthroughs in terms of registration

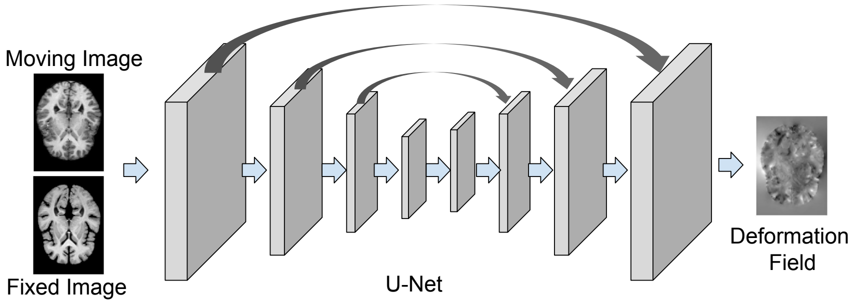

Figure 2.2: An example of the U-Net framework used for brain MRI image registration. The moving image and fixed image are concatenated at first. A U-Net takes it as input and predicts the deformation field. The U-Net is an encode-decoder network.

accuracy compared to traditional methods, they have provided a means to accelerate registration many times. To offer a basis for understanding deep learning-based registration methods, we briefly introduce and discuss three fundamental and widely used components of image registration networks, namely an encoder-decoder Convolutional Neural Network (CNN), a Spatial Transformer Network (STN) [29], and a Generative Adversarial Network (GAN) [30].

The success of deep learning in visual recognition tasks is mainly due to convolutional neural networks (CNNs). This type of deep learning network has a hierarchical structure of replicated feature detectors or, in other words, successive layers of "convolution" that are used to automatically learn multiscale features specific to tasks. Several CNN architectures have been proposed in recent years, each with specific architectural modifications to address the issue of vanishing/exploding gradients common to deep networks, such as AlexNet [31], VGG [32], ResNet [33], and DenseNet [34]. Among these, in medical image segmentation and registration, the most widely used architecture is the U-Net [35] - an encoder-decoder-style network with skip connections between the encoding and decoding paths (as depicted in Figure 2.2). The encoder contains several convolutional layers and pooling layers, which downsample the input image to a low resolution. However, the decoder is made up of deconvolution layers with a number of layers that correspond to the encoder. Through the decoder, the feature maps are reconstructed to the original size of the input images. The U-Net uses several down-sampling and up-sampling layers to learn features at different resolutions,

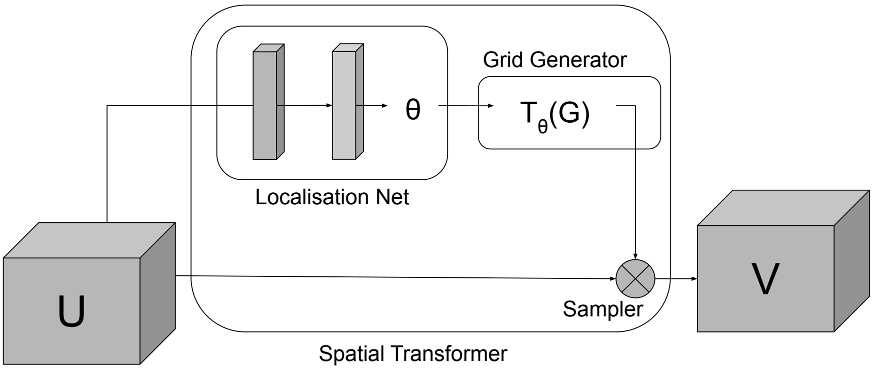

Figure 2.3: Illustration of spatial transfer network, which consists of three sub-blocks: localisation net, grid generator and sampler. U and V mean the input and output respectively. Localisation net is used to learn the transformation parameter θ from U. With θ, the grid generator can generate transformation grids. Then a bilinear/trilinear sampling is applied to sample coordinates from U to V. The whole network is differential so the back-propagation could update the parameters automatically.

at the limited expense of computational resources. It has been widely applied in various medical imaging applications (e.g. segmentation), and due to its flexibility, most state-of-the-art Deep Learning-based medical Image Registration (DLIR) methods use it as well in some components of the overall framework.

Another core component of most DLIR approaches is STN, proposed in 2015 [29], which learns to spatially transform feature maps in a manner beneficial to the task of interest. Although they were not explicitly designed for image registration but rather to imbue networks with the means to learn features in a manner invariant to rigid and deformable transformations, they have become the basis for most unsupervised registration methods. As shown in Figure 2.3, STN includes three components: a localisation network, a grid generator, and a sampler. The localisation network is a CNN, which takes feature maps as input and outputs the parameters of a suitable/user-specified spatial transformation. The transformation parameters are then used to create a resampling grid with the grid generator. After that, a linear sampler is used to carry out differentiable image sampling based on the grid created in the prior step. For 3D rigid registration, the spatial transformation is composed of just six parameters, namely, three rotation parameters and three translation parameters. In the case of nonrigid registration, the localisation network estimates a deformation field represented in a parametric or non-parametric form, as defined by the user, of the same size as the input. Most DLIR methods could be seen as expanded STNs, which look to improve the performance of the localisation network to generate more accurate deformation fields for warping the moving image. As with conventional image registration algorithms, the objective function optimised in image registration networks is a similarity/dissimilarity metric computed between the warped moving image and the fixed image, in addition to suitable regularisation terms which ensure that the problem is suitably constrained and well-posed. The latter also controls the smoothness of the estimated deformation field.

As STN gives neural networks the ability to spatially transform feature maps, it has become the basis of most of the DLIR methods, especially unsupervised/weakly-supervised DLIR methods. The generator in Figure 2.4 could be seen as a general DLIR framework, which consists of a CNN (U-Net) and a spatial transform block (refer to STN). The CNN takes the moving image and fixed image as input and predicts a deformation field (deformable registration), and then the spatial transform block deforms the moving image based on the predicted deformation field. The registration networks are thus formulated as end-to-end networks that utilise CNN and STN to jointly estimate the desired deformation fields and warp the moving images. The losses of similarity/dissimilarity (between warped moving images and fixed images) and regularisation (on deformation fields) would be computed to update the parameters in the CNN. Once the network is trained, the registration between new image pairs is just one forward prediction.

Generative adversarial networks [30] are also a common component of DLIR approaches. They are the most widely used generative models for image synthesis [36, 37, 38] and have found use in the medical domain as tools for data augmentation [39], and for applications requiring image-to-image translation [40], and segmentation [41], among others. It contains two parts, a generator and a discriminator, both of which are typically convolutional neural networks. The former constitutes the generative model in the network, which learns to sample from the data distribution and can be used to synthesise new instances. The latter, on the other hand, is used to distinguish between synthesised and real samples, thus competing with the generator, or in other words, acting as its “adversary”. Essentially, GANs are trained in a minimax two-player game,

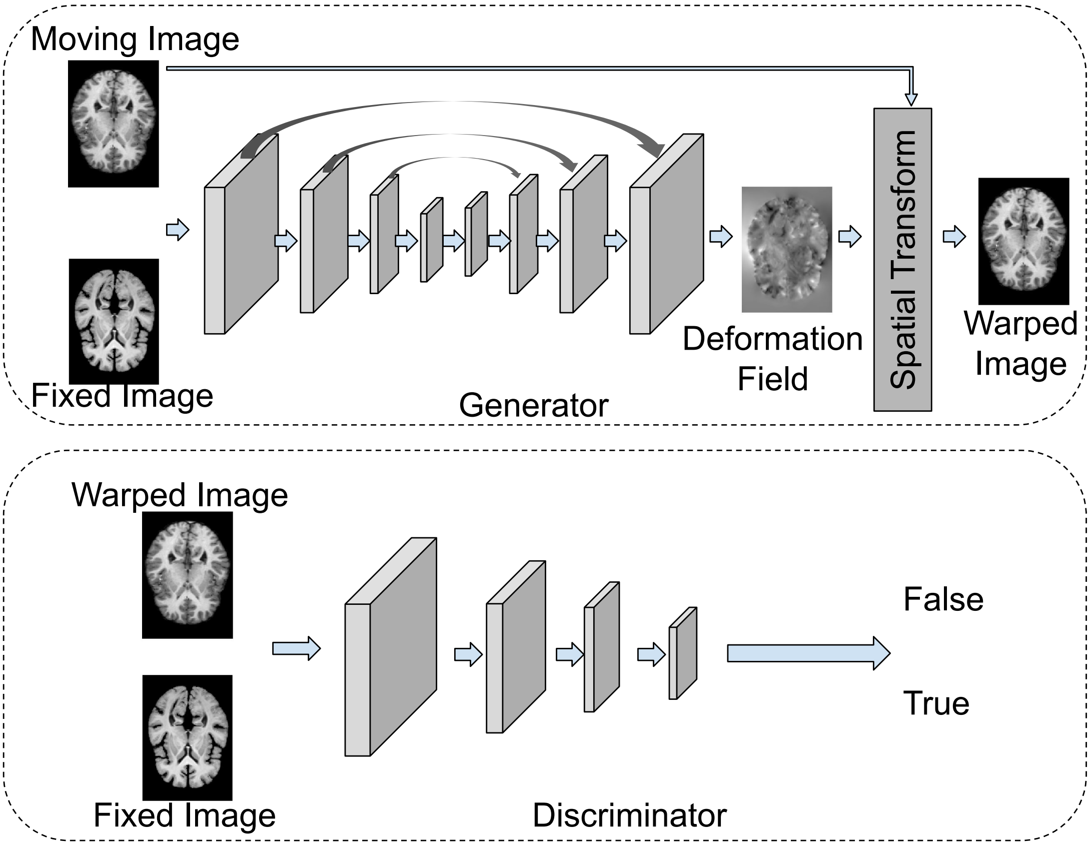

Figure 2.4: An example of GAN-based image registration. The generator combines a U-Net and an STN to synthesise the deformation field and the warped moving image simultaneously. The discriminator is used to discriminate the difference between the warped moving image and the fixed image, urging the generator to predict a high-similarity warped moving image to the fixed image.

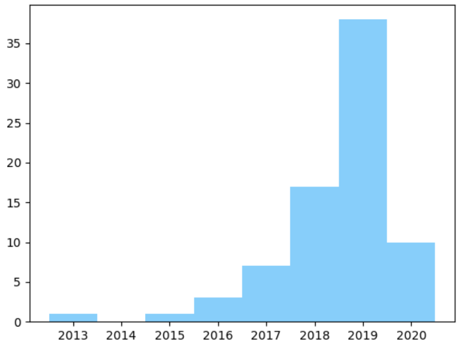

Figure 2.5: An overview of the number of papers published from 2013 to 2020 about DLIR methods.

where the generator looks to maximise the probability of the discriminator mistaking a synthesised sample as a real one from the training data. This leads to both networks learning hierarchical representations of the training data in an unsupervised fashion. A generic GAN-based registration framework is shown in Figure 2.4. With the fixed image and moving image as input, the generator predicts the warped image. Then the discriminator evaluates how similar the warped image is to the fixed image. The discriminator in GANs offers a novel learnable mechanism to evaluate the similarity between two images. This property has significant potential to build adaptable and learnable similarity metrics, especially relevant for multimodal image registration. In numerous multimodal registration approaches, GAN-based image translation networks (e.g. Cycle-GANs [37]) learn to map the appearance shift between domains, i.e. between images from different modalities. This simplifies the task of choosing a suitable similarity metric by transforming the multimodal registration problem into a monomodal one. Consequently, GAN-based networks are widely used in medical image registration, which we discuss in more detail in Section 2.2.

The aim of this chapter is to provide a critical overview of existing literature on DL-based image registration, by highlighting innovations from a methodological and functional perspective, discussing current trends, challenges, and limitations, and providing insights to the possible directions for future research. While several review papers have recently been published on DL-based medical image registration [42, 43, 44, 45], they mainly focus on the architecture of networks proposed for DL-based medical image registration, grouping, and discussing them according to their design and learning paradigms (i.e., supervised, weakly supervised, or unsupervised, for example). Consequently, in this chapter, we provide an up-to-date detailed account of both the methodological and functional contributions of the DLIR techniques proposed thus far. To facilitate benchmarking of existing DLIR approaches and provide future work with a frame of reference for comparison, we also present a comprehensive summary of publicly available datasets used to design and validate numerous DLIR methods and provide links for all methods with publicly available code. In this chapter, we include 77 papers that focus on DLIR, with the majority published after 2016. The increasing adoption of DL for medical image registration is highlighted by the yearly count described in Figure 2.5. Throughout the review, we provide statistics on the number of DLIR papers published, grouped according to their methodological and functional characteristics. We restrict the focus of this chapter to publications concerned with medical image registration alone. To identify relevant publications, PubMed and Web of Science were queried for papers using combinations of terms such as — DL, medical image registration, deformable image registration, image fusion, multimodal image registration, motion tracking, among others. In addition to these databases, other sources such as Google Scholar[1], ArXiv[2] and Semantic Scholar[3] were also searched in the same way, and publications with significant contributions to the community were selected for review.

The remainder of this chapter is organised as follows. In Section 2.2, we discuss how DL networks are applied in the registration of medical images. Section 2.3 describes those methods from the perspective of applications. Sections 2.4 and 2.5 discuss development trends, main challenges/limitations, and possible directions of innovation for DL-based medical image registration.

2.2 Deep Learning Based Medical Image Registration

The fundamental building blocks of image registration are identical in both traditional and DL-based approaches, comprising a similarity metric, transformation model, and optimiser. Neural networks have been integrated into this framework, repla-

[1] https://scholar.google.co.uk/

[2] https://arxiv.org/

[3] https://www.semanticscholar.org/

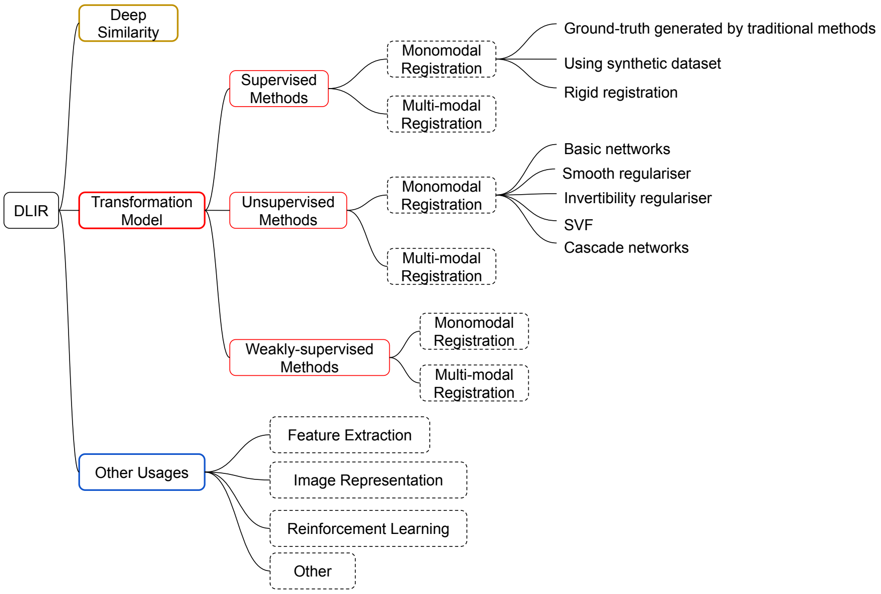

Figure 2.6: Classes of deep learning-based image registration methods.

cing/enhancing the role played by one or several of these components. We categorise DLIR methods into three parent classes, namely, approaches that (a) use neural networks as a similarity metric (often called deep similarity); (b) parameterise the transformation model using neural networks; and (c) employ neural networks to facilitate other operations (such as feature extraction or learning new image representations, referred to as other usages in this chapter) that improve registration quality. Each of these categories can be further divided into subgroups as described by Figure 2.6, and will be discussed in subsequent sections.

2.2.1 Deep Learning for Similarity Metrics

In traditional medical image registration methods, studies often focus on improving the similarity metric to obtain a higher registration accuracy. Various similarity metrics have been used in previous studies, such as cross-correlation (CC), mutual information (MI), and dice similarity coefficient (DSC), corresponding to different scenarios, without sufficient justification for their choice in many cases. That is, these similarity metrics were not application or immorality specific, as they were learnt from or designed for the images to be registered. Visual recognition and perception tasks have benefited substantially from the ability of deep neural networks (specifically, convolutional neural networks) to extract features and combine them across multiple scales, providing the possibility to evaluate the distance between images from different modalities, in a common feature space. Several studies [46, 47] have used neural networks as data-driven, learnable interpretations of similarity metrics, providing a framework that adapts to different applications and image modalities.

DL-based similarity metrics are usually employed for multimodal image registration due to the substantial variation in the appearance and intensity distributions of the moving and fixed images. For example, Haskins et al. [46] proposed a similarity metric based on a regression CNN to register Magnetic Resonance Imaging (MRI) and Transrectal Ultrasound (TRUS) images, which demonstrated promising performance compared to MI, and several other conventional similarity metrics. Deep CNN-based similarity metrics have also been demonstrated to be useful for monomodal image registration. For instance, Zhu et al. [47] used a pre-trained CNN as a similarity metric for ultrasound (US) image registration, showing comparable or better performance than manual registration.

Additionally, the formulation of the discriminator in GANs naturally lends itself to use as a similarity metric, as its role in distinguishing between generated and real images can easily be reformulated as one of computing the difference between the warped and fixed images. Such metrics are often referred to as adversarial similarity and have been used in several unsupervised image registration networks [48, 49, 50, 51].

Although deep neural networks employed in this context offer more robustness and flexibility than traditional similarity metrics, the image registration process is still iterative. Therefore, while methods within this category can achieve registration accuracy that is similar to or better than conventional approaches, they are still time-consuming during inference.

2.2.2 Deep Learning for Transformation Models

In this section, we discuss approaches that parameterise spatial transformations using deep neural networks. As described by Figure 2.6, this category of approaches can be further divided into supervised, weakly supervised, and unsupervised approaches, based on the learning paradigm used to train the networks. The fundamental advantage of this group of techniques over conventional approaches and deep similarity networks is the substantial acceleration they afford during inference, enabling real-time rigid and nonrigid image registration.

Supervised Registration

This sub-group of techniques employs deep neural networks to estimate the spatial transformation parameters necessary to register two (or a group of) images, in a "supervised" fashion, i.e. using ground-truth/target values for the parameters to guide the learning process. As with other supervised learning approaches common to medical image analysis tasks such as segmentation or classification, such techniques depend on the availability of ground-truth/target values for the transformation parameters. In general, there are two methods to obtain these target parameters: (a) by estimating them using traditional registration methods; or (b) by using simulated images with known ground-truth transformations. Supervised registration networks thus estimate the parameters associated with the transformation model adopted (rigid or non-rigid) to warp the moving image to the fixed image space, and subsequently, compute the loss between predicted parameters and ground-truth values. This loss over the transformation parameters, in turn, is used to compute its gradients with respect to the weights of the network, which parameterise the spatial transformations, and is used to guide the training of the network. Following training, the registration of two or more images is achieved as a single forward pass through the network, substantially reducing the execution time relative to iterative approaches.

Table 2.1 summarises the most relevant supervised DL-based medical image registration methods that we identified for this chapter. To provide readers with operationally useful information, we also provide links to repositories for all methods that have made their code publicly available. We further group supervised methods into monomodal registration and multimodal registration. Monomodal registration, also called unimodal registration, aims to register moving images and fixed images from the same modality such as MRI, computed tomography (CT), and X-ray. Multimodal registration is applied to register images from different modalities (e.g. CT to MRI, X-ray to MRI). We found that a large proportion of existing supervised DLIR methods are monomodal (refer to Table 2.1). As obtaining ground-truth transformations is a key problem for supervised registration methods, we further classify the monomodal registration methods into three classes: (a) generating them using traditional regis-

Table 2.1: A summary of supervised DL-based registration methods. All the supervised methods are firstly classified into two general classes: monomodal registration and multimodal registration. Then monomodal registration methods are further categorised according to the methodology of obtaining ground truth. Hyperlinks are given for those works with code publicly available.

Registration	Reference	Network	Modality	Dimension	Organ	Code
Monomodal	**Ground-truth generated by traditional methods**					
	Yang et al., 2016 [52]	Encoder-decoder	MRI	2D, 3D patch	Brain	link
	Cao et al., 2017 [53]	Similarity-steered CNN regression	MRI	2D	Brain	-
	Cao et al., 2018 [54]	Cue-aware deep regression network	MRI	3D patch	Brain	-
	Fan et al., 2019 [55]	U-Net+hierarchical dual-supervision	MRI	3D	Brain	-
	Synthetic deformable datasets					
	Rohe et al., 2017 [56]	SVF-Net(U-Net)	MRI	3D	Heart	-
	Eppenhof et al., 2018 [57]	U-Net	CT	3D	Lung	-
	Eppenhof et al., 2019 [58]	Progressive U-Net network	CT	3D	Lung	-
	Sokooti et al., 2017 [59]	RegNet	CT	3D patch	Chest	link
	Synthetic rigid/affine datasets					
	Mohseni et al., 2019 [60]	18-layer residual CNN	MRI	3D	Fetal Brain	-
	Xia et al., 2019 [61]	Cascaded CNN	Plantar pressure image(PPI)	2D	Plantar	-
	Zhao et al., 2015 [62]	10-layer CNN	MRI, CT	2D, 3D	Brain, Lung	-
multimodal	Yang et al., 2017 [63]	Bayesian encoder-decoder network	MRI	3D patch	Brain	-
	Yang et al., 2017 [64]	Encoder-decoder	MRI	3D patch	Brain	link
	Yan et al., 2018 [65]	GAN	MRI, TRUS	3D	Prostate	-
	Sedghi et al., 2019 [66]	3D classification CNN	MRI	3D	Brain	-
	Yao et al, 2019. [67]	CIR	CT, CBCT	3D	Head, Abdomen, Chest, Pelvic	-
	Liao et al., 2019 [68]	POINT	X-ray, CBCT	2D, 3D	Whole Body	-
	Liao et al., 2020 [69]	MSReg	MRI, TRUS	3D	Prostate	-

tration methods; (b) using synthetic datasets with known ground-truth deformation fields (for nonrigid registration); and (c) generating synthetic datasets with rigid/affine transformations (for rigid/affine registration).

Ground-truth generated by traditional methods: In 2016, Yang et al. [52] proposed a supervised encoder-decoder network for large-deformation diffeomorphic metric mapping (LDDMM) registration, which used PyCA[1] LDDMM to generate ground-truth deformations. Their approach was shown to substantially accelerate registration and achieve a lower registration error, compared with traditional methods. Similarly, Cao et al. [53] designed a 3D patch similarity-steered CNN regression network for brain MRI registration, which used Symmetric Normalisation (SyN) and diffeomorphic Demons to generate ground truth deformation fields. Their final registration results obtained a higher DSC than SyN and Demons. They also proposed a key-point truncated-balanced sampling strategy and a cue-aware deep regression net-

[1] https://bitbucket.org/scicompanat/pyca

work to enhance registration generalisation, which tackled various registration tasks in different databases [54]. With ground truth generated by Advanced Normalisation Tools (ANTs) [70] and LCC-Demons [71], Fan et al. [55] proposed a dual-guidance network BIRNET which involved two losses to guide the training process: the distance between generated deformation fields and ground truth, and the dissimilarity between fixed image and warped moving image.

Synthetic deformable datasets: Instead of generating ground-truth deformations using traditional registration methods, Sokooti et al. [59] used artificially generated displacement vector fields (DVF) as ground-truth, and designed a network "RegNet" for chest CT image registration. They proved that the trained model could be applied to real data and obtained registration results on par with a conventional B-spline registration approach. Eppenhof et al. [57] proposed a U-Net-based registration network trained on synthetically deformed clinical images, with augmentation transformations to aid in generalisation. Similarly, they generated a large number of ground truth data by applying random synthetic transformations to a training set of images and proposed a progressive learning network, which enabled training in large and small transformations within the same CNN [58]. Rohe et al. [56] proposed to derive a reference Stationary Velocity Field (SVF) deformation using segmented shapes. Using the obtained reference SVF as the ground truth, they designed a 3D U-Net-based network SVF-Net for cardiac MRI image registration.

Synthetic rigid/affine datasets: The ground truth for rigid/affine registration is much easier to synthesise as random combinations of operations such as rotation, translation, and scaling would be sufficient to generate the data required to train a network. Besides, unlike the non-rigid transformations, most rigid transformation parameters could be obtained manually. Though this task is much easier than non-rigid registration, a few studies have investigated the use of DLIR for rigid registration. For example, Salehi et al. [60] proposed an 18-layer residual CNN regression model for 3D pose estimation and rigidly registered reconstructed foetal brain MRI images to a standard (atlas) space. However, based on images generated by the four transformations (i.e. scaling, horizontal or vertical shift, and rotation), Xia et al. [61] proposed a two-level cascade CNN for plantar pressure image registration. To capture large and complex deformations, Zhao et al. [62] proposed a 10-layer CNN to estimate the rotation parameters (360 classes) and initialise the subsequent registration step. They utilised the

Demons algorithm for nonrigid registration, and achieved substantial improvements in registration accuracy over previous approaches.

Multimodal registration: Supervised DL networks have also been used for multimodal image registration. As in their previous study [52], Yang et al. [63] utilised PyCA to obtain ground truth deformation fields and proposed a 3D Bayesian encoder-decoder network to estimate momentum fields for the registration of multimodal brain MRI images. Furthermore, they developed an approach applicable to both monomodal and multimodal registration called "Quicksilver" [64], which combined a registration and correction network for the LDDMM registration. Using images aligned manually by experts as ground truth, Yan et al. [65] proposed a GAN-based multimodal image registration method called "AIR-Net", which estimated the transformation parameters directly with an efficient forward pass of the generator and additionally evaluated the quality of registration using the discriminator. Unlike general DL methods that predict the displacement field directly, Sedghi et al. [66] used a deep multiclass classifier to predict a collection of discrete displacements between patches. They obtained the final registration results by iterations.

Several approaches have also focused on rigid multimodal image registration, for example - Yao et al. [67] used a regression CNN for coarse rigid registration, which subsequently initialised a conventional intensity-based registration method for fine-grained registration. This approach combined CNNs with conventional methods to align 3D CT and CBCT images. Liao et al. [68] proposed a novel multiview 2D-3D rigid registration method based on learning that directly measured 3D misalignment using a Point-Of-Interest Network for Tracking (POINT), and found the point-to-point correspondence between two images. To tackle the task of rigid MRI-TRUS registration in prostate images, Guo et al. [69] proposed a new strategy to generate augmented datasets, and designed a coarse-to-fine multi-stage network, which significantly reduced the registration error compared to previous methods.

Unsupervised Learning Methods

Although supervised DLIR methods have been shown to substantially accelerate registration and achieve accuracy comparable to traditional methods, the difficulty in obtaining plausible ground-truth transformations is a fundamental challenge and limitation of this group of methods. Methods used to obtain ground-truth transformations typically result in implausible or oversimplified transformations, or are constrained by the performance of the traditional registration methods used to estimate the same. Consequently, in either scenario, the performance of DLIR methods on real data may be limited by the quality of ground-truth transformations available for training. Therefore, researchers have explored unsupervised learning and weakly supervised learning methods to ameliorate the need for ground-truth. Unsupervised registration networks require only the moving and fixed images for training, while weakly supervised approaches (discussed in Subsection 2.2.2) require some additional information such as segmentation masks or landmarks, which are much easier to obtain than ground-truth transformations.

Currently, unsupervised methods are a hot topic in medical image registration, as they can predict the deformation fields and warped moving images in just one forward pass, and do not require ground-truth transformations for training. Similarly to supervised methods, Table 2.2 gives a summary of the most relevant unsupervised medical image registration methods. As before, we first classify all methods as mono- or multimodal. The monomodal methods are further categorised according to the type of regularisation used. Without ground-truth deformation fields, it is difficult for DLIR methods to guarantee diffeomorphic transformations. Therefore, several approaches have been proposed to constrain the estimation of deformation fields and improve their smoothness. To provide an overview of the types of regularisation techniques employed thus far, we group the monomodal unsupervised methods into several subclasses: (1) basic networks, (2) smoothness regulariser, (3) invertibility regulariser, (4) SVF, and (5) cascade networks.

Basic networks: As no ground-truth data are available/used, the first problem to tackle with training unsupervised registration networks is to formulate a loss function that can be optimised to train the network. Using STN, DL networks can generate deformation fields to warp the moving image. The dissimilarity between the warped moving image(s) and fixed image(s) can subsequently be used to calculate the loss function for backpropagation. This measure of dissimilarity (or similarity) is typically estimated using metrics such as Mean Square Error (MSE) and MI, in traditional registration approaches, and can be employed for DLIR methods as well. This group of networks, which we refer to as “CNN+STN”, form the basis for most DL-based image registration networks.

Table 2.2: A summary of unsupervised deep learning-based registration methods. Methods are first classified as monomodal or multimodal. Monomodal approaches are then further classified into several sub-classes.

Registration	Reference	Networks	Modality	Dimension	Organ	Code
Monomodal	**Basic networks**					
	De Vos et al., 2017 [72]	DIRNet	Cine MRI	2D	Heart	link
	Jun et al., 2018 [73]	CNN+STN	MRI	2D patch, 2D	Abdomen	-
	Smoothness regulariser					
	Li et al., 2018 [74]	Multi-resolution FCN	X-ray, MRI	3D	Brain	link
	Balakrishnana et al., 2018[75], 2019[76]	VoxelMorph	MRI	3D	Brain	link
	Fan et al., 2018 [48]	GAN-based registration network	MRI	3D	Brain	-
	Zhu et al., 2020 [77]	Affine subnetwork+Deformable subnetwork	MRI	3D	Brain	-
	Fu et al., 2020 [78]	LungRegNet	CT	3D patch	Lung	-
	Stergios et al., 2018 [79]	CNN+STN	MRI	3D	Lung	link
	Kuang et al., 2019 [80]	FAIM	MRI	3D	Brain	link
	Ali et al., 2019 [81]	Conv2Warp	CT, MRI	3D,4D patch	Lung, Brain	-
	Hu et al., 2019 [82]	Dual-PRNet	MRI	3D	Brain	-
	Bhalodia et al., 2019 [83]	U-Net+Cooperative auto-encoder(CAE)	MRI	2D,3D	Brain, Heart	-
	Sang et al., 2020 [84]	CNN+Convolution auto-encoder	MRI	2D,3D	Heart	-
	Invertibility regulariser					
	Fechter et al., 2020 [85]	U-Net+STN	CT, MRI	3D,4D	Lung, Heart	link
	Mahapatra et al., 2019 [51]	SARNet	X-ray, MRI	2D,3D	Chest, Brain	-
	Gu et al., 2020 [86]	SCC-Net	MRI	3D	Brain	-
	Kim et al., 2019 [87]	Cycle-Consistent CNN	CT	3D	Liver	-
	SVF					
	Dalca et al., 2018[1], 2019[2]	Voxelmorph-diff(Probabilistic Model)+SVF	MRI	3D	Brain	link
	Krebs et al., 2019 [88]	CVAE+SVF	Cine MRI	3D	Heart	-
	Liu et al., 2019 [89]	CNN(Feature-level Probabilistic Model)	MRI	3D	Brain	-
	Shen et al., 2019 [90]	AVSM	MRI	3D	Knee, Femoral, Tibial Cartilage	link
	shen et al., 2019 [91]	3D U-Net+SVF	MRI, CT	2D,3D	Knee, Lung	link
	niethammer et al., 2019 [92]	CNN+vSVF+CNN regulariser	MRI	2D,3D	Brain	link
	Cascade networks					
	De Vos et al., 2019 [93]	DLIR (multi-stage ConvNets)	Cine MRI, CT	3D patch	Heart, Chest	-
	Zhao et al., 2019 [94]	Recursive cascade architecture	CT, MRI	3D	Liver, Brain	link
	Zhao et al., 2019 [95]	Cascading VTN	CT, MRI	3D	Liver, Brain	link
multimodal	Cao et al., 2018 [96]	U-Net+STN	CT, MRI	3D patch	Prostate, Bladder, Rectum	-
	Qin et al., 2019 [50]	UMDIR+cross-cycle reconstruction	CT, MRI	3D	Lung, Brain	-
	Fan et al., 2019 [49]	GAN-based registration network	MRI, CT	3D	Brain, Prostate, Bladder, Rectum	-
	Jiang et al., 2020 [97]	MJ-CNN	CT, CBCT	3D	Lung	-

In 2017, De Vos et al. [72] were the first to propose an unsupervised end-to-end network, based on CNN and STN, to register 2D cardiac cine MRI images. The accuracy of their approach in the registration was demonstrated to be comparable to SimpleElastix[1]. Similarly, Jun et al. [73] proposed a "CNN+STN"network for 2D abdomen MRI registration, which was the first CNN-based registration method for abdominal images.

Smoothness regulariser: Although similarity metrics can guide the training of unsupervised registration networks, previous studies have demonstrated that estimated deformation fields may contain several regions with "folds", where the determinant of

[1] https://simpleelastix.github.io/

the Jacobian (of the deformation field) is negative. The proportion of voxels with negative values for the Jacobian determinant (or the number of folds) is an important criterion used in most DLIR methods to evaluate the smoothness of the predicted deformation fields. Ideally, deformation fields should be diffeomorphic, and hence smooth and invertible. To enforce the estimated deformation fields to be spatially smooth, several researchers [74, 79] have employed various forms of regularisation within the loss function during training. Li et al. [74] employed the total variation (TV) loss as a smoothness regulariser and designed a multi-resolution FCN to estimate dense deformation fields. Instead of the TV loss, Stergios et al. [79] proposed a network similar to "CNN+STN" with L1 regularisation for 3D lung MRI image registration.

Regularisation using L2-norm derivatives of the deformation fields has also been previously proposed [76, 75]. Here, the proposed approach (called "Voxelmorph") was based on a "U-Net+STN" framework with different traditional similarity metrics (MSE and CC) for 3D brain MRI image registration. The approach was shown to outperform several traditional registration methods such as SyN [98] and NiftyReg[1]. Following Voxelmorph, Hu et al. [82] designed a two-stream 3D encoder-decoder network that computed two convolutional feature pyramids separately and included a pyramid registration module to predict multiscale registration fields. Similarly, Ali et al. [81] proposed a novel end-to-end CNN that comprised sequential linear and deformable convolutions along with a learnt nonlinear sampler. With the same smoothness regulariser, Fan et al. [48] proposed an adversarial similarity network (combining a registration network and a discrimination network) for brain MRI registration. They also learnt a meaningful metric for effective training of the registration network, using the discrimination network. Using a similar smooth loss, Zhu et al. [77] designed an end-to-end network comprising an affine alignment subnetwork and deformable subnetwork, which did not require additional preprocessing of affine registration prior to registration. Similarly, Fu et al. [78] proposed a LungRegNet based on two GAN-based networks to register lung CT images from coarse to fine, where the adversarial network in the GANs was used to enforce additional DVF regularisation. Kuang et al. [80] designed a fast image registration network (FAIM), with two explicit anti-folding regularisation terms to force the generated deformation field to be smooth: regularisation for overall smoothness of the predicted displacements and regularisation for negative

[1] https://cmiclab.cs.ucl.ac.uk/mmodat/niftyreg

Jacobian determinants in the transformation.

In addition to adopting a smoothness enforcing loss, Bhalodia et al. [83] proposed to simultaneously learn and use the population-level statistics of the spatial transformations to regularise the neural networks. To do this task, they employed a Cooperative Auto-encoder (CAE) on the predicted deformation fields to urge them to lie in the vicinity of a low-dimensional manifold, and then the reconstruction loss of the CAE was used as a regulariser term. Similarly, Sang [84] pre-trained a convolutional autoencoder on 3,000 DVF samples obtained by SimpleElastix and applied it as a regulariser, which improved the physical and physiological feasibility of DVF.

Invertibility regulariser: Although the aforementioned smooth losses contribute to improving the smoothness of deformation fields, they are unable to guarantee an invertible deformation. Therefore, several studies have focused on designing invertible frameworks and appropriate losses to tackle this problem. Using a cyclic constraint in loss, Fechter et al. [85] presented an approach to calculate DVF for periodic motion tracking in 3D and 4D medical image datasets. This approach was able to calculate the forward and inverse transformation simultaneously. Similarly, using a cycle consistency loss, Mahapatra et al. [51] proposed a GAN-based registration network in combination with segmentation information (learnt automatically), which could directly transfer the registration model trained on one type of images to another type of images (for example, training on lung X-ray images while registering brain MRI on testing). To improve the consistency of the registration, Gu et al. [86] designed a Symmetric Cycle Consistency Network (SCC-Net), which introduced pairwise and groupwise constraints on the consistency of the deformation by losses in inverse consistency and cycle consistency. Some researchers also proposed improving the invertibility by network design. Kim et al. [87] designed a novel registration framework containing two invertible registration networks, where the fixed image and the moving image were deformed/warped to match each other, and subsequently deformed back to the original fixed and moving images.

SVF: Smoothness and invertibility regularisation enhance the diffeomorphic properties of spatial transformations. However, they cannot guarantee the prediction of diffeomorphic transformation fields. In theory, SVF and LDDMM can guarantee diffeomorphism [88]. Therefore, instead of predicting regular dense displacement fields, previous studies have opted to predict SVF to guarantee diffeomorphic transformations. Krebs et al. [88] designed a multiscale Conditional Variational Auto-encoder (CVAE) to estimate stationary velocity fields, which enabled accurate registration of two images and analysis of deformations. Similarly, Dalca et al. [1, 2] proposed a network Voxelmorph-diff, combining diffeomorphic transformations with DL networks, and provided a framework for quantifying registration uncertainty. Following the structure in Voxelmorph-diff to estimate SVF, Liu et al. [89] developed feature-level probabilistic models to estimate the deformation fields for feature maps/images from multiple layers of two convolutional neural networks, which provided direct regularisation for hidden CNN layers. Shen et al. [90] developed an end-to-end registration method Affine-vSVF-Mapping (AVSM), using a multistep Affine-Net to obtain an initial transformation map and a U-Net-like network to generate initial momentum. Subsequently, these two outputs were used as input to the registration component, vSVF, to obtain the final registration fields. Experiments showed that their method achieved higher accuracy and smoother (fewer foldings) fields than Voxelmorph-diff. Based on a vector momentum SVF model, Niethammer et al. [92] were the first to propose a CNN-based local regulariser for registration, generating deformation fields without foldings. The initial momentum could be obtained using various methods, including DLIR methods. For simplicity, we categorised it as an unsupervised DL method. Similar to the method proposed in [90] to obtain the deformation fields, Shen et al. [91] proposed a region-specific diffeomorphic metric mapping registration technique. They obtained large diffeomorphic deformations with a spatiotemporal regulariser and achieved higher accuracy than AVSM [90]. Rather than estimating displacement fields, these methods predict SVF/LDDMM and generate smoother fields than previous methods. The benefit of such approaches is that the estimated deformation fields contain only a few foldings, or in some cases are perfectly smooth.

Cascade networks: Cascade networks combine several registration networks to obtain the final registration results, often obtaining higher accuracy after several rounds of registration. However, these networks do not guarantee diffeomorphic transformations. De Vos et al. [93] proposed a novel registration framework comprising several ConvNets to solve the problem of unsupervised affine and deformable registration. They demonstrated that stacking multiple ConvNets into a more extensive architecture facilitated coarse-to-fine image registration. Zhao et al. [94] presented a deep recursive cascade architecture for deformable image registration, which could be used to cascade other state-of-the-art networks to improve registration quality. In addition, they further designed a registration framework called the Volume Tweening Network (VTN) and incorporated an additional loss of invertibility into the training process [95]. They showed that cascaded registration subnetworks improved performance for registering images with large deformations, with minimal increase in computational cost.

Multimodal registration: Unsupervised registration methods, especially GAN-based methods, are also widely used for multimodal image registration. A common problem in multimodal image registration is to choose/formulate a suitable metric to evaluate the dissimilarity between images from different modalities. Cao et al. [96] designed a "CNN+STN" network for image registration between CT and MRI images. With a prealigned CT and MRI dataset (fixed and moving images are CT-MRI pairs), they proposed an intramodality similarity metric, turning the dissimilarity between MRI and CT images into a combination of two intramodality dissimilarities in MRI and CT. Qin et al. [50] presented a multimodal deformable image registration method (UMDIR), which learnt a bidirectional registration function based on the representation of the disentangled shape. They pre-trained an image-to-image translation network with unpaired data, then used it to train the multimodal registration network and GAN discriminator (to calculate the dissimilarity between images). This method reduced the registration of multimodal images to monomodal images. Fan et al. [49] designed a GAN-based network for multimodal and monomodal image registration between 3D MRI and CT images, designing an adversarial similarity network to learn a meaningful metric for network training. Focussing on pulmonary CT-CBCT and CBCT-CBCT registration, Jiang et al. [97] proposed a multiscale framework called "MJ-CNN" to prevent the registration network from being trapped in a local minimum, which contained three subnetworks at different scale levels (from coarse to fine). They trained these three sub-networks separately first, then jointly trained them in a whole framework.

Compared with traditional registration methods, the unsupervised DLIR methods are significantly faster. Additionally, unsupervised registration networks do not need ground truth transformations for training, addressing a fundamental limitation of supervised image registration methods. Moreover, numerous approaches [1, 2] have shown that unsupervised methods achieve similar or sometimes better registration performance than traditional state-of-the-art registration methods. Consequently, current research in the field is predominantly focused on improving the performance and expanding the capabilities of unsupervised image registration techniques.

Weakly-supervised Learning Methods

As discussed previously, supervised image registration methods require ground-truth deformation fields, which are generally difficult to obtain. In contrast, unsupervised image registration methods disregard all available information and utilise only fixed and moving images. Consequently, useful information that may help guide image registration is not exploited. To utilise such information (typically encoded as anatomical cues) and improve the image registration performance of unsupervised approaches, several weakly supervised learning methods have been proposed. Table 2.3 summarises all deep learning-based weakly supervised registration methods published to date. It is relevant to note that several studies have proposed both unsupervised registration networks and their weakly supervised counterparts simultaneously [76, 2]. As done previously for supervised and unsupervised methods, we categorise this group of approaches into monomodal and multimodal registration, and discuss them accordingly.

Table 2.3: A summary of weakly-supervised DL methods (categorised as monomodal and multimodal registration).

Registration	Reference	Network	Modality	Dimension	Organ	Code
Monomodal	Hering et al., 2019 [99]	CNN	Cine MRI	2D	Heart	-
	Balakrishnana et al., 2019 [76]	Voxelmorph	MRI	3D	Brain	link
	Dalca et al., 2019 [2]	Voxelmorph-diff	MRI	3D	Brain	link
	Heinrich et al., 2019 [100]	PDD-Net	CT	3D	Abdominal	link
	Xu et al., 2019 [101]	DeepAtlas (segmentation and registration CNN)	MRI	3D	Knee, Brain	link
	Chen et al., 2020 [102]	Segmentation+Two-stage registration network	CT	3D	Lung	-
	Ha et al., 2020 [103]	U-Net+Two-stage registration network	MRI	3D	Heart	link
	Mansilla et al., 2020 [104]	AC-RegNet	X-ray	2D	Chest	link
multimodal	Hu et al., 2018 [105, 106]	Global-Net, Local-Net CNN	MRI, TRUS	3D	Prostate Gland	link
	Hering et al., 2019 [107]	U-Net	MRI, CT	3D	Heart	-

Monomodal registration: Most weakly-supervised registration networks are similar to unsupervised networks, with the exception that additional information is utilised during training. This additional information is typically encoded as region-wise labels/masks or landmarks and is only utilised during training. The labels are spatially aligned jointly with the images, by minimising a loss function of the warped moving label and the fixed label. The intuition here is that the labels help preserve anatomical coherence between tissue/organ boundaries by acting as attention maps that guide the estimation of spatial transformations. These label pairs for the fixed and moving images might include solid organs, ducts, vessels, point landmarks, and other ad hoc structures that are deemed relevant to guiding registration. In the reviewed literature, there are mainly two types of labels utilised to guide registration, segmentation masks, and landmarks. Both types of labels are used to construct a combined loss that is optimised to match both labels and images and estimate the desired deformation field. Hering et al. [99] advanced the state-of-the-art in CNN-based deformable registration by combining a square difference loss between fixed segmentation and warped moving segmentation with the similarity between fixed and warped moving images. Following Voxelmorph, Balakrishnan et al. [76] proposed an extension that incorporated a segmentation loss during training, calculated as the Dice score between the fixed and warped moving segmentation masks. Similarly, Dalca et al. [2] also built a weakly supervised version of Voxelmorph-diff by incorporating the surface distance between the segmentation results. With an MSE loss on segmentation, Heinrich et al. [100] designed PDD-Net for the registration of monomodal abdominal CT image, which combined probabilistic dense displacements with differentiable mean field regularisation. This approach was shown to outperform previous DL approaches, achieving an improvement of 15% in Dice overlap.

Instead of using segmentation masks as just additional terms to match in the loss function, Xu et al. [101] proposed the first approach to jointly learn two deep neural networks for simultaneous image registration and segmentation. The registration network and segmentation network can guide each other's training on unlabelled data based on anatomy similarity loss, therefore, the proposed method only required a few manual segmentation samples. With a similar idea, Chen et al. [102] proposed using semantic information (lung lobes and airway masks obtained from a pre-trained segmentation network) to guide registration. They designed a two-stage registration network, where the first predicted coarse deformation in the segmentation masks, while the second was fine registration in the vessel structures. Instead of registering images directly, Ha et al. [103] proposed a semantically guided registration network, which applied a U-Net to the extracted semantic features and used a two-stage registration network to predict the final deformation fields based on the extracted semantic features, under the guidance of two losses in segmentation. As applying the Dice score on the segmentation results does not consider the global context of the anatomical structures, to tackle this issue, Mansilla et al. [104] proposed to use an auto-encoder to extract the global anatomical features from fixed and warped moving masks, then computed the squared Euclidean distance on them as an additional global loss, which helped to predict more realistic and accurate results.

Multimodal registration: Weakly-supervised registration methods have also been employed for multimodal registration. Hu et al. [105] introduced a flexible framework that could use all types of anatomical labels for multimodal T2W-TRUS registration. They proposed a network that combined global net (affine registration) and local net (deformable registration), which significantly outperformed a separate global net or local net. Based on the reviewed literature, this is the first DLIR method to use weak labels to guide image registration. Using segmentation masks for the entire heart in CT and MRI, Hering et al. [107] combined three 2D networks to construct a 2.5D registration approach for cardiac MRI-CT registration. They demonstrated that their approach achieved a higher Dice score than previous state-of-the-art unsupervised registration methods.

2.2.3 Other Usages

Besides predicting similarity metrics and transformation fields, deep neural networks have been used in other ways to facilitate image registration, such as feature extraction, learning new image representations, and reinforcement learning, among others. Table 2.4 summarises these other usages of DL networks for medical image registration. The majority of approaches thus far have employed DL networks to either: (1) learn feature maps for the input moving images and fixed images; or (2) learn new image representations (transfer the original images to new images which are more convenient for registration, for example, learn a clean image from the noisy image, or transfer fixed and moving images to same modality in multimodal registration) for the original fixed images and moving images. We discuss the details of these methods in subsequent sections.

Feature extraction: As DL networks have been proven to be efficient at feature extraction, a few early studies [108, 109] first used DL networks for feature extraction, and subsequently applied traditional registration methods using the obtained features. Wu et al. [108] built a convolutional independent stacked subspace analysis network to learn the hierarchical basis filters from several image patches in brain MRI. They applied HAMMER [123] for registration, achieving better registration perform-

Table 2.4: A summary of other reviewed uses of DL networks for medical image registration, including 3 main classes and several interesting works that are not included in former classes.

Reference	Network	Modality	Dimension	Organ	Usage	Code
Wu et al., 2013 [108]	2-layer ISA	MRI	3D patch	Brain	Feature extraction	-
Wu et al., 2016 [109]	SAE	MRI	3D patch	Brain		-
Kearney et al., 2018 [110]	DCIGN	CBCT, CT	3D patch	Head, Neck		-
Zhu et al., 2018 [111]	PCANet	CT, MRI	2D patch	Brain		-
Blendowsk et al., 2019 [112]	CNN	CT	3D	Lung		-
Zheng et al., 2018 [113]	PDA module	X-ray, DRR	2D,3D	Spine		-
Canalini et al., 2019 [114]	3D U-Net	US	3D	Brain		-
Yang et al., 2016 [115]	Encoder-decoder	MRI	2D	Brain	Image representation	-
Liu et al., 2019 [116]	10-layer FCN	MRI	2D patch	Brain		-
Liu et al., 2019 [117]	IB-cGAN	MV-DRs, KV-DRRs	2D	Head, Neck, Chest, Pelvis		-
Lee et al.,2019 [118]	ISTN	MRI	3D	Brain		link
Tang et al., 2019 [40]	Cycle-GAN	MRI	3D	Brain		-
Blendowski et al., 2019 [119]	Shape encoder-decoder	CT, MRI	3D	Heart		-
Liao et al., 2017 [120]	3D classification CNN	CT, CBCT	3D	Spine, Heart	Reinforcement learning	-
Toth et al., 2018 [121]	CNN	CT, MRI, X-ray	2D,3D	Heart		-
Miao et al., 2018 [122]	FCN+MDP	X-ray, CBCT	2D,3D	Spine		-

ance than other HAMMER-based methods. Based on a similar idea, they also designed a stackable auto-encoder to learn latent feature representations for 3D medical image patches [109]. Kearney et al. [110] proposed a Deep Convolutional Inverse Graphics Network (DCIGN) to extract hierarchical features as input channels to a sparse Deformable Image Registration (DIR) algorithm for registering CBCT to CT images. Blendowski et al. [112] proposed a CNN-based approach for learning discriminative 3D binary descriptors. Focussing on multimodal registration, Zhu et al. [111] designed a novel structural representation method based on PCANet [124] to automatically learn intrinsic image features. Subsequently, the spline-based Free-Form Deformation (FFD) was applied to register the images, obtaining lower Target Registration Error (TRE) than traditional state-of-the-art methods. In addition, Canalini et al. [114] first proposed a segmentation-based registration method, combining a 3D U-Net for segmentation and a traditional registration method, which registered US volumes acquired at different surgical stages. To transfer the model trained in the source domain (i.e. synthetic data) to the target domain (i.e. clinical data), Zheng et al. [113] proposed a pairwise domain adaptation module (PDA) to tackle the domain-shifting problem for CNN-based 2D-3D registration, which learnt domain invariant features using only a few paired real and synthetic data. Experiments showed that they obtained better performance than fine-tuning, using the same pre-trained registration model.

