UNIVERSITY OF LEEDS

Bayesian Deep Learning for Cardiac Motion Modelling and Analysis

Ning BI

University of Leeds

School of Computing

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

March, 2024

Intellectual Property Statement

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

Acknowledgements

First and foremost, I extend my deepest gratitude to my supervisors, Dr Arezoo Zakeri, Dr Zeike Taylor, Professor Alex Frangi, and Dr Nishant Ravikumar, for their invaluable guidance, patience, and unwavering support throughout this journey. Their expertise and insightful feedback have been fundamental in shaping my research, and their encouragement has been a constant source of motivation.

My heartfelt appreciation goes to my colleagues and friends Pascale Chalmers-ArnoldView, Michael Macraild, Dr Soodeh Kalaie, Dr Shuang Song, Yash Deo, Nina Cheng, Xiaoyang Sun, Dr Behnaz Elhaminia, Rodrigo Bonazzola, Kun Wu for surviving through the hard times and celebrating the good times together. I'll always miss these most precious four years we spent together. Special thanks to Haoran Dou and Dr Yan Xia for the stimulating discussions, and the sleepless nights working on projects together. I'd also like to thank my climbing crew, Ming Feng, Shenghao Qiu, Yuki Gauchan, and Alexander Ryan Parkinson for climbing with me and boosting my mental health. At least either the brain or muscles are getting gains!

A special mention goes to Professor Shenghua Gao from ShanghaiTech University, Dr Kang Chou from The Chinese University of Hong Kong, and Dr Minye Wu from Katholieke Universiteit Leuven, for their collaboration, valuable contributions to my doctoral journey, and guidance in my professional development.

I am also thankful to the School of Computing, for providing the financial support which made this research possible.

On a personal note, I would like to thank my family, Ms Liqin Chen and Mr Ende Bi, for their unconditional love and support, not only during my PhD but throughout my life. No words can suffice to express my gratitude to them. They always strive to support and inspire me, encouraging their daughter to explore the world, to be creative, and to be brave. I look forward to making you proud.

Finally, I am forever indebted to all those who, directly or indirectly, contributed to the successful completion of this thesis. Thank you all for your support and belief in my work.

Abstract

Cardiovascular diseases (CVDs) remain a primary cause of mortality globally, with an estimated 17.9 million deaths in 2019, accounting for 32% of all global fatalities. In recent decades, non-invasive imaging, particularly Magnetic Resonance Imaging (MRI), has become pivotal in diagnosing CVDs, offering high-resolution, multidimensional, and sequential cardiac data. However, the interpretation of cardiac MRI data is challenging, due to the complexities of cardiac motion and anatomical variations. Traditional manual methods are time-consuming and subject to variability. Deep learning (DL) methods, notably generative models, have recently advanced medical image analysis, offering state-of-the-art solutions for segmentation, registration, and motion modelling.

This thesis encapsulates the development and validation of deep-learning frameworks in the field of cardiac motion modelling and analysis from sequential cardiac MRI scans. At its core, it introduces a probabilistic generative model for cardiac motion modelling, underpinned by temporal coherence, capable of synthesising new CMR sequences. Three models are derived from this foundational probabilistic model, each contributing to different aspects.

Firstly, through the innovative application of gradient surgery techniques, we address the dual objectives of attaining high registration accuracy and ensuring the diffeomorphic characteristics of the predicted motion fields. Subsequently, we introduce the joint operation of ventricular segmentation and motion modelling. The proposed method combines anatomical precision with the dynamic temporal flow to enhance both the accuracy of motion modelling and the stability of sequential segmentation. Furthermore, we introduce a conditional motion transfer framework that leverages variational models for the generation of cardiac motion, enabling anomaly detection and the augmentation of data, particularly for pathologies that are less commonly represented in datasets. This capability to transfer and transform cardiac motion across healthy and pathological domains is set to revolutionize how clinicians and researchers understand and interpret cardiac function and anomalies.

Collectively, these advancements present novelty and application poten-

tials in cardiac image processing. The methodologies proposed herein have the potential to transform routine clinical diagnostics and interventions, allowing for more nuanced and detailed cardiac assessments. The probabilistic nature of these models promises to deliver not only more detailed insights into cardiac health but also to foster the development of personalised medicine approaches in cardiology.

Contents

List of Figures

List of Tables

Abbreviations

CVD	Cardiovascular disease
MRI	Magnetic resonance imaging
CMR	Cardiac cine-MRI
DL	Deep Learning
ROI	Region of interest
RV	Right ventricle
LV	Left ventricle
LVendo	Left ventricular endocardium
LVmyo	Left ventricular myocardium
LVepi	Left ventricular epicardium
LVEDV	Left ventricular end-diastolic volume
LVESV	Left ventricular end-systolic volume
RVEDV	Right ventricular end-diastolic volume
RVESV	Right ventricular end-systolic volume
ED	End diastolic
ES	End systolic
CAD	Coronary artery disease
ECG	Electrocardiography
CT	Computed tomography
MI	myocardial infarction
DCM	Dilated cardiomyopathy
HCM	Hypertrophic cardiomyopathy
ARV	Abnormal right ventricle
HF	Heart failure
CFI	Cardiac functional index
CO	Cardiac output

SV	Stroke volume
EF	Ejection fraction
GWM	Global wall motion
VM	Ventricular mass
EDV	End-diastolic volume
ESV	End-systolic volume
ESS	End-systolic strain
MS	Myocardium strain
SAX	Short-axis
LAX	Long-axis
GR	Gyromagnetic ratio
RF	Radio-frequency
NMR	Nuclear magnetic resonance
GE	Gradient-echo
SE	Spin-echo
CSF	Cerebrospinal fluid
b-SSFP	Balanced steady-state free precession
FID	Free induction decay
t-MRI	Tagging MRI
ANNs	Artificial Neural Networks
DNN	Deep Neural Network
CNN	Convolutional Neural Network
FCN	Fully Convolutional Network
RNN	Recurrent Neural Network
GMM	Gaussian Mixture Models
HMM	Hidden Markov Model
BN	Batch Normalisation
ReLU	Rectified Linear Unit
Tanh	Hyperbolic transient
ELU	Exponential Linear Unit
PReLU	Parametric ReLU
LSTM	Long Short-Term Memory
FC-LSTM	Fully connected LSTM

ConvLSTM	Convolutional LSTM
GRU	Gated Recurrent Units
TCN	Temporal Convolutional Network
NLP	Natural Langrage Processing
PCA	Principal Component Analysis
BoF	Bag of Features
AE	Auto Encoder
GAN	Generative Adversarial Network
VAE	Variational Auto Encoder
HVAE	Hierarchical VAE
CVAE	Conditional VAE
VMD	Variational Diffusion Model
ELBO	Evidence Lower Bound
KL-Divergence	Kullback-Leibler Divergence
DSC	Dice coefficient
HD	Hausdorff Distance
LCC-Demons	Local Correlation Coefficient Demons
LDDMM	Large Deformation Diffeomorphic Metric Mapping
SyN	Symmetric Normalisation
DVF	Displacement vector fields
DIR	Deformable image registration
STN	Spatial Transformer Network
NCC	Normalised Cross Correlation
DLIR	Deep Learning Image Registration
VRNN	Variational RNN
UKBB	UK Biobank
ACDC	Automated Cardiac Diagnosis Challenge
M&M	Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Segmentation Challenge

Chapter 1

Introduction

1.1 Background and motivation

Cardiovascular diseases (CVDs) continue to be a leading cause of mortality worldwide. According to the World Health Organisation (WHO)[1] fact sheet on CVDs, an estimated 17.9 million people died from CVDs in 2019, representing 32% of all local deaths [14]. This fact highlights the critical importance of accurate and comprehensive cardiac imaging and analysis techniques.

In the past few decades, with the help of the fast development of medical imaging techniques, non-invasive imaging hardware has become the main-stream screening method for CVDs and has gained appealing progress. Among the various imaging techniques Magnetic Resonance Imaging (MRI) has emerged as a powerful and non-invasive imaging modality for the assessment of cardiac function and anatomical structures. With its ability to provide high-resolution, multi-dimensional, and sequential information, cardiac MRI has become an invaluable tool in understanding cardiac physiology and pathology [15].

The interpretation and analysis of cardiac MRI data are primarily focused on two major branches, modelling and functional analysis. The cardiac-related modelling tasks include motion modelling and structure modelling. The motion modelling usually performs on sequential data, capturing the heart motion between pairs [16–19] or through a period of time (i.e. a cardiac cycle) [20]. The structure modelling extracts the spatial information from 3D scans and converts it into a 3D atlas. The functional analysis covers a wide range of more specified tasks including but not limited to medical indices estimation, regional analysis [21], and anomaly detection [22]. Despite the variations in the objectives of these research tasks, image registration and segmentation play fundamental roles in the downstream tasks. Image registration refers to finding the optimal transformation to align the moving image to the fixed image. And image segmentation assigns categorical classes at the pixel level. In medical images, the categorical classes usually represent the different anatomical regions of interest (ROI), i.e. left ventricular endocardium (LVendo), myocardium (LVmyo), and right ventricle (RV). Various specific tasks are performed on top of these two fundamental tasks.

Despite its potential, the interpretation and analysis of cardiac MRI data remain challenging due to the complexities associated with cardiac motion and anatomical variations. Traditional methods often rely on manual segmentation and analysis, which are

[1] World Health Organisation website: https://www.who.int/

time-consuming, subjective, and prone to intra- and inter-observer variability. In recent years, deep learning has revolutionized various fields of medical image analysis, providing state-of-the-art solutions for segmentation [23–27], registration [18, 28, 29], and motion modelling [20] tasks. Among the various deep learning techniques, generative deep learning has gained significant attention for its ability to provide uncertainty estimates, making it particularly suitable for medical applications where robustness and interpretability are crucial.

The overarching goal of this thesis is to propose and develop novel DL-based frameworks towards more accurate and robust cardiac motion modelling and regional analysis of sequential MRI data. The author aims to leverage the representational power of deep learning along with the uncertainty quantification of Bayesian methods to enhance the reliability and accuracy of cardiac image analysis.

1.2 Contributions

The successful implementation of Bayesian deep learning for cardiac motion modelling and analysis on sequential MRI in this thesis offers several key contributions in the realm of medical imaging and beyond.

- **Exploration of probabilistic modelling of cardiac motion:** This thesis explores a probabilistic way of cardiac motion modelling, enhancing the understanding and analysis of cardiac motion. This method not only provides a more accurate depiction of cardiac dynamics but also contributes to the broader field of medical image analysis by offering a robust framework that could potentially be adapted for other anatomical regions.

- **Innovations in cardiac motion modelling with gradient-surgery:** The development of an improved cardiac motion modelling method, using a layer-wise gradient-surgery backpropagation scheduler, marks a substantial advancement. This technique not only refines the accuracy of cardiac motion modelling but also preserves the diffeomorphism characteristic of the predicted deformation field, contributing to hyperparameter tuning-free deformable image registration.

- **Achieving full sequence segmentation with spatio-temporal information:** Utilising the spatio-temporal feature from motion modelling to aid the full

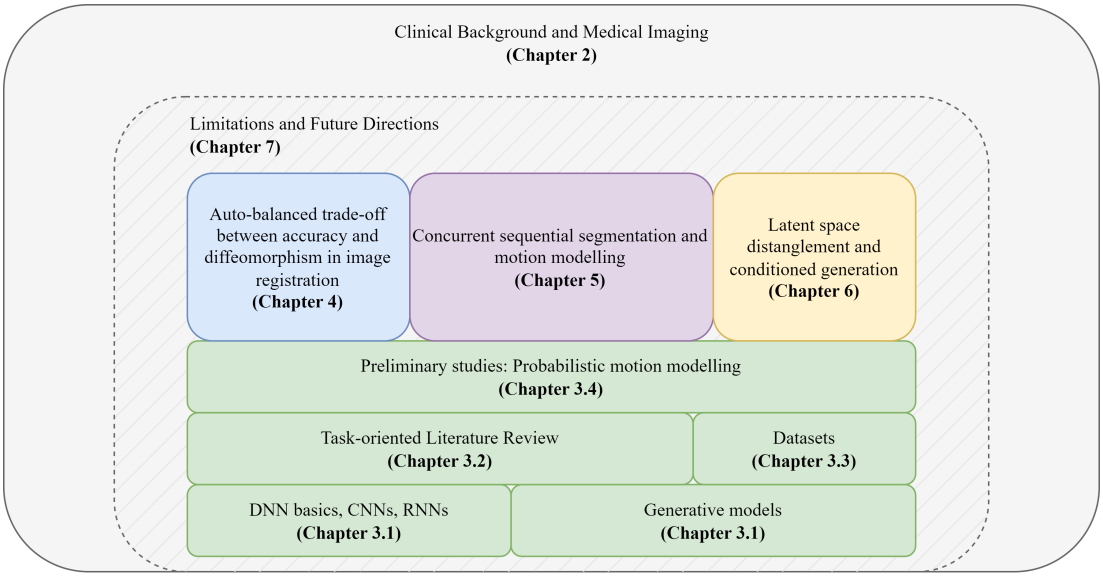


Figure 1.1: **Diagram of thesis structure.**

sequence segmentation addresses the challenge of limited annotation in cine-CMR sequence segmentation. This approach enhances the depth and accuracy of segmentation, which is crucial for detailed cardiac studies and can substantially aid in clinical decision-making.

- **Conditioned motion generation through motion transfer:** This research has pioneered a method to derive more normal samples from typical appearances via motion transfer. This contribution is particularly valuable for creating comprehensive datasets, especially when dealing with conditions where normal samples are limited.

Overall, this thesis offers substantial improvements in the efficiency, accuracy, and reliability of cardiac motion modelling and regional analysis, contributing to diagnosis and treatment planning, which is crucial for patient care. Additionally, the exploration of generative properties in probabilistic generative models for uncertainty assessment not only aids in generating additional samples for in-silico trials but also contributes to the reliability and robustness of medical imaging analysis.

1.3 Thesis structure

Chapter 2 provides the clinical background of this study, offering a concise introduction to relevant medical imaging concepts. The clinical background covers cardiac anatomy and motion, cardiovascular disease, and the cardiac functional indices used for assessing cardiac conditions. The medical imaging background focuses on cardiac magnetic resonance imaging including its theory, physics, and various sequencing techniques in image acquisition.

Chapter 3 presents a literature review from technical and methodological perspectives. **Section 3.1** provides a brief introduction to Deep Neural Networks (DNNs) and generative models, elucidating their foundational components and theoretical bases, with a special emphasis on Recurrent Neural Networks (RNNs) and Variational Auto-Encoders (VAEs). **Section 3.2** reviews cardiac sequential analysis literature, focusing on cardiac ventricular segmentation and motion modelling, primarily exploring machine learning and deep learning approaches. **Section 3.3** introduces three cardiac cine-MRI datasets utilised in this thesis. **Section 3.4** outlines the problem formulation and development of a deep probabilistic model for cardiac motion modelling, laying the foundation for subsequent chapters.

Chapter 4 delves into the trade-off between registration accuracy and the preservation of a deformable deformation field in general image registration tasks. It proposes a deformable registration model, GSMorph, which incorporates layer-wise gradient surgery to simplify hyperparameter tuning. Instead of introducing extra hyperparameters, GSMorph reformulates the optimisation process, projecting the gradient of similarity loss orthogonally to the plane associated with the smoothness constraint.

Chapter 5 extends the foundational spatio-temporal probabilistic generative model to perform concurrent sequential segmentation and motion modelling on SAX cine-CMR. The proposed model, SegMorph, establishes a recurrent latent space for multitask inference and synthesis, capturing spatio-temporal features from cine-MRI sequences. It adopts a learned prior from temporal inputs and demonstrates that motion estimation can enrich sequential segmentation tasks with pseudo-ground truth supervision.

Chapter 6 explores the latent space disentanglement and conditioned generation of the probabilistic motion modelling model via performing motion transfer between healthy patients and patients with cardiac motion disorders. The proposed model,

cMT-VAE, rooted in a conditioned Variational Auto-Encoder, is capable of cardiac motion modelling and cross-domain cardiac motion generation. It shows great potential in position-wise cardiac motion anomaly detection and data augmentation, especially for minority pathology groups. cMT-VAE leverages a conditional variational model to transfer the motions between healthy cardiac cine-MRI (CMR) samples and anomalous ones while preserving respective anatomical structures. This enables a diverse generation of the CMR sequences with different anomalous patterns, which are usually reflected in cardiac contraction motions.

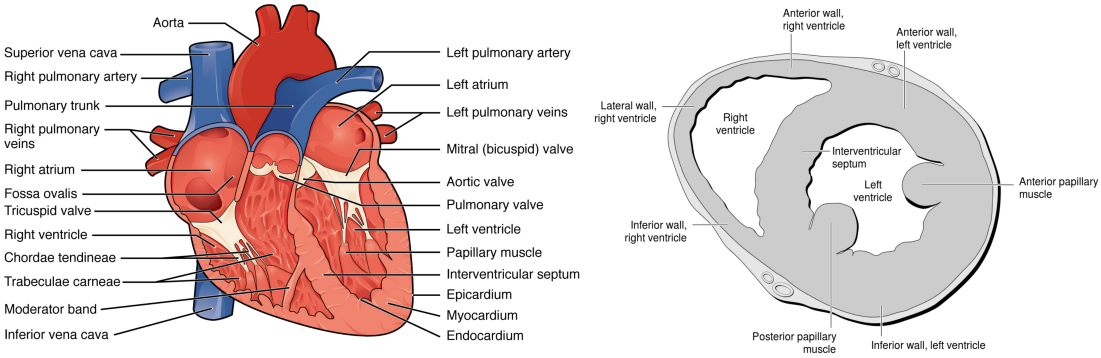
Chapter 7 concludes the thesis, summarising methodologies and findings from the preceding chapters. It discusses the limitations of this research and suggests potential directions for future work.

Chapter 2

Clinical Background and Medical Imaging

2.1 Cardiac anatomy and cardiac cycle

The human heart is a vital organ in the circulatory system and is an intricate structure that orchestrates the rhythmic propulsion of blood throughout the body. As shown in Fig.2.1, the heart consists of four chambers: two atria and two ventricles. The left and right ventricles, positioned at the base of the heart, are pivotal in driving the systemic and pulmonary circulations, respectively. The left ventricle (LV) is responsible for propelling oxygen-rich blood into the systemic circulation through the aorta. In contrast, the right ventricle (RV) serves as the driving force behind pulmonary circulation, propelling deoxygenated blood to the lungs for oxygenation. The LV's muscular wall is notably thicker than that of the RV, reflecting its greater workload in overcoming systemic vascular resistance. Conversely, the RV, adapted for pulmonary circulation, boasts a more delicate structure, ensuring optimal function in pumping blood through the lower-resistance pulmonary vasculature.



(a) (b)

Figure 2.1: **The anatomy of the human heart.** (a) illustrates the internal structures of the heart from the anterior view. It shows the four chambers, the major vessels and the valves. The image is from DeSaix *et al.* [1]. (b) gives a normal short-axis echocardiography view of the heart, showing the LV, RV, and myocardium. Compared with the relatively thinner RV myocardium, the LV myocardium comprises three distinct layers: the epicardium, myocardium, and endocardium, located from the internal surface to the outermost layer of the LV. The image is adapted from Patrick J. Lynch [2].

The cardiac cycle, as shown in Fig.2.2, is a rhythmic sequence of events of the hu-

man heart from the end of one heartbeat to the beginning of the next. It consists of two periods: one during which the heart muscles relax and refill with blood, called diastole, following a period of robust contraction and pumping of blood, dubbed systole. Each cardiac cycle, or heartbeat, takes about 0.8 seconds to complete the cycle. The cycle commences with atrial systole: the atria contract to push the remaining blood into the ventricles. The ventricles which are in diastole, relax to receive the incoming blood. As the atria relax, the ventricles initiate contraction, leading to an increase in intraventricular pressure. The atrioventricular valves close, marking the start of isovolumic ventricular contraction. Once ventricular pressure surpasses the pressure in the pulmonary artery and aorta, the semilunar valves open. This allows blood to be ejected from the ventricles into the pulmonary artery and aorta (ejection phase), driving circulation throughout the body. The ventricles then enter a relaxation phase, during which the semilunar valves close. This marks the onset of isovolumic ventricular relaxation, characterised by a drop in ventricular pressure without any change in volume. With the ventricular pressure lower than that of the atria, the atrioventricular valves open, enabling passive filling of the ventricles. Blood flows into the ventricles from the atria, aided by the atrial kick from the preceding atrial contraction.

Throughout the cardiac cycle, the end-diastole (ED) and end-systole (ES) are two critical phases as the ED phase corresponds to the ventricle's maximal filling, reflecting its capacity to accommodate blood during relaxation, and the ES phase marks the minimal volume remaining after contraction, indicating the amount of blood ejected into circulation. These indices hold immense diagnostic significance, aiding in the assessment of cardiac performance, identification of abnormalities, and formulation of appropriate treatment strategies.

2.2 Cardiovascular disease (CVD)

Cardiovascular disease, is an umbrella term for conditions that affect the heart or circulation. It can also be associated with damage to arteries in organs such as the brain, heart, kidneys, and eyes. Some of the most common forms of CVD include coronary artery disease (CAD), heart failure, hypertension, stroke, and arrhythmias.

CVD are considered to be the leading cause of death worldwide [30, 31]. According to WHO's fact sheet on CVD 2021, an estimated 17.9 million people died from CVD in 2019, representing 32% of all global deaths [14]. Focusing specifically on the

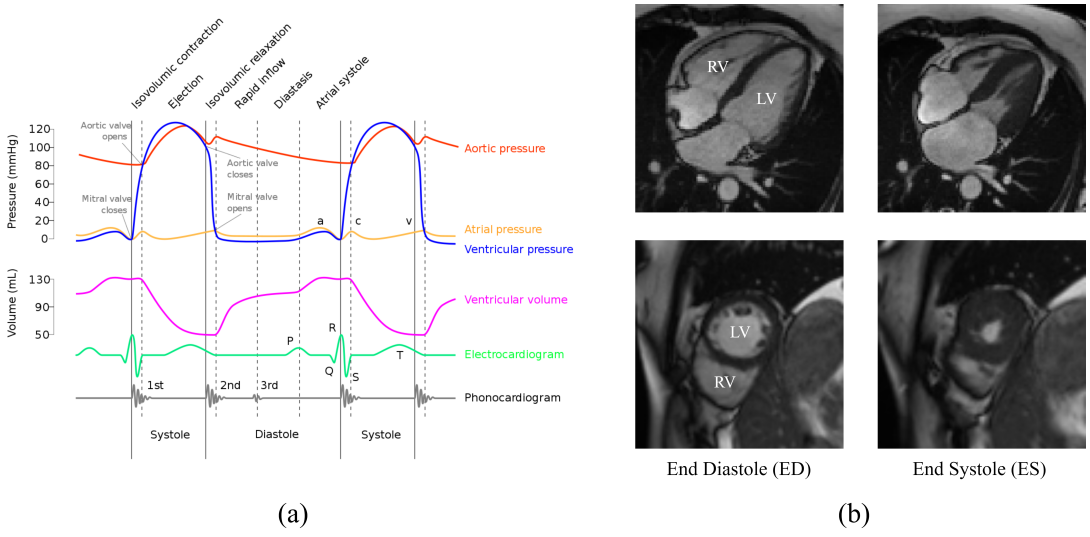


Figure 2.2: **The cardiac cycle.** The Wiggers diagram [3] (a) provides a visual representation of the cardiac cycle. Each event and stage is indicated by the dashed lines from left to right. The horizontally distributed trace lines show the changes in each parameter value (Aortic/atrial/ventricular pressure and ventricular volume) within a cardiac cycle indicated by ECG. (b) shows the 4-chamber long-axis and short-axis view of CMR at ED and ES phase per row from a healthy volunteer.

UK, CVD continued to be a significant health burden as of 2015, accounting for over 28% of all deaths in the country, [32]. Many factors contribute to the development of CVD, including unhealthy diets, physical inactivity, tobacco use, excessive alcohol consumption, obesity, and genetic predisposition.

CVD is diagnosed through an array of laboratory tests and medical imaging studies which include but are not limited to medical and family histories, blood tests, electrocardiogram, stress testing, echocardiography, coronary angiography, cardiac catheterization, chest x-ray, computed tomography (CT), and cardiac MRI analysis. Advances in medical imaging techniques, including CMR, CT, and ECGs, have significantly improved the non-invasive diagnosis and assessment of cardiovascular diseases [33]. CMR, in particular, has gained popularity in the diagnosis of cardiomyopathies [34–36], i.e. myocardial infarction (MI), dilated cardiomyopathy (DCM), hypertrophic cardiomyopathy (HCM), and abnormal right ventricle (ARV), and motion-related abnormalit-

ies [4, 37], e.g. ventricular wall motion assessment, heart failure (HF). These imaging modalities allow for the evaluation of cardiac structure, function, and blood flow, aiding in the early detection and monitoring of CVD.

2.3 Cardiac functional index

The Cardiac Functional Index (CFI) is a comprehensive set of metrics used in cardiology to assess the overall performance of the heart. It provides insights into the heart's functional ability and efficiency, aiding in the diagnosis, monitoring, and management of various cardiovascular conditions. CFI combines several global and regional functional indices, each offering unique information about different aspects of cardiac function, allowing clinicians to make informed decisions regarding patient care and treatment strategies.

2.3.1 Global functional indices

GFIs which predominantly focus on ventricular characteristics, offer insights into the pathological aspects of a heart by examining its overall behaviour and features. These indicators serve as reference points for evaluating myocardial contractility and identifying ventricular abnormalities such as hypertrophy. Commonly used GFIs are usually straight measured via medical devices or calculated based on ROIs in medical images [38]. They encompass metrics such as ventricular volume and mass, cardiac output (CO), stroke volume (SV), ejection fraction (EF), global wall motion (GWM) etc.

The ventricular volume of the left and right ventricles at different phases is a common metric used to analyse cardiac motion and compute many other indices like VM, SV and EF. It involves relating the total chamber volume to the area of a pixel and the ratio of total counts within the chamber to the counts observed [39]. Ventricular mass (VM) is obtained via the subsequent multiplication of the myocardial volume with the tissue (myocardium) density [40]. The two primary parameters calculated are the End-Diastolic Volume (EDV), representing the maximum ventricular capacity during relaxation, and the End-Systolic Volume (ESV), signifying the minimum volume at the peak of contraction. Stroke volume defines the difference between EDV and ESV, measuring the amount of blood pumped through the heart during each stroke. Another critical parameter used to evaluate the systolic function and assess the overall cardiac

pump performance is ejection fraction. EF counts the percentage of blood ejected from the left ventricle with each heartbeat. The EF is defined as the ratio between SV and EDV:

$$EF = \frac{SV}{EDV} \times 100\% = \frac{EDV - ESV}{EDV} \times 100\% \tag{2.1}$$

A healthy heart typically has an EF value between 50% to 70%, indicating efficient pumping of blood. These GFIs are also indicators for early screening of CVD. For example, an increase in VM shows a potential pathological thickening of the heart muscle (hypertrophy), which can signify various cardiac disorders, including hypertension, aortic stenosis, and hypertrophic cardiomyopathy.

2.3.2 Regional functional indices

Beyond the global insights provided by the GFIs, extensive research has shown that regional myocardial functions enable early identification of cardiac dysfunctions, especially for the cardiac disorders that affect localised regions, e.g. myocardium ischemic. Regional functional indices focus on specific areas or segments of the heart to evaluate their individual performance. Most of the regional assessments, including wall thickness and motion, are carried out based on a well-recognised left ventricle myocardium division model proposed by Cerqueira *et al.* [41] shown in Fig. 2.3 (a), which divides the LVmyo into 17 segments based on apex, apical, mid and basal slices. These regional indices are especially useful in identifying regional abnormalities and diagnosing conditions affecting specific heart regions. For example, in Fig. 2.3 (b), the patient with MI shows an observable diffused strain reduction on cine-MRI and the circumferential end-systolic strain (ESS) assessment illustrates the regional abnormality of the LVmyo.

Wall thickness (WT) and wall thickening (WN)

Wall thickness (WT) refers to the distance between the endocardial and epicardial surfaces of the heart's muscular walls. It is commonly measured perpendicular to the inner and outer surfaces of the myocardium. The myocardial wall contracts and thickens during systole and relaxes during diastole. On the other hand, wall thickening (WN), is used to describe the biggest change in myocardial thickness during the cardiac cycle,

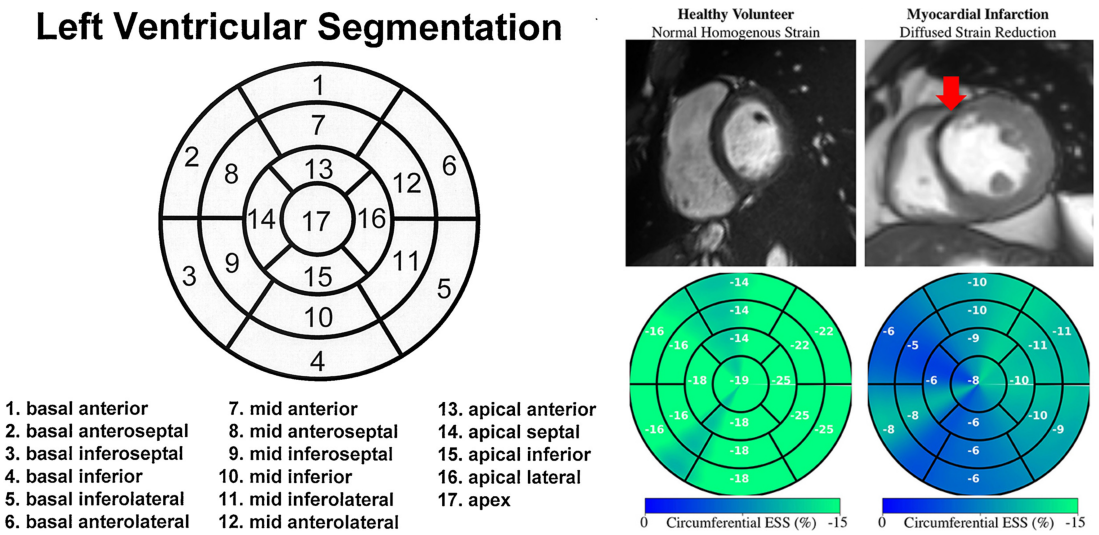


Figure 2.3: **Standardised myocardial segmentation and nomenclature.** (a) shows the Bullseys diagram of the 17 segments of the left ventricular myocardium. It is constructed by segmenting the LVmyo at apex, apical, mid and basal slices, then dividing each segmentation of LVmyo into 4 to 8 sub-segments in anterior, inferior, septal and lateral directions. (b) shows the anatomical (top) and regional (bottom) circumferential end-systolic strain (ESS) for healthy and MI subjects from Morales *et al.* [4]

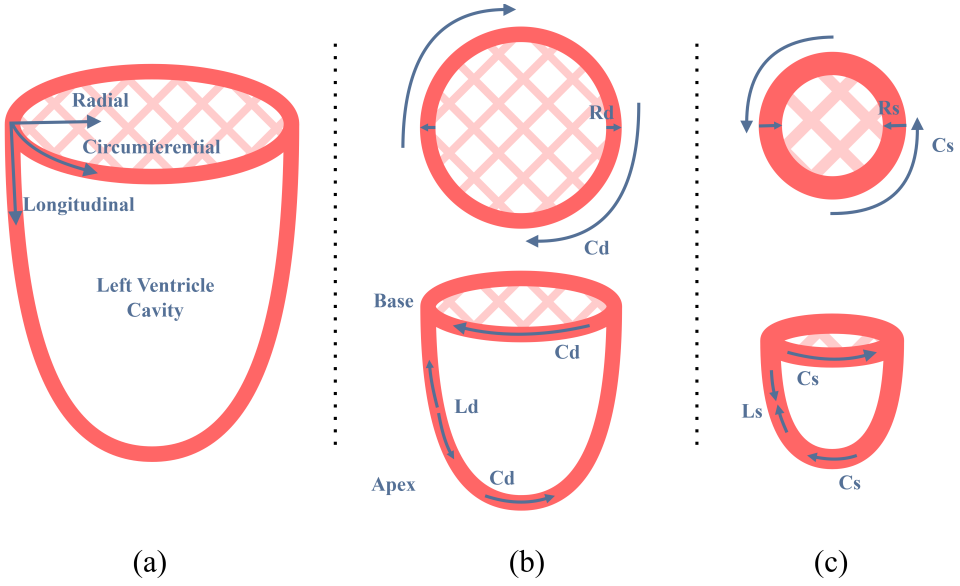


Figure 2.4: **Schematic diagram of myocardium strain (MS).** (a) illustrates three types of myocardium strains caused by different movements of myocardial fibres: the longitudinal, circumferential and radial strains. (b) and (c) shows the deviation of three strains at ED and ES from short-axis and long-axis views. The figure is adapted from Zhang *et al.* [5].

i.e. a percentage change of WT from ED to ES:

$$\mathrm{WN} = \frac{\eta_{\mathrm{ES}} - \eta_{\mathrm{ED}}}{\eta_{\mathrm{ED}}} \times 100\% \tag{2.2}$$

where η represents the average WT measured at specific phases.

Variations in WT can indicate changes in myocardial viability. Reduced thickness may point to areas of scar tissue or fibrosis resulting from conditions like MI. Altered WT is a hallmark of various cardiomyopathies, including HCM with increased and DCM with reduced left ventricular wall thickness.