Image representation: Given the fixed and moving images, most previous studies focus on improving the performance of a component in the registration algorithm, and often overlook the quality of the given images. However, even in several well-curated publicly available datasets, low-quality images resulting from tissue, motion, or scanner-related artefacts are prevalent. This adversely affects the accuracy of the final registration, unless addressed adequately. Consequently, given such low-quality images, generating new image representations with prominent distinguishable anatomical features is essential to ensure high registration accuracy. Additionally, in the context of multimodal registration, shifting the domain of the fixed and moving images to a single modality would simplify the registration task. To this end, several studies have proposed using DL networks to learn new representations of the images to be registered. Yang et al. [115] proposed an encoder-decoder network to learn the mapping from pathological images to quasinormal images. Subsequently, they utilised NiftyReg for registration and demonstrated superior registration performance compared with other state-of-the-art approaches. Lee et al. [118] proposed an image-and-spatial transformer network to learn a new image representation for the downstream registration task (using STNs). They showed that their approach outperformed both unsupervised and supervised STNs.

Using DL networks to learn new image representations also attracts much attention in multimodal registration. Liu et al. [116] designed a 10-layer FCN for image synthesis, which learnt a direct image-to-image/patch-to-patch mapping between different modalities and turned multimodal image registration into monomodal registration. With a similar idea, Liu et al. [117] presented a novel modality synthesis approach IB-cGAN to synthesise Kilovoltage Digital Reconstructed Radiographs (KV-DRR) images from Megavoltage Digital Radiographs (MV-DR), and built a multimodal image registration method combining IB-cGAN with a traditional registration approach. Rather than converting images (generally fixed images) from one modality to another, Blendowski et al. [119] built a shared space for images from different modalities. In contrast, Tang et al. [40] designed a multiatlas registration framework, using a Cycle-GAN to synthesise multimodal average atlases.

Reinforcement learning: Reinforcement learning networks are also explored in medical image registration, where the key idea is to provide a reward for every registration action. This class of approaches is mainly employed for rigid registration, mimicking a manual registration process. In 2017, Liao et al. [120] first decomposed the 3D rigid registration task into a sequence of classification problems. They trained the intelligent agent in a greedy supervised fashion and proposed a hierarchical registration framework relying on the trained networks. Subsequent studies also explored a multi-agent system [122] and multimodal registration [121]. Miao et al. [122] formulated 2D-3D registration as a Markov Decision Process (MDP) with observations, actions, and rewards defined according to X-ray imaging systems, and proposed a multi-agent system to solve this challenging problem. Similarly, Toth et al. [121] proposed a novel solution to register 3D preoperative models with 2D intraoperative images. They used a CNN to predict the optimal action with the highest reward and demonstrated clinical feasibility through the robustness and efficiency of their framework.

In summary, DLIR methods have been demonstrated to outperform traditional registration methods in two main aspects, registration speed and accuracy. After training, the registration of DLIR methods (supervised/unsupervised/weakly supervised methods) is just one forward prediction, generally less than 1 second for an image pair. It is significantly faster than traditional methods, because several iterations are necessary for traditional registration methods. Furthermore, most studies have shown that DLIR methods are capable of achieving higher registration accuracy than traditional methods, by utilising large training datasets. The introduction of deep neural networks has significantly improved image registration technologies, from their use for deriving novel representations of transformation models to augmenting the execution of existing traditional image registration methods. In the next section, we further introduce DLIR methods from the point of view of application.

2.3 Applications

In this section, we discuss DLIR methods from a different perspective, analysing them based on their applications. Medical image registration is essential for various clinical applications, such as disease diagnosis and treatment planning, image-guided therapy and surgical interventions, treatment evaluation and patient prognostication, among others. The primary advantage of DLIR methods is their ability to compensate for soft tissue and patient motion in real-time, setting them apart from iterative traditional registration approaches. For instance, Krebs et al. [125] designed an unsuper-

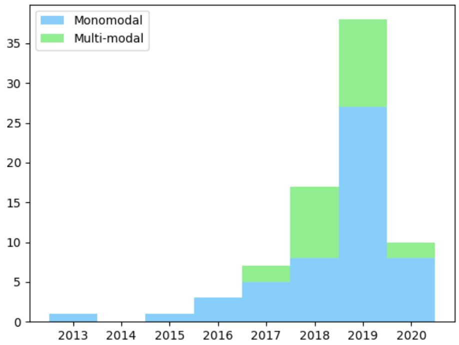

Figure 2.7: The number of papers for monomodal and multimodal image registration methods in recent years.

vised generative deformation model within a temporal convolutional network to learn a probabilistic motion model from a sequence of images, which could be applied for both spatiotemporal registration of cardiac cine MRI and motion analysis. This approach could be used for real-time cardiac motion analysis, providing the basis for the discovery of novel motion-based disease biomarkers. DLIR methods can also be applied to estimate population-averaged atlases from medical images. Dalca et al. [126] described a probabilistic spatial deformation model based on diffeomorphisms, which enabled the generation of atlases conditioned on several attributes of interest, such as age and gender. Such approaches could be employed to generate virtual populations of anatomical structures of interest, which would be useful for conducting in silico clinical trials of medical devices. In addition, they provide a structured framework for assessing anatomical variability across populations, conditioned on relevant covariates. Image registration can also be used to directly facilitate image segmentation. By transforming images from a labelled atlas, Dalca et al. [127] proposed a Bayesian segmentation method for 3D brain MRI based on an unsupervised DLIR framework, removing the need for laborious manual segmentation of numerous images. These studies highlight the versatility in the application of DLIR methods and present several promising directions for future research.

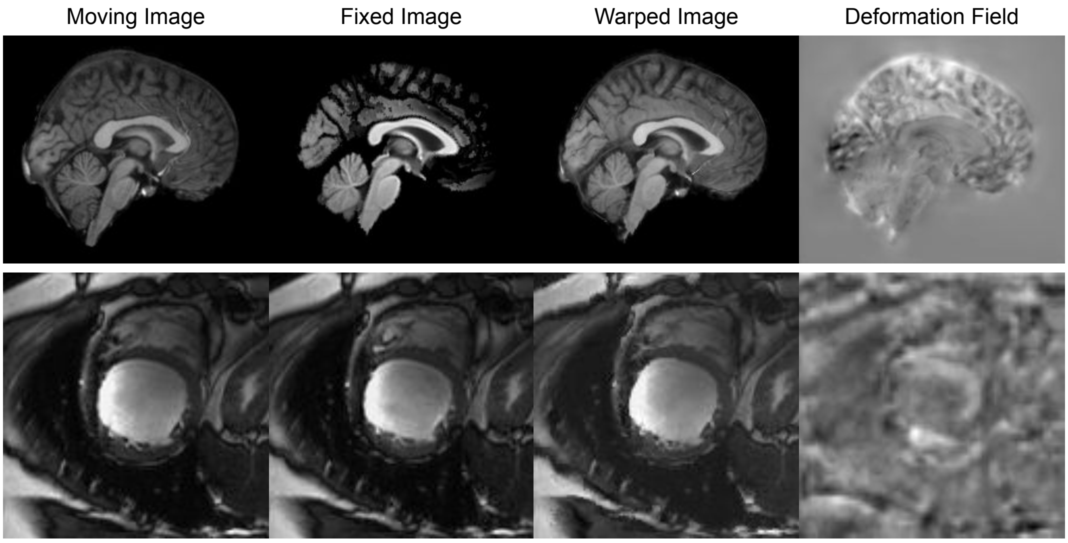

Figure 2.8: An example of brain and cardiac MRI image registration with Voxelmorph-diff [1, 2]. The first and second rows are brain MRI registration and cardiac MRI registration, respectively.

2.3.1 Monomodal Registration

To facilitate and improve future research on DLIR, we summarise all publicly available datasets used to develop registration methods in Table 2.5, with links to each. Figure 2.7 summarises the number of articles published on monomodal and multimodal registration methods in recent years. We observe that most of the studies conducted thus far have focused on monomodal registration, with a substantial increase over the past year. The rate of development of DL-based multimodal registration techniques is relatively slow compared to the above, but the observed trend indicates that it is likely to increase substantially over the next couple of years. In this section, we review monomodal DLIR methods, focussing on the most common image modalities used in the clinic, namely, MRI, CT, US, and X-ray.

MRI registration: MRI is the most widely used modality for developing image registration techniques, with a special focus on brain MRIs, due to the availability of numerous large-scale public datasets (an example of brain and cardiac MRI registration is shown in Figure 2.8). Therefore, a large proportion of recent DLIR methods are

Table 2.5: Overview of the data sets used for medical image registration. We list some basic information (organ, registration type, name, and image modality) of every dataset and the corresponding link and references which exemplify their methods on it. Note that some brain MRI datasets containing various modalities (e.g. T1W, T2W) could also be applied for multimodal registration.

Organ	Registration	Datasets	Modality	Reference
Brain	Monomodal	ADNI [128]	MRI	[108, 109, 53, 75, 1, 54, 74, 126, 2, 94, 51, 95, 86]
		IXI	MRI	[108, 54, 55, 116, 66]
		OASIS [129]	MRI	[115, 52, 64, 75, 1, 76, 126, 127, 2, 83, 77]
		BRATS2015 [130]	MRI	[115, 40]
		LPBA40 [131]	MRI	[109, 64, 53, 54, 74, 48, 55, 49, 40, 89, 94, 95, 81, 82, 92, 86, 77]
		IBIS [132]	MRI	[64, 63].
		IBSR18 [131]	MRI	[64, 48, 55, 49, 92, 86]
		MGH10 [131]	MRI	[64, 48, 55, 49, 81, 92, 86]
		CUMC12 [131]	MRI	[64, 48, 55, 49, 81, 92, 86]
		ABIDE [133]	MRI	[75, 1, 76, 126, 127, 2, 94, 95]
		ADHD200 [134]	MRI	[75, 1, 76, 126, 127, 2, 94, 95]
		MCIC [135]	MRI	[75, 1, 76, 126, 127, 2]
		PPMI [136]	MRI	[75, 1, 76, 126, 127, 2]
		HABS [137]	MRI	[75, 1, 76, 126, 127, 2]
		Harvard GSP [138]	MRI	[75, 1, 76, 126, 127, 2]
		FreeSurfer Buckner40 [139]	MRI	[76]
		Mindboggle101 [140]	MRI	[80, 89, 101, 82, 77]
		BraTS2017 [141]	MRI	[50]
		BrainWeb [142]	Simulated MRI	[62, 111, 116]
	Multimodal	RIRE	CT, MRI	[111]
		BITE [143]	US, MRI	[114]
		RESECT [144]	US, MRI	[114, 145, 146]
Heart	Monomodal	Sunnybrook [147]	Cine MRI	[72, 93, 85, 84]
		ACDC [148]	Cine MRI	[88, 125, 99, 103]
	Multimodal	MM-WHS [12]	CT, MRI	[107, 119]
Knee	Multimodal	OAI	MRI, X-ray	[90, 101, 91]
Liver	Monomodal	MICCAI 2007 Grand Challenge [149]	CT	[95]
		MSD	CT	[94]
		SLIVER [150]	CT	[94]
		LiTS	CT	[94, 95]
Chest	Monomodal	COPDGen [151]	CT	[50]
		NLST [152]	CT, X-ray	[93]
		DIR-Lab-COPDgen [153]	CT	[112]
		DIR-Lab-4DCT [154]	CT	[93, 85, 81, 57, 58, 97, 78]
		SPARE [155]	CT, CBCT	[97]
		POPI [156]	CT	[85, 81, 57, 58]
		LIDC-IDRI [157]	CT	[121, 58]
		Empire 10 lung datasets	CT	[62]
		NIH ChestXray14 dataset [158]	X-ray	[51]
		JSRT [159]	X-ray	[104]
		Montgomery County X-ray database [160]	X-ray	[104]
		Shenzhen Hospital X-ray database [160]	X-ray	[104]
Several Organs	Multimodal	UK Biobank Imaging Study	MRI	[118]
Whole Body	Multimodal	VISCERAL Anatomy3 [161]	CT, MRI	[100]

validated in brain MRIs, in order to compare performance with previous state-of-the-art methods, such as Voxelmorph [76, 75], VTN [95], and Conv2warp [81]. Several brain MRI datasets are also used to develop multimodal image registration methods [64, 40], with T1W and T2W modalities available in most brain MRI datasets. In addition to neuroimaging, cine MRI is the primary modality used for cardiac image registration and cardiac motion estimation [88, 125], with two available public datasets, Sunnybrook Cardiac Data (SCD) [147] and Automatic Cardiac Diagnosis Challenge (ACDC) [148].

CT registration: CT images are widely used to scan organs in the chest (lungs, heart) and abdomen (liver, kidneys, and pancreas). Specifically, as shown in Table 2.5, there are four liver CT image data sets (MICCAI 2007 Grand Challenge [149], MSD, SLIVER [150], LiTS) and eight thoracic CT data sets (LIDC-IDRI [157], POPI [156], Empire 10 lung datasets, COPDGen [151], NLST [152], DIR-Lab-COPDgen [153], DIR-Lab-4DCT [154]). In addition, there are also several multimodal datasets containing CT images, VISCERAL Anatomy3 [161], MM-WHS [12] and RIRE respectively. We found that CT image registration is the second largest domain used to develop medical image registration methods, with numerous recent studies on the topic [67, 93, 85, 81, 57, 58]. Compared with brain MRI registration, CT image registration is more challenging to some extent, due to limited soft-tissue contrast, and greater variability in image quality.

Ultrasound registration and X-ray registration: In contrast to the modalities discussed so far, there are few publicly available datasets for US and X-ray images. Correspondingly, the number of papers focussing on the registration of US and X-ray images is also limited. There are two brain datasets, RESECT and BITE, containing US images, and only one article focussing on monomodal US image registration [114] using publicly available datasets. Regarding X-ray images, there are six publicly available datasets, NLST [152], NIH ChestXray14 [158], OAI, JSRT [159], Montgomery County X-ray database [160] and Shenzhen Hospital X-ray database [160]. However, there are relatively few studies on X-ray image registration [51, 104], compared to MRI and CT.

2.3.2 Multimodal Registration

With the ability to calculate the dissimilarity between images of different modalities, DL has been widely applied in multimodal registration. However, in contrast to monomodal registration, there is limited availability of public datasets for multimodal registration. Based on the reviewed literature, we found only three publicly available multimodal data sets for developing registration approaches, namely, RIRE, VISCERAL Anatomy3 benchmark [161] and multimodal Whole Heart Segmentation dataset (MM-WHS) [12] respectively. Although there are numerous studies that focus on multimodal registration, most of them collect and use independent, private datasets to develop and validate their algorithms. In this section, we discuss several typical multimodal registration applications, for example, T1W-T2W registration, CT-MRI registration, CT-CBCT registration, and 2D-3D registration.

T1W-T2W registration: T1W-T2W registration aims to learn a mapping between T1-weighted MRI images and T2-weighted MRI images. It is a common multimodal registration task in neuroimaging, with many publicly available brain MRI datasets. Yang et al. [63] proposed a 3D Bayesian encoder-decoder network for multimodal registration T1W-T2W based on the IBIS 3D autism brain image dataset. Qin et al. [50] proposed a GAN-based UMDIR network for this task based on the BraTS2017 dataset. Liu et al. [116] tested their methods on several multimodal registration tasks, T2W vs proton density (PD), T1W vs PD, and T1W vs T2W, respectively. Tang et al. [40] used a Cycle-GAN to synthesise multimodal atlases (T1W, T1 contrast-enhanced, T2W, FLAIR), building a bridge between different modalities.

CT-MRI registration: CT-MRI matching is another common multimodal registration application. The three public multimodal registration datasets we mentioned previously all contain both CT and MRI images for the same subjects, useful for developing multimodal registration approaches. Zhu et al. [111] proposed a PCANet to learn structural representations of FFD in the RIRE data set. Using a private dataset, Cao et al. [96] proposed a “CNN+STN” network for registering CT and MRI images. In addition to these, GAN-based networks have also been used for pelvic [49], and other studies have proposed approaches to register cardiac CT and MRI images based on the MM-WHS dataset [107].

MRI-TRUS registration: Several papers have also explored the registration of MRI and TRUS images. From our reviewed research, two datasets RESECT [144] and BITE [143] are publicly available for this registration task, and several methods were developed based on them [145, 146]. However, most of these studies are based on private datasets. Guo et al. [69] proposed a supervised network to tackle rigid MRI-TRUS registration on prostate images. Hu et al. [106, 105] proposed a global subnetwork, for affine registration, with a local subnetwork for deformable registration of T2W MRI and TRUS images. Yan et al. [65] designed a GAN-based adversarial image registration network (AIR-Net) to address this task. Haskins et al. [43] used CNN to calculate the similarity between the MRI and TRUS images.

CT-CBCT registration: Recently, image registration between CBCT and CT images has also drawn some attention [110, 67, 97]. Focussing on CT-CBCT deformable registration on head and neck images, Kearney et al. [110] proposed DCIGN to learn hierarchical characteristics, which outperformed intensity-corrected Demons and landmark-guided DIR. To achieve CT-CBCT rigid registration in image-guided radiotherapy (IGRT), Yao [67] proposed a CNN to predict an initial rough transformation, then used traditional intensity-based registration to refine the registration. This shortened the prediction time while ensuring high registration accuracy.

2D-3D registration: In most multimodal registration applications discussed thus far, the dimensions of the fixed and moving images are identical. Publicly available datasets provide 3D image volumes, which can also be employed for slice-wise 2D-2D registration. Therefore, studies to date have focused primarily on 2D-2D and 3D-3D image registration. In addition to these, 2D-3D image registration is also useful for a variety of clinical applications and forms a major part of ongoing research in DL-based multimodal image registration. This task is even more challenging, due to the difference in dimensionality and the issue of overlapping tissues and contrast common to 2D images such as X-rays. Studies on 2D-3D registration have focused mainly on registering X-ray images with other 3D modality images, such as MRI/US [121], CT [113], and CBCT [122, 68]. Additionally, slice-to-volume registration has also received some attention in recent years [60].

2.4 Discussion

Previous sections have introduced and discussed the most relevant DLIR published to date. In this section, we present current trends in the development of DLIR methods and discuss the main challenges that are yet to be addressed. Finally, a summary of the possible directions for future research in the field is presented.

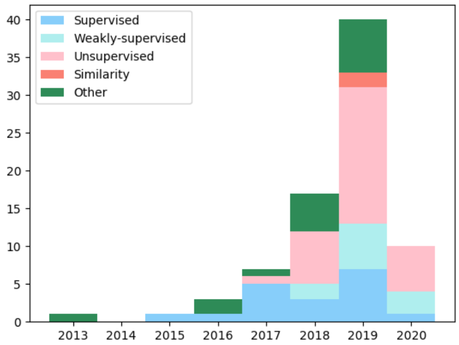

Figure 2.9: Histogram depicts the number of DLIR papers published until 2020, grouped according to the categories defined in Section 2.2. “Similarity” refers to the category of deep similarity.

2.4.1 Development Trends

As discussed previously, recent years have seen a dramatic increase in the number of papers published on DLIR methods. Unsurprisingly, this follows wider trends in the use of DL for various tasks in medical image analysis and computer vision. The development of DL experienced a boom after 2015, with the release of several open-source deep learning software libraries (e.g. Tensorflow, Keras, and Pytorch). This provided a convenient and easy-to-use environment for rapid prototyping of DL networks. We found that the development of DLIR began in 2015. The first two methods proposed in 2013 and 2015 applied CNNs for feature extraction. DLIR methods with high impact in this domain were first proposed in 2016, where DL networks were used to predict deformation fields. Subsequent years have seen a continuous increase in the number of DLIR papers, with several significant and innovative contributions making a strong case for their superiority over traditional, iterative registration approaches.

Although it has only been a few years since DL networks were applied to medical image registration, the use of DL for medical image registration has seen several changes. The evolution in the development of DLIR methods is described by the histogram plot shown in Figure 2.9. We characterise this evolution over four stages. The first stage attempted to use deep neural networks for feature extraction, which in turn were used to guide traditional registration algorithms, by providing more discriminative information than the original images. The studies then focused on addressing a crucial limitation of iterative traditional registration approaches, viz. long execution times. By learning the space of desired spatial transformations, given suitable training data, the aim of several supervised networks proposed in this stage was primarily to speed up registration during inference. Models trained in this fashion on suitable image pairs are many times faster than iterative registration approaches during testing/inference. However, supervised methods require ground-truth spatial transformations to be available for training samples, which are difficult to obtain in most real-world applications, thereby limiting their applicability.

To circumvent the need for ground-truth deformation fields, at the third stage, unsupervised and weakly supervised methods were proposed. These approaches demonstrated comparable registration accuracy and speed with supervised methods, while requiring just weak labels or no labels at all. Specifically, weakly supervised registration methods were proposed a little later than unsupervised methods. At this stage, there was no noticeable improvement in accuracy. In contrast, the deformation fields generated by DL networks were sometimes non-smooth and unrealistic. The final stage was aimed at improving the accuracy of the registration and making the deformation fields smoother. Several additional types of information (e.g. segmentation masks) were incorporated into networks using weakly supervised learning frameworks, and various forms of regularisation were introduced during training. These four stages are not strictly separated. However, we could see a clear line in the development of DLIR methods, as evidenced by the graph shown in Figure 2.9.

We note that the dimensionality of images used to train DLIR networks is gradually tending towards the natural space of deformations or organs of interest, as powerful computing hardware becomes available to handle the high computational and memory requirements. Initially, the input data used to train the DL registration networks were mainly 2D images [115, 52, 72, 111, 117, 125, 61, 99] or 2D image patches [116, 73, 47]. They gradually became 3D image patches [108, 52, 109, 64, 53, 63, 110, 96, 55, 93], and finally whole 3D image volumes and 4D images/patches [85, 81]. In fact, it is natural to perform 3D registration for most medical images, as most organ motions take place in 3D. For most medical image registration applications, 3D is enough for registration tasks. However, for some special applications such as cardiac motion estimation, researchers are exploring 3D+t or 4D image registration techniques, which are less common in other computer vision applications.

2.4.2 Main Challenges

Though DLIR methods have addressed many challenging problems in medical image registration and have achieved faster and more accurate registration than traditional methods, there are several challenges that must be addressed in this domain.

Preprocessing: Preprocessing is an integral part of image registration, which generally consists of several operations geared towards simplifying the data to be registered. Different preprocessing steps may lead to different registration results, even using the same datasets. In other computer vision tasks, such as image classification and image segmentation, researchers demonstrate their methods on public datasets, where the preprocessing is easy to realise and shared by all researchers. However, in medical image registration, although there are many publicly available datasets, the preprocessing steps tend to vary between studies. For example, in brain MRI image registration, there are many publicly available datasets, such as OASIS [129], ADNI [128], IXI and MGH10 [131]. Furthermore, there are several well-acknowledged preprocessing steps, such as skull-stripping, affine registration, spatial resampling, image enhancement, intensity normalisation, and cropping. However, studies often use different datasets for training and testing and employ different preprocessing procedures with adapted parameters for each step (e.g. voxel size, smoothing factor, etc.). Therefore, in some earlier DLIR studies, specifically, before Voxelmorph, methods were usually only compared with traditional state-of-the-art registration approaches (e.g. ANTs [70], Elastix [162], Demons [163, 164]).

Clinical applications: Clinical applications are the final destination for all medical image processing and analysis methods. Until now, numerous DL-based image registration methods have proved their efficiency and superiority compared to classical methods. However, we are yet to see a DL-based tool deployed in a clinical setting, such as ANTs and Elastix in classical methods. It is challenging for clinicians and clinical researchers to use DL networks in clinical applications without the right tools. Furthermore, since DL networks are challenging to interpret, even though a trained model shows high accuracy in the test datasets, clinicians are still wary of using them regularly to analyse patient data. Several studies have attempted to quantify the uncertainty of some predicted registration results of clinicians with useful information to provide the validity of the registration [52, 63, 115, 1, 2]. However, more research and a systematic assessment of registration uncertainty are required to build community trust and accelerate the adoption of DLIR methods in clinical settings.

Limited data: The lack of suitable public data sets is another fundamental problem limiting the development of DLIR methods. To obtain accurate and robust models, DL networks must be trained on large-scale datasets. Although unsupervised learning registration methods do not require ground-truth data, currently the primary publicly available datasets are focused solely on brain MRI images, with just a few datasets containing other organs/modality images. Besides, for supervised methods and weakly supervised methods, sourcing high-quality ground-truth data remains a challenge. We also observe that several studies only exemplify their method on their private datasets due to a lack of publicly available datasets, which is not convenient for benchmarking and comparing state-of-the-art methods. With the increase in datasets, a more fair comparison will be possible, facilitating greater innovation in DLIR.

2.4.3 Possible Directions

In this section, we outline possible directions for future research in DLIR to address the challenges discussed so far. The first step towards identifying these is to consider the aims of DLIR. Accuracy, robustness, and speed are common goals for all registration methods. DLIR methods trained to predict the spatial transformation matching a pair or group of images have not shown a significant difference in registration speed. Therefore, the obvious focus of future approaches on DLIR should be on improving the accuracy and generalisation capability of the networks and ensuring that the estimated deformation fields are more realistic and smooth.

Combining the superiority of traditional methods with DL: A possible direction is to combine the advantages of traditional methods with deep learning networks. Although DLIR methods have significantly improved registration speed and accuracy compared with classical methods, the superiority of classical methods (e.g. diffeomorphic attributes and robust registration) can not be overlooked. The trend to make deformation fields smoother is to combine the diffeomorphic transformation in traditional methods with DL networks.

Boosting performance with priors: As discussed previously, medical image registration differs greatly from other medical image analysis tasks. Future research should introduce more registration priors to DL networks, making DL networks more specific to image registration, and more application-specific. To improve registration performance, DLIR networks could be imbued with prior information related to the expected type of deformation, the spatial relationship between anatomical structures, and the topology and morphology of anatomical structures. For example, although ground-truth spatial transformations are seldom available, other labels could serve as ground-truth to guide the training process. Several methods on weakly supervised image registration have been proposed, which generally achieve better performance than its corresponding unsupervised variant (at no additional cost in terms of execution speed). More informative priors combined with synthetically modified training data, such as blackening pixels in the moving image, or generating adversarial examples [165], could enhance the ability of networks to generalise to unseen data, while remaining robust to variable image quality. Consequently, combining different types of spatial and temporal priors with DL networks is a promising direction for future research in the field.

2.5 Conclusion

In this chapter, we comprehensively summarised the evolution of deep learning-based medical image registration. We discussed the existing challenges and potential directions for future research and presented a thorough summary of publicly available datasets and links to the code of published papers, to facilitate benchmarking of algorithms and enhance future research. The development of DL-based image registration methods has experienced a similar trend to the development of DL. Image registration networks increasingly operate in the natural space of the organs or deformations of interest, i.e., gradually evolving from processing 2D images to 3D/4D (dynamic) volumes. Recent contributions range from speeding up registration in higher dimensions to reducing the need for ground-truth during training, or advanced regularisation constraints to retrieve plausible deformation fields and preserve anatomical topology. Due to the difficulty in obtaining ground-truth data for training, DLIR networks gradually turned to unsupervised learning from supervised learning.

The lack of available data is a major impediment to the advancement of DLIR techniques. Furthermore, the various preprocessing steps used in different studies make it hard to compare the most recent approaches and conduct thorough benchmarking studies. Although DLIR networks have made considerable progress in terms of registration speed and accuracy for most tasks, some tasks still have accuracy levels that are only comparable to traditional methods. Additionally, there is a lack of studies that demonstrate the clinical applicability of DLIR methods, similar to what has been done for several traditional registration tools (e.g. ANTs, Demons). We have yet to observe this trend in DLIR methods, but we anticipate that this will be the next area of research in the field. Accuracy, generalisation, realistic and smooth deformation will likely remain the main research focus for medical image registration in the near future. Alongside an increased availability of multimodal datasets, we expect an increased focus on multimodal registration using DL approaches.

Image registration is a fundamental task to understand and capture cardiac motion. Deep learning-based registration networks have achieved comparable or better registration performance than traditional approaches in most scenarios, while few studies have considered the priors of cardiac motion in DL-based image registration networks. In the following chapter, we would introduce two specific DL-based image registration networks, incorporating cardiac motion priors in the network to achieve more accurate and realistic cardiac image registration.

Chapter 3

Joint Segmentation and Discontinuity-preserving Image Registration

Segmentation and image registration are two fundamental tasks in medical image analysis. The former is to find the region of interest (in cardiac image segmentation, LV blood pool, LVM and RV) in the original image, while the latter aims to find the point correspondence between the moving and the fixed images (pairwise registration) and predict deformation fields to deform the moving image to the fixed image. These two tasks are relevant and beneficial to each other. Previous research has explored the deformation of a template to achieve anatomy-plausible segmentation and the incorporation of segmentation masks in image registration to improve registration performance. Several approaches have also shown that considering segmentation and registration simultaneously can achieve better performance than achieving them separately.

Recently, deep learning-based methods have been widely applied in medical image registration, achieving much faster and comparable results compared with traditional methods. However, most of the deep learning-based registration methods assume that the deformation fields are smooth and continuous everywhere, which is not always true, especially in medical image registration (e.g. cardiac and abdominal images). Due to the different motion patterns (e.g. sliding) and different properties of different organs/tissues, discontinuity may occur on the boundary between different regions. Consequently, assuming totally smooth deformation would lead to unrealistic deformation and sub-optimal registration performance. However, this issue is ignored by most of the current deep learning-based registration methods.

To tackle this issue, in this chapter, we first introduce a deep discontinuity-preserving registration method, named DDIR, which can preserve the discontinuity in deformation fields and achieve more accurate registration performance. As ground-truth segmentation is required in both training and inference of DDIR, we further propose to achieve segmentation and discontinuity-preserving registration in a single network, which only requires moving and fixed images as inputs and predicts accurate segmentation masks and discontinuity-preserving registration results simultaneously. With the predicted segmentation masks and deformation fields, clinical indices and some motion biomarkers can be computed for subsequent CVD prediction and diagnosis tasks.

3.1 Deep Discontinuity-preserving Image Registration Network

Image registration aims to establish spatial correspondence across pairs or groups of images, and is a cornerstone of medical image computing and computer-assisted interventions. Currently, most deep learning-based registration methods assume that the desired deformation fields are globally smooth and continuous, which is not always valid for real-world scenarios, especially in medical image registration (e.g. cardiac imaging and abdominal imaging). Such a global constraint can lead to artefacts and increased errors at discontinuous tissue interfaces. To tackle this issue, we propose a weakly-supervised Deep Discontinuity-preserving Image Registration network (DDIR), to obtain better registration performance and realistic deformation fields. We demonstrate that our method achieves significant improvements in registration accuracy and predicts more realistic deformations, in registration experiments on cardiac magnetic resonance (MR) images from UK Biobank Imaging Study (UKBB), than state-of-the-art approaches.

3.1.1 Introduction

Image registration is a fundamental component of several applications in medical imaging. Recent years have seen a shift from traditional iterative methods to deep learning (DL)-based registration approaches. Although training DL-based approaches is time-consuming, the inference is rapid, involving just a single forward pass through the network. Consequently, DL-based approaches offer substantial acceleration for pair-/group-wise image registration relative to traditional approaches, achieving near-real-time performance in certain applications.

Most existing DL-based registration methods constrain deformation fields to be globally smooth and continuous, through various means [76, 127, 88]. However, this assumption is often violated in medical image registration applications, as tissue boundaries are naturally discontinuous. This is especially pronounced in cardiac or abdominal imaging, which involves large deformations of multiple tissue-types, and organ motion/sliding at tissue boundaries. Variability in the physical properties of different tissue-types results in discontinuities at native tissue boundaries [166, 167]. Hence, enforcing deformation fields to be globally smooth can generate unrealistic deformations and lead to increased errors near these boundaries.

Discontinuity-preserving image registration is an active area of research in the context of traditional registration methods [168, 169, 170, 166]. For example, Hua et al. [166] proposed a discontinuous registration approach that utilised enriched B-spline basis functions at control points near discontinuous tissue boundaries, achieving significant improvement in registration accuracy, relative to other existing discontinuity-preserving registration methods. In contrast, only one study thus far has proposed a discontinuous DL-based image registration framework. Ng et al. [171] proposed a custom discontinuity-preserving regulariser on the deformation fields (used with a typical unsupervised registration network), to preserve discontinuities, while ensuring local smoothness within specific regions. They formulated a regularisation term based on the unsigned area of the parallelogram spanned by two displacement vectors associated with moving image voxels. However, without additional boundary information for guidance, such a discontinuity regularisation term alone is insufficient to preserve strong discontinuities in deformation fields.

This work assumes that the desired deformation fields are locally smooth, but discontinuities may exist between different regions/organs at tissue interfaces. Therefore, we generate distinct smooth deformation fields for different regions of interest and add them to obtain the final registration field, used to warp the moving image. Such a locally-smooth and globally-discontinuous registration scheme is achieved using a novel Deep Discontinuity-preserving Image Registration network, or DDIR. The contributions of this work are two-fold: (1) we designed a novel framework, DDIR, for discontinuous DL-based image registration. This is the first study to incorporate discontinuity in DL network structure and training strategy, and not only in terms of a custom regularisation term in the loss function. (2) Our proposed DDIR achieves significant improvement in registration accuracy over state-of-the-art registration methods, and preserves key cardiac morphological indices post-registration, not afforded by the latter.

3.1.2 Method

Pair-wise image registration aims to establish spatial correspondence between the moving image $\mathbf{I}_M$ and fixed image $\mathbf{I}_F$ and is formulated as,

$$\phi(\mathbf{x}) = \mathbf{x} + u(\mathbf{x}), \tag{3.1}$$

where, $\mathbf{x}$ represents voxels/pixels in the moving image $\mathbf{I}_M$, $u(\mathbf{x})$ denotes the displacement field, and $\phi(\circ)$ represents the deformation function.

To generate deformation fields that are locally smooth and discontinuous at the boundaries of different organs/regions, we propose to generate deformation fields for different sub-regions, and add them to obtain the final deformation field. Sub-regions in the images to be registered must first be segmented either manually or automatically. With short-axis (SAX) cardiac cine-magnetic resonance (CMR) images, manual and automatic segmentation results for left ventricle blood pool (LVBP), left ventricle myocardium (LVM) and right ventricle (RV) are generally available in public data sets, large-scale imaging initiatives (e.g. UK Biobank) and from previous studies on automatic CMR segmentation [172]. As the focus of this work is on SAX-CMR image registration, we explicitly model discontinuities along cardiac boundaries by splitting the images into four sub-regions, namely, LVBP, LVM, RV, and background. These sub-regions are subsequently used to train our DDIR approach and register CMR images in a manner that preserves discontinuities at their boundaries.

Network Architecture

Multi-channel Encoder-decoder. Most previous DL-based registration methods apply an encoder-decoder network (generally U-Net [35]) to extract feature maps from the concatenated input moving image and fixed image. However, as shown in Figure 3.1, in DDIR the original moving image and fixed image (at 128 × 128 × 32) are divided into four image pairs, i.e. LVBP, LVM, RV and background, using segmentation masks for the corresponding regions. In each of these pairs, voxels in corresponding regions are preserved while the rest are set at zero. Each pair is concatenated and fed as input to a distinct U-Net block, which extracts region-specific feature maps. These four U-Nets have the same architecture, including four down-sampling layers and three corresponding up-sampling layers. Using this multi-channel encoder-decoder structure, we obtain four sets of feature maps ($64 \times 64 \times 16$) corresponding to different sub-regions. We use the same U-Net architecture (with identical hyper-parameters) in all DL-based registration approaches investigated in this study.

Discontinuity Addition. Using the region-specific feature maps learnt by U-Nets, we first predict four different smooth deformation fields (corresponding to each region) and then add them to obtain the final deformation field, to preserve local smoothness

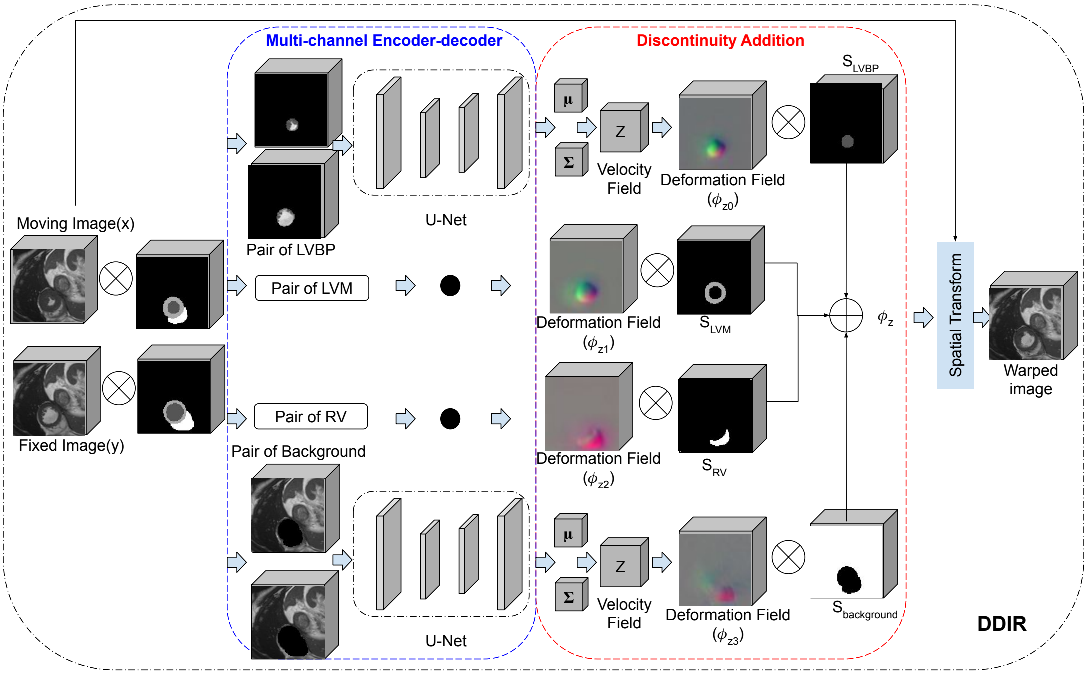

Figure 3.1: Schema of DDIR. The registration network applies four different channels extracting features from pairs of LVBP, LVM, RV and background. Based on them, we obtain four sub-deformation fields for different regions. The final deformation field is obtained by adding these four deformation fields with corresponding segmentation. The cardiac MR images were reproduced by kind permission of UK Biobank ©.

and discontinuity at the interfaces. Similar to previous papers [127, 88], we assume the transformation function (denoted as ϕ_z) is parametrised by stationary velocity fields (SVF) ($z_i, i \in [0, 3]$), which are sampled from a multivariate Gaussian distribution. With the predicted feature map, we compute the mean μ_i and variance Σ_i of z_i (using two different convolution layers). Based on them, four SVFs (z_0, z_1, z_2, z_3) corresponding to different regions (LVBP, LVM, RV and background) are sampled. With the corresponding integration layer and up-sampling layer, we obtain four diffeomorphic deformation fields ϕ_{z_0}, ϕ_{z_1}, ϕ_{z_2} and ϕ_{z_3}. As before, we use region-specific segmentation masks to extract each region of interest from the obtained deformation fields (setting the remaining voxels to zero) and add them to generate the final deformation field. Denoting the segmented regions of LVBP, LVM, RV and background as S_{LVBP}, S_{LVM},

S_{RV} and $S_{background}$ respectively, the addition can be formulated as,

$$\phi_z = \phi_{z_0} \times S_{LVBP} + \phi_{z_1} \times S_{LVM} + \phi_{z_2} \times S_{RV} + \phi_{z_3} \times S_{background}. \tag{3.2}$$

Loss Function

The loss function includes two terms, a dissimilarity and a regularisation term. The former is the distance between the warped moving image and the fixed image, while, the latter constrains the estimated deformation fields to be locally smooth (i.e. within each region), to avoid unrealistic deformations. The dissimilarity loss in DDIR captures the dissimilarity on both images and segmentations. We use normalised cross-correlation (NCC) L_{NCC} to evaluate the similarity between the warped moving image and the fixed image. As the region-wise segmentation masks are available, we also compute the region-wise dice loss, denoted L_{Dice} as in [173].

To preserve discontinuity at the interfaces of the organs/regions while ensuring local smoothness, a global smoothness constraint is not enforced on the final deformation field. The addition of different deformation fields preserves discontinuities at interfaces, therefore, we only need to guarantee the deformation field of each sub-region is smooth. This is achieved by regularising each sub-deformation field. Following Voxelmorph-diff [127], we calculate the Kullback-Leibler (KL) divergence between the approximate posterior $q_\psi(z|\mathbf{I}_F;\mathbf{I}_M)$ and the prior $p(z)$ ($p(z) = \mathcal{N}(z;\mu_z,\Sigma_z)$) of each velocity field z, formulated as,

$$\begin{aligned} R &= KL(q_\psi(z|\mathbf{I}_F;\mathbf{I}_M)||p(z|\mathbf{I}_F;\mathbf{I}_M)), \\ L_R &= \frac{1}{4}(R_{LVBP} + R_{LVM} + R_{RV} + R_{background}), \end{aligned} \tag{3.3}$$

where R denotes the regularisation for each deformation field and L_R is the combined regularisation term. The $q_\psi(z|\mathbf{I}_F;\mathbf{I}_M) = N(z;\mu_{\mathbf{z}|\mathbf{I}_F,\mathbf{I}_M},\Sigma_{\mathbf{z}|\mathbf{I}_F,\mathbf{I}_M})$ is a multivariate normal, where, $\mu_{\mathbf{z}|\mathbf{I}_F,\mathbf{I}_M}$ and $\Sigma_{\mathbf{z}|\mathbf{I}_F,\mathbf{I}_M}$ are the mean and variance of the distribution, learnt by convolution layers. The complete loss function used to train the network is, $L_{total} = \lambda_0 \times L_{NCC} + \lambda_1 \times L_{Dice} + \lambda_2 \times L_R$, where, λ_0, λ_1 and λ_2 are used to weight the importance of each loss term.

3.1.3 Experiments and Results

Data and Implementation

The registration performance of the proposed approach is evaluated on SAX-CMR images (spatial resolution at $\sim 1.8 \times 1.8 \times 10mm^3$), available from UKBB. We chose images from 2,000 subjects at random, and used images at end-diastole (ED) and end-systole (ES) for intra-subject registration. Among these, 1,600 subjects were chosen at random for training DDIR, equating to 3,200 image pairs (ED-to-ES or ES-to-ED registration). Image pairs from the remaining 400 subjects were used for testing. All CMR images were resampled to $1.50 \times 1.50 \times 3.15mm^3$ using bi-cubic interpolation, and cropped to a size of $128 \times 128 \times 32$ (with zero-padding for images with fewer than 32 slices). The region-wise segmentation masks for all CMR images were obtained automatically using the segmentation method proposed in [172]. DDIR was implemented using Python and Keras on a Tesla M60 GPU machine. The Adam optimiser was used for training, with a learning rate of $1e^{-4}$. The batch size was set to 2, and the hyper-parameters λ_0, λ_1 and λ_2 were set to $20, 200, 0.1$ (determined empirically), respectively. The source code will be publicly available on the Github [1].

Quantitative Comparison and Analysis

To demonstrate the superiority of our approach, we compare DDIR with both traditional registration and DL-based registration methods. For the former, we choose Symmetric Normalisation (SyN) registration (3 resolution level, with 100 iterations in each sampling level) in ANTs [70], Demons (Fast Symmetric Forces Demons [174] with 800 iterations and standard deviations 1.0) in SimpleITK and B-spline registration (max iteration step is 2,000, sampling 6,000 random points per iteration) in SimpleElastix [175], for comparison. For the latter, DDIR is compared with Voxelmorph-diff [127]. As DDIR uses segmentation masks during training and inference, it is a weakly-supervised registration method. For a fair comparison, we build three weakly-supervised versions of Voxelmorph - VM-Dice, VM(img+seg) and VM-Dice(img+seg). VM-Dice uses a Dice loss L_{Dice} term and binary cardiac segmentation masks for the fixed and moving images during training, but does not require the latter for inference. In VM(img+seg), we concatenate the fixed and moving images with their correspond-

[1] https://github.com/cistib/DDIR

ing multi-class masks (i.e. distinct labels for each region) and use these to train the network. While, VM-Dice(img+seg) is a combination of the previous two methods. We did not compare with the DL-based discontinuity-preserving method proposed in [171], as there is no corresponding source code publicly available. This strategy to register different sub-regions and add corresponding deformation fields is also applicable to the aforementioned networks. Hence, we also apply this strategy during inference, for trained Voxelmorph-diff and VM-Dice models (as they only require sub-images as input on the inference), for comparison with DDIR. These are denoted Voxelmorph-diff(add) and VM-Dice(add). These two approaches are different from DDIR as the addition of sub-deformation fields is not learnt end-to-end during training (as in DDIR).

To demonstrate the advantage of incorporating discontinuity in the DL-based registration network, we also build a baseline for DDIR, DDIR(baseline), where the predicted feature maps from the four different channels are concatenated and used to compute a single diffeomorphic deformation field (instead of four sub-deformation fields, as in DDIR).

Qualitative Results. Registration results obtained using DDIR and the other methods investigated are assessed visually in Figure 3.2. Here, the moving and fixed images are shown in the first column. The corresponding warped moving images, deformation fields, and Jacobian determinants (rows 1-3) obtained following registration using SyN, B-spline, Voxelmorph-diff, DDIR(baseline) and DDIR, are shown in columns 2-6. The warped moving images obtained by both traditional registration methods are distinctly different to the fixed image, although the B-spline result appears visually more similar than those obtained by SyN. All warped moving images obtained using DL-based methods look more similar to the fixed image, than the former. The deformation fields and their corresponding Jacobian determinants estimated using each approach indicate that distinct boundaries for the left and right ventricle are retained using DDIR, not afforded by the rest.

Quantitative Results. To quantitatively evaluate the performance of our approach, we compare DDIR with previous methods using the Dice score (DS) and the Hausdorff Distance (HD). DS is computed for LVBP, LVM and RV. These values and the average DS and HD across all regions are reported in Table 3.1. Besides, to demonstrate the clinical value of DDIR, we also compute two clinical indices, LV end-diastolic volume (LVEDV) and LV myocardial mass (LVMM). The former is computed using

ED segmentations, while the latter, is computed using ED and ES segmentations, pre- and post-registration. Pre-registration, LVEDV and LVMM are computed based on the moving and fixed segmentations (used as reference values). Post-registration, we compute them based on the warped moving segmentation. Therefore, as we perform both ED-to-ES and ES-to-ED registration for each subject, the LVMM values reported in Table 3.1 represent the average computed at both ED and ES, across all subjects. Thus the closer LVEDV and LVMM (post-registration) are to the reference values, the better the registration performance.

DL-based approaches outperform traditional registration methods in terms of both DS and HD. The weakly-supervised variants of Voxelmorph-diff provide improvements over Voxelmorph-diff, consistent with previous research[127]. Using segmentation masks

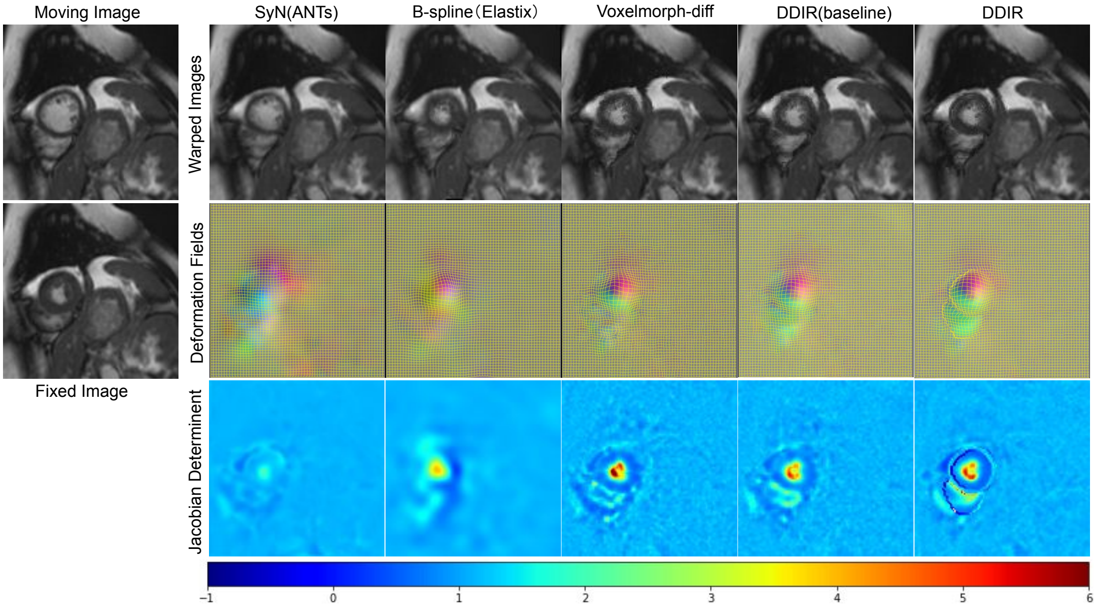

Figure 3.2: Visual comparison of deformation fields estimated using DDIR and state-of-the-art methods. Left column: Moving and fixed images; Right column: corresponding warped moving image (first row), deformation fields (second row) and Jacobian determinant (last row). Colours in the Jacobian determinant images, from blue to red represent the intensity from low to high. The cardiac MR images were reproduced by kind permission of UK Biobank ©.

Table 3.1: Quantitative comparison between DDIR and state-of-the-art methods using the DS of LVBP, LVM, RV and average Dice (denoted as Avg. DS) and HD. Statistically significant improvements in registration accuracy (DS and HD) are highlighted in bold. Besides, LVEDV and LVMM indices with no significant difference from the reference are also highlighted in bold.

Methods	LVBP DS (%)	LVM DS (%)	RV DS (%)	Avg. DS (%)	HD (mm)	LVEDV (ml)	LVMM (g)
before Reg	57.68 ± 6.21	30.88 ± 8.68	55.13 ± 7.51	47.90 ± 6.33	12.91 ± 2.48	143.76 ± 32.13	83.67 ± 21.06
B-spline	74.44 ± 11.50	68.06 ± 7.20	61.76 ± 12.05	68.09 ± 8.76	13.72 ± 3.57	131.14 ± 40.64	81.11 ± 22.60
Demons	80.29 ± 10.00	69.96 ± 5.50	64.86 ± 9.67	71.70 ± 6.96	13.06 ± 3.12	138.00 ± 34.15	80.00 ± 21.25
SyN	70.92 ± 9.36	57.88 ± 10.59	60.30 ± 8.35	63.03 ± 8.29	12.98 ± 2.68	120.09 ± 41.83	**83.12± 21.20**
Voxelmorph-diff	81.73 ± 8.71	72.04 ± 4.65	65.73 ± 9.62	73.16 ± 6.26	12.96 ± 3.14	137.16 ± 32.59	78.65 ± 21.68
VM-Dice	82.28 ± 8.75	72.53 ± 4.59	66.30 ± 9.67	73.70 ± 6.28	13.00 ± 3.24	139.58 ± 32.79	78.98 ± 21.57
VM(img+seg)	82.54 ± 8.50	72.66 ± 4.80	66.69 ± 9.64	73.96 ± 6.28	12.68 ± 3.21	138.29 ± 33.00	80.83 ± 21.62
VM-Dice(img+seg)	81.97 ± 8.53	71.23 ± 4.79	70.20 ± 12.05	74.47 ± 6.79	11.28 ± 4.35	**144.33 ± 32.93**	80.17 ± 22.02
Voxelmorph-diff(add)	78.82 ± 6.38	67.41 ± 8.80	75.10 ± 6.97	73.78 ± 6.10	11.74 ± 3.08	119.30 ± 38.71	91.39 ± 23.07
VM-Dice(add)	79.59 ± 5.91	68.81 ± 7.81	**77.93 ± 6.63**	75.44 ± 5.36	11.14 ± 3.12	120.90 ± 38.14	94.89 ± 25.96
DDIR(baseline)	84.25 ± 8.63	75.02 ± 4.50	71.42 ± 10.32	76.90 ± 6.58	11.85 ± 3.38	**141.73 ± 32.29**	79.01 ± 21.40
DDIR	84.63 ± 8.07	75.27 ± 5.03	74.07 ± 8.73	**77.99 ± 5.47**	**10.65 ± 3.51**	**141.84 ± 32.59**	**81.92 ± 21.86**

as additional input channels to the network (VM(img+seg)) yields better results than using them just to compute the loss and drive gradient updates (VM-Dice) (73.96% vs 73.70%). However, conversely the former requires segmentation masks during inference, while the latter does not. The combination of these two strategies (VM-Dice(img+seg)) further improves registration performance ($\sim$ 0.5% in terms of average DS). Adding sub-deformation fields also improves the registration accuracy of the trained networks, with Voxelmorph-diff (add) achieving 0.6% higher average DS than Voxelmorph-diff (73.78% vs 73.16%), and VM-Dice (add) achieving $\sim$ 1.7% higher average DS than VM-Dice (75.44% vs 73.70%). We found that the DDIR(baseline) achieves $\sim$ 1% higher average DS than VM-Dice(img+seg) (76.90% vs 75.93%), which highlights the advantage of using a multi-channel encoder-decoder network. Compared with DDIR, we found that incorporating discontinuity further improves the average DS (77.99% vs 76.90%). Correspondingly, DDIR also obtains the best performance in terms of the DS for LVBP, LVM and HD, while its RV DS is lower than VM-Dice(add). We evaluated the statistical significance of these results using paired t-tests and found that DDIR significantly outperforms Voxelmorph-diff, VM-Dice, VM(img+seg) and VM-Dice(img+seg) on all DS and HD metrics (P-value<0.05). DDIR also significantly outperforms DDIR(baseline) in terms of average DS, RV DS and HD. Each sub-deformation field generated by DDIR is smooth (without foldings). After composing, the discontinuity only exists at the interface of different sub-regions, which demonstrates that DDIR can generate locally-smooth but globally-discontinuous deformation fields.

The clinical indices, LVEDV and LVMM, show no significant differences (P-value >0.05) post-registration using DDIR to the reference values, not afforded by other approaches. This demonstrates the superiority and clinical value of our method. To analyse the discontinuity on the deformation fields, we visualise the deformation fields generated using DDIR and DDIR (baseline) (presented in Figure A.1 in Appendix A), where the discontinuity is observed for the former along the LV and RV boundaries. To further demonstrate the robustness and generalisability of our approach, we apply the models trained on UKBB data, to the publicly available Automatic Cardiac Diagnosis Challenge (ACDC) data set. The qualitative and quantitative results are included in Figure A.2 and Table A.1 for brevity. As cardiac motion in ACDC images is not as pronounced as in UKBB (in some cases, the images in ED are very similar to ES), only marginal differences in registration performance are observed between DDIR and the other addition-based methods in terms of DS and HD. However, as before, DDIR outperforms Voxelmorph-diff and traditional state-of-the-art methods. Additionally, the clinical indices quantified (LVEDV, LVMM) post-registration using DDIR show no significant differences to the reference, not afforded by any of the other methods investigated. This demonstrates the potential for applying DDIR in real clinical scenarios.

3.1.4 Conclusion

We proposed a novel weakly-supervised discontinuity-preserving registration network, DDIR, which significantly outperformed the state-of-the-art, in intra-patient CMR registration. DDIR preserves LV clinical indices post-registration, not afforded by other approaches. This makes it compelling as a tool for use in clinical applications as it ensures that common diagnostic biomarkers for LV are preserved post-registration.

3.2 Joint Segmentation and Discontinuity-preserving Registration Network

Medical image registration is a challenging task involving the estimation of spatial transformations to establish anatomical correspondence between pairs or groups of images. Recently, deep learning-based image registration methods have been widely explored, and demonstrated to enable fast and accurate image registration in a variety of applications. However, most deep learning-based registration methods assume that the deformation fields are smooth and continuous everywhere in the image domain, which is not always true, especially when registering images whose fields of view contain discontinuities at tissue/organ boundaries. In such scenarios, enforcing smooth, globally continuous deformation fields leads to incorrect/implausible registration results. We propose a novel discontinuity-preserving image registration method to tackle this challenge, which ensures globally discontinuous and locally smooth deformation fields, leading to more accurate and realistic registration results. The proposed method leverages the complementary nature of image segmentation and registration and enables joint segmentation and pair-wise registration of images. A co-attention block is proposed in the segmentation component of the network to learn the structural correlations in the input images, while a discontinuity-preserving registration strategy is employed in the registration component of the network to ensure plausibility in the estimated deformation fields at tissue/organ interfaces. We evaluate our method on the task of intra-subject spatio-temporal image registration using large-scale cine cardiac magnetic resonance image sequences, and demonstrate that our method achieves significant improvements over the state-of-the-art for medical image registration, and produces high-quality segmentation masks for the regions of interest.

3.2.1 Introduction

Image registration involves establishing spatial correspondence between a given pair or group of images, which is fundamental for many downstream medical imaging applications (e.g. image fusion, atlas-based segmentation, image-guided interventions, organ motion tracking and strain analysis, amongst others). Recently, deep learning-based methods have found widespread use in medical image registration, achieving comparable or better performance than traditional registration methods, and yielding substantial speed-ups in execution relative to the latter. Among them, unsupervised and weakly-supervised methods are the most popular as they do not require ground-truth deformation fields to be available. Unsupervised methods [176, 75, 1] do not need any ground-truth annotations (e.g. segmentation, landmarks and ground-truth deformation fields) for training and rely just on the information available in the pair/group of images to be registered. These approaches have been shown to achieve similar or better registration performance than traditional registration methods [177], at a fraction of the execution time. Weakly-supervised methods [76, 127], however, require some annotations (e.g. landmarks, segmentation masks) for training, but have been shown to improve registration performance relative to their unsupervised counterparts.

Currently, most deep learning-based registration methods assume globally smooth and continuous deformation fields throughout the image domain, using regularisation like the L2 norm of deformation fields to ensure that. However, this assumption is not appropriate for all medical image registration applications, especially when there are physical discontinuities resulting in sliding motion between organs/soft tissues that must be estimated to register the input images. For example, respiratory motion resulting from inflation and deflation of the lungs during breathing contains discontinuities between the lungs, the pleural sac encompassing the lungs and the surrounding rib cage. The pleural sac itself contains two layers that slide over one another as the lung inflates and deflates resulting in what is perceived as a sliding motion at the boundaries of the lungs. Enforcing deformation fields to be completely smooth when registering thoracic images of any given individual to recover breathing motion would result in physically unrealistic deformation fields and artefacts near lung boundaries.

Generally, the sliding of organs and different material properties of different subregions in the input images may cause the deformation fields to be locally smooth but globally discontinuous [167]. In previous research, many traditional methods have been proposed to achieve discontinuity-preserving image registration [167, 168, 169, 170, 166, 178, 179, 180, 181, 182]. The fundamental goal of discontinuity-preserving registration is to predict deformation fields which are locally smooth, i.e. within each sub-region, while, discontinuous globally, such as at the interface between different regions/organs. A simple solution to this problem is to register the corresponding sub-organs in the input images independently and then add them to obtain the final deformation field [168]. Another approach is to reformulate the regularisation constraint enforced on the estimated deformation field/functions, to allow for discontinuities at points near the interface between different tissues/organs [169, 170]. However, for those methods, the label information (segmentation masks/landmark) is generally required to delineate where the discontinuity may occur, which may not always be available in realistic scenarios. Therefore, some research has explored achieving discontinuity-preserving registration without requiring *a priori* definition of segmentation masks/landmarks, like using vectorial total variation regularisation [179] or bounded formation theory [180].

Most existing deep learning-based image registration methods do not tackle the problem of estimating deformation fields that preserve discontinuities at tissue/organ boundaries and generally regularise the estimated deformations to be globally smooth and continuous across the image domain. Ng et al.[171] was the first to propose a custom discontinuity-preserving regulariser to constrain the estimation of deformation fields and guide the training of a deep registration network. They assumed that the motion vectors should be parallel to each other, and achieved it by minimising the unsigned area of the parallelogram spanned by two displacement vectors associated with moving image voxels. The advantage of this method was that it did not require label information like segmentation/landmarks. However, it was unable to locate the accurate position of discontinuity that may occur and thereby did not show significant improvement than traditional methods. Previously, we proposed a deep neural network [183] to register pairs of corresponding anatomical structures in the input images separately, and add the deformation fields of each pair to obtain the final deformation field used to warp the source/moving image to the target/fixed image. Instead of using a globally smooth regularisation, the smooth regularisation is applied to each sub-deformation field, which ensures the final deformation fields are locally smooth while globally discontinuous. The proposed approach is shown to significantly outperform state-of-the-art traditional and deep learning-based registration methods. However, it requires segmentation masks to split both the moving and fixed images into corresponding pairs of anatomical regions/structures, during both training and testing, which limits its utility in real scenarios (e.g. segmentation masks may not be readily available for the regions/structures of interest and trained segmentation models to supplement the same may not available either).

This work is an extension of our previous work, namely, the deep discontinuity-preserving registration method (DDIR) presented at the Medical Image Computing and Computer Assisted Intervention (MICCAI) 2021 conference [183]. In [183], the segmentation masks were required during both training and testing to split the original moving and fixed images into pairs of corresponding anatomical regions, limiting its application to scenarios where segmentation masks are readily available for the anatomical regions of interest or can be predicted automatically using a suitable segmentation approach. In this section, we propose a joint registration and segmentation approach wherein, a segmentation module is incorporated within DDIR, which we refer to as SDDIR. This ameliorates the need for segmenting the fixed and moving images prior to registering them. Instead of using ground-truth segmentation, SDDIR applies the predicted segmentation masks to split the moving and fixed images into corresponding pairs of anatomical regions/structures. A co-attention block is used within the segmentation module to learn the structural correlation between the moving and fixed images, and further improve the segmentation performance. A comprehensive set of experiments conducted using a large-scale cardiac cinematic magnetic resonance (cine-MR) imaging dataset, available in the UK Biobank (UKBB) study [184], is used to demonstrate that the proposed approach outperforms competing methods in terms of registration accuracy, whilst also yielding high-quality segmentation masks of the cardiac structures of interest in the fixed and moving images. Additionally, we also demonstrate the generalisation of our method by transferring the pre-trained network on UKBB to two external cardiac magnetic resonance (MR) image datasets, Automatic Cardiac Diagnosis Challenge (ACDC [148]) and Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image Segmentation Challenge [185] (M&M).

3.2.2 Related Work

Jointly Segmentation and Registration

Image segmentation and image registration are both fundamental tasks in computer vision and medical image analysis, which share similarities with each other in terms of the visual cues/features that are relevant for solving either task. Some traditional methods have considered these two independent tasks simultaneously, leading to improved performance in both tasks [186, 187], due to their complementary nature. For example, Droske et al. [186] presented a variational approach to achieve multi-tasks: the detection of corresponding edges, edge-preserving denoising, and morphological registration. They demonstrated that the edge detection and registration tasks were beneficial to the other, with the local weak edge detection improving registration performance and vice versa. Similarly, Dong et al. [187] designed a joint segmentation and registration method for infant brain image registration and found the segmentation and registration steps were mutually beneficial.

In deep learning based methods, several previous research have proposed to achieve segmentation by registration (atlas-based segmentation [188]) or use segmentation to improve registration (e.g. weakly-supervised registration [183]). Most recently, studies have proposed to tackle these tasks jointly in a single end-to-end framework [101, 189, 190], which generally includes two parallel sub-networks, the segmentation sub-network and registration sub-network. A composite loss function is used to train these networks, comprising four terms, namely, the intensity similarity loss, the segmentation loss on moving/fixed images, the regularisation term enforcing estimated deformation fields to be globally smooth, and a segmentation consistency loss term. The segmentation consistency loss is computed on warped moving segmentation and ground-truth fixed segmentation, and is shared by both sub-networks, which helps to improve both segmentation and registration performance. In [101], Xu et al. proposed a novel joint segmentation and registration network, named DeepAtlas, which was flexible and could be applied in samples without label segmentation (segmentation masks). In the training stage, the registration sub-network and segmentation sub-network are trained alternately, with the ground-truth segmentation missing in some of the training samples. To train the registration sub-network, when the ground-truth segmentation was available, the segmentation consistency loss was computed based on the ground-truth segmentation masks, otherwise using the predicted moving and fixed segmentation from the segmentation sub-network. They demonstrated that their method could achieve significant improvement in segmentation and registration than sole segmentation or registration networks. Different from [101], Li et al. [189] trained both sub-networks simultaneously, with the same loss function. They only segmented the moving image in the network and used segmentation accuracy (computed between the predicted moving segmentation and the ground truth moving segmentation mask) and segmentation consistency (computed between the warped moving segmentation and the fixed segmentation mask) for network training. Subsequent studies have also explored removing the requirements of ground-truth segmentation on the segmentation part based on Bayesian inference with probabilistic atlas [190], or extending the idea of jointly learning segmentation and registration to multimodal image registration [191].

Segmentation sub-networks in existing joint segmentation and registration approaches generally segment fixed and moving images independently (or only segment the moving images), and ignore the inherent correlations that exist between them. To exploit this correlated structural information, with a view to enhance joint segmentation and registration performance, we employ a “co-attention” based segmentation sub-network within the proposed approach to jointly segment the fixed and moving input images.

Co-attention based Segmentation

Co-attention based segmentation aims to improve segmentation performance by sufficiently leveraging the structural correlations that exist between multiple input images to be segmented. Generally, there are at least two input images, which contain the same type of objects to segment. By learning common/correlated features from multiple images containing the same objects, the co-attention block has been shown to improve segmentation robustness and accuracy for the objects of interest [192, 193, 194, 195, 196, 197].

The co-attention block is generally used to automatically establish correspondence between correlated regions in input images/feature representations through training on large-scale data, where, the correlated regions would be enhanced while other parts of the images are suppressed. The most popular type of co-attention is spatial co-attention [193, 195, 198], where the co-attention establishes correspondence within the spatial domain of the input images. Sometimes an additional channel co-attention is also used prior to the spatial co-attention [192, 196, 197]. Spatial co-attention has been predominantly applied to image features, however, recent studies have also applied it to graph features learnt in graph neural networks [194]. Li et al. [192] utilised co-attention within a recurrent neural network architecture to learn correlated structural information across a group of images and improve segmentation performance by suppressing the influence of uncorrelated/noisy information. In the group-wise training objective, they used the cross-image similarity between the co-occurring objects and figure-ground distinctness (i.e. distinctness between the detected co-occurring objects and the rest of the images like background) as additional supervision. Additionally, co-attention has also been used for the segmentation of the same object in different time frames, in a video sequence for example. Ahn et al. [193] proposed a multi-frame attention network to learn highly correlated spatio-temporal features in a sequence. Experiments demonstrated that their method significantly outperformed other competing deep learning-based methods. Furthermore, Yang [198] proposed a zero-shot object detection approach for analysing video sequences, using co-attention to learn motion patterns. They empirically demonstrated that their approach outperformed previous zero-shot video object segmentation approaches, while requiring fewer training data.

In this study, we tackle the problem of intra-subject registration of cardiac cine-MR images, acquired at different phases/time points in the cardiac cycle. In intra-subject cardiac image registration, the moving and fixed images are different/deformed representations of the same heart, acquired at different time points in the cardiac cycle (e.g. at end-diastole (ED) to end-systole (ES)). Hence, we hypothesise that joint segmentation of both the fixed and moving images using co-attention can yield more consistent segmentations of the cardiac regions/structures of interest, and in turn, improve the overall registration performance of the proposed approach.

Discontinuity-preserving Image Registration

Discontinuity-preserving image registration has been widely explored in traditional iterative optimisation-based registration approaches but remains relatively unexplored in the context of deep learning-based image registration. In medical image registration, due to various material properties between different tissues/organs, and the physical discontinuities that exist at their boundaries, the underlying deformations that must be recovered to register the images are often locally smooth and globally discontinuous (e.g. at the boundaries between different organs [167]). Consequently, registration methods which constrain estimated deformation fields to be globally smooth, lead to implausible deformations at tissue/organ boundaries. Traditional discontinuity-preserving registration methods can be roughly divided into two categories, those that use additional weak labels such as contours/segmentation masks to guide image registration and others that do not. Methods that require segmentation masks or contour key points delineating the discontinuities of interest between structures/organs in the images, can be further summarised into two categories - (1) registering different sub-regions in the images independently [168, 199, 200]; (2) using custom regularisation constraints to preserve global discontinuities [169, 170, 201] or revising the interpolation function [167, 166] at boundaries between different image sub-regions. These methods have been demonstrated to generate more realistic deformation fields and achieve more accurate results than registration methods that assume globally smooth deformation fields [168, 169, 170, 166].

The aforementioned approaches generally require segmentation masks/landmarks delineating the boundaries of objects/structures of interest, which are not always readily available. Several approaches addressed this issue by revising the regularisation term in the loss/energy function [178, 202, 179, 180, 181, 182], either using some label information like segmentation masks or contour points/landmarks computed pre-registration, or based on different assumptions for physical property of the deformation fields (e.g. isotropic total variation or bounded deformation). Li et al. [181] designed a two-stage registration framework to tackle this issue. They predicted a coarse segmentation mask based on the motion fields predicted during the first stage of image registration, which used mask-free regularisation. Subsequently, in the second stage, the smoothness constraint was relaxed at object/structure boundaries with discontinuities, using masked regularisation and masked interpolation. Similarly, Sandkuhler [202] proposed an adaptive edge weight function based on local image intensities and transformation fields to detect the sliding organ boundaries, then applied an adaptive anisotropic graph diffusion regularisation in the Demons registration to achieve discontinuity-preserving image registration. Some previous approaches do not need to compute any weak label information, prior to registering images [178, 179, 180, 203]. Demirovic et al. [178] proposed to replace the Gaussian filter of the accelerated Demons with a bilateral filter, using information from both displacement and image intensity. By adjusting two tunable parameters, they could obtain more realistic deformations in the presence of discontinuities. Vishnevskiy et al. [179] designed an isotropic total variation regularisation approach for B-splines based image registration, to enable non-smooth deformation fields and used the Alternating Directions Method of Multipliers to solve it. Their method did not require organ masks and could estimate the motion of organs/structures on either side of the discontinuous boundary separating them. By assuming the desired deformation field to be a function of bounded deformation/bounded generalised deformation (referring to [204]), Nie et al [180, 203] built novel variational frameworks to allow possible discontinuities of displacement fields in images, outperforming [179].

Most deep learning-based image registration methods assume the desired deformation fields to be globally smooth and continuous, and do not consider the presence or relevance of discontinuities at structure/organ boundaries and their impact on the image registration task. To our best knowledge, only two previous studies have attempted to preserve discontinuities at object/structure boundaries in deep learning-based image registration [171, 183]. Ng et al. [171] addressed this issue in an unsupervised manner. They proposed a discontinuity-preserving regularisation term by comparing local displacement vectors with neighbouring displacement vectors individually, which was able to tackle specific behaviours on the discontinuous deformation fields. Without ground-truth information specifying the locations of discontinuous boundaries, their registration performance did not show significant improvements than traditional approaches. In contrast, we previously [183] proposed a deep discontinuity-preserving image registration (DDIR) approach, to generate locally smooth sub-deformation fields for each image sub-region, which were then added to obtain locally smooth and globally discontinuous deformation fields. Although [183] was shown to outperform the state-of-the-art, its need for segmentation masks delineating the objects/structures of interest during inference, limits its application in real-world scenarios. Therefore, in this section, to tackle this issue, a joint segmentation and registration approach is proposed for discontinuity-preserving registration, which only requires ground-truth masks in the training process.

3.2.3 Method

In this work, we focus on pair-wise image registration, aiming to establish spatial correspondence between the moving image $\mathbf{I}_M$ and fixed image $\mathbf{I}_F$. This task can be formulated as,

$$\phi(\mathbf{x}) = \mathbf{x} + u(\mathbf{x}), \tag{3.4}$$

where, $\mathbf{x}$ is the coordinate of voxels/pixels in the moving image $\mathbf{I}_M$, $u(\mathbf{x})$ and $\phi(\circ)$ represents the displacement field and the deformation function, respectively.

To preserve the discontinuities during image registration, similar to our previous study [183], we decompose the fixed and moving images into corresponding pairs of image sub-regions, register each pair and combine the obtained sub-deformation fields to obtain the final deformation field used to warp the moving image. Different from [183], in this study we propose a joint segmentation and discontinuity-preserving image registration (SDDIR) approach, which includes an additional segmentation sub-network in DDIR [183], to jointly segment the fixed and moving images. The primary motivation for the approach proposed in this study is to ameliorate the need for separately sourcing segmentation masks (either manually or automatically) for the images to be registered, as required by DDIR. As shown in Figure 3.3, our SDDIR includes a segmentation sub-network and a registration sub-network. The input fixed and moving images are first fed into the segmentation branch and tissue/organ specific segmentation masks are predicted for each image. For example, the focus of this study is on intra-subject cardiac cine-MR image registration, and given input cine-MR images,

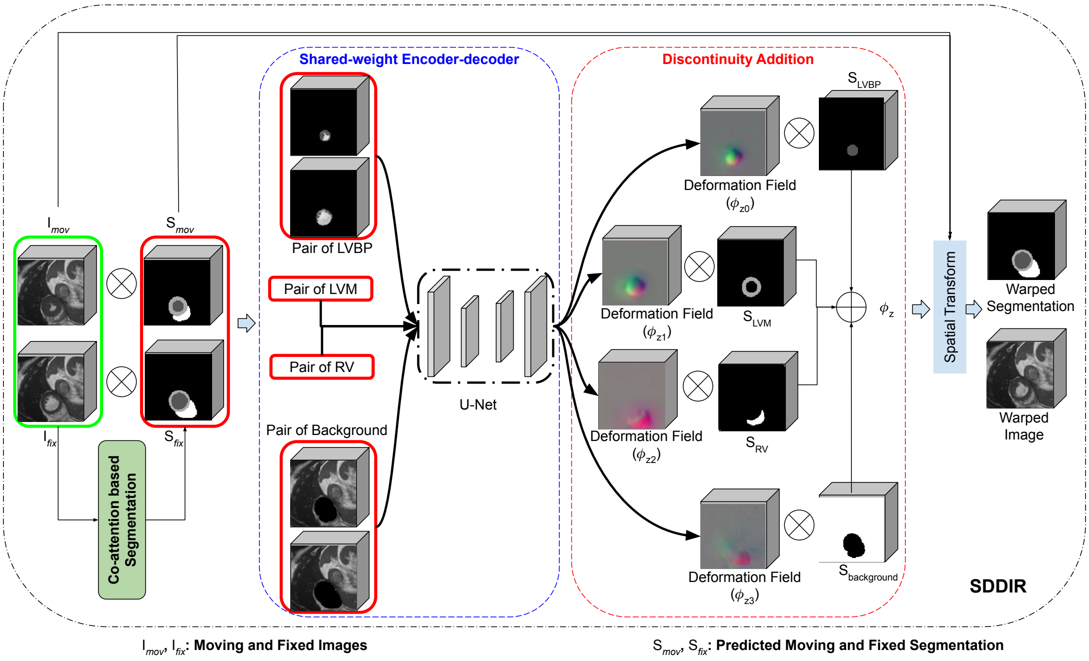
Shared-weight Encoder-decoder
Discontinuity Addition
I_{mov}
S_{mov}
I_{fix}
S_{fix}
Co-attention based Segmentation
Pair of LVBP
Pair of LVM
Pair of RV
Pair of Background
U-Net
Deformation Field (ϕ_{z0})
Deformation Field (ϕ_{z1})
Deformation Field (ϕ_{z2})
Deformation Field (ϕ_{z3})
S_{LVBP}
S_{LVM}
S_{RV}
$S_{background}$
ϕ_z
Spatial Transform
Warped Segmentation
Warped Image
SDDIR
I_{mov}, I_{fix}: Moving and Fixed Images
S_{mov}, S_{fix}: Predicted Moving and Fixed Segmentation

Figure 3.3: Schema of SDDIR. The proposed network includes a co-attention based segmentation block, a shared-weight encoder-decoder and a discontinuity addition block. The details of the co-attention based segmentation block can be found in Figure 3.4, which provides segmentation masks for subsequent registration tasks. The shared-weight encoder-decoder is to extract features from pairs of LVBP, LVM, RV and background. Based on them, we obtain four sub-deformation fields for different regions. The final deformation field is obtained by composing these four sub-deformation fields with corresponding segmentation, through a discontinuity addition block. The cardiac MR images were reproduced by kind permission of UK Biobank ©.

four-class segmentation masks are predicted, delineating the following regions - left ventricle myocardium, left ventricle blood pool, right ventricle blood pool and background tissue. Subsequently, the fixed and moving images, and their predicted segmentation masks are fed into the discontinuity-preserving image registration branch, which predicts region/structure-specific sub-deformation fields, and adds them into a final deformation field used to warp the moving image. The segmentation and registration branches are trained jointly end-to-end, as a single network, using a combined composite loss function. In subsequent sections, we describe the co-attention based segmentation sub-network, discontinuity-preserving image registration sub-network, and the composite loss function used in the proposed approach.

Co-attention Based Segmentation

The segmentation sub-network in SDDIR is based on a 3D U-Net [205], with a co-attention block in the bottleneck layer designed to learn structural correlations between the fixed and moving images, as shown in Figure 3.4. In the segmentation sub-network, the encoder and decoder branches each comprise two pairs of downsampling and upsampling convolution blocks, respectively. The encoder contains two separate channels to encode the original moving and fixed images from $R^{H \times W \times D}$ into features of $R^{\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4}}$. Each encoder channel uses two downsampling blocks (comprising a convolution layer, an activation and an average-pooling layer). The bottleneck layer contains the co-attention block which takes fixed and moving image features extracted by the encoder as inputs and predicts corresponding attention maps. Similarly, in the decoder, two separate channels comprising two upsampling convolution blocks each, are used to predict segmentation masks for the fixed and moving images in their original size/resolution, given their corresponding attention feature maps as inputs. Here, each upsampling block comprises an upsampling layer, a convolution layer and an activation layer. To improve network training and performance, skip-connections are used to concatenate features from each layer in the encoder with its corresponding layer (i.e. at the same spatial resolution) in the decoder.

The co-attention block predicts task-specific attention feature maps for both the fixed and moving images, where relevant pixels are enhanced while the rest are suppressed. The moving and fixed image feature maps $\mathbf{F}_{mov}, \mathbf{F}_{fix} \in \mathcal{R}^{W \times H \times D \times C}$ (C, W, H and D are channels, width, height and depth of feature maps, respectively) are first

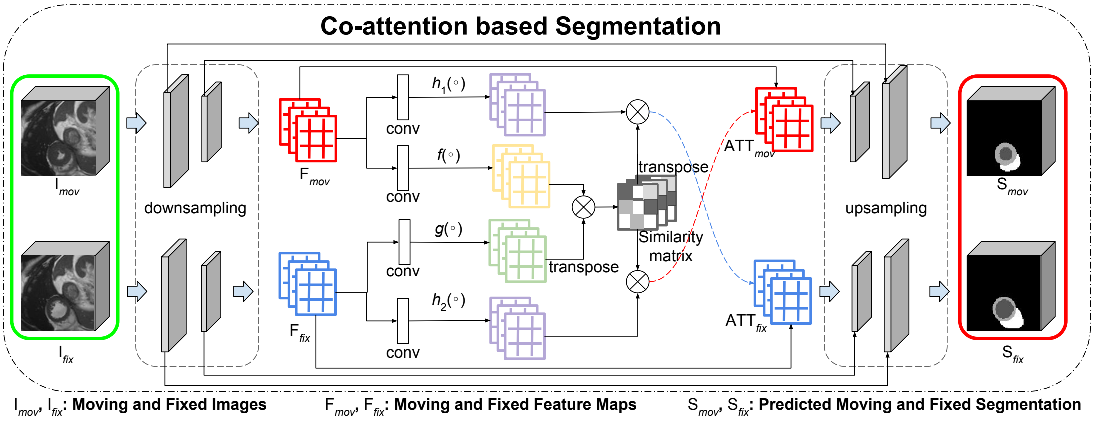
Co-attention based Segmentation
I_mov
I_fix
downsampling
F_mov
F_fix
conv
conv
conv
conv
h_1(∘)
f(∘)
g(∘)
h_2(∘)
transpose
Similarity matrix
transpose
ATT_mov
ATT_fix
upsampling
S_mov
S_fix
I_mov, I_fix: Moving and Fixed Images
F_mov, F_fix: Moving and Fixed Feature Maps
S_mov, S_fix: Predicted Moving and Fixed Segmentation

Figure 3.4: The design of co-attention based segmentation. Two downsampling blocks are used to extract features from original moving and fixed images. With the co-attention, the corresponding moving and fixed attention maps can be learnt, which are concatenated with the input features in the same size as the input of upsampling blocks. After two upsampling blocks and a Softmax operation, the segmentation masks of moving and fixed images are obtained. The cardiac MR images were reproduced by kind permission of UK Biobank ©.

transformed into two different feature spaces (denoted as $f(\circ)$ and $g(\circ)$) by two $1\times1\times1$ convolution layers and flattened (from $\mathcal{R}^{W\times H\times D\times C}$ to $\mathcal{R}^{N\times C}, N = W\times H\times D$) to calculate similarity matrix $\mathbf{S}\in\mathcal{R}^{N\times N}$. With the similarity matrix $\mathbf{S}$ and the feature maps learnt from the inputs using two additional $1\times1\times1$ convolution layers (denoted $h_1(\circ)$ and $h_2(\circ)$), the fixed attention maps $\mathbf{ATT}_{fix}$ and moving attention maps $\mathbf{ATT}_{mov}$ are computed. The process of co-attention can be formulated as,

$$
\begin{aligned}
&\mathbf{S} = f(\mathbf{F}_{mov}) \times g(\mathbf{F}_{fix})^T,\\
&\mathbf{ATT}_{mov} = Softmax(\mathbf{S}) \times h_2(\mathbf{F}_{fix}),\\
&\mathbf{ATT}_{fix} = Softmax(\mathbf{S}^T) \times h_1(\mathbf{F}_{mov}),\\
&\mathbf{O}_{mov} = Concat(\mathbf{F}_{mov}, \sigma(\mathbf{ATT}_{mov}) \cdot \mathbf{ATT}_{mov}),\\
&\mathbf{O}_{fix} = Concat(\mathbf{F}_{fix}, \sigma(\mathbf{ATT}_{fix}) \cdot \mathbf{ATT}_{fix}),
\end{aligned}
\tag{3.5}
$$

where, $\mathbf{O}_{mov}$ and $\mathbf{O}_{fix}$ are the output feature maps (learnt new representations) of $\mathbf{F}_{mov}$ and $\mathbf{F}_{fix}$, following application of their estimated co-attention maps, respectively. $Softmax(\circ)$ is the Softmax function, applied to the last channel of the similarity matrix $\mathbf{S}$. $\sigma(\circ)$ denotes the Sigmoid function, which is a $1 \times 1 \times 1$ convolution layer followed by a Sigmoid activation. $Concat(\circ)$ is to concatenate the input feature with the corresponding attention maps, comprising a concatenation, a $1 \times 1 \times 1$ convolution layer, a batch-normalisation and an activation layer.

Following the two upsampling blocks in the decoder, the co-attention feature maps of the fixed and moving images are recovered to the original size and resolution of the input images. A $3 \times 3 \times 3$ convolution followed by a Softmax activation function is used to predict the segmentation masks of the moving and fixed images. The focus of this study is on intra-subject spatiotemporal registration of cine-MR image sequences, i.e. pair-wise registration of images acquired at different time points in the cardiac cycle. We train and evaluate the performance of SDDIR on cardiac cine-MR images available from the UKBB database. We focus on segmenting and decomposing the fixed and moving images into four sub-regions, namely, the left ventricle blood pool (LVBP), left ventricle myocardium (LVM), right ventricle (RV) and background. It is important to note that while the focus of this study is on intra-subject cardiac MR image registration, the proposed method is agnostic to imaging modality, organ(s) of interest and application. SDDIR may be used for joint segmentation (into regions/organs of interest) and registration of other types of images (e.g. computed tomography, computed tomography angiography, x-ray, MR angiography, etc.).

The co-attention block is inspired by [206], but differs in the following ways - (1) Skip-connections are applied in our implementation (as shown in Figure 3.4), which helps ensure better flow of gradients during training and helps improve overall segmentation performance; (2) In addition to the standard segmentation loss (e.g. Dice loss between the predicted segmentation and ground-truth segmentation), we also compute the cross-entropy between the warped predicted segmentation and the ground-truth fixed segmentation, as a segmentation “consistency” loss, to ensure that the predicted segmentations for the fixed and moving images and the deformation field mapping the latter to the former, are consistent with each other.

Discontinuity-preserving Registration

The segmentation masks predicted for the input fixed and moving images (by the segmentation sub-network) are passed as inputs along with their corresponding original images, to the registration sub-network. To estimate the desired locally smooth and globally discontinuous deformation field mapping the moving image to the fixed image, we first predict four different smooth sub-deformation fields for each of the four sub-regions of interest in the cardiac MR images (i.e. LVBP, LVM, RV and background), and then add them to obtain the final deformation field.

Network Architecture The segmentation masks predicted by the segmentation sub-network are used to split the original pair of fixed and moving images into four different image pairs, comprising, the LVBP, LVM, RV and background sub-regions. As shown in Figure 3.3, in each pair, the pixel/voxel values within the mask are retained, while those from the surrounding regions are set to zero. Then, a shared-weight U-Net (comprising four downsampling and three upsampling blocks) is used to learn features from all four image pairs. Therefore, we obtain features at $64 \times 64 \times 8$ from the original image pairs ($128 \times 128 \times 16$). A shared-weight convolution layer followed by a scaling and squaring layer is used to process the learnt features and estimate their corresponding diffeomorphic sub-deformation fields. The predicted moving segmentation masks are used again to extract the corresponding regions in the estimated sub-deformation fields and combine them to obtain the final globally discontinuous deformation field. Finally, a spatial transform layer is used to warp the moving image and predicted segmentation using the discontinuous deformation field.

Discontinuity Addition The addition of deformation fields estimated for relevant image sub-regions is essential to ensure locally smooth and globally discontinuous deformations fields are used for registering images. Similar to previous papers [127, 88, 183], we assume the transformation function (denoted as ϕ_z) is parameterised by stationary velocity fields (SVF) ($z_i, i \in [0, 3]$). With the predicted feature map, we obtain four SVFs (z_0, z_1, z_2, z_3) corresponding to different regions (LVBP, LVM, RV and background) using a shared-weight convolution layer of size $3 \times 3 \times 3$, whose weights are sampled from a Normal distribution. The SVFs are integrated by scaling and squaring layers (referring to [127]) to diffeomorphic deformation fields. After an upsampling operation, we obtain four diffeomorphic deformation fields ϕ_{z_0}, ϕ_{z_1}, ϕ_{z_2} and ϕ_{z_3}. Similarly, we use the predicted fixed segmentation masks to extract each region of interest from the obtained deformation fields and add them to generate the final deformation field. Let the segmentation masks of LVBP, LVM, RV and background be

S_{LVBP}, S_{LVM}, S_{RV} and S_B respectively, the addition can be formulated as,

$$\phi_z = \phi_{z_0} \times S_{LVBP} + \phi_{z_1} \times S_{LVM} + \phi_{z_2} \times S_{RV} + \phi_{z_3} \times S_B. \tag{3.6}$$

Loss Function

The segmentation and registration sub-networks within the proposed approach are trained jointly using the same loss function L_{total}. The loss function L_{total} includes four terms: segmentation accuracy loss, image similarity loss, segmentation consistency loss and a discontinuity-preserving regularisation term. The segmentation accuracy loss includes two parts, the accuracy loss for the moving image and the fixed image. We use cross-entropy to compute the distance between the predicted segmentation and their respective ground-truth segmentation masks. Denoting the cross-entropy loss as CE, the segmentation accuracy loss L_{seg} is formulated as,

$$L_{seg} = CE(S_{pre}^{mov}, S_{gt}^{mov}) + CE(S_{pre}^{fix}, S_{gt}^{fix}), \tag{3.7}$$

where S_{pre}^{mov},S_{gt}^{mov},S_{pre}^{fix},S_{gt}^{fix} are the predicted and ground-truth segmentation of moving and fixed images, respectively.

The image similarity loss evaluates the dissimilarity between the warped moving image and the fixed image. We use the mean squared error to evaluate the distance between them, formulated as,

$$L_{MSE} = \frac{1}{W \times H \times D} \sum (I_{mov} \circ \phi - I_{fix})^2. \tag{3.8}$$

The segmentation consistency loss links the segmentation and registration sub-networks, allowing them to be jointly optimised. This loss term is computed as the Dice overlap [173] between the predicted fixed segmentation and the warped predicted moving segmentation, formulated as,

$$L_{Dice} = 1 - \frac{2|(S_{pre}^{mov} \circ \phi) \cap S_{pre}^{fix}|}{|S_{pre}^{mov} \circ \phi| + |S_{pre}^{fix}|}. \tag{3.9}$$

The discontinuity-preserving regularisation must ensure that estimated deformation fields are locally smooth and globally discontinuous. Specifically, discontinuous at boundaries between structures/regions of interest (i.e. in our case at boundaries between LVBP, LVM, RV and background). Therefore, we cannot enforce a global smoothness constraint on the final deformation field. As the addition of different deformation fields preserves discontinuities at interfaces, we only need to guarantee the deformation field of each sub-region is smooth. This is achieved by applying L_2-regularisation on each sub-deformation field, also referred to as the diffusion regulariser [127, 76], denoted R, on the spatial gradients of each sub-displacement field $\mathbf{u}$. R is formulated as,

$$\begin{aligned} R(\phi) &= || \bigtriangledown \mathbf{u}||^2, \\ L_R &= \frac{1}{4}(R_{LVBP} + R_{LVM} + R_{RV} + R_{background}), \end{aligned} \tag{3.10}$$

where L_R denotes the combined regularisation terms for each sub-region.

The complete loss function used to train the network is,

$$L_{total} = \lambda_0 \times L_{seg} + \lambda_1 \times L_{MSE} + \lambda_2 \times L_{Dice} + \lambda_3 \times L_R, \tag{3.11}$$

where, λ_0, λ_1, λ_2 and λ_3 are hyper-parameters used to weight the importance of each loss term.

3.2.4 Experiments and Results

Data and Implementation

The proposed approach, SDDIR, is trained and evaluated on three publicly available cardiac MR image datasets, namely, the UKBB [184], ACDC [148] and M&M [185]. We choose 1437 subjects from the UKBB dataset, each including short-axis (SAX) image stacks at ED and ES. The pixel spacing of images in the UKBB is $\sim 1.8 \times 1.8 \times 10mm^3$. To train the network, we pre-process all image volumes by cropping and padding (with zeros) them to a fixed size of $128 \times 128 \times 16$. In this work, we focus on intra-subject deformable image registration, specifically, registering images from ED to ES and ES to ED. We split the UKBB data into a training set (1080 subjects), a validation set (157 subjects), and a test set (200 subjects). This resulted in a total of 2160, 304, and 400 samples (i.e. pair of fixed and moving images for each subject) that were used for training, validation and testing. The ground-truth segmentation masks for the UKBB dataset were manually annotated by experts, as part of a previous study [207]. To verify the generalisation and robustness of the proposed approach, we also apply the trained model (i.e. trained on UKBB data) to images from the ACDC and M&M datasets. Similarly, we choose the ED and ES SAX images from 100 subjects (total of 200 samples) in the ACDC dataset, whose ground-truth segmentation masks are available. In the M&M dataset, 300 samples (registration from ED to ES and from ES to ED) are extracted from 150 subjects. Each image volume in ACDC and M&M is pre-processed similarly to the UKBB data, resulting in images of size 128 × 128 × 16 using resampling, cropping and padding. Note that, to reduce the domain gap between different datasets, histogram-matching is applied to the ACDC and M&M images, using a random image volume from UKBB as the reference.

The SDDIR was implemented in Python using PyTorch, on a Tesla M60 GPU machine. The Adam optimiser with a learning rate of $1e^{-3}$ was used to optimise the network. We set the batch-size to 3, due to limitations in GPU memory available. The hyper-parameters in the total loss L_{total} λ_0, λ_1, λ_2 and λ_3 were tuned empirically and were set to 0.1, 1, 0.1, 0.01, respectively, throughout all experiments presented in this study. The source code will be publicly available on Github (https://github.com/cistib/DDIR).

Competing Methods and Evaluation Metrics

To highlight the benefits of the proposed approach, we quantitatively compare the registration performance of SDDIR against both traditional and state-of-the-art deep learning-based registration methods. Three traditional registration methods are compared against SDDIR, namely, the Symmetric Normalisation (SyN, using 3 resolution levels, with 100, 80, 60 iterations respectively) in ANTs [70], Demons (Fast Symmetric Forces Demons [174] with 100 iterations and standard deviations 1.0) available in SimpleITK, and B-splines registration (max iteration step is 4000, sampling 4000 random points per iteration), available in SimpleElastix [175]. State-of-the-art deep learning-based registration methods chosen for quantitative comparison against the proposed approach include, Voxelmorph (VM [127]), the weakly-supervised version of VM (denoted as VM-Dice), and a baseline joint segmentation and registration network, named Baseline. VM-Dice essentially trains the original VM approach in a weakly supervised manner, with a Dice loss L_{Dice} on the warped moving segmentation and fixed segmentation (using the ground-truth segmentation masks). We implement the Baseline network by referring to [101, 189] based on our setting, which uses a general U-Net for segmentation and a VM-like architecture for registration. It is trained with the same loss function as SDDIR, where the only connection between the segmentation sub-network and the registration sub-network is the segmentation consistency loss. All the networks are trained until convergence on the training dataset, and the hyper-parameters and final models are selected based on their performance on the validation set.

We also compare the proposed approach, against other discontinuity-preserving registration methods, namely, DDIR [183] and two other sub-deformation field addition methods investigated previously in [183], denoted VM(add) and VM-Dice(add). In these addition-based methods, the original MR images are first split into four different pairs, using the ground-truth segmentation masks. Then the trained network (VM or VM-Dice) is used to register those pairs independently. The obtained sub-deformation fields are added into the final deformation field, which is used to warp the moving image and segmentation. This strategy is a simple and conventional approach to enabling the estimation of discontinuity-preserving deformation fields. In contrast with SDDIR, in this strategy networks are not trained end-to-end and they require segmentation masks to be available during inference.

Registration performance of each method investigated is quantitatively evaluated and compared using the following metrics - Dice scores (computed between the warped moving segmentation and fixed segmentation) on LVBP, LVM and RV, the average Dice score (denoted as Avg. DS) across all cardiac structures, and Hausdorff distance (95%) (HD95), where, higher Dice score and lower HD95 indicate better registration performance. Additionally, two clinical cardiac indices, the LV end-diastolic volume (LVEDV) and LV myocardial mass (LVMM), are also computed to demonstrate that the proposed registration approach preserves clinically relevant volumetric indices post image registration. They are calculated based on the warped moving segmentation of LVBP and LVM post ES-to-ED registration ($LVEDV = V_{LVBP} \times SP_x \times SP_y \times SP_z$, $LVMM = V_{LVM} \times SP_x \times SP_y \times SP_z \times Den$, where V, SP, Den are the volume of structures, spacing of images, and an assumed myocardial density (1.05 g/ml) [208, 209], respectively. V_{LVBP} indicates the left ventricle blood pool volume computed by all the voxels in the LVBP mask and V_{LVM} denotes the left ventricle myocardium volume computed by all the voxels in the LVM mask). The closer the clinical indices are to the reference (i.e. the clinical indices computed based on ground-truth ED segmentation, presented in the row “before Reg”), the better the registration performance. To assess segmentation performance, the average Dice scores (denoted as Seg DS) across all cardiac structures for the predicted moving and fixed segmentation masks, with respect to their corresponding ground-truth segmentation masks, are also calculated. To evaluate the smoothness of deformation fields, we also compute the percentage of voxels with a non-positive Jacobian determinant (denoted as $|J| \leq 0$) for each method, where a lower percentage means smoother deformation fields. Note that, for discontinuity-preserving registration approaches (with four sub-deformation fields), we compute the average $|J| \leq 0$ of sub-deformation fields.

Registration Results: UKBB Data

Quantitative registration results obtained for the unseen test set from UKBB are summarised in Table 3.2(the corresponding P-values are shown in Table 3.3). The Baseline joint segmentation and registration network achieves higher Dice scores on the registration results than those solely designed for image registration (e.g. Demons, B-spline and VM). The addition-based methods VM(add) and VM-Dice(add) do not show any improvements over VM and VM-Dice, and perform consistently worse than VM-Dice across all metrics evaluated. While the Baseline network significantly outperforms VM across all metrics, its average Dice score (computed across all three cardiac structures, LVBP, LVM and RV), only marginally outperforms VM-Dice, and it performs worse than VM-Dice in terms of HD95. Both the Baseline and VM-Dice networks use the Dice loss to guide the training of their constituent registration networks. The architectures of the constituent registration networks are almost identical, leading to similar performance of the Baseline and VM-Dice networks. SDDIR significantly outperforms the Baseline network in terms of both the Dice score and HD95, highlighting the superior registration performance of the proposed approach. The DDIR approach, which uses manually annotated ground truth segmentation masks for decomposing the input images into regions of interest during inference, achieves the best registration performance of all methods investigated. DDIR achieves an average Dice score 5% higher than SDDIR and an HD95 score that is on average 2 mm lower than SDDIR. In addition, its deformation fields are most smooth over all approaches (lowest percentage of $|J| \leq 0$). In terms of the clinical indices, the LVMM values of both DDIR and SDDIR show no significant differences with respect to the reference, and the LVEDV of SDDIR also makes no significant difference to the reference (which is not achieved by DDIR). While DDIR shows some improvements over SDDIR in terms of registration accuracy, the latter does not require segmentation masks to be provided as inputs during inference (as required by DDIR). SDDIR thus has more flexibility in its utility/application and is better suited to real-world scenarios where segmentation masks for regions of interest may not be available prior to registering the input images.