Strain analysis

Strain imaging provides dynamic information on myocardial deformation. It is a measure of the deformation of cardiac tissues during systole and diastole, capturing nuanced changes that static regional measurements might overlook [42]. Myocardial strain measures the degree of deformation of a myocardial segment from its initial length, L_0, at

ED to its maximum length, L, at ES [43]. The strain is represented in different forms based on the degree of deformation. The MS is expressed using the Green strain form as the myocardium tissue undergoes a relatively large deformation in the cardiac cycle:

$$\epsilon_G = \frac{L^2 - L_0^2}{2L_0^2} = \frac{1}{2}(\boldsymbol{F}^T\boldsymbol{F} - \boldsymbol{I}) \tag{2.3}$$

where $\boldsymbol{F}$ is the gradient of the deformation, and $\boldsymbol{I}$ represents the identity matrix.

Based on the nature of the myocardial fibre deforming pattern, measurements from three different directions are used to describe the motion of the myocardium, the longitudinal strain, circumferential strain and radial strain. As demonstrated in Fig. 2.4, longitudinal strain represents the longitudinal shortening from the base to the apex, which is usually calculated using a long-axis view CMR and expressed by negative values. The longitudinal strain can be obtained through the Lagrangian strain tensor formula [44] and motion tracking [45]. Circumferential strain derives from LV myocardial fibre shortening along the circular perimeter observed on a short-axis view, and it consequently yields negative values. While the radial strain is the radially directed myocardial deformation towards the centre of the LV cavity and indicates the LV thickening and thinning motion during the cardiac cycle; it is usually represented as positive values [43].

By measuring the strain in specific regions of the heart, clinicians can identify subtle abnormalities that may not be evident through traditional imaging methods [46–49]. Additionally, MS can be used to assess the risk of arrhythmias and major adverse cardiovascular events in patients with structural heart disease [50, 51].

The key to a precise strain measurement is the accurate localisation of the myocardium over time. To this end, tagging MRI is considered the gold standard technique for measuring myocardial strain and validating other strain measurement techniques due to its tagging gradient patterns facilitating the key points localisation. However, it has limitations such as tag fading and low spatial resolution [7]. For image modalities with higher resolution but less direct key points, like cine-MRI, CT and ultrasound images, strain analysis is usually performed in a multi-stage manner, which mainly involves two steps: extracting spatial-temporal consistent features and keeping track of the displacement of features over the cardiac cycle to calculate the strain. With the rapid development of medical imaging, feature tracking analysis can be performed using different automated or semi-automated software packages, such as TomTec, Medis,

and Circle Cardiovascular Imaging. In recent years, DL techniques have been exploited to aid the feature extraction [52] and automate the whole strain analysis pipeline [4].

2.4 Cardiac MRI

MRI is a non-invasive medical imaging technique that provides qualitative and quantitative assessments of the heart. To form an image of a certain area, MRI scanners first form strong magnetic fields around the patient's body. Then the oscillating magnetic field is temporarily applied to a certain area with appropriate resonance frequency. The hydrogen nuclei in body tissues emit signals when absorbing energy which is collected and processed to form an image of the body in terms of the density of those nuclei in a specific region. Comparing with other widely used imaging techniques i.e. CT, x-rays, MRI stands out in many ways which make MRI widely exploited in disease diagnosis and quantification. MRI is capable to derive detailed images of soft-tissues that other imaging techniques cannot achieve. Besides, MRI can form an image from any direction and in any orientation under a long-range of scales. For cardiac imaging, a CMRI volume is usually formed along two axes, short axis (SAX) and long-axis (LAX), which involves several slices/volumes of images along each axis. Thus all the volumes within a cardiac cycle will form a $3D + time$ data series, which offers a more comprehensive structure for cardiac shape and motion analysis.

This section of the thesis focuses on two essential techniques within cardiac MRI: Cine-MRI and Tagging-MRI. These techniques play a crucial role in providing detailed insights into cardiac morphology, contractility, and myocardial mechanics, enabling better diagnosis and management of various cardiovascular diseases.

2.4.1 Construction and physics

The human body is made up of around 65% water. MRI employs magnetic fields and radio waves to gauge the water content in various tissues of the body. It maps the distribution of this water and utilizes this data to generate detailed images.

The water molecule (H_2O) is made up of two hydrogen atoms and one oxygen atom. The hydrogen nuclei, or protons (^1H), when placed in a strong external magnetic field B_0, align their intrinsic angular momentum (spins) with the direction of the magnetic field. The magnetic dipole can be described with a magnetisation vector M_0 precessing

around the direction of the external magnetic field B_0 with precession frequency equal to the Larmor frequency, ω_0. The magnitude of M_0 is proportional to the external magnetic field strength $M_0 = \gamma B_0$, where γ is a constant called gyromagnetic ratio (GR). The GR of hydrogen nuclei is approximated as 42.58 MHz/T.

When a radio-frequency (RF) pulse is applied perpendicular to the magnetic field, it can flip the spins of these nuclei away from their aligned positions. This is also known as Nuclear Magnetic Resonance (NMR) phenomenon. After the RF pulse is turned off, the aligned nuclei gradually return to their original alignment with the magnetic field. The process is also referred to as the relaxation process. MRI uses gradients in the magnetic field to spatially encode the NMR signals. During the relaxation process, the radiofrequency coils receive the signals emitted as the aligned nuclei relax back to their equilibrium states. The signal contains information about the tissue properties, and the spatial information is encoded in the signal's frequency and phase. The detected signals are converted from the frequency and phase domain into a spatial domain via mathematical techniques and image reconstruction algorithms [53].

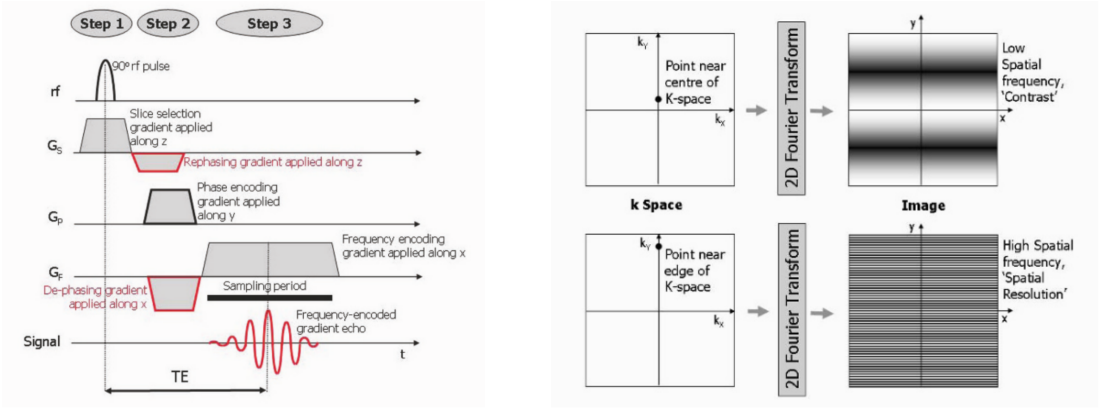
2.4.2 Spatial encoding and k-space

Spatial encoding in MRI involves determining the location of nuclear spins within the body to construct a detailed anatomical image. This is achieved by manipulating magnetic gradients during the imaging process. The primary magnetic field, B_0, establishes a uniform magnetic environment, but spatial information is encoded by applying additional gradient fields (G_x, G_y, and G_z) that vary linearly across the imaging volume. These gradients cause a variation in the resonant frequency of nuclear spins along each axis, thereby allowing for spatial differentiation. The effect of a linear magnetic gradient in one dimension can be expressed as:

$$\omega_x = \omega_0 + \gamma G_x x \tag{2.4}$$

where ω_d is the RF at position x; ω_0 is the RF at the isocentre; γ is constant GR of the nuclear species; G_x is the gradient strength in the x-direction; x represents the spatial position along the gradient direction. The equation applies to y and z gradient directions as well.

In MRI acquisition, the digitised MR signals are stored in k-space, a frequency domain. K-space is filled with raw data one line per time repetition during a conventional



(a) (b)

Figure 2.5: **MRI formation and k-space.** (a) The pulse sequence diagram shows the relative time of the RF and gradient pulses from z, y, and x directions in a single MRI acquisition process. Note that the frequency-encoded gradient echo received during the sampling period will fill one line in k-space.(b) demonstrates the relationship between k-space (frequency domain) and image-space (spatial domain). Each point in the k-space is represented as a wave in the image space. Moreover, the points located near the centre of the k-space represent low-frequency signals with a longer wavelength and ones near the edge of the k-space represent high-frequency signal, e.g. the edge or boundary area of the object. The figures are adapted from [6].

gradient-echo (GE) or spin-echo (SE) acquisition. Each repetition of the sequence acquires a full row of data in the frequency-encoded direction. While the phase-encode gradient is altered for each repetition so that each line in k-space has a different position in the phase-encode direction. The central region of k-space corresponds to low spatial frequencies and contributes to the overall image contrast, while the outer regions contribute to finer details and edges. The k-space data is acquired in the form of raw data points that are then transformed into an image using the Fourier transformation.

2.4.3 MRI sequence

MRI sequences are specific combinations of RF pulses and gradients applied during the imaging process, each designed to highlight different tissue properties and characteristics. These sequences help to generate images with varying contrasts, allowing healthcare professionals to visualise different types of tissue and structures within the body.

T1-weighted sequence

T1-weighted images show the measuring of the spin-lattice relaxation (T1 relaxation). T1 relaxation is the process by which excited protons return to their equilibrium states aligned with the main magnetic field. Tissues with longer T1 relaxation times recover slower to their equilibrium states, resulting in higher signal intensities in T1-weighted images. In T1-weighted imaging, tissues with short T1 relaxation times, such as fat, appear bright, while tissues with longer T1 relaxation times, such as cerebrospinal fluid (CSF), appear darker. This contrast is particularly useful for highlighting anatomical structures and detecting abnormalities like tumours or inflammation.

T2-weighted sequence

T2-weighted images emphasise the differences in spin-spin relaxation (T2 relaxation) times between tissues. T2 relaxation is the process by which excited protons lose their phase coherence due to interactions with neighbouring protons. Tissues with shorter T2 relaxation times will lose their coherence more rapidly, leading to lower signal intensities in T2-weighted images. In T2-weighted imaging, tissues with longer T2 relaxation times, such as CSF, appear bright, while tissues with shorter T2 relaxation times, such

as muscle or bone, appear darker. This contrast is valuable for visualizing fluid-filled structures, oedema, and certain pathologies like multiple sclerosis lesions.

Balanced Steady-State Free Precession (b-SSFP) sequence

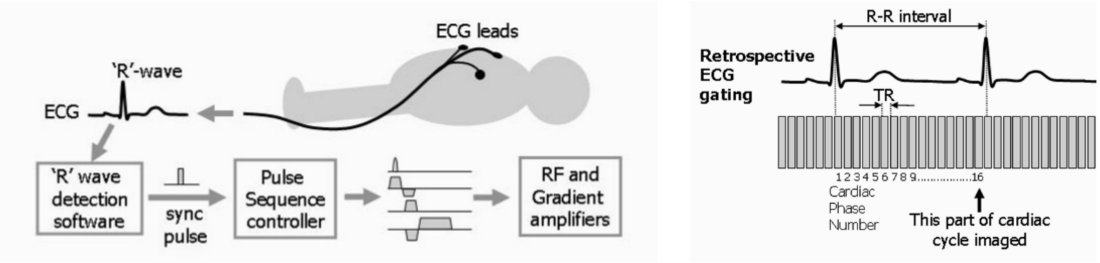
Each type of GE sequence consists of a train of excitation pulses separated by a constant time interval, i.e. TR. After each RF pulse, free induction decay (FID) signals will occur and produce stimulated echoes during dephasing. The transverse magnetisation decays between two RF pulses, leaving a time gap between FID and SEs signals. The SSFP refers to a situation that the time gap is eliminated, forming a continuous signal of varying amplitude, by applying RF-pulses with a sufficient frequency (i.e. TR $\ll$ T1, T2) [54].

Balanced SSFP sequence is one of the rapid GE sequence methods. It shows a strong contrast between tissues with different rations of T2 and T1, for example, between blood and muscle (or myocardium), between fat and muscle, or between liquid compartments and surrounding tissue. B-SSFP is thus perfectly suited for morphological imaging such as cardiac or vessel imaging [55].

In conclusion, by varying the timing parameters and combining different types of MRI pulse sequences, radiologists can create images with specific contrasts tailored to the characteristics of the tissues they want to visualise. This ability to manipulate tissue contrast is a major strength of MRI, allowing for detailed and comprehensive anatomical and pathological assessments without the need for ionising radiation.

2.4.4 cine-MRI

Due to its fast-pace acquisition and ability to show a high myocardial-blood pool contrast, the b-SSFP sequence is also used to acquire cine-MRI [15]. The cine imaging of the heart is achieved by acquiring scans for each slice location at multiple time points throughout the cardiac cycle. A 4D (3D + t) data structure is formed after a repeated acquisition process over slice locations. The number of slices varies from 15 to 50 depending on specific acquisition devices. The overall procedure takes 15-20 breath holds discarding the arithmetic sequences [6]. During the cine-MRI acquisition, it is important to keep the time consistency throughout the repetitions. For this purpose, an ECG signal is used to coordinate the cine image acquisition and detect the arrhythmia. As shown in Fig.2.6, the cardiac cycle interval, i.e. the R-R interval, is determined accord-



(a) (b)

Figure 2.6: **Schematic diagram of cine-MRI acquisition.** Cine imaging is achieved by acquiring MRI data slice-by-slice at multiple time points throughout the cardiac cycle. (a) illustrates the ECG-guided data acquisition pipeline. During the data acquisition, the patient is asked to hold their breath to achieve a repeated and stable ECG measurement. An algorithm is used to detect the steady state and generate the synchronisation pulse, which triggers the cine imaging process shown in (b). (b) shows the retrospective ECG gating cine imaging process, where repeated pulse sequences are applied to capture the specific slice throughout the cardiac cycle and the phase number of each scan is assigned based on its temporal location in the ECG signal. The figure is adapted from [6]

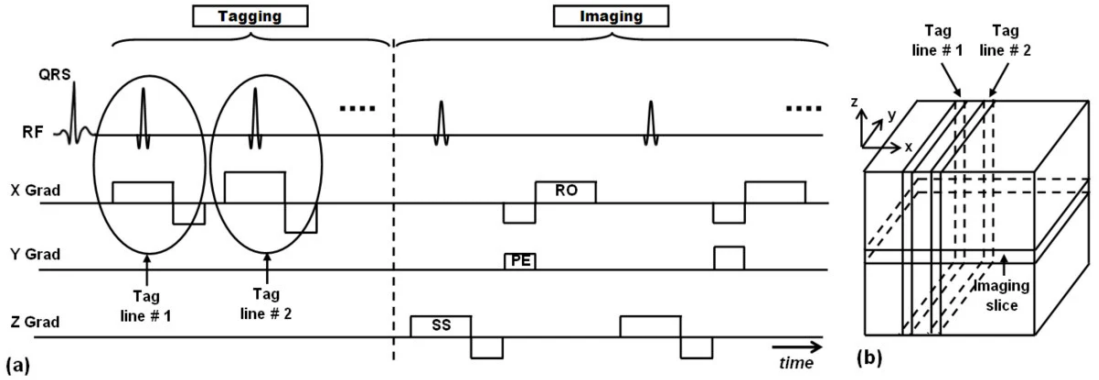


Figure 2.7: **Schematic diagram of cardiac tagging-MRI data acquisition.** (a) shows the two-stage process in t-MRI data acquisition. In tagging preparation, RF pulses are applied perpendicular to the imaging plane to perturb the longitudinal magnetisation at specified locations. During the imaging stage, the tagged areas show darker signal intensity than non-tagged tissues due to the magnetisation saturation they previously experienced. (b) demonstrates the tagging planes and imaging planes which are perpendicular to each other. The image is adapted from Ibrahim *et al.* [7].

ing to ECG signals. To capture the whole cardiac cycle, the illustrated retrospective ECG gating method continuously fills the k-space (repeating the MRI acquisition process illustrated in Fig.2.5) for a sufficient number of time points. The acquired scans are later sorted according to their temporal position relative to the R-wave in the ECG.

Cardiac cine-MRI is a pivotal cornerstone in contemporary cardiology, enabling precise evaluation of cardiac structure and function. Its non-invasive nature, exceptional spatial and temporal resolution, and comparative advantages over other imaging modalities position it as an indispensable tool for clinicians and researchers alike, facilitating improved patient care and advancing our understanding of cardiovascular physiology and pathology. The high frame rate achieved by cine MRI allows for precise assessment of parameters such as ventricular volumes, ejection fraction, and wall motion abnormalities [56, 57]. This temporal accuracy is crucial for diagnosing conditions such as congestive heart failure, MI, and various valvular disorders.

2.4.5 tagging-MRI (t-MRI)

Cardiac tagging-MRI (t-MRI) has been an essential technique for regional myocardial function measurement. Compared with cine-MRI that primarily focuses on the global dynamics of the heart chambers, t-MRI allows the qualification of local intra-myocardial motion measures, like strain, strain rate, and torsion, without the need for physical markers [7]. This capability facilitates the identification of regional abnormalities, ischemic segments, and even early signs of cardiomyopathies that might go undetected in cine imaging.

T-MRI employs a unique tagging pattern to label specific regions of the myocardium, allowing for the precise tracking of tissue motion over time. It is thus considered the reference standard for strain qualification [58, 59]. Myocardial tagging technology has been developing for decades since Zerhouni *et al.* [60] first introduced the non-invasive technique to create tagging markers using CMR. A comprehensive explanation of its theory and applications is summarised by Ibrahim *et al.* [7]. In general, as is shown in Fig.2.7, the t-MRI data acquisition process involves two stages: tagging preparation and imaging. During tagging preparation, RF pulses are applied perpendicular to the imaging plane to perturb the longitudinal magnetisation at specified locations. The rest of the imaging slice is not affected by the tagging pulses. Then during the imaging stage, the tagged areas show darker signal intensity than non-tagged tissues due to magnetisation saturation they previously experienced. The acquired image shows visual evidence of tissue deformation that occurred since the time of tagging pulses application. In its simplest form, the imaging stage consists of a series of slice-selective RF pulses, each followed by phase encoding and readout gradients for k-space filling as described in Section 2.4.2.

T-MRI has several clinical applications in CAD diagnosis and treatment. It can be used to measure regional myocardial function [61–63], assess myocardial viability, detect myocardial ischemia [64, 65], and study the pathophysiology of various CVD [66–68].

Chapter 3

Literature Review and Theory

3.1 Deep neural networks and generative models

Deep Neural Networks (DNNs), inspired by the intricate web of neurons in the human brain, are a subclass of Artificial Neural Networks (ANNs) characterised by their deep architectures with multiple layers. These layers progressively transform input data into a representation that makes the desired task easier [69]. The power of DNNs lies in their ability to learn intricate patterns from large datasets, and they've been applied to a range of tasks in medical imaging. DNNs is a relatively broad category covering a broad and diverse of specific architectures and types of neural networks. Convolutional Neural Networks (CNN) [70], with their hierarchical structure, are adept at capturing intricate patterns and features in images, making them particularly suited for medical images where subtle details are paramount. Recurrent Neural Networks (RNN) [71], on the other hand, is specialised in capturing features from sequential data, which makes it widely applicable in sequential medical data like ECGs and cine-MRI.

Generative models are a class of statistical models that can generate new data samples similar to the input data. Historically, classical generative techniques such as Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM) were prevalent. However, with the booming of deep learning, more sophisticated architectures like Generative Adversarial Networks (GANs) [72–75] and Variational Autoencoders (VAEs) [76] have taken centre stage. A deep generative model is usually built with a composition of different DNNs. In the area of medical image processing, deep generative models are proudly adopted in the realm of data synthesis and anomaly detection due to their capabilities of generating new samples from the learned distribution and resembling real observations.

In the following subsections, we explain more details of DNNs and generative models closely related to the research topic.

3.1.1 Convolutional Neural Network

Various architectures of DNNs have been proposed in the recent decade, along with the immersive studies conducted in Deep learning-related fields. At the heart of different networks lie multiple layers, each designed to capture and transform different features from the input data, ultimately leading to a desired output. This subsection delves into the definitions and underlying mathematics of various layers commonly found in neural networks.

Most of the DNN frameworks can be divided into three processes: encoding inputs, where the inputs are mapped from image/data space to feature space; feature fusion, where features from different sources are fused together in the feature space; decoding features, where the high-level features are mapped to desired space varies according to the objections.

The convolution block is essential in the encoding stage, mapping the inputs to the feature space. Each convolution block consists of a Convolution (Conv) layer, a Batch Normalisation (BN) layer and an activation layer. The Conv layer performs a convolution operation on its input with learnable kernels. Given an input matrix I and a kernel K, the convolution operation at position (i, j) is:

$$(I * K)(i, j) = \sum_m \sum_n I(m, n) \cdot K(i - m, j - n) \tag{3.1}$$

Note that to map an input with C_{in} to C_{out} channels, we need C_{out} number of learnable kernels, each with size of $h \times w \times C_{in}$, which ended up with overall $h \times w \times C_{in} \times C_{out}$ learnable parameters. Moreover, as shown in Fig.3.1, by applying the convolution operation with a stride greater than one, the strided convolution layer can reduce spatial dimensions on feature maps and help in increasing the receptive field, allowing the network to capture broader contextual information from the input [8]. After each

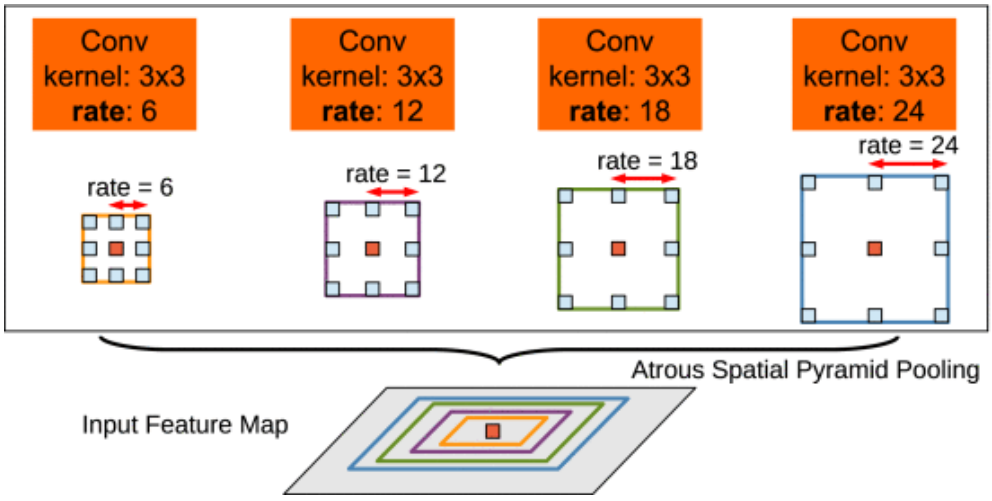


Figure 3.1: **Dilated convolution layers and receptive field.** The figure is adapted from Chen *et al.* [8]

convolution operation, a BN layer is often added to normalise the activations. This helps

stabilise and accelerate the training process by reducing internal covariate shift [77]. For a given mini-batch $\mathcal{B}$ of size m with mean $\mu_{\mathcal{B}}$ and variance $\sigma^2_{\mathcal{B}}$ the normalised output is:

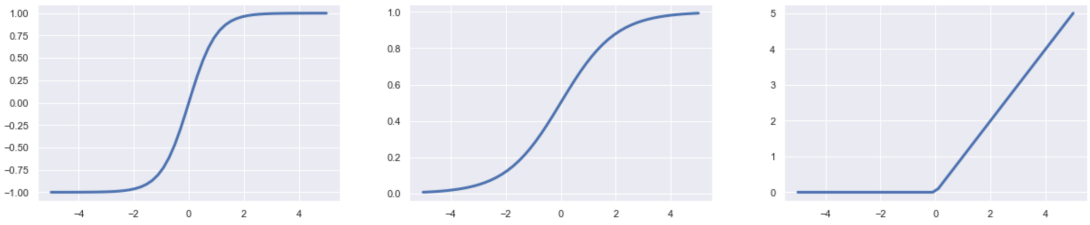
$$\hat{x}_i = \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma^2_{\mathcal{B}+\epsilon}}} \tag{3.2}$$

where ϵ is a small constant for numerical stability.

The activation layer is applied to introduce non-linearity into the network, enabling it to learn complex relationships in the data. The choice of the activation functions varies on the objectives. Historically, the Sigmoid and Hyperbolic transient (Tanh) functions, shown in Eq.3.3, are the activation functions of choice.

$$\begin{aligned} \sigma(x) &= \frac{1}{1+e^{x}} \\ \text{Tanh}(x) &= \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \end{aligned} \tag{3.3}$$

As shown in Fig.3.2, the sigmoid function maps the input value to an output ranging from 0 to 1 while tanh function has an output range spanning negative values, giving an output range $[-1, 1]$. However, with an input value either very large or very small, the sigmoid and tanh function saturates, leading to vanishing gradients, which can slow down learning and more frequently appear in deeper network architecture. The Rectified Linear Unit (ReLU) function is proposed by Nair *et al.* [78] to solve the vanishing gradient problem during training and is later adopted in the famous backbone model AlexNet [79]. More variations of ReLU are proposed to address the "dying



(a) (b) (c)

Figure 3.2: **Activation functions.** (a) Hyperbolic transient (tanh), (b) Sigmoid function, and (c) Rectified linear unit (ReLU) activation function.

ReLU" problem, where neurons can sometimes become inactive and cease to update or adjust during training, especially if a large gradient flows through a ReLU neuron,

updating the weights in such a way that the neuron will always output zero. Seen from Eq.3.4, Leaky ReLU allows a small gradient, determined by α, which prevents the neurons from dying out entirely. Exponential Linear Unit (ELU) mitigates the vanishing gradient problem for negative input values. ELU introduces an exponential curve, which can help the network push mean unit activations closer to zero, making the learning process more robust. Later, a parametric variation PReLU proposed by He *et al.* [80] generalised the Leaky ReLU by introducing a learnable parameter that allows the activation function to adaptively learn the slop for negative values during the training process. PReLU also claims to be the key factor in surpassing human-level performance on ImageNet [81] classification task.

$$\begin{aligned} \mathrm{ReLU}(x) &= \max(0, x) \\ \mathrm{LeakyReLU}(x) &= \max(\alpha x, x) \\ \mathrm{ELU}(x) &= \begin{cases} x & \text{if } x > 0 \\ \alpha(e^x - 1) & \text{if } x \leq 0 \end{cases} \end{aligned} \tag{3.4}$$

By stacking several convolution blocks, dense layers, pooling layers and soft-max layers in specific orders, different backbone networks are formed to extract hierarchical features from the input data, which are then used by the subsequent layers or modules tailored for specific tasks. In medical image processing, CNNs are used to extract hierarchical features from the input images to perform various tasks including lesion detection [22, 82, 83], classification [84–86], segmentation [25, 87–89], etc. A comparative analysis of prevalent backbone networks is presented in Table 3.1.1. These networks are largely employed as pre-trained feature extractors across various applications. The criteria for comparison include input dimensions, count of convolutional blocks, convolutional kernel dimensions, quantity of trainable parameters, and the associated memory footprint of the model.

3.1.2 Recurrent Neural Networks

Recurrent neural network is a class of neural networks that allow the previous outputs to contribute as the input and have hidden states. These features enable RNNs to outperform other DNNs in sequential data processing, like Natural Language Processing (NLP) [93], video captioning [94] and sequential image registration [95]. In medical imaging, recurrent modules are widely adopted in constructing a more complex model

Table 3.1: Comparison of different fundamental CNN architectures.

Backbones	AlexNet [79]	VGG-16 [90]	GoogleNet [91]	ResNet-50(v1) [92]
Input size	227 × 227	224 × 224	224 × 224	224 × 224
Number of Conv blocks	5	16	21	50
Filter Size	3,5,11	3	1,3,5,7	1,3,7
Number of FC layers	3	3	1	1
Total Weights	61M	138M	7M	25.5M
Total MACs	724 M	15.5G	1.43G	3.9G

for handling temporal data. In [86, 96, 97] recurrent modules are adopted in U-Net and its variations to conduct image segment brain, blood vessel, and cardiac ventricles on CT and MRI. Dileep *et al.* [98] combine bi-directional LSTM and CNN to perform heart disease detection on patient datasets.

Figure 3.3 illustrates the computational graph of a vanilla RNN. Unlike traditional

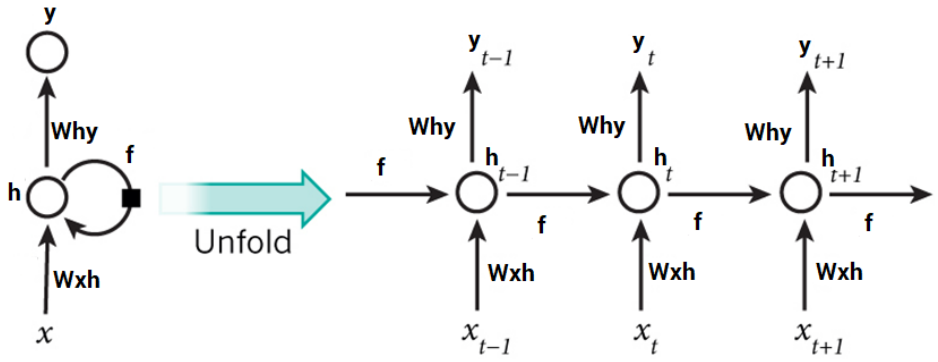


Figure 3.3: **Computational graph of vanilla RNN.** The RNN with recurrent connections on the *left* can be unfolded *w.r.t.* time as the computational graph shown on the *right*, where each node is associated with its particular time instance. The input sequence $\{x_1, x_2, ..., x_T\}$, an RNN updates its hidden state h_t at each time step t using the current input x_t and the previous knowledge stored in h_{t-1}. The output of each time step y_t is derived from its hidden state h_t. The calculation details are explained in Eq.3.5

feed-forward neural networks, RNNs have connections that loop back on themselves, allowing information to persist. This characteristic makes them particularly well-suited

for tasks where temporal dynamics and the context from earlier steps are important. RNNs maintain a form of memory by using their internal state, which captures information about previous time steps. This allows them to remember and use past information to influence the current state output. Moreover, unlike the common DNNs, which require a fixed input size, RNNs can handle inputs and outputs of variable lengths, making them versatile for tasks like language translation, where the input and output sequences might have different lengths. Given an input sequence $x = \{x_1, x_2, ..., x_T\}$, an RNN updates its hidden state h_t at each time step t using the following equations:

$$\begin{aligned} h_t &= \sigma\left(W_{hh}h_{t-1} + W_{xh}x_t + b_h\right) \\ y_t &= W_{hy}h_t + b_y \end{aligned} \tag{3.5}$$

where W_{hh}, W_{hy}, and W_{hx} are weight matrices; b_h and b_y are bias vectors; $\sigma(\cdot)$ is an activation function, often the tanh is adopted.

Despite their potential, vanilla RNNs are often plagued by issues like vanishing and exploding gradient problems when handling long sequences [99]. To address these challenges, more advanced RNN architectures like Long Short-Term Memory (LSTM) [100–102] and Gated Recurrent Units (GRU) [103] have been developed.

Long Short-Term Memory (LSTM)

Graves *et al.* [100] introduce the fully connected Long Short-Term Memory (FC-LSTM) with a set of gating mechanisms that regulate the flow of information into and out of the memory cell. Fig.3.4 shows the schematic diagram of LSTM and Convolutional LSTM (ConvLSTM). The forget gate, f_i decides which information from the cell state should be thrown away or kept. It looks at the previous hidden state, h_{t-1}, and the current input, x_t, and outputs a weight with range $[0, 1]$ for each number in the cell state, indicating the amount of the history state is taken into account in updating the next hidden state, C_t. The input gate, i_i, updates the cell state with new information. It has two parts: A sigmoid layer, called the "input gate layer," that decides which values will be updated. A tanh layer that creates a vector of new candidate values. The output gate, o_i, decides what the next hidden state should be. It takes the current input, x_t, and the previous hidden state, h_{t-1} and outputs a value between 0 and 1 for each number in the cell state. The cell state is passed through a tanh function to push values between -1 and 1 and then multiplied by the output of the sigmoid gate so that only the parts determined by the gate are outputted.

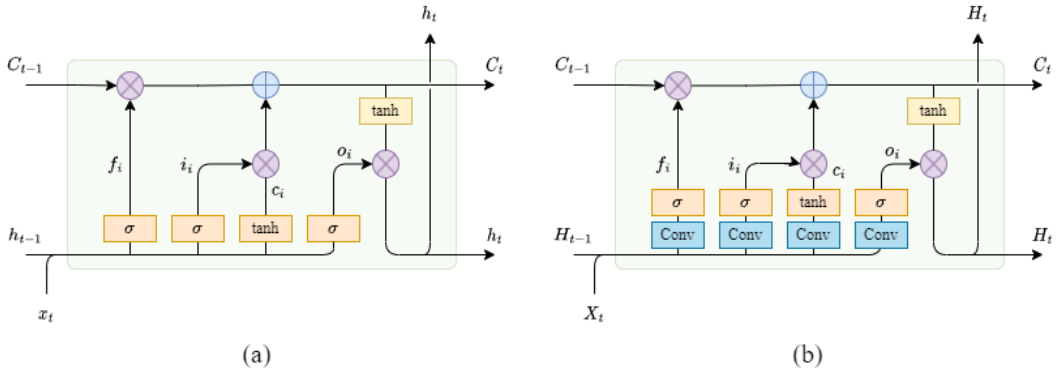


Figure 3.4: **FC-LSTM and ConvLSTM.** (a) An illustration of the Fully-connected Long Short-Term Memory (FC-LSTM) architecture, showcasing its intricate gating mechanisms, input i_i, forget f_i, and output o_i gates, that enable the effective modeling of long-term dependencies in sequential data. (b) Convolutional LSTM (ConvLSTM) replaces the recurrent connections in LSTM with convolutional connections, enabling ConvLSTM to capture the feature in spatial-temporal data.

The FC-LSTM updates its cell state C_t and hidden state h_t at each time step using the following equations:

$$\begin{aligned} f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf} \circ c_{t-1} + b_f) \\ i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci} \circ c_{t-1} + b_f) \\ \tilde{c}_t &= \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\ c_t &= f_t \circ c_{t-1} + i_t \circ \tilde{c}_t \\ o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co} \circ c_t + b_o) \\ h_t &= o_t \circ \tanh(c_t) \end{aligned} \tag{3.6}$$

referring to Fig. 3.4(a), f_t, i_t and o_t are the forget, input, and output gates, respectively. $\tilde{c}_t$ is the candidate cell state. W matrices and b vectors are learnable parameters. $\sigma(\cdot)$ is the activation function, usually sigmoid, and $\circ$ denotes the Hadamard product.

Although the LSTM layer has shown better performance in handling temporal correlation compared to vanilla RNNs, it carries too much redundancy when comes to spatial-temporal inputs, like temporal image sequence, due to its possession of fully connected layers in the base structure. To tackle this problem, Shi *et al.* propose Con-

volutional LSTM [102] to improve the LSTM efficiency and performance on spatial-temporal data.

As shown in Fig.3.4(b), ConvLSTM extends the FC-LSTM to have convolutional structures in both the input-to-state and state-to-state transitions. Note that under ConvLSTM setting, the inputs $\{X_1, ..., X_T\}$, cell outputs $\{C_1, ..., C_T\}$, hidden states $\{H_1, ..., H_T\}$, and all the gate outputs are tensors instead of vectors. Replacing the fc layers with convolutional structures reduces the spatial redundancy and the number of parameters in optimization. The forward path in ConvLSTM is described by the following equations:

$$\begin{aligned} f_t &= \sigma(W_{xf} * X_t + W_{hf} H_{t-1} + W_{cf} \circ C_{t-1} + b_f) \\ i_t &= \sigma(W_{xi} * X_t + W_{hi} H_{t-1} + W_{ci} \circ C_{t-1} + b_f) \\ \tilde{C}_t &= \tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c) \\ C_t &= f_t \circ C_{t-1} + i_t \circ \tilde{C}_t \\ o_t &= \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \circ C_t + b_o) \\ H_t &= o_t \circ \tanh(C_t) \end{aligned} \tag{3.7}$$

where f_t, i_t and o_t denote the forget, input, and output gates, respectively. $*$ denotes the convolution operator and $\circ$ denotes the Hadamard product.

CNNs and RNNs have greatly boosted the potential of DL models in medical image analysis. However, unlike the more readily accessible and feasible natural data, medical data present a unique set of challenges, predominantly stemming from the stringent and confidential processes enveloped in data acquisition. The sensitivity and privacy of health-related data necessitate rigorous ethical and legal considerations, often resulting in a constrained availability of datasets for research purposes. Furthermore, the diversity in modalities and acquisition machines introduces additional complexities, instigating domain gaps in the acquired data due to variations in image quality, resolution, and format across different medical imaging technologies. Moreover, the inherent rarity of lesion samples in medical datasets often poses a formidable challenge, as obtaining a sufficient volume of data to train DL models becomes an intricate task, potentially hindering the development and validation of robust predictive models.

To circumvent these challenges, generative models have emerged as pivotal players in the realm of medical image analysis. These models have been instrumental in data augmentation, where they have been employed to synthetically enlarge datasets by

generating new, plausible samples, thereby ameliorating the issues posed by limited data availability. Some generative models, with their initial property of being probabilistic models, also enable the analysis of uncertainty, providing a nuanced understanding of the data generation process and potential variations therein.

3.1.3 Generative models

Unlike discriminative models, which focus on distinguishing between different classes of data, generative models aim to create new samples consistent with observed data. These models have profound implications, from synthesising realistic images and sounds to facilitating drug discovery and understanding complex systems.

Generative models aim to capture and replicate the underlying data distribution, generating new samples by transforming latent space representations back into the original data space [72]. Transitioning between data and latent spaces involves various compression and mapping algorithms. In machine learning, algorithms like Principal Component Analysis (PCA) [104] and Bag of Features (BoF) [105] are employed. In deep learning, this often involves neural networks that map input from data space to latent space (Encoders) and modules that reconstruct or generate data from latent codes (Decoders). Auto-encoder (AE) are a notable example in this context [106].

Auto-encoder is designed to learn a compact latent representation from data without any supervision from the ground truth. A typical auto-encoder consists of two networks: an encoder network for the representation of the input and a decoder network for the reconstruction of the input. For each input $x^{(i)}$ from a dataset $\{x^{(1)}, ..., x^{(N)}\}$ of N samples, the encoder, paramdterised by θ_e, maps $x^{(i)}$ to a latent code denoted $z^{(i)}$. Conversely, the decoder, parameterised by θ_d, maps $z^{(i)}$ back to the original data space, resulting in the reconstruction, $\hat{x}^{(i)}$. The encoding and decoding processes are mathematically represented as:

$$\begin{aligned} z^{(i)} &= f(x^{(i)}; \theta_e) \\ \hat{x}^{(i)} &= g(z^{(i)}; \theta_d) \end{aligned} \tag{3.8}$$

where $f_{\theta_e}(\cdot)$ denotes the encoder and $f_{\theta_d}(\cdot)$ denotes the decoder. The primary goal in training basic AE is to minimise the reconstruction error by finding the values for the parameters $\{\theta_e, \theta_d\}$ that minimises reconstruction error. Besides, a regularisation term for the latent vector is usually exploited to improve the generalisation and prevent the

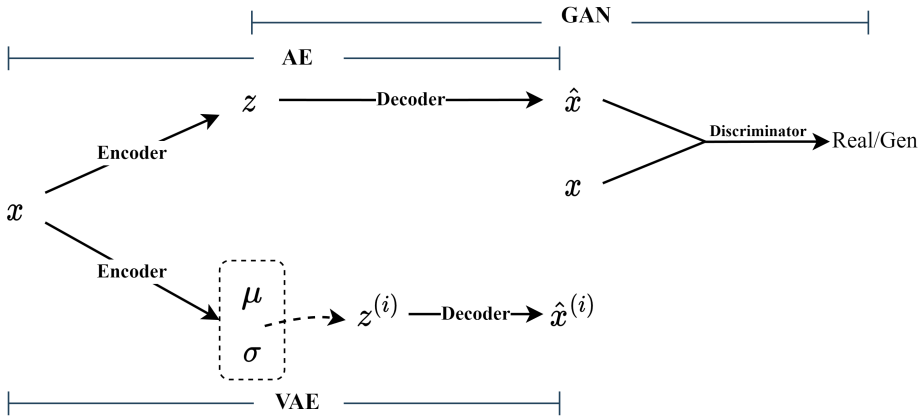


Figure 3.5: **Schematic diagram of Auto Encoder (AE), Generative Adversarial Network (GAN), and Variational Auto Encoder (VAE).** Although sharing a similar Encoder-Decoder network architecture, AE learns to compress data into a lower-dimensional latent space and then reconstruct it back to the original space, focusing on minimising reconstruction errors. VAE, an extension of AE, introduces a probabilistic approach by encoding inputs as distributions in the latent space usually parameterised by μ and σ, enhancing the generation of diverse and novel samples. GAN, on the other hand, consist of two competing networks: a generator that creates samples and a discriminator that evaluates their authenticity, leading to the generation of high-quality, realistic data. While AEs and VAEs are primarily focused on encoding and decoding mechanisms, GANs emphasise the adversarial process for sample generation.

AE from overfitting [107]. The reconstruction and regularisation losses are defined as:

$$\begin{aligned} \mathcal{L}_{\text{recon}} &= \frac{1}{n}\sum_{i=1}^{n} \|x_i - \hat{x}_i\|^2 \\ \mathcal{L}_{\text{reg}} &= -\frac{1}{2}\sum_{j=1}^{m}(1 + \log\left(\sigma_j^2\right) - \mu_j^2 - \sigma_j^2) \end{aligned} \tag{3.9}$$

Where m is the dimension of the latent space, and μ, σ are the mean and standard deviation of the latent distribution. Combines the reconstruction loss and regularisation term, the overall loss is as follows:

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \lambda \mathcal{L}_{\text{reg}} \tag{3.10}$$

where λ is a weighting hyperparameter that balances the two terms.

With its potential to compress and reconstruct the original data, AEs have been widely used in the field of medical imaging. Schlemper et al.[108], Oktay et al.[109], Yue et al.[110], have employed auto-encoders to extract general semantic features or the shape information from the input images or labels and then use these features to guide the cardiac image segmentation.

However, AE focuses on a deterministic encoding of data to a latent representation, which limits its capability to produce new samples from a given training population. The deep generative models [76, 111] unlock the generation power of the Encoder-Decoder structure by introducing the stochasticity and probabilistic modeling of the latent space. This probabilistic approach enables these models to not only learn complex data distributions but also to generate novel, diverse samples that are representative of the underlying dataset.

Generative Adversarial Network (GAN)

The concept of Generative adversarial networks is proposed by Goodfellow et al. [111] for image synthesis from noise. GANs are a type of generative models utilises adversarial training to model the sampling procedure of a complex distribution from real data and thus can create new examples. A GAN consists of a generator network (G) and a discriminator network (D). G takes a random noise vector z as input to synthesis data samples, aiming to produce indistinguishable samples from real data. Simultaneously, the D evaluates both real data samples x and fake (generated) samples $\hat{x}$ produced by the Generator, classifying them as real or fake. Their interaction is defined by a minimax objective function:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))] \tag{3.11}$$

where $p_{\text{data}}(x)$ is the distribution of real data, and $p_z(z)$ is the noise distribution. The adversarial training scheme of GANs, where G and D are iteratively updated through backpropagation, marks its most significant legacy. The training objectives can be mathematically expressed as:

$$\begin{aligned} \mathcal{L}_D^{GAN} &= \max_D \mathbb{E}_{x\sim p_{\text{data}}(x)}\left[\log D(x)\right] + \mathbb{E}_{\hat{x}\sim p_{\text{gen}}(x)}\left[\log(1-D(\hat{x}))\right] \\ \mathcal{L}_G^{GAN} &= \min_G \mathbb{E}_{\hat{x}\sim p_{\text{gen}}(x)}\left[1-\log D(\hat{x})\right] \end{aligned} \tag{3.12}$$

GANs, leveraging unsupervised learning, have been employed in various fields, including image-to-image translation with models like Pix2Pix [112] and CycleGAN [113],

and in medical image processing for tasks such as deep augmentation [114], reconstruction [115–118], segmentation [119, 120], and anomaly detection [121]. However, challenges such as training difficulty, resource consumption, and performance evaluation complicate their use, particularly in medical imaging where interpretability and stability are crucial. Thus, Variational Auto Encoders (VAEs) present a compelling alternative, especially in medical image processing, due to the emphasis on interpretability and structured latent space manipulation in this field of research.

Variational Auto Encoder (VAE)

Variational Auto Encoder, introduced by Kingma and Welling [76], represents a paradigm shift in generative models, grounded in Bayesian inference. VAEs aim to learn a model that assigns a high likelihood $p(x)$ to the observed data samples x. Unlike direct likelihood modelling, which is challenging due to data complexity, VAEs use lower-dimensional latent variables z, typically formulated as Gaussian distribution parameterised by μ and σ. This approach is analogous to viewing data space as the manifestation of underlying 'genetic' codes in the latent space.

The likelihood is obtained by marginalising the latent variable from the joint distribution or applying the chain rule of probability, expressed as:

$$\begin{aligned} p(x) &= \int p(x,z)\ dz \\ p(x) &= \frac{p(x,z)}{p(z|x)} = \frac{p(x|z)p(z)}{q(z|x)} \end{aligned} \tag{3.13}$$

In this formulation, $p(x,z)$ is the joint distribution, $p(x)$ the marginal, $p(x|z)$ the likelihood, $p(z)$ the prior, and $q(z|x)$ the intractable true posterior.

However, as it is not feasible to marginalise over all the possible latent variables in the latent space nor get access to a real latent encoder $q(z|x)$. VAEs introduces a posterior distribution of z based on the observation x is introduced, $q_\phi(z|x)$, to approximate $q(z|x)$. Optimising the posterior allows for marginalising z by sampling from $q_\phi(z|x)$, bypassing $q(z|x)$. The objective is to maximise the marginal log-likelihood,

$\log p(x)$, or equivalently, its Lower Bound (ELBO):

$$\begin{aligned}\log p(x) &= \log \int p(x,z)\ dz \\ &= \log \int \frac{p(x,z)q_\phi(z|x)}{q_\phi(z|x)}\ dz \\ &= \log \mathbb{E}_{q_\phi(z|x)}\left[\frac{p(x,z)}{q_\phi(z|x)}\right] \\ &\geq \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p(x,z)}{q_\phi(z|x)}\right]\end{aligned} \tag{3.14}$$

Decomposing ELBO from Eq.5.3 reveals its two components: the reconstruction likelihood and the Kullback-Leibler (KL) Divergence between the approximated posterior and the prior:

$$\begin{aligned}\mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p(x,z)}{q_\phi(z|x)}\right] &= \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p(x|z)p(z)}{q_\phi(z|x)}\right] \\ &= \mathbb{E}_{q_\phi(z|x)}\left[\log p(x|z)\right] + \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p(z)}{q_\phi(z|x)}\right] \\ &= \mathbb{E}_{q_\phi(z|x)}\left[\log p(x|z)\right] - \mathrm{KL}\left(q_\phi(z|x)\|p(z)\right)\end{aligned} \tag{3.15}$$

Observing the decomposed ELBO from Eq.3.15, in the best case, when the ELBO is powerfully parameterised and perfectly optimised, it becomes exactly equivalent to the evidence. Maximising ELBO entails enhancing reconstruction accuracy while aligning the learned latent distribution with the prior, in other words, maximising the reconstruction likelihood term and minimising the KL Divergence term.

The VAE architecture encompasses an encoder (parameterised by ϕ) approximating the posterior, $q_\phi(z|x)$, and a decoder (parameterised by θ) mapping latent vectors to reconstructions. The prior $p(z)$ is typically formulated as a standard Gaussian, $\mathcal{N}(0, \boldsymbol{I})$. The approximated posterior and likelihood are usually modelled as MVGs parameterised by mean and covariance. The covariance however is usually adopted in diagonal form for simplification. Their mathematical forms are expressed as follows:

$$\begin{aligned}q_\phi(z|x) &= \mathcal{N}(z; \mu_\phi(x), \sigma^2_\phi(x)\boldsymbol{I}) \\ p_\theta(x|z) &= \mathcal{N}(x; \mu_\theta(z), \sigma^2_\theta(z)\boldsymbol{I})\end{aligned} \tag{3.16}$$

The optimisation involves the ELBO, with the reconstruction term estimated via Monte Carlo sampling and the KL divergence explicitly computed. The loss function depicts

the above-mentioned optimisation process can be expressed as follows:

$$\begin{aligned}\mathcal{L}(\theta, \phi; x) &= -\mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] + KL\left(q_\phi(z|x)\|p(z)\right) \\ &\approx -\frac{1}{L}\sum_{l=1}^{L}\log p_\theta(x|z^{(l)}) + KL\left(q_\phi(z|x)\|p(z)\right)\end{aligned} \tag{3.17}$$

Note that the reconstruction term is approximated via the Monte Carlo approach, where L samples $\{z^{(l)}\}_{l=1}^{L}$ are samples from the approximated latent distribution $q_\phi(z|x)$, obtaining L reconstructions. Then the reconstruction term is calculated by averaging the L reconstruction errors. However, stochastic sampling is generally non-differentiable, posing a challenge for backpropagation through the neural network layers. To address this, VAEs employ the reparameterisation trick, which enables gradient-based optimisation despite the stochastic nature of the latent space.

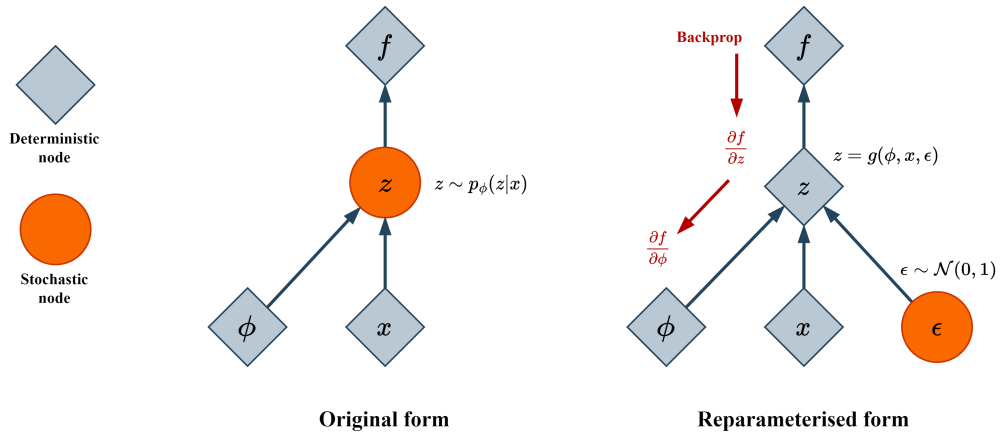


Figure 3.6: **Reparameterisation trick in VAE.** In the original form, the direct sampling makes the computational node for the latent variable a stochastic node, which then blocks the backpropagation flow from the decoder to the encoder. After reparameterisation, the noise variable $\epsilon \sim \mathcal{N}(\epsilon; 0, 1)$ takes the stochastic move. Since ϵ is a leaf node in the computational graph and has no parameters to be optimised, it enables the backpropagation to flow through the network and does not bring in new parameters.

The reparameterisation trick rewrites a random variable as a deterministic function of a noise variable ϵ, allowing the optimisation of the non-stochastic terms through gradient descent. As shown in Fig.3.6, in the original form, direct sampling from the latent distribution creates a stochastic node for the latent variable, impeding the flow

of backpropagation from the decoder to the encoder. By implementing reparameterisation, the noise variable ϵ, drawn from a standard normal distribution $\mathcal{N}(\epsilon; 0, 1)$, assume the stochastic component in the computational graph. This transformation ensures that ϵ remains a leaf node in the computational graph, devoid of any parameters needing optimisation, thereby enabling backpropagation to traverse through the network unobstructed, without introducing additional parameters. The stochastic sampling of the latent variable z is represented as follows:

$$\begin{aligned} z^{(l)} &= \mu + \sigma\epsilon, \text{ with } \epsilon \sim \mathcal{N}(\epsilon; 0, 1) \\ &= \mu_\phi(x) + \sigma_\phi(x) \odot \epsilon^{(l)} \end{aligned} \tag{3.18}$$

where μ_ϕ and σ_ϕ are the outputs of the encoder, given the input x; $\epsilon^{(l)}$ represents the l-th sample from a standard Gaussian distribution; and $\odot$ denotes element-wise multiplication. After training a VAE, the generation of new samples is achieved by direct sampling from the prior distribution $p(z)$ and subsequently processing these samples through the decoder.

In conclusion, Variational Autoencoders are instrumental in learning compact representations, primarily because the dimensionality of the latent space is often significantly lower than that of the input space. This encoded latent space can be effectively utilised in a variety of tasks beyond mere reconstruction, including segmentation and classification. In contrast to other stochastic generative models, VAEs are adept at generating credible new samples, owing to the relatively more constrained latent space (modelled in standard Gaussian). Besides, the probabilistic modelling of the latent space also gives VAEs to assess the uncertainty of its generation, thereby augmenting the interpretability of these samples.

The fundamental concepts of VAEs have spurred a multitude of variations, notably in analysing sequential data [122, 123], conditioned generation [76], and interpretable factorised representation [124]. Moreover, advanced generative models that demonstrate immense promise in producing vast quantities of high-quality outputs, such as Hierarchical Variational Autoencoder (HVAE) [125, 126], Variational Diffusion Models (VMD) [127, 128], and score-based generative models [129, 130], are predicated on the generalised principles of variational auto-encoding.

3.2 Literature review on cardiac sequential analysis

Cardiac sequential analysis encompasses an array of critical tasks that collectively aim to elucidate dynamic aspects of cardiac physiology. This includes but is not limited to, analyses of myocardial motion, blood flow, and volumetric changes within the cardiac chambers across various phases of the cardiac cycle. In recent decades, the advent of deep learning has further refined the accuracy and efficiency of these analyses, paving the way for personalised medical treatment and advanced AI-aided cardiac disease management [131].

The ensuing sections will delve into literature reviews of two primary groups of cardiac sequential analysis. One concentrates on structural and regional analysis, encompassing cardiac ventricular segmentation. The other focuses on cardiac motion modelling, understanding the functional dynamics of the heart. Together, ventricular segmentation and cardiac motion modelling lay the foundation for further motion and regional analysis.

3.2.1 Ventricular segmentation

Medical image segmentation, a combination of the fundamental Computer Vision task of semantic segmentation with medical images, requires more than merely applying general semantic segmentation methods to medical data [132]. It necessitates a profound understanding of segment subjects coupled with a flexible adaptation of semantic segmentation techniques.

Semantic segmentation involves assigning a label to each pixel in an image, essentially performing classification tasks at the pixel level. In cardiac image segmentation, the region of interest is partitioned into regions with semantic (i.e., anatomical) significance, i.e. right ventricle, left ventricle endocardium, and left ventricle epicardium. Accurate segmentation is vital for downstream cardiac analysis tasks like 3D shape reconstruction and cardiac clinical indices estimation, based on quantitative measures extracted from images, such as myocardial mass, wall thickness, left and right ventricle volumes, and ejection fraction. The segmentation maps are important for deriving clinical indices, including left ventricular end-diastolic volume (LVEDV), left ventricular end-systolic volume (LVESV), right ventricular end-diastolic volume (RVEDV), right ventricular end-systolic volume (RVESV), and EF. In addition, these segmentation maps are essential for 3D shape analysis [133, 134], 3D + time cardiac motion analysis

[135], as well as survival prediction [136].

Compared to natural images, the semantic ROIs in CMR images exhibit pronounced correlations both temporally and spatially. For example, most CMR volumes display consistent variation patterns across slices and throughout the cardiac cycle, with the position of ROIs being relatively fixed within the image. However, CMR images acquired from different medical equipment may vary in resolution and contrast. Additionally, the unusual movements during data acquisition, such as breathing, can significantly impact image quality, presenting a unique set of challenges to the segmentation process [137].

Deep learning based approaches

In the past few decades, advancements in computer hardware, notably graphical processing units (GPUs) and tensor processing units (TPUs), combined with an increased availability of training data, have been instrumental in the ascendancy of DL-based segmentation algorithms. These algorithms have gradually surpassed traditional machine learning approaches, such as K-means [138–140], and Level-set based methodologies [141, 142].

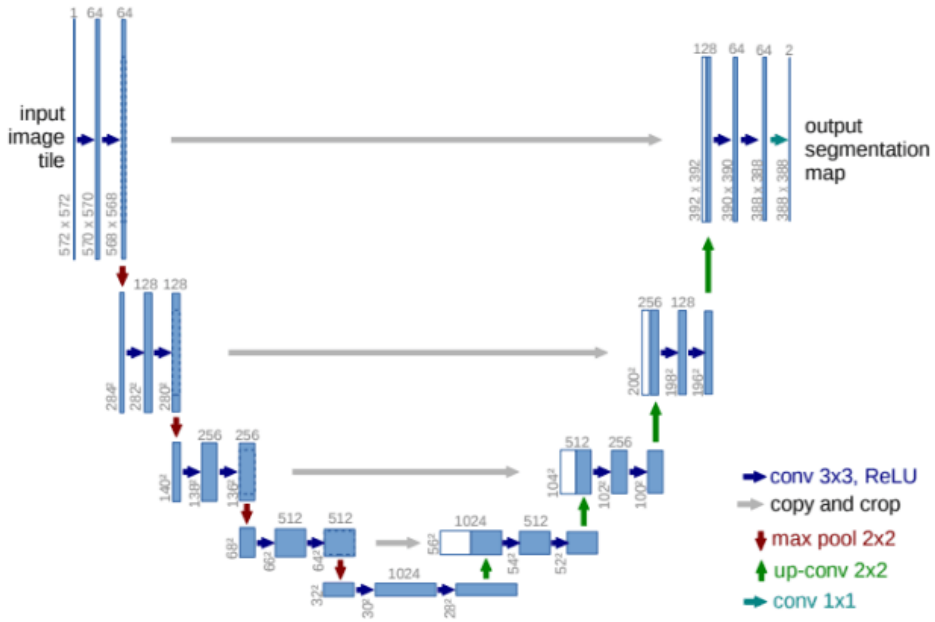


Figure 3.7: **The network architecture of U-Net [9].** Deriving from Fully Convolutional Netowrks [10], the U-Net exploits multi-scale feature extraction and the skipping connections between the encoder and decoder to recover the spatial context loss in the down-sampling path, yielding a more precise segmentation result. The left side of the U-Net is a contracting path, similar to a typical convolutional neural network. It consists of repeated application of two 3×3 convolutions. With each downsampling step, the network doubles the number of feature channels. The right side of the U-Net, the expansive path, includes several up-convolutional layers which increase the resolution of the output. Each step in the expansive path consists of an upsampling of the feature map followed by a 2×2 deconvolution, a concatenation with the correspondingly cropped feature map from the contracting path, and two 3×3 convolutions, each followed by a ReLU. The figure is adapted from [9].

This shift towards DL-based methods in medical image segmentation has been comprehensively reviewed in several studies. Garcia-Garcia *et al.* [23] provide an overview of general semantic segmentation, while Icsin *et al.* [143] focus on general medical image segmentation. The broader applications of DL methods in medical image analysis were detailed by Greenspan *et al.* [144] and Litjens *et al.* [145]. Specifically, in the context of cardiovascular imaging, Gandhi *et al.* [146] and Mazurowski *et al.* [147] have

conducted surveys dedicated to exploring DL applications designed for cardiovascular image analysis. Chen *et al.* [24] provide an in-depth review of DL-based methods in cardiac image segmentation.

Tran *et al.*[148] were among the first ones to apply a Fully Convolutional Network (FCN) [10] for segmenting the left ventricle, myocardium, and right ventricle directly in short-axis cardiac MRIs. Their end-to-end FCN-based approach achieved notable segmentation performance, significantly outperforming traditional methods in both speed and accuracy. Subsequent studies based on FCNs aimed to enhance segmentation further. Some focused on optimising network structures to improve feature learning for segmentation [149–155], such as Khened *et al.* [149], who developed a dense U-Net with inception modules combining multi-scale features for robust segmentation with large anatomical variability. Additionally, Additionally, Jang *et al.* [150], Yang *et al.* [151], Sander *et al.* [156], and Chen *et al.* [157] explored different loss functions, like weighted cross-entropy, weighted Dice loss, deep supervision loss, and focal loss, to enhance segmentation performance. Notably, the majority of these FCN-based methods favoured 2D networks over 3D, mainly due to the typical low through-plane resolution and motion artefacts (miss alignment) in most CMR scans, which limits the applicability of 3D networks [95].

Later researches [9, 10, 158] focus on utilising CNNs to maximise semantic information extraction from training samples. In medical imaging, Ronneberger *et al.* [9] enhanced the fully convolutional layer to function effectively with limited training images and provide more accurate segmentation, particularly in its application to biomedical images. This model, introduced as U-Net, shown in Fig.3.7, comprises a contracting path, an expansive path, and skip connections. The skipping connections between the encoder and decoder recover the spatial context loss in the down-sampling path. By capturing information at different scales, U-Net can segment regions with varying sizes, which is crucial for detailed biomedical imaging tasks. The design and features have set a milestone in this research field, inspiring numerous further developments and adaptations [25, 159, 160]. Yu *et al.* [161] introduce dilated convolutions, blending multi-scale context information without resolution loss to increase the field of perception in convolution operation. Chen *et al.* [8] incorporates atrous spatial pyramid pooling (ASPP) exploiting multi-scale features, which further advanced segmentation precision.

In medical imaging research, semi-supervised and unsupervised approaches are in-

creasingly favoured over purely supervised methodologies. This preference stems, on one hand, from the inherent difficulty in accessing medical images as compared to natural images. On the other hand, the labour-intensive nature of labelling medical images for training purposes often results in a limited pool of labelled data. Moreover, the effectiveness of supervised approaches is heavily dependent on both the volume and quality of the annotated training data. In the specific context of CMR sequence segmentation, the scant number of annotations available, typically confined to only two frames at ED and ES phases, poses significant challenges in developing robust models capable of maintaining spatio-temporal consistency throughout CMR sequences. Additionally, disregarding the potential utility of unlabeled data in the training process can be considered a wasted opportunity, further underscoring the need for methods that can leverage such data effectively.

Recent studies concentrate on achieving anatomically accurate and robust segmentation in various challenging CMR scenarios. For instance, Chen *et al.* [162] and Zotti *et al.* [163] have developed methods to construct a diverse latent space using contextual information from adjacent slices, multi-view images, and offline shape priors, thereby enhancing segmentation performance on difficult slices. In addressing sequential data, Yan *et al.* [164] have integrated optical flow into the U-Net architecture to augment the temporal coherence of left ventricle segmentation. Du *et al.* [165] proposed a framework for CMR sequence segmentation that combines a ConvLSTM [102] with U-Net. Additionally, Bai *et al.* [166] have advanced a sequential segmentation method utilising sparse temporal annotations, constructing pseudo-labels by deforming masks from frames where annotations are available to supervise those without ground truth.

Evaluation metrics for ventricular segmentation

To quantitatively evaluate the performance of segmentation algorithms, three types of metrics are commonly used: (a) volume-based metrics (e.g., Dice coefficient, Jaccard similarity index); (b) surface distance-based metrics (e.g., mean contour distance, Hausdorff distance); (c) clinical performance metrics (e.g., ventricular volume and mass) [167].

The Dice coefficient (DSC) is defined as twice the area of overlap between the predicted and ground truth segmentations, divided by the total number of pixels in

both images. The Dice Loss, which is one minus the Dice coefficient, is formulated as:

$$\text{Dice Loss} = 1 - \frac{2 \times |X \cap Y|}{|X| + |Y|} \tag{3.19}$$

where X and Y represent the predicted and ground truth segmentations, respectively. Hausdorff Distance (HD) is a measure of the distance between two subsets of a metric space. In the context of medical image segmentation, it is used to determine the largest distance from a point in one segmentation to the closest point in the other segmentation. It is defined as:

$$\text{HD}(X, Y) = \max \left\{ \sup_{x \in X} \inf_{y \in Y} d(x, y), \sup_{y \in Y} \inf_{x \in X} d(y, x) \right\} \tag{3.20}$$

where X and Y are two non-empty subsets, representing the predicted and ground truth segmentations. The function $d(x, y)$ denotes the distance between points x and y. The Hausdorff Distance is computed as the maximum of all the distances from a point in one set to the closest point in the other set.

3.2.2 Cardiac motion modelling

Motion modelling is the process of creating a mathematical or computational model to represent the movement of objects or entities. This involves understanding and representing the dynamics and kinematics of motion, often with the aim of predicting future movements or understanding the underlying mechanics. Spatio-temporal motion modelling has been widely used in medical applications such as motion management in radiation therapy, tumour localisation, treatment planning, assessing organ motion in different image modalities [168–170], as well as analysing the heart motion along a cardiac cycle via motion indices [20, 45, 171, 172].

Machine Learning based motion modelling

A common direction of motion modelling is through image registration, which aims to find an optimal spatial transformation aligning two or more images based on specific similarity metrics. Traditional deformable registration algorithms iteratively solve an optimisation problem for each image pair. Prominent examples of such algorithms include Local Correlation Coefficient Demons (LCC-Demons) [16], Large Deformation Diffeomorphic Metric Mapping (LDDMM) [17], and Symmetric Normalisation (SyN)

[173], all of which have demonstrated considerable effectiveness in medical image registration and have been widely adopted in clinical applications. LCC-Demons utilises a local correlation coefficient to improve the registration of images with intensity variations, making it particularly effective for scenarios involving multi-modal image registration. LDDMM models deformations in a diffeomorphic space, ensuring topological preservation and smooth, invertible mappings. This method is renowned for its ability to capture complex anatomical variations, making it invaluable for longitudinal studies and morphological analysis in medical imaging. SyN adopts the symmetric considerations into image registration, which ensures that the transformation from the moving image the the target image is consistent with the transformation from the target back to the moving image. This symmetric approach effectively preserved the diffeomorphic mapping properties, particularly beneficial in inter-subject and inter-modality registrations. However, these traditional techniques are computationally intensive and rely on run-time optimisation (parameter fine-tuning) to achieve optimal performance. In a real clinical setting, where time efficiency and robustness are paramount, more robust and effective frameworks are expected.

Deep Learning based motion modelling

Deep learning techniques have revolutionised the field of image registration by significantly expediting the process for unseen images through the use of trained networks [174]. These techniques are broadly classified into supervised and unsupervised categories. The supervised methods rely on ground-truth displacement vector fields (DVFs), which are often provided by random transformation generation [175–177], conventional ML-based registration methods [178–182], or model-based DVF generation techniques [183, 184]. However, these methods often suffer from noisy labels or flawed ground truth from non-perfect conventional methods, and random transformations often deviate from actual physiological motion, leading to biases in training and subsequent performance degradation. Sokooti *et al.* [184] have shown that training a supervised model for 3D-CT lung image registration using realistic, model-based DVF generation yields more accurate results compared to using random transformations, particularly in terms of registration accuracy. In most medical applications, the lack of training datasets with known ground-truth DVFs limits the utility of the supervised registration algorithms. Besides, the lack of availability of training images of a certain kind

is a big challenge. Uzunova *et al.* [183] proposed a model-based data augmentation scheme to allow for deep learning on small training populations. They adapted the supervised FlowNet [185] architecture for CNN-based optical flow estimation in cardiac images, allowing the augmentations of more training samples, however, with known correspondences.