Table 3.2: Quantitative comparison on UKBB between SDDIR and state-of-the-art methods. Statistically significant improvements of SDDIR over previous methods (excluding DDIR) in registration and segmentation accuracy (DS and HD95) are highlighted in bold. Besides, LVEDV and LVMM indices with no significant difference from the reference (shown in the row of "before Reg") are also highlighted in bold.

Methods	Avg. DS (%)	LVBP DS (%)	LVM DS (%)	RV DS (%)	HD95 (mm)	Seg DS (%)	LVEDV(ml)	LVMM(g)	% of $\lvert J\rvert \leq 0$
before Reg	43.8 ± 5.5	63.1 ± 14.2	52.9 ± 13.1	68.1 ± 15.4	15.4 ± 4.6	-	157.5 ± 32.8	99.5 ± 27.9	-
B-spline	66.8 ± 6.6	75.1 ± 9.4	62.7 ± 7.9	62.7 ± 7.8	15.7 ± 4.9	-	140.3 ± 41.1	**100.3 ± 30.3**	$4.6e^{-3} \pm 4.5e^{-3}$
Demons	68.1 ± 5.7	77.6 ± 7.8	63.1 ± 7.2	63.6 ± 7.3	14.1 ± 4.5	-	139.6 ± 37.3	102.5 ± 30.3	$2.5e^{-3} \pm 3.7e^{-3}$
SyN	64.5 ± 6.1	77.5 ± 8.8	56.2 ± 6.8	59.7 ± 10.9	13.5 ± 3.9	-	149.8 ± 32.4	93.1 ± 27.7	$2.8e^{-2} \pm 9.7e^{-3}$
VM	74.2 ± 4.9	85.5 ± 6.4	69.4 ± 6.6	67.6 ± 7.4	12.7 ± 4.6	-	151.9 ± 33.8	97.1 ± 29.3	$2.0e^{-3} \pm 1.3e^{-3}$
VM-Dice	79.8 ± 4.4	86.9 ± 5.8	71.7 ± 6.8	80.7 ± 5.9	8.7 ± 4.6	-	162.1 ± 34.4	**100.4 ± 28.6**	$4.1e^{-3} \pm 2.0e^{-3}$
Baseline	79.9 ± 4.3	88.4 ± 5.1	73.6 ± 6.4	77.9 ± 6.8	9.7 ± 4.1	87.9 ± 2.7	158.6 ± 33.9	97.7 ± 28.5	$3.0e^{-3} \pm 1.9e^{-3}$
VM(add)	70.6 ± 10.2	80.0 ± 13.3	55.4 ± 10.9	76.3 ± 10.9	10.3 ± 3.7	-	150.9 ± 33.5	84.9 ± 31.3	$1.4e^{-3} \pm 6.9e^{-4}$
VM-Dice(add)	72.9 ± 11.5	83.1 ± 14.9	57.8 ± 11.4	77.7 ± 13.1	9.0 ± 4.3	-	156.9 ± 33.1	81.9 ± 32.5	$1.9e^{-3} \pm 1.2e^{-3}$
DDIR	92.7 ± 4.9	94.9 ± 4.6	90.9 ± 6.1	92.4 ± 6.2	4.4 ± 5.2	-	157.9 ± 33.2	**99.1 ± 27.5**	$3.8e^{-4} \pm 4.3e^{-4}$
SDDIR	**87.7 ± 3.4**	**92.7 ± 3.9**	**83.1 ± 5.2**	**87.2 ± 5.5**	**6.7 ± 4.9**	**88.6 ± 2.1**	**157.6 ± 32.9**	**98.7 ± 26.1**	$3.5e^{-3} \pm 1.6e^{-3}$
SDDIR(-DA)	80.5 ± 4.3	88.5 ± 5.0	74.2 ± 6.3	78.6 ± 6.6	9.4 ± 4.1	88.2 ± 2.3	156.4 ± 33.4	**98.9 ± 28.4**	$2.6e^{-3} \pm 1.4e^{-3}$

Table 3.3: P-values between SDDIR and state-of-the-art methods in Table 3.2. Regarding LVEDV and LVMM, the P-values are computed between the results post-registration and reference. All P-values larger than 0.05 are highlighted in bold.

Methods	Avg. DS (%)	LVBP DS (%)	LVM DS (%)	RV DS (%)	HD95 (mm)	Seg DS (%)	LVEDV(ml)	LVMM(g)	% of $\lvert J\rvert \leq 0$
B-spline	$5.6e^{-216}$	$1.2e^{-137}$	$4.9e^{-195}$	$3.4e^{-195}$	$1.2e^{-122}$	-	$6.8e^{-34}$	**0.17**	-
Demons	$7.3e^{-241}$	$4.2e^{-160}$	$1.2e^{-207}$	$2.0e^{-218}$	$1.5e^{-112}$	-	$8.2e^{-46}$	$1.6e^{-8}$	-
SyN	$2.2e^{-256}$	$1.6e^{-154}$	$1.3e^{-256}$	$4.8e^{-176}$	$2.3e^{-95}$	-	$2.8e^{-23}$	$4.4e^{-25}$	-
VM	$2.0e^{-214}$	$2.9e^{-105}$	$9.6e^{-192}$	$1.3e^{-190}$	$8.2e^{-100}$	-	$3.3e^{-18}$	$3.6e^{-5}$	
VM-Dice	$9.9e^{-176}$	$3.3e^{-99}$	$2.7e^{-168}$	$1.2e^{-107}$	$3.0e^{-35}$	-	$9.1e^{-23}$	**0.08**	
Baseline	$9.6e^{-175}$	$6.2e^{-79}$	$2.7e^{-156}$	$1.5e^{-130}$	$5.0e^{-57}$	$3.3e^{-25}$	0.03	$3.2e^{-4}$	-
VM(add)	$7.1e^{-140}$	$9.7e^{-78}$	$2.9e^{-174}$	$2.1e^{-90}$	$2.6e^{-64}$	-	$1.0e^{-33}$	$3.5e^{-47}$	-
VM-Dice(add)	$1.4e^{-104}$	$5.2e^{-43}$	$1.4e^{-158}$	$8.7e^{-59}$	$2.7e^{-31}$	-	$1.4e^{-4}$	$6.3e^{-52}$	-
DDIR	$4.3e^{-109}$	$1.3e^{-35}$	$4.4e^{-96}$	$1.2e^{-81}$	$2.4e^{-27}$	-	$3.0e^{-3}$	**0.09**	-
SDDIR	-	-	-	-	-	-	**0.75**	**0.13**	-
SDDIR(-DA)	$3.0e^{-167}$	$2.1e^{-77}$	$8.8e^{-151}$	$9.5e^{-128}$	$4.8e^{-55}$	$1.8e^{-16}$	0.01	**0.29**	-

Registration results visualised in Figure 3.5 indicate that the warped moving image predicted by SDDIR is more similar to the fixed image than those predicted by the other methods investigated, which is consistent with the quantitative results obtained. This is especially evident along the boundaries of the right ventricle and the left ventricle myocardium. Additionally, Figure 3.5 indicates that compared with all

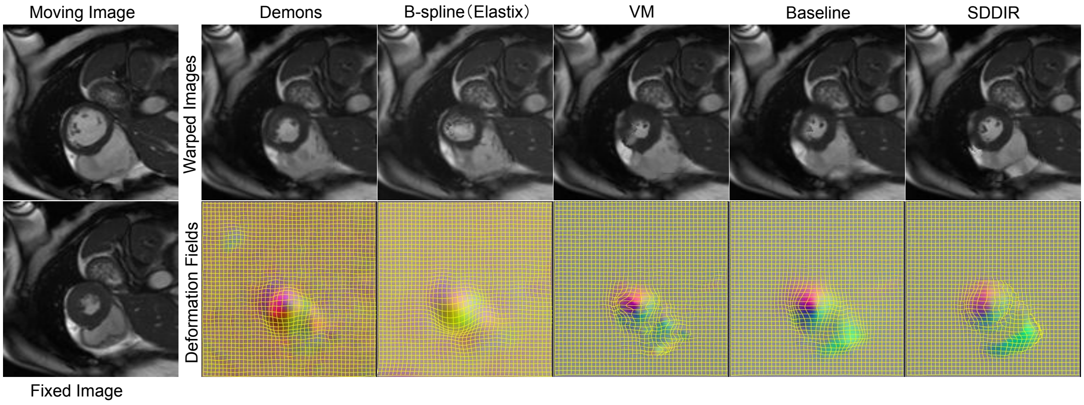

Figure 3.5: Visual comparison of results on UKBB estimated using SDDIR and state-of-the-art methods. Left column: moving and fixed images; Right column: corresponding warped moving image (first row), deformation fields (second row). The cardiac MR images were reproduced by kind permission of UK Biobank ©.

other approaches, the deformation field estimated by SDDIR captures discontinuities at boundaries between different structures/sub-regions (such as between the left and right ventricle, for example) more strongly.

Registration Results: ACDC and M&M Data

To assess the ability of the proposed approach to generalise to unseen data representative of real-world data acquired routinely in clinical examinations, we apply the SDDIR model pre-trained on UKBB data, to other external cardiac MR data sets, namely, ACDC and M&M. Data available in ACDC and M&M were acquired at multiple different imaging centres distributed across different countries, using different types of MR scanners, and from patients diagnosed with different types of cardiac diseases/abnormalities (e.g. myocardial infarction, dilated cardiomyopathy, hypertrophic cardiomyopathy, abnormal right ventricle). Generalising to such unseen data is challenging due to domain shifts in the acquired images, relative to the UKBB data used for training SDDIR. The quantitative and qualitative results obtained for data from ACDC are shown in Table 3.4 (the corresponding P-values are shown in Table 3.5) and Figure 3.6, respectively. For the results on ACDC, consistent with results in

UKBB, VM outperforms the traditional registration approaches (B-spline, Demons and SyN). The VM-Dice consistently performs better than VM, while its addition version, VM-Dice(add), achieves lower registration accuracy than the addition version of VM (VM(add)). As Baseline only uses the predicted segmentations to compute the Dice loss (rather than using it to partition the original images as in SDDIR), the registration quality of the Baseline network is less dependent on segmentation quality. As a result, although the segmentation performance of Baseline is significantly worse, it performs comparably to SDDIR in terms of registration quality on ACDC, achieving 1% higher Dice, and a marginally lower average HD95 score. Using ground-truth segmentation masks during inference, DDIR performs the best out of all models investigating, achieving an average Dice score of 87% Dice score. The registration performance of SDDIR drops significantly relative to the results obtained for UKBB data, with SDDIR performing marginally worse than the Baseline and VM-Dice networks in terms of the Dice and HD95 metrics (compared with Baseline, no significant difference on HD95 and RV Dice, while average Dice, LV Dice and LVM Dice significantly decreased) used to evaluate registration performance (see columns 2-6 in Table 3.4). Conversely, SDDIR obtains 9% improvement in segmentation accuracy, evaluated using the Dice score, relative to the Baseline network. This is mainly because the registration sub-network in SDDIR is highly-dependent on the segmentation sub-network, to split the original MR images into pairs of sub-regions. Consequently, segmentation errors are propagated to the subsequent registration step and overall registration performance drops significantly when predicted segmentations are of poor quality. Although the SDDIR(-DA) (smooth version of SDDIR) significantly outperforms previous approaches on Dice score, there are significant differences between its clinical indices with the reference. As shown in Figure 3.6, the Demons and B-spline tend to predict warped moving images with over-smoothed image features and object boundaries, losing local details (e.g. the papillary muscles in the left ventricle). The deep learning-based registration methods obtain more such localised anatomical details more consistently.

The quantitative results on the M&M dataset are shown in Table 3.6 (the corresponding P-values are shown in Table 3.7). Similar to the results on ACDC, VM outperforms B-spline, Demons and SyN. VM-Dice and Baseline obtain higher average Dice scores than the traditional methods and VM. Different from ACDC, the addition approaches, VM(add) and VM-Dice(add) achieve better registration performance

Table 3.4: Quantitative comparison on ACDC between SDDIR and the state-of-the-art methods. Statistically significant improvements of our method over previous methods (excluding DDIR) in registration accuracy (DS and HD95) and segmentation performance are highlighted in bold. Besides, LVEDV and LVMM indices with no significant difference from the reference are also highlighted in bold.

Methods	Avg. DS (%)	LVBP DS (%)	LVM DS (%)	RV DS (%)	HD95 (mm)	Seg DS (%)	LVEDV(ml)	LVMM(g)	% of $\lvert J\rvert \leq 0$
before Reg	60.2 ± 11.2	65.7 ± 16.2	51.9 ± 14.5	63.1 ± 14.2	10.6 ± 3.9	-	165.1 ± 73.6	130.1 ± 50.6	-
B-spline	74.9 ± 9.6	80.4 ± 14.2	75.7 ± 7.8	68.8 ± 15.5	10.9 ± 4.7	-	152.5 ± 82.1	134.4 ± 51.4	$1.8e^{-3} \pm 3.5e^{-3}$
Demons	73.5 ± 8.8	78.4 ± 12.6	72.3 ± 8.9	69.9 ± 13.8	10.5 ± 3.9	-	147.0 ± 81.0	137.1 ± 53.4	$1.3e^{-4} \pm 3.3e^{-4}$
SyN	70.1 ± 7.5	79.7 ± 10.1	65.9 ± 8.6	64.6 ± 14.7	10.6 ± 3.5	-	156.1 ± 76.4	132.9 ± 51.5	$2.2e^{-2} \pm 5.3e^{-3}$
VM	76.0 ± 7.8	83.9 ± 10.4	74.2 ± 8.3	70.1 ± 13.9	9.9 ± 4.4	-	156.6 ± 75.9	$\mathbf{130.5 \pm 51.9}$	$9.8e^{-4} \pm 1.0e^{-3}$
VM-Dice	77.8 ± 7.4	84.7 ± 9.9	73.9 ± 8.2	74.7 ± 12.4	8.0 ± 3.9	-	$\mathbf{164.7 \pm 77.2}$	134.8 ± 54.0	$1.7e^{-3} \pm 1.3e^{-3}$
Baseline	78.7 ± 6.9	86.6 ± 9.1	76.7 ± 7.5	72.8 ± 12.9	9.2 ± 3.9	63.8 ± 21.6	161.5 ± 74.9	$\mathbf{131.0 \pm 52.1}$	$1.4e^{-3} \pm 1.4e^{-3}$
VM(add)	78.2 ± 11.2	83.6 ± 15.5	72.2 ± 11.9	78.7 ± 15.9	7.8 ± 4.6	-	159.1 ± 73.9	123.5 ± 53.8	$8.9e^{-4} \pm 7.4e^{-4}$
VM-Dice(add)	77.7 ± 11.4	84.5 ± 14.9	70.3 ± 12.0	78.1 ± 16.4	7.2 ± 4.0	-	161.3 ± 73.9	120.7 ± 53.3	$9.7e^{-4} \pm 8.9e^{-4}$
DDIR	92.6 ± 5.6	94.8 ± 5.9	92.7 ± 5.3	90.2 ± 9.9	4.5 ± 4.9	-	164.3 ± 73.9	130.9 ± 51.9	$3.1e^{-4} \pm 5.2e^{-4}$
SDDIR	77.0 ± 9.7	86.2 ± 10.1	72.9 ± 12.1	71.9 ± 14.5	9.5 ± 3.4	$\mathbf{72.7 \pm 13.7}$	161.0 ± 72.5	123.5 ± 53.4	$4.1e^{-3} \pm 2.0e^{-3}$
SDDIR(-DA)	$\mathbf{79.9 \pm 6.5}$	$\mathbf{87.1 \pm 8.3}$	$\mathbf{77.7 \pm 6.7}$	74.9 ± 12.2	8.4 ± 4.0	70.9 ± 14.9	160.9 ± 75.6	133.8 ± 52.9	$1.2e^{-3} \pm 1.1e^{-3}$

Table 3.5: P-values between SDDIR(-DA) and state-of-the-art methods in Table 3.4. Regarding LVEDV and LVMM, the P-values are computed between the results post-registration and reference. All P-values larger than 0.05 are highlighted in bold.

Methods	Avg. DS (%)	LVBP DS (%)	LVM DS (%)	RV DS (%)	HD95 (mm)	Seg DS (%)	LVEDV(ml)	LVMM(g)	% of $\lvert J\rvert \leq 0$
B-spline	$5.0e^{-22}$	$1.1e^{-16}$	$5.8e^{-9}$	$6.0e^{-19}$	$6.2e^{-21}$	-	$3.5e^{-12}$	$1.7e^{-7}$	
Demons	$1.7e^{-41}$	$1.2e^{-34}$	$2.8e^{-37}$	$9.3e^{-22}$	$1.1e^{-23}$	-	$4.9e^{-21}$	$6.4e^{-14}$	-
SyN	$4.1e^{-82}$	$8.7e^{-54}$	$4.6e^{-77}$	$1.1e^{-48}$	$3.3e^{-32}$	-	$7.4e^{-14}$	$3.0e^{-3}$	
VM	$4.7e^{-48}$	$7.7e^{-31}$	$1.7e^{-39}$	$6.3e^{-27}$	$6.0e^{-17}$	-	$1.0e^{-14}$	**0.55**	
VM-Dice	$2.0e^{-21}$	$1.9e^{-24}$	$3.4e^{-34}$	**0.29**	$3.8e^{-3}$	-	**0.57**	$1.1e^{-6}$	
Baseline	$3.0e^{-24}$	$6.8e^{-6}$	$2.7e^{-11}$	$3.0e^{-18}$	$4.2e^{-12}$	$6.2e^{-18}$	$1.1e^{-6}$	**0.20**	-
VM(add)	$2.5e^{-4}$	$1.5e^{-7}$	$3.6e^{-17}$	$8.8e^{-10}$	$5.4e^{-4}$	-	$1.2e^{-16}$	$5.0e^{-13}$	-
VM-Dice(add)	$1.5e^{-5}$	$2.8e^{-5}$	$3.2e^{-24}$	$1.1e^{-6}$	$3.9e^{-13}$	-	$2.3e^{-19}$	$4.6e^{-20}$	-
DDIR	$1.6e^{-107}$	$8.1e^{-60}$	$9.1e^{-103}$	$7.9e^{-76}$	$1.3e^{-37}$	-	$7.7e^{-5}$	0.01	-
SDDIR	$4.7e^{-7}$	0.04	$4.7e^{-10}$	$2.5e^{-4}$	$1.4e^{-6}$	$1.5e^{-4}$	$1.2e^{-3}$	$2.0e^{-9}$	
SDDIR(-DA)	-	-	-	-	-	-	$1.4e^{-7}$	$3.3e^{-7}$	

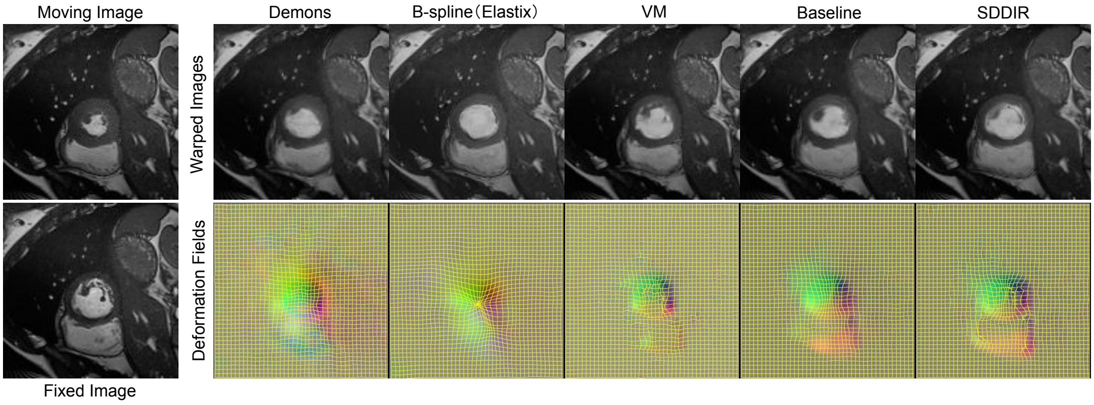

Figure 3.6: Visual comparison of results on ACDC estimated using SDDIR and state-of-the-art methods. Left column: moving and fixed images; Right column: corresponding warped moving images (first row), deformation fields (second row).

than VM and VM-Dice. To our analysis, this is because the MR images in M&M are significantly different from UKBB, and thereby the addition-based methods using ground-truth segmentation would lead to better registration results. DDIR obtains the highest registration performance, while the performance of SDDIR is significantly decreased to 72.12%, due to poor performance of the segmentation sub-network (66.78%). SDDIR(-DA) obtains significantly better results on the Dice score of LVM than the rest approaches (exclude DDIR). Despite the drop in registration performance, SDDIR predicts registered/warped images that show no statistically significant differences from the reference in terms of LVEDV (Paired samples t-test, P-value > 0.05), not afforded by other methods.

The aforementioned results were achieved by directly applying our method to ACDC and M&M, without any fine-tuning steps. To explore the performance of the fine-tuning strategy and further demonstrate the generalisation of our method, we also conducted fine-tuning experiments on ACDC and M&M datasets. We fine-tuned our SDDIR network for 200 epochs using 10, 20, and 40 samples for training, 10 other samples for validation and the remaining samples as the unseen test set. As shown in Figure 3.7, increasing the number of samples used for fine-tuning the network generally improves registration accuracy (average Dice score) for SDDIR on ACDC and M&M. This im-

Table 3.6: Quantitative comparison on M&M between SDDIR and the state-of-the-art methods. Statistically significant improvements of our method over previous methods (excluding DDIR) in registration accuracy (DS and HD95) and segmentation performance are highlighted in bold. Besides, LVEDV and LVMM indices with no significant difference from the reference are also highlighted in bold.

Methods	Avg. DS (%)	LVBP DS (%)	LVM DS (%)	RV DS (%)	HD95 (mm)	Seg DS (%)	LVEDV(ml)	LVMM(g)	% of $\|J\| \leq 0$
before Reg	54.1 ± 10.4	60.4 ± 12.5	44.3 ± 14.2	57.8 ± 10.6	15.0 ± 4.9	-	161.4 ± 63.9	124.4 ± 46.2	-
B-spline	65.6 ± 10.9	70.7 ± 14.9	67.0 ± 9.3	59.1 ± 13.3	18.8 ± 7.3	-	144.6 ± 72.6	121.5 ± 44.8	$3.1e^{-4} \pm 8.6e^{-4}$
Demons	65.8 ± 10.0	71.7 ± 12.9	65.0 ± 9.7	60.5 ± 12.5	16.9 ± 6.7	-	142.3 ± 69.7	**125.2 ± 45.9**	$8.5e^{-4} \pm 2.1e^{-3}$
SyN	58.4 ± 9.9	66.9 ± 13.5	58.4 ± 10.1	49.8 ± 12.9	17.1 ± 5.3	-	138.7 ± 69.2	116.7 ± 44.5	$1.9e^{-2} \pm 6.4e^{-3}$
VM	68.9 ± 8.2	76.7 ± 10.0	67.9 ± 8.6	62.2 ± 11.3	15.7 ± 5.9	-	144.8 ± 65.9	115.4 ± 45.3	$9.3e^{-4} \pm 1.6e^{-3}$
VM-Dice	72.7 ± 7.6	79.3 ± 9.3	67.6 ± 7.9	71.1 ± 11.5	11.8 ± 4.0	-	164.7 ± 66.2	**125.6 ± 46.2**	$2.5e^{-3} \pm 2.3e^{-3}$
Baseline	73.4 ± 7.9	81.5 ± 9.3	70.4 ± 8.1	68.5 ± 11.4	12.6 ± 4.5	59.9 ± 23.2	154.7 ± 63.8	118.6 ± 44.8	$1.6e^{-3} \pm 2.2e^{-3}$
VM(add)	73.0 ± 11.6	78.0 ± 14.5	65.0 ± 12.5	75.9 ± 14.7	10.4 ± 3.9	-	148.5 ± 66.3	109.0 ± 49.3	$1.5e^{-3} \pm 1.1e^{-3}$
VM-Dice(add)	74.4 ± 13.9	81.9 ± 15.9	67.1 ± 11.8	74.2 ± 19.1	11.0 ± 6.9	-	156.9 ± 63.8	110.5 ± 48.6	$1.9e^{-3} \pm 1.4e^{-3}$
DDIR	91.3 ± 6.7	93.8 ± 6.3	90.6 ± 6.2	89.4 ± 10.4	5.6 ± 4.9	-	160.7 ± 64.3	122.5 ± 45.8	$6.7e^{-4} \pm 1.0e^{-3}$
SDDIR	72.2 ± 10.2	80.9 ± 11.5	67.9 ± 10.9	66.7 ± 16.2	12.1 ± 4.1	**66.8 ± 16.4**	**162.1 ± 64.4**	120.9 ± 43.7	$6.1e^{-3} \pm 3.1e^{-3}$
SDDIR(-DA)	75.0 ± 7.6	82.4 ± 8.8	**72.1 ± 7.5**	70.5 ± 11.6	12.1 ± 4.5	65.1 ± 18.9	159.2 ± 65.8	122.8 ± 45.9	$2.0e^{-3} \pm 3.3e^{-3}$

Table 3.7: P-values between SDDIR(-DA) and state-of-the-art methods in Table 3.6. Regarding LVEDV and LVMM, the P-values are computed between the results post-registration and reference. All P-values larger than 0.05 are highlighted in bold.

Methods	Avg. DS (%)	LVBP DS (%)	LVM DS (%)	RV DS (%)	HD95 (mm)	Seg DS (%)	LVEDV(ml)	LVMM(g)	% of $\|J\| \leq 0$
B-spline	$2.0e^{-58}$	$2.9e^{-49}$	$1.0e^{-31}$	$1.9e^{-52}$	$9.9e^{-56}$	-	$2.6e^{-21}$	$1.1e^{-4}$	-
Demons	$3.3e^{-80}$	$5.5e^{-68}$	$1.2e^{-59}$	$6.7e^{-61}$	$1.4e^{-48}$	-	$1.1e^{-29}$	**0.34**	-
SyN	$8.9e^{-100}$	$1.1e^{-66}$	$9.9e^{-93}$	$3.5e^{-85}$	$1.4e^{-52}$	-	$4.9e^{-24}$	$5.9e^{-17}$	-
VM	$6.9e^{-79}$	$2.3e^{-63}$	$1.1e^{-49}$	$3.0e^{-60}$	$1.7e^{-39}$	-	$3.0e^{-34}$	$2.6e^{-22}$	-
VM-Dice	$1.3e^{-27}$	$8.0e^{-37}$	$4.0e^{-51}$	0.03	0.03	-	$6.3e^{-4}$	**0.22**	-
Baseline	$1.8e^{-25}$	$1.5e^{-8}$	$3.2e^{-20}$	$3.0e^{-20}$	$3.5e^{-5}$	$1.3e^{-18}$	$9.1e^{-17}$	$1.5e^{-11}$	-
VM(add)	$4.1e^{-7}$	$4.2e^{-17}$	$1.7e^{-28}$	$1.2e^{-25}$	$3.0e^{-10}$	-	$7.1e^{-30}$	$4.7e^{-36}$	-
VM-Dice(add)	**0.13**	**0.21**	$1.4e^{-18}$	$3.7e^{-7}$	$2.9e^{-4}$	-	$6.5e^{-26}$	$1.5e^{-37}$	-
DDIR	$1.6e^{-183}$	$3.2e^{-122}$	$2.7e^{-146}$	$1.7e^{-148}$	$4.6e^{-94}$	-	$4.4e^{-4}$	$5.1e^{-8}$	-
SDDIR	$6.1e^{-8}$	0.01	$9.2e^{-10}$	$7.1e^{-5}$	**0.49**	$7.5e^{-4}$	**0.55**	$3.7e^{-4}$	-
SDDIR(-DA)	-	-	-	-	-	-	$7.1e^{-4}$	0.04	-

provement in registration performance is more pronounced for M&M than ACDC, with the latter showing only marginal improvements when the number of training samples used for fine-tuning is increased from 20 to 40. Furthermore, for ACDC, the results fine-tuned with 20 samples show no significant difference in registration performance from those fine-tuned with 40 samples (P-value=0.26). This demonstrates that the fine-tuning strategy can be applied to those unseen data when the performance of our SDDIR is sub-optimal, only requiring limited samples (∼20).

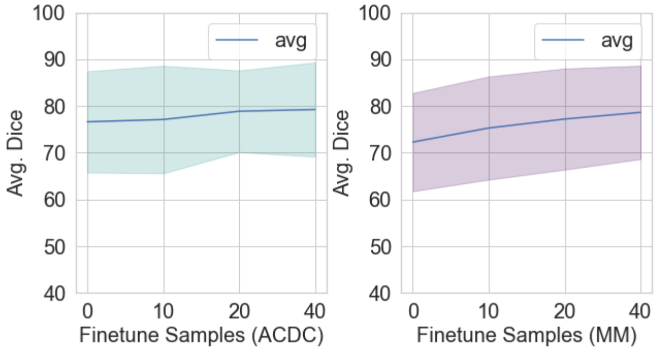

Figure 3.7: Fine-tuning experiments on ACDC and M&M.

Segmentation Analysis

In this section, we analyse the segmentation performance of SDDIR and the Baseline network. Examples of segmentation masks predicted using either approach are shown in Figure 3.8, and quantitative results summarising the segmentation accuracy of both approaches are presented in Table 3.2, Table 3.4 and Table 3.6. The segmentation results obtained for the UKBB, summarised in Table 3.2, show that SDDIR achieves 1% higher Dice score (significantly better, P-value=$3.3e^{-25}$) than the Baseline network across the unseen test data. This improvement in segmentation accuracy is more pronounced for the ACDC and M&M datasets, with SDDIR achieving $>$ 6% improvement in the Dice score for both datasets, relative to the Baseline network (72.72% vs 63.80% and 66.78% vs 59.94%). These results demonstrate that SDDIR significantly outperforms the Baseline network in terms of segmentation accuracy (for the input fixed and moving images), consistently across multiple datasets.

Figure 3.8 highlights the ability of both the Baseline network and SDDIR to predict

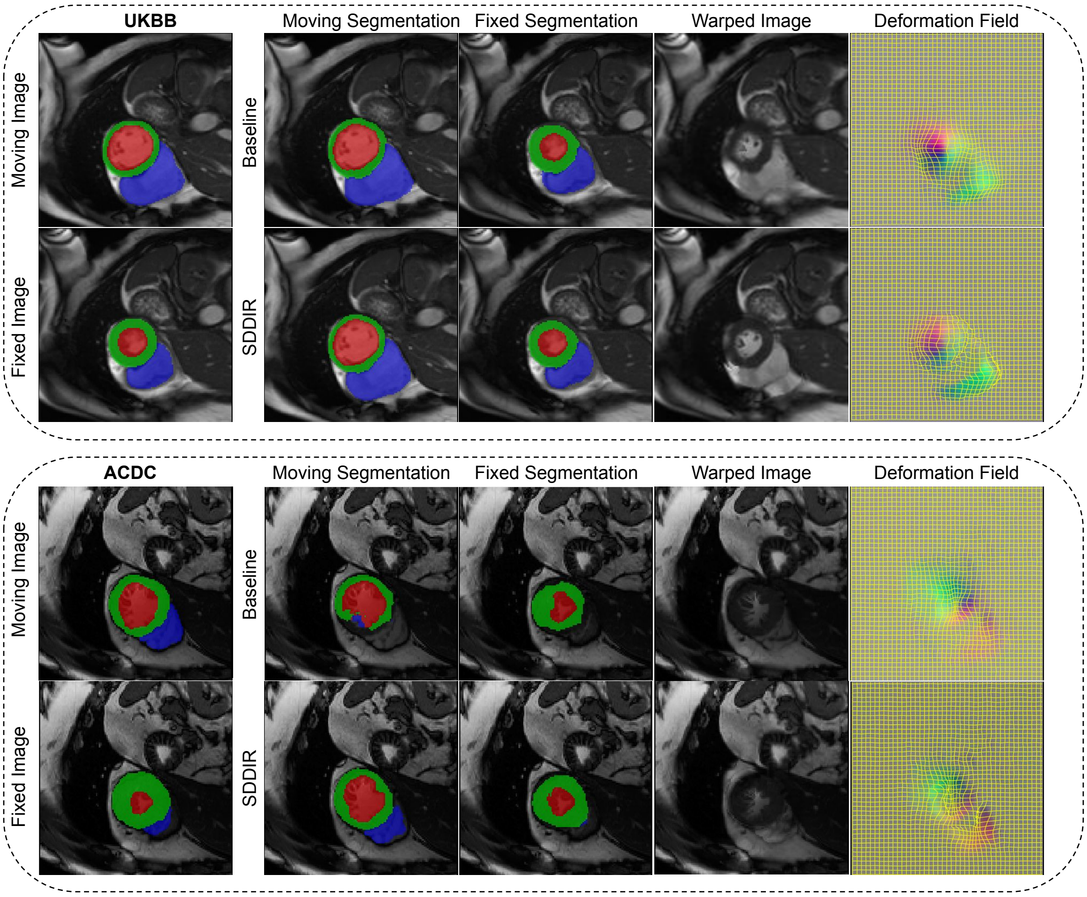

Figure 3.8: Visual comparison of segmentation estimated using SDDIR and Baseline. The first two rows are the results on UKBB and the bottom two rows are the results on ACDC. Left column: moving and fixed images with the corresponding segmentation; Right column: corresponding moving segmentation, fixed segmentation, warped moving image, and deformation fields. The cardiac MR images were reproduced by kind permission of UK Biobank ©.

high-quality segmentation masks for the input image pairs (fixed and moving) from the UKBB dataset. Their performance on data from ACDC however, degrades, producing masks that are visually different to the ground-truth (e.g. the segmentation of the right ventricle, plotted in blue in Figure 3.8). In the UKBB dataset, the results of SDDIR are more similar to the ground-truth, but there is over-segmentation in the results of the Baseline (see the region of the right ventricle in the moving segmentation). There is a domain gap from the UKBB to the ACDC dataset, which leads to the decreased segmentation Dice for both methods. However, SDDIR offers some improvements over the Baseline, for example, by capturing the right ventricle in cases where it is entirely missed by the latter. In summary, due to the co-attention block, SDDIR is more robust (than the Baseline) for segmenting the input pairs of images in the presence of domain shifts.

Ablation Study

To analyse the contribution of each block in SDDIR, we conducted an ablation study on the proposed network, using UKBB data. The results are shown in Table 3.8 (the corresponding P-values are shown in Table 3.9), where, SDDIR(-DA), SDDIR(-CA), SDDIR(-Seg) and SDDIR(-Reg) denote removing the discontinuity addition block, co-attention block, segmentation sub-network and registration sub-network in SDDIR, respectively. By comparing these variants of SDDIR, we can assess the contribution of the proposed joint segmentation and registration sub-networks, co-attention block and discontinuity-preserving strategy. Without the discontinuity addition block, the SDDIR(-DA) is essentially a globally smooth registration method. SDDIR(-CA) applies the same segmentation sub-network as Baseline, whilst still ensuring discontinuity-preserving registration. To compare the performance of co-attention and self-attention, we also build an SDDIR variation with self-attention [210] in the segmentation sub-network, denoted as SDDIR(SA). By removing the segmentation sub-network, the SDDIR(-Seg) turns into a registration network similar to VM-Dice. Correspondingly, the SDDIR(-Reg) is a sole co-attention based segmentation network. To explore the contribution of joint segmentation and registration in registration performance, we further compare the performance of SDDIR with the results combining DDIR and automatic segmentation approaches, denoted as DDIR(+CoSeg) and DDIR(+SDDIR). The former utilises the predicted segmentation masks from SDDIR(-Reg) as the input segmentation masks of

DDIR, while the latter feeds the predicted segmentation masks from SDDIR to DDIR.

Table 3.8: Quantitative comparison on UKBB between different versions of SDDIR. Statistically significant improvements in registration and segmentation accuracy (DS and HD95) are highlighted in bold. Besides, LVEDV and LVMM indices with no significant difference from the reference are also highlighted in bold.

Methods	Avg. DS (%)	LVBP DS (%)	LVM DS (%)	RV DS (%)	HD95 (mm)	Seg DS (%)	LVEDV(ml)	LVMM(g)	% of $\|J\| \leq 0$
before Reg	43.8 ± 5.5	63.1 ± 14.2	52.9 ± 13.1	68.1 ± 15.4	15.4 ± 4.6	-	157.5 ± 32.8	99.5 ± 27.9	-
SDDIR(-DA)	80.5 ± 4.3	88.5 ± 5.0	74.2 ± 6.3	78.6 ± 6.6	9.4 ± 4.1	88.2 ± 2.3	156.4 ± 33.4	**98.9 ± 28.4**	$2.6e^{-3} \pm 1.4e^{-3}$
SDDIR(-CA)	85.7 ± 3.9	91.4 ± 4.5	80.7 ± 5.7	84.9 ± 6.2	7.5 ± 4.8	87.2 ± 2.7	159.0 ± 33.2	**99.3 ± 27.1**	$8.9e^{-3} \pm 3.4e^{-3}$
SDDIR(SA)	86.4 ± 3.6	91.3 ± 4.3	81.5 ± 5.6	86.3 ± 5.5	7.3 ± 4.9	87.6 ± 2.3	156.3 ± 32.8	**98.8 ± 27.0**	$2.8e^{-3} \pm 1.8e^{-3}$
SDDIR(-Reg)	-	-	-	-	-	87.4 ± 4.3	-	-	-
SDDIR(-Seg)	79.8 ± 4.4	86.9 ± 5.8	71.7 ± 6.8	80.7 ± 5.9	8.7 ± 4.6	-	162.1 ± 34.4	**100.4 ± 28.6**	$4.1e^{-3} \pm 2.0e^{-3}$
DDIR(+CoSeg)	84.1 ± 4.8	89.7 ± 5.5	79.0 ± 6.0	83.6 ± 6.8	8.3 ± 4.7	-	**157.5 ± 32.9**	**100.0 ± 26.3**	$3.8e^{-4} \pm 4.2e^{-4}$
DDIR(+SDDIR)	84.8 ± 4.6	90.6 ± 5.2	79.7 ± 5.9	84.0 ± 6.5	8.0 ± 4.7	-	156.2 ± 32.9	**99.3 ± 26.3**	$4.0e^{-4} \pm 4.4e^{-4}$
SDDIR	**87.7 ± 3.4**	**92.7 ± 3.9**	**83.1 ± 5.2**	**87.2 ± 5.5**	**6.7 ± 4.9**	**88.6 ± 2.1**	**157.6 ± 32.9**	**98.7 ± 26.1**	$3.5e^{-3} \pm 1.6e^{-3}$

Table 3.9: P-values between SDDIR and state-of-the-art methods in Table 3.8. Regarding LVEDV and LVMM, the P-values are computed between the results post-registration and reference. All P-values larger than 0.05 are highlighted in bold.

Methods	Avg. DS (%)	LVBP DS (%)	LVM DS (%)	RV DS (%)	HD95 (mm)	Seg DS (%)	LVEDV(ml)	LVMM(g)	% of $\|J\| \leq 0$
SDDIR(-DA)	$3.0e^{-167}$	$2.1e^{-77}$	$8.8e^{-151}$	$9.5e^{-128}$	$4.8e^{-55}$	$1.8e^{-16}$	0.01	**0.29**	-
SDDIR(-CA)	$2.7e^{-77}$	$1.6e^{-23}$	$9.5e^{-71}$	$1.4e^{-46}$	$2.2e^{-13}$	$2.0e^{-66}$	$1.1e^{-6}$	**0.68**	-
SDDIR(SA)	$4.9e^{-54}$	$2.4e^{-31}$	$3.1e^{-47}$	$1.2e^{-13}$	$7.4e^{-8}$	$1.1e^{-58}$	0.01	**0.13**	-
SDDIR(-Reg)	-	-	-	-	-	$2.4e^{-14}$	-	-	-
SDDIR(-Seg)	$9.9e^{-176}$	$3.3e^{-99}$	$2.7e^{-168}$	$1.2e^{-107}$	$3.0e^{-35}$	-	$9.1e^{-23}$	**0.08**	-
DDIR(+CoSeg)	$4.1e^{-96}$	$5.4e^{-55}$	$3.6e^{-83}$	$6.5e^{-66}$	$2.6e^{-29}$	-	**0.99**	**0.29**	-
DDIR(+SDDIR)	$7.2e^{-91}$	$1.3e^{-41}$	$9.3e^{-75}$	$9.3e^{-70}$	$1.0e^{-27}$	-	$1.0e^{-3}$	**0.65**	-
SDDIR	-	-	-	-	-	-	**0.75**	**0.13**	-

According to Table 3.8, self-attention improves performance with respect to conventional convolutional segmentation network (SDDIR(SA) vs SDDIR(-CA)) but performs worse than co-attention (SDDIR). For LVEDV, all three approaches are close to the reference, while only the results of SDDIR make no significant difference to the reference (P-value>0.05). When considering a 0.01 significance level, SDDIR and SDDIR(SA) both show no significant difference relative to the reference. In addition, regardless of which segmentation sub-network is used, their LVMM all make no significant differences to the reference (P-value>0.05). Therefore, it can be found that the registration sub-network plays a central role in preserving the clinical indices post-image registration, while a better segmentation sub-network would lead to better clinical indices. After removing the discontinuity addition block, the average registration Dice score of SDDIR(-DA) is significantly decreased, as it is unable to ensure globally discontinuous and locally smooth deformation fields. Comparing SDDIR with the corresponding variants of the network that tackle purely segmentation (SDDIR(-Reg)) and registration tasks (SDDIR(-Seg)), we find that the joint segmentation and registration framework improves the performance of each sub-network for each corresponding task. This indicates that the two tasks are mutually beneficial to each other. With the ground-truth segmentation as input, DDIR achieves state-of-the-art registration performance, while it may significantly decrease when using automatic segmentation as the input. Comparing the results of DDIR(+CoSeg) and SDDIR, it can be found that, using the same separate segmentation network as SDDIR (co-attention based segmentation network) to predict segmentation for DDIR leads to significantly worse performance than the performance of joint segmentation and registration. Even with the same segmentation masks as the input for registration (DDIR(+SDDIR)), the registration accuracy of DDIR is significantly lower than SDDIR, which highlights the superiority of joint segmentation and registration.

Although the overall registration performance of SDDIR is worse than DDIR, the former outperforms other state-of-the-art methods, and ensures that cardiac clinical indices derived from the warped/registered images show no statistically significant differences to the reference (derived from the original images). Furthermore, SDDIR does not require high-quality segmentation masks to be available a priori for the input images to be registered, unlike DDIR (which was trained and evaluated using segmentation masks delineated manually by experts). SDDIR thus lends itself to use real clinical applications where high-quality segmentation masks are seldom available, and has the added benefit of producing good quality segmentation masks for both input images to be registered, as auxiliary outputs. Although the globally smooth version of SDDIR (i.e. without discontinuity addition), SDDIR(-DA), incurs significantly higher registration errors than SDDIR for the UKBB data set, it consistently outperforms other state-of-the-art approaches in terms of registration accuracy and joint registration and segmentation performance, across all three datasets. Considering those scenarios where globally smooth registration is required/appropriate, SDDIR(-DA) can be employed in place of SDDIR to reduce the dependency of registration accuracy on segmentation quality, and improve overall registration performance. For example, results reported for ACDC and M&M data sets in Tables 3.4, 3.6 show that SDDIR(-DA) outperforms SDDIR in terms of registration accuracy, at the cost of enforcing global smoothness on the estimated deformation fields.

3.2.5 Discussion and Conclusion

The proposed approach, SDDIR, is versatile and can be employed in various clinical applications requiring pair-wise image registration. For example, the ability to simultaneously segment and register cardiac MR images means that SDDIR can facilitate real-time quantification of cardiac clinical indices across the cardiac cycle and quantitative analysis of cardiac motion. In addition, as SDDIR can predict both fixed segmentation and warped moving segmentation, a more anatomical structure-plausible segmentation results can be obtained by jointly considering those two predictions. SDDIR is agnostic to image modality and the organs/structures visible within the field of view of the images being registered. Hence, SDDIR may also be used to jointly segment and register thoracic or abdominal computed tomography images, where strong discontinuities exist between organ structures due to their relative motion (i.e. sliding at organ boundaries) resulting from respiration.

Although the proposed approach is demonstrated to jointly segment and register input pairs of images accurately, outperforming several state-of-the-art approaches, one main limitation remains. The registration performance of SDDIR is highly dependent on the performance of the segmentation sub-network, i.e. on the quality of the segmentation masks predicted for the input pair of images to be registered. As the registration sub-network requires the predicted segmentation masks to split the original MR images into pairs of corresponding regions, the registration sub-network performs poorly when the quality of the segmentation masks is poor. Thus, SDDIR performs well on the UKBB dataset as accurate segmentation masks are predicted for the input pairs of images and used to effectively guide the discontinuity-preserving registration. As SDDIR was trained using UKBB data, the segmentation sub-network was able to generalise well to unseen data from UKBB due to homogeneity/consistency in appearance across images from different subjects. Conversely, SDDIR's performance significantly degraded when the trained model (on UKBB data) was used to register images from other datasets (e.g. ACDC and M&M). This is due to domain shifts resulting from variations in imaging scanners and protocols, used to acquire images in different datasets. The presented results indicate that SDDIR outperforms the Baseline network in terms of registration accuracy only when good quality segmentation masks are predicted by its constituent segmentation sub-network. Specifically, we found that segmentation accuracy (in terms of Dice) on the fixed and moving images had to be over 76% for ACDC and 70% for M&M, for the subsequent registration accuracy of SDDIR to be better than the Baseline network. Therefore, future work in the field should look to improve the robustness of discontinuity-preserving image registration methods to domain shifts that are commonly found in medical images. This may be achieved by imbuing SDDIR with recent approaches to domain generalisation, for example, to mitigate the drop in segmentation and registration performance resulting from domain shifts (relative to the training data). Additionally, the over-dependence of registration quality on the quality of the segmentation masks predicted by SDDIR could be relaxed by modelling object/tissue boundaries as weak discontinuities (as opposed to strong discontinuities used currently in SDDIR) that are incorporated into the regularisation of the deformation field to ensure locally smooth and globally discontinuous deformation fields.

In this section, we propose a novel weakly-supervised discontinuity-preserving registration network, SDDIR. The proposed approach is applied to the task of intra-patient spatio-temporal cardiac MR registration, to jointly segment the input pair of images to be registered and predict a locally smooth but globally discontinuous deformation field that warps the source/moving image to the fixed/target image. Compared with previous discontinuity-preserving registration methods, SDDIR provides improvements in terms of execution speed (relative to traditional iterative approaches), and does not require segmentation masks to be available prior to registering images (unlike some state-of-the-art deep learning-based registration approaches such as DDIR). We demonstrate the registration performance of SDDIR on three cardiac MR datasets, and prove that it can significantly outperform both traditional and deep learning-based state-of-the-art registration methods. Future works will explore domain generalisation techniques to mitigate the drop in performance observed with SDDIR due to domain shifts and will look to weaken the dependency on segmentation quality to ensure accurate image registration.

3.3 Conclusion

In this chapter, two deep learning-based cardiac image registration methods are proposed to achieve discontinuity-preserving registration.

We first proposed a novel weakly-supervised discontinuity-preserving registration network, DDIR, which significantly outperformed the state-of-the-art, in intra-patient CMR registration. DDIR preserves LV clinical indices post-registration, not afforded by the other approaches. This makes it compelling as a tool for use in clinical applications as it ensures that common diagnostic biomarkers for LV are preserved post-registration.

While DDIR can achieve significant improvements over previous traditional and deep learning-based registration methods, it requires segmentation masks on both training and inference, limiting its application in realistic scenarios. To tackle this issue, we further proposed a joint segmentation and registration network, SDDIR to achieve segmentation and discontinuity-preserving registration simultaneously. Given the input moving and fixed images as input, our SDDIR can predict corresponding accurate segmentation masks and deformation fields at the same time.

Image segmentation and registration are both fundamental tasks in cardiac image analysis and CVD prediction/diagnosis. Using the segmentation masks predicted by our proposed method, we can compute clinical indices and reconstruct corresponding cardiac meshes. In the next chapter, we will introduce learning accurate cardiac shape representations from cardiac MR images/contours.

Chapter 4

Deep Learning in 3D Cardiac Shape Reconstruction

Shape reconstruction from sparse point clouds/images is a challenging and relevant task required for a variety of applications in computer vision and medical image analysis (e.g. surgical navigation, cardiac motion analysis, augmented/virtual reality systems). A subset of such methods, viz. 3D shape reconstruction from 2D contours, is especially relevant for computer-aided diagnosis and intervention applications involving meshes derived from multiple 2D image slices, views or projections. We propose a deep learning architecture, coined Mesh Reconstruction Network (MR-Net), which tackles this problem. MR-Net enables accurate 3D mesh reconstruction in real-time despite missing data and with sparse annotations. Using 3D cardiac shape reconstruction from 2D contours defined on short-axis cardiac magnetic resonance image slices as an exemplar, we demonstrate that our approach consistently outperforms state-of-the-art techniques for shape reconstruction from unstructured point clouds. Our approach can reconstruct 3D cardiac meshes to within 2.5-mm point-to-point error, concerning the ground-truth data (the original image spatial resolution is $\sim 1.8 \times 1.8 \times 10mm^3$). We further evaluate the robustness of the proposed approach to incomplete data, and contours estimated using an automatic segmentation algorithm. MR-Net is generic and could reconstruct shapes of other organs, making it compelling as a tool for various applications in medical image analysis.

4.1 Introduction

Reconstructing plausible 3D shapes (represented as parametric surface meshes) from sparse, unstructured point clouds (PCs) extracted from single- or multi-view images, is an active problem in computer vision (CV) and medical image analysis. 3D shape reconstruction helps visualise the spatial structure of 3D objects, and is relevant to several applications such as, computer-aided diagnosis, surgical planning, image-guided interventions, and computational simulations, to name a few [211, 212].

Generally, traditional cardiac shape reconstruction comprises two steps: (1) cardiac image segmentation; and (2) mesh generation from the estimated segmentations. Cardiac image segmentation (manual/automatic segmentation) aims to find the region of interest in the original magnetic resonance (MR)/computed tomography (CT) images (e.g. left ventricle (LV), right ventricle (RV)). The mesh generation process then takes the segmentation results as input and generates the corresponding meshes. Marching Cubes [213] is the most widely used algorithm for generating meshes from segmented image volumes, but generally requires dense segmentation volumes for reconstructing 3D shapes as triangulated surfaces/meshes. Such an approach is ill-suited to reconstructing 3D shapes from sparse, stacked 2D contours. Therefore, in cardiac shape reconstruction, previous studies have approached the problem as one of mesh adaptation. In this context, a template mesh is first generated (either using an isosurfacing technique or directly from an existing statistical atlas/template), and then deformed under the guidance of contours or points (extracted from segmentation results) [214, 215, 216, 217]. Using segmented contours to deform the template mesh, [215] proposed to reconstruct specific 4D meshes (spatial-temporal mesh) for patients. [217] proposed a method to reconstruct geometrical surface meshes from sparse, heterogeneous, non-coincidental contours. They used contours to guide the deformation of an initial mesh to obtain the target mesh, using a smoothness term while maximising the data fitting. However, those methods are all time-consuming, which limits mesh reconstruction for real-time applications in surgical guidance and navigation.

Deep learning-based methods have also been explored for this task. As inference using a trained deep neural network is just one forward pass through the network, such methods can significantly speed up the process of cardiac shape reconstruction. Few studies have explored the application of deep learning methods on this task. For example, [218] proposed to tackle this task as a volumetric mapping problem followed by isosurface estimation using the generated volume. Their approach generated three dense 3D volumes, LV myocardium, LV cavity and RV cavity, from sparse volumes of contours. Then marching cube was used to reconstruct the bi-ventricle cardiac meshes. This approach was able to accurately predict cardiac meshes even with discrepancies between intersecting slices (short-axis (SAX) view and long-axis (LAX) view slices). [219] viewed shape reconstruction as a regression problem, building a deep regression network to predict the cardiac shape parameters in Principal Component Analysis (PCA) space from image data (from the UK Biobank (UKBB) cohort), using both short and long axis views and patient metadata. Using a cardiac statistical shape model (SSM) estimated *a priori* and its associated mean template mesh and principal eigenvectors, during inference, they reconstructed the bi-ventricle cardiac meshes for each unseen image volume using the PCA parameters estimated by their network. Instead of using traditional methods to generate the final shape, some studies have proposed to predict cardiac PCs [220] or meshes [221] directly using deep neural networks, enabling cardiac shape reconstruction in real-time. [220] firstly proposed to apply a deep learning network in cardiac point cloud reconstruction, which could reconstruct RV from a single image (in the LAX view). Similarly, [221] designed a deep learning network, Instantiation-Net, to reconstruct 3D RV mesh based on a single LAX view image. However, reconstructing a 3D object from the image in a single view is ill-posed due to the large proportions of missing information, making it difficult to generate accurate meshes.

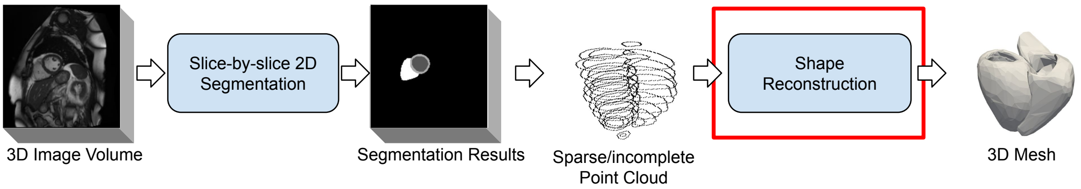

Figure 4.1: The proposed pipeline for 3D cardiac shape reconstruction from MR images (The cardiac image presented were reproduced with the permission of UK Biobank ©). Note that, the slice-by-slice segmentation methods can be both manual segmentation and automatic segmentation algorithms.

To reconstruct plausible and high-quality meshes from cardiac images, multiple images with boundary information (e.g. contours) would be a better input choice. They are usually available from manual/semi-automatic contours derived from most medical image segmentation tools. As those tools do not provide full 3D reconstructions, the mesh reconstruction method could be a supplement of these tools in return. Previous research [214, 215, 216, 217] have also proved that deforming a template mesh under the guidance of contours facilitates the generation of high-quality personalised meshes (fitted to the contours). Therefore, in this chapter, we focus on cardiac mesh reconstruction from a point cloud of contours. We design a novel approach, MR-Net, to achieve the task of reconstructing 3D bi-ventricle cardiac shapes from stacked 2D contours, viewing it as a DL-based template-to-PC fitting task. An overview of the proposed framework is presented in Figure 4.1. Given SAX cine-cardiac MR image stacks, we first manually/automatically segment the cardiac structures of interest in each 2D slice. Next, PCs of stacked contours are extracted from these segmentations. Finally, MR-Net is applied to predict high-quality meshes from PCs of contours. With deep learning-based segmentation methods and MR-Net, we can reconstruct accurate 3D cardiac shapes from the MR images accurately, robustly, and in real-time.

Recently, many deep learning-based methods have been proposed for meshes/PCs reconstruction and analysis [222, 223, 224]. Among them, the most popular task is to reconstruct 3D mesh from single-/multi-view image(s). [224] firstly proposed a network Pixel2mesh based on graph convolutional network (GCN) [225, 226, 227] for mesh reconstruction from a 2D image (a projection of the original 3D object on to one view). They used an ellipsoid mesh as the template, then applied the GCN blocks to deform it with the guidance of features extracted from the input image using VGG 16-like architecture [32]. Based on it, [228] proposed an improved network, Pixel2mesh++, to tackle the problem of 3D mesh reconstruction from multi-view images, reconstructing more accurate surfaces of 3D objects. Instead of GCN, [229] proposed to apply a multilayer perceptron (MLP) as the deformation module followed by topology modification blocks, and finally designed a boundary refinement block to improve the visual quality of reconstructed meshes further. These approaches were developed and validated on publicly available datasets for the reconstruction of general objects (e.g. plane, chair). Using approaches like Pixel2mesh, recent studies have also explored the reconstruction of human hand [230] or body [231] meshes from 2D images. In addition, several deep learning-based methods have been proposed for mesh reconstruction from dense PCs, which rely on predicting the surface normal vector for every point in the input PCs [232], or predicting the skinned multi-person linear model (SMPL, i.e. a parametric human body model [233]) parameters of the target mesh, then using the off-the-shelf SMPL model to reconstruct meshes from parameters [234]. However, in our case, the input PCs are sparse contours with large proportions of missing information relative to dense point cloud-based representations of shapes. And these contour points differ in number and spatial distribution to the vertices of the surface (our target/output) that they implicitly represent.

To this end, considering the nature of the traditional cardiac mesh reconstruction methods and the context of deep learning-based mesh reconstruction methods, we propose to use a deep learning network to deform a cardiac template mesh to obtain the target meshes under the guidance of contours. The key idea behind mesh reconstruction from single/multiple images is to find a mapping from the input image(s) to the template mesh, and subsequently, to use the learnt features in the input image to guide the deformation of the template mesh. Generally, a 2D projective transformation is applied to find the corresponding pixels in the 2D image for every vertex in the template mesh, before transferring the features of 2D pixels to the corresponding vertex. However, in our case, the inputs are contour PCs in 3D coordinate space. Applying a single 2D projection of the input PCs would cause a loss of structural information. Therefore, we design a PC-to-PC mapping going from the 3D contour point cloud to a 3D volume, and correspondingly, from the 3D volume to the vertices of the 3D template mesh (i.e. a PC-volume-PC mapping), which addresses the challenge of mapping features between unstructured data sets that lack spatial correspondence.

The main contribution of our work is a hybrid graph convolutional neural network for 3D mesh reconstruction, MR-Net, which approaches the problem as a template deformation task conditioned on the sparse point cloud data (stacked 2D contours in our case). To the best of our knowledge, this is the first study to employ deep learning for registering a 3D mesh to sparse PCs (or stacked 2D contours), enabling real-time 3D shape reconstruction. Although we focus on 3D cardiac shape reconstruction from stacked 2D contours in this study, MR-Net is generic and flexible, and can be employed for various PC-to-PC/mesh reconstruction tasks (e.g. PC/mesh reconstruction, PC/mesh completion and correction) within the medical imaging or CV domain. To sum up, the contributions of this chapter are as follows,

- We propose a novel cardiac mesh reconstruction framework, which can predict accurate cardiac meshes from original MR images in a fast and robust manner, assisted by existing deep learning-based segmentation methods.
- We demonstrate that MR-Net can generate accurate and high-quality meshes even from incomplete contours, a challenge that often arises in clinical scenarios.

The rest of the chapter is organised as follows: In Section 4.2, each component of the proposed approach is described. Section 4.3 exemplifies our proposed MR-Net on UKBB dataset. Finally, Section 4.4 is the conclusion of this chapter.

4.2 Method

Traditional 3D shape reconstruction approaches have relied on iterative deformation of a template mesh to sparse contours/PC, using the latter to guide the former, with various penalty terms to ensure the estimated deformation is smooth. To eliminate the requirement of several iterations during inference (which can be time-consuming), in this chapter, a deep learning-based network, MR-Net, is designed to mimic such a process. After training, unseen contours/PCs are reconstructed into 3D shapes (represented as triangulated surface meshes) via a simple forward pass through the network. This can significantly speed up 3D shape reconstruction while predicting high-quality meshes. In subsequent sub-sections, we first introduce the overall network architecture of MR-Net, and then provide details of — the feature extraction module, deformation module, 3D PC-to-PC mapping, and the loss function formulated for effective training of the proposed approach.

4.2.1 Network Architecture

The task of our MR-Net is to reconstruct personalised meshes from sparse contours under the guidance of a template mesh. To accomplish this, we design two modules: the feature extraction module and the deformation module (comprising three GCN blocks), as shown in Figure 4.2. The purpose of the feature extraction module is to extract features from the input point cloud of stacked contours that are beneficial for the deformation module, while the latter utilises this information to deform the template mesh to the personalised target mesh under the guidance of the features from the feature extraction module. The feature extraction module consists again of two parts: direct PC feature extraction (in PC domain), and 3D convolutional neural network (CNN [235]) feature extraction (in image domain). The former is to extract features directly from point clouds whereas the latter extracts features from a voxel-based representation of the contours.

Generally, the meshes can be presented by vertices and connectivity. Following Pixel2mesh [224], we assume the connectivity in the target meshes is fixed (the same as template mesh), and thereby the mesh reconstruction from PC could be simplified to learn the mapping between input PC and vertices of target meshes. To achieve this mapping, two problems must be addressed: (1) how to learn the shape priors from the input PC; (2) how to find the point-to-point correspondence between the input PC and vertices of the template, in order to apply the graph convolution. The main contributions of our proposed MR-Net lie to tackle these two challenges.

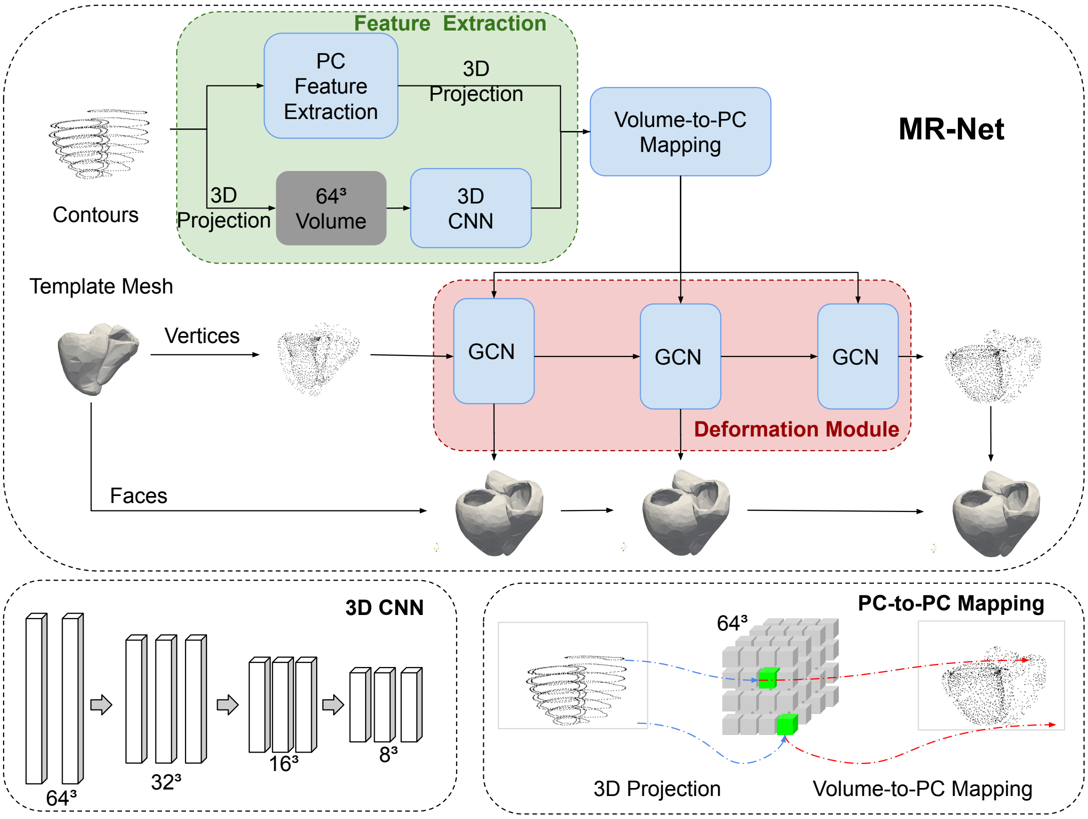

Figure 4.2: Schema of the proposed method, MR-Net. The overall network is displayed in the top row, with the details of the 3D CNN and our proposed PC-to-PC mapping blocks (between input PCs and vertices of template mesh) presented in the bottom row. The feature extraction module extracts features from the input contours, and then the deformation module deforms the template mesh to the target mesh under the guidance of the learnt features in the features extraction module.

Feature Extraction

Due to the large proportion of missing inter-slice information, 3D shape reconstruction from sparse 2D contours is a challenging task. A template mesh is randomly selected from the training dataset to supply the missing information in the reconstruction process. The input PCs of contours serve as the guidance of template deformation. All the input PCs and corresponding target meshes are normalised to a standard sphere (centred at $(0, 0, 0)$ with a radius of 1) before training the network. To learn the guidance information, feature extraction from the input contours is decomposed into two paths.

The first path is a point cloud feature extraction block based on PointNet++ [223], which predicts two new PCs using sampling and grouping. In our experiments, the number of points in input PCs is 3,000, and these two new PCs contain 2,000 and 1,578 points respectively (the number of points is set empirically, sampling and grouping the original point clouds of contours gradually from 3,000 to 1,578). After obtaining these two new PCs, a 3D projection (i.e. a mapping from vertices' coordinate to index of voxels, projecting points in 3D space to voxels in 3D volume, see in Formula. 4.2) is applied to transfer them with the original point cloud of contours into three 64^3 features, where each voxel is a feature vector with dimension 1×4.

In the other path, we first apply a 3D projection to turn the unstructured input point cloud into a structured volume with 64^3 voxels in the image domain. Then a 3D CNN (4 layers, downsampling from 64^3 to 8^3) is used to extract features from the 3D volume projected from the input point cloud, where the extracted features contain feature maps in all four resolutions (64^3,32^3,16^3,8^3), where corresponding feature dimensions are 64, 128, 256, 500 respectively.

With volume-to-PC mapping, we can map the features in voxels of volumes back to points in the template mesh and guide its deformation. Therefore, we finally obtain a feature of $(64 + 128 + 256 + 500 + 4 \times 3) = 960 \times 1$ for every point in the vertices of template mesh, which is concatenated with the coordinate (3×1) of the template mesh (or the coordinate predicted in the previous GCN block) and taken as input by the GCN blocks. Although it would cause little information missing in the process of 3D projection, the multi-layer 3D CNN learns rich structured features (across different resolutions) from the original PCs, which is essential for extracting features from the input PCs. With feature extraction in both the point cloud domain and the image domain, we can obtain a proper understanding of the input contours and use it to guide the deformation of the template mesh.

Deformation Module

With the features learnt from input contours as guidance, we design a deformation module to deform the template mesh gradually, which helps to preserve the topology and the connectivity of meshes, following deformation. The deformation module includes three GCN blocks (referring to Pixel2mesh [224]), each comprising 14-15 graph convolution layers (the first is 14, while the next two are 15). Note that, the number of layers in MR-Net is set empirically and tuned based on results obtained on the validation set.

3D meshes comprise vertices, edges and faces. The vertices are the coordinates of the nodes on the mesh, which is generally an $N \times 3$ array (the three columns stand for x,y,z coordinates respectively). Edge denotes the connectivity between two vertices. In our case, the face of the mesh is defined by surface triangles, whereby, every face in the mesh comprises the indices of three vertices (connected by edges to form a triangle). Let $\mathbf{F} = \{\mathbf{f}_i\}_i^N$ be the features on every vertex of the mesh, the graph convolution layer can be formulated as,

$$\mathbf{f}_{\mathbf{p}}^{l+1} = \omega_0 \mathbf{f}_{\mathbf{p}}^{l} + \sum_{\mathbf{q} \in N(\mathbf{p})} \omega_1 \mathbf{f}_{\mathbf{q}}^{l}, \tag{4.1}$$

where $\mathbf{f}_{\mathbf{p}}^{l+1} \in \mathbb{R}^{d_{l+1}}$ is the output feature of vertex $\mathbf{p}$ after l-th graph convolution layer, and $\mathbf{f}_{\mathbf{p}}^{l} \in \mathbb{R}^{d_l}$ is the corresponding input feature in l-th layer. $N(\mathbf{p})$ are the neighbour points of vertex $\mathbf{p}$. Both ω_0 and ω_1 are parameters $(d_l \times d_{l+1})$ automatically learnt during training. The ω_1 is shared by all edges, and thereby the graph convolution layer can be applied to meshes with irregular shapes (i.e. nodes with different vertex degrees).

The structure of GCN blocks mainly follows Pixel2mesh [224]. In the first GCN block, the first graph convolution layer takes the concatenation of the learnt feature (1×960) and the original vertices (1×3) of template mesh as input and predicts hidden features at 1×256, followed by 12 hidden graph convolution layers (the input is 1×256 and the output is 1×256) and a graph convolution layer to predict the coordinate of each vertex (1×3). The next two GCN blocks are the same, where the first graph convolution layer takes the concatenation of learnt contour features (1×960), the predicted coordinates (1×3) and the learnt features (1×256) in hidden layers of the previous GCN block as input and predicts features at 1×256. This is followed by 13 hidden graph convolution layers (both input and output are 1×256) and a graph convolution layer to predict the coordinates (1×3). Therefore, each GCN block predicts an output of the target mesh, while the template mesh is deformed gradually to fit the contours. Further details about GCN can be found in [224, 225].

3D PC-to-PC Mapping

To apply deformation based on GCN, point-level features are required for the vertices in the template mesh. However, as the input point cloud and the template mesh are both unstructured and have different cardinalities, there is no point-to-point correspondence between them. To transfer the learnt shape information from the input point cloud to the vertices of the template mesh, we build a PC-to-PC mapping module comprising 3D projection and volume-to-PC mapping, where the 3D volume is used as the bridge between the input point cloud and template. The 3D projection aims to map 3D PCs to 3D volumes, which can be formulated as follows (using a volume of 64^3 voxels as an example),

$$V_{x,y,z} = \begin{cases} 0, & (x,y,z) \neq \lfloor(\mathbf{P}_i) \times 32\rfloor + 32 \\ 1, & (x,y,z) = \lfloor(\mathbf{P}_i) \times 32\rfloor + 32 \end{cases} \tag{4.2}$$

where $\mathbf{P}_i$ is the coordinate of i-th point in PCs, which has been normalised before the training. $V_{x,y,z}$ is the corresponding voxel in projected 3D volumes. We project the point cloud into a 64^3 volume. If there is a corresponding point in the point cloud, the voxel in 3D volume would be 1, otherwise 0. To implement our MR-Net, fixed-size inputs are required. Therefore, we randomly replicate the points in the original point clouds of contours to obtain point clouds of the same cardinality (3,000). As 3D projection maps the point cloud into a 3D volume based on the appearance of the input point cloud alone (where points with identical coordinates are presented as one point/voxel in the 3D space or volume), our MR-Net is invariant to duplicates in point clouds.

Correspondingly, the volume-to-PC mapping is the inverse process of 3D projection,

$$\mathbf{f}_i = \mathbf{VF}_{x,y,z}, \ \ s.t.(x,y,z) = \lfloor(\mathbf{P}_i) \times 32\rfloor + 32, \tag{4.3}$$

where $\mathbf{f}_i$ is the obtained feature for point i in template mesh, and $\mathbf{VF}_{x,y,z}$ is the corresponding feature in 3D volume. With these two mappings, we finally obtain point-level features for the template mesh, which serve as the input of GCN blocks. Note that, there is a coordinate scale missing (from float coordinate to integer index) in the process of 3D PC-to-PC mapping. Generally, larger volumes would enable more accurate reconstruction results, although requiring more memory. For a trade-off between the accuracy and computational complexity (GPU memory), we choose a 64^3 volume as the bridge between the input PC and template mesh.

4.2.2 Loss Functions

We employ deep supervision with a multi-term mesh loss function to train our proposed MR-Net. The mesh loss is designed following Pixel2mesh [224], including Chamfer distance (CD), edge loss, normal loss and Laplacian loss. CD is applied to capture an overall distance between the predicted vertices and vertices of ground-truth. It does not require the point number/order to be the same in the two PCs. Denoting $\mathbf{p}$ and $\mathbf{q}$ as the predicted and ground-truth vertices, Chamfer distance L_{CD} is written as,

$$L_{CD} = \sum_{\mathbf{p}} min_{\mathbf{q}} ||\mathbf{p} - \mathbf{q}||_2^2 + \sum_{\mathbf{q}} min_{\mathbf{p}} ||\mathbf{p} - \mathbf{q}||_2^2. \tag{4.4}$$

Edge loss is a regularisation to penalise high edge length. We use the sum of all edge lengths in the predicted mesh as the edge loss L_{edge},

$$L_{edge} = \sum_{\mathbf{p}} \sum_{\mathbf{k} \in N(\mathbf{q})} ||\mathbf{p} - \mathbf{k}||_2^2, \tag{4.5}$$

where $N(\mathbf{q})$ is the neighbour vertices of $\mathbf{q}$.

Normal loss L_{normal} is computed on surface normals, which helps preserve mesh topology and retain fine structural details, and is formulated as,

$$L_{normal} = \sum_{\mathbf{p}} \sum_{\mathbf{q} = argmin_{\mathbf{q}}(||\mathbf{p} - \mathbf{q}||_2^2)} || < \mathbf{p} - \mathbf{k}, \mathbf{n_q} > ||_2^2, s.t.\ \mathbf{k} \in N(\mathbf{p}), \tag{4.6}$$

where $< \cdot, \cdot >$ is the inner product of two vectors, $\mathbf{k}$ belongs to the neighbour point of $\mathbf{p}$ (denoted by $N(\mathbf{p})$), and $\mathbf{n}_q$ is the surface normal of ground-truth. In the predicted/target meshes, the vectors (edges) from each vertex to its neighbour vertices should be perpendicular to its normal. If the predicted vertices of meshes are exactly the same as the target mesh, the normal loss becomes zero. Therefore, this loss is to guarantee the normal of the predicted mesh is close to the normal in the target mesh.

Similar to edge loss, Laplacian loss $L_{Laplacian}$ is also a regularisation term. Let $\delta_{\mathbf{p}}$ be the Laplacian coordinate of vertex $\mathbf{p}$. The $L_{Laplacian}$ is as follows,

$$\begin{aligned} \delta_{\mathbf{p}} &= \mathbf{p} - \sum_{k \in N(\mathbf{p})} \frac{1}{||N(p)||}\mathbf{k}, \\ L_{Laplacian} &= \sum_{\mathbf{p}} ||\delta_{\mathbf{p}}{}' - \delta_{\mathbf{p}}||_2^2, \end{aligned} \tag{4.7}$$

where $\delta_{\mathbf{p}}$ and $\delta_{\mathbf{p}}{}'$ are the Laplacian coordinates of vertex $\mathbf{p}$ before and after deformation.

The mesh loss has been proven to be useful in mesh reconstruction [224, 228]. However, in our task, we found it is inadequate to generate accurate vertex coordinates, as there is no exact point-to-point loss. To tackle this issue, we further apply an additional $L1$ loss between the predicted and ground-truth vertices. This term ($L1$) urges MR-Net to predict more accurate vertices for the reconstructed cardiac mesh, and it is formulated:

$$L1 = \frac{1}{M}\sum_{i}^{M} |\mathbf{p}_i - \mathbf{q}_i|, \tag{4.8}$$

where M is the number of points in the predicted mesh. $\mathbf{p}_i$ and $\mathbf{q}_i$ are coordinates of the i-th point in the predicted and target mesh, respectively.

Therefore, the complete mesh loss L_{mesh} we propose is as follows,

$$L_{mesh} = L_{CD} + L_{edge} + L_{norm} + L_{Laplacian} + \lambda_0 \times L1, \tag{4.9}$$

where λ_0 is a hyper-parameter that needs to be tuned empirically.

As there are three outputs in MR-Net from coarse to fine, we compute the mesh loss on all three outputs. Therefore, the final loss function L_{total} is computed as follows,

$$L_{total} = \lambda_1 L_{mesh1} + \lambda_2 L_{mesh2} + \lambda_3 L_{mesh3}. \tag{4.10}$$

In the loss function, λ_0, λ_1, λ_2 and λ_3 are hyper-parameters that weight the relative influence of each structural loss term, on the overall gradient backpropagated through the network to update the constituent weights. These weights are also tuned empirically.

4.3 Experiments

4.3.1 Data and Implementation

All experiments conducted to validate MR-Net are performed using 7,870 stacks of 2D contours, available from the manual delineation of SAX view cardiac MR images (at the end of systole and diastole), within the UKBB dataset. The spacing for cardiac MR images in UKBB is 1.8×1.8 mm^2 with a slice thickness of 8.0 mm and a slice gap of 2 mm. Manual contouring was performed by a team of cardiac imaging experts [207] and the corresponding 3D bi-ventricle cardiac reference shapes were available from a previous study [219]. We randomly split the dataset into training (6,000), validation (935) and test sets (935). Each training sample comprises a source-target pair, where the former is the sparse 2D contour points to be reconstructed, while the latter is the corresponding bi-ventricle surface mesh (i.e. the target shape). We pre-process all source PCs to maintain the same cardinality (3,000 points used in all experiments) across all samples. This is done by replicating points at random for samples with less than 3,000 points. The target mesh vertices, however, all have the same cardinality (i.e. 1,578 points) and consequently need not be resampled. In the training dataset, all input PCs and target mesh vertices are normalised before training the network, using their centroid and radius (fixed as 100.00 mm). Therefore, all the PCs and meshes used for training are normalised to a sphere centred at $(0, 0, 0)$ with a radius of 1. Correspondingly, during testing, the input PCs are also normalised before shape reconstruction, such that the predicted meshes can be transformed to their original size using the same values for the centroid and radius.

We use the Adam optimiser, with a learning rate of $1e^{-05}$ and a batch size of 1 to train MR-Net, in all experiments conducted. The hyper-parameters λ_0, λ_1, λ_2 and λ_3 for the total structural loss are 1,000, 0.1, 0.3 and 0.6 respectively, which are determined empirically. Note that, these parameters are the same in all experiments. The network is implemented using Python and TensorFlow, and all experiments are streamlined and executed on Tesla M60 GPUs, accessed over the MULTI-X platform [1] [236]. All networks are trained until convergence on the training set, taking ∼3-4 days. Our source code is available on the Github [2].

[1] www.multi-x.org

[2] https://github.com/cistib/MR-Net

4.3.2 Comparison with the State-of-the-art

To the best of our knowledge, no previous deep learning-based method for mesh reconstruction from (stacked) contour PCs exists in the literature. However, various techniques such as point cloud up-sampling, point cloud segmentation and mesh reconstruction from a single image could be modified to partially address the reconstruction problem. Consequently, we build three baselines, using state-of-the-art networks for comparison, namely, PointNet++ [223], PU-Net [237], and Pixel2mesh [224]. PointNet++ is a popular network originally proposed for point cloud classification and segmentation. We build an encoder-decoder network based on its kernel block, with a feature integration component, to obtain our baseline network PointNet++. PU-Net is a state-of-the-art network for point cloud up-sampling. We adapt it to point cloud reconstruction by incorporating a sampling layer at the end of the original network. These two networks can predict PCs only with similar structures to the ground-truth, but cannot recover the cardiac mesh as the order of predicted points differs from the ground-truth. To compare our approach with mesh reconstruction methods, we project the input PCs onto 2D images, and then reconstruct 3D cardiac meshes from them using Pixel2mesh [224]. In addition to deep learning-based methods, we also compare our MR-Net with two traditional point set registration methods, coherent point drift (CPD [238]) and GMMREG [239], where the template mesh is the same as MR-Net and the hyper-parameters are tuned based on samples from the training and validation set.

Qualitative Results

A visual comparison of the generated reconstructions using the proposed method against the baseline networks is depicted in Figure 4.3. For Pixel2mesh, CPD and our proposed MR-Net, both predicted meshes and the corresponding vertices (PCs) are presented, while, only PCs are available for PointNet++ and PU-Net results. For GMMREG, only the mesh is presented, due to limited space. As shown in Figure 4.3, the PointNet++ and PU-Net reconstructions still contain several “contour-like” distributions of points and lack the inlet to the pulmonary artery at the top of the RV. The reconstruction of Pixel2mesh just learns a coarse representation of the cardiac shape, and the corresponding mesh does not preserve bi-ventricle topology and is thus significantly different from the ground-truth. The main reason for this is that 2D projection

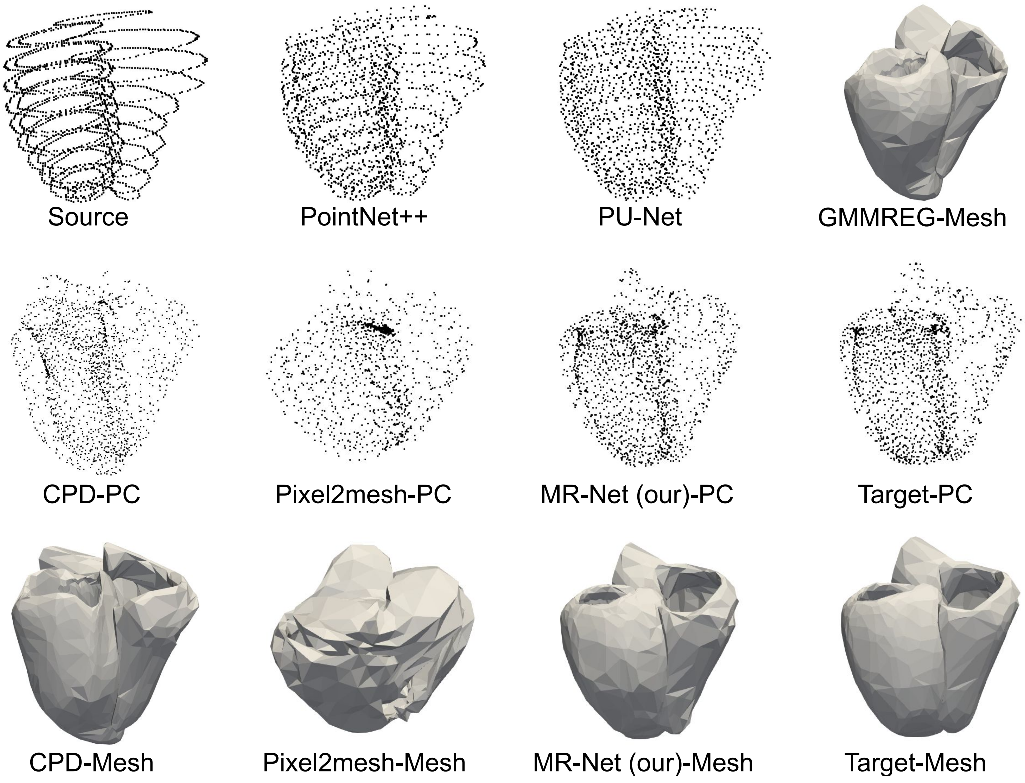

Figure 4.3: Qualitative results for our MR-Net and baseline networks viz. PointNet++, PU-Net, GMMREG, CPD and Pixel2mesh. In the second and third rows, PCs and meshes computed using MR-Net, CPD and Pixel2mesh are presented.

causes a significant loss of information, resulting in erroneous reconstructions. It is difficult for Pixel2mesh to reconstruct meshes with holes using 2D information only. Both traditional point set registration methods, CPD and GMMREG can reconstruct smooth cardiac meshes, whilst preserving topology. In our task, the performance of CPD is better than GMMREG. However, the mesh obtained using CPD is significantly different to the ground-truth mesh, failing to capture several local details (mainly on the top and bottom of the ventricles). MR-Net can reconstruct evenly distributed PCs without contour-like artefacts, while preserving bi-ventricle topology and retaining fine structural details such as the inlet to the pulmonary artery. The reconstructed mesh is of high quality and more closely matches the target shape, compared with the other approaches. This is further supported by the quantitative results summarised in the next section.

Table 4.1: Quantitative comparison between MR-Net and the baseline networks using the CD, EMD, HD and PC-to-PC error. Statistically significant differences in reconstruction accuracy are highlighted in bold. MR-Net (automatic) represents the mesh reconstruction results from contours extracted using automatic segmentation methods (see Section 4.3.4).

Methods	CD (mm)	EMD (mm)	HD (mm)	$\epsilon_{PC-to-PC}$ (mm)	Inference Time(s)
PointNet++	13.03 ± 2.96	17.94 ± 2.07	17.04 ± 3.57	-	< 0.1
PU-Net	12.15 ± 2.88	14.94 ± 2.02	15.74 ± 3.37	-	< 0.1
Pixel2mesh	19.38 ± 5.54	25.27 ± 4.48	16.20 ± 3.30	50.63 ± 7.29	< 0.1
CPD	12.10 ± 6.63	12.49 ± 5.46	13.05 ± 7.74	7.03 ± 2.94	37.45
GMMREG	20.90 ± 7.18	17.58 ± 4.85	15.87 ± 3.04	8.36 ± 1.85	60.90
MR-Net (No L1)	255.08 ± 94.54	36.61 ± 5.49	47.80 ± 7.35	39.12 ± 5.21	< 0.1
MR-Net (Only L1)	6.14 ± 1.61	7.01 ± 1.48	8.10 ± 1.79	3.34 ± 0.65	< 0.1
MR-Net (No PC feature)	6.84 ± 1.69	8.07 ± 1.64	8.78 ± 1.92	3.87 ± 0.65	< 0.1
MR-Net (No 3D CNN)	80.71 ± 39.28	32.03 ± 7.27	29.63 ± 7.06	18.54 ± 2.68	< 0.1
MR-Net	**4.39 ± 1.48**	**5.05 ± 1.41**	**6.89 ± 1.88**	**2.48 ± 0.63**	< 0.1
MR-Net (automatic)	7.57 ± 3.59	8.19 ± 2.87	9.31 ± 2.86	3.45 ± 0.98	< 0.1
MR-Net (small dataset)	6.89 ± 1.76	8.12 ± 1.71	8.83 ± 2.00	3.92 ± 0.79	< 0.1

Quantitative Results

The reconstruction performance of MR-Net is also quantitatively evaluated and compared with other baseline networks. Following previous shape reconstruction research [220, 240, 224], reconstruction accuracy was measured using CD, earth mover distance (EMD), Hausdorff distance (HD) [240, 237, 224] and point cloud to point cloud (PC-to-PC) error [220, 221], which could capture the distance between two PCs from different perspectives. The CD, EMD, and HD are well-known metrics to evaluate the distance between two PCs, while PC-to-PC error is computed as,

$$\epsilon_{PC-to-PC} = \frac{1}{M}\sum_{m=1}^{M}\sqrt{\sum_{i=1}^{3}(\mathbf{p}_{m,i} - \mathbf{q}_{m,i})^2}, \tag{4.11}$$

where **p** and **q** are vertices of predicted meshes and ground-truth, and the M is the number of points in predicted meshes (in our experiments is 1,578). For all evaluation metrics, lower values signify better performance. The average reconstruction accuracy (expressed as mean±std) across all test samples is summarised in Table 4.1, for each approach, as mentioned earlier. Paired sample t-tests were used to assess statistical significance, by comparing the reconstruction accuracy of each baseline network with that of MR-Net. MR-Net consistently outperformed the others, achieving the best results across all metrics. Note that, CPD is also the method used to generate the ground-truth meshes (as mentioned in [219]), requiring a time-consuming process of tuning hyper-parameters for each sample. In this chapter, for comparison, we tune the hyper-parameters based on several samples from the training and validation set, and use the same hyper-parameters for all testing samples. This is why the meshes obtained using CPD in this study are different to the target meshes (generated in a previous study [219]). As the inference of MR-Net is much faster (< 0.1 s vs 37.45/60.90 s) and more accurate than traditional methods, there is potential for its use in real-time applications.

To further demonstrate the clinical potential and superiority of our approach, we extract the corresponding segmentations from the predicted and ground-truth meshes (using the SAX-planes from the original cardiac MR images), and compute five clinical indices based on the obtained segmentation results - LV end-diastolic volume (LVEDV), end-systolic volume (LVESV), LV stroke volume (LVSV), LV ejection fraction (LVEF) and LV myocardial mass (LVM) respectively. The clinical indices are shown in Table 4.2 (as topology is not preserved in meshes predicted by pixel2mesh and MR-Net(No L1), we did not include their clinical indices), where, values showing no statistically significant difference to the clinical indices computed on the ground-truth meshes are highlighted in bold (P-value ≥ 0.05). While the meshes predicted by MR-Net incur an average point-to-point error of 2.48mm to the ground-truth, we found that the com-

Table 4.2: Clinical indices (LVEDV, LVESV, LVSV, LVEF, LVM) computed based on the segmentation obtained from the predicted meshes. Those clinical indices make no statistically significant difference to the ground-truth (GT) are highlighted in bold (P-value≥ 0.05).

Methods	LVEDV (ml)	LVESV (ml)	LVSV (ml)	LVEF(%)	LVM(g)
CPD	77.66 ± 17.57	43.79 ± 11.86	33.87 ± 9.05	43.71 ± 7.09	161.71 ± 36.47
GMMREG	84.33 ± 19.15	48.19 ± 12.04	36.14 ± 9.00	42.94 ± 5.23	175.96 ± 38.87
MR-Net (Only L1)	**132.63 ± 30.49**	**40.23 ± 15.55**	**92.40 ± 19.86**	**70.12 ± 5.87**	85.81 ± 19.55
MR-Net (No PC feature)	128.33 ± 29.13	**38.88 ± 14.41**	89.44 ± 19.27	**70.05 ± 5.65**	**87.37 ± 19.16**
MR-Net (No 3D CNN)	80.70 ± 13.63	**38.99 ± 16.47**	41.71 ± 13.43	52.42 ± 17.37	56.40 ± 16.58
MR-Net	**131.69 ± 30.59**	**39.69 ± 12.71**	**92.00 ± 20.10**	**70.12 ± 4.33**	**88.36 ± 19.97**
MR-Net (automatic)	131.50 ± 30.81	**39.76 ± 12.81**	91.74 ± 20.28	**70.01 ± 4.54**	88.70 ± 20.15
GT Clinical Indices	132.24 ± 30.25	39.61 ± 11.92	92.63 ± 20.24	70.27 ± 3.88	87.78 ± 20.26

puted clinical indices for MR-Net show no significant difference to the latter. All other approaches investigated on the other hand, show significant differences to the ground truth, in terms of the clinical indices evaluated. This further demonstrates the superiority of our approach at preserving key clinical indices that are routinely used to assess cardiac function.

We also explore the performance of MR-Net when trained with a limited number of samples, as 6,000 samples are not easy to obtain in real clinical applications. We randomly choose 200 samples from the original training set to train MR-Net and evaluate its performance with the same test set. The results are shown in Table 4.1(MR-Net(small dataset)). These results indicate that MR-Net performs well in the small data regime, and outperforms other state-of-the-art methods which were trained on a significantly larger sample size (6,000).

These quantitative results follow the visual assessment (cf. Sec 4.3.2) of the bi-ventricle shapes reconstructed using each approach. This highlights further the efficacy of our proposed MR-Net for 3D shape reconstruction from stacked 2D contours.

4.3.3 Shape Reconstruction from Incomplete Contours

Typical artefacts encountered during cardiac MR image acquisition include missing slices between the base and apical of the heart, and low signal-to-noise (SNR) ratio in parts of the myocardium, resulting in blurred boundaries for the left and right ventricles. Correspondingly, these errors are propagated to the manually or automatically extracted contours from such image volumes, which might cause missing contours at intermediate points across the heart. A 3D cardiac shape reconstruction framework robust to the presence of such irregularities, would be of significant clinical value as it would enable accurate quantification of cardiac functional indices, despite such artefacts. For that reason, the robustness of MR-Net to incomplete data in sparse 2D contours, used for 3D shape reconstruction, is also evaluated.

Incomplete samples are generated by retaining the basal and apical contours and randomly removing the contours between them. This process is used to generate four new samples with 2 to 5 slices each, for every sample in the original dataset. Additionally, to tackle the common issue encountered in routine CMR imaging, of missing apical/basal slices, two new samples with one/two pairs of base and apical slices missing are also generated. The resulting dataset, comprising 42,000 training samples, is used to re-train MR-Net and evaluate its robustness to incomplete data.

Table 4.3: Quantitative results for our MR-Net with incomplete data (with 2-5 slices and original input). The -2 slices and -4 slices denote the results with contours removing one/two pairs of apical and basal slices.

Criterion	2 Slices	3 Slices	4 Slices	5 Slices	-2 Slices	-4 Slices	Original Input
CD (mm)	13.54 ± 14.65	7.94 ± 3.02	7.91 ± 3.36	6.97 ± 2.54	6.51 ± 1.98	9.97 ± 2.68	**5.22 ± 1.78**
EMD (mm)	12.17 ± 4.93	8.78 ± 2.65	8.74 ± 2.79	7.92 ± 2.31	7.74 ± 2.03	9.13 ± 2.40	**6.16 ± 1.75**
HD (mm)	11.83 ± 4.12	9.58 ± 2.83	9.34 ± 2.73	8.73 ± 2.40	9.00 ± 2.50	10.20 ± 2.89	**7.48 ± 1.96**
$\epsilon_{PC-to-PC}$ (mm)	5.46 ± 2.44	3.94 ± 1.11	3.88 ± 1.12	3.55 ± 0.96	3.46 ± 0.84	4.07 ± 1.36	**2.87 ± 0.73**

The quantitative and qualitative results in Table 4.3 and Figure 4.4, respectively, indicate that MR-Net can generate accurate reconstructions of cardiac shape even in the presence of missing information (i.e. missing slices).

In the extreme scenario (reconstruction from 2 slices), only bottom and apical slices are given, our proposed MR-Net can still reconstruct high-quality meshes, although small misalignments exist between the reconstructed mesh and input contours. We observe that the reconstruction performance progressively improves with including more slices/contours, with a proportional decrease in the variance. When 5 slices are given for mesh reconstruction, the reconstruction performance is close to the results obtained using a complete stack of slices, across all metrics. Although the reconstruction accuracy of MR-Net for extreme scenarios is significantly lower than that of the original input (unmodified 2D contours), the values summarised in Table 4.3 indicate its performance is still comparable to/better than the baseline networks' performance on complete data

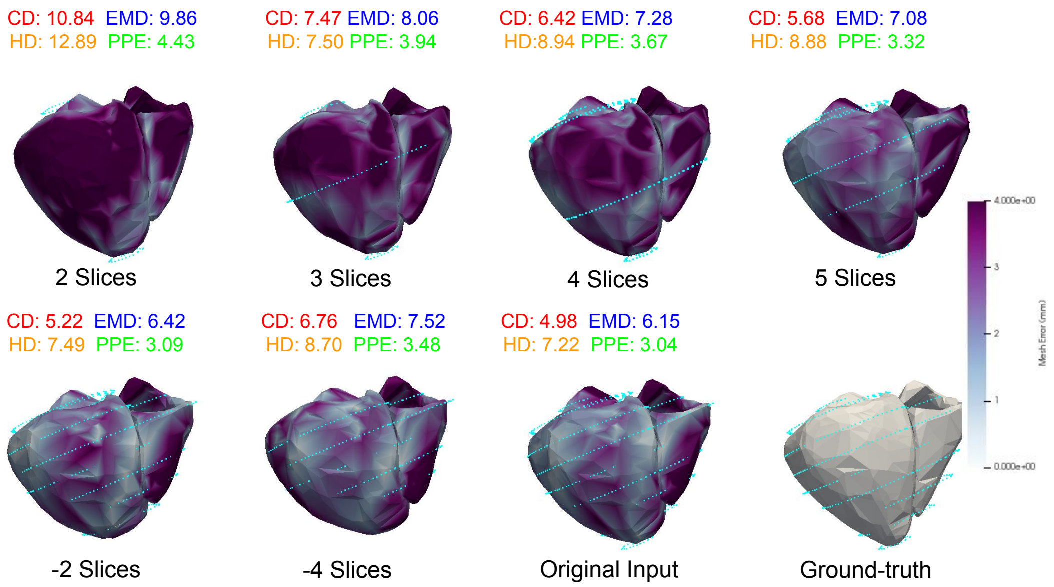

Figure 4.4: 3D cardiac shape reconstruction with incomplete input. The CD, EMD, HD and PC-to-PC error (denoted by PPE in the figure) are shown on the left-top with red, blue, yellow and green numbers, respectively. The mesh colours indicate the PC-to-PC error from the predicted meshes to the target meshes (colours corresponding to distances between 0.00 mm and 4.00 mm are shown in the colour bar).

(cf. Table 4.1).

To further evaluate the robustness of our approach, we employed the trained model to reconstruct meshes in the absence of apical and basal slices. As the apical and basal slices are essential to provide the network with contextual information regarding cardiac size, removing them significantly affects the quality of mesh reconstruction. Therefore, the results of removing apical and basal slices (-2 or -4 slices) are generally worse than removing the same number of slices between the apical and basal slices. However, our approach can still generate high-quality cardiac meshes with the basal/apical slices missing, as shown in Table 4.3 and Figure 4.4.

The robustness of our approach to missing slices implies we can reconstruct high-quality cardiac meshes using fewer annotated (manually/semi-automatically) slices and from sparse SAX cine-MR images. This provides avenues to reduce scan time in the future. Hence, the proposed approach could be of significant value in a clinical setting, especially for applications requiring real-time shape reconstruction (e.g. surgical navigation).

4.3.4 Shape Reconstruction from Autocontouring

To further validate the robustness and efficacy of MR-Net in realistic scenarios, in this section, we exemplify our method on shape reconstruction with contours extracted from automatic segmentation results instead of manual segmentations. Compared to manual segmentation, automatic segmentation results may contain several errors, posing a challenge for accurate 3D shape reconstruction. To be viable for a real clinical setting, however, a shape reconstruction method should be able to cope with such errors and facilitate accurate shape reconstruction from the original cardiac MR images.

For the samples in our testing dataset, the original MR images, the corresponding PCs of contours from manual segmentations and their target meshes are all available. Therefore, we use a deep learning-based cardiac segmentation method [172] to segment the original MR images, and then extract PCs of contours from the segmentation results. Finally, we apply our pre-trained MR-Net to reconstruct 3D cardiac meshes from them. As the target mesh for every MR image is available, we compare the predicted meshes with the former (shown in Table. 4.1 and Figure 4.5).