Unsupervised motion estimation techniques have successfully addressed many challenges associated with supervised models in medical image analysis. Typically, these techniques employ a series of 2D or 3D convolutional layers followed by a spatial transformer layer, forming the architecture of unsupervised deformable image registration (DIR) networks [186–192].

Jaderberg *et al.* [193] introduced the Spatial Transformer Network (STN), a component frequently incorporated in unsupervised registration models [186–190]. The STN enables the deformation of the moving image in a fully differentiable manner, thereby allowing for optimising image similarity during training.

For unsupervised CNN-based models, Balakrishnan *et al.* [186] proposed VoxelMorph, an auto-encoder structured model for pairwise 3D brain MRI volume registration. VoxelMorph utilises a composite loss comprising similarity loss, leveraging Normalised Cross Correlation (NCC) to effectively handle intensity and contrast variations between images; and regularization loss to ensure the smoothness and invertibility of predicted deformations. VoxelMorph demonstrates performance comparable to traditional non-learning-based methods such as ANTs SyN [194, 195] and NiftyReg, particularly in terms of the DICE coefficient for multiple anatomical structures, while offering significantly faster operation, achieving real-time testing run-time [186].

De Vos *et al.* [190] develope the 2D DIRNet model, which includes a CNN regressor, a spatial transformer, and a resampler, tested on MNIST and short-axis cardiac MRI slices. Subsequently, they introduced the 3D Deep Learning Image Registration (DLIR) framework, a multi-stage unsupervised model for affine and deformable image registration, comprising a stack of CNNs [191]. This model, evaluated on cardiac MRIs and chest CT images, achieved results comparable to the conventional SimpleElastix method [196] but with faster processing times [191]. Additionally, FAIM, a CNN model designed for 3D Brain MR image registration, incorporates a penalty loss on negative Jacobian determinants to reduce regions of non-invertibility [192]. Pursuing similar objectives, Zhang *et al.* [197] proposed the Inverse-Consistent Deep Network (ICNet) for

T1-weighted brain MRI, which employs inverse-consistent and anti-folding constraints to maintain the diffeomorphic properties of the transformation.

While learning-based DIR models are invaluable in numerous aspects, they are inherently limited by their deterministic framework, restricting their capacity to generate synthetic motion beyond the scope of registration tasks. In contrast, probabilistic models offer several advantages that extend well beyond mere registration, including:

- Probabilistic frameworks allow for the quantification of uncertainty in registration results. This is crucial in medical applications where decisions are often made under uncertainty, and knowing the confidence level of registration can guide clinical decision-making.
- Probabilistic methods are more robust to the variations, existing in image modalities and patient anatomies, more effectively by learning distributions of high-level feature space instead of producing a single deterministic output.
- Probabilistic models can generate realistic deformations and synthetic images, which can be used for data augmentation. This is particularly valuable in medical imaging, where the availability of annotated data is limited.

Probabilistic motion modelling

Recently, probabilistic frameworks have been proposed for these purposes [172, 187–189, 198]. Dalca *et al.* [13, 18] propose a probabilistic framework that achieves competitive registration results and enables uncertainty estimation on DVFs. They model the latent velocity field as a Multivariate Gaussians distribution and regularise it with standard Gaussian prior. Krebs *et al.* [20, 199] propose an unsupervised cine-MRI registration model based on the conditional variational auto-encoder (CVAE)[200]. The authors use a Gaussian smoothness kernel followed by a differentiable exponentiation layer to obtain diffeomorphism transformations, using symmetric local cross-correlation criterion as the similarity loss function. However, these models lack the constraints of deploying temporal dependencies in the loss function to regulate the continuous motion in a sequence.

More recently, Krebs *et al.* [189] extended their probabilistic model to a spatio-temporal registration method for short-axis cine-MRI images. In this model, time dependencies are modelled using a temporal convolutional network (TCN) [201] and

a temporal dropout (TD) scheme to capture local dependencies over time. The authors performed motion simulation and motion transport by applying the recovered motion from one subject to another [172, 189]. Although the temporal dependencies were elegantly captured via a Gaussian process in the low dimensional latent space, no pixel-wise explicit probability distributions for the deformations were specified. The uncertainties in the estimated deformations remained unexplored and only cardiac cycle generation was demonstrated in terms of heart volumetric variations. Zakeri *et al.* [202] proposed a probabilistic variational recurrent neural network (VRNN) [122] framework for sequential registration on long-axis CMR sequences. The authors adopt a learnable prior on the latent variable and ConvLSTM [102] to capture the spatio-temporal features through the sequence. However, these appearance-similarity-oriented approaches often face challenges in accurately capturing anatomical motion. Emphasis on matching pixel-level intensity similarity between registration pairs results in the neglect of motion at anatomical boundaries, thus failing to preserve anatomical shape after applying DVFs to categorical-level representations, such as semantic masks, and topological representations.

Another advantage of the probabilistic view over other learning-based methods is analytical uncertainty estimation. Clinicians benefit from this information in terms of data analysis and confidence in the model for decision-making. Data-related uncertainty (also referred to as aleatoric uncertainty) and uncertainty in the model parameters and structure (epistemic uncertainty) induce the predictive uncertainty (i.e., the confidence we have in a prediction) [203]. However, they are difficult to assess in a high-dimensional complex model, needing an uncertainty quantification approach proposed in Bayesian neural networks [203, 204]. One practical approach for approximate inference of the uncertainties is to execute stochastic forward pass when applying dropouts to weights [205–207]. However, this strategy increases inconsistent outputs [208]. By sampling from the learned velocity fields, propagating them through the diffeomorphic layers to calculate the deformation fields, and calculating the empirical diagonal covariance across samples, Dalca *et al.* [13] describe an empirical method for estimating uncertainty for motion fields.

In summary, while the field of cardiac motion modelling has seen remarkable advancements through machine learning and DL-based approaches, there remain significant challenges. These include the need for more effective handling of temporal depend-

encies, better modelling of pixel-level uncertainties, and the preservation of anatomical shape in medical imaging.

3.3 Datasets

3.3.1 UK Biobank (UKBB)

The UK Biobank dataset (UKBB) was established with the foresight to provide a comprehensive data platform to facilitate multifaceted research, UKBB has meticulously collected and curated data from approximately 500,000 participants, aged between 40 and 69 years, across the UK during the recruitment period from 2006 to 2010, adhering to a detailed data collection protocol [11]. This extensive dataset includes health questionnaires, physical measurements, biological samples, and a wealth of imaging data, offering a comprehensive view of factors influencing diseases in middle and old age [209]. The imaging data comprises MRI, dual-energy X-ray absorptiometry (DXA), and CT

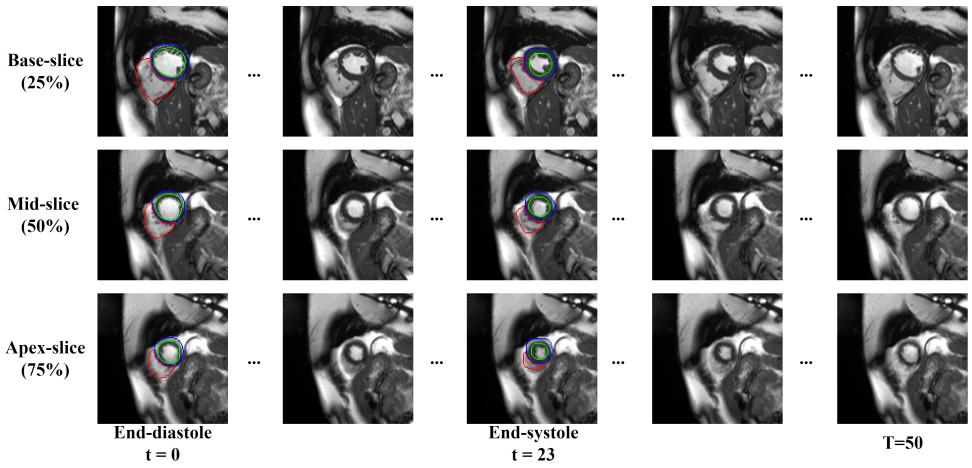


Figure 3.8: **Showcase of short-axis CMR sequence in UKBB.** As per the UKBB data collection protocol [11], therapists and clinicians acquire 50 frames per patient to cover a full cardiac cycle in cine-CMR scans. Each short-axis (SAX) stack consists of 8 to 13 slices of scenes. Each SAX scan slice is captured with a 208 *times* 187 matrix size and an in-plane resolution of 1.8 $\times$ 1.8 mm^2. Ground truth annotations for the LVendo, LVepi, and RV are available for ED and ES phases, with ED at the 0-th frame and ES varying from the 21-st to 26-th frame, depending on the patient. The reference CMR images were reproduced with the permission of UK Biobank©.

scans, among others, in various modalities like the brain, heart, bones, and abdomen. Focusing on CMR data, Petersen *et al.* [210] manually analysed the CMR of 5,065 consecutive UKB participants. The authors manually segmented all slices of each 3D CMR scan at the ED and ES phases. The annotation and assessment were performed across two core laboratories based in London and Oxford. Through CISTIB's collaboration under UK Biobank Application 1135, we accessed around 5,000 CMR samples with manually drawn contours for the anatomical structures in the left ventricle (LV) and right ventricle (RV). As illustrated in Fig.3.8, As part of CISTIB's collaboration under UK Biobank Application 1135, we have access to around 5,000 CMR samples with manually drawn contours for LV and RV. Referring to the showcase shown in Fig.3.8, for each patient, 50 frames of SAX stacks are collected covering a full cardiac cycle under the b-SSFP. Each SAX stack comprises 8 to 13 slices. The SAX scan slices have a 208 *times* 187 matrix size and an in-plane resolution of $1.8 \times 1.8\ mm^2$. Ground truth annotations for the left ventricle endocardium, epicardium, and right ventricle are available for ED and ES phases, with ED at the 0-th frame and ES varying from the 21-st to 26-th frame, depending on the patient.

3.3.2 Automated Cardiac Diagnosis Challenge (ACDC)

The Automated Cardiac Diagnosis Challenge (ACDC) dataset has become a crucial public resource in cardiac imaging, addressing the intricate challenges of CMR image analysis. Introduced during the ACDC challenge at the Medical Image Computing and Computer Assisted Intervention (MICCAI) 2017, this dataset has been carefully curated to advance research in automated cardiac pathology analysis using CMR data [34].

The dataset encompasses 150 patients, divided into five subgroups (four pathological and one healthy), with defined classification criteria. Additional data on weight, height, the number of frames collected, and the timing of diastolic and systolic phases are also provided.

According to the data collection protocol [34], the data were collected using two MRI scanners of varying magnetic strengths: 1.5 Tesla (Siemens Area, Siemens Medical Solutions, Germany) and 3.0 Tesla (Siemens Trio Tim, Siemens Medical Solutions, Germany). Cine-MRI were acquired in a breath-hold using retrospective or prospective gating and a Steady-State Free Precession (SSFP) sequence in a SAX orientation. Notably, a series of SAX slices cover the LV and RV from base to apex, with slice

thicknesses of 5 mm or occasionally 8 mm, and sometimes an interslice gap of 5 mm. The spatial resolution ranges from 1.37 to 1.68 mm^2/pixel, and 28 to 40 images cover the entire or partial cardiac cycle. With prospective gating, 5 to 10% of the end of the cardiac cycle may be omitted, depending on the patient.

The test set of ACDC has 100 SAX CMR sequences throughout a cardiac cycle, out of which 20 are normal and 80 cases cover various well-defined pathologies such as previous myocardial infarction (MINF), dilated cardiomyopathy (DCM), hypertrophic cardiomyopathy (HCM), and abnormal right ventricle (ARV). The anatomical structure segmentations are provided at end-diastolic and end-systolic volumes.

The dataset is stratified for balanced representation across different pathologies, offering a robust platform for developing and validating algorithms in CMR analysis, especially in segmentation and cardiac motion analysis.

3.3.3 Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Segmentation (M&Ms) Challenge

In the realm of cardiac image analysis, the Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Segmentation (M&Ms) Challenge has emerged as a pivotal initiative, aiming to address the inherent complexities and variations introduced by multi-center and multi-vendor data [12].

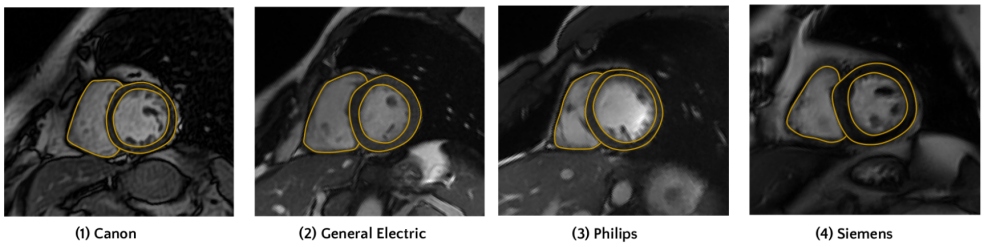


Figure 3.9: **Showcase of short-axis CMR sequence in M&M.** Visual appearance of a CMR short-axis middle slice for anatomically similar subjects in the four different vendors considered. The figure is adapted from [12].

The M&Ms dataset is characterised by its diversity, encompassing cardiac magnetic resonance (CMR) images from multiple centres, acquired using scanners from different vendors, and representing a wide array of cardiac pathologies, including coronary heart disease, cardiomyopathies, abnormal right ventricle, or myocarditis, and a variety of

imaging protocols and cardiology units. M&M covers 375 cases from four vendors, six centres, and three countries. A showcase of CMR samples from four different vendors is shown in Fig.3.9. This diversity introduces variations in image quality, resolution, and contrast, posing significant challenges for algorithm generalisation and robustness. The dataset includes a rich set of annotations, providing valuable ground truth for the segmentation of the LV, RV, and LVmyo, thereby facilitating research in automated cardiac segmentation across varied and challenging conditions.

3.4 Preliminary study: A probabilistic model for cardiac sequential motion modelling

The following section provides a brief introduction to the preliminary study, a probabilistic model for cardiac sequential motion modelling proposed in DragNet [211]. This framework has laid a profound foundation for the works introduced in the following chapters.

DragNet leverages the generative nature of the VAE in combination with RNN [122] for conducting motion modelling in LAX CMR sequences and generating synthetic datasets. As a probabilistic spatio-temporal registration framework, DragNet models the temporal motion across CMR sequence via capturing the temporal dependencies in the latent space. As shown in Fig.3.10, DragNet framework comprises five modules: a shared feature extractor followed by two neural networks to compute the parameters of the prior and the posterior distributions of the latent variables. Subsequently, a decoder block predicts the posterior distributions of the displacement field. The deterministic recurrent parameters are managed via a Conv-LSTM layer, alongside a spatial transformer network (STN) layer that warps the moving image from the previous frame (or a specific reference frame) to the fixed image for each time step. The optimisation process of DragNet follows the previous learning-based methods [28, 212], which maximises the reconstruction similarity with a weighted normalisation term over the spatial smoothness of the predicted displacement fields.

DragNet demonstrates significantly faster and is capable of generating smoother temporal deformations compared to conventional image registration techniques. Besides, it can generate synthetic motion sequences, which can be utilised for recovering missing frames in a cardiac sequence, validation of supervised DIR algorithms, and

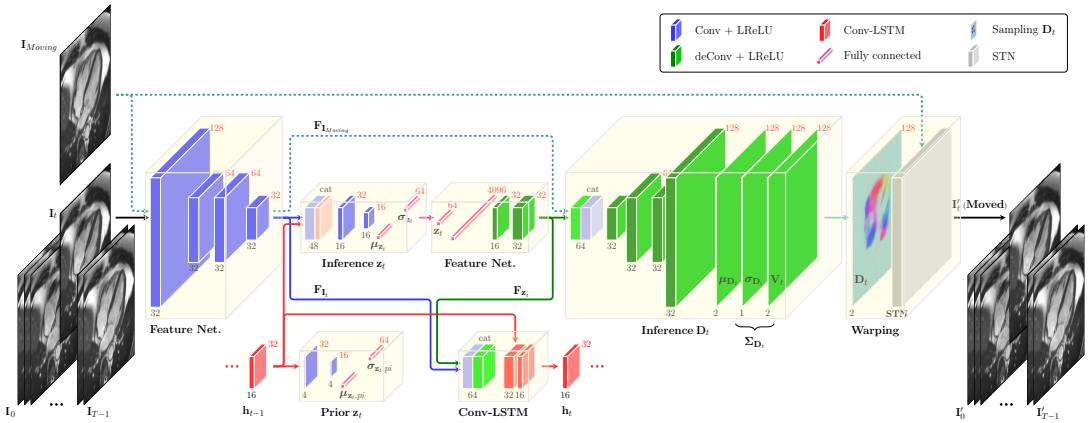


Figure 3.10: The architecture of the proposed DragNet is illustrated. The model comprises five main modules: neural networks to compute the parameters of the prior and the posterior distributions of the latent variables ($\mathbf{z}_t$), posterior distributions of the displacement field ($\mathbf{D}_t$), the deterministic recurrent parameters ($\mathbf{h}_t$) via Conv-LSTM layer, and a spatial transformer network (STN) layer that warps the moving image from the previous frame (or a specific reference frame) to the fixed image at time t. The CMR images were reproduced with a permission of UK Biobank©.

even enabling in-silico trials modelling a specific moving organ.

The formulation of probabilistic modelling of cardiac motion rooted in VAE and utilisation of RNNs in capturing spatio-temporal features across CMR sequences has laid a foundation for further explorations in this thesis. In the following chapters, three directions of exploration are derived from this DL design of probabilistic cardiac motion modelling: (1) **Chapter 4** explores the auto-balance of the trade-off between registration accuracy and the smoothness of predicted DVFs. (2) **Chapter 5** extends the spatio-temporal probabilistic generative model to conduct concurrent sequential segmentation and motion modelling using SAX cine-CMR. (3) **Chapter 6** explores the latent space disentanglement and conditioned generation to perform motion transfer between healthy patients and patients with cardiac motion disorders.

Chapter 4

GSMorph: Balancing accuracy and diffeomorphism with gradient surgery

4.1 Introduction

Image registration is fundamental to many medical image analysis applications, e.g., motion tracking, atlas construction, and disease diagnosis [213]. Conventional registration methods usually require computationally expensive iterative optimisation, making it inefficient in clinical practice [214, 215]. Deep learning has recently been widely exploited in the registration domain due to its superior representation extraction capability and fast inference speed [13, 216]. Deep-learning-based registration (DLR) formulates registration as a network learning process minimising a composite objective function comprising one similarity loss to penalise the difference in the appearance of the image pair and a regularisation term to ensure the smoothness of the deformation field. Typically, to balance the registration accuracy and smoothness of the deformation field, a hyperparameter is introduced in the objective function. However, performing hyperparameter tuning is labour-intensive, time-consuming, and *ad-hoc*; searching for the optimal parameter setting requires extensive ablation studies and hence training tens of models and establishing a reasonable parameter search space. Therefore, alleviating, even circumventing, hyperparameter search to accelerate the development and deployment of DLR models remains challenging.

Recent advances [217–219] in DLR have primarily focused on network architecture design to boost registration performance. However, a complex architecture may lead to further requirements of time and resources in hyperparameter tuning. Few studies [220, 221] investigated the potential in preventing hyperparameter searching by hypernetwork [222] and conditional learning [223]. Hoopes *et al.* [220] leveraged a hypernetwork that takes the hyperparameter as input to generate the weight of the DLR network. Although effective, it introduces numerous additional parameters to the DLR network, making the framework computationally expensive. In parallel, Mok *et al.* [221] proposed to learn the effect of the hyperparameter and condition it on the feature statistics (usually illustrated as *style* in computer vision [223]) to manipulate the smoothness of the deformation field in the inference phase. Both methods can avoid hyperparameter tuning while training the DLR model. However, they still require a reasonable sampling of the hyperparameter space and strategy, which can be empirically dependent.

Gradient surgery (GS) projects conflicting gradients of different loss functions during the optimisation process of the model to mitigate gradient interference. This has

proven useful in multi-task learning [224] and domain generalization [225]. Motivated by these studies, we propose utilising the GS to moderate the discordance between similarity and regularisation loss. Furthermore, our method can avert searching hyper-parameters in training the DLR network by only involving gradient regularisation in direction rather than magnitude. Our contributions are as follows:

- We propose GSMorph, a gradient-surgery-based hyperparameter-free DLR model. Our method can circumvent hyperparameter tuning with a gradient-level reformulated objective function to reach the trade-off between registration accuracy and the smoothness of the deformation field.
- Existing GS approaches have operated the parameters' gradients independently or integrally. We propose a layer-wise GS to group by the parameters for optimisation to ensure the flexibility and robustness of the optimisation process.
- Our method is model-agnostic and can be integrated into any DLR network without extra parameters or losing inference speed.

4.2 Methodology

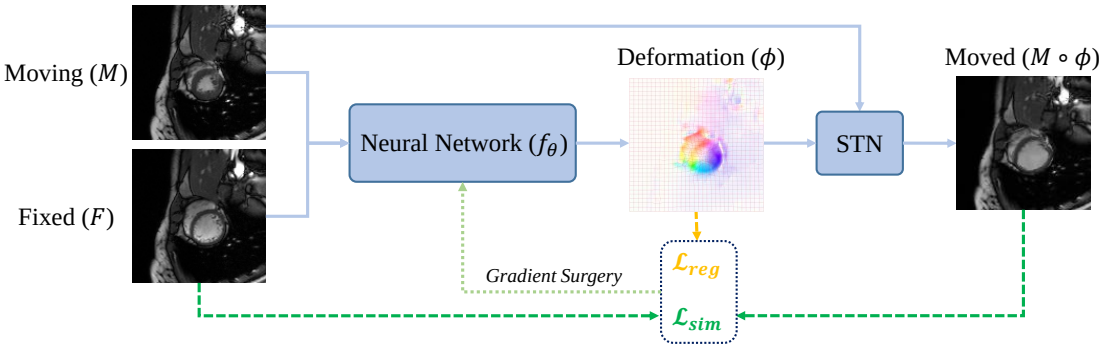


Figure 4.1: Schematic illustration of our proposed GSMorph. GS modifies the gradients computed by similarity loss $\mathcal{L}_{sim}$ and regularization loss $\mathcal{L}_{reg}$, then updates the parameters θ. The GSMorph is akin to the architecture of VoxelMorph [13], which takes the moving and fixed images as input and generates the deformations fields. And STN refers to the spatial transform network.

Deformable image registration estimates the non-linear correspondence (Fig. 6.1) field ϕ between the moving, M, and fixed, F, images. Such procedure is mathematically formulated as $\phi = f_\theta(F, M)$. For learning-based registration methods, f_θ (usually adopted as a neural network) takes the fixed and moving image pair as input and outputs the deformation field via the optimal parameters θ. Typically, θ can be updated using standard mini-batch gradient descent as follows:

$$\theta := \theta - \alpha \nabla_\theta \left(\mathcal{L}_{sim}(\theta; F, M \circ \phi) + \lambda \mathcal{L}_{Reg}(\theta; \phi) \right) \tag{4.1}$$

where α is the learning rate; $\mathcal{L}_{sim}$ is the similarity loss to penalise differences in the appearance of the moving and fixed images (e.g., mean square error, mutual information or local negative cross-correlation); $\mathcal{L}_{reg}$ is the regularisation loss to encourage the smoothness of the deformation field (this can be computed by the gradient of the deformation field); λ is the hyperparameter balancing the trade-off between $\mathcal{L}_{sim}$ and $\mathcal{L}_{reg}$ to achieve desired registration accuracy while preserving the smoothness of the deformation field in the meantime. However, hyperparameter tuning is time-consuming and highly experience-dependent, making it tough to reach the optimal solution.

Insight into the optimisation procedure in Eq. 4.1, as registration accuracy and spatial smoothness are potentially controversial in model optimisation, the two constraints might have different directions and strengths while going through the gradient descent. Based on this, we provide a geometric view to depict the gradient changes for θ based on the *gradient surgery* technique. The conflicting relationship between two controversial constraints can be geometrically projected as orthogonal vectors. Depending on the orthogonal relationship, merely updating the gradients of the similarity loss would automatically associate with the updates of the regularisation term. In this way, we avoid tuning the hyperparameter λ to optimise θ. The Eq. 4.1 can then be rewritten into a non-hyperparameter pattern:

$$\theta := \theta - \alpha \Phi(\nabla_\theta \mathcal{L}_{sim}(\theta; F, M \circ \phi), \nabla_\theta \mathcal{L}_{Reg}(\theta; \phi)) \tag{4.2}$$

where $\Phi(\cdot)$ is the operation of proposed GS method.

4.2.1 Layer-wise Gradient Surgery

Figure 4.2 illustrates the two scenarios of gradients while optimising the DLR network via vanilla gradient descent or gradient surgery. We first define that the gradient

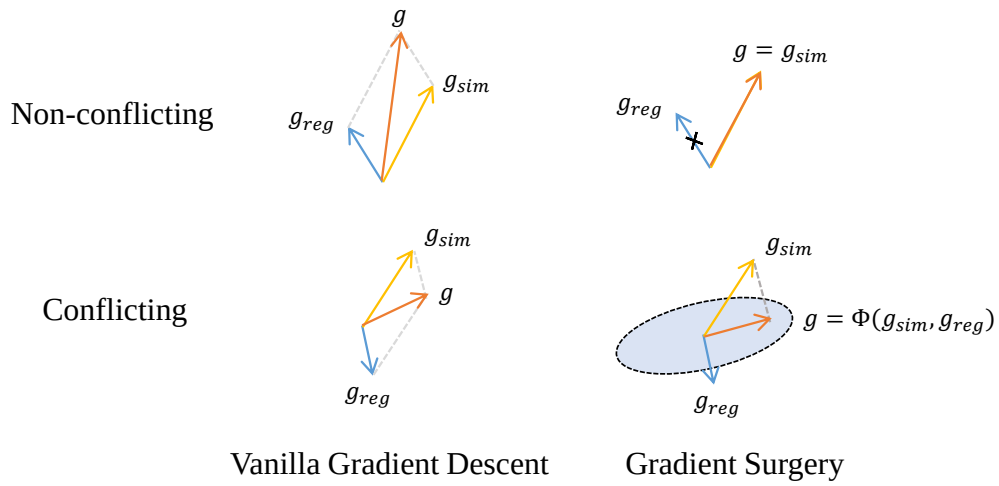


Figure 4.2: Visualization of vanilla gradient descent and gradient surgery for non-conflicting and conflicting gradients. Regarding vanilla gradient descent, the gradient, g, is computed based on the average of g_{sim} and g_{reg}. Our GS-based approach projects the g_{sim} onto the normal vector of g_{reg} to prevent disagreements between the similarity loss and regularisation loss. On the other hand, we only update the g_{sim} in non-conflicting scenarios.

of similarity loss, g_{sim}, and that of regularisation loss, g_{reg}, are conflicting when the angle between g_{sim} and g_{reg} is the obtuse angle, viz. $\langle g_{sim}, g_{reg} \rangle < 0$. In this study, we propose updating the parameters of neural networks by the original g_{sim} independently, when g_{sim} and g_{reg} are non-conflicting because g_{sim} has no incompatible component of the gradient along the direction of g_{reg}. Consequently, optimisation with sole g_{sim} within a non-conflicting scenario can inherently facilitate the spatial smoothness of deformations.

Conversely, as shown in Fig. 4.2, conflicting gradients are the dominant reason associated with non-smooth deformations. Hence, deconflicting gradients in the optimisation of the DLR network to ensure high registration accuracy, as well as smooth deformation, is the primary goal of our study. Following a simple and intuitive procedure, we project the g_{sim} onto the normal plane of the g_{reg}, where the projected similarity gradient g and g_{reg} are non-conflicting along each gradient's direction.

Existing studies [224, 225] performed the GS in terms of independent parameters or the entire network. Despite the effectiveness, these can be either unstable or inflexible. Considering that a neural network usually extracts features through the collaboration

of each parameter group in the convolution layers, we introduce a layer-wise GS to ensure its stability and flexibility. The parameter updating rule is detailed in the Algorithm 1. Specifically, in each gradient updating iteration, we first compute the gradients of two losses for the parameter group in each layer separately. Then, the conflicting relationship between the two gradients is calculated based on their inner product. Once the two gradients are non-conflicting, the gradients used to update its corresponding parameter group will be only the original gradients of similarity loss; on the contrary, the gradients will be the projected similarity gradients orthogonal to the gradients of regularisation, which can be calculated as $g^i_{sim} - \frac{\langle g^i_{sim}, g^i_{reg} \rangle}{\|g^i_{reg}\|^2} g^i_{reg}$. After performing GS on all layer-wise parameter groups in the network, the final gradients will be used to update the model's parameters.

Algorithm 1 Gradient surgery

Require: Parameters θ_i in ith layer of the network; Number of layers in the network N.

$g_{sim} \leftarrow \nabla_\theta \mathcal{L}_{sim}$
$g_{reg} \leftarrow \nabla_\theta \mathcal{L}_{reg}$
for $i = 1 \rightarrow N$ **do**
    if $\langle g^i_{sim}, g^i_{reg} \rangle > 0$ **then**
        $g_i = g^i_{sim}$
    else
        $g_i = g^i_{sim} - \frac{\langle g^i_{sim}, g^i_{reg} \rangle}{\|g^i_{reg}\|^2} g^i_{reg}$
    end if
    $\Delta\theta_i = g_i$
end for
Update θ

4.2.2 Network Architecture

Our network architecture (seen in Fig. 6.1) is similar to VoxelMorph [13] that comprises naive U-Net [9] and spatial transform network (STN) [226]. The U-Net takes the moving and fixed image pair as inputs and outputs the deformation field, which is used to warp the moving image via STN. The U-Net consists of an encoder and a decoder with skip connections, which forward the features from each layer in the encoder to the corresponding layer in the decoder by concatenation to enhance the feature aggregation

and prevent gradient vanishing. The number of feature maps in the encoder part of the network is 16, 32, 64, 128, and 256, increasing the number of features as their size shrinks, and vice versa in the decoder part. Each convolutional block in the encoder and decoder has two sequential convolutions with a kernel size of 3, followed by a batch normalisation and a leaky rectified linear unit.

4.3 Experiments and Results

4.3.1 Datasets and Implementations

Datasets

In this study, we used the public ACDC dataset [34], which contains 100 subjects, for investigation and comparison. We randomly split them into 75, 5, and 20 for training, validation, and testing, respectively. We selected the image from the cine-MRI cardiac sequence at the End Systole (ES) time point of the cardiac cycle as the moving image, and that at the End Diastole (ED) as the fixed one. All images were cropped into the size of 128×128 centralised to the heart. We normalised the intensity of images into the range from 0 to 1 before inputting them into the model.

Implementation details

We implemented our model in PyTorch [227], using a standard PC with an NVIDIA GTX 2080ti GPU. We trained the network through Adam optimizer [228] with a learning rate of 5e-3 and a batch size of 32 for 500 epochs, which cost ∼3 hours. We also implemented and trained alternative methods for comparison with the same data and similar hyper-parameters for optimisation. Our source code will be available online soon.

4.3.2 Alternative Methods

To demonstrate the advantages of our proposed method in medical image registration, we compared it with two conventional deformable registration methods, i.e., **Demons** [214] and **SyN** [215], and the widely-used DLR model, **VoxelMorph** [13]. These methods usually need laborious effort in hyperparameter tuning. Meanwhile, we compared our approach to two alternative DLR models based on the hyperpara-

meter learning, i.e., **HyperMorph** [220] and **CIR-DM** [221]. Those methods only require additional validations in searching the optimal hyperparameter without necessarily tuning it from scratch. Finally, we reformulated two variations of GS based on our concept for further comparison. Specifically, **GS-Agr** [225] treats the gradient of each parameter independently. It updates the parameter with the gradient of similarity loss in the non-conflicting scenario, and a random gradient sampled from the Gaussian distribution when conflicting. While **GS-PCGrad** [224] uses the same GS strategy as ours, but with respect to the whole parameters of the entire network. The **Initial** represents the results without any deformation.

4.3.3 Evaluation Criteria

In this study, we used six criteria to evaluate the efficacy and efficiency of the investigated methods in terms of similarity in appearance, comparability in anatomy, smoothness of deformation fields, and practicality in computational cost. Concretely, Dice score (Dice) and 95% Hausdorff distance (HD95) were used to validate the registration accuracy of the regions of interest; Mean square error (MSE) was leveraged to evaluate the pixel-level appearance difference between the moved and fixed image-pairs; the percentage of pixels with negative Jacobian determinant (NJD) values was utilised to compare the smoothness and diffeomorphism of the deformation field; the number of parameters (Param) of the neural network and inference speed (Speed) were adopted to investigate the efficiency.

4.3.4 Results

As summarised in Table 4.1, our method outperformed other investigated methods in terms of almost all evaluation metrics, which gives average Dice of 87.45%, HD95 of 1.34 mm, and MSE of 0.31×10^{-2} with only 0.87% pixels of NJD. The Dice and HD95 reported in Table 4.1 were averaged over three anatomical regions of interest in the heart, i.e., Left ventricle, Myocardium, and Right Ventricle (LV, Myo, and RV). Consequently, the proposed model achieved superior registration accuracy and spatial regularisation with faster inference speed compared with the two conventional registration methods. Meanwhile, it has made an average of 0.82% and 0.05% improvement over VoxelMorph in terms of Dice and NJD. We also observed that our approach gained higher registration performance compared with two alternative models (HpyerMorph and CIR-DM).

Table 4.1: Quantitative comparison of investigated methods on ACDC dataset. (mean±std; The best results are shown in **bold**.

Methods	Dice(%)	HD95(mm)	MSE(10^{-2})	NJD(%)	Params	Speed(sec)
Initial	61.81±8.68	4.40±1.33	1.58±0.52	-	-	-
Demons [214]	85.38±3.52	1.67±0.75	0.46±0.21	1.31±0.59	-	16.04±6.15
SyN [215]	79.28±8.23	2.24±1.28	0.65±0.21	0.30±0.27	-	8.66±3.14
VoxelMorph [216]	86.63±2.29	**1.34±0.29**	0.36±0.14	1.79±0.92	1.96M	2.29±0.83
HyperMorph [220]	83.44±3.55	1.75±0.64	0.47±0.20	1.60±0.86	126M	2.96±1.09
CIR-DM [221]	87.21±2.43	1.41±0.44	0.34±0.12	**0.15±0.16**	4.91M	3.03±1.15
GS-Agr [225]	63.40±9.15	4.20±1.35	1.33±0.43	0	1.96M	2.29±0.83
GS-PCGrad [224]	84.59±3.53	1.62±0.53	0.51±0.16	0.11±0.17	1.96M	2.29±0.83
GSMorph	**87.45±2.27**	**1.34±0.40**	**0.31±0.11**	0.87±0.52	1.96M	2.29±0.83

The CIR-DM was built on a multi-resolution strategy through the composed multi-scale deformation and deep supervision [229], while our network architecture was only constructed on a naive U-Net. Regarding the GS-based methods, GS-Agr totally collapsed, as the conflicting gradients accounted for most have been replaced by random noise. On the other hand, GS-PCGrad only yielded an inadequate registration performance with an inclination of over-regularisation. The comparison in the GS-based method shows the flexibility and robustness of our approach.

Figure 4.3 illustrates the registration accuracy of the compared methods on three different anatomical structures. It shows that the proposed model has achieved the best performance on LV as well as comparable accuracy to other state-of-the-art methods in Myo and RV. Overall, the results of the comparisons in Table 4.1 and Fig. 4.3 indicate that our method performed the best among all the techniques that we implemented and examined, showing the effectiveness of our model in balancing the trade-off between registration accuracy and smoothness of deformations.

In Table 4.1, we have also reported the number of parameters and inference speed. We observed that DLR methods could obtain faster speed compared with conventional ones in general. As our proposed approach only modified the optimisation procedure of the backbone network, it could maintain the original inference speed and the number of parameters. Conversely, the hyperparameter-learning-based methods (CIR-DM and HyperMorph) introduced extra parameters and loss of inference speed as they adopted the secondary network to generate the conditions or weights of the main network architecture.

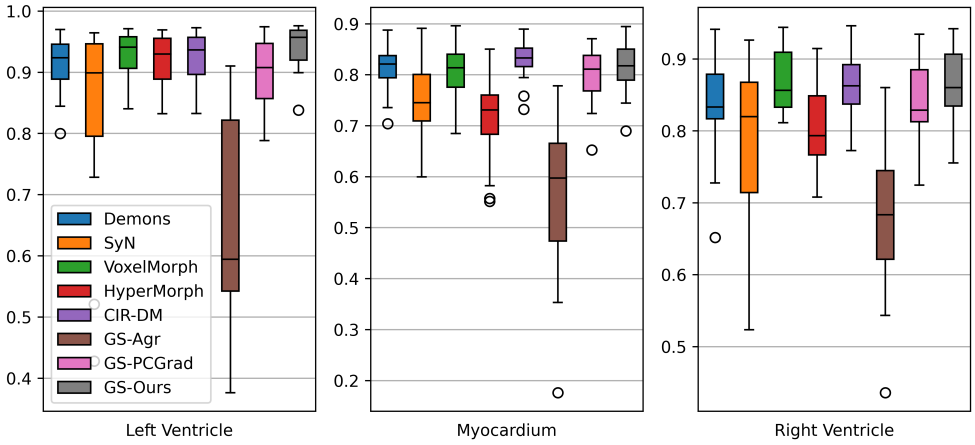


Figure 4.3: Boxplots of Dice scores of the LV, Myo and RV for all the investigated methods.

4.4 Discussion

The trade-off between registration accuracy and smoothness of deformation is essential in learning-based medical image registration methods. Typically, a hyperparameter λ is leveraged in the objective function to balance the effects of the similarity loss and regularisation item. However, tuning hyperparameters inevitably depends on extensive trials, which is labour-intensive and time-consuming. Meanwhile, it is usually empirical-dependent because the performances associated with different hyperparameters are sensitive to the choice of the search space. Therefore, for the first time, we propose a GS-based learning scheme that optimises the parameters only through the projected gradient computed by similarity loss, thus bypassing the requirement of the hyperparameter λ. Besides, our approach is model-agnostic and can be a plug-and-play to any DLR model without introducing extra parameters or computational cost.

Demons and SyN are two powerful conventional registration methods. However, they both require repeating the inefficient iterative optimisation for each fixed and moving image pair. The long-running time (seen in Table 4.1) makes them less practical. Benefiting from the GPU acceleration, Voxelmorph enjoys fast inference speed but is often trapped in the dilemma of balancing registration accuracy and spatial smoothness. It has to take numerous trials to search for an optimal model. Our proposed method projects the gradients from similarity loss onto the orthogonal relation

to those from regularisation constraint, which enables the updated similarity gradients to implicitly contain the characteristic of regularisation, thus avoiding introducing the hyperparameter into the optimisation process.

Prior studies [220, 221] have investigated the potential of bypassing the hyperparameter tuning in the medical image registration network. Those models make the hyperparameter λ an additional input to enable the network to perceive the effect of λ in the objective function. Specifically, HyperMorph leverages multiple fully connected layers to generate the weights of the registration network. However, such a design introduces numerous parameters in this model (enlarging more than 50 times as reported in Table 4.1), bringing a burden to model optimisation. On the other hand, CIR-DM proposes embedding λ into the statistic characteristic of the feature maps, which requires fewer parameters than HyperMorph. Although effective, the performance improvements of CIR-DM are mainly attributed to its multi-scale designed network architecture, rather than its conditional learning scheme. Furthermore, both HyperMorph and CIR-DM require multiple validations to determine the final optimal λ for the test dataset, which has to interact with clinical experts once lacking annotations. More importantly, they require setting empirical searching space for sampling the potential hyperparameters. Concretely, CIR-DM sets the sampling space as [0, 10] when using the NCC as similarity loss; HyperMorph sets the sampling space as [0, 1] when using the MSE loss but pre-scales the loss by estimated image noise. Compared to the aforementioned methods, our GSMorph improves the learning procedure without changing the network architecture. This enables our proposed approach to be model-agnostic and easily plugged into other registration networks.

Gradient surgery has been previously promoted in the domain of generalisation and multi-task learning. Our method exploits the concept of GS in deconflicting gradients from multi-task/domain to modify the gradients from similarity loss to enable it with less irregularity. Due to the inherent disagreement of similarity and regularisation loss, most of their gradients are in a conflicting scenario. It results in the model collapse of GS-Agr, where only a few parameters are updated in limited iterations during optimisation. In other words, the network is similar to being imposed by the dropout with a high dropout rate. Regarding GS-PCGrad, the scenario mentioned above will become more obvious, because it computes the cosine similarity over the entire parameter space. However, instead of replacing the gradient with random noise

like GS-Agr, GS-PCGrad uses a similar gradient projecting strategy as ours, making it still trainable, but inclined to give the over-regularised deformation.

4.5 Conclusion

This work presents a hyperparameter-free registration framework for medical images with a grouped gradient-surgery strategy. To the best of our knowledge, this is the first study to employ gradient surgery to refine the optimisation procedure in learning the deformation fields. In our GSMorph, the gradients from the similarity constraint were projected onto the plane orthogonal to those from the regularisation term. In this way, merely updating the gradients in optimising the registration accuracy would result in a joint updating of the gradients from the similarity and regularity constraints. Then, no additional regularisation loss is required in the network optimization and no hyperparameter is further required to explicitly trade off between registration accuracy and spatial smoothness. Our model outperformed the conventional registration methods and the alternative state-of-the-art DLR models. Finally, the proposed method is model-agnostic and can be integrated into any DLR network without introducing extra parameters or compromising the inference speed. We believe GSMorph will facilitate the development and deployment of DLR models and alleviate the influence of hyperparameters on performance.

Chapter 5

SegMorph: Concurrent Motion Estimation and Segmentation for Cardiac MRI Sequences

5.1 Introduction

Image segmentation and motion estimation are fundamental tasks in medical image processing and play essential roles in downstream tasks, including motion modelling [168, 170], medical quantitative medical imaging [169], and population-based functional association studies [230] in many different modalities. As magnetic resonance imaging (MRI) is increasingly deployed in various clinical situations to provide a wealth of qualitative and quantitative information, Cine MRI, in particular, gives spatio-temporal scans over time, which has great potential in clinical assessments of cardiac morphology and function[146, 147].

Deep learning (DL) based segmentation and motion estimation in MRI analysis have been explored extensively in recent years [18, 20, 166]. However, the mono-segmentation approaches usually can not incorporate all data as providing training ground truth is labour-intensive. Besides, adopting additional training data from offline registration methods might involve further noise disturbing the training. On the other hand, existing motion estimation approaches are mainly constrained by appearance similarity, resulting in transformations lacking knowledge of anatomical structures in the images.

Recently, there have been attempts to perform concurrent segmentation and motion estimation in a deterministic way [231–233]. Although they improve the performance in both tasks, despite the staged training process, they struggle with capturing the inherent uncertainty in the sequential medical data and showing the model confidence in their predictions.

We propose a novel recurrent probabilistic model, SegMorph, conducting concurrent segmentation and motion estimation on the cine-MRI sequence. The proposed framework extracts shared multitask features from the temporal inputs and model them as low-dimensional Multivariate Gaussians (MVGs) in the latent space. The semantic entities and motion are inferred by parallel decoding the embedding sampled from the latent space. By probabilistic modelling, our framework supports inference uncertainty assessments in both tasks. The variety in the latent space also allows our model to generate sequences and corresponding segmentation masks, given the moving image and a starting frame.

We evaluate the proposed method through quantitative and qualitative experiments on segmentation and motion estimation tasks. The proposed method shows superior results over the baseline and state-of-the-art approaches. Through comprehensive ex-

periments, we demonstrate that multi-task formulation benefits both tasks in producing anatomically-informed free-form DVFs, and temporally robust anatomical masks while achieving state-of-the-art performance. Besides, SegMorph is a general model towards 'nD+t' data, showing potential in sequential motion estimation and anatomical structure extraction. The generative nature of the probabilistic model also allows our model to provide data augmentation by generating not only the spatio-temporal images, but also the corresponding segmentation masks and DVFs. Our contributions are:

- We introduce a recurrent multi-task model that uses spatio-temporal information from cine-MRI sequences to achieve temporally robust segmentation and anatomically aware motion estimation.
- We explore a probabilistic formulation of concurrent segmentation and motion estimation on cine-MRI sequence, enabling probabilistic assessments in both tasks.
- We introduce a scheme that enables end-to-end multi-task training and establishes a mutual benefit between high-level representations and motion estimation.
- Through extensive experiments on the cardiac MR UK biobank dataset, we showed that our method improves temporal robustness compared to pure segmentation methods and outperforms pure motion estimation methods with smoother and better temporally-consistent DVFs.

5.2 Related Works

5.2.1 Towards anatomically-informed cardiac image segmentation

Correct segmentation of CMR images is essential for downstream cardiac analysis tasks such as 3D shape reconstruction, cardiac index calculation, etc. The fast inference speed and adaptive feature extraction of DL-based methods have drawn significant attention in medical image segmentation, the conventional approaches[167, 234] rely on feature engineering. DL-based methods usually use convolutional neural networks (CNNs) to extract features from the image and then predict pixel-level categories. Early approaches exploiting fully convolutional neural networks (FCNs) to segment ventricles [148, 235] surpass conventional approaches in accuracy and running efficiency. This group of approaches is further refined regarding network architecture and framework

design. U-Net and its variations [9, 25] reform FCNs to an auto-encoder (AE) architecture and use skip connections between encoder and decoder to broadcast multi-scale features in mask estimation. However, the performance of these supervised approaches highly depends on the amount and the quality of annotated training data. In CMR sequence segmentation, the limited number of annotations (usually only two frames at end-diastolic (ED) and end-systolic (ES)) adds a burden to training a robust model that handles spatio-temporal consistency in CMR sequences.

Recent studies have made great efforts towards anatomically plausible and robust segmentation in many challenging CMR scenarios. [162, 163] construct a diverse latent space from contextual information from adjacent slices, multi-view images, and offline shape priors, thus improving segmentation performance on challenging slices. As for sequential data, Yan *et al.* [164] aggregate optical flow into U-Net architecture to improve the temporal coherence of LV segmentation. Du *et al.* [165] propose a CMR sequence segmentation framework combining convolutional Long short-term memory model (ConvLSTM) [102] and U-Net. Bai *et al.* [166] sequentially segment with sparse temporal annotations. They construct pseudo-labels by deforming the masks from the annotation-available frames via the estimated motions to supervise frames without ground truth. However, fully supervised methods [164, 165] require annotations at each time step, which can be labour-intensive. Furthermore, [166] requires an additional offline registration model to generate motion fields to propagate masks in sequence as training supervision.

5.2.2 Deformable motion estimation

Medical image motion estimation aims to find an optimal spatial transformation aligning the anatomical structure in the moving image to those in the fixed image. Various methods have been proposed for registering deformable medical images. Conventional diffeomorphic approaches, including LCC-Demons [16], LDDMM [17], and SyN [173] have shown substantial performance in medical image motion estimation and are widely applied. However, conventional approaches are time-consuming as they often require run-time optimisation, and can be sensitive to the choice of parameters among cases.

In recent years, more attempts have exploited DL-based methods in deformable motion estimation tasks. However, the ground truth for training the deep neural network, i.e., the exact deformations in image registration problems, is often unavailable.

Supervised approaches [180, 181] and data augmentation-based methods [182] exploit results from traditional approaches as training supervision. These methods often suffer from noisy labels or flawed ground truth from conventional methods. Unsupervised and semi-supervised methods have been introduced to eliminate the reliance on ground truth. Balakrishnan *et al.* [216] present an AE-based unsupervised model on pair-wise 3D brain MRI volumes. The authors use the grid sampler in Spatial Transform Network (STN) [226] to warp the moving volume given the predicted deformation field. A similarity loss function is introduced to compute the distance between the moving volumes and the fixed ones, thus circumventing the need for ground truth. In their later work [28], the authors extend the method with semi-supervision from available segmentation masks. They also add an extra Dice loss between the moved and fixed masks to the framework to provide supervision.

Beyond deterministic models, probabilistic models have also shown strength in medical image registration. Dalca *et al.* [13, 18] propose a probabilistic framework that achieves competitive registration results and offers uncertainty estimation on DVFs. They model the latent velocity field as a Multivariate Gaussians (MVGs) distribution and regularize it with standard normal MVGs prior. Krebs *et al.* [20, 199] propose an unsupervised cine-MRI registration model based on the conditional variational autoencoder (CVAE)[200]. The authors model the spatial transformation parameterised by a low-dimensional latent vector. The decoder predicts the displacement field based on latent space samples and the encoderâ€™s multi-scale design.

Beyond pairwise registration, some approaches use the temporal information in the sequential data to improve the effectiveness and temporal consistency in motion estimation. Krebs *et al.* [20] present a spatio-temporal motion estimation on cardiac sequences. The authors add a temporal convolutional network (TCN) [201] in the latent space to capture the temporal information through the CMR sequence. Zakeri *et al.* [202] proposed a probabilistic variational recurrent neural network (VRNN) [122] framework for sequential registration on long-axis CMR sequences. The authors adopt a learnable prior on the latent variable and ConvLSTM [102] to capture the spatio-temporal features through the sequence.

However, these appearance-similarity-oriented approaches usually suffer from capturing the anatomical motion. Emphasis on matching pixel-level intensity similarity between registration pairs results in ignoring motion at anatomical boundaries and

preserving anatomical shape when applying DVFs to categorical-level representations.

5.2.3 Concurrent segmentation and registration

The general intention of combining both tasks is to improve the analysis of specific anatomical structures. Segmentation and motion estimation are fundamental tasks in medical image processing and are closely related. It is common to see approaches utilise available segmentation ground truth to help with motion estimation [28] and vice versa [166].

Before the advent of DL-based methods, concurrent segmentation and motion estimation were often performed based on active contour [236], Bayesian [237], Markov random fields (MRFs) [238], etc. These non-learning-based approaches require online optimisation on each test case, which can be computationally costly.

Xu *et al.* [232] and Estienne *et al.* [233] propose weakly supervised joint registration and segmentation models on paired brain MRI scans to improve the accuracy of both tasks. Both works bridged segmentation and registration with an anatomy similarity loss between the target and warped masks warped by STN. In [232], the segmentation and motion estimation are two parallel branches, each with its own hidden space, while in [233], two branches share the encoder block. They have shown that two tasks can boost each other in coherent training. However, in both works, the loss term shared by the two tasks will have an impact on the network optimisation only if the ground truth for either task is feasible. Qin *et al.* [231] proposed a framework for joint registration and segmentation on the CMR sequence, consisting of an unsupervised registration branch and a semi-supervised segmentation branch. They also find that the segmentation branch improves the registration results, especially around the boundaries. However, the temporal information captured by the recurrent unit is only leveraged in the registration branch rather than directly shared with both tasks. Furthermore, all the methods mentioned above are deterministic, which cannot provide uncertainty estimation on their predictions.

5.3 Methodology

This work aims to achieve concurrent segmentation and motion estimation on CMR sequences. Given a CMR sequence $\{\boldsymbol{I}_0, ..., \boldsymbol{I}_\mathrm{T}\}$ of $T+1$ time points and its segmentation masks on the ED and ES, $\{\boldsymbol{m}_\mathrm{ED}, \boldsymbol{m}_\mathrm{ES}\}$, the concurrent segmentation and motion estimation model estimates the motion sequence $\{\boldsymbol{d}_1, ..., \boldsymbol{d}_\mathrm{T}\} \in \mathbb{R}^{2\times H\times W}$ described as the DVFs from the reference frame $\boldsymbol{I}_0$ to the fixed images $\{\boldsymbol{I}_1, ..., \boldsymbol{I}_\mathrm{T}\}$, and the mask predictions of C anatomical labels in the input sequence $\{\boldsymbol{m}'_0, ..., \boldsymbol{m}'_\mathrm{T}\} \in \mathbb{R}^{C\times H\times W}$. Following a fundamental design of recurrent variational autoencoder (RVAE), we form a low-dimensional latent space $\{\boldsymbol{z}_1, ..., \boldsymbol{z}_\mathrm{T}\} \in \mathbb{R}^L$ conditioned on spatial-temporal features $\{\boldsymbol{h}_1, ..., \boldsymbol{h}_\mathrm{T}\}$ that is updated during the sequential processing. The decoder comprises two branches to achieve multi-task learning, where one predicts the anatomical masks and the other conducts motion estimation via predicting the DVFs between the moving and fixed frames. To establish the mutual benefits between the two tasks, we exploit the predicted DVFs by warping the corresponding mask of the moving image (the moving mask), which serves as an additional supervision signal for segmenting those unlabelled intermediate frames between ED and ES. Concurrently, the registration branch produces more plausible deformations by utilising the shared recurrent latent space that stores anatomical information and temporal contexture features. The network architecture is illustrated in Fig.5.1.

Below we introduce our framework, starting with its overall formulation and then explaining each part. Next, we introduce the variational constraints used in the training process, followed by an explanation of the network architecture.

5.3.1 Learn a recurrent temporal conditioned latent space for multi-tasking

The proposed model is developed based on a probabilistic motion estimation model, DragNet [202]. Beyond the probabilistic motion modelling in DragNet, we exploit multi-task learning and feature fusion. The proposed model consists of three major parts: an encoder to approximate the posterior $q_\theta(\boldsymbol{z}_{1:T}|\boldsymbol{I}_{1:T}, \boldsymbol{h}_{1:T-1})$, a recurrent unit to compute state variable $p_\tau(\boldsymbol{h}_{1:T}|\boldsymbol{I}_{1:T-1}, \boldsymbol{z}_{1:T-1}, \boldsymbol{h}_0)$, and a multi-branch decoder for computing the conditional distribution of $p_\omega\left(\boldsymbol{m}_{1:T}, \boldsymbol{I}_{1:T}, \boldsymbol{d}_{1:T}|\boldsymbol{z}_{1:T}, \boldsymbol{h}_{1:T-1}, \boldsymbol{I}_0\right)$, parameterzed by θ, τ and ω, respectively.

In forward propagation, the encoder first processes the input sequence to produce

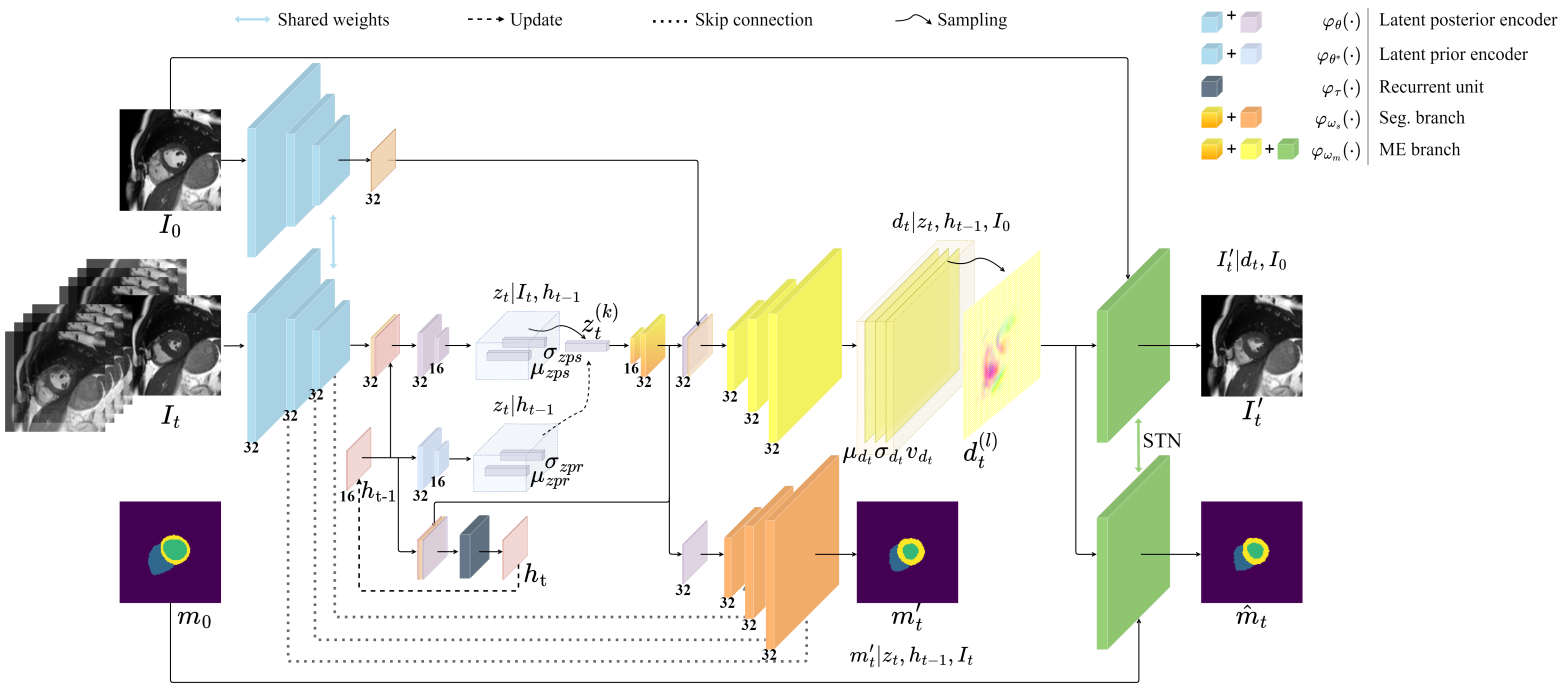


Figure 5.1: Overview of the proposed model, SegMorph, inherits an encoder-decoder architecture. From sequential image inputs, the encoder extracts features from the fixed sequence and blends them with the temporal information preserved in the hidden state vectors $\boldsymbol{h_t}$ to form the latent space. The sampled latent vector $\boldsymbol{z_t}$ is fed to the dual-branch decoder for concurrent segmentation and motion tasks, and the recurrent block to update the hidden state. The diagram is better viewed in colour.

high-level semantic features. Those features are then aggregated via concatenation with the spatio-temporal feature from the recurrent unit to describe a comprehensive latent space for multi-task learning. On the decoder ends, two parallel branches are designed to handle sequential segmentation and motion estimation tasks, respectively. The segmentation branch projects the latent vector to the categorical spaces. In contrast, the motion estimation branch maps the latent vector to form DVFs, which are used to reconstruct the fixed frames via pixel-wise warping.

Spatio-temporally conditioned latent space

In SegMorph, we model the multi-task feature $\boldsymbol{z}$ in the latent space as MVGs parameterised by $[\boldsymbol{\mu}, \boldsymbol{\sigma}]$. The latent space shall describe a balanced blend of features from the current input frame and the spatio-temporal information from earlier frames for both tasks. For sequential data, extracting spatio-temporal information directly from the high-dimensional inputs involves too much noise and is inefficient. A better way is to extract the spatio-temporal feature in the latent space with reduced dimensionality and high-level features. We adopt a recurrent unit in the latent space to preserve spatio-temporal features from sequential inputs. The recurrence of the recurrent unit at each time-point t is formulated as $\boldsymbol{h}_t = \varphi_\tau\left(\boldsymbol{I}_t, \boldsymbol{z}_t, \boldsymbol{h}_{t-1}\right)$, where φ_τ is the RNN model parameterised by τ. The spatio-temporal conditioned latent space is described by a Gaussian distribution $p_{\theta^*}\left(\boldsymbol{z}_t|\boldsymbol{h}_{t-1}\right)$ with the prior inference network predicting its mean and variance $[\boldsymbol{\mu}_{\text{zpr}}, \boldsymbol{\sigma}_{\text{zpr}}] = \varphi_{\theta*}(\boldsymbol{h}_{t-1})$ and the approximate MVGs posterior $q_\theta(\boldsymbol{z}_t|\boldsymbol{I}_t, \boldsymbol{h}_{t-1})$ through the inference encoder computing the corresponding parameters $[\boldsymbol{\mu}_{\text{zps}}, \boldsymbol{\sigma}_{\text{zps}}] = \varphi_\theta(\boldsymbol{I}_t, \boldsymbol{h}_{t-1})$. By building such a recurrent learning scheme, our model can learn more discriminative spatio-temporal representations, enabling better simultaneous cardiac segmentation and motion estimation. Specifically, we use the ConvLSTM [102] as the RNN model's architecture because of its superiority in leveraging spatio-temporal information.

The variational decoder for multi-tasking

On top of spatio-temporally conditioned latent space, we conduct concurrent probabilistic segmentation and motion estimation via a multi-branch decoder. In the segmentation branch, the categorical masks $\boldsymbol{m}$ are modelled as multinoulli distributions: $\prod_{\text{i}=1}^{C} \boldsymbol{m}'_{\text{t,i}}{}^{\boldsymbol{m}_{\text{t,i}}}$, where $\boldsymbol{m}'_{\text{t,i}}$ is a C-channel matrix and each pixel-wise classification vec-

tor is a one-hot vector approximated through Softmax layer. The motion estimation branch follows the design in [202]. The decoder first predicts the displacement field characterised by $[\boldsymbol{\mu}_d, \boldsymbol{\Sigma}_d]$, the mean and covariance of pexel-wise distributions, from the sampled latent vector. Then explicit $\boldsymbol{d}_t$ is sampled to deform the moving image via STN grid sampler [226]. The parameterisation of the multi-branch decoder results in and was motivated by the factorisation of the decoding process w.r.t. time t as:

$$\begin{aligned} &p_\omega\left(\boldsymbol{m}_t, \boldsymbol{I}_t, \boldsymbol{d}_t | \boldsymbol{I}_0, \boldsymbol{z}_t, \boldsymbol{h}_{t-1}\right) \\ &= p_{\omega_s}\left(\boldsymbol{m}_t | \boldsymbol{I}_t, \boldsymbol{z}_t, \boldsymbol{h}_{t-1}\right) p_{\omega_m}\left(\boldsymbol{I}_t, \boldsymbol{d}_t | \boldsymbol{z}_t, \boldsymbol{h}_{t-1}, \boldsymbol{I}_0\right) \end{aligned} \tag{5.1}$$

where two branches are parameterised by ω_s and ω_m and presented as $\varphi_{\omega_s}(\cdot)$ and $\varphi_{\omega_m}(\cdot)$ in Fig.5.1.

Generation phase

As a variational model, given the starting frame $\boldsymbol{I}_0$ and fist state variable $\boldsymbol{h}_0$, SegMorph can generate new sequences with corresponding anatomical masks by sampling $\boldsymbol{z_t}$ from the spatio-temporally conditioned prior p_{θ^*}. The generation phase can be factorised as:

$$\begin{aligned} &p_\omega\left(\boldsymbol{m}_{1:T}, \boldsymbol{I}_{1:T}, \boldsymbol{d}_{1:T} | \boldsymbol{I}_0, \boldsymbol{h}_0\right) \\ &= \prod_{t=1}^{T} \int_{\boldsymbol{z}_t} p_\omega\left(\boldsymbol{m}_t, \boldsymbol{I}_t, \boldsymbol{d}_t | \boldsymbol{I}_0, \boldsymbol{z}_t, \boldsymbol{h}_{t-1}\right) p_{\theta^*}\left(\boldsymbol{z}_t | \boldsymbol{h}_{t-1}\right) d\boldsymbol{z}_t \end{aligned} \tag{5.2}$$

The testing phase of SegMorph operates in two modes: direct concurrent segmentation and motion modelling (inference mode), and generative mode. As shown in Fig.5.2, in the inference mode, the model makes predictions based on the encoding of the input CMR sequence. Thus the segmentation branch runs parallel to the motion branch. While in the generation mode, the segmentation branch runs after the motion branch, as the generated frame needs to go through the feature extractor to update the hierarchical feature maps and pass the feature to the segmentation branch via skip connection.

5.3.2 Variational constraints towards mutual-beneficial multi-task training

Like other variational models, we optimize the performance by optimising the evidence lower bound (ELBO) using stochastic gradient methods. Specifically, for each input

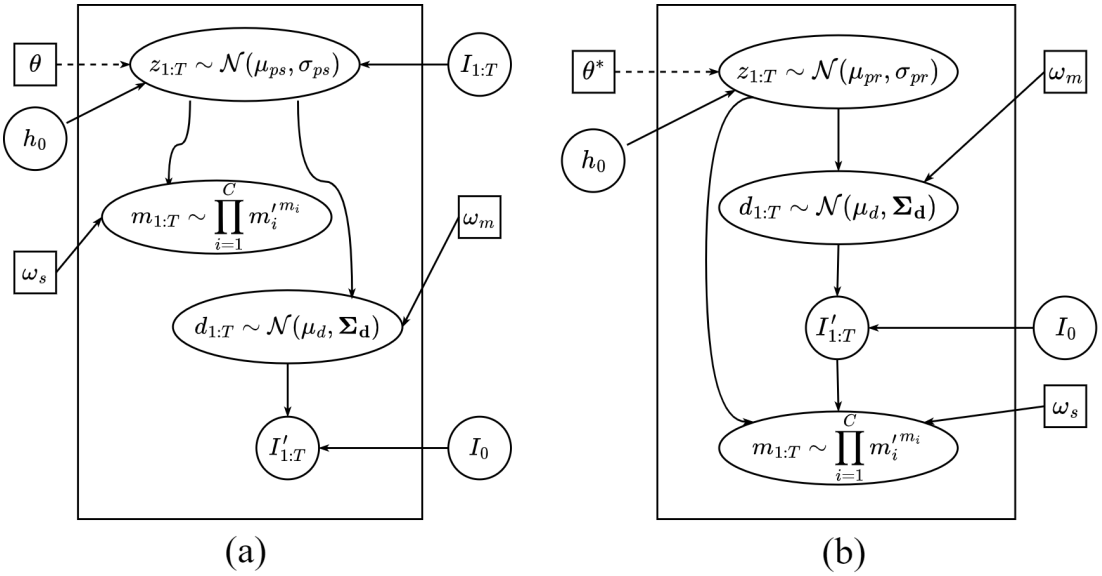


Figure 5.2: Illustration of two modes of the proposed framework during testing. (a) In the inference mode, the latent variable blends temporal information and the new input frame. The registration branch infers the DVFs given the sample from latent space and the reference (moving) image. Then STN applies DVFs on the moving image to obtain the warped image. Similarly, the segmentation masks are obtained through the latent sample and multi-scale features passing through the skip connection. (b) In the generation mode, a latent prior is formed based on the temporal knowledge from the hidden variable. Then, multiple samples from latent space generate the new moving image and corresponding semantic masks.

batch $\{\boldsymbol{I}_{1:T}, \boldsymbol{m}_{\text{ED, ES}}\}$, the optimisation causes the following loss:

$$\begin{aligned} &\mathcal{L}(\theta^*, \theta, \omega, \tau; \boldsymbol{I}, \boldsymbol{m}) = \\ &- \mathbb{E}_{\shortparallel_\theta} \left[\log p_\omega \left(\boldsymbol{m}, \boldsymbol{I}, \boldsymbol{d} | \boldsymbol{z}, \boldsymbol{h}, \boldsymbol{I}_0\right)\right] + \mathrm{KL}\left[q_\theta(\boldsymbol{z}|\boldsymbol{I}, \boldsymbol{h}) \| p_{\theta^*}(\boldsymbol{z}|\boldsymbol{h})\right] \end{aligned} \tag{5.3}$$

where the first term measures the reconstruction similarity and the second term encourages the similarity in spatio-temporal prior and approximated posterior. The time footnotes of variables are omitted for brevity. For a sequential input with T frames, the variational loss over the latent space is *Kullback-Leibler* Divergence (KLD) summed up over time as:

$$\mathcal{L}_{\boldsymbol{z}} = \sum_{t=1}^{T} \mathrm{KL}\left[q_\theta\left(\boldsymbol{z}_{\mathrm{t}} | \boldsymbol{h}_{t-1}, \boldsymbol{I}_t\right) \| p_{\theta^*}\left(\boldsymbol{z}_{\mathrm{t}} | \boldsymbol{h}_{\mathrm{t\text{-}1}}\right)\right] \tag{5.4}$$

where KL$(\cdot)$ denotes the KLD between two MVGs, which is calculated:

$$\begin{aligned}\mathcal{D}_{\mathrm{KL}}[q\|p] =& \frac{1}{2}\left[\log\frac{|\boldsymbol{\Sigma}_p|}{|\boldsymbol{\Sigma}_q|} - k + tr\{\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\Sigma}_q\}\right. \\ &\left. +(\boldsymbol{\mu}_p-\boldsymbol{\mu}_q)^T\boldsymbol{\Sigma}_p^{-1}(\boldsymbol{\mu}_p-\boldsymbol{\mu}_q)\right]\end{aligned} \tag{5.5}$$

where q and p stand for posterior and prior distribution, respectively. $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are corresponding mean and covariance matrix. k is the dimension of vector space.