As shown in Figure 4.5, there are small differences between the input contours extracted from automatic segmentation results and manual segmentation results in terms of the number of contours, location and shape. However, even with those differences, our proposed MR-Net can still reconstruct accurate and high-quality meshes, achieving comparable performance to the reconstruction from manually segmented contours. This is further demonstrated by the results in Table. 4.1, where we see that mesh reconstruction accuracy using automatically segmented contours (MR-Net (automatic)) is a little worse than the results of mesh reconstruction from manual segmentation (MR-Net), but significantly better (achieving an average 3.5 mm PC-to-PC error about the ground-truth) than the other baseline networks investigated. During inference, MR-Net can reconstruct the shape of a sample less than 0.1s on average, and $\sim$ 1s or less duration is required for the estimation of bi-ventricle contours using the deep learning-based segmentation method. Therefore, with the automatic segmentation method and MR-Net, we can reconstruct accurate, high-quality, 3D cardiac meshes from original

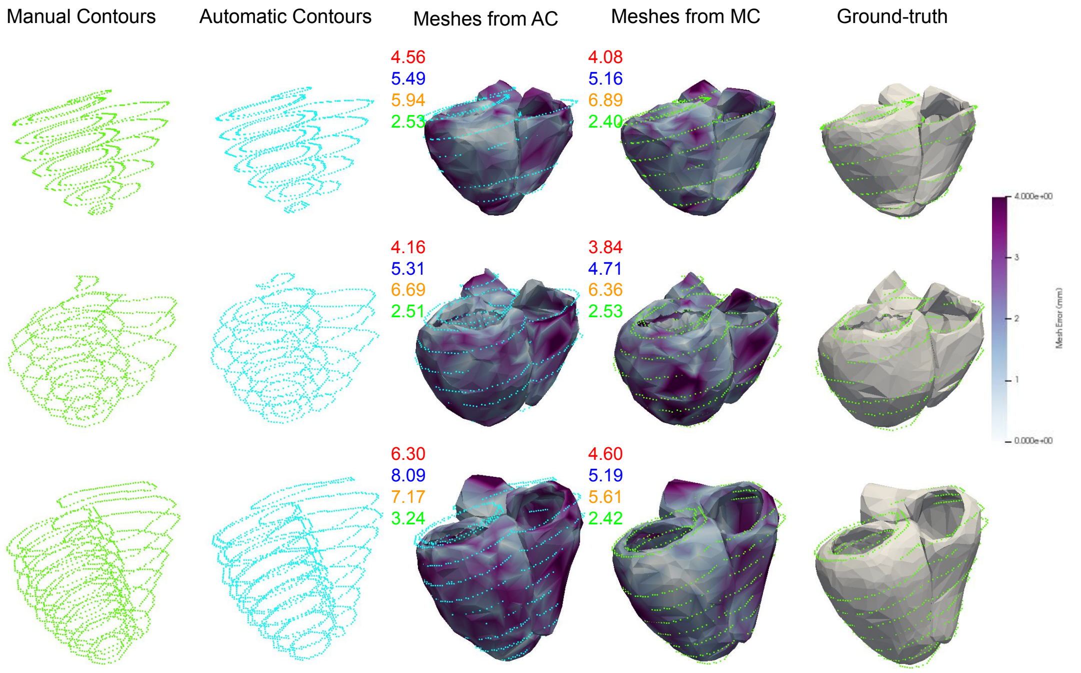

Figure 4.5: Samples of 3D cardiac shape reconstruction using automatic and manual annotated contours. Each row is one sample. Columns from left to right are: Manual annotated contours (MC), automatic annotated contours (AC), reconstructed 3D meshes from both AC and MC, and ground-truth. The colours of metrics (between predicted meshes and ground-truth) are the same as predicted meshes in Figure 4.4.

cardiac MR images very quickly ($\sim$ 1s), which is adequate for their use in real-time applications.

Compared with traditional 3D cardiac shape reconstruction approaches, MR-Net achieves a significant improvement in the inference time, without compromising the accuracy of the reconstructed 3D shapes. Additionally, as demonstrated, the proposed approach outperforms existing state-of-the-art deep learning approaches in terms of shape reconstruction accuracy. Assisted by deep learning-based segmentation methods, MR-Net can be further applied for direct 3D shape reconstruction from original MR/CT images. MR-Net can be applied to (1) guide other clinical image tasks in return (e.g. segmentation and registration) as it provides a continuous shape in 3D space, (2) in several real-time applications (e.g. surgical navigation), and (3) as an extension of clinical tools for visualising the 3D shape of anatomical structures. Although our proposed MR-Net can reconstruct highly similar meshes to the ground-truth, currently, the reconstruction accuracy is still constrained by the size of the 3D volume, which is the fundamental building block of PC-to-PC mapping. The reconstruction accuracy can be further improved with larger volume (e.g. $128 \times 128 \times 128$ or $256 \times 256 \times 256$ voxels) as the bridge for PC-to-PC mapping.

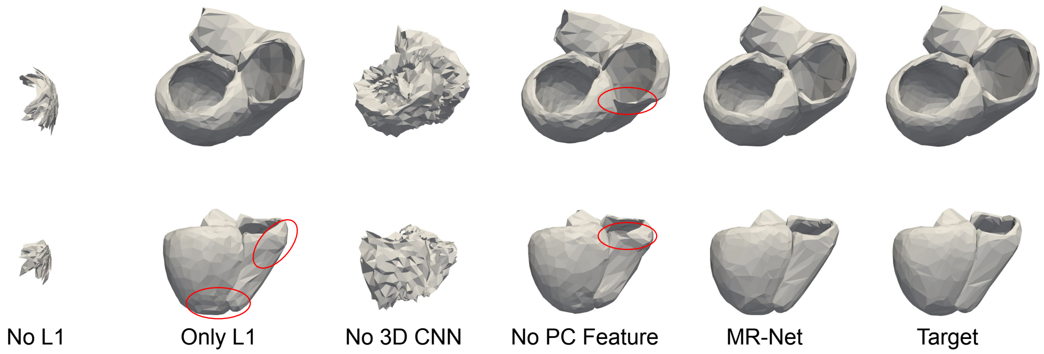

Figure 4.6: The results are predicted by different versions of MR-Net, where the first and second rows are the meshes from two different orientations.

4.3.5 Ablation Study

To analyse the contribution of different components in MR-Net, an ablation study is performed, as shown in Table 4.1, Table 4.2 and Figure 4.6. MR-Net (No L1), MR-Net (Only L1), MR-Net (No PC feature), and MR-Net (No 3D CNN) refer to training MR-Net without the L1 loss, with just the L1 loss, without the PC feature extraction block, and without the 3D CNN feature extraction block, respectively. MR-Net achieves statistically significant improvements (evaluated using paired t-tests) to the aforementioned variations of MR-Net on all metrics (P-value$<<$ 0.01). Comparing the results between MR-Net (No L1) and MR-Net, we found that the L1 loss plays a key role in the network training, without which the network fails to reconstruct cardiac shapes. The other losses (except L1 loss) bring marginal improvements to the reconstruction accuracy, help better preserve fine structural details (viz. top and bottom of the right ventricle in Figure 4.6) and facilitate the generation of smoother meshes. Similarly, the lack of a PC feature extraction block weakens the reconstruction accuracy of MR-Net, while, the lack of a 3D CNN feature extraction block significantly affects mesh reconstruction quality. Therefore, we can conclude that the L1 loss and 3D CNN feature extraction block are the key contributors to the reconstruction accuracy of MR-Net. The remaining components (other losses and the PC feature extraction block) help further refine mesh reconstruction accuracy.

4.4 Conclusion

A novel deep learning-based approach for 3D shape reconstruction from stacked 2D contours is proposed in this chapter. Our approach, MR-Net, can accurately reconstruct 3D shapes from sparse and incomplete 2D contour data, outperforming three state-of-the-art point cloud/mesh reconstruction networks. We further prove that our proposed approach can reconstruct accurate 3D cardiac meshes using contours generated by an automatic segmentation approach. This demonstrates that our model is robust to the segmentation errors induced by the latter. Using 2D automatic segmentation methods and our MR-Net, it is possible to reconstruct high-quality 3D cardiac meshes in real-time. The versatile and robust nature of the proposed framework highlights its potential for application in several diagnostic and interventional settings. MR-Net is a supervised method, requiring ground-truth meshes during training. To alleviate the burden of curating high-quality ground truth meshes, which can be non-trivial in several applications, the problem of shape reconstruction from sparse contour/point cloud data can be tackled in an unsupervised manner. This could be achieved by approaching the problem in a manner similar to unsupervised deep learning-based image registration techniques, using the template mesh as the moving image and the point clouds of contours as the fixed image. This will be the subject of future work.

Cardiac mesh has drawn more and more attention in cardiac motion analysis and disease diagnosis, as it provides a more efficient and intuitive 3D representation of the heart. Prior to this, how to obtain plausible cardiac mesh is essentially important, because it is unable to obtain directly by imaging. With our proposed MR-Net, we can reconstruct plausible and accurate 3D cardiac shapes from the segmentation masks/contours almost in real-time. However, for this approach, manual or automatic segmentation is required prior to the mesh reconstruction. This additional pre-process step would delay the mesh reconstruction, and at the same time may introduce additional segmentation interference when using automatic segmentation. Therefore, we further propose a deep learning-based method to reconstruct accurate 3D four-chamber cardiac meshes directly from original images (details can be found in [23]).

With the aforementioned approaches (i.e. SDDIR, MR-Net), we can generate two additional cardiac representations, the segmentation masks and cardiac meshes, from raw cardiac MR images. In the next chapter, we will focus on CVD prediction and diagnosis, using the biomarkers extracted from the raw MR images, segmentation masks and meshes.

Chapter 5

Deep Learning in Cardiovascular Disease Prediction and Diagnosis

In previous chapters, we have described approaches to obtain the cardiac deformation fields, segmentation and 3D cardiac meshes from given cardiac MR images. With these cardiac representations (including original MR images), we can extract adequate biomarkers regarding the cardiac anatomical structures and motion functions. In this chapter, we introduce how to incorporate those available features/predictors into the prediction and diagnosis of CVDs. Specifically, we designed a novel multi-channel variational auto-encoder (MCVAE), named MIVAE, to learn a joint representation of the paired mesh and image. After training, the shape-aware image representation (SAIR) can be learnt directly from the raw images and then applied for further CVD prediction and diagnosis. We demonstrate our proposed method on the data from the UK Biobank (UKBB) study and two other publicly available datasets via extensive experiments. We show that our proposed method can reconstruct high-quality images and meshes from the latent embedding, even with a single input. It can be applied for 3D cardiac mesh reconstruction from the corresponding image. Using the learnt SAIR as a novel biomarker in subsequent prediction/diagnosis of CVDs, we find it leads to better performance than traditional biomarkers (e.g. clinical indices), and can be applied as an efficient supplement to them, which is of significant potential in CVD analysis and prediction/diagnosis.

5.1 Introduction

Cardiovascular disease is the leading cause of global mortality. As non-invasive methods, medical imaging techniques such as magnetic resonance (MR), computed tomography (CT) and ultrasound (US), followed by computer vision techniques, have become more and more popular in the analysis and diagnosis of heart-related diseases. MR image is generally considered the gold standard for disease diagnosis among those image modalities, due to its high contrast in anatomical structures and lack of ionizing radiation [241]. Previous research has demonstrated the feasibility and efficiency of image-based diagnosis on various cardiovascular diseases (e.g. heart failure, ischemic heart disease, congenital heart disease, pulmonary hypertension, dilated cardiomyopathy) [24].

To analyse and predict/diagnose the CVDs from the given images, many machine learning and deep learning (DL)-based approaches have been proposed, solving various medical image analysis tasks [242]. Automatic segmentation approaches [172] are widely studied to get rid of the time-consuming manual delineation work. Based on the predicted segmentation masks at end-diastole (ED) and end-systole (ES) of the cardiac cycle, clinical indices like left ventricle (LV) and right ventricle (RV) ejection fraction, ES and ED volume, and myocardial mass can be estimated. In addition, image registration can obtain the deformation fields between different time frames in the cardiac cycle, for cardiac motion tracking and strain estimation [177, 183]. Cardiac shape analysis based on the 3D cardiac mesh is also popular, which provides an intuitive way to observe and capture cardiac motion [23].

Research on machine learning/DL-based CVD analysis can be roughly divided into three classes, direct disease diagnosis [243], disease/survival prediction [24](i.e. predict the probability of disease/death in a specific period from now), and association analysis between cardiac motion and diseases/genomes/other factors [242, 244, 245]. The direct disease diagnosis generally extracts biomarkers or feature descriptors from the original images/deformation fields/cardiac meshes, then uses classifiers (e.g. support vector machine (SVM), random forest, artificial neural network (ANN)) for disease diagnosis [243, 21]. For this type of method, the extraction of biomarkers is essentially important, where some basic information (such as sex and age) and cardiac clinical indices derived from images/segmentation/deformation fields, are generally used. Recently DL-based approaches have been demonstrated to overcome traditional machine learning-based methods in various tasks [177], however, there are still few works that have attempted to apply DL methods for direct cardiac disease diagnosis, because of the lack of interpretability. Disease/survival prediction has similar feature extraction steps to disease diagnosis, aiming to predict the probability of getting CVDs in specific years or the survival time of CVD patients. For example, Bello et al. [24] proposed a novel auto-encoder for time-resolved 3D meshes to learn the task-specific latent representation and survival time of patients with pulmonary hypertension, significantly outperforming human benchmarking and Cox proportional hazards model [246]. Instead of directly applying DL-based methods for classification, previous studies [242, 244, 245] have proposed to use DL networks to predict segmentation, deformation fields and 3D cardiac meshes from given MR images, based on which some cardiac motion patterns (e.g. strain and myocardial wall thickness) can be calculated automatically. Then, the correlation between the cardiac motion phenotype and specific factors (e.g. genetic & environmental factors) can be studied.

However, current CVD prediction/diagnosis is generally solely based on the image domain (MR image) or spatial domain (3D mesh). For image domain-based analysis, cardiac MR images provide high-quality local details, but certain limitations like large slice thickness (for cardiac cine-MR image), slice misalignment, interference of background tissues and inability to visualise 3D shape weaken the interpretability and performance of CVD diagnosis/analysis methods. In contrast, in the spatial domain-based analysis, cardiac mesh provides a more intuitive way to present cardiac shape, and facilitates the assessment of cardiac motion. Nevertheless, the accuracy of analysis relies heavily on the quality of reconstructed meshes (derived from cardiac images), and may introduce inaccurate results on local details (due to the nature of mesh reconstruction). Therefore, a natural idea is to synergistically leverage the advantages of both cardiac representations to enhance subsequent prediction and diagnosis, attaining optimal performance.

To obtain explainable and efficient representations from cardiac MR images and improve CVD prediction/diagnosis performance, we propose a novel MCVAE [247], mesh-image variational auto-encoder (MIVAE), to learn the joint latent representations of cardiac meshes and cine-MR images. Following training, using images alone as input, the learnt latent embedding of images, which we named shape-aware image representation (SAIR), is fed into a machine learning classifier (e.g. SVM) for downstream tasks like cardiovascular disease prediction and diagnosis. Once the MIVAE is trained, it can be applied as a mesh reconstruction approach by only feeding images as input. We demonstrate the CVD prediction/diagnosis performance of our method on a large dataset UKBB [184], and also evaluate the trained MIVAE on another two datasets, Automatic Cardiac Diagnosis Challenge (ACDC) [148] and Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image Segmentation Challenge (M&M) [185], without any fine-tuning steps.

The contributions of this chapter are summarised as follows,

- We propose a novel MCVAE network for cross-modality data (mesh and image), comprising a convolutional encoder-decoder and a graph encoder-decoder. To the best of our knowledge, it is the first paper to learn the joint latent variable from images and 3D surface meshes.

- A novel, efficient and robust cardiac feature representation, SAIR, is learnt by our MIVAE. We demonstrate that incorporating this feature can improve predictive performance relative to existing biomarkers and can correspondingly, supplement the latter in predictive diagnostic and prognostic tasks.

- Using the learnt MIVAE, we are able to automatically reconstruct cardiac bi-ventricle meshes from corresponding MR images, which provides a novel method for cardiac mesh reconstruction from the corresponding images.

5.2 Related Work

This work is mainly related to joint representation learning and CVD prediction/diagnosis.

5.2.1 Joint Representation Learning

Joint representation learning is a process of learning a parametric mapping from different data in multi-domains (e.g. images, video, sound, text) to feature vectors/tensors in a shared latent space, with the aim of apprehending the latent correlation between multi-modality data and extracting more refined and valuable vectors/features, which can enhance performance across various downstream tasks. It has drawn attention from various cross-modality applications in different domains such as population clustering [248] and disease diagnosis [247, 249]. The architectural designs of joint representation learning may significantly vary from each other due to differences in inputs. Nonetheless, they generally encompass several distinct encoders, which encode data from different domains into the joint latent space. Corresponding to different inputs, the encoders are generally different (for example, convolutional layers for images/videos, fully connected (FC) layers for vectors, and graph convolutional layers [250] for graph), aiming to convert the redundant input into a low-dimensional vector/tensor that can enhance performance in downstream tasks.

MCVAE [247] is a popular structure used in joint representation learning, including multiple encoder-decoder pairs, which can encode data from different modalities into the same latent distribution. MCVAE can reconstruct missing channels when the given input is incomplete, and the learnt latent representations which contain sufficient information from the input multimodal data can be applied for subsequent analysis (e.g. disease diagnosis [247, 249] and separating cell populations [248]). For example, Diaz et al. [249] proposed a two-channel MCVAE to predict cardiac MR images from the retinal images of the same subject, and used it for left ventricle end-diastolic volume (LVEDV) and left ventricle mass estimation and prediction of myocardial infarction. Ternes et al. [248] proposed an MCVAE in single-cell image analysis to extract transform-invariant biologically meaningful features, which helped to advance the understanding of complex cell biology and enable discoveries previously hidden behind image complexity (like clustering populations).

Note that, although the previous approaches on MCVAE learn the correlation between different data modalities, they are generally different sub-modalities belonging to the same modality (e.g. different contrasts of MR images), or several image modalities with a quite simple modality (e.g. a feature vector). To our best knowledge, no previous research has explored the joint correlation between two significantly different complex modalities like 3D image and mesh.

5.2.2 Image-based Cardiovascular Disease Analysis

As a non-invasive technology, cardiac imaging has been widely used in CVD diagnosis and to improve our understanding of the structure and function of the heart. To obtain a fast and accurate CVD prediction/diagnosis, numerous machine learning/DL-based approaches [251, 21] have been developed, feeding the features extracted from the image and corresponding non-image data to a classifier/regressor and achieving the prediction/diagnosis of CVDs. The non-image data generally includes demographic data (e.g. sex, age), conventional risk factors (e.g. smoking, hypertension, high cholesterol, diabetes) and other available data in the electronic health record. In this chapter, we simply refer to them all as metadata. The classifier can be general machine learning classifiers like the random forest, SVM, and ANN, also including recent DL networks. It is essentially important in CVD prediction/diagnosis to extract discriminative features/biomarkers that are representative of the patient data and their underlying target class of interest. In general, the features can be divided into four categories, metadata, clinical indices, radiomic features [252, 251, 21] and automatic features extracted by deep neural networks. The first is available in the electronic health record, while the rest three are derived from the images.

In most previous research [253, 254, 255], the metadata and clinical indices were widely used and achieved reasonable results. Standard cardiac clinical indices include LVEDV, left ventricle end-systole volume (LVESV), left ventricle ejection fraction (LVEF), left ventricle myocardium mass, right ventricle end-diastole volume (RVEDV), right ventricle end-systole volume (RVESV), right ventricle ejection fraction (RVEF).

Note that, the clinical indices are computed based on the segmentation masks at end-diastolic (ED) and end-systolic (ES) frames of the cardiac cycle, and thereby segmentation masks are required for this biomarker. Radiomic features are also derived from raw images, referring to a collection of handcrafted features derived based on first and second-order statistics of local intensity patterns in images and based on other types of filter responses resulting from processing image patches/local neighbourhoods of pixels [251, 256, 257, 258, 25, 26]. It was originally widely used for cancer diagnosis [258], and recently used in diagnosis/prediction of CVDs [25, 26, 21], leading to promising results. Similar to clinical indices, corresponding segmentation masks are generally required to calculate radiomic features.

Due to a large amount of data in the raw cardiac image cycle, early researchers tend to not directly use the raw image for CVD prediction/diagnosis. Instead, they proposed to extract several biomarkers (e.g. radiomic features and clinical indices) from the raw images or corresponding segmentation at ED and ES frames, and then fed them to classification/regression approaches. With the advent of DL, researchers have started to explore deep neural networks for CVD prediction/diagnosis, using automatic feature extraction (by a CNN) instead of manual-designed feature extraction [259, 260, 261, 262]. Lu et al. [261] proposed a deep regression network to estimate clinical measurements from B-Mode echocardiography images, and used it for abnormality detection, achieving better results than using a direct classification network. Similarly, Kusunose et al. [262] designed an end-to-end deep CNN for automated diagnosis of myocardial ischemia using echocardiography images, achieving comparable results to that produced by cardiologists and sonographer readers.

However, there are some limitations to end-to-end DL-based prediction/diagnosis approaches. Firstly, end-to-end classification/regression networks generally lack interpretability. Although DL-based networks can provide fast and accurate classification results, it is difficult to interpret learnt features in a clinical sense and how they contributed to the diagnosis of different CVDs, because of the nature of end-to-end DL networks. In addition, CVD refers to a group of complex diseases that affect cardiac structure and motion. Consequently, more than one frame in the cardiac cycle is required for accurate diagnosis. However, it would bring a huge computation burden to incorporate all frames (each is a 3D image) of the cardiac cycle into DL networks.

A possible solution is to use DL approaches to learn explainable representations of the heart and utilise them for subsequent analysis and diagnosis. In this chapter, we design an MCVAE to learn a joint latent representation of cardiac image- and shape-based features, where the impacts of each variable in the latent vector can be assessed by varying it and visualising the resulting reconstructions. Using the learnt latent representation of the image (SAIR), as a biomarker for subsequent CVD prediction/diagnosis, we can achieve the task of CVD prediction/diagnosis. Compared with radiomic features and clinical indices, the extraction of SAIR does not require detailed segmentation masks. Consequently, our proposed method does not propagate segmentation errors incurred to the features that are extracted/learnt (which is the problem with clinical indices and radiomic features) and can fit more complex scenarios where segmentation masks may not be available.

5.3 Method

The study and experiments in this chapter are designed and reported in adherence with the guidelines in CLAIM checklist [263]. In this section, we first introduce data preparation and the network architecture of MIVAE, then describe CVD prediction and diagnosis with the SAIR learnt from MIVAE.

5.3.1 Mesh Preparation

In this work, all the patient-specific cardiac meshes are obtained by registering a template mesh to the contours of each subject (the details can be found in our previous paper [23]). To capture the critical shape information and ensure that learnt representations do not capture differences in pose (i.e. position and orientation) between the individual shape instances, we remove all differences in pose (by spatially normalising the meshes with respect to translation, rotation and scale) between the cardiac meshes by rigidly aligning them to the original cardiac mesh. The resulting rigidly registered patient-specific meshes present the critical shape information of each heart, named normalised mesh. To balance the point/voxel value between cardiac meshes and corresponding MR images, and enhance the network training, all coordinates of vertices in the normalised mesh are divided by a radius of k mm ($k = 100$ in this work, following [22]) to make sure the mesh is within a sphere centre at (0,0,0) with radius 1, which is used as the input mesh ($\overline{\mathbf{S}}_\mathbf{i}$) in MIVAE. Therefore, denoting the original

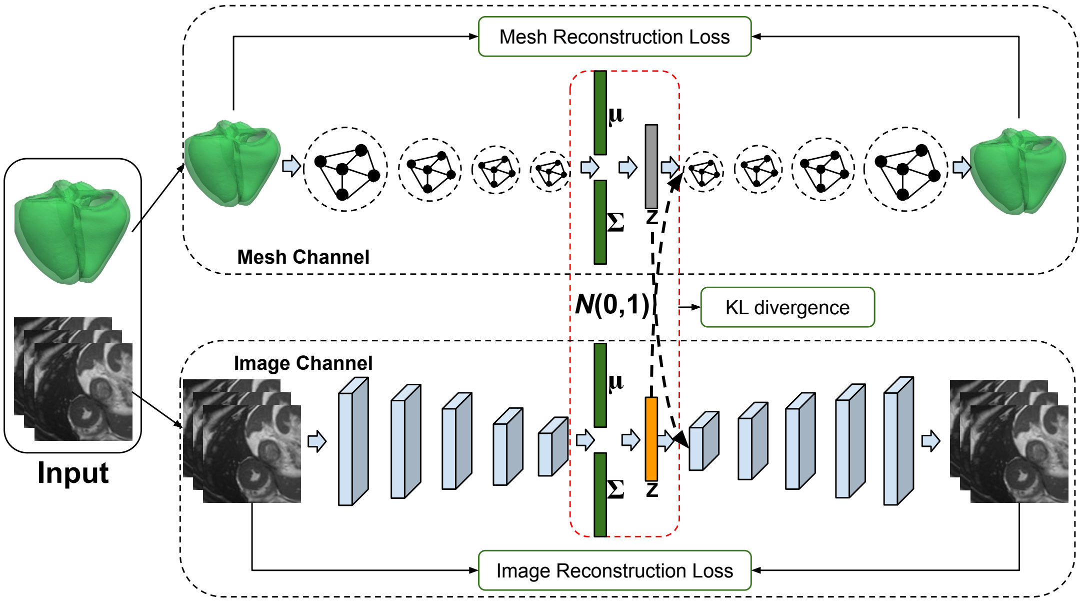

Figure 5.1: Schema of MIVAE. Our MIVAE includes two channels, the mesh encoder-decoder and image encoder-decoder respectively. Note that, the latent variable learnt in the image channel is exactly the SAIR used for subsequent CVD diagnosis (highlighted in orange). The cardiac MR images were reproduced by kind permission of UK Biobank ©.

mesh of a subject as $\mathbf{O_i}$, the recovery process from the $\mathbf{\overline{S}_i}$ to the original patient-specific cardiac mesh $\mathbf{O_i}$ is formulated as,

$$\mathbf{z_i} = c_i \times (\mathbf{\bar{s}_i} \times k) \times \mathbf{r_i} + \mathbf{t_i}, \tag{5.1}$$

where c_i, $\mathbf{R_i}$, $\mathbf{t_i}$ are the corresponding scale, rotation and translation parameters for each patient-specific mesh $\mathbf{O_i}$. Corresponding to $\mathbf{\overline{S}_i}$, all the input MR image slices are cropped, scaled (to 128×128) and normalised (the intensity value are normalised into $[-1, 1]$).

5.3.2 Mesh-image Variational Auto-encoder (MIVAE)

To learn the joint latent embedding of cardiac image and mesh, we design a mesh-image variational auto-encoder, MIVAE, as shown in Figure 5.1. MIVAE is an MCVAE [247], consisting of two channels of encoder-decoder, the mesh channel and image channel respectively. Given the input (i.e. cardiac image and mesh pairs), denoted as $\mathbf{x} = \{\mathbf{x}_{mesh}, \mathbf{x}_{img}\}$, the corresponding encoder (image encoder or mesh encoder) would encode them into l-dimensional latent vectors $\mathbf{z}$. Subsequently, two corresponding decoders (image decoder and mesh decoder) are applied to decode the latent vectors $\mathbf{z}$ to obtain the reconstructed results, denoted as $\mathbf{x}' = \{\mathbf{x}'_{mesh}, \mathbf{x}'_{img}\}$. The generative process for the observation is formulated as follows,

$$\mathbf{z} \sim p(\mathbf{z}), \tag{5.2}$$

$$\mathbf{x}_c \sim p(\mathbf{x}_c \mid \mathbf{z}, \boldsymbol{\theta}_c), \qquad \text{for } c \text{ in } \{1, 2\}, \tag{5.3}$$

where $p(\mathbf{z})$ is the prior distribution of latent vector $\mathbf{z}$ and $p(\mathbf{x}_c \mid \mathbf{z}, \boldsymbol{\theta}_c)$ is a likelihood distribution for the observations conditioned on the latent variable. The likelihood functions belong to a distribution family P parameterised by the set $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{mesh}, \boldsymbol{\theta}_{img}\}$.

As deriving the posterior $p(\mathbf{z} \mid \mathbf{x}, \boldsymbol{\theta})$ is not always computable analytically, variational inference is used to compute an approximate posterior. We approximate the posterior distribution with $q(\mathbf{z} \mid \mathbf{x}_c, \boldsymbol{\phi}_c)$ (conditioned on single channel $\mathbf{x}_c$ and corresponding variational parameters $\boldsymbol{\phi}_c$), which belong to a distribution family Q parameterised by the set of parameters $\boldsymbol{\phi} = \{\boldsymbol{\phi}_{mesh}, \boldsymbol{\phi}_{img}\}$. Therefore, the MIVAE is trained by maximizing the variational lower bound $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{x})$,

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{x}) = \mathbb{E}_c \left[L_c - \mathcal{D}_{\mathrm{KL}}\left(q(\mathbf{z} \mid \mathbf{x}_c, \boldsymbol{\phi}_c) \mid\mid p(\mathbf{z}) \right) \right], \tag{5.4}$$

where $\mathcal{D}_{\mathrm{KL}}$ is Kullback-Leibler (KL) divergence, used to impose a constraint enforcing each $q(\mathbf{z} \mid \mathbf{x}_c, \boldsymbol{\phi}_c)$ to be as close as possible to the target posterior distribution. Here, L_c is the expected log-likelihood of decoding each channel from the latent representation of channel $\mathbf{x}_c$, generally formulated as,

$$L_c = \mathbb{E}_{q(\mathbf{z} \mid \mathbf{x}_c, \boldsymbol{\phi}_c)} \sum_{i=1}^{C} \ln p(\mathbf{x}_i \mid \mathbf{z}, \boldsymbol{\theta}_i). \tag{5.5}$$

As there is a channel for mesh encoder-decoder in MIVAE, in addition to the log-likelihood, we further incorporate a mesh loss to ensure high-quality mesh reconstructions. The details about the loss function can be found in Section 5.3.3.

Mesh Channel Encoder-decoder comprises mesh encoder and decoder subnetworks which encodes the input mesh into a latent vector and then decodes it back to reconstruct the input 3D mesh. A mesh $\mathbf{M}(\mathbf{V}, \mathbf{F})$ is constructed by vertices $\mathbf{V}$ and faces **F**. In this chapter, all the meshes are obtained by registering a template mesh, sharing the same faces. Therefore, we only need to predict the vertices of each cardiac mesh in the mesh encoder-decoder.

Both mesh encoder and mesh decoder are built using Chebyshev graph convolution [250] layers. The former is composed of four down-sampling blocks (sampling $\frac{1}{16}, \frac{1}{8}, \frac{1}{4}, \frac{1}{4}$ points from the points in the previous layer, respectively), each comprising a graph convolution layer, a down-sampling layer and an exponential linear unit (ELU) activation layer. Then a flatten operation followed by an FC layer is used to map the resulting features to an l-dimensional vector. The encoder predicts a distribution of an l-dimensional variable, parameterised by the mean $\boldsymbol{\mu}$ and standard variation $\boldsymbol{\sigma}$. Following the general MCVAE [247], the reparameterisation trick is used, and thereby an l-dimensional latent vector $\mathbf{z}_1$ is sampled from the distribution.

In the decoder, similarly, an FC layer followed by a reshape operation is utilised to recover the latent vector into a graph structure (i.e. shape like $N \times M$, where N is the number of points and M is the dimension of feature for each vertex). After that, corresponding to the encoder, four up-sampling blocks (comprising an up-sampling layer, a graph convolution layer, and an ELU activation layer) are used to up-sample the graph structure back to the original size of the input mesh.

Image Channel Encoder-decoder is a general convolution-based encoder-decoder. The input images are cardiac MR images in short-axis view (SAX), including a stack of slices. Due to the large gap between slices (the image spacing for cardiac MR images in UKBB is generally $1.8 \times 1.8 \times 10mm^3$), we use 2D convolution instead of 3D convolution in the image encoder-decoder.

In the image encoder, five down-sampling blocks are used, each comprising a convolution layer, a batch-normalisation layer and an activation layer (Leaky ReLU). Similarly, a flatten operation with an FC layer is used to turn the down-sampled features into an l-dimensional latent vector.

Similar to mesh encoder-decoder, the reparameterisation trick is also used and an l-dimensional vector $\boldsymbol{z}_2$ is sampled from the latent image distribution. In the image decoder, the $\boldsymbol{z}_2$ is fed into an FC layer followed by a reshape operation, turning into feature maps with the size of 4×4. Corresponding to the image encoder, five up-sampling blocks, comprising the convolution layer, batch-normalisation layer and Leaky ReLU activation layer, are used to recover the feature back to the original images. The output layer is a convolution layer, followed by a $tanh()$ activation. Note that, both $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ are fed into the mesh decoder and image decoder, predicting the reconstructed mesh and image of each channel.

5.3.3 Loss Functions

In most variational auto-encoder based networks, the reconstruction error (generally negative log-likelihood) and KL divergence are directly used as the loss function. However, different from the image, the mesh represented by vertices in the MIVAE is sparse and discontinuous in the space, where even a small difference in the coordinates may lead to broken faces in the mesh, without the constraints of face regularisation. Hence, in addition to the log-likelihood, we include a mesh loss as an additional regularisation in the final loss function.

The mesh loss follows Pixel2mesh [224, 22], including two regularisation losses, the edge loss and normal loss, and a point-to-point loss. Edge loss is a regularisation to penalise too-long edges. Denoting $\mathbf{p}$ and $\mathbf{q}$ as the predicted and ground-truth vertices, We use the sum of all edge lengths in the predicted mesh as the edge loss L_{edge},

$$L_{edge} = \sum_{\mathbf{p}} \sum_{\mathbf{k} \in N(\mathbf{q})} ||\mathbf{p} - \mathbf{k}||_2^2, \tag{5.6}$$

where $N(\mathbf{q})$ is the neighbour vertices of $\mathbf{q}$.

Normal loss L_{normal} is computed using the point-wise surface normal vectors, which helps preserve mesh topology and retain fine structural details, and is formulated as,

$$L_{normal} = \sum_{\mathbf{p}} \sum_{\mathbf{q} = argmin_{\mathbf{q}}(||\mathbf{p}-\mathbf{q}||_2^2)} || < \mathbf{p} - \mathbf{k}, \mathbf{n_q} > ||_2^2, \tag{5.7}$$

where $< \cdot, \cdot >$ is the inner product of two vectors, $\mathbf{k}$ belongs to the neighbour point of $\mathbf{p}$ ($\mathbf{k} \in N(\mathbf{p})$) , and $\mathbf{n}_q$ is the surface normal of ground-truth. In the predicted and target meshes, the vectors (edges) from each vertex to its neighbour vertices should be perpendicular to its normal. If the predicted vertices of meshes are exactly the same as the target mesh, the normal loss becomes zero. Therefore, this loss is to guarantee the normal vectors of the predicted mesh are as close as possible to the normal vectors of the target mesh.

To improve the performance of mesh reconstruction, we further apply an L1 loss between the coordinates of the predicted mesh and ground-truth mesh, in addition to the negative log-likelihood loss. Then, the complete mesh loss L_{mesh} is as follows,

$$L_{mesh} = L_{edge} + L_{norm} + \lambda_0 \times L1, \tag{5.8}$$

where λ_0 is a hyper-parameter that needs to be tuned empirically.

Therefore, the final loss function L_{total} to train MIVAE is computed as follows,

$$L_{total} = \lambda_1 L_{KL} + \lambda_2 L_{NLL} + \lambda_3 L_{mesh}, \tag{5.9}$$

where the L_{KL}, L_{NLL} are the KL divergence and negative log-likelihood of reconstructed images/meshes (for both channels), as mentioned in Eqn. 5.4 and Eqn. 5.5. λ_1, λ_2 and λ_3 are hyper-parameters that weigh the relative influence of each loss term, which are tuned empirically.

With MIVAE, we can learn the latent representations of both cardiac images and shapes. One important benefit of MCVAE is that, it enables training and inference with missing channels of information, allowing for the missing information to be imputed during inference. In most realistic applications, the cardiac mesh may not always be available. In such a scenario, the MIVAE can still extract latent representations for cardiac MR images. As the representation SAIR is sampled from a joint latent distribution of cardiac image and mesh, it captures sufficient shape-aware information conditioned on the input image and can be used as a biomarker for downstream analyses such as disease prediction/diagnosis. Moreover, due to the nature of MCVAE, MIVAE can provide a mapping from image to mesh, or from mesh to image, in addition to the encoder-decoder of image/mesh. Therefore, it further offers a novel approach for fast and accurate mesh reconstruction from images.

5.3.4 CVD Prediction/diagnosis

Most CVDs affect both cardiac structure and motion, and thereby it is not sufficient to use a single volume in the cardiac cycle for CVD prediction/diagnosis. With MIVAE, we can extract the SAIR features for all the cardiac images from the whole cardiac cycle. Considering the fact that most previous CVD features/biomarkers (e.g. clinical indices) are extracted from cardiac images at ED and ES, in this work, we also compute SAIR at both ED and ES frames of the cardiac cycle for subsequent analysis. To demonstrate the efficiency of SAIR, we compare it to traditional CVD features like clinical indices and radiomic features.

Using the learnt SAIR as predictors, we can achieve CVD prediction/diagnosis with any available classifiers. In this work, we use a traditional machine learning approach, SVM, as the classifier. Limited by available datasets, we evaluate the prediction performance of SAIR in UKBB and diagnosis performance in ACDC and M&M. Note that, MIVAE are only trained on UKBB. The SAIR features extracted using the trained model are used for CVD prediction/diagnosis using a test set from UKBB and external test sets from ACDC and M&M without any re-training/fine-tuning on the external data. In all three datasets, the samples are limited, and thereby we use 10-fold cross-validation to validate the performance of classification. In addition, the SAIR feature and radiomic features are both high dimensional (over 500) feature vectors, which are easy to over-fit on limited samples. Therefore, following [25, 21], we use sequential forward feature selection to identify the most relevant ones for CVD prediction/diagnosis in each feature composition.

5.4 Experiments

5.4.1 Data and Implementation

Our MIVAE is trained using cardiac SAX images from UKBB. To train the MIVAE, 1,176 image-mesh pairs at ED/ES time points of the cardiac cycle from the UKBB are used, each with the corresponding SAX image and 3D bi-ventricle mesh. We split the dataset into training and testing sets, with 1,052 and 124 samples respectively. The input meshes are obtained following our previous paper [23], with the preprocessing steps described in 5.3.1. There are 96,749 coordinates in each mesh, comprising the left ventricle and right ventricle. All input MR images are cropped, resized and padded into $128 \times 128 \times 15$, and then their intensities are normalised into $[-1, 1]$.

For CVD prediction, we choose 442 subjects from UKBB, of which 221 got acute myocardial infarction (AMI) within 10 years after cardiac MR scanning while the remaining 221 did not (until December 15, 2022). The segmentation masks for all cardiac MR images were obtained automatically using the segmentation method in [23]. The metadata of each subject is a 24-dimensional vector, including four types, biological factors (e.g. age, sex), lifestyle (e.g. ethnicity and smoking status), diagnoses (e.g. diabetes) and treatments (details can be found in [264, 265]). Following previous biomarkers (e.g. clinical indices), we extract SAIR features (a 512-dimensional feature vector) at ED and ES for each subject. The performance of SAIR is compared with traditional biomarkers and metadata.

To further demonstrate the robustness and generalisation of our MIVAE in CVD diagnosis, another two datasets ACDC and M&M are also used for inference. In these two datasets, cardiac meshes are no longer available, but we can still use the images from them to validate the performance of CVD diagnosis. Different from UKBB, in these two datasets, patients are grouped into more than two disease classes. Hence, multi-class classification is applied in these two datasets. In ACDC, five classes, normal subjects (NOR), myocardial infarction (MINF), dilated cardiomyopathy (DCM), hypertrophic cardiomyopathy (HCM), and abnormal right ventricle (ARV), are available, each containing 20 samples. Similarly, M&M dataset also includes five classes, DCM, HCM, NOR, ARV, and hypertensive heart disease (HHD), containing 54, 49, 38, 8, and 1 samples, respectively. To reduce the domain gap between different datasets, all the MR images in ACDC and M&M are pre-processed by re-sampling (to $1.8 \times 1.8 \times 10mm^3$), histogram-matching (to the average histogram of 100 random subjects in UKBB), cropping, scaling, padding and intensity normalisation into a size of $128 \times 128 \times 15$.

We use the Adam optimiser, with a learning rate of $1e^{-3}$ and a batch size of 7 to train MIVAE, in all experiments. The hyper-parameters λ_0, λ_1, λ_2 and λ_3 for the total structural loss are 100, 1, 1, 1 and 15 respectively, which are determined empirically and the same in all experiments. In MIVAE, the input meshes and images are both encoded into a 512-dimensional latent vector ($l = 512$, determined empirically). The network is implemented in Python using the PyTorch library, and is trained using Tesla M60 GPUs. All networks are trained until convergence on the training set.

5.4.2 Comparison and Evaluation Metrics

In the following sections, we first demonstrate that our learnt SAIR is able to capture discriminative information useful for CVD prediction/diagnosis from the given images/meshes, and then apply it in CVD prediction and diagnosis tasks. To demonstrate the performance of mesh reconstruction, we compare our proposed MIVAE with a previous mesh reconstruction approach, MCSI-Net [23].

For the CVD prediction and diagnosis, we compare the SAIR features learnt by MIVAE with several traditional biomarkers/features, including metadata, clinical indices, and radiomic features (extracted by the Python package Pyradiomics). For radiomic features, we extract seven types of features (the shape-based (3D), shape-based (2D), first-order statistics, gray level co-occurence matrix, gray level run length matrix, gray level size zone matrix, neighbouring gray tone difference matrix, gray level dependence matrix) at ED and ES frames of the cardiac cycle, a total of 2104-dimensional features. For the comparison in UKBB, we also compare SAIR with Qrisk3 [266, 267], a popular risk to estimate the 10-year risk of having a cardiovascular event for a subject based on his/her metadata.

To evaluate the performance of mesh/image reconstruction, we use point-to-point error e_{mesh} (following [22]) and mean absolute error e_{image} between the reconstructed data and the original inputs. In addition, to further evaluate the anatomical structure accuracy of the reconstructed meshes, we also compute the Dice score (for the left ventricle (LV), left ventricle myocardium (LVM), right ventricle (RV) and the average of them) and Hausdorff distance between the segmentation masks delineated from predicted meshes and ground-truth meshes. For the prediction/diagnosis performance, we calculate four classification evaluation metrics, accuracy, precision, F1 score and recall to obtain a thorough evaluation. In UKBB, the area under the curve (AUC) of the receiver operating characteristic (ROC) is also computed.

5.4.3 Evaluation of Image/mesh Reconstruction

Before applying MIVAE to learn the joint latent representation of cardiac images and meshes, it is important to ensure that the latent embedding can present sufficient information from the input data, which can be reflected by the reconstruction quality. The quantitative results of reconstruction are shown in Table. 5.1. In MIVAE, each input channel has two outputs: the reconstructed image and the reconstructed mesh. Therefore, in the inference, we can use MR images/meshes alone as input, and reconstruct the corresponding cardiac images and meshes. We find that using either image or mesh alone as input can reconstruct high-quality meshes. The reconstructed meshes using either meshes or images as input have low point-to-point error to target meshes (same as input meshes); even the meshes reconstructed from the image input are with ~3.6 mm point-to-point error to target meshes. Applying the reconstructed meshes for cardiac MR image segmentation, the results of MIVAE when given images as input are comparable to the MCSI-Net, with no significant difference in the LV Dice score.

Considering the nature of the variational auto-encoder and no ground-truth contour information is needed, SAIR captures sufficient information and is deemed suitable for subsequent CVD diagnosis.

In most realistic applications, there are only the original cardiac images, without the corresponding meshes. Our proposed MIVAE can reconstruct the corresponding accurate meshes from given images for multiple subsequent analysis tasks (e.g. segmentation). Meanwhile, the obtained latent embedding (SAIR) from MIVAE can be further applied for CVD prediction and diagnosis.

Table 5.1: Reconstruction error of mesh and image. The first and second rows are the results using only mesh or image as input, respectively. Point-to-point error e_{mesh} is used to evaluate the distance between reconstructed meshes and the ground-truth mesh. We segment the original image with predicted mesh, and plot the segmentation performance with Dice score and Hausdorff distance (HD). For image reconstruction error, we simply use the mean absolute error e_{image}. In the results of MIVAE using the image as the sole input, the bold highlights the results of MIVAE making no significant difference to MCSI-Net (P-value larger than 0.05).

Methods	e_{mesh} (mm)	Average Dice	LV Dice	LVM Dice	RV Dice	HD (mm)	e_{image}
MIVAE(mesh as input)	0.67 ± 0.09	97.76 ± 0.37	98.01 ± 0.52	97.25 ± 0.50	98.01 ± 0.46	6.67 ± 3.28	0.172 ± 0.247
MIVAE(image as input)	3.56 ± 1.02	88.31 ± 3.07	$\mathbf{90.87} \pm \mathbf{2.79}$	85.81 ± 3.70	88.24 ± 3.58	17.65 ± 9.65	0.155 ± 0.227
MCSI-Net [23](image as input)	2.77 ± 1.23	90.28 ± 5.51	91.63 ± 5.75	88.76 ± 5.96	90.48 ± 5.19	14.35 ± 10.06	-

5.4.4 Segmentation with Predicted Meshes

To further demonstrate the accuracy of our MIVAE in mesh reconstruction, we rigidly transform the predicted meshes from both channels in MIVAE back to the space of original meshes with corresponding transformation parameters, then overlay the mesh back into the original MR images to obtain the segmentation masks. We compute the Dice score and Hausdorff distance between the segmentation masks from predicted meshes and segmentation from ground-truth meshes, as shown in Figure 5.2 and Table. 5.1. It can be observed that, the meshes reconstructed of MIVAE using meshes alone as input are with a high Dice score (97.76%) to the ground-truth, since the embedding is learnt from the input mesh. For the results of MIVAE using image alone as input, while it is marginally lower than MCSI-Net [23] on the average Dice score (88.31% vs 90.28%), it makes no significant difference to MCSI-Net on the LV

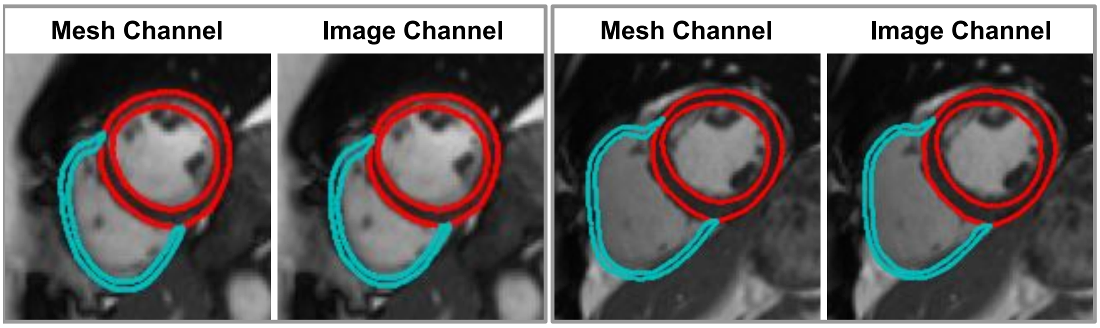

Figure 5.2: Segmentation results with the predicted cardiac meshes. Two subjects are presented here, where the left is the result of the mesh channel and the right is the result of the image channel. The cardiac MR images were reproduced by kind permission of UK Biobank ©.

Dice score. Note that, MCSI-Net requires both SAX images and long-axis view images as input, while MIVAE only uses SAX images (less in formation in input) and achieves comparable performance to MCSI-Net.

5.4.5 AMI Prediction on UKBB

With the learnt SAIR from MIVAE, we are able to implement CVD prediction/diagnosis by feeding it into any available classifiers (here we use the SVM in scikit-learn package, with radial basis function kernel and $C = 1.0$). We first demonstrate its performance in the prediction of AMI, using cardiac MR images from UKBB. Note that, here we only use the SAIR at ED and ES frames of the cardiac cycle for each subject, as most traditional biomarkers (e.g. clinical indices) only use these two frames. We compare the performance of SAIR with traditional biomarkers used in CVD prediction, including metadata, clinical indices and radiomic features. In addition, we also compare our SAIR with Qrisk3 score (using the Qrisk3 score as a feature for classification), and explore the result of combining all features (including metadata, clinical indices, radiomic features, Qrisk3 and SAIR).

Note that, we only have 442 samples for the training and testing, which is a small number for a SAIR feature (1024-dimension) or radiomic features (2104-dimension). Considering the curse of dimensionality, feature selection is needed. Considering the

Table 5.2: Quantitative comparison of AMI prediction results on UKBB between SAIR and the baseline features using Accuracy, Precision, F1 and Recall. For each feature composition, 8 features are selected from the original feature using Mlxtend package with 10-fold cross-validation.

Methods	Accuracy(%)	Precision(%)	F1(%)	Recall(%)	AUC
Metadata	65.79 ± 3.40	66.89 ± 3.66	65.17 ± 3.15	65.75 ± 3.17	66.93 ± 3.72
Qrisk3	65.41 ± 2.81	65.91 ± 2.91	65.26 ± 2.78	65.63 ± 2.80	69.86 ± 3.59
Qrisk3+Metadata	67.67 ± 2.45	67.84 ± 2.29	67.62 ± 2.44	67.81 ± 2.29	71.69 ± 2.89
Clinical Indices	57.52 ± 2.26	57.79 ± 1.98	57.35 ± 2.26	57.70 ± 2.01	60.55 ± 3.84
Radiomic features	79.47 ± 3.76	79.73 ± 3.63	79.38 ± 3.77	79.52 ± 3.74	82.49 ± 3.17
SAIR	81.43 ± 2.93	81.47 ± 2.93	81.38 ± 2.91	81.46 ± 2.87	84.52 ± 2.88
All features (w/o SAIR)	82.18 ± 2.73	82.54 ± 2.77	82.09 ± 2.72	82.23 ± 2.76	84.91 ± 3.79
All features	83.38 ± 2.53	83.62 ± 2.64	83.32 ± 2.49	83.42 ± 2.50	85.10 ± 3.14

curse of dimension, we apply a sequential forward feature selection method (following [25], using the Mlxtend package with 10-fold cross-validation) to select 8 features from the given features, as the input to SVM. The classification results after feature selection are shown in Table. 5.2, with the corresponding receiver operating characteristic (ROC) curve in Figure 5.3. We observe that, the results of metadata outperform the results of clinical indices (65.79% vs 57.52%). Qrisk score performs similarly to the metadata, as it is derived from the metadata. It is interesting to see that the combination of metadata and Qrisk3 obtain a higher accuracy (67.67%) than when either one is used independently. Radiomic features outperform those traditional biomarkers, with 79.47% accuracy. However, our SAIR leads to a significantly better performance than radiomic features (81.43% vs 79.47%). Combining all the traditional features with SAIR (All features) results in better AMI prediction accuracy (82.18% vs 83.38%), which demonstrates our SAIR can provide complementary information for traditional biomarkers.

In addition, an accuracy curve along the SAIR feature dimension is also provided in Figure 5.4. In the beginning, with more features, the classification accuracy increases. However, after 80 dimensions (where the accuracy is 87.59%), the accuracy tends to drop. This demonstrates the importance of feature selection. To understand how the SAIR features contribute to the CVD diagnosis and the ventricle shape changes these features encode, we adopt a method for explainability using latent traversals proposed recently in [268]. Each of the selected latent SAIR features is varied about

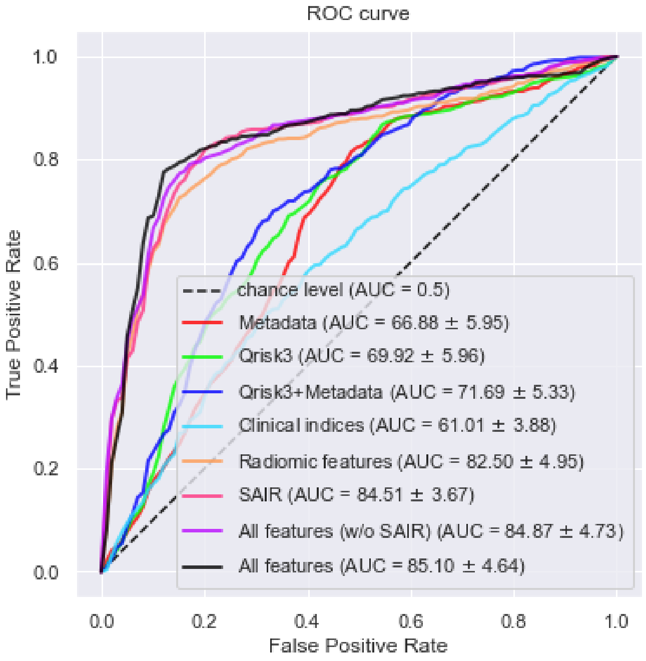

Figure 5.3: Classification ROC curve of SAIR and baseline features on AMI prediction.

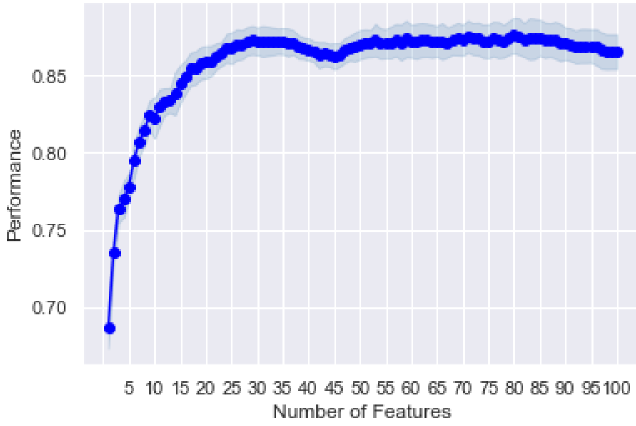

Figure 5.4: Classification accuracy (mean with standard variation) with the increase of SAIR feature dimension, in UKBB. When selecting 80 features from the 1024-dimensional SAIR, it achieves the highest accuracy (87.59%).

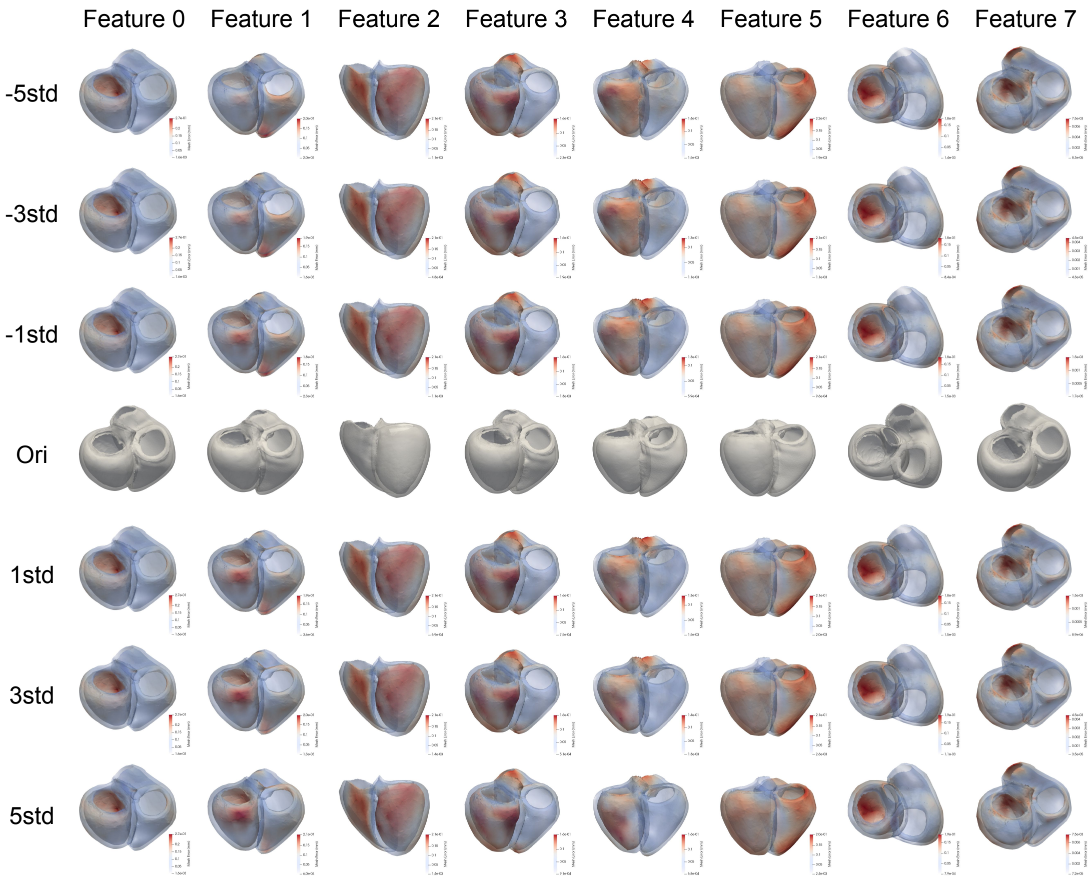

Figure 5.5: Visualisation of feature dimension. Those feature dimensions are visualised in cardiac meshes, by interpolating the original SAIR vector with specific values (from -5 times standard variation (std) to 5 times std) in the corresponding position.

the estimated mean (referred to as latent traversals) to reconstruct new ventricle shapes, and the observed shape changes are visualised in Figure 5.5. We find that most of the critical features for AMI prediction are located on the left ventricle wall, especially on the endocardium (see features 0, 1, 2, 6&7). Part of the features also focus on the connection between the ventricle and artery (the pulmonary valve and aortic valve, see features 2, 3, 4 &7), and both walls of the left and right ventricles (see feature 5).

5.4.6 CVD Diagnosis on ACDC and M&M

To demonstrate the generalisation of our MIVAE and its applications in realistic scenarios, we further apply the MIVAE trained on UKBB data, to unseen images from two publicly available, external cardiac MR datasets, ACDC and M&M. Note that, we directly test our MIVAE trained from UKBB on ACDC and M&M, without any fine-tuning. Similarly, we also select 8 features from the original feature. These two datasets have multiple classes. In ACDC each class contains 20 samples, and in M&M the number of samples for each class is not the same. There is a class in M&M containing only 1 sample, and thereby we simply remove it from the original 150 samples. The metadata in these two datasets both contains two features (weight and height in ACDC, sex and age in M&M).

Table 5.3: Quantitative comparison on ACDC between SAIR and the baseline features using Accuracy, Precision, F1 and Recall. For each biomarker (except metadata), 8 features are selected from the original feature using the Mlxtend package with 10-fold cross-validation.

Methods	Accuracy(%)	Precision(%)	F1(%)	Recall(%)
Metadata	16.67 ± 5.16	20.10 ± 10.86	15.66 ± 5.41	20.30 ± 7.61
Clinical Indices	92.67 ± 3.59	93.09 ± 3.52	92.07 ± 4.03	92.50 ± 4.00
Radiomic features	61.67 ± 10.57	65.85 ± 9.27	61.19 ± 9.93	65.74 ± 7.91
SAIR	66.00 ± 7.12	68.90 ± 8.02	65.66 ± 6.95	68.98 ± 7.66
All features (w/o SAIR)	93.33 ± 3.65	93.27 ± 3.51	93.37 ± 3.32	94.41 ± 2.88
All features	94.33 ± 2.60	94.72 ± 2.96	94.55 ± 2.97	95.41 ± 2.88

The results on ACDC are shown in Table. 5.3. As there are only two features, height and weight of subjects, available in metadata, the classification accuracy of metadata in ACDC is quite low (16.67%). In contrast, the traditional clinical indices obtain a high accuracy (92.67%). Our SAIR outperforms radiomic features, with a 4% increase (66.00% vs 61.67%). Comparing the results of All features with/without SAIR(94.33% vs 93.33%), it can be observed that incorporating SAIR can further improve the accuracy of CVD diagnosis.

In Table. 5.4, we present the diagnosis results on M&M. Similar to the findings in ACDC, the metadata leads to the lowest classification performance, with 43.56% accuracy, due to the limited information provided by age and sex. Radiomic features achieve the best classification results (with 87.78% accuracy). While clinical indices outperform all other biomarkers in ACDC, in M&M, the accuracy of clinical indices is 62.44%, lower than our SAIR (71.78%). In addition, we find the incorporation of our SAIR leads to a better classification performance than without (88.67% vs 87.78%), which demonstrates our SAIR can provide further information apart from the existing traditional biomarkers for CVD prediction/diagnosis.

Table 5.4: Quantitative comparison after feature selection on M&M between features learnt by MIVAE and the baseline features using the Accuracy, Precision, F1 and Recall. For each feature (except metadata), 8 features are selected from the original feature using the Mlxtend package with 10-fold cross-validation.

Methods	Accuracy(%)	Precision(%)	F1(%)	Recall(%)
Metadata	46.22 ± 7.08	28.62 ± 8.58	30.56 ± 7.17	35.24 ± 5.77
Clinical Indices	62.44 ± 5.48	48.77 ± 7.41	48.30 ± 7.65	50.65 ± 7.47
Radiomic features	87.78 ± 3.18	67.68 ± 8.21	69.50 ± 7.67	71.82 ± 7.10
SAIR	71.78 ± 4.94	56.18 ± 6.39	56.38 ± 5.94	58.33 ± 5.19
All features (w/o SAIR)	87.78 ± 3.18	67.68 ± 8.21	69.50 ± 7.67	71.82 ± 7.10
All features	88.67 ± 4.03	68.43 ± 8.20	69.96 ± 7.82	72.13 ± 7.31

We observe that there is a decrease in classification/predictive performance when applying the learnt MIVAE directly to unseen images from ACDC and M&M. This is attributed to the domain shift between images from UKBB and the images available in ACDC and M&M, leading to reduced predictive performance on ACDC and M&M. Nevertheless, SAIR features extracted by our model outperform specific biomarkers (for example, outperform radiomic features in ACDC, and clinical indices in M&M) and serve as an effective supplement for existing biomarkers.

5.4.7 Discussion

Due to the nature of MCVAE, our proposed MIVAE can be applied to cross-domain reconstruction, reconstructing meshes from corresponding images, or reconstructing images from corresponding meshes. In realistic scenarios, the raw images are generally available, thus our MIVAE can be used as a novel image-to-mesh reconstruction approach. We have demonstrated that it can achieve comparable segmentation performance to the previous mesh reconstruction approach using less input information (only SAX images), but there is a limitation for MIVAE that, the predicted mesh of MIVAE is not aligned to the input images, and thus an additional approach like [23] to predict transformation parameters is required.

In CVD diagnosis, the learnt SAIR shows significantly better results than traditional biomarkers in UKBB, and comparable or better performance than traditional features/biomarkers in unseen images from other datasets. Therefore, it can be used as a useful augmentation for existing features. Although in non-UKBB data our SAIR perform worse than specific biomarkers, it does not require detailed ground-truth segmentation as radiomic features/clinical indices, which means our SAIR has more general and robust applications.

Similar to other DL-based approaches, a main limitation of SAIR is the generalisation between different datasets. While we can directly apply the trained MIVAE to extract SAIR of images from other sources, its performance would be weakened if the input images have significantly different appearances from images in UKBB. Pre-processing techniques like re-sampling, and histogram matching can be applied to alleviate this decrease, and the exploration of domain adaptation or generalisation techniques could provide a solution to this problem. In addition to CVD diagnosis, our learnt SAIR can be also applied to find potential sub-types of diseases using clustering techniques.

5.5 Conclusion

In this chapter, to obtain efficient representations from the raw MR images for subsequent CVD prediction and diagnosis of CVDs, we designed a novel two-channel MCVAE, MIVAE, to learn joint latent representations of cardiac images and corresponding meshes. After training, given MR images alone as input, our MIVAE can reconstruct high-quality bi-ventricle meshes and learn shape-aware image representations, useful for subsequent CVD prediction/diagnosis. Through experiments on UKBB, we demonstrate that the segmentation performance using meshes predicted by our approach is comparable to previous approaches. The learnt novel biomarker SAIR by MIVAE captures efficient representations from raw images for CVD prediction/diagnosis, leading to better performance than traditional biomarkers. Also, the learnt SAIR features capture information that is not contained within existing biomarkers (e.g. metadata, clinical indices), and thereby it helps supplement existing biomarkers and improves overall predictive performance. We further demonstrate the robustness of SAIR on non-UKBB data (ACDC and M&M) and show that it can enhance the predictive performance of traditional predictors.

Chapter 6

Conclusions

In this chapter, we make a summary of the main achievements of this thesis which advance the field of cardiac image analysis and the prediction/diagnosis of the corresponding CVDs through the application of deep learning techniques. Furthermore, this chapter discusses certain inherent limitations of the existing methodology and outlines potential directions for further research to enhance the approaches proposed in this thesis.

6.1 Summary and Achievements

The motivation of this thesis is to improve the performance of automatic CVD analysis by promoting its critical tasks, image registration and segmentation, mesh reconstruction and CVD prediction/diagnosis. The main contributions of the thesis are outlined below.

In Chapter 2, a detailed review of deep learning-based image registration methods is provided. It reviews all the deep learning-based registration methods since 2013, points out the existing drawbacks in this domain, and summarises several possible future directions. The thoroughness of this review provides a solid basis for novice and experienced researchers in the field of medical image registration.

In Chapter 3, we design two image registration methods to address discontinuity-preserving registration problems in deep learning-based image registration. We first propose a novel weakly-supervised registration network, DDIR, significantly outperforming the state-of-the-art, in intrapatient cardiac MR image registration, while achieving discontinuity-preserving registration. Furthermore, to eliminate the requirements of segmentation masks in DDIR, a joint segmentation and discontinuity-preserving registration network, SDDIR, is proposed. Using only moving and fixed images as input, SDDIR can accurately predict segmentation masks and deformation fields that preserve discontinuities, outperforming existing methods.

In Chapter 4, we propose a novel deep learning-based mesh reconstruction network, MR-Net, to archive rapid, precise, and robust cardiac bi-ventricle mesh reconstruction from cardiac contours, significantly outperforming previous mesh reconstruction approaches. We demonstrated that MR-Net can reconstruct accurate and plausible meshes even with contours missing in the input, applicable for various complex scenarios. Furthermore, in conjunction with an established deep learning-based segmentation approach, our MR-Net can reconstruct accurate 3D cardiac meshes directly from the original MR images in real time.

Finally, in Chapter 5, we build a novel multichannel VAE, named MIVAE, to learn the joint latent distribution of cardiac MR images and meshes, enabling the extraction of efficient and explainable biomarkers from cardiac MR images. The new biomarker learnt, named Shape-Aware Image Representation (SAIR), exhibits significant potential for use in CVD prediction and diagnosis. With only learnt SAIR as a feature in CVD diagnosis, we observe highly promising prediction and diagnosis results that outperform those of popular traditional biomarkers. Furthermore, the learnt SAIR features capture information that is not contained within existing biomarkers (e.g. metadata, clinical indices), thereby helping supplement existing biomarkers and improving overall predictive performance. In addition to learning explainable biomarkers, MIVAE can also work as a potential mesh reconstruction approach, which can reconstruct accurate cardiac meshes directly from the corresponding images.

6.2 Limitations and Future Research Directions

In this thesis, we investigate three principal tasks in CVD analysis: cardiac image registration/segmentation, cardiac mesh reconstruction, and CVD prediction/diagnosis. While we have made notable strides and achieved promising outcomes in these areas, surpassing previous approaches, certain challenges remain that require further attention to facilitate better cardiac image analysis and CVD prediction/diagnosis.

In cardiac image registration, the discontinuity-preserving registration problem has not been completely solved. As mentioned above, our proposed DDIR requires segmentation masks for both training and inference, which may not always be applicable due to the unavailability of such masks. SDDIR can achieve accurate segmentation and discontinuity-preserving registration with only moving and fixed images as input, while its performance may be suboptimal when applied to other datasets where segmentation performance may be poor. Further research into robust discontinuity-preserving registration approaches that do not require segmentation masks is warranted to overcome this challenge.

In 3D cardiac mesh reconstruction, unsupervised cardiac mesh reconstruction is a promising direction worthy of exploration. Our MR-Net provides a fast, robust, and accurate cardiac mesh reconstruction approach, while it requires ground-truth cardiac mesh on the network training. Considering the nature of registration, it is possible to build an unsupervised network to reconstruct a 3D cardiac mesh from the given contours. On the other hand, MR-Net needs additional segmentation approaches when reconstructing meshes from the original images, which may introduce additional interference. In our recent work [23], we have achieved cardiac mesh reconstruction directly from the original images and applied it to cardiac image segmentation. This is achieved by using a point distribution model and two deep learning networks to predict shape and transformation parameters, respectively. However, while it can reconstruct smooth cardiac meshes and predict anatomical structure-preserved segmentation results, its segmentation performance is comparable to or slightly inferior to segmentation-specific networks. Furthermore, it requires two separate networks to predict the shape parameters and transformation parameters separately, which leads to redundant architecture and less efficiency. Therefore, there is still a need to design a more efficient and accurate network that can directly reconstruct meshes from the original images, without requiring additional segmentation approaches.

In deep learning-based CVD diagnosis, more efficient image latent representations and more time-relevant features should be considered. In MIVAE, we design a two-channel VAE to learn the joint latent space of cardiac meshes and images. Considering the objective of learning latent representation, it is interesting to try building a three-channel VAE to learn joint latent representations of cardiac meshes, cardiac images, and cardiac segmentation masks. As the cardiac mesh and cardiac segmentation can be derived from each other, such an architecture may provide further promotion than MIVAE. However, most CVD prediction/diagnosis approaches use only MR images in the ED and ES frames, while the remaining frames in the cardiac cycle are not fully exploited. Incorporating these additional frames in CVD image analysis can result in more accurate prediction/diagnosis performance.

A more general challenge is domain adaptation between different datasets. This is a common problem in current deep learning-based approaches, where a network works very well in data similar to the training data, while its performance may significantly decrease when applied to other unseen data. This challenge is particularly relevant to medical image analysis, where images from different scanners and centres can exhibit marked differences in appearance. In our experiments, the networks trained on UKBB also suffer from domain gaps when applied to other datasets such as ACDC and M&M. To address this issue, three directions are worth trying: (1) data augmentation. This is to generate a range of input appearances during the training stage, enabling the model to be more robust to unseen data. (2) pre-processing of the test data. Preprocessing techniques can be used to reduce the gap between the distribution of the training data and that of the unseen data by preprocessing the unseen data into similar attributes (e.g. spacing, light, voxel distribution) to the training data. We have found that resampling and histogram matching can be effective when evaluating the performance of the UKBB-trained model on ACDC and M&M datasets. (3) build domain-invariant representations for subsequent tasks. This should be the essential way to solve the domain adaptation problem, while it is very challenging and few works have explored this aspect.

Another general challenge is the limitation of datasets. In the context of CVD analysis, even medical image analysis, current research tends to rely on small-scale datasets. For example, different from general classification problems in the computer vision domain, there are limited positive samples in medical image prediction/diagnosis. Publicly available data sets typically comprise a limited number of samples (for CVD, generally ~ 100), leading to less reliable and unconvincing results when applied to realistic scenarios. Given that it is infeasible to substantially expand the scale of such datasets to obtain sufficient training data in the near future, it is imperative to pursue the development of explainable diagnosis approaches, rather than simply relying on end-to-end classification networks.

In summary, although deep learning-based approaches have demonstrated remarkable success across a wide range of computer vision applications, successful and precise cardiac image analysis and prediction/diagnosis of CVDs require specific attention to the inherent limitations of DL methods and relevant priors of cardiac imaging, such as cardiac motion, anisotropic image spacing, and the cardiac cycle. Taking these factors into account is both required and critical to achieving more efficient and accurate results in the context of CVD analysis.

6.3 Code and Results Availability

The source codes of our work on discontinuity-preserving registration (DDIR) and mesh reconstruction (MR-Net) are already publicly available on GitHub.

The data used to train, validate, and test our networks in this thesis are mainly from UKBB, and we also test our networks on two publicly available datasets, ACDC and M&M.

Appendix A

Supplementary Material for DDIR in Chapter 3

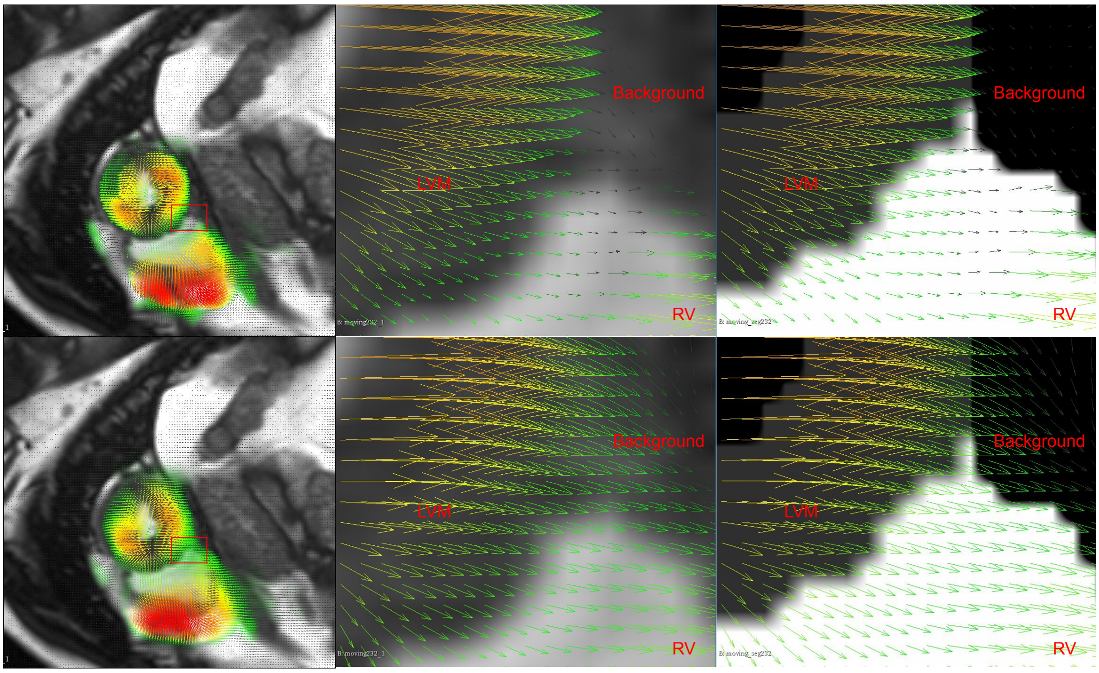

Figure A.1: Visualisation of discontinuity on deformation fields. The first row and second row are the deformation fields predicted by DDIR and DDIR(baseline). The first column is the original deformation fields (vector arrows) overlay moving images. The red box marks the zoom-in regions on the right columns(the second column is deformation arrows on zoom-in images, while the third column is on the corresponding segmentation results). The discontinuity can be found on the interface of LVM, RV and background. The cardiac MR images were reproduced by kind permission of UK Biobank ©.

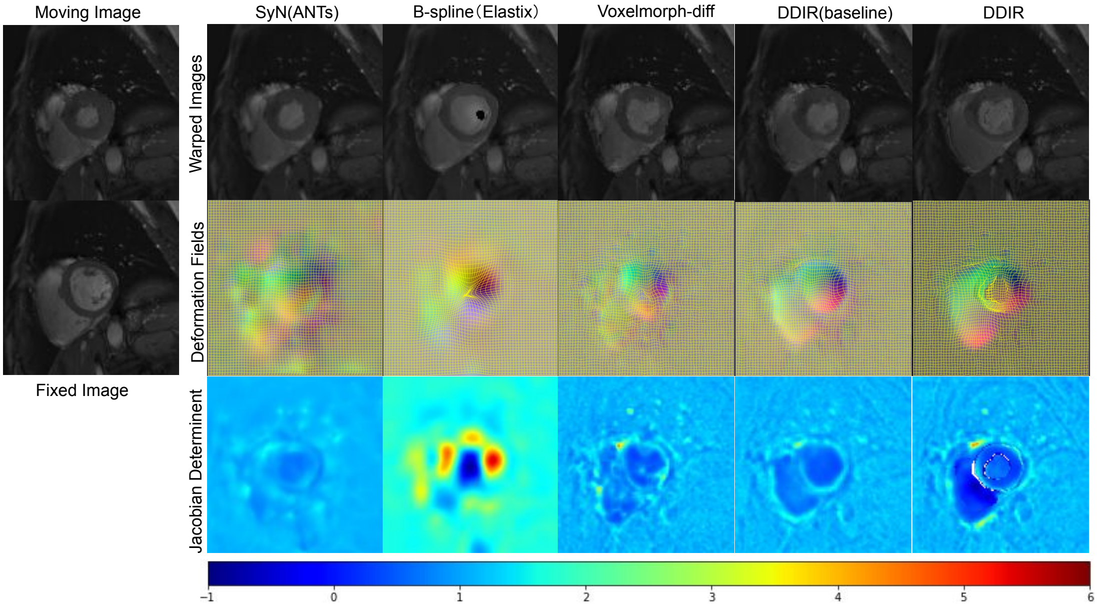

Figure A.2: Visual comparison of results estimated using DDIR and state-of-the-art methods on the ACDC dataset.

Table A.1: Quantitative comparison between DDIR and state-of-the-art methods on the ACDC dataset (the metrics are the same as on the UKBB dataset).

Methods	LVBP DS (%)	LVM DS (%)	RV DS (%)	Avg. DS (%)	HD (mm)	LVEDV	LVMM
before Reg	69.08 ± 14.56	52.50 ± 14.68	66.00 ± 16.36	62.53 ± 11.35	9.67 ± 3.10	51.80 ± 20.88	39.28 ± 15.66
B-spline	77.15 ± 14.38	78.13 ± 7.28	82.73 ± 12.52	79.34 ± 8.79	8.51 ± 3.38	46.38 ± 22.26	**40.52 ± 16.16**
SyN	77.43 ± 14.04	69.82 ± 10.91	77.26 ± 14.19	73.84 ± 10.12	9.18 ± 3.19	44.53 ± 21.99	**42.16 ± 16.74**
Demons	76.30 ± 13.94	76.75 ± 8.45	84.12 ± 10.47	79.05 ± 9.04	8.72 ± 3.39	46.50 ± 21.53	**39.56± 15.89**
Voxelmorph-diff	77.78 ± 12.49	76.17 ± 7.77	84.81 ± 9.79	79.58 ± 7.34	8.74 ± 3.36	46.25 ± 21.41	**40.36 ± 16.41**
VM-Dice	77.56 ± 12.53	76.62 ± 7.76	85.00 ± 9.76	79.73 ± 7.25	8.84 ± 3.34	46.05 ± 21.52	**40.46 ± 16.43**
VM(img+seg)	76.98 ± 12.57	77.23 ± 7.36	85.23 ± 9.72	79.82 ± 7.14	8.82 ± 3.38	45.58 ± 21.62	**41.01 ± 16.89**
VM-Dice(img+seg)	77.38 ± 12.51	77.65 ± 6.59	84.94 ± 10.06	79.99 ± 7.20	8.66 ± 3.34	45.84 ± 21.46	**40.83 ± 16.62**
Voxelmorph-diff(compose)	84.13 ± 9.45	80.62 ± 10.68	83.76 ± 12.29	82.84 ± 8.48	7.82 ± 8.03	46.57 ± 21.46	43.09 ± 17.75
VM-Dice(compose)	84.27 ± 9.18	80.99 ± 9.94	84.66 ± 11.11	83.31 ± 7.86	7.66 ± 7.90	46.69 ± 21.31	42.92 ± 117.57
DDIR(baseline)	80.69 ± 11.84	78.15 ± 7.59	84.26 ± 9.91	81.03 ± 7.00	7.66 ± 3.19	**47.85 ± 21.36**	**41.35 ± 17.20**
DDIR	82.20 ± 11.20	77.04 ± 8.15	84.36 ± 12.20	81.20 ± 8.11	7.92 ± 3.24	**48.23 ± 21.03**	**39.76 ± 16.56**

Appendix B

List of Publications

B.1 Journal Papers

1. Chen X, Diaz-Pinto A, Ravikumar N, Frangi AF. Deep learning in medical image registration[J]. Progress in Biomedical Engineering, 2021, 3(1): 012003.
2. Chen X, Ravikumar N, Xia Y, Attar R, Diaz-Pinto A, Piechnik SK, Neubauer S, Petersen SE, Frangi AF. Shape registration with learned deformations for 3D shape reconstruction from sparse and incomplete point clouds[J]. Medical Image Analysis, 2021, 74: 102228.
3. Xia Y*, Chen X*, Ravikumar N, Christopher K, Attar R, Aung N, Neubauer S, Petersen SE, Frangi AF. Automatic 3D+ t Four-Chamber CMR Quantification of the UK Biobank: integrating imaging and non-imaging data priors at scale[J]. Medical Image Analysis, 2022: 102498 (* denotes joint-first author).
4. Chen X, Xia Y, Ravikumar N, Frangi AF. "Joint Segmentation and Discontinuity-preserving Image Registration using Deep Learning." under review.
5. Chen X, Xia Y, Dall'Armellina E, Ravikumar N, Frangi AF. "Joint shape/texture representation learning for cardiovascular disease diagnosis from MRI." under review.
6. Xia Y, Chen X, Ravikumar N, Frangi AF. "Multi-Contrast MR Image Synthesis from Incomplete Input Data using Multi-channel Adversarial VAEs." under review.

B.2 Conference Papers

1. Chen X, Xia Y, Ravikumar N, Frangi AF. A Deep Discontinuity-Preserving Image Registration Network[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2021: 46-55.

Appendix C

Biosketch

Xiang Chen was born in Hunan, China, in 1994. He earned his Bachelor of Science in Electronics and Information Engineering from Sichuan University in Chengdu in 2016 and his Master of Science in Communication and Information System from the same university in 2019. During his postgraduate studies from 2015 to 2019, he was part of the Computer Vision (CV) group at the Image Information Institute of Sichuan University. During this time, he was the main contributor to several projects in the fields of artificial intelligence and computer vision, such as the Identification of Unsound Wheat Kernels and the Construction of Knowledge Base. In September 2019, he began his PhD studies at the Centre for Computational Imaging and Simulation Technologies in Biomedicine (CISTIB) at the School of Computing, University of Leeds. His research was supervised by Prof. Alejandro Frangi, Dr. Nishant Ravikumar, and Dr. Yan Xia, and he was granted a School of Computing Full-Time Fees and Maintenance PhD Scholarship.

The main aim of his PhD project is to develop automated medical image (mainly cardiac images) analysis methods comprising image registration/segmentation, cardiac mesh reconstruction, and CVD prediction/diagnosis in an accurate, robust and efficient manner, which has led to some other collaborations with clinicians at the Leeds Institute of Cardiac and Metabolic Medicine (LICAMM) at the School of Medicine in Leeds or other external collaborators such as Prof. Stefan Neubauer and Prof. Stefan K. Piechnik (University of Oxford) and Prof. Steffen Petersen (Queen Mary University of London).

He has completed several academic/industrial projects allowing him to publish several journal and conference papers in the field of computer vision, and medical image analysis. Click here to see the complete list of his publications available on Google Scholar (currently, total citations = 200 and h-index = 6). He has served as a reviewer for several leading journals including IEEE Transactions on Medical Imaging and Medical Image Analysis. Since 2019, he has also had various teaching responsibilities as a teaching assistant in several BSc/MSc courses such as Data Mining and Text Analytics, Python Programming, Programming Project, Artificial Intelligence, and Machine Learning.

His main research interests include image synthesis, computer vision, medical image registration, mesh reconstruction, and medical image analysis.

References

[1] A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu, "Unsupervised Learning for Fast Probabilistic Diffeomorphic Registration," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 729–738, Springer, 2018.

[2] A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu, "Unsupervised Learning of Probabilistic Diffeomorphic Registration for Images and Surfaces," *Medical Image Analysis*, vol. 57, pp. 226–236, 2019.

[3] C. W. Tsao, A. W. Aday, Z. I. Almarzooq, A. Alonso, A. Z. Beaton, M. S. Bittencourt, A. K. Boehme, A. E. Buxton, A. P. Carson, Y. Commodore-Mensah, *et al.*, "Heart Disease and Stroke Statistics 2022 Update: a Report from the American Heart Association," *Circulation*, vol. 145, no. 8, pp. e153–e639, 2022.

[4] D. J. Pennell, S. R. Underwood, C. C. Manzara, R. H. Swanton, J. M. Walker, P. J. Ell, and D. B. Longmore, "Magnetic Resonance Imaging During Dobutamine Stress in Coronary Artery Disease," *The American journal of cardiology*, vol. 70, no. 1, pp. 34–40, 1992.

[5] K. R. Nandalur, B. A. Dwamena, A. F. Choudhri, M. R. Nandalur, and R. C. Carlos, "Diagnostic Performance of Stress Cardiac Magnetic Resonance Imaging in the Detection of Coronary Artery Disease: a Meta-analysis," *Journal of the American College of Cardiology*, vol. 50, no. 14, pp. 1343–1353, 2007.

[6] H. Chao, H. Shan, F. Homayounieh, R. Singh, R. D. Khera, H. Guo, T. Su, G. Wang, M. K. Kalra, and P. Yan, "Deep Learning Predicts Cardiovascular Disease Risks from Lung Cancer Screening Low Dose Computed Tomography," *Nature Communications*, vol. 12, no. 1, p. 2963, 2021.

[7] F. S. Villanueva and W. R. Wagner, "Ultrasound Molecular Imaging of Cardiovascular Disease," *Nature Clinical Practice Cardiovascular Medicine*, vol. 5, no. Suppl 2, pp. S26–S32, 2008.

[8] I. Aly, A. Rizvi, W. Roberts, S. Khalid, M. W. Kassem, S. Salandy, M. du Plessis, R. S. Tubbs, and M. Loukas, "Cardiac Ultrasound: an Anatomical and Clinical Review," *Translational Research in Anatomy*, vol. 22, p. 100083, 2021.

[9] R. J. Jaszczak, R. E. Coleman, and C. B. Lim, "SPECT: Single Photon Emission Computed Tomography," *IEEE Transactions on Nuclear Science*, vol. 27, no. 3, pp. 1137–1153, 1980.

[10] O. O. Sogbein, M. Pelletier-Galarneau, T. H. Schindler, L. Wei, R. G. Wells, T. D. Ruddy, *et al.*, "New SPECT and PET Radiopharmaceuticals for Imaging Cardiovascular Disease," *BioMed Research International*, vol. 2014, pp. 1–26, 2014.

[11] X. Zhuang, "Challenges and Methodologies of Fully Automatic Whole Heart Segmentation: a Review," *Journal of Healthcare Engineering*, vol. 4, no. 3, pp. 371–407, 2013.

[12] X. Zhuang and J. Shen, "Multi-scale Patch and Multi-modality Atlases for Whole Heart Segmentation of MRI," *Medical Image Analysis*, vol. 31, pp. 77–87, 2016.

[13] D. Kassop, M. S. Donovan, M. K. Cheezum, B. T. Nguyen, N. B. Gambill, R. Blankstein, and T. C. Villines, "Cardiac Masses on Cardiac CT: a Review," *Current Cardiovascular Imaging Reports*, vol. 7, pp. 1–13, 2014.

[14] D. J. Pennell, "Cardiovascular Magnetic Resonance," *Circulation*, vol. 121, no. 5, pp. 692–705, 2010.

[15] P. Scarborough, K. Wickramasinghe, P. Bhatnagar, and M. Rayner, "Trends in Coronary Heart Disease 1961–2011," *London: British Heart Foundation*, vol. 2011, 2011.

[16] W. H. Organization, *The World Health Report 2002: Reducing Risks, Promoting Healthy Life*. World Health Organization, 2002.

[17] W. H. Organization, *World Health Statistics 2008.* World Health Organization, 2008.

[18] M. A. Morales, M. Van den Boomen, C. Nguyen, J. Kalpathy-Cramer, B. R. Rosen, C. M. Stultz, D. Izquierdo-Garcia, and C. Catana, “DeepStrain: a Deep Learning Workflow for the Automated Characterization of Cardiac Mechanics,” *Frontiers in Cardiovascular Medicine*, vol. 8, p. 730316, 2021.

[19] P. Lu, H. Qiu, C. Qin, W. Bai, D. Rueckert, and J. A. Noble, “Going Deeper into Cardiac Motion Analysis to Model Fine Spatio-temporal Features,” in *Annual Conference on Medical Image Understanding and Analysis*, pp. 294–306, Springer, 2020.

[20] Y. Tsadok, Z. Friedman, B. A. Haluska, R. Hoffmann, and D. Adam, “Myocardial Strain Assessment by Cine Cardiac Magnetic Resonance Imaging Using Non-rigid Registration,” *Magnetic Resonance Imaging*, vol. 34, no. 4, pp. 381–390, 2016.

[21] E. R. Pujadas, Z. Raisi-Estabragh, L. Szabo, C. McCracken, C. I. Morcillo, V. M. Campello, C. Martín-Isla, A. M. Atehortua, H. Vago, B. Merkely, *et al.*, “Prediction of Incident Cardiovascular Events Using Machine Learning and CMR Radiomics,” *European Radiology*, pp. 1–13, 2022.

[22] X. Chen, N. Ravikumar, Y. Xia, R. Attar, A. Diaz-Pinto, S. K. Piechnik, S. Neubauer, S. E. Petersen, and A. F. Frangi, “Shape Registration with Learned Deformations for 3D Shape Reconstruction from Sparse and Incomplete Point Clouds,” *Medical Image Analysis*, vol. 74, p. 102228, 2021.

[23] Y. Xia, X. Chen, N. Ravikumar, C. Kelly, R. Attar, N. Aung, S. Neubauer, S. E. Petersen, and A. F. Frangi, “Automatic 3D+ t Four-chamber CMR Quantification of the UK biobank: Integrating Imaging and Non-imaging Data Priors at Scale,” *Medical Image Analysis*, vol. 80, p. 102498, 2022.

[24] G. A. Bello, T. J. Dawes, J. Duan, C. Biffi, A. De Marvao, L. S. Howard, J. S. R. Gibbs, M. R. Wilkins, S. A. Cook, D. Rueckert, *et al.*, “Deep-learning Cardiac Motion Analysis for Human Survival Prediction,” *Nature Machine Intelligence*, vol. 1, no. 2, pp. 95–104, 2019.

[25] I. Cetin, G. Sanroma, S. E. Petersen, S. Napel, O. Camara, M.-A. G. Ballester, and K. Lekadir, "A Radiomics Approach to Computer-aided Diagnosis with Cardiac Cine-MRI," in *International Workshop on Statistical Atlases and Computational Models of the Heart*, pp. 82–90, Springer, 2017.

[26] U. Neisius, H. El-Rewaidy, S. Nakamori, J. Rodriguez, W. J. Manning, and R. Nezafat, "Radiomic Analysis of Myocardial Native T1 Imaging Discriminates between Hypertensive Heart Disease and Hypertrophic Cardiomyopathy," *JACC: Cardiovascular Imaging*, vol. 12, no. 10, pp. 1946–1954, 2019.

[27] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based Learning Applied to Document Recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[28] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A Fast Learning Algorithm for Deep Belief Nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[29] M. Jaderberg, K. Simonyan, A. Zisserman, *et al.*, "Spatial Transformer Networks," in *Advances in Neural Information Processing Systems*, pp. 2017–2025, 2015.

[30] 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*, pp. 2672–2680, 2014.

[31] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

[32] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-scale Image Recognition," in *International Conference on Learning Representations*, May 2015.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[34] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, 2017.

[35] 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*, pp. 234–241, Springer, 2015.

[36] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image Translation with Conditional Adversarial Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1125–1134, 2017.

[37] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-image Translation Using Cycle-consistent Adversarial Networks," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2223–2232, 2017.

[38] X. Chen, L. Qing, X. He, J. Su, and Y. Peng, "From Eyes to Face Synthesis: A New Approach for Human-centered Smart Surveillance," *IEEE Access*, vol. 6, pp. 14567–14575, 2018.

[39] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Synthetic Data Augmentation Using GAN for Improved Liver Lesion Classification," in *IEEE International Symposium on Biomedical Imaging*, pp. 289–293, IEEE, 2018.

[40] Z. Tang, P.-T. Yap, and D. Shen, "A New Multi-atlas Registration Framework for Multimodal Pathological Images Using Conventional Monomodal Normal Atlases," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2293–2304, 2018.

[41] Z. Han, B. Wei, A. Mercado, S. Leung, and S. Li, "Spine-GAN: Semantic Segmentation of Multiple Spinal Structures," *Medical Image Analysis*, vol. 50, pp. 23–35, 2018.

[42] N. Andrade, F. A. Faria, and F. A. M. Cappabianco, "A Practical Review on Medical Image Registration: From Rigid to Deep Learning Based Approaches," in *SIBGRAPI Conference on Graphics, Patterns and Images*, pp. 463–470, IEEE, 2018.

[43] G. Haskins, U. Kruger, and P. Yan, “Deep Learning in Medical Image Registration: A Survey,” *Machine Vision and Applications*, vol. 31, pp. 1–18, 2020.

[44] Y. Fu, Y. Lei, T. Wang, W. J. Curran, T. Liu, and X. Yang, “Deep Learning in Medical Image Registration: a Review,” *Physics in Medicine & Biology*, vol. 65, no. 20, p. 20TR01, 2020.

[45] N. J. Tustison, B. B. Avants, and J. C. Gee, “Learning Image-based Spatial Transformations via Convolutional Neural Networks: A Review,” *Magnetic Resonance Imaging*, vol. 64, pp. 142–153, 2019.

[46] G. Haskins, J. Kruecker, U. Kruger, S. Xu, P. A. Pinto, B. J. Wood, and P. Yan, “Learning Deep Similarity Metric for 3D MR–TRUS Image Registration,” *International Journal of Computer Assisted Radiology and Surgery*, vol. 14, no. 3, pp. 417–425, 2019.

[47] N. Zhu, M. Najafi, B. Han, S. Hancock, and D. Hristov, “Feasibility of Image Registration for Ultrasound-Guided Prostate Radiotherapy Based on Similarity Measurement by A Convolutional Neural Network,” *Technology in Cancer Research & Treatment*, vol. 18, pp. 1–11, 2019.

[48] J. Fan, X. Cao, Z. Xue, P.-T. Yap, and D. Shen, “Adversarial Similarity Network For Evaluating Image Alignment in Deep Learning Based Registration,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 739–746, Springer, 2018.

[49] J. Fan, X. Cao, Q. Wang, P.-T. Yap, and D. Shen, “Adversarial Learning for Mono-or Multi-modal Registration,” *Medical Image Analysis*, vol. 58, p. 101545, 2019.

[50] C. Qin, B. Shi, R. Liao, T. Mansi, D. Rueckert, and A. Kamen, “Unsupervised Deformable Registration for Multi-modal Images via Disentangled Representations,” in *International Conference on Information Processing in Medical Imaging*, pp. 249–261, Springer, 2019.

[51] D. Mahapatra and Z. Ge, “Training Data Independent Image Registration with GANs Using Transfer Learning and Segmentation Information,” in *IEEE International Symposium on Biomedical Imaging*, pp. 709–713, IEEE, 2019.

[52] X. Yang, R. Kwitt, and M. Niethammer, “Fast Predictive Image Registration,” in *Deep Learning and Data Labeling for Medical Applications*, pp. 48–57, Springer, 2016.

[53] X. Cao, J. Yang, J. Zhang, D. Nie, M. Kim, Q. Wang, and D. Shen, “Deformable Image Registration Based on Similarity-steered CNN Regression,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 300–308, Springer, 2017.

[54] X. Cao, J. Yang, J. Zhang, Q. Wang, P.-T. Yap, and D. Shen, “Deformable Image Registration Using A Cue-aware Deep Regression Network,” *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 9, pp. 1900–1911, 2018.

[55] J. Fan, X. Cao, P.-T. Yap, and D. Shen, “BIRNet: Brain Image Registration Using Dual-supervised Fully Convolutional Networks,” *Medical Image Analysis*, vol. 54, pp. 193–206, 2019.

[56] M.-M. Rohé, M. Datar, T. Heimann, M. Sermesant, and X. Pennec, “SVF-Net: Learning Deformable Image Registration Using Shape Matching,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 266–274, Springer, 2017.

[57] K. A. Eppenhof and J. P. Pluim, “Pulmonary CT Registration Through Supervised Learning with Convolutional Neural Networks,” *IEEE Transactions on Medical Imaging*, vol. 38, no. 5, pp. 1097–1105, 2018.

[58] K. A. Eppenhof, M. W. Lafarge, M. Veta, and J. P. Pluim, “Progressively Trained Convolutional Neural Networks for Deformable Image Registration,” *IEEE Transactions on Medical Imaging*, vol. 39, no. 5, pp. 1594–1604, 2019.

[59] H. Sokooti, B. de Vos, F. Berendsen, B. P. Lelieveldt, I. Išgum, and M. Staring, “Nonrigid Image Registration Using Multi-scale 3D Convolutional Neural Networks,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 232–239, Springer, 2017.

[60] S. S. M. Salehi, S. Khan, D. Erdogmus, and A. Gholipour, “Real-time Deep Pose Estimation with Geodesic Loss for Image-to-template Rigid Registration,” *IEEE Transactions on Medical Imaging*, vol. 38, no. 2, pp. 470–481, 2018.

[61] Y. Xia, Y. Li, L. Xun, Q. Yan, and D. Zhang, "A Convolutional Neural Network Cascade for Plantar Pressure Images Registration," *Gait & Posture*, vol. 68, pp. 403–408, 2019.

[62] L. Zhao and K. Jia, "Deep Adaptive Log-demons: Diffeomorphic Image Registration with Very Large Deformations," *Computational and Mathematical Methods in Medicine*, vol. 2015, pp. 836202:1–836202:16, 2015.

[63] X. Yang, R. Kwitt, M. Styner, and M. Niethammer, "Fast Predictive Multimodal Image Registration," in *IEEE International Symposium on Biomedical Imaging*, pp. 858–862, IEEE, 2017.

[64] X. Yang, R. Kwitt, M. Styner, and M. Niethammer, "Quicksilver: Fast Predictive Image Registration–A Deep Learning Approach," *NeuroImage*, vol. 158, pp. 378–396, 2017.

[65] P. Yan, S. Xu, A. R. Rastinehad, and B. J. Wood, "Adversarial Image Registration with Application for MR and TRUS Image Fusion," in *International Workshop on Machine Learning in Medical Imaging*, pp. 197–204, Springer, 2018.

[66] A. Sedghi, T. Kapur, J. Luo, P. Mousavi, and W. M. Wells, "Probabilistic Image Registration via Deep Multi-class Classification: Characterizing Uncertainty," in *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging and Clinical Image-Based Procedures*, pp. 12–22, Springer, 2019.

[67] Z. Yao, H. Feng, Y. Song, S. Li, Y. Yang, L. Liu, and C. Liu, "A Supervised Network for Fast Image-guided Radiotherapy (IGRT) Registration," *Journal of Medical Systems*, vol. 43, pp. 1–8, 2019.

[68] H. Liao, W.-A. Lin, J. Zhang, J. Zhang, J. Luo, and S. K. Zhou, "Multiview 2D/3D Rigid Registration via A Point-Of-Interest Network for Tracking and Triangulation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12638–12647, 2019.

[69] H. Guo, M. Kruger, S. Xu, B. J. Wood, and P. Yan, "Deep Adaptive Registration of Multi-modal Prostate Images," *Computerized Medical Imaging and Graphics*, vol. 84, p. 101769, 2020.

[70] B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein, and J. C. Gee, "A Reproducible Evaluation of ANTs Similarity Metric Performance in Brain Image Registration," *Neuroimage*, vol. 54, no. 3, pp. 2033–2044, 2011.

[71] M. Lorenzi, N. Ayache, G. B. Frisoni, X. Pennec, A. D. N. I. (ADNI, *et al.*, "LCC-Demons: A Robust and Accurate Symmetric Diffeomorphic Registration Algorithm," *NeuroImage*, vol. 81, pp. 470–483, 2013.

[72] B. D. de Vos, F. F. Berendsen, M. A. Viergever, M. Staring, and I. Išgum, "End-to-end Unsupervised Deformable Image Registration with A Convolutional Neural Network," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 204–212, Springer, 2017.