As for the reconstruction loss, referring to the factorization for decoding in Eq.5.1, we split the first term in Eq.5.3 into two parts for each task. The part for the segmentation branch compares the similarity between the predicted masks and the reference masks. To bypass the need for annotated ground-truth sequences, we semi-supervise the segmentation process using the ground-truth at ES $\boldsymbol{m}_{\mathrm{ES}}$ and the warped masks $\hat{\boldsymbol{m}}_{1:T,t\neq t_{\mathrm{ES}}}$ elsewhere ground-truth is unavailable. We obtain the warped mask by warping the moving mask $\boldsymbol{m}_{\mathrm{ED}}$ w.r.t. the deformation field $\boldsymbol{d}_t$ via STN as $\hat{\boldsymbol{m}}_t = \boldsymbol{m}_{\mathrm{ED}} \circ \boldsymbol{d}_t$. Thus, given the predicted masks sequence $\boldsymbol{m}'_{1:T}$ and its semi-supervision $\boldsymbol{m} = \{\boldsymbol{m}_{\mathrm{ES}}, \hat{\boldsymbol{m}}_{1:T,t\neq t_{\mathrm{ES}}}\}$, the loss term for segmentation branch is:

$$\mathcal{L}_{\mathrm{msk}} = \sum_{t=1}^{T}\sum_{i=1}^{K} \boldsymbol{H}(\boldsymbol{m}_t, \boldsymbol{m}'_{t,\boldsymbol{z}_t^{(i)}})/K + \boldsymbol{H}(\boldsymbol{m}_{\mathrm{ES}}, \hat{\boldsymbol{m}}_{t=t_{\mathrm{ES}}}) \tag{5.6}$$

where $\boldsymbol{H}(\cdot)$ denotes the cross entropy (CE) function, and K denotes the number of samples drawn from the latent distribution. The first term in $\mathcal{L}_{\mathrm{msk}}$ informs the segmentation branch with the motion estimation branch. The second term uses anatomical structure information to help the motion estimation branch achieve anatomically plausible motion estimation.

In the motion estimation branch, we assign a standard MVG prior to the transformation parameter distributions for regularisation purposes [239]. Thus, the negative log-likelihood of the motion estimation branch is split into two, $\mathcal{L}_{\mathrm{sim}}$ encourages the similarity between the fixed image and the warped image and $\mathcal{L}_{\boldsymbol{d}}$ encourages a smooth and sparse deformation. Sampling K latent vectors and L DVFs from their distribution, we obtain the loss functions as follows:

$$\mathcal{L}_{\mathrm{sim}} = \sum_{t=1}^{T}\sum_{i=1}^{K}\sum_{j=1}^{L}\left\|\boldsymbol{I}_t - \boldsymbol{I}_0 \circ \boldsymbol{d}^{(j)}_{t,\boldsymbol{z}_t^{(i)}}\right\|^2 /2\sigma_{\boldsymbol{I}}^2 KL + C \tag{5.7}$$

where $\circ$ denotes the grid sampler, $\sigma_{\boldsymbol{I}}$ is the fixed standard deviation assigned to the warped images with a value equal to 1, C is the constant values in the equation. The

regularisation term over the deformation field is calculated as:

$$\mathcal{L}_{\boldsymbol{d}} = \sum_{t=1}^{T}\sum_{i=1}^{K} \mathrm{KL}\left[q_{\omega_m}\left(\boldsymbol{d}_{\mathrm{t}}|\boldsymbol{z}_{\mathrm{t}}^{(i)}, \boldsymbol{I}_0, \boldsymbol{h}_{t-1}\right) \| \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{\boldsymbol{2\times 2}})\right]/K \tag{5.8}$$

Thus, we form the overall constraints for the optimisation with the four terms mentioned above in Eq.5.3 as:

$$\mathcal{L}(\theta^*, \theta, \omega, \tau; \boldsymbol{I}, \boldsymbol{I_0}, \boldsymbol{m}) = \mathcal{L}_{\mathrm{msk}} + \mathcal{L}_{\mathrm{sim}} + \mathcal{L}_{\boldsymbol{d}} + \mathcal{L}_{\boldsymbol{z}} \tag{5.9}$$

5.3.3 Network architecture

Referring to Fig.5.1, we select the ED frame from CMR sequence as the moving image $\boldsymbol{I}_0$ and the rest of the sequence as the fixed images $\{\boldsymbol{I}_1, ..., \boldsymbol{I}_{\mathrm{T}}\}$. The proposed framework follows an encoder-decoder architecture with a recurrent latent space. It takes the moving and fixed images as inputs and the moved frames and corresponding segmentation masks are the output. The network first extracts features from each frame via weight-sharing encoders. The number of feature channels in each convolution block of the encoder part is 32, increasing the depth of the network as their size shrinks. The convolution block consists of two convolution layers with a kernel size of 3. The first one is followed by a LeakyReLu layer as an activation function, and the other convolutional layer downsamples the features with stride 2.

As described in Sec. 5.3.1, the posterior latent distribution is inferred by recurrently forwarding the aggregation of the features of the fixed frame $\boldsymbol{F}_{\mathrm{fix}}$ and the hidden state $\boldsymbol{h_{t-1}}$ into the latent encoder. On the other hand, the temporal conditioned prior is inferred by only inputting the hidden state $\boldsymbol{h_{t-1}}$ to the latent prior encoder. The aforementioned latent encoders are lightweight and have the same architecture as the convolution block in the encoder. In the latent space, the latent vector $\boldsymbol{z}_t$ is sampled from the latent distribution using reparameterisation trick [76]. The latent vector sampled from the latent distribution is up-sampled via transpose convolutional layers with a kernel size of 4 and stride of 2 to form the latent feature map $F_{\boldsymbol{F}_{\boldsymbol{z}_t}}$. This latent feature map is processed through ConvLSTM to learn the spatio-temporal representations, which are leveraged to update the hidden variable.

The decoder comprises two branches for simultaneous segmentation and motion estimation. The motion estimation branch takes the feature of the moving image $\boldsymbol{F}_0$ and the latent feature $\boldsymbol{F}_{\boldsymbol{z}_t}$ as inputs to predict the distribution of the deformation field

($\boldsymbol{\mu}_{\boldsymbol{d}_t}, \boldsymbol{\Sigma}_{\boldsymbol{d}_t}$). This branch is constructed by 3 convolution blocks, each of which has a transpose convolution layer with a kernel size of 4 and stride of 2 for upsampling, a LeakyReLu layer for non-linearity, and a convolution layer with a kernel size of 3 and stride of 2. A Gaussian smoothing kernel is applied on $\boldsymbol{\mu}_{\boldsymbol{d}_t}$ to maintain local smoothness. Then the grid sampler from STN warps the moving frame $\boldsymbol{I}_0$ and its mask $\boldsymbol{m}_0$ given $\boldsymbol{\mu}_{\boldsymbol{d}_t}$ to produce the warped image $\boldsymbol{I}'_t$ and the warped mask $\hat{\boldsymbol{m}}_t$. The segmentation branch shares the same network architecture as the motion estimation branch but only takes the latent feature $\boldsymbol{F}_{\boldsymbol{z}_t}$ as input. In contrast to the motion estimation branch, skip connections between the segmentation branch and encoder enrich the representations with multi-scale features. We conduct a Softmax layer at the end of the decoder to approximate the categorical distribution on mask prediction $\boldsymbol{m}'_t$.

5.4 Experiments and Results

This section evaluates our registration and segmentation framework on short axis (SAX) cine-CMR sequences. We conducted qualitative and quantitative comparison experiments with various methods to show our method's superiority in sequential segmentation and motion estimation. The end-diastolic (ED) frame was used in all experiments as the moving image $\boldsymbol{I}_0$.

We built our model using Python and Pytorch. The network is trained in an End-to-End manner. We used Adam optimiser [228] with a batch size of 16 and a fixed learning rate of 0.001. The training time is ∼8h on an NVIDIA GTX TITAN X GPU.

5.4.1 Data and annotations

The proposed framework is trained and tested on a partial dataset acquired from UK Biobank (UKBB) dataset [240]. The dataset consists of 3600 short-axis (SAX) cine-CMR sequences. For data acquisition, a complete SAX stack of balanced steady-state free procession (bSSFP), including the left ventricle (LV) and the right ventricle (RV) acquired through a cardiac cycle. For each case, 50 time-point of SAX scan stacks are acquired covering a full cardiac cycle. Each SAX scan slice is captured with a 208 $\times$ 187 matrix size and an in-plane resolution of 1.8 mm $\times$ 1.8 mm. The ground-truth contour notations are available at the ED and ES phases, including the left ventricular myocardium (LVmyo), left ventricular endocardium (LVendo), and right ventricular endocardium (RVendo).

In this study, we used 2700 cine-CMR sequences for training, 300 sequences for validation, and 600 sequences for testing. Each sequence comprises a 10-slice SAX stack of cine-CMR image sequences sampling at 50 frames an average cardiac cyle. A SAX cine-CMR stack consists of 6-slice scans avoiding the base and apical slices. In pre-processing, all the input images are centrally cropped with the size of 128×128 in pixels, and the image intensity is normalised and re-scaled to $[0, 1]$. Assuming a periodic cardiac motion, we arrange each CMR sequence starting with ED frame followed by 4 volumes from ED to ES frame. Then the ES volume locates in the middle of the sequence, followed by the rest 4 volumes from ES to ED.

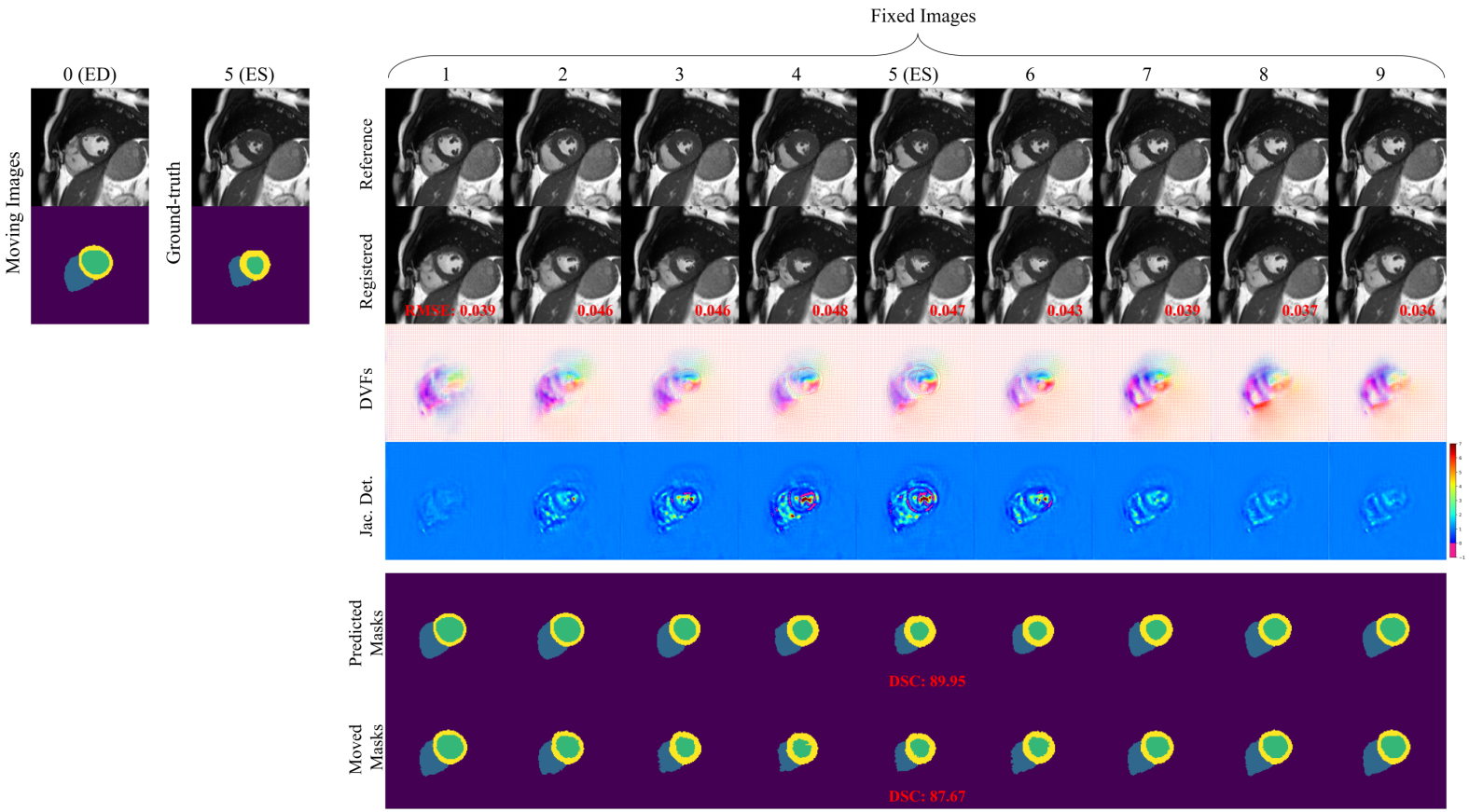


Figure 5.3: Concurrent segmentation and motion estimation on a SAX cine-CMR sequence. The moving image $\boldsymbol{I}_0$ and its mask $\boldsymbol{m}_0$ are shown at the top left. On the RHS, from the top, we listed the fixed images, the warped images, the corresponding DVFs, and Jacobian determinant maps, respectively. In the last two rows, we show the predicted masks from the segmentation branch and the warped mask using predicted DVFs. The registered images show a good agreement compared with reference regarding reported RMSE and SSIM. The RMSE is reported in red at the bottom right of the registrated images. The corresponding DVFs indicate most of the motion is concentrated on the edge of the ventricles, and the mostly positive Jacobian Determinant maps support the diffeomorphic property of DVFs. The predicted masks and warped masks show good consistency throughout the sequence. This indicates a precise motion estimation on the anatomical edge from the proposed model. Furthermore, the predicted masks are smoother than the moved ones, especially on the contour area of each anatomical part. The CMR images are reproduced with permission of UK Biobank©.

Table 5.1: Quantitative comparison on segmentation performance among SegMorph, U-Net, and JMS. Metrics include Dice Similarity Coefficient (DSC), and 95%-tile Hausdorff Distance (HD95) on different anatomical structures. All metrics are computed over all test subjects. Bold values indicate significant differences in the comparison ($p \ll 0.001$).

Method	ED						ES					
	RV		LVendo		LVmyo		RV		LVendo		LVmyo	
	DSC(%)	HD95(mm)	DSC(%)	HD95(mm)	DSC(%)	HD95(mm)	DSC(%)	HD95(mm)	DSC(%)	HD95(mm)	DSC(%)	HD95(mm)
U-Net [9]	74.45(22.80)	8.48(12.87)	86.40(11.96)	3.88(6.97)	65.52(19.04)	4.23(7.31)	74.89(21.32)	6.68(10.25)	84.12(17.25)	4.00(10.61)	73.32(23.70)	6.13(15.40)
JMS [231]	**88.02(12.11)**	4.84(6.45)	**93.92(8.03)**	2.77(6.49)	**74.43(20.22)**	3.23(5.94)	71.28(21.03)	11.19(12.80)	83.83(12.71)	5.49(10.13)	67.81(21.16)	5.70(7.80)
Ours (seg)	85.52(13.53)	**3.92(2.55)**	92.02(10.16)	**2.00(1.87)**	71.18(18.35)	**2.67(2.32)**	**76.09(19.36)**	**5.96(2.91)**	**86.15(12.28)**	**3.30(2.50)**	**76.94(15.46)**	**4.10(3.03)**

Table 5.2: Quantitative comparison of segmentation performance in terms of clinical indices. The medical indices include the volume of the LV and RV at ED and ES (LVEDV, LVESV, RVEDV, RVESV, in mL), the left ventricular muscle mass (LVMM, in g), the stroke volume of LV and RV (LVSV, RVSV, in mL) and the ejection fraction of LV and RV (LVEF, RVEF in percentage). Indices show no significant difference (p-value > 0.05) from the reference are highlighted in bold.

Method	LVEDV(mL)	LVESV(mL)	LVSV(mL)	LVEF(%)	LVMM(g)	RVEDV(mL)	RVESV(mL)	RVSV(mL)	RVEF(%)
GT	109.31(20.07)	57.71(14.28)	51.60(10.88)	47.37(6.44)	58.06(12.08)	118.19(24.39)	70.52(18.72)	47.66(16.45)	40.14(10.96)
U-Net	87.44(17.11)	52.39(13.36)	35.04(8.90)	40.21(7.62)	75.17(12.96)	85.17(23.65)	64.24(18.82)	20.93(15.77)	22.23(21.50)
JMS	**107.99(18.91)**	63.04(14.08)	44.94(9.67)	41.76(5.99)	**57.17(10.65)**	112.84(24.33)	97.48(22.25)	15.36(7.89)	13.55(6.65)
Ours	**108.22(18.41)**	**56.54(12.40)**	**51.68(8.70)**	**47.97(4.71)**	63.01(10.98)	113.55(24.95)	**69.15(19.40)**	44.40(15.11)	**39.06(10.14)**

5.4.2 Sequential segmentation

We evaluate the sequential segmentation effectiveness of our model in terms of segmentation accuracy and temporal consistency with the representative baseline method, U-Net[9] and the state-of-the-art approach in joint cardiac segmentation and motion estimation (JMS) [231].

Both compared models are re-trained using the public implementations and adjusted training parameters. U-Net has been trained using all available end-systolic volumes in the training set. JMS was trained using both CMR sequences and the ground-truth mask at ED and ES. To quantitatively evaluate the performance of the proposed approach, we compare the Dice Similarity Coefficient (DSC) [241] and 95%-tile Hausdorff Distance (HD95, in mm) on the three anatomical structures, LVendo, LVmyo, and RVendo. From a clinical practicality viewpoint, we evaluated the clinical indices calculated from the segmentation results, including the volume of the LV and RV at ED and ES (LVEDV, LVESV, RVEDV and RVESV), the stroke volume of LV

and RV (LVSV, RVSV), the left ventricular muscle mass (LVMM), and the ejection fraction of the LV and RV (LVEF and RVEF).

Table 5.1 presents the segmentation results of the three investigated methods on each anatomical structure. Both multi-tasking approaches (JMS and SegMorph) significantly outperform U-Net's supervised approach in all segmentation metrics. We attribute this as the contribution of the motion estimation task to the segmentation task. Our method achieves comparable results compared to JMS at ED frame and outperforms JMS substantially at ES, even though JMS has supervision at both ED and ES frames while our segmentation branch is only supervised at the ES frame.

Our model shows good grasp of temporal variations on both ED and ES frames and throughout the sequence. A complete result of a test sequence is shown in Fig. 5.3. In terms of anatomical masks, the predicted masks are generally more precise and smoother throughout the sequence (comparing the last two rows in Fig. 5.3). Among compared segmentation methods (mask contours comparison illustrated in Fig. 5.4), U-Net tends to under-segment the ED frame and JMS over-segment the ES frame, particularly on the right ventricles.

In Table 5.2, we compared the clinical indices calculated from our predictions with the clinical reference values. We found in our prediction, the majority of the clinical indices (LVEDV, LVESV, LVSV, LVEF, RVESV, RVEF) show no significant differences (p-value > 0.05) compared to the reference values, which is not afforded by other approaches. All three approaches perform less effectively in segmenting the RV, as RV shape varies dramatically in a cardiac cycle. This suggests the segmentation prediction from our model yields better coherence in the clinical scenario and potentially applies to clinical analysis. Notice that all the measurements listed in Table 5.2 are consistently lower than those reported in analogous studies [210, 230]. This discrepancy can be attributed to the truncation applied to the number of slices in each scan.

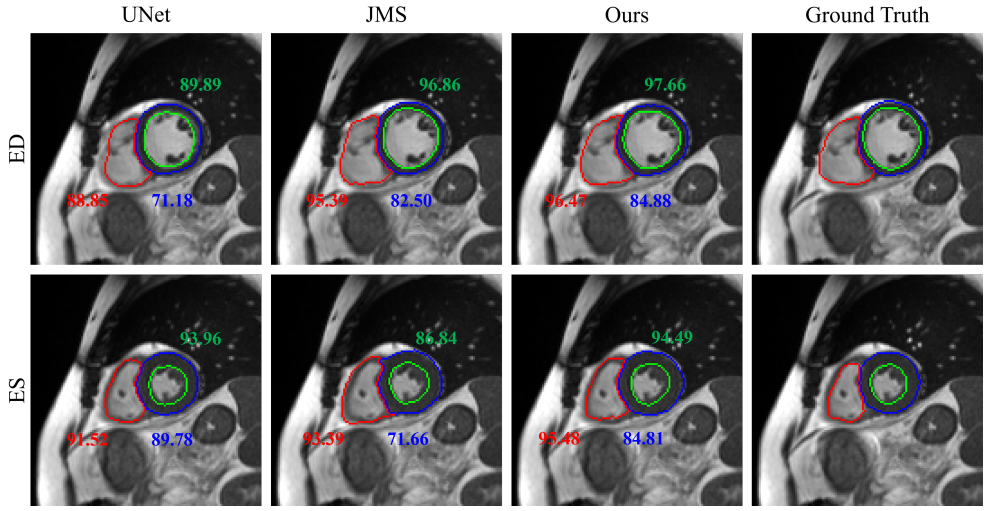


Figure 5.4: Showcases of mask contour comparison. Each column presents the mask contour predictions on ED and ES from each compared method. The DSC of each anatomical structure is reported with corresponding contour colours. As a fully supervised method, U-Net undersegments the ED frame, especially for the RV, as it was only trained with ES frames. Despite only using ES frame ground-truth supervision, our approach achieves comparable segmentation performance compared to JMS, which has supervision at both ES and ED . The CMR images are reproduced with permission of UK Biobank©.

Table 5.3: Quantitative comparison on registration performance between SegMorph and LCC-Demons, ANTs SyN, VoxelMorph, JMS, DragNet. The comparing metrics include average Root Mean Square Error (RMSE), average Structure Similarity (SSIM), average Non-Positive Jacobian Determinant (NJD), Dice Similarity Coefficient (DSC), and 95% Hausdorff Distance (HD95).

Method	RMSE(10^{-2})	SSIM	NJD	DSC-RV (%)	DSC-LVendo(%)	DSC-LVmyo(%)	HD95-RV(mm)	HD95-LVendo(mm)	HD95-LVmyo(mm)
LCC-Demons[16]	4.26(2.16)	0.96(0.03)	43.47(29.76)	62.32(18.75)	63.21(9.42)	46.48(15.30)	10.93(7.89)	8.10(1.88)	6.81(2.62)
SyN[173]	4.49(1.37)	0.95(0.02)	-	68.60(18.29)	67.19(19.22)	**78.48(11.50)**	**2.95(1.05)**	5.67(2.65)	3.70(1.64)
VoxelMorph[13]	**3.18(1.31)**	0.96(0.03)	86.39(106.64)	68.23(19.21)	77.98(13.95)	61.60(19.37)	9.66(8.11)	4.63(2.15)	5.07(2.54)
JMS[231]	4.24(1.70)	0.95(0.03)	67.01(72.23)	69.11(19.11)	83.29(10.51)	52.64(19.04)	5.18(2.51)	**2.84(1.51)**	3.28(1.23)
DragNet	3.91(1.26)	0.94(0.03)	33.88(58.58)	69.87(18.73)	79.78(10.08)	68.96(14.36)	9.14(7.87)	4.75(1.73)	4.72(2.54)
Ours (reg)	4.24(1.48)	0.94(0.03)	**23.72(54.81)**	**75.61(16.82)**	**86.79(9.55)**	76.51(14.97)	5.12(2.62)	2.94(1.40)	**3.06(1.43)**

5.4.3 Registration and motion estimation

To show the efficacy of our method in image registration, we compare our framework with state-of-the-art conventional free-form deformation approaches, Symmetric Normalisation (SyN) in ANTs implementation [173] and LCC-Demons[16]. And learning-based approaches including VoxelMorph[13], DragNet[202], and the joint learning segmentation and motion estimation model, JMS[231].

SyN and LCC-Demons are manually tuned on a few training images to determine the parameter settings. The learning-based methods are re-trained using official implementations with adjusted hyper-parameters to achieve comparable registration performance on the given dataset.

We evaluate the model performance for motion estimation regarding registration accuracy and the diffeomorphic property of DVFs. Root mean square error (RMSE) and Structural Similarity Index Measure (SSIM) [242] are used to evaluate registration accuracy. Additionally, we compared the DSC and HD95 between the warped mask volume and the ground truth at ES. Diffeomorphic image registration is evaluated by analysing the Jacobian Determinant maps of DVFs. The average count of locations with a negative Jacobian determinant (NJD) in the DVF measures its diffeomorphism. The smaller NJD count indicates better diffeomorphism.

Table 5.3 presents the average RMSE, SSIM computed for all test subjects and DSC, HD95 on each anatomical structure at the ES frame of all compared methods. In terms of appearance similarity, there is no significantly superior method under SSIM metric, and VM achieves the best performance under RMSE but with more negative Jacobian determinant pixels. Our method produces displacement fields with the lowest average NJD and significantly outperforms other methods in terms of DSC on RV and LVendo, and HD95 on LVmyo. This indicates that our proposed method produces more anatomically plausible displacement fields.

Fig.5.5 presents the visual comparison results for different techniques compared to the proposed method for a single time point registration from ED to ES, where the largest displacements occur. The first columns list the ground-truth CMR scans at ED and ES with corresponding masks. LCC-Demons achieve acceptable registration results but with a noisy jacobian determinant map and DVFs with less anatomical awareness results in under-warped masks. SyN achieves smooth DVFs and warped mask while losing appearance details on the left ventricle. As for DL-based methods, there are

Figure 5.5: Visual comparison of registration performance between proposed SegMorph and LCC-Demons, ANTs SyN, VoxelMorph, JMS, and DragNet. The **Left half** listed the moving image at the ED frame, the fixed image at ES frames, and corresponding ground-truth masks. The **Right half** lists results from different methods in each column. We list warped images, warped masks, DVFs with grids, and Jacobian Determinant maps in each row. The negative values on Jacobian Determinant maps are stated in pink. The CMR images are reproduced with permission of UK Biobank©.

noticeable motions at the boundaries of the ventricles. VoxelMorph and DragNet are biased on matching the appearance as the supervisions mainly come from appearance loss. Thus, the warped masks have indentation and discontinuity at the ventricular boundaries. JMS involves a segmentation branch to facilitate training. However, due to its training scheme, the segmentation branch had a limited impact on the motion estimation branch. In comparison, our model achieves a good balance between matching the appearance and mask warping with smooth and reasonable boundaries.

5.4.4 Uncertainty assessments on segmentation and motion estimation

We also performed uncertainty assessments for segmentation and motion estimation by computing the entropy of the outputs. For the segmentation branch, we sampled several latent vectors propagating them through the decoder to output the mask predictions and use the mean and variance to compute the uncertainty map. We followed the description in [13] to perform an uncertainty assessment on the deformation field by computing the differential entropy of the DVFs predictions.

Fig.5.6 shows the segmentation and deformation field uncertainty maps for one test input. For segmentation, the uncertainty is relatively higher on the structure boundaries than on the region within them. For motion estimation, the uncertainty primarily focuses on the ventricular area and is lower near the structure boundaries (indicated as a dashed contour) than within the boundaries.

5.5 Discussions and Conclusion

We presented a unified recurrent variational probabilistic model that achieves concurrent segmentation and motion estimation on cine-CMR sequences. The proposed model intends to construct a shared, collaborative low-dimensional latent space for a multi-task model learned from CMR sequences with sparse mask annotations. We introduced losses to let the two branches support each other and achieve mutually beneficial end-to-end training. Through a thorough set of experiments, we showed that the multi-task formulation benefits both tasks in producing anatomically meaningful free-form deformation fields and temporally robust anatomical masks while achieving state-of-the-art performance.

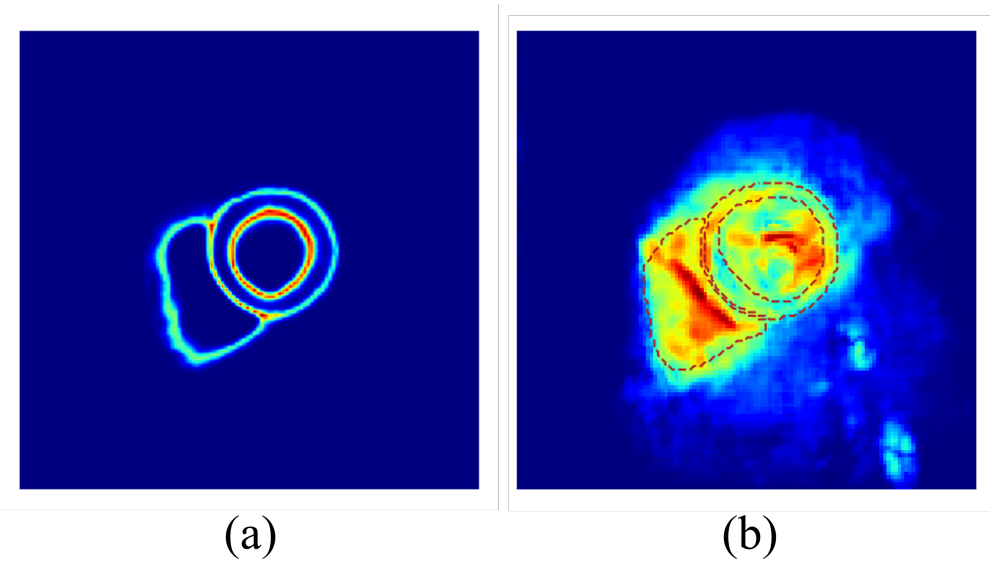


Figure 5.6: Uncertainty assessment results on segmentation and motion estimation. In (a), the higher uncertainty mostly locates at the structure boundaries, which is more challenging for the model to distinguish between classes. Besides, the variety of the latent space also leads to variant outputs when sampling. In (b), the uncertainty of motion estimation primarily focuses on the ventricular areas as they are the main deformed regions, and relatively smaller at the boundaries.

Our model has shown superior segmentation and motion estimation accuracy compared with state-of-the-art algorithms. We improved the anatomical awareness in motion estimation and the temporal consistency in sequential segmentation. In the proposed model, the registration branch can obtain smooth DVFs comparable to the approaches with regularisation terms. This is attributed to the fact that, in part, the segmentation branch introduces a spatial attention mechanism in the shared encoder. The introduction of the warping loss enhances the accuracy of DVFs predictions, especially at the anatomical boundary regions, by using the labels from the segmentation branch. For the segmentation branch, the warped masks from the motion estimation branch provide more reference for the sequential segmentation branch boosting its temporal performance. On the other hand, through an end-to-end training scheme, two branches contribute evenly to form the spatio-temporal latent space. We credit the robustness in segmentation to involving temporal information in the segmentation prediction decoder and the mutual assistance from the motion estimation branch. However, our way of obtaining the 3D motion fields and segmentation is vanilla. In terms of the choice of base network structure, a 3D encoder might encode more spatial features from out-of-plane slices, but it also increases the searching domain, which can create difficulties in training. The 3D encoder also requires a fixed input size in the vertical direction, but the number of slices can vary in different scan volumes. Alternatively, introducing recurrent blocks in the vertical direction could potentially include out-of-plane dependencies without adding extra computational burden.

In future work, we aim to improve the method of including out-of-plane data and explore the optimal sizing of the latent variable based on the length of the input sequence to capture comprehensive and balanced features in multi-task situations.

Chapter 6

cMT-VAE: Content Conditioned Variational Model for Cardiac cine-MRI Motion Transfer

6.1 Introduction

Cardiac motion refers to the spatial and temporal movement of the heart as it contracts and relaxes within a cardiac cycle. Understanding and modelling cardiac motion plays an essential role in the diagnosis of heart pathology [230, 243, 244]. For instance, capturing the motion of heart contraction is the fundamental step of assessing cardiac regional abnormality [245]. Typically, registration between different time-point images in the cardiac cycle is the standard approach to access the motion, which is usually represented as the displacement field containing the spatial movement of each pixel.

Conventional approaches for cardiac motion modelling rely on physical models [16, 173], which are computationally expensive and time-consuming, making them less practical. With the emergence of large-scale datasets, recent deep learning-based methods earn credits in terms of accuracy and running speed [28, 213], among which the variational models [13, 20, 211] have gained increasing attention due to their generative nature. However, those methods can only generate intra-subject motion variations and hardly achieve cross-domain inter-subject generation. Meanwhile, the limited size of the dataset with pathology makes these methods easily model collapse and cannot guarantee the diversity of generated motions. Recent advances in motion transfer (MT) [246] look to address those limitations by learning motion representations to achieve inter-subject transfer. This technique can generate diverse samples by combing the motion and content from different subjects to achieve conditional motion modeling, which is more flexible compared with the registration-based method.

In this study, we propose a conditional variational model to achieve content-controllable motion synthesis on cine-MRI (CMR) data. The proposed model aims to transfer the motion characteristics of source data to target data while preserving the visual appearance of the target. To achieve this goal, the proposed model maps the input sequence into two disentangled latent spaces learning the representations of motion and content separately. The deformation flow is then reconstructed and conditioned on the content feature. Through the swapping of latent representations from different subjects, the motion can be easily transferred correspondingly. We summarise the contributions of this work as follows:

- This study presents a novel approach to modelling cardiac motion through motion transfer, introducing a content-conditioned variational model.

- The proposed model effectively embeds and disentangles content and motion features from the inputs, enabling within-domain reconstruction through motion modelling and cross-domain conditional generation.
- Through experimentation, this study demonstrates the efficacy of the proposed model in capturing and generating realistic cardiac motion, highlighting its potential for clinical applications.

6.2 Methodology

Fig. 6.1 overviews our proposed method. Consider a 2D CMR pair, denoted by $\boldsymbol{s_i} \in \boldsymbol{S}$, which consists of two scans acquired at end-diastole (ED) and end-systole (ES), respectively. The pair is represented by $\boldsymbol{s_i} = \{\boldsymbol{x}_i^{\mathrm{ED}}, \boldsymbol{x}_i^{\mathrm{ES}}\}$, where $\boldsymbol{x_i} \in \mathbb{R}^{1\times H\times W}$, with H and W being the dimensions of the scans. In the proposed model, the input from the image domain is encoded into a content vector $\boldsymbol{c_i} \in \mathbb{R}^N$ in a determinant content domain, and a motion capturing variable $\boldsymbol{z}_{\boldsymbol{m_i}} \in \mathbb{R}^M$ in a probabilistic motion domain, where $N, M \ll H \times W$. Subsequently, a motion flow $\boldsymbol{d_i} \in \mathbb{R}^{2\times H\times W}$is predicted from the motion decoder given the concatenated latent code that is composed of the content vector and the sample from the motion domain. Moreover, the proposed model can perform motion transfer between any two input pairs $\boldsymbol{s_i}$ and $\boldsymbol{s_{j\neq i}}$ by swapping their content vectors in the latent space during the reconstruction phase. Specifically, the concatenation of $\boldsymbol{c_j}$ and $\boldsymbol{z}_{\boldsymbol{m_i}}$ yields a motion prediction $\boldsymbol{d}_{\boldsymbol{j\to i}}$, which leverages the anatomical structure of $\boldsymbol{s_j}$ while adopting the motion pattern from $\boldsymbol{s_i}$. This process is symmetric, and a similar approach is applied to obtain $\boldsymbol{d}_{\boldsymbol{i\to j}}$. In the following sections, we first introduce conditional variational modelling in our proposed approach and follow up by explaining the latent space disentanglement techniques and constraints in order to achieve efficient conditional MT.

6.2.1 Conditional Variational Modelling

Conditional Variational Autoencoder (CVAE) [247] offers an approach to control the data-generating process of a VAE with auxiliary covariates. Instead of adopting external covariates, the latent condition is acquired directly from the input data in the proposed model. A variational auto-encoder (VAE) and an auto-encoder (AE) structures are combined to extract both the texture information and the motion patterns

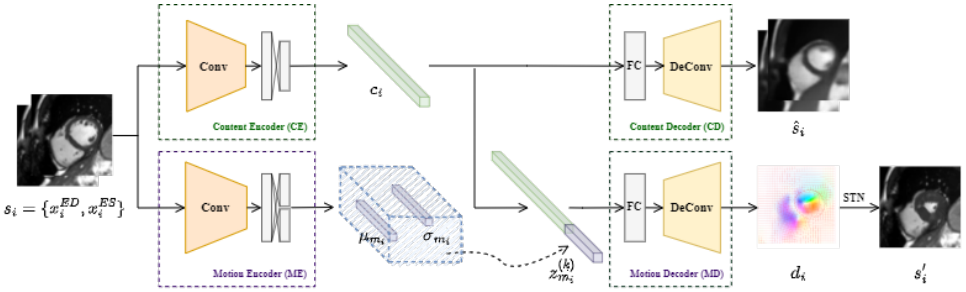


Figure 6.1: Illustration of the network architecture of the proposed framework. It includes parallel content and motion encoders composed of strided convolutional blocks and linear layers. The inputs are mapped to a deterministic content space and a probabilistic motion latent space. The content decoder reconstructs the inputs while the motion flow is generated based on the content feature and concatenated motion vector sampled from the approximated posterior. The concatenated feature vector is then decoded to a flow field, which is applied to the moving frame to generate the moved frame.

from the input data. The content variable, represented as $\boldsymbol{c_i} = \Phi_{\text{CE}}(\boldsymbol{s_i})$, is determined deterministically and serves as an instance-specific condition for predicting the deformation flow, $\boldsymbol{d_i}$. In contrast, the motion pattern variable $\boldsymbol{z_{m_i}}$ is modelled probabilistically. Specifically, the prior distribution of the motion variable is defined as a standard Gaussian distribution, i.e., $p(\boldsymbol{z_{m_i}}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. The approximated posterior of the motion variable follows multi-variate Gaussian (MVGs) distribution parameterized by mean and covariance:

$$q(\boldsymbol{z_{m_i}}|\boldsymbol{s_i}) = \mathcal{N}\left(\boldsymbol{\mu_{m_i}}, diag(\boldsymbol{\sigma_{m_i}^2})\right);\ [\boldsymbol{\mu_{m_i}}, \boldsymbol{\sigma_{m_i}}] = \Phi_{\text{ME}}(\boldsymbol{s_i}) \tag{6.1}$$

where $\boldsymbol{\mu_{m_i}}$ and $\boldsymbol{\sigma_{m_i}}$ are inferred from the motion encoder $\Phi_{\text{ME}}(\cdot)$. Subsequently, the prediction of the deformation flow $\boldsymbol{d_i}$ is made based on the concatenation of the content embedding $\boldsymbol{c_i}$ and the motion vector $\boldsymbol{z_{m_i}}$ sampled from the motion latent space. The input pair $\boldsymbol{s'i}$ is then reconstructed by warping each frame $\boldsymbol{x_i^{\text{ED}}}$ and $\boldsymbol{x_i^{\text{ES}}}$ using the corresponding deformation flow $\boldsymbol{d_i^{\text{fwd}}}$ and $\boldsymbol{d_i^{\text{bwd}}}$, respectively, through the utilization of STN grid sampler [226]. The formulation for the reconstruction phase is as follows:

$$p(\boldsymbol{s_i}|\boldsymbol{z_{m_i}^{(k)}}, \boldsymbol{c_i}) = \mathcal{N}(\boldsymbol{\mu_{s_i}}, \boldsymbol{\sigma_{s_i}});\ \boldsymbol{\mu_{s_i}} = \text{STN}\left(\boldsymbol{s_i}, \Phi_{\text{MD}}(\boldsymbol{z_{m_i}^{(k)}}, \boldsymbol{c_i})\right), \boldsymbol{\sigma_{s_i}} = 1 \tag{6.2}$$

where $\boldsymbol{z}_{\boldsymbol{m}_i}^{(\boldsymbol{k})}$ represents the k-th sample from the latent space, $\Phi_{\mathrm{MD}}(\cdot)$ denotes the motion decoder and $\mathrm{STN}(\cdot)$ is the grid sampler for warping. The covariance term $\boldsymbol{\sigma}_{\boldsymbol{s}_i}$ is manually set to 1 to stabilise the training process. Furthermore, to enforce the content vector embedding the texture feature from the input, we reconstruct the input alone using the content vector, $\hat{\boldsymbol{s}}_{\boldsymbol{i}} = \Phi_{\mathrm{CD}}(\boldsymbol{c}_{\boldsymbol{i}})$, where $\Phi_{\mathrm{CD}}(\cdot)$ denotes the content decoder.

The overall objective function in the proposed CVM results in the composition of the negative variational lower bound for motion prediction, and the reconstruction loss for encouraging sufficient content embedding:

$$\mathcal{L}_{cvm} = \mathbb{E}_{q(\boldsymbol{z}_{\boldsymbol{m}_i}|\boldsymbol{s}_i)}\left[-\log p(\boldsymbol{s}_i|\boldsymbol{z}_{\boldsymbol{m}_i}, \boldsymbol{c}_i) + \mathrm{KL}\left(q(\boldsymbol{z}_{\boldsymbol{m}_i}|\boldsymbol{s}_i)\|p(\boldsymbol{z}_{\boldsymbol{m}_i})\right)\right] + |\boldsymbol{s}_i - \hat{\boldsymbol{s}}_i| \tag{6.3}$$

where the first term represents the negative log-likelihood with respect to the approximated posterior distribution $q_{\boldsymbol{z}_{\boldsymbol{m}_i}}$, which is computed as the mean square error (MSE) between the input $\boldsymbol{s}_i$ and the reconstruction through warping, $\boldsymbol{s}'_i = \{\boldsymbol{x}'^{\mathrm{ED}}_i, \boldsymbol{x}'^{\mathrm{ES}}_i\}$. The second term is the Kullback-Leibler divergence, representing as $\mathrm{KL}(\cdot)$, between the approximated posterior and the prior distribution, which is used to regularize the latent variables $\boldsymbol{z}_{\boldsymbol{m}_i}$ during the inference process. Besides, an L1 constraint is imposed on the content reconstruction term to ensure a sharp reconstruction with detailed information. The network architecture of the proposed model is shown in Fig.6.1.

6.2.2 Content-conditioned Motion Transfer

As shown in Fig.6.2, the inputs from different domains (e.g. healthy cases and cases with dilated cardiomyopathy) can encode and reconstruct themselves through separate forward propagation. By applying the reconstruction stage described in Eq.6.2, the motion-transferred images for $\boldsymbol{s}_{\boldsymbol{i}\to\boldsymbol{j}}$ and $\boldsymbol{s}_{\boldsymbol{j}\to\boldsymbol{i}}$ are obtained as follows:

$$\boldsymbol{\mu}_{\boldsymbol{s}_{i\to j}} = \mathrm{STN}\left(\boldsymbol{s}_i, \Phi_{\mathrm{MD}}(\boldsymbol{z}^{(\boldsymbol{k})}_{\boldsymbol{m}_j}, \boldsymbol{c}_i)\right);\ \boldsymbol{\mu}_{\boldsymbol{s}_{j\to i}} = \mathrm{STN}\left(\boldsymbol{s}_j, \Phi_{\mathrm{MD}}(\boldsymbol{z}^{(\boldsymbol{k})}_{\boldsymbol{m}_i}, \boldsymbol{c}_j)\right) \tag{6.4}$$

where the transferred images have motion patterns of the motion templates and preserve their own texture content in the meantime. To prevent either content or motion dominant the motion generation and further encourage feature disentanglement, we introduce the conditional decoding consistency term:

$$\mathcal{L}_{cdc} = \mathrm{KL}\left(q(\boldsymbol{z}_{\boldsymbol{m}_i}|\boldsymbol{s}'_i)\|q(\boldsymbol{z}_{\boldsymbol{m}_i}|\boldsymbol{s}_i)\right) + max\left(D(\boldsymbol{c}_i, \boldsymbol{c}^{\mathrm{pos}}) - D(\boldsymbol{c}_i, \boldsymbol{c}^{\mathrm{neg}}) + \epsilon, 0\right) \tag{6.5}$$

where the first term encourages motion consistency in the latent space. For the second term, sit is observed that minimizing the distance between two content features will

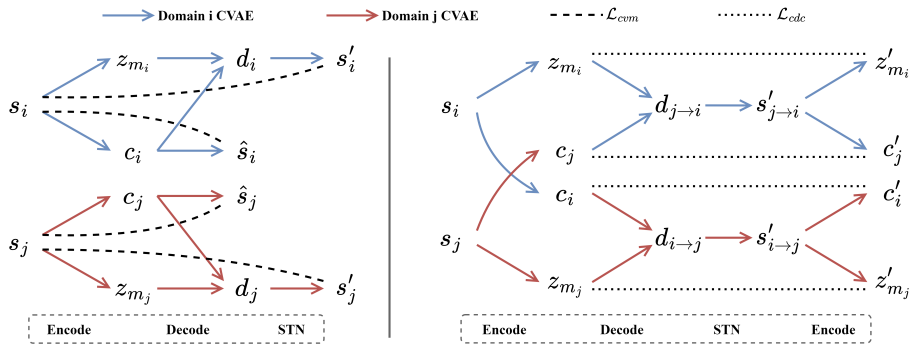


Figure 6.2: The reconstruction and motion transfer phases of our proposed framework. **(a) The reconstruction phase**. Each input undergoes forward propagation reconstructing the input via content embedding and flows via the content-conditioned latent sample. The reconstruction objectives is introduced to encourage content preservation and motion capturing (dashed line). **(b) The motion transfer phase**. Two cases swap their content vectors in the latent space and obtain the transferred flows. The newly formed pairs ($\boldsymbol{s'_{i\to j}}$ and $\boldsymbol{s'_{j\to i}}$) are then re-encoded to content and motion space to establish the consistency constrain (dotted line) and encourage the disentanglement of two latent spaces. In the training stage, two phases run in parallel.

result in meaningless content features. Instead, we adopt the triplet loss to enlarge the distance between content features from different inputs while maintaining the re-encoded content feature in close proximity. c^{pos} is the re-encoded content $\boldsymbol{c'}$, while c^{neg} is the content feature of a negative input $\boldsymbol{s}^{\text{neg}}$, which is derived from the random flipping of the batch dimension of $\boldsymbol{s_i}$. In the final objective function, we empirically weight $\mathcal{L}_{cdc}$ with 10^{-3}. The KLD term in $\mathcal{L}_{cvm}$ is weighted by 10^{-6} to provide a certain degree of freedom for motion embedding.

6.3 Experimental Results

We evaluate the proposed model on cardiac cine-MRI. We present an in-depth evaluation of two tasks, within and cross-domain motion transfer. We compare the performance of our proposed method, referred to as cMT-VAE, with a state-of-the-art CMR registration method, DSNet, which claims to have the ability to transfer motion [20]. Additionally, we perform experiments using two ablation study models: one utilizing a shared shallow feature encoder with limited convolutional down-sampling layers before

separate encoders (cMT-shared), and the other employing the same network architecture as the proposed method but trained without the decoding consistency constraint (cMT-ncdc).

6.3.1 Dataset

The present study involved the performance of experiments using the mid-slice of short-axis (SAX) view cardiac cine-MRI obtained from UK BioBank (UKB) [240], and ACDC [34] datasets. Each case in UKB contains a sequence of 50 SAX stacks throughout the cardiac cycle. Each SAX scan slice within the stack is captured with a 208×187 matrix size and an in-plane resolution of 1.8×1.8 mm^2. On the other hand, the ACDC dataset consists of 100 CMR sequences throughout a cardiac cycle, out of which 20 are normal and 80 cases cover various well-defined pathologies such as previous myocardial infarction (MINF), dilated cardiomyopathy (DCM), hypertrophic cardiomyopathy (HCM), and abnormal right ventricle (ARV). We selected 3000 cases from the UKB dataset for training and testing, with 2000 cases for training, 200 cases for validation, and 800 cases for testing. Additionally, we used 57 cases from the ACDC dataset (15 with DCM, 11 with MINF, 15 with ARV, and 16 normal cases) for testing. Notably, each case contains two mid-slices of end-diastolic (ED) and end-systolic (ES) stacks.

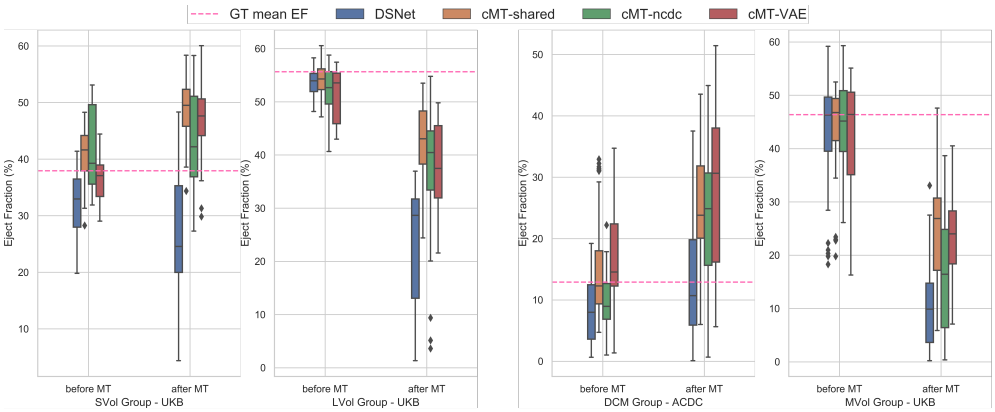
6.3.2 Implementation Details

In pre-processing, the input images are central-cropped with the size of 128×128 in pixels, and the image intensity is normalised to $[0, 1]$. The proposed framework and all experiments are implemented using Python 3.8 with PyTorch. The model is trained over 100 epochs with a batch size of 8, a learning rate of 0.0005, and optimised using Adam optimiser [228]. The learning rate is decayed by 0.0001 after every 20 epochs. The training process takes $\sim 6h$ on an NVIDIA GTX TITAN X GPU 8 GB.

6.3.3 Results and Analysis

Ejection Fraction (EF) measures the percentage of the total amount of blood that is pumped out with each heart contraction, which is considered a physiological parameter in various anomalous motion identifications. In the experiment, we compare the EF factor to quantitatively demonstrate the effectiveness of the proposed model. We first

group the UKB test set into three EF groups according to their ground-truth EF, SVol (33.9% $\sim$ 40.2%), MVol (40.2% $\sim$ 52.8%), and LVol (52.8% $\sim$ 59.1%). For ACDC dataset, we group the data based on their abnormality labels. For within-domain motion transfer, we randomly sampled 50 cases from SVol and LVol and swap the motion between each other, which results in 2500 motion-transferred outcomes for each group. For the cross-domain motion transfer case, we sample the same amount of samples from MVol and Then we calculate the EF of before and after motion transfer and compare among methods. For a fair comparison, we pre-trained a segmentation model (U-Net [9]) using all the available data from both UKBB and ACDC datasets to segment cardiac anatomical parts for EF calculation.



(a) Within-domain Motion Transfer (b) Cross-domain Motion Transfer

Figure 6.3: The box-plots of ejection fraction (EF) values in within-domain and cross-domain motion transfer. (a) Demonstration of the MT between SVol and LVol groups. For an efficient MT, the EF value after MT should increase for SVol and decrease for LVol. (b) Demonstration of the MT between DCM samples from ACDC and MVol samples from UKB. For DCM samples, after receiving the normal motion from MVol group, its EF value after EF should increase, and normal sample from MVol will decrease its contraction motion and result in smaller EF values after MT.

The boxplots in Fig.6.3 display the distribution of EF values of each group's EF variation before and after the motion transfer. The EF value before the MT is calculated based on the moved ED and ES warped by the reconstructed flows, and the after ones are calculated from the MT images applied with the transferred flows. For within-domain motion transfer Fig.6.3(a), the proposed model demonstrates efficient

and balanced EF value shift before and after MT, where SVol group's EF value increased and Lvol group's decreased with a similar amount. While the methods that did not contain a cross-domain consistency constraint, cMT-ncdc and DsNet, tend to be blunter to an increased contraction motion, resulting in giving non-obvious EF gain when it comes to an increased motion change. This might be caused by the entangle of content and motion latent domains. The proposed model encodes a marginally higher change in motion compared to its ablation models. In Fig.6.3(b), we illustrate the EF changes in cross-domain MT between the healthy MVol group and DCM group from ACDC dataset. The DCM group has a small EF value due to its pathological properties. The proposed models are still capable to encode the large contraction from MVol group to the DCM group and enlarging its EF values. We also observe the increase of diversity in EF values after MT, which also supported the model's generation capability. In Fig.6.3, we demonstrated qualitative results from MT experiments. From the within-domain MT results from the first two rows, we can see that the MT flow preserves the region of motion from its content input, in other words, the MT flow keeps a similar shape to the Recon flow. In the meantime, by comparing the saturation level of MT flow and Recon flow, it becomes evident that the MT flow adopts the strength of the motion from the motion source. For cross-domain MT, we observe a clear change between the Recon flow and MT flow. The MT flow indicates more contraction within the anatomical boundaries of the DCM sample, especially from the left ventricle. Besides, the differences in flows are focused on the centre of contraction, while they spread out to a wider area when the MT applies to anomalous samples.

6.4 Conclusion

In this paper, we have presented a conditional variational model to perform motion transfer on CMR images. The proposed framework has demonstrated promising results in revealing pixel-wise motion anomalies and enhancing data augmentation for downstream tasks, particularly those affected by class imbalance. The experiments have validated the effectiveness of the proposed approach, and we aim to extend it to volume data motion transfer and CMR sequential motion modelling in future research.

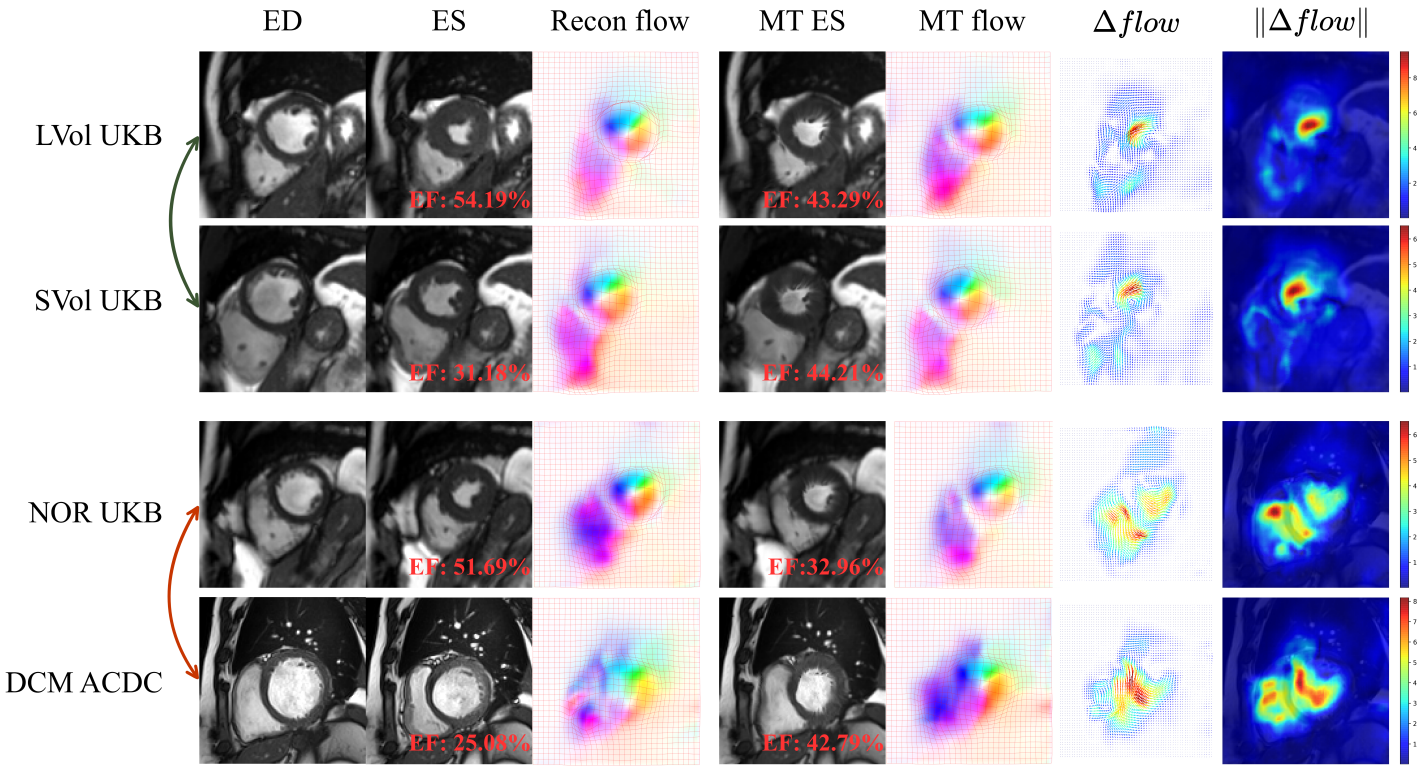


Figure 6.4: Demonstration of qualitative results in within-domain and cross-domain motion transfer. Two samples from LVol and SVol groups in UKB demonstrated within-domain MT in the first two rows. In this experiment, the LVol sample lost 10.90% EF after MT and the SVOl one gained 13.03% EF. For cross-domain MT, the healthy sample is taken from UKB and the anomalous one is taken from ACDC with DCM. The first three columns on the left display the original ED, ES frames and reconstructed flow describing the motion from ED to ES. In the following columns, we display the moved ED frame and its corresponding MT flow. The last two columns are the quiver field and the magnitude of changes before and after MT.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

Motivated by the intricate challenges of interpreting cardiac MRI data and the absence of spatio-temporal analysis of cardiac motion and topological structure, this research sought to revolutionise the way cardiac motion is modelled and segmented. This thesis explores motion modelling and regional analysis of cardiac cine-MRI via deep-learning techniques.

Rooted in Bayesian inference, we developed the base model for probabilistic modelling of cardiac motion. This approach not only provided a more sophisticated depiction of cardiac dynamics but also established a versatile framework potentially applicable to various anatomical regions. Three follow-up works are proposed based on this probabilistic generative modelling.

Firstly, we introduced a layer-wise gradient-surgery backpropagation scheduler to refine the accuracy of cardiac motion modelling and preserve the diffeomorphic characteristic of the predicted deformation field, leading to more efficient deformable image registration without the require of hyperparameter tuning. The proposed model, GS-Morph, demonstrates superior performance compared to state-of-the-art registration approaches. Moreover, the proposed layer-wise gradient-surgery algorithm is model-agnostic and can be integrated into general DL-based registration frameworks without introducing extra parameters or slowing down the inference.

Furthermore, this research made significant strides in full sequence segmentation, utilising spatio-temporal features from motion modelling to overcome the limitations posed by sparse annotations in cine-CMR sequence segmentation. This approach not only shows supreme motion modelling and segmentation accuracy but also demonstrates its ability to generate new cine-CMR sequence motions and corresponding semantic masks, which can potentially be used to reconstruct the 3D cardiac atlas across the cardiac cycle and thus contribute to deep atlas augmentation and analysis.

Additionally, the conditioned motion generation through motion transfer, cMT-VAE, explores feature disentanglement and can generate new motion sequences based on a given cine-CMR sequence. This is particularly useful in creating comprehensive datasets in scenarios where anomalous samples are limited. The proposed model also shows great potential in personalised modelling and pathology development assessment.

These contributions collectively have improved the efficiency, accuracy, and reliability of cardiac motion modelling and regional analysis, directly impacting patient care

and clinical decision-making. The exploration of generative properties in probabilistic generative models for uncertainty assessment not only facilitated the generation of additional samples for in-silico trials but also contributed to the overall reliability and robustness of medical imaging analysis.

7.2 Limitations and directions

While the advancements presented in this thesis are significant, it is essential to acknowledge its limitations. The frameworks proposed in Chapters 4, 5, and 6, though robust, may face challenges in their generalisability across diverse datasets and different pathological conditions. This limitation is crucial, as the effectiveness of any model in medical imaging is largely determined by its applicability to a wide range of real-world scenarios.

Moreover, the actual motion correspondence in heart muscle might not align precisely with the pixel displacement in cine-MRI. This discrepancy highlights a fundamental limitation in appearance-oriented motion modelling, which aims to minimise the appearance difference between reference and moved images but may fail to capture the true cardiac motion mechanism. To address this, future work should focus on integrating the results of semantic segmentation and motion to produce a more clinically relevant atlas-based motion modelling. This approach could bridge the gap between computational predictions and actual physiological movements, enhancing the clinical applicability of the research.

Another notable limitation is the lack of quality control assessment for the generative models, particularly in the context of deep augmentation in medical image analysis. While the proposed models demonstrate impressive generative capabilities, their practical utility is contingent on the quality and reliability of the generated samples. Future research should, therefore, include comprehensive quality control protocols to ensure that the generated data are not only diverse but also clinically valid. Regarding the motion transfer model, there is a need for more quantitative regional analysis experiments to assess its effectiveness thoroughly. Such experiments would provide valuable insights into the model's performance and help identify areas for improvement.

7.2.1 Future directions

Looking ahead, future research could aim at enhancing the generalisability of these models across varied cardiac conditions and datasets. This would involve not only refining the models to adapt to different cardiac anomalies but also expanding their applicability to a broader range of patient data. Conducting population analysis would be another vital area of exploration. By analysing large datasets, researchers can gain deeper insights into cardiac conditions, potentially leading to more personalised and effective treatments. Improving conditioned generation for deep augmentation is another promising avenue. This approach could lead to more robust models that can generate a diverse range of data samples, thereby addressing issues like data scarcity and imbalance in medical imaging.

Additionally, incorporating real-world clinical feedback into the development and refinement process is essential. This integration can ensure that these models align more closely with the practical demands of medical scenarios. Feedback from clinicians and patients can provide valuable insights that are often overlooked in a purely data-driven approach. This can lead to the development of models that are not only technically sound but also clinically relevant and user-friendly.

7.2.2 Expanding the data cohorts

Most investigations in this thesis were conducted on short-axis cardiac cine-MRI, primarily using healthy samples. While this was beneficial for proof-of-concept experiments and maintaining a consistent theme in comparison, it is crucial to expand these studies to include multi-centred data and other modalities, such as MRI of different organs. This expansion would test the robustness of the proposed probabilistic generative model in a more varied clinical setting. Future studies should aim to enhance the adaptability of these models across various cardiac conditions and datasets. Developing more general model architectures or exploring transfer learning techniques could facilitate the application of these models to different patient populations and medical imaging modalities, thus broadening their clinical utility.

7.2.3 Population studies with spatio-temporal analysis

The efficient framework proposed in Chapter 5 opens up possibilities for obtaining segmentation from spatio-temporal sequences. This means that beyond analysing car-

diac motion, variations in clinical indices such as left- and right-ventricle volume and myocardium mass can be monitored across the cardiac cycle. A detailed statistical population analysis of these clinical indices across the cardiac cycle would add significant value to the UK Biobank community. It would provide deeper insights into various pathologies and could lead to more targeted and effective treatment strategies.

7.2.4 Conditioned generation for deep augmentation

The motion transfer model proposed, cMT-VAE, enables the augmentation of new samples from a sample sequence. Implementing effective quality control of the generated samples is crucial. Involving the generated anomalous group in the dataset could address the issue of data imbalance and benefit the modelling of minority classes and diverse pathologies. Rather than using an entire sequence as the source for generation conditions, a more refined approach could be to distil the generation conditions down to specific clinical indices. This refinement would likely yield more potent and applicable results in clinical practice scenarios, leading to more personalised and precise medical interventions.

References

[1] P. DeSaix, G. J. Betts, E. Johnson, J. E. Johnson, K. Oksana, D. H. Kruse, B. Poe, J. A. Wise, and K. A. Young, "Anatomy & physiology (openstax)," 2013.

[2] P. J. Lynch, "Heart normal short axis echocardiography view," 2006.

[3] J. R. Mitchell and J.-J. Wang, "Expanding application of the wiggers diagram to teach cardiovascular physiology," *Advances in physiology education*, vol. 38, no. 2, pp. 170–175, 2014.

[4] 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.

[5] X. Zhang, R. V. Alexander, J. Yuan, and Y. Ding, "Computational analysis of cardiac contractile function," *Current cardiology reports*, vol. 24, no. 12, pp. 1983–1994, 2022.

[6] J. P. Ridgway, "Cardiovascular magnetic resonance physics for clinicians: part i," *Journal of cardiovascular magnetic resonance*, vol. 12, no. 1, pp. 1–28, 2010.

[7] E.-S. H. Ibrahim, "Myocardial tagging by cardiovascular magnetic resonance: evolution of techniques–pulse sequences, analysis algorithms, and applications," *Journal of Cardiovascular Magnetic Resonance*, vol. 13, pp. 1–40, 2011.

[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[10] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[11] S. E. Petersen, P. M. Matthews, F. Bamberg, D. A. Bluemke, J. M. Francis, M. G. Friedrich, P. Leeson, E. Nagel, S. Plein, F. E. Rademakers *et al.*, "Imaging in population science: cardiovascular magnetic resonance in 100,000 participants of uk biobank-rationale, challenges and approaches," *Journal of Cardiovascular Magnetic Resonance*, vol. 15, no. 1, pp. 1–10, 2013.

[12] V. M. Campello, P. Gkontra, C. Izquierdo, C. Martin-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.

[13] 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*. Springer, 2018, pp. 729–738.

[14] W. H. Organisation, "Cardiovascular diseases fact sheet," July 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[15] P. S. Rajiah, C. J. François, and T. Leiner, "Cardiac mri: state of the art," *Radiology*, vol. 307, no. 3, p. e223008, 2023.

[16] 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.

[17] M. F. Beg, M. I. Miller, A. Trouvé, and L. Younes, "Computing large deformation metric mappings via geodesic flows of diffeomorphisms," *International journal of computer vision*, vol. 61, no. 2, pp. 139–157, 2005.