[73] J. Lv, M. Yang, J. Zhang, and X. Wang, "Respiratory Motion Correction for Free-breathing 3D Abdominal MRI Using CNN-based Image Registration: A Feasibility Study," *The British Journal of Radiology*, vol. 91, p. 20170788, 2018.

[74] H. Li and Y. Fan, "Non-rigid Image Registration Using Self-supervised Fully Convolutional Networks Without Training Data," in *IEEE International Symposium on Biomedical Imaging*, pp. 1075–1078, IEEE, 2018.

[75] 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*, pp. 9252–9260, 2018.

[76] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "Voxelmorph: A Learning Framework for Deformable Medical Image Registration," *IEEE Transactions on Medical Imaging*, vol. 38, no. 8, pp. 1788–1800, 2019.

[77] Z. Zhu, Y. Cao, C. Qin, Y. Rao, D. Ni, and Y. Wang, "Unsupervised 3D End-to-end Deformable Network for Brain MRI Registration," in *International Conference of the IEEE Engineering in Medicine & Biology Society*, pp. 1355–1359, IEEE, 2020.

[78] Y. Fu, Y. Lei, T. Wang, K. Higgins, J. D. Bradley, W. J. Curran, T. Liu, and X. Yang, "LungRegNet: An Unsupervised Deformable Image Registration Method for 4D-CT Lung," *Medical Physics*, vol. 47, no. 4, pp. 1763–1774, 2020.

[79] C. Stergios, S. Mihir, V. Maria, C. Guillaume, R. Marie-Pierre, M. Stavroula, and P. Nikos, “Linear and Deformable Image Registration with 3D Convolutional Neural Networks,” in *Image Analysis for Moving Organ, Breast, and Thoracic Images*, pp. 13–22, Springer, 2018.

[80] D. Kuang and T. Schmah, “Faim–A Convnet Method for Unsupervised 3D Medical Image Registration,” in *International Workshop on Machine Learning in Medical Imaging*, pp. 646–654, Springer, 2019.

[81] S. Ali and J. Rittscher, “Conv2Warp: An Unsupervised Deformable Image Registration with Continuous Convolution and Warping,” in *International Workshop on Machine Learning in Medical Imaging*, pp. 489–497, Springer, 2019.

[82] X. Hu, M. Kang, W. Huang, M. R. Scott, R. Wiest, and M. Reyes, “Dual-Stream Pyramid Registration Network,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 382–390, Springer, 2019.

[83] R. Bhalodia, S. Y. Elhabian, L. Kavan, and R. T. Whitaker, “A Cooperative Autoencoder for Population-Based Regularization of CNN Image Registration,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 391–400, Springer, 2019.

[84] Y. Sang and D. Ruan, “Enhanced Image Registration With a Network Paradigm and Incorporation of a Deformation Representation Model,” in *International Symposium on Biomedical Imaging*, pp. 91–94, IEEE, 2020.

[85] T. Fechter and D. Baltas, “One Shot Learning for Deformable Medical Image Registration and Periodic Motion Tracking,” *IEEE Transactions on Medical Imaging*, vol. 39, no. 7, pp. 2506–2517, 2020.

[86] D. Gu, X. Cao, S. Ma, L. Chen, G. Liu, D. Shen, and Z. Xue, “Pair-Wise and Group-Wise Deformation Consistency in Deep Registration Network,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 171–180, Springer, 2020.

[87] B. Kim, J. Kim, J.-G. Lee, D. H. Kim, S. H. Park, and J. C. Ye, “Unsupervised Deformable Image Registration Using Cycle-Consistent CNN,” in *International*

Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 166–174, Springer, 2019.

[88] J. Krebs, H. Delingette, B. Mailhé, N. Ayache, and T. Mansi, "Learning A Probabilistic Model for Diffeomorphic Registration," *IEEE Transactions on Medical Imaging*, vol. 38, no. 9, pp. 2165–2176, 2019.

[89] L. Liu, X. Hu, L. Zhu, and P.-A. Heng, "Probabilistic Multilayer Regularization Network for Unsupervised 3D Brain Image Registration," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 346–354, Springer, 2019.

[90] Z. Shen, X. Han, Z. Xu, and M. Niethammer, "Networks for Joint Affine and Non-parametric Image Registration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4224–4233, 2019.

[91] Z. Shen, F.-X. Vialard, and M. Niethammer, "Region-specific Diffeomorphic Metric Mapping," *Advances in Neural Information Processing Systems*, vol. 32, pp. 1–11, 2019.

[92] M. Niethammer, R. Kwitt, and F.-X. Vialard, "Metric Learning for Image Registration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8463–8472, 2019.

[93] B. D. de Vos, F. F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, and I. Išgum, "A Deep Learning Framework for Unsupervised Affine and Deformable Image Registration," *Medical Image Analysis*, vol. 52, pp. 128–143, 2019.

[94] S. Zhao, Y. Dong, E. I. Chang, Y. Xu, *et al.*, "Recursive Cascaded Networks for Unsupervised Medical Image Registration," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 10600–10610, 2019.

[95] S. Zhao, T. Lau, J. Luo, I. Eric, C. Chang, and Y. Xu, "Unsupervised 3D End-to-End Medical Image Registration with Volume Tweening Network," *IEEE journal of biomedical and health informatics*, vol. 24, no. 5, pp. 1394–1404, 2019.

[96] X. Cao, J. Yang, L. Wang, Z. Xue, Q. Wang, and D. Shen, "Deep Learning Based Inter-modality Image Registration Supervised by Intra-modality Similarity," in

International Workshop on Machine Learning in Medical Imaging, pp. 55–63, Springer, 2018.

[97] Z. Jiang, F.-F. Yin, Y. Ge, and L. Ren, "A Multi-scale Framework with Unsupervised Joint Training of Convolutional Neural Networks for Pulmonary Deformable Image Registration," *Physics in Medicine & Biology*, vol. 65, no. 1, p. 015011, 2020.

[98] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee, "Symmetric Diffeomorphic Image Registration with Cross-correlation: Evaluating Automated Labeling of Elderly and Neurodegenerative Brain," *Medical Image Analysis*, vol. 12, no. 1, pp. 26–41, 2008.

[99] A. Hering, S. Kuckertz, S. Heldmann, and M. P. Heinrich, "Enhancing Label-driven Deep Deformable Image Registration with Local Distance Metrics for State-of-the-art Cardiac Motion Tracking," in *Bildverarbeitung für die Medizin 2019*, pp. 309–314, Springer, 2019.

[100] M. P. Heinrich, "Closing the Gap Between Deep and Conventional Image Registration Using Probabilistic Dense Displacement Networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 50–58, Springer, 2019.

[101] Z. Xu and M. Niethammer, "DeepAtlas: Joint Semi-supervised Learning of Image Registration and Segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 420–429, Springer, 2019.

[102] L. Chen, X. Cao, L. Chen, Y. Gao, D. Shen, Q. Wang, and Z. Xue, "Semantic Hierarchy Guided Registration Networks for Intra-subject Pulmonary CT Image Alignment," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 181–189, Springer, 2020.

[103] I. Y. Ha, M. Wilms, and M. Heinrich, "Semantically Guided Large Deformation Estimation with Deep Networks," *Sensors*, vol. 20, no. 5, pp. 1392:1–1392:13, 2020.

[104] L. Mansilla, D. H. Milone, and E. Ferrante, “Learning Deformable Registration of Medical Images with Anatomical Constraints,” *Neural Networks*, vol. 124, pp. 269–279, 2020.

[105] Y. Hu, M. Modat, E. Gibson, N. Ghavami, E. Bonmati, C. M. Moore, M. Emberton, J. A. Noble, D. C. Barratt, and T. Vercauteren, “Label-driven Weakly-supervised Learning for Multimodal Deformable Image Registration,” in *IEEE International Symposium on Biomedical Imaging*, pp. 1070–1074, IEEE, 2018.

[106] Y. Hu, M. Modat, E. Gibson, W. Li, N. Ghavami, E. Bonmati, G. Wang, S. Bandula, C. M. Moore, M. Emberton, *et al.*, “Weakly-supervised Convolutional Neural Networks for Multimodal Image Registration,” *Medical Image Analysis*, vol. 49, pp. 1–13, 2018.

[107] A. Hering, S. Kuckertz, S. Heldmann, and M. P. Heinrich, “Memory-efficient 2.5 D Convolutional Transformer Networks for Multi-modal Deformable Registration with Weak Label Supervision Applied to Whole-heart CT and MRI Scans,” *International Journal of Computer Assisted Radiology and Surgery*, vol. 14, no. 11, pp. 1901–1912, 2019.

[108] G. Wu, M. Kim, Q. Wang, Y. Gao, S. Liao, and D. Shen, “Unsupervised Deep Feature Learning for Deformable Registration of MR Brain Images,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 649–656, Springer, 2013.

[109] G. Wu, M. Kim, Q. Wang, B. C. Munsell, and D. Shen, “Scalable High-performance Image Registration Framework by Unsupervised Deep Feature Representations Learning,” *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1505–1516, 2015.

[110] V. Kearney, S. Haaf, A. Sudhyadhom, G. Valdes, and T. D. Solberg, “An Unsupervised Convolutional Neural Network-based Algorithm for Deformable Image Registration,” *Physics in Medicine & Biology*, vol. 63, no. 18, p. 185017, 2018.

[111] X. Zhu, M. Ding, T. Huang, X. Jin, and X. Zhang, “PCANet-Based Structural Representation for Nonrigid Multimodal Medical Image Registration,” *Sensors*, vol. 18, no. 5, pp. 1477:1–1477:11, 2018.

[112] M. Blendowski and M. P. Heinrich, “Combining MRF-based Deformable Registration and Deep Binary 3D-CNN Descriptors for Large Lung Motion Estimation in COPD Patients,” *International Journal of Computer Assisted Radiology and Surgery*, vol. 14, no. 1, pp. 43–52, 2019.

[113] J. Zheng, S. Miao, Z. J. Wang, and R. Liao, “Pairwise Domain Adaptation Module for CNN-based 2-D/3-D Registration,” *Journal of Medical Imaging*, vol. 5, no. 2, pp. 021204–021204, 2018.

[114] L. Canalini, J. Klein, D. Miller, and R. Kikinis, “Segmentation-based Registration of Ultrasound Volumes for Glioma Resection in Image-guided Neurosurgery,” *International Journal of Computer Assisted Radiology and Surgery*, vol. 14, no. 10, pp. 1697–1713, 2019.

[115] X. Yang, X. Han, E. Park, S. Aylward, R. Kwitt, and M. Niethammer, “Registration of Pathological Images,” in *International Workshop on Simulation and Synthesis in Medical Imaging*, pp. 97–107, Springer, 2016.

[116] X. Liu, D. Jiang, M. Wang, and Z. Song, “Image Synthesis-based Multi-modal Image Registration Framework by Using Deep Fully Convolutional Networks,” *Medical & Biological Engineering & Computing*, vol. 57, no. 5, pp. 1037–1048, 2019.

[117] C. Liu, Z. Lu, L. Ma, L. Wang, X. Jin, and W. Si, “A Modality Conversion Approach to MV-DRs and KV-DRRs Registration Using Information Bottlenecked Conditional Generative Adversarial Network,” *Medical Physics*, vol. 46, no. 10, pp. 4575–4587, 2019.

[118] M. C. Lee, O. Oktay, A. Schuh, M. Schaap, and B. Glocker, “Image-and-spatial Transformer Networks for Structure-guided Image Registration,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 337–345, Springer, 2019.

[119] M. Blendowski, N. Bouteldja, and M. P. Heinrich, “Multimodal 3D Medical Image Registration Guided by Shape Encoder–decoder Networks,” *International Journal of Computer Assisted Radiology and Surgery*, vol. 15, pp. 269–276, 2020.

[120] R. Liao, S. Miao, P. de Tournemire, S. Grbic, A. Kamen, T. Mansi, and D. Comaniciu, "An Artificial Agent for Robust Image Registration," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, pp. 4168–4175, 2017.

[121] D. Toth, S. Miao, T. Kurzendorfer, C. A. Rinaldi, R. Liao, T. Mansi, K. Rhode, and P. Mountney, "3D/2D Model-to-image Registration by Imitation Learning for Cardiac Procedures," *International Journal of Computer Assisted Radiology and Surgery*, vol. 13, no. 8, pp. 1141–1149, 2018.

[122] S. Miao, S. Piat, P. Fischer, A. Tuysuzoglu, P. Mewes, T. Mansi, and R. Liao, "Dilated FCN for Multi-Agent 2D/3D Medical Image Registration," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 4694–4701, 2018.

[123] D. Shen, "Image Registration by Local Histogram Matching," *Pattern Recognition*, vol. 40, no. 4, pp. 1161–1172, 2007.

[124] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, "PCANet: A Simple Deep Learning Baseline for Image Classification?," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5017–5032, 2015.

[125] J. Krebs, T. Mansi, N. Ayache, and H. Delingette, "Probabilistic Motion Modeling from Medical Image Sequences: Application to Cardiac Cine-MRI," in *Statistical Atlases and Computational Models of the Heart. Multi-Sequence CMR Segmentation, CRT-EPiggy and LV Full Quantification Challenges*, pp. 176–185, Springer International Publishing, 2020.

[126] A. Dalca, M. Rakic, J. Guttag, and M. Sabuncu, "Learning Conditional Deformable Templates with Convolutional Networks," in *Advances in Neural Information Processing Systems*, pp. 804–816, 2019.

[127] A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu, "Unsupervised Learning of Probabilistic Diffeomorphic Registration for Images and Surfaces," *Medical Image Analysis*, vol. 57, pp. 226–236, 2019.

[128] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. R. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett, "Ways Toward an Early Dia-

gnosis in Alzheimer's Disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI)," *Alzheimer's & Dementia*, vol. 1, no. 1, pp. 55–66, 2005.

[129] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults," *Journal of Cognitive Neuroscience*, vol. 19, no. 9, pp. 1498–1507, 2007.

[130] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, *et al.*, "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.

[131] A. Klein, J. Andersson, B. A. Ardekani, J. Ashburner, B. Avants, M.-C. Chiang, G. E. Christensen, D. L. Collins, J. Gee, P. Hellier, *et al.*, "Evaluation of 14 Nonlinear Deformation Algorithms Applied to Human Brain MRI Registration," *NeuroImage*, vol. 46, no. 3, pp. 786–802, 2009.

[132] H. C. Hazlett, H. Gu, B. C. Munsell, S. H. Kim, M. Styner, J. J. Wolff, J. T. Elison, M. R. Swanson, H. Zhu, K. N. Botteron, *et al.*, "Early Brain Development in Infants at High Risk for Autism Spectrum Disorder," *Nature*, vol. 542, no. 7641, pp. 348–351, 2017.

[133] A. Di Martino, C.-G. Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, J. S. Anderson, M. Assaf, S. Y. Bookheimer, M. Dapretto, *et al.*, "The Autism Brain Imaging Data Exchange: Towards A Large-scale Evaluation of the Intrinsic Brain Architecture in Autism," *Molecular Psychiatry*, vol. 19, no. 6, pp. 659–667, 2014.

[134] M. P. Milham, D. Fair, M. Mennes, S. H. Mostofsky, *et al.*, "The ADHD-200 Consortium: A Model to Advance the Translational Potential of Neuroimaging in Clinical Neuroscience," *Frontiers in Systems Neuroscience*, vol. 6, pp. 62:1–62:5, 2012.

[135] R. L. Gollub, J. M. Shoemaker, M. D. King, T. White, S. Ehrlich, S. R. Sponheim, V. P. Clark, J. A. Turner, B. A. Mueller, V. Magnotta, *et al.*, "The MCIC Collection: A Shared Repository of Multi-modal, Multi-site Brain Image Data from A Clinical Investigation of Schizophrenia," *Neuroinformatics*, vol. 11, no. 3, pp. 367–388, 2013.

[136] K. Marek, D. Jennings, S. Lasch, A. Siderowf, C. Tanner, T. Simuni, C. Coffey, K. Kieburtz, E. Flagg, S. Chowdhury, *et al.*, “The Parkinson Progression Marker Initiative (PPMI),” *Progress in Neurobiology*, vol. 95, no. 4, pp. 629–635, 2011.

[137] A. Dagley, M. LaPoint, W. Huijbers, T. Hedden, D. G. McLaren, J. P. Chatwal, K. V. Papp, R. E. Amariglio, D. Blacker, D. M. Rentz, *et al.*, “Harvard Aging Brain Study: Dataset and Accessibility,” *NeuroImage*, vol. 144, pp. 255–258, 2017.

[138] A. J. Holmes, M. O. Hollinshead, T. M. O’Keefe, V. I. Petrov, G. R. Fariello, L. L. Wald, B. Fischl, B. R. Rosen, R. W. Mair, J. L. Roffman, *et al.*, “Brain Genomics Superstruct Project Initial Data Release with Structural, Functional, and Behavioral Measures,” *Scientific Data*, vol. 2, p. 150031, 2015.

[139] B. Fischl, “Freesurfer,” *NeuroImage*, vol. 62, no. 2, pp. 774–781, 2012.

[140] A. Klein and J. Tourville, “101 Labeled Brain Images and A Consistent Human Cortical Labeling Protocol,” *Frontiers in Neuroscience*, vol. 6, pp. 171:1–171:12, 2012.

[141] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos, “Advancing the Cancer Genome Atlas Glioma MRI Collections with Expert Segmentation Labels and Radiomic Features,” *Scientific Data*, vol. 4, p. 170117, 2017.

[142] C. A. Cocosco, V. Kollokian, R. K.-S. Kwan, G. B. Pike, and A. C. Evans, “Brainweb: Online Interface to A 3D MRI Simulated Brain Database,” in *NeuroImage*, Citeseer, 1997.

[143] L. Mercier, R. F. Del Maestro, K. Petrecca, D. Araujo, C. Haegelen, and D. L. Collins, “Online Database of Clinical MR and Ultrasound Images of Brain Tumors,” *Medical Physics*, vol. 39, no. 6, pp. 3253–3261, 2012.

[144] Y. Xiao, M. Fortin, G. Unsgård, H. Rivaz, and I. Reinertsen, “RE troSpective Evaluation of Cerebral Tumors (RESECT): A Clinical Database of Pre-operative MRI and Intra-operative Ultrasound in Low-grade Glioma Surgeries,” *Medical Physics*, vol. 44, no. 7, pp. 3875–3882, 2017.

[145] L. Sun and S. Zhang, "Deformable MRI-ultrasound Registration Using 3D Convolutional Neural Network," in *Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation*, pp. 152–158, Springer, 2018.

[146] J. Hong and H. Park, "Non-linear Approach for MRI to Intra-operative US Registration Using Structural Skeleton," in *Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation*, pp. 138–145, Springer, 2018.

[147] P. Radau, Y. Lu, K. Connelly, G. Paul, A. Dick, and G. Wright, "Evaluation Framework for Algorithms Segmenting Short Axis Cardiac MRI," *The MIDAS Journal-Cardiac MR Left Ventricle Segmentation Challenge*, vol. 49, 2009.

[148] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester, *et al.*, "Deep Learning Techniques for Automatic MRI Cardiac Multi-structures Segmentation and Diagnosis: Is the Problem Solved?," *IEEE Transactions on Medical Imaging*, vol. 37, no. 11, pp. 2514–2525, 2018.

[149] K. Murphy, B. Van Ginneken, J. M. Reinhardt, S. Kabus, K. Ding, X. Deng, K. Cao, K. Du, G. E. Christensen, V. Garcia, *et al.*, "Evaluation of Registration Methods on Thoracic CT: the EMPIRE10 Challenge," *IEEE Transactions on Medical Imaging*, vol. 30, no. 11, pp. 1901–1920, 2011.

[150] T. Heimann, B. Van Ginneken, M. A. Styner, Y. Arzhaeva, V. Aurich, C. Bauer, A. Beck, C. Becker, R. Beichel, G. Bekes, *et al.*, "Comparison and Evaluation of Methods for Liver Segmentation from CT Datasets," *IEEE Transactions on Medical Imaging*, vol. 28, no. 8, pp. 1251–1265, 2009.

[151] E. A. Regan, J. E. Hokanson, J. R. Murphy, B. Make, D. A. Lynch, T. H. Beaty, D. Curran-Everett, E. K. Silverman, and J. D. Crapo, "Genetic Epidemiology of COPD (COPDGene) Study Design," *COPD: Journal of Chronic Obstructive Pulmonary Disease*, vol. 7, no. 1, pp. 32–43, 2011.

[152] N. L. S. T. R. Team, "Reduced Lung-cancer Mortality with Low-dose Computed Tomographic Screening," *New England Journal of Medicine*, vol. 365, no. 5, pp. 395–409, 2011.

[153] R. Castillo, E. Castillo, D. Fuentes, M. Ahmad, A. M. Wood, M. S. Ludwig, and T. Guerrero, "A Reference Dataset for Deformable Image Registration Spatial Accuracy Evaluation Using the COPDgene Study Archive," *Physics in Medicine & Biology*, vol. 58, no. 9, pp. 2861–2877, 2013.

[154] R. Castillo, E. Castillo, R. Guerra, V. E. Johnson, T. McPhail, A. K. Garg, and T. Guerrero, "A Framework for Evaluation of Deformable Image Registration Spatial Accuracy Using Large Landmark Point Sets," *Physics in Medicine & Biology*, vol. 54, no. 7, pp. 1849–1870, 2009.

[155] C.-C. Shieh, Y. Gonzalez, B. Li, X. Jia, S. Rit, C. Mory, M. Riblett, G. Hugo, Y. Zhang, Z. Jiang, *et al.*, "SPARE: Sparse-view Reconstruction Challenge for 4D Cone-beam CT from A 1-min Scan," *Medical Physics*, vol. 46, no. 9, pp. 3799–3811, 2019.

[156] J. Vandemeulebroucke, S. Rit, J. Kybic, P. Clarysse, and D. Sarrut, "Spatiotemporal Motion Estimation for Respiratory-correlated Imaging of the Lungs," *Medical Physics*, vol. 38, no. 1, pp. 166–178, 2011.

[157] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, *et al.*, "The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans," *Medical Physics*, vol. 38, no. 2, pp. 915–931, 2011.

[158] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-supervised Classification and Localization of Common Thorax Diseases," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2097–2106, 2017.

[159] 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.

[160] 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.

[161] O. Jimenez-del Toro, H. Müller, M. Krenn, K. Gruenberg, A. A. Taha, M. Winterstein, I. Eggel, A. Foncubierta-Rodríguez, O. Goksel, A. Jakab, *et al.*, "Cloud-based Evaluation of Anatomical Structure Segmentation and Landmark Detection Algorithms: VISCERAL Anatomy Benchmarks," *IEEE Transactions on Medical Imaging*, vol. 35, no. 11, pp. 2459–2475, 2016.

[162] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. Pluim, "Elastix: A Toolbox for Intensity-based Medical Image Registration," *IEEE Transactions on Medical Imaging*, vol. 29, no. 1, pp. 196–205, 2009.

[163] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache, "Symmetric Log-domain Diffeomorphic Registration: A Demons-based Approach," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 754–761, Springer, 2008.

[164] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache, "Diffeomorphic demons: Efficient Non-parametric Image Registration," *NeuroImage*, vol. 45, no. 1, pp. S61–S72, 2009.

[165] X. Ma, Y. Niu, L. Gu, Y. Wang, Y. Zhao, J. Bailey, and F. Lu, "Understanding Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems," *Pattern Recognition*, vol. 110, p. 107332, 2021.

[166] R. Hua, J. M. Pozo, Z. A. Taylor, and A. F. Frangi, "Multiresolution eXtended Free-Form Deformations (XFFD) for Non-rigid Registration with Discontinuous Transforms," *Medical Image Analysis*, vol. 36, pp. 113–122, 2017.

[167] R. Hua, *Non-rigid Medical Image Registration with Extended Free Form Deformations: Modelling General Tissue Transitions.* PhD thesis, University of Sheffield, 2016.

[168] Z. Wu, E. Rietzel, V. Boldea, D. Sarrut, and G. C. Sharp, "Evaluation of Deformable Registration of Patient Lung 4DCT with Subanatomical Region Segmentations," *Medical Physics*, vol. 35, no. 2, pp. 775–781, 2008.

[169] A. Schmidt-Richberg, R. Werner, H. Handels, and J. Ehrhardt, "Estimation of Slipping Organ Motion by Registration with Direction-dependent Regularization," *Medical Image Analysis*, vol. 16, no. 1, pp. 150–159, 2012.

[170] D. F. Pace, S. R. Aylward, and M. Niethammer, "A Locally Adaptive Regularization based on Anisotropic Diffusion for Deformable Image Registration of Sliding Organs," *IEEE Transactions on Medical Imaging*, vol. 32, no. 11, pp. 2114–2126, 2013.

[171] E. Ng and M. Ebrahimi, "An Unsupervised Learning Approach to Discontinuity-Preserving Image Registration," in *International Workshop on Biomedical Image Registration*, pp. 153–162, Springer, 2020.

[172] W. Bai, M. Sinclair, G. Tarroni, O. Oktay, M. Rajchl, G. Vaillant, A. M. Lee, N. Aung, E. Lukaschuk, M. M. Sanghvi, *et al.*, "Automated Cardiovascular Magnetic Resonance Image Analysis with Fully Convolutional Networks," *Journal of Cardiovascular Magnetic Resonance*, vol. 20, no. 1, pp. 1–12, 2018.

[173] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," in *International Conference on 3D Vision*, pp. 565–571, IEEE, 2016.

[174] T. Vercauteren, X. Pennec, A. Perchant, N. Ayache, *et al.*, "Diffeomorphic Demons Using ITK Finite Difference Solver Hierarchy," *The Insight Journal*, vol. 1, pp. 1–8, 2007.

[175] K. Marstal, F. Berendsen, M. Staring, and S. Klein, "SimpleElastix: A User-friendly, Multi-lingual Library for Medical Image Registration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 134–142, 2016.

[176] B. D. d. Vos, F. F. Berendsen, M. A. Viergever, M. Staring, and I. Išgum, "End-to-end Unsupervised Deformable Image Registration with A Convolutional Neural

Network," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 204–212, Springer, 2017.

[177] X. Chen, A. Diaz-Pinto, N. Ravikumar, and A. F. Frangi, "Deep Learning in Medical Image Registration," *Progress in Biomedical Engineering*, vol. 3, no. 1, p. 012003, 2021.

[178] D. Demirović, A. Šerifović-Trbalić, N. Prljača, and P. C. Cattin, "Bilateral Filter Regularized Accelerated Demons for improved Discontinuity Preserving Registration," *Computerized Medical Imaging and Graphics*, vol. 40, pp. 94–99, 2015.

[179] V. Vishnevskiy, T. Gass, G. Szekely, C. Tanner, and O. Goksel, "Isotropic Total Variation Regularization of Displacements in Parametric Image Registration," *IEEE Transactions on Medical Imaging*, vol. 36, no. 2, pp. 385–395, 2016.

[180] Z. Nie and X. Yang, "Deformable Image Registration Using Functions of Bounded Deformation," *IEEE Transactions on Medical Imaging*, vol. 38, no. 6, pp. 1488–1500, 2019.

[181] D. Li, W. Zhong, K. M. Deh, T. D. Nguyen, M. R. Prince, Y. Wang, and P. Spincemaille, "Discontinuity Preserving Liver MR Registration with Three-dimensional Active Contour Motion Segmentation," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 7, pp. 1884–1897, 2018.

[182] J. Zhang, K. Chen, and B. Yu, "An Improved Discontinuity-preserving Image Registration Model and Its Fast Algorithm," *Applied Mathematical Modelling*, vol. 40, no. 23-24, pp. 10740–10759, 2016.

[183] X. Chen, Y. Xia, N. Ravikumar, and A. F. Frangi, "A Deep Discontinuity-Preserving Image Registration Network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 46–55, Springer, 2021.

[184] S. E. Petersen, P. M. Matthews, J. M. Francis, M. D. Robson, F. Zemrak, R. Boubertakh, A. A. Young, S. Hudson, P. Weale, S. Garratt, *et al.*, "UK Biobanks Cardiovascular Magnetic Resonance Protocol," *Journal of Cardiovascular Magnetic Resonance*, vol. 18, no. 1, pp. 1–7, 2015.

[185] V. M. Campello, P. Gkontra, C. Izquierdo, C. Martín-Isla, A. Sojoudi, P. M. Full, K. Maier-Hein, Y. Zhang, Z. He, J. Ma, *et al.*, "Multi-centre, Multi-vendor and Multi-disease Cardiac Segmentation: the M&Ms Challenge," *IEEE Transactions on Medical Imaging*, vol. 40, no. 12, pp. 3543–3554, 2021.

[186] M. Droske and M. Rumpf, "Multiscale Joint Segmentation and Registration of Image Morphology," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2181–2194, 2007.

[187] P. Dong, L. Wang, W. Lin, D. Shen, and G. Wu, "Scalable Joint Segmentation and Registration Framework for Infant Brain Images," *Neurocomputing*, vol. 229, pp. 54–62, 2017.

[188] M. Sinclair, A. Schuh, K. Hahn, K. Petersen, Y. Bai, J. Batten, M. Schaap, and B. Glocker, "Atlas-ISTN: Joint Segmentation, Registration and Atlas Construction with Image-and-spatial Transformer Networks," *Medical Image Analysis*, p. 102383, 2022.

[189] B. Li, W. J. Niessen, S. Klein, M. de Groot, M. A. Ikram, M. W. Vernooij, and E. E. Bron, "A Hybrid Deep Learning Framework for Integrated Segmentation and Registration: Evaluation on Longitudinal White Matter Tract Changes," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 645–653, Springer, 2019.

[190] L. Qiu and H. Ren, "U-RSNet: An Unsupervised Probabilistic Model for Joint Registration and Segmentation," *Neurocomputing*, vol. 450, pp. 264–274, 2021.

[191] Y. Chen, L. Xing, L. Yu, W. Liu, B. Pooya Fahimian, T. Niedermayr, H. P. Bagshaw, M. Buyyounouski, and B. Han, "MR to Ultrasound Image Registration with Segmentation-based Learning for HDR Prostate Brachytherapy," *Medical Physics*, vol. 48, no. 6, pp. 3074–3083, 2021.

[192] B. Li, Z. Sun, Q. Li, Y. Wu, and A. Hu, "Group-wise Deep Object Co-segmentation with Co-attention Recurrent Neural Network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8519–8528, 2019.

[193] S. S. Ahn, K. Ta, S. Thorn, J. Langdon, A. J. Sinusas, and J. S. Duncan, "Multi-frame Attention Network for Left Ventricle Segmentation in 3D Echocardiography," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 348–357, Springer, 2021.

[194] S. Mo, M. Cai, L. Lin, R. Tong, Q. Chen, F. Wang, H. Hu, Y. Iwamoto, X.-H. Han, and Y.-W. Chen, "Mutual Information-Based Graph Co-Attention Networks for Multimodal Prior-Guided Magnetic Resonance Imaging Segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 2512–2526, 2021.

[195] Y.-C. Chen, Y.-Y. Lin, M.-H. Yang, and J.-B. Huang, "Show, Match and Segment: Joint Weakly Supervised Learning of Semantic Matching and Object Co-segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3632–3647, 2020.

[196] H. Chen, Y. Huang, and H. Nakayama, "Semantic Aware Attention Based Deep Object Co-segmentation," in *Asian Conference on Computer Vision*, pp. 435–450, Springer, 2018.

[197] B. Lei, S. Huang, H. Li, R. Li, C. Bian, Y.-H. Chou, J. Qin, P. Zhou, X. Gong, and J.-Z. Cheng, "Self-co-attention Neural Network for Anatomy Segmentation in Whole Breast Ultrasound," *Medical Image Analysis*, vol. 64, p. 101753, 2020.

[198] S. Yang, L. Zhang, J. Qi, H. Lu, S. Wang, and X. Zhang, "Learning Motion-Appearance Co-Attention for Zero-Shot Video Object Segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1564–1573, 2021.

[199] L. Risser, F.-X. Vialard, H. Y. Baluwala, and J. A. Schnabel, "Piecewise-diffeomorphic Image Registration: Application to the Motion Estimation between 3D CT Lung Images with Sliding Conditions," *Medical Image Analysis*, vol. 17, no. 2, pp. 182–193, 2013.

[200] M. von Siebenthal, G. Szekely, U. Gamper, P. Boesiger, A. Lomax, and P. Cattin, "4D MR Imaging of Respiratory Organ Motion and Its Variability," *Physics in Medicine & Biology*, vol. 52, no. 6, pp. 1547–1564, 2007.

[201] C. Jud, R. Sandkühler, N. Möri, and P. C. Cattin, "Directional Averages for Motion Segmentation in Discontinuity Preserving Image Registration," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 249–256, Springer, 2017.

[202] R. Sandkühler, C. Jud, S. Pezold, and P. C. Cattin, "Adaptive Graph Diffusion Regularisation for Discontinuity Preserving Image Registration," in *International Workshop on Biomedical Image Registration*, pp. 24–34, Springer, 2018.

[203] Z. Nie, C. Li, H. Liu, and X. Yang, "Deformable Image Registration Based on Functions of Bounded Generalized Deformation," *International Journal of Computer Vision*, vol. 129, no. 5, pp. 1341–1358, 2021.

[204] R. Temam and G. Strang, "Functions of Bounded Deformation," *Archive for Rational Mechanics and Analysis*, vol. 75, no. 1, pp. 7–21, 1980.

[205] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation," in *International Conference on Medical Image Computing and Computer-assisted Intervention*, pp. 424–432, Springer, 2016.

[206] X. Lu, W. Wang, C. Ma, J. Shen, L. Shao, and F. Porikli, "See More, Know More: Unsupervised Video Object Segmentation with Co-attention Siamese Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3623–3632, 2019.

[207] S. E. Petersen, N. Aung, M. M. Sanghvi, F. Zemrak, K. Fung, J. M. Paiva, J. M. Francis, M. Y. Khanji, E. Lukaschuk, A. M. Lee, *et al.*, "Reference Ranges for Cardiac Structure and Function Using Cardiovascular Magnetic Resonance (CMR) in Caucasians from the UK Biobank Population Cohort," *Journal of Cardiovascular Magnetic Resonance*, vol. 19, no. 1, pp. 1–19, 2017.

[208] C. Petitjean, M. A. Zuluaga, W. Bai, J.-N. Dacher, D. Grosgeorge, J. Caudron, S. Ruan, I. B. Ayed, M. J. Cardoso, H.-C. Chen, *et al.*, "Right Ventricle Segmentation from Cardiac MRI: a Collation Study," *Medical Image Analysis*, vol. 19, no. 1, pp. 187–202, 2015.

[209] J. Bogaert, S. Dymarkowski, A. M. Taylor, and V. Muthurangu, *Clinical Cardiac MRI*. Springer Science & Business Media, 2012.

[210] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, “Self-attention Generative Adversarial Networks,” in *International Conference on Machine Learning*, pp. 7354–7363, PMLR, 2019.

[211] A. Suinesiaputra, P. Ablin, X. Alba, M. Alessandrini, J. Allen, W. Bai, S. Cimen, P. Claes, B. R. Cowan, J. D’ hooge, *et al.*, “ Left Ventricle: Myocardial Infarct Classification Challenge,” *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 2, pp. 503–515, 2017.

[212] H. Lehmann, R. Kneser, M. Neizel, J. Peters, O. Ecabert, H. Kühl, M. Kelm, and J. Weese, “Integrating Viability Information into a Cardiac Model for Interventional Guidance,” in *International Conference on Functional Imaging and Modeling of the Heart*, pp. 312–320, Springer, 2009.

[213] W. E. Lorensen and H. E. Cline, “Marching Cubes: A High Resolution 3D Surface Construction Algorithm,” *ACM Siggraph Computer Graphics*, vol. 21, no. 4, pp. 163–169, 1987.

[214] P. Medrano-Gracia, B. R. Cowan, D. A. Bluemke, J. P. Finn, J. A. Lima, A. Suinesiaputra, and A. A. Young, “Large Scale Left Ventricular Shape Atlas Using Automated Model Fitting to Contours,” in *International Conference on Functional Imaging and Modeling of the Heart*, pp. 433–441, Springer, 2013.

[215] C. W. Lim, Y. Su, S. Y. Yeo, G. M. Ng, V. T. Nguyen, L. Zhong, R. San Tan, K. K. Poh, and P. Chai, “Automatic 4D Reconstruction of Patient-specific Cardiac Mesh with 1-to-1 Vertex Correspondence from Segmented Contours Lines,” *PloS one*, vol. 9, no. 4, pp. e93747:1–e93747:14, 2014.

[216] M. Zou, M. Holloway, N. Carr, and T. Ju, “Topology-constrained Surface Reconstruction from Cross-sections,” *ACM Transactions on Graphics*, vol. 34, no. 4, pp. 1–10, 2015.

[217] B. Villard, V. Grau, and E. Zacur, “Surface Mesh Reconstruction from Cardiac MRI Contours,” *Journal of Imaging*, vol. 4, no. 1, pp. 16–36, 2018.

[218] H. Xu, E. Zacur, J. E. Schneider, and V. Grau, “Ventricle Surface Reconstruction from Cardiac MR Slices Using Deep Learning,” in *International Conference on Functional Imaging and Modeling of the Heart*, pp. 342–351, Springer, 2019.

[219] R. Attar, M. Pereañez, C. Bowles, S. K. Piechnik, S. Neubauer, S. E. Petersen, and A. F. Frangi, “3D Cardiac Shape Prediction with Deep Neural Networks: Simultaneous Use of Images and Patient Metadata,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 586–594, Springer, 2019.

[220] X.-Y. Zhou, Z.-Y. Wang, P. Li, J.-Q. Zheng, and G.-Z. Yang, “One-Stage Shape Instantiation from A Single 2D Image to 3D Point Cloud,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 30–38, Springer, 2019.

[221] Z.-Y. Wang, X.-Y. Zhou, P. Li, C. Theodoreli-Riga, and G.-Z. Yang, “Instantiation-Net: 3D Mesh Reconstruction from Single 2D Image for Right Ventricle,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 680–691, Springer, 2020.

[222] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, “PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 652–660, 2017.

[223] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, “PointNet++: Deep Hierarchical Feature Learning on Point Sets in A Metric Space,” in *Advances in Neural Information Processing Systems*, pp. 5099–5108, 2017.

[224] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang, “Pixel2mesh: Generating 3D Mesh Models from Single RGB Images,” in *Proceedings of the European Conference on Computer Vision*, pp. 52–67, 2018.

[225] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, “Geometric Deep Learning: Going beyond Euclidean Data,” *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 18–42, 2017.

[226] M. Defferrard, X. Bresson, and P. Vandergheynst, “Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering,” in *Advances in Neural Information Processing Systems*, pp. 3844–3852, 2016.

[227] T. N. Kipf and M. Welling, “Semi-Supervised Classification with Graph Convolutional Networks,” in *International Conference on Learning Representations*, pp. 1–14, 2017.

[228] C. Wen, Y. Zhang, Z. Li, and Y. Fu, “Pixel2mesh++: Multi-view 3D Mesh Generation via Deformation,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1042–1051, 2019.

[229] J. Pan, X. Han, W. Chen, J. Tang, and K. Jia, “Deep Mesh Reconstruction from Single RGB Images via Topology Modification Networks,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9964–9973, 2019.

[230] L. Ge, Z. Ren, Y. Li, Z. Xue, Y. Wang, J. Cai, and J. Yuan, “3D Hand Shape and Pose Estimation from a Single RGB Image,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10833–10842, 2019.

[231] N. Kolotouros, G. Pavlakos, and K. Daniilidis, “Convolutional Mesh Regression for Single-image Human Shape Reconstruction,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4501–4510, 2019.

[232] T. Hashimoto and M. Saito, “Normal Estimation for Accurate 3D Mesh Reconstruction with Point Cloud Model Incorporating Spatial Structure,” in *CVPR Workshops*, pp. 54–63, 2019.

[233] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, “SMPL: A Skinned Multi-person Linear Model,” *ACM Transactions on Graphics*, vol. 34, no. 6, pp. 1–16, 2015.

[234] H. Jiang, J. Cai, and J. Zheng, “Skeleton-Aware 3D Human Shape Reconstruction From Point Clouds,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5431–5441, 2019.

[235] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based Learning Applied to Document Recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[236] M. H. de Vila, R. Attar, M. Pereañez, and A. F. Frangi, "MULTI-X, A State-of-the-art Cloud-based Ecosystem for Biomedical Research," in *IEEE International Conference on Bioinformatics and Biomedicine*, pp. 1726–1733, IEEE, 2018.

[237] L. Yu, X. Li, C.-W. Fu, D. Cohen-Or, and P.-A. Heng, "PU-Net: Point Cloud Upsampling Network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2790–2799, 2018.

[238] A. Myronenko and X. Song, "Point Set Registration: Coherent Point Drift," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 12, pp. 2262–2275, 2010.

[239] B. Jian and B. C. Vemuri, "Robust Point Set Registration Using Gaussian Mixture Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1633–1645, 2011.

[240] R. Li, X. Li, C.-W. Fu, D. Cohen-Or, and P.-A. Heng, "PU-GAN: A Point Cloud Upsampling Adversarial Network," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 7203–7212, 2019.

[241] A. Khalil, S.-C. Ng, Y. M. Liew, and K. W. Lai, "An Overview on Image Registration Techniques for Cardiac Diagnosis and Treatment," *Cardiology Research and Practice*, vol. 2018, pp. 1–15, 2018.

[242] M. Thanaj, J. Mielke, K. A. McGurk, W. Bai, N. Savioli, A. de Marvao, H. V. Meyer, L. Zeng, F. Sohler, R. T. Lumbers, *et al.*, "Genetic and Environmental Determinants of Diastolic Heart Function," *Nature Cardiovascular Research*, vol. 1, no. 4, pp. 361–371, 2022.

[243] E. Sarmiento, J. Pico, and F. Martinez, "Cardiac Disease Prediction from Spatiotemporal Motion Patterns in Cine-MRI," in *International Symposium on Biomedical Imaging*, pp. 1305–1308, IEEE, 2018.

[244] W. Bai, W. Shi, A. de Marvao, T. J. Dawes, D. P. O'Regan, S. A. Cook, and D. Rueckert, "A Bi-ventricular Cardiac Atlas Built from 1000+ High Resolution MR Images of Healthy Subjects and an Analysis of Shape and Motion," *Medical Image Analysis*, vol. 26, no. 1, pp. 133–145, 2015.

[245] W. Bai, H. Suzuki, J. Huang, C. Francis, S. Wang, G. Tarroni, F. Guitton, N. Aung, K. Fung, S. E. Petersen, *et al.*, "A Population-based Phenome-wide Association Study of Cardiac and Aortic Structure and Function," *Nature Medicine*, vol. 26, no. 10, pp. 1654–1662, 2020.

[246] D. R. Cox, "Regression Models and Life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.

[247] L. Antelmi, N. Ayache, P. Robert, and M. Lorenzi, "Sparse Multi-channel Variational Autoencoder for the Joint Analysis of Heterogeneous Data," in *International Conference on Machine Learning*, pp. 302–311, PMLR, 2019.

[248] L. Ternes, M. Dane, S. Gross, M. Labrie, G. Mills, J. Gray, L. Heiser, and Y. H. Chang, "A Multi-encoder Variational Autoencoder Controls Multiple Transformational Features in Single-cell Image Analysis," *Communications Biology*, vol. 5, no. 1, pp. 1–10, 2022.

[249] A. Diaz-Pinto, N. Ravikumar, R. Attar, A. Suinesiaputra, Y. Zhao, E. Levelt, E. Dallâ€™Armellina, M. Lorenzi, Q. Chen, T. D. Keenan, *et al.*, "Predicting Myocardial Infarction Through Retinal Scans and Minimal Personal Information," *Nature Machine Intelligence*, vol. 4, no. 1, pp. 55–61, 2022.

[250] A. Ranjan, T. Bolkart, S. Sanyal, and M. J. Black, "Generating 3D Faces Using Convolutional Mesh Autoencoders," in *Proceedings of the European Conference on Computer Vision*, pp. 704–720, 2018.

[251] C. Martin-Isla, V. M. Campello, C. Izquierdo, Z. Raisi-Estabragh, B. Baeßler, S. E. Petersen, and K. Lekadir, "Image-based Cardiac Diagnosis with Machine Learning: A Review," *Frontiers in Cardiovascular Medicine*, vol. 7, pp. 1–19, 2020.

[252] J. J. Van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. Aerts, "Computational Radiomics System to Decode the Radiographic Phenotype," *Cancer Research*, vol. 77, no. 21, pp. e104–e107, 2017.

[253] R. Chen, A. Lu, J. Wang, X. Ma, L. Zhao, W. Wu, Z. Du, H. Fei, Q. Lin, Z. Yu, *et al.*, "Using Machine Learning to Predict One-year Cardiovascular Events in

Patients with Severe Dilated Cardiomyopathy," *European Journal of Radiology*, vol. 117, pp. 178–183, 2019.

[254] M. Khened, V. Alex, and G. Krishnamurthi, "Densely Connected Fully Convolutional Network for Short-axis Cardiac Cine MR Image Segmentation and Heart Diagnosis Using Random Forest," in *International Workshop on Statistical Atlases and Computational Models of the Heart*, pp. 140–151, Springer, 2017.

[255] L. E. Juarez-Orozco, R. J. Knol, C. A. Sanchez-Catasus, O. Martinez-Manzanera, F. M. Van der Zant, and J. Knuuti, "Machine Learning in the Integration of Simple Variables for Identifying Patients with Myocardial Ischemia," *Journal of Nuclear Cardiology*, vol. 27, no. 1, pp. 147–155, 2020.

[256] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. G. Van Stiphout, P. Granton, C. M. Zegers, R. Gillies, R. Boellard, A. Dekker, *et al.*, "Radiomics: Extracting More Information from Medical Images Using Advanced Feature Analysis," *European Journal of Cancer*, vol. 48, no. 4, pp. 441–446, 2012.

[257] V. Kumar, Y. Gu, S. Basu, A. Berglund, S. A. Eschrich, M. B. Schabath, K. Forster, H. J. Aerts, A. Dekker, D. Fenstermacher, *et al.*, "Radiomics: the Process and the Challenges," *Magnetic Resonance Imaging*, vol. 30, no. 9, pp. 1234–1248, 2012.

[258] H. J. Aerts, E. R. Velazquez, R. T. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, *et al.*, "Decoding Tumour Phenotype by Noninvasive Imaging Using A Quantitative Radiomics Approach," *Nature Communications*, vol. 5, no. 1, pp. 4006–4014, 2014.

[259] J. Betancur, L.-H. Hu, F. Commandeur, T. Sharir, A. J. Einstein, M. B. Fish, T. D. Ruddy, P. A. Kaufmann, A. J. Sinusas, E. J. Miller, *et al.*, "Deep Learning Analysis of Upright-supine High-efficiency SPECT Myocardial Perfusion Imaging for Prediction of Obstructive Coronary Artery Disease: A Multicenter Study," *Journal of Nuclear Medicine*, vol. 60, no. 5, pp. 664–670, 2019.

[260] J. M. Wolterink, T. Leiner, B. D. de Vos, R. W. van Hamersvelt, M. A. Viergever, and I. Išgum, "Automatic Coronary Artery Calcium Scoring in Cardiac

CT Angiography Using Paired Convolutional Neural Networks," *Medical Image Analysis*, vol. 34, pp. 123–136, 2016.

[261] A. Lu, E. Dehghan, G. Veni, M. Moradi, and T. Syeda-Mahmood, "Detecting Anomalies from Echocardiography Using Multi-view Regression of Clinical Measurements," in *International Symposium on Biomedical Imaging*, pp. 1504–1508, IEEE, 2018.

[262] K. Kusunose, T. Abe, A. Haga, D. Fukuda, H. Yamada, M. Harada, and M. Sata, "A Deep Learning Approach for Assessment of Regional Wall Motion Abnormality from Echocardiographic Images," *Cardiovascular Imaging*, vol. 13, no. 2_Part_1, pp. 374–381, 2020.

[263] J. Mongan, L. Moy, and C. E. Kahn Jr, "Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a Guide for Authors and Reviewers," *Radiology: Artificial Intelligence*, vol. 2, no. 2, p. e200029, 2020.

[264] A. R. Carter, D. Gill, G. D. Smith, A. E. Taylor, N. M. Davies, and L. D. Howe, "Cross-sectional Analysis of Educational Inequalities in Primary Prevention Statin Use in UK Biobank," *Heart*, vol. 108, no. 7, pp. 536–542, 2022.

[265] C. Yang, F. Starnecker, S. Pang, Z. Chen, U. Güldener, L. Li, M. Heinig, and H. Schunkert, "Polygenic Risk for Coronary Artery Disease in the Scottish and English Population," *BMC Cardiovascular Disorders*, vol. 21, no. 1, pp. 1–9, 2021.

[266] Y. Li, M. Sperrin, and T. van Staa, "R Package "QRISK3": an Unofficial Research Purposed Implementation of ClinRisk's QRISK3 Algorithm into R," *F1000Research*, vol. 8, pp. 2139–2164, 2020.

[267] J. Hippisley-Cox, C. Coupland, and P. Brindle, "Development and Validation of QRISK3 Risk Prediction Algorithms to Estimate Future Risk of Cardiovascular Disease: Prospective Cohort Study," *BMJ*, vol. 357, pp. 1–21, 2017.

[268] R. Harkness, A. F. Frangi, K. Zucker, and N. Ravikumar, "Learning Disentangled Representations for Explainable Chest X-ray Classification Using Dirichlet VAEs," in *Medical Imaging 2023: Image Processing*, vol. 12464, pp. 208–220, SPIE, 2023.