[18] 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.

[19] 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.

[20] J. Krebs, H. Delingette, N. Ayache, and T. Mansi, "Learning a generative motion model from image sequences based on a latent motion matrix," *IEEE Transactions on Medical Imaging*, vol. 40, no. 5, pp. 1405–1416, 2021.

[21] E. Ferdian, A. Suinesiaputra, K. Fung, N. Aung, E. Lukaschuk, A. Barutcu, E. Maclean, J. Paiva, S. K. Piechnik, S. Neubauer *et al.*, "Fully automated myocardial strain estimation from cardiovascular mri–tagged images using a deep learning framework in the uk biobank," *Radiology: Cardiothoracic Imaging*, vol. 2, no. 1, p. e190032, 2020.

[22] S. Kermani, M. G. Oghli, A. Mohammadzadeh, and R. Kafieh, "Nf-rcnn: Heart localization and right ventricle wall motion abnormality detection in cardiac mri," *Physica Medica*, vol. 70, pp. 65–74, 2020.

[23] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," *arXiv preprint arXiv:1704.06857*, 2017.

[24] C. Chen, C. Qin, H. Qiu, G. Tarroni, J. Duan, W. Bai, and D. Rueckert, "Deep learning for cardiac image segmentation: A review," *Frontiers in Cardiovascular Medicine*, vol. 7, p. 25, 2020.

[25] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, 2018, pp. 3–11.

[26] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. DollÃ¡r, and R. Girshick, "Segment anything," 2023.

[27] M. B. Calisto and S. K. Lai-Yuen, "Adaen-net: An ensemble of adaptive 2d–3d fully convolutional networks for medical image segmentation," *Neural Networks*, vol. 126, pp. 76–94, 2020.

[28] 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.

[29] A. Sotiras, C. Davatzikos, and N. Paragios, "Deformable medical image registration: A survey," *IEEE transactions on medical imaging*, vol. 32, no. 7, pp. 1153–1190, 2013.

[30] E. Wilkins, L. Wilson, K. Wickramasinghe, P. Bhatnagar, J. Leal, R. Luengo-Fernandez, R. Burns, M. Rayner, and N. Townsend, "European cardiovascular disease statistics 2017," 2017.

[31] H. Ritchie, F. Spooner, and M. Roser, "Causes of death," *Our World in Data*, 2018. [Online]. Available: https://ourworldindata.org/causes-of-death

[32] P. Bhatnagar, K. Wickramasinghe, J. Williams, M. Rayner, and N. Townsend, "The epidemiology of cardiovascular disease in the uk 2014," *Heart*, vol. 101, no. 15, pp. 1182–1189, 2015.

[33] 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, p. 1, 2020.

[34] 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.

[35] B. Baessler, M. Mannil, S. Oebel, D. Maintz, H. Alkadhi, and R. Manka, "Subacute and chronic left ventricular myocardial scar: accuracy of texture analysis on nonenhanced cine mr images," *Radiology*, vol. 286, no. 1, pp. 103–112, 2018.

[36] N. Zhang, G. Yang, Z. Gao, C. Xu, Y. Zhang, R. Shi, J. Keegan, L. Xu, H. Zhang, Z. Fan *et al.*, "Deep learning for diagnosis of chronic myocardial infarction on nonenhanced cardiac cine mri," *Radiology*, vol. 291, no. 3, pp. 606–617, 2019.

[37] B. Baessler, C. Luecke, J. Lurz, K. Klingel, A. Das, M. Von Roeder, S. de Waha-Thiele, C. Besler, K.-P. Rommel, D. Maintz *et al.*, "Cardiac mri and texture analysis of myocardial t1 and t2 maps in myocarditis with acute versus chronic symptoms of heart failure," *Radiology*, vol. 292, no. 3, pp. 608–617, 2019.

[38] F. J. Olsen, K. G. Skaarup, M. C. H. Lassen, N. D. Johansen, M. Sengeløv, G. B. Jensen, P. Schnohr, J. L. Marott, P. Søgaard, G. Gislason *et al.*, "Normal values for myocardial work indices derived from pressure-strain loop analyses: from the cchs," *Circulation: Cardiovascular Imaging*, vol. 15, no. 5, p. e013712, 2022.

[39] T. Massardo, R. A. Gal, R. P. Grenier, D. H. Schmidt, and S. C. Port, "Left ventricular volume calculation using a count-based ratio method applied to multigated radionucide angiography," *Journal of Nuclear Medicine*, vol. 31, no. 4, pp. 450–456, 1990.

[40] M. J. Götte, A. C. Van Rossum, J. W. Twisk, J. P. Kuijer, J. T. Marcus, and C. A. Visser, "Quantification of regional contractile function after infarction: strain analysis superior to wall thickening analysis in discriminating infarct from remote myocardium," *Journal of the American College of Cardiology*, vol. 37, no. 3, pp. 808–817, 2001.

[41] A. H. A. W. G. on Myocardial Segmentation, R. for Cardiac Imaging:, M. D. Cerqueira, N. J. Weissman, V. Dilsizian, A. K. Jacobs, S. Kaul, W. K. Laskey, D. J. Pennell, J. A. Rumberger, T. Ryan *et al.*, "Standardized myocardial segmentation and nomenclature for tomographic imaging of the heart: a statement for healthcare professionals from the cardiac imaging committee of the council on clinical cardiology of the american heart association," *Circulation*, vol. 105, no. 4, pp. 539–542, 2002.

[42] P. S. Rajiah, K. Kalisz, J. Broncano, H. Goerne, J. D. Collins, C. J. François, E.-S. Ibrahim, and P. P. Agarwal, "Myocardial strain evaluation with cardiovascular mri: physics, principles, and clinical applications," *RadioGraphics*, vol. 42, no. 4, pp. 968–990, 2022.

[43] A. Scatteia, A. Baritussio, and C. Bucciarelli-Ducci, "Strain imaging using cardiac magnetic resonance," *Heart failure reviews*, vol. 22, pp. 465–476, 2017.

[44] A. Elen, H. F. Choi, D. Loeckx, H. Gao, P. Claus, P. Suetens, F. Maes, and J. D'hooge, "Three-dimensional cardiac strain estimation using spatio–temporal elastic registration of ultrasound images: A feasibility study," *IEEE transactions on medical imaging*, vol. 27, no. 11, pp. 1580–1591, 2008.

[45] E. Puyol-Antón, B. Ruijsink, W. Bai, H. Langet, M. De Craene, J. A. Schnabel, P. Piro, A. P. King, and M. Sinclair, "Fully automated myocardial strain estimation from cine mri using convolutional neural networks," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 1139–1143.

[46] B. Baeßler, F. Schaarschmidt, A. Dick, G. Michels, D. Maintz, and A. C. Bunck, "Diagnostic implications of magnetic resonance feature tracking derived myocardial strain parameters in acute myocarditis," *European journal of radiology*, vol. 85, no. 1, pp. 218–227, 2016.

[47] J. Weigand, J. C. Nielsen, P. P. Sengupta, J. Sanz, S. Srivastava, and S. Uppu, "Feature tracking-derived peak systolic strain compared to late gadolinium enhancement in troponin-positive myocarditis: a case–control study," *Pediatric cardiology*, vol. 37, pp. 696–703, 2016.

[48] C. McComb, D. Carrick, J. D. McClure, R. Woodward, A. Radjenovic, J. E. Foster, and C. Berry, "Assessment of the relationships between myocardial contractility and infarct tissue revealed by serial magnetic resonance imaging in patients with acute myocardial infarction," *The international journal of cardiovascular imaging*, vol. 31, pp. 1201–1209, 2015.

[49] J. N. Khan, A. Singh, S. A. Nazir, P. Kanagala, A. H. Gershlick, and G. P. McCann, "Comparison of cardiovascular magnetic resonance feature tracking and tagging for the assessment of left ventricular systolic strain in acute myocardial infarction," *European journal of radiology*, vol. 84, no. 5, pp. 840–848, 2015.

[50] T. J. Moon, N. Choueiter, T. Geva, A. M. Valente, K. Gauvreau, and D. M. Harrild, "Relation of biventricular strain and dyssynchrony in repaired tetralogy of

fallot measured by cardiac magnetic resonance to death and sustained ventricular tachycardia," *The American Journal of Cardiology*, vol. 115, no. 5, pp. 676–680, 2015.

[51] S. Orwat, G.-P. Diller, A. Kempny, R. Radke, B. Peters, T. Kühne, D. Boethig, M. Gutberlet, K.-O. Dubowy, P. Beerbaum *et al.*, "Myocardial deformation parameters predict outcome in patients with repaired tetralogy of fallot," *Heart*, vol. 102, no. 3, pp. 209–215, 2016.

[52] K. Hammouda, F. Khalifa, H. Abdeltawab, A. Elnakib, G. Giridharan, M. Zhu, C. Ng, S. Dassanayaka, M. Kong, H. Darwish *et al.*, "A new framework for performing cardiac strain analysis from cine mri imaging in mice," *Scientific reports*, vol. 10, no. 1, p. 7725, 2020.

[53] D. W. McRobbie, E. A. Moore, M. J. Graves, and M. R. Prince, *MRI from Picture to Proton*. Cambridge university press, 2017.

[54] H. Carr, "Steady-state free precession in nuclear magnetic resonance," *Physical Review*, vol. 112, no. 5, p. 1693, 1958.

[55] K. Scheffler and S. Lehnhardt, "Principles and applications of balanced ssfp techniques," *European radiology*, vol. 13, pp. 2409–2418, 2003.

[56] R. J. van der Geest and J. H. Reiber, "Quantification in cardiac mri," *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 10, no. 5, pp. 602–608, 1999.

[57] N. Benameur, M. A. Mohammed, R. Mahmoudi, Y. Arous, B. Garcia-Zapirain, K. H. Abdulkareem, and M. H. Bedoui, "Parametric methods for the regional assessment of cardiac wall motion abnormalities: comparison study," *CMC-Computers, Materials & Continua*, vol. 69, no. 1, pp. 1233–1252, 2021.

[58] N. F. Osman, S. Sampath, E. Atalar, and J. L. Prince, "Imaging longitudinal cardiac strain on short-axis images using strain-encoded mri," *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 46, no. 2, pp. 324–334, 2001.

[59] D. Kim, W. D. Gilson, C. M. Kramer, and F. H. Epstein, "Myocardial tissue tracking with two-dimensional cine displacement-encoded mr imaging: development and initial evaluation," *Radiology*, vol. 230, no. 3, pp. 862–871, 2004.

[60] E. A. Zerhouni, D. M. Parish, W. J. Rogers, A. Yang, and E. P. Shapiro, "Human heart: tagging with mr imaging–a method for noninvasive assessment of myocardial motion." *Radiology*, vol. 169, no. 1, pp. 59–63, 1988.

[61] S. S. Klein, T. P. Graham, and C. H. Lorenz, "Noninvasive delineation of normal right ventricular contractile motion with magnetic resonance imaging myocardial tagging," *Annals of biomedical engineering*, vol. 26, pp. 756–763, 1998.

[62] N. R. Saber and H. Wen, "Construction of the global lagrangian strain field in the myocardium using dense mri data," in *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 2. IEEE, 2004, pp. 3670–3673.

[63] I. Rodriguez, D. B. Ennis, and H. Wen, "Noninvasive measurement of myocardial tissue volume change during systolic contraction and diastolic relaxation in the canine left ventricle," *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 55, no. 3, pp. 484–490, 2006.

[64] D. L. Kraitchman, S. Sampath, E. Castillo, J. A. Derbyshire, R. C. Boston, D. A. Bluemke, B. L. Gerber, J. L. Prince, and N. F. Osman, "Quantitative ischemia detection during cardiac magnetic resonance stress testing by use of fastharp," *Circulation*, vol. 107, no. 15, pp. 2025–2030, 2003.

[65] G. Korosoglou, S. Lehrke, A. Wochele, B. Hoerig, D. Lossnitzer, H. Steen, E. Giannitsis, N. F. Osman, and H. A. Katus, "Strain-encoded cmr for the detection of inducible ischemia during intermediate stress," *JACC: Cardiovascular Imaging*, vol. 3, no. 4, pp. 361–371, 2010.

[66] E. Castillo, N. F. Osman, B. D. Rosen, I. El-Shehaby, L. Pan, M. Jerosch-Herold, S. Lai, D. A. Bluemke, and J. A. Lima, "Quantitative assessment of regional myocardial function with mr-tagging in a multi-center study: interobserver and intraobserver agreement of fast strain analysis with harmonic phase (harp) mri," *Journal of Cardiovascular Magnetic Resonance*, vol. 7, no. 5, pp. 783–791, 2005.

[67] C. M. Kramer, N. Reichek, V. A. Ferrari, T. Theobald, J. Dawson, and L. Axel, "Regional heterogeneity of function in hypertrophic cardiomyopathy." *Circulation*, vol. 90, no. 1, pp. 186–194, 1994.

[68] Y. J. Kim, B. W. Choi, J. Hur, H.-J. Lee, J. S. Seo, T. H. Kim, K. O. Choe, and J.-W. Ha, "Delayed enhancement in hypertrophic cardiomyopathy: comparison with myocardial tagging mri," *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 27, no. 5, pp. 1054–1060, 2008.

[69] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[70] 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.

[71] J. L. Elman, "Finding structure in time," *Cognitive science*, vol. 14, no. 2, pp. 179–211, 1990.

[72] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[73] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2794–2802.

[74] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Advances in neural information processing systems*, vol. 29, 2016.

[75] Y. Qin, D. Song, H. Chen, W. Cheng, G. Jiang, and G. Cottrell, "A dual-stage attention-based recurrent neural network for time series prediction," *arXiv preprint arXiv:1704.02971*, 2017.

[76] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[77] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. pmlr, 2015, pp. 448–456.

[78] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[79] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[80] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[81] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211–252, 2015.

[82] V. R. Allugunti, "Breast cancer detection based on thermographic images using machine learning and deep learning algorithms," *International Journal of Engineering in Computer Science*, vol. 4, no. 1, pp. 49–56, 2022.

[83] F. Schwendicke, K. Elhennawy, S. Paris, P. Friebertshäuser, and J. Krois, "Deep learning for caries lesion detection in near-infrared light transillumination images: A pilot study," *Journal of dentistry*, vol. 92, p. 103260, 2020.

[84] R. Cui, M. Liu, A. D. N. Initiative *et al.*, "Rnn-based longitudinal analysis for diagnosis of alzheimerâ€™s disease," *Computerized Medical Imaging and Graphics*, vol. 73, pp. 1–10, 2019.

[85] J. Zhang, B. Zheng, A. Gao, X. Feng, D. Liang, and X. Long, "A 3d densely connected convolution neural network with connection-wise attention mechanism for alzheimer's disease classification," *Magnetic Resonance Imaging*, vol. 78, pp. 119–126, 2021.

[86] J. A. Fries, P. Varma, V. S. Chen, K. Xiao, H. Tejeda, P. Saha, J. Dunnmon, H. Chubb, S. Maskatia, M. Fiterau *et al.*, "Weakly supervised classification of aortic valve malformations using unlabeled cardiac mri sequences," *Nature communications*, vol. 10, no. 1, p. 3111, 2019.

[87] J. Ma and B. Wang, "Segment anything in medical images," *arXiv preprint arXiv:2304.12306*, 2023.

[88] F. Isensee, P. F. Jäger, P. M. Full, P. Vollmuth, and K. H. Maier-Hein, "nnu-net for brain tumor segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers, Part II 6*. Springer, 2021, pp. 118–132.

[89] H. Tang, C. Zhang, and X. Xie, "Nodulenet: Decoupled false positive reduction for pulmonary nodule detection and segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*. Springer, 2019, pp. 266–274.

[90] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[91] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[92] 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*, 2016, pp. 770–778.

[93] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *Eleventh annual conference of the international speech communication association*, 2010.

[94] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu, "Video paragraph captioning using hierarchical recurrent neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4584–4593.

[95] Q. Zheng, H. Delingette, N. Duchateau, and N. Ayache, "3-d consistent and robust segmentation of cardiac images by deep learning with spatial propagation," *IEEE transactions on medical imaging*, vol. 37, no. 9, pp. 2137–2148, 2018.

[96] K. K. Wong, W. Xu, M. Ayoub, Y.-L. Fu, H. Xu, R. Shi, M. Zhang, F. Su, Z. Huang, and W. Chen, "Brain image segmentation of the corpus callosum by combining bi-directional convolutional lstm and u-net using multi-slice ct and mri," *Computer Methods and Programs in Biomedicine*, vol. 238, p. 107602, 2023.

[97] D. Zhang, I. Icke, B. Dogdas, S. Parimal, S. Sampath, J. Forbes, A. Bagchi, C.-L. Chin, and A. Chen, "A multi-level convolutional lstm model for the segmentation of left ventricle myocardium in infarcted porcine cine mr images," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 470–473.

[98] P. Dileep, K. N. Rao, P. Bodapati, S. Gokuruboyina, R. Peddi, A. Grover, and A. Sheetal, "An automatic heart disease prediction using cluster-based bi-directional lstm (c-bilstm) algorithm," *Neural Computing and Applications*, vol. 35, no. 10, pp. 7253–7266, 2023.

[99] S. Hochreiter, Y. Bengio, P. Frasconi, J. Schmidhuber *et al.*, "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies," 2001.

[100] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[101] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning precise timing with lstm recurrent networks," *Journal of machine learning research*, vol. 3, no. Aug, pp. 115–143, 2002.

[102] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, vol. 28, 2015.

[103] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[104] I. T. Jolliffe, "Principal component analysis," *Technometrics*, vol. 45, no. 3, p. 276, 2003.

[105] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on statistical learning in computer vision, ECCV*, vol. 1, no. 1-22. Prague, 2004, pp. 1–2.

[106] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[107] H. Steck and D. G. Garcia, "On the regularization of autoencoders," *arXiv preprint arXiv:2110.11402*, 2021.

[108] J. Schlemper, O. Oktay, W. Bai, D. C. Castro, J. Duan, C. Qin, J. V. Hajnal, and D. Rueckert, "Cardiac mr segmentation from undersampled k-space using deep latent representation learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 259–267.

[109] O. Oktay, W. Bai, M. Lee, R. Guerrero, K. Kamnitsas, J. Caballero, A. de Marvao, S. Cook, D. Oâ€™Regan, and D. Rueckert, "Multi-input cardiac image super-resolution using convolutional neural networks," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2016, pp. 246–254.

[110] Q. Yue, X. Luo, Q. Ye, L. Xu, and X. Zhuang, "Cardiac segmentation from lge mri using deep neural network incorporating shape and spatial priors," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 559–567.

[111] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[112] 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*, 2017, pp. 1125–1134.

[113] 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*, 2017, pp. 2223–2232.

[114] Y. Chen, X.-H. Yang, Z. Wei, A. A. Heidari, N. Zheng, Z. Li, H. Chen, H. Hu, Q. Zhou, and Q. Guan, "Generative adversarial networks in medical image augmentation: A review," *Computers in Biology and Medicine*, vol. 144, p. 105382, 2022.

[115] Y. Xia, L. Zhang, N. Ravikumar, R. Attar, S. K. Piechnik, S. Neubauer, S. E. Petersen, and A. F. Frangi, "Recovering from missing data in population imaging–cardiac mr image imputation via conditional generative adversarial nets," *Medical Image Analysis*, vol. 67, p. 101812, 2021.

[116] Y. Xia, N. Ravikumar, and A. F. Frangi, "Learning to complete incomplete hearts for population analysis of cardiac mr images," *Medical Image Analysis*, vol. 77, p. 102354, 2022.

[117] H. Yang, J. Sun, A. Carass, C. Zhao, J. Lee, Z. Xu, and J. Prince, "Unpaired brain mr-to-ct synthesis using a structure-constrained cyclegan," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*. Springer, 2018, pp. 174–182.

[118] R. Kalantar, C. Messiou, J. M. Winfield, A. Renn, A. Latifoltojar, K. Downey, A. Sohaib, S. Lalondrelle, D.-M. Koh, and M. D. Blackledge, "Ct-based pelvic t1-weighted mr image synthesis using unet, unet++ and cycle-consistent generative adversarial network (cycle-gan)," *Frontiers in Oncology*, vol. 11, p. 665807, 2021.

[119] 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.

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

[121] C. Han, L. Rundo, K. Murao, T. Noguchi, Y. Shimahara, Z. Á. Milacski, S. Koshino, E. Sala, H. Nakayama, and S. Satoh, "Madgan: Unsupervised medical anomaly detection gan using multiple adjacent brain mri slice reconstruction," *BMC bioinformatics*, vol. 22, no. 2, pp. 1–20, 2021.

[122] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio, "A recurrent latent variable model for sequential data," *Advances in neural information processing systems*, vol. 28, 2015.

[123] I. Sutskever, O. Vinyals, and Q. Le, "Sequence to sequence learning with neural networks," *Advances in NIPS*, 2014.

[124] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," in *International conference on learning representations*, 2016.

[125] S. Webb, J. P. Chen, M. Jankowiak, and N. Goodman, "Improving automated variational inference with normalizing flows," in *ICML, Workshop on Automated Machine Learning*, 2019.

[126] C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, and O. Winther, "Ladder variational autoencoders," *Advances in neural information processing systems*, vol. 29, 2016.

[127] D. Kingma, T. Salimans, B. Poole, and J. Ho, "Variational diffusion models," *Advances in neural information processing systems*, vol. 34, pp. 21 696–21 707, 2021.

[128] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *ACM Computing Surveys*, 2022.

[129] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *Advances in neural information processing systems*, vol. 32, 2019.

[130] ——, "Improved techniques for training score-based generative models," *Advances in neural information processing systems*, vol. 33, pp. 12 438–12 448, 2020.

[131] C. Krittanawong, H. Zhang, Z. Wang, M. Aydar, and T. Kitai, "Artificial intelligence in precision cardiovascular medicine," *Journal of the American College of Cardiology*, vol. 69, no. 21, pp. 2657–2664, 2017.

[132] X. Liu, L. Song, S. Liu, and Y. Zhang, "A review of deep-learning-based medical image segmentation methods," *Sustainability*, vol. 13, no. 3, p. 1224, 2021.

[133] C. Biffi, O. Oktay, G. Tarroni, W. Bai, A. De Marvao, G. Doumou, M. Rajchl, R. Bedair, S. Prasad, S. Cook *et al.*, "Learning interpretable anatomical features through deep generative models: Application to cardiac remodeling," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2018, pp. 464–471.

[134] W. Xue, G. Brahm, S. Pandey, S. Leung, and S. Li, "Full left ventricle quantification via deep multitask relationships learning," *Medical image analysis*, vol. 43, pp. 54–65, 2018.

[135] Q. Zheng, H. Delingette, and N. Ayache, "Explainable cardiac pathology classification on cine mri with motion characterization by semi-supervised learning of apparent flow," *Medical image analysis*, vol. 56, pp. 80–95, 2019.

[136] 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, p. 95, 2019.

[137] E. P. Anton, M. Pop, C. Martín-Isla, M. Sermesant, A. Suinesiaputra, O. Camara, K. Lekadir, and A. Young, *Statistical Atlases and Computational Models of the Heart. Multi-Disease, Multi-View, and Multi-Center Right Ventricular Segmentation in Cardiac MRI Challenge: 12th International Workshop, STACOM 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Revised Selected Papers*. Springer Nature, 2022, vol. 13131.

[138] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.

[139] H. Ng, S. Ong, K. Foong, P. Goh, and W. Nowinski, "Medical image segmentation using k-means clustering and improved watershed algorithm," in *2006 IEEE*

southwest symposium on image analysis and interpretation. IEEE, 2006, pp. 61–65.

[140] N. Dhanachandra, K. Manglem, and Y. J. Chanu, "Image segmentation using k-means clustering algorithm and subtractive clustering algorithm," *Procedia Computer Science*, vol. 54, pp. 764–771, 2015.

[141] C. Li, C. Xu, C. Gui, and M. D. Fox, "Distance regularized level set evolution and its application to image segmentation," *IEEE transactions on image processing*, vol. 19, no. 12, pp. 3243–3254, 2010.

[142] L. A. Vese and T. F. Chan, "A multiphase level set framework for image segmentation using the mumford and shah model," *International journal of computer vision*, vol. 50, no. 3, pp. 271–293, 2002.

[143] A. Işın, C. Direkoğlu, and M. Şah, "Review of mri-based brain tumor image segmentation using deep learning methods," *Procedia Computer Science*, vol. 102, pp. 317–324, 2016.

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

[145] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[146] S. Gandhi, W. Mosleh, J. Shen, and C.-M. Chow, "Automation, machine learning, and artificial intelligence in echocardiography: a brave new world," *Echocardiography*, vol. 35, no. 9, pp. 1402–1418, 2018.

[147] M. A. Mazurowski, M. Buda, A. Saha, and M. R. Bashir, "Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on mri," *Journal of magnetic resonance imaging*, vol. 49, no. 4, pp. 939–954, 2019.

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

[149] M. Khened, V. A. Kollerathu, and G. Krishnamurthi, "Fully convolutional multi-scale residual densenets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers," *Medical image analysis*, vol. 51, pp. 21–45, 2019.

[150] Y. Jang, Y. Hong, S. Ha, S. Kim, and H.-J. Chang, "Automatic segmentation of lv and rv in cardiac mri," in *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer, 2017, pp. 161–169.

[151] A. Guala, G. Teixido-Tura, L. Dux-Santoy, C. Granato, A. Ruiz-Muñoz, F. Valente, L. Galian-Gay, L. Gutiérrez, T. González-Alujas, K. Johnson *et al.*, "Decreased rotational flow and circumferential wall shear stress as early markers of descending aorta dilation in marfan syndrome: a 4d flow cmr study," *Journal of Cardiovascular Magnetic Resonance*, vol. 21, no. 1, p. 63, 2019.

[152] J. Li, Z. L. Yu, Z. Gu, H. Liu, and Y. Li, "Dilated-inception net: multi-scale feature aggregation for cardiac right ventricle segmentation," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 12, pp. 3499–3508, 2019.

[153] X.-Y. Zhou and G.-Z. Yang, "Normalization in training u-net for 2-d biomedical semantic segmentation," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1792–1799, 2019.

[154] J. Zhang, J. Du, H. Liu, X. Hou, Y. Zhao, and M. Ding, "Lu-net: an improved u-net for ventricular segmentation," *IEEE Access*, vol. 7, pp. 92 539–92 546, 2019.

[155] C. Cong and H. Zhang, "Invert-u-net dnn segmentation model for mri cardiac left ventricle segmentation," *The Journal of Engineering*, vol. 2018, no. 16, pp. 1463–1467, 2018.

[156] J. Sander, B. D. de Vos, J. M. Wolterink, and I. Išgum, "Towards increased trustworthiness of deep learning segmentation methods on cardiac mri," in *Medical Imaging 2019: Image Processing*, vol. 10949. International Society for Optics and Photonics, 2019, p. 1094919.

[157] M. Chen, L. Fang, and H. Liu, "Fr-net: Focal loss constrained deep residual networks for segmentation of cardiac mri," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 764–767.

[158] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 11–19.

[159] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*. Springer, 2016, pp. 424–432.

[160] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[161] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[162] C. Chen, C. Biffi, G. Tarroni, S. Petersen, W. Bai, and D. Rueckert, "Learning shape priors for robust cardiac mr segmentation from multi-view images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 523–531.

[163] C. Zotti, Z. Luo, A. Lalande, and P.-M. Jodoin, "Convolutional neural network with shape prior applied to cardiac mri segmentation," *IEEE journal of biomedical and health informatics*, vol. 23, no. 3, pp. 1119–1128, 2018.

[164] W. Yan, Y. Wang, Z. Li, R. J. Van Der Geest, and Q. Tao, "Left ventricle segmentation via optical-flow-net from short-axis cine mri: preserving the temporal coherence of cardiac motion," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 613–621.

[165] X. Du, S. Yin, R. Tang, Y. Zhang, and S. Li, "Cardiac-deepied: Automatic pixel-level deep segmentation for cardiac bi-ventricle using improved end-to-end

encoder-decoder network," *IEEE journal of translational engineering in health and medicine*, vol. 7, pp. 1–10, 2019.

[166] W. Bai, H. Suzuki, C. Qin, G. Tarroni, O. Oktay, P. M. Matthews, and D. Rueckert, "Recurrent neural networks for aortic image sequence segmentation with sparse annotations," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2018, pp. 586–594.

[167] P. Peng, K. Lekadir, A. Gooya, L. Shao, S. E. Petersen, and A. F. Frangi, "A review of heart chamber segmentation for structural and functional analysis using cardiac magnetic resonance imaging," *Magnetic Resonance Materials in Physics, Biology and Medicine*, vol. 29, no. 2, pp. 155–195, 2016.

[168] A. Giger, R. Sandkühler, C. Jud, G. Bauman, O. Bieri, R. Salomir, and P. C. Cattin, "Respiratory motion modelling using cgans," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 81–88.

[169] X. Teng, Y. Chen, Y. Zhang, and L. Ren, "Respiratory deformation registration in 4d-ct/cone beam ct using deep learning," *Quantitative Imaging in Medicine and Surgery*, vol. 11, no. 2, p. 737, 2021.

[170] T. Mezheritsky, L. V. Romaguera, W. Le, and S. Kadoury, "Population-based 3d respiratory motion modelling from convolutional autoencoders for 2d ultrasound-guided radiotherapy," *Medical image analysis*, vol. 75, p. 102260, 2022.

[171] M.-M. Rohé, M. Sermesant, and X. Pennec, "Automatic multi-atlas segmentation of myocardium with svf-net," in *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer, 2017, pp. 170–177.

[172] J. Krebs, T. Mansi, N. Ayache, and H. Delingette, "Probabilistic motion modeling from medical image sequences: application to cardiac cine-MRI," in *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer, 2019, pp. 176–185.

[173] 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.

[174] 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.

[175] S. S. M. Salehi, S. Khan, D. Erdogmus, and A. Gholipour, "Real-time deep registration with geodesic loss," *arXiv preprint arXiv:1803.05982*, 2018.

[176] 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.

[177] K. A. Eppenhof, M. W. Lafarge, P. Moeskops, M. Veta, and J. P. Pluim, "Deformable image registration using convolutional neural networks," in *Medical Imaging 2018: Image Processing*, vol. 10574. International Society for Optics and Photonics, 2018, p. 105740S.

[178] T. Sentker, F. Madesta, and R. Werner, "GDL-FIRE 4D: Deep learning-based fast 4D CT image registration," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 765–773.

[179] 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.

[180] 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*. Springer, 2017, pp. 300–308.

[181] 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.

[182] 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.

[183] H. Uzunova, M. Wilms, H. Handels, and J. Ehrhardt, "Training CNNs for image registration from few samples with model-based data augmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 223–231.

[184] H. Sokooti, B. de Vos, F. Berendsen, M. Ghafoorian, S. Yousefi, B. P. Lelieveldt, I. Isgum, and M. Staring, "3D convolutional neural networks image registration based on efficient supervised learning from artificial deformations," *arXiv preprint arXiv:1908.10235*, 2019.

[185] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2758–2766.

[186] 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.

[187] 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*. Springer, 2018, pp. 729–738.

[188] J. Krebs, T. Mansi, B. Mailhé, N. Ayache, and H. Delingette, "Unsupervised probabilistic deformation modeling for robust diffeomorphic registration," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 101–109.

[189] J. Krebs, H. Delingette, N. Ayache, and T. Mansi, "Learning a generative motion model from image sequences based on a latent motion matrix," *IEEE Transactions on Medical Imaging*, vol. 40, no. 5, pp. 1405–1416, 2021.

[190] 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*. Springer, 2017, pp. 204–212.

[191] 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.

[192] D. Kuang and T. Schmah, "FAIM–a convnet method for unsupervised 3D medical image registration," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2019, pp. 646–654.

[193] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," *Advances in neural information processing systems*, vol. 28, 2015.

[194] 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.

[195] B. B. Avants, N. L. 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.

[196] 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*, 2016, pp. 134–142.

[197] J. Zhang, "Inverse-consistent deep networks for unsupervised deformable image registration," *arXiv preprint arXiv:1809.03443*, 2018.

[198] 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.

[199] J. Krebs, T. Mansi, B. Mailhé, N. Ayache, and H. Delingette, "Unsupervised probabilistic deformation modeling for robust diffeomorphic registration," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 101–109.

[200] D. P. Kingma, S. Mohamed, D. Jimenez Rezende, and M. Welling, "Semi-supervised learning with deep generative models," *Advances in neural information processing systems*, vol. 27, 2014.

[201] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[202] A. Zakeri, A. Hokmabadi, N. Bi, I. Wijesinghe, M. G. Nix, S. E. Petersen, A. F. Frangi, Z. A. Taylor, and A. Gooya, "Dragnet: Learning-based deformable registration for realistic cardiac mr sequence generation from a single frame," *Medical Image Analysis*, p. 102678, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841522003061

[203] A. F. Psaros, X. Meng, Z. Zou, L. Guo, and G. E. Karniadakis, "Uncertainty quantification in scientific machine learning: Methods, metrics, and comparisons," *arXiv preprint arXiv:2201.07766*, 2022.

[204] A. G. Wilson and P. Izmailov, "Bayesian deep learning and a probabilistic perspective of generalization," *Advances in neural information processing systems*, vol. 33, pp. 4697–4708, 2020.

[205] Y. Gal and Z. Ghahramani, "Bayesian convolutional neural networks with Bernoulli approximate variational inference," *arXiv preprint arXiv:1506.02158*, 2015.

[206] A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," *arXiv preprint arXiv:1511.02680*, 2015.

[207] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *Advances in neural information processing systems*, vol. 30, 2017.

[208] S. Kohl, B. Romera-Paredes, C. Meyer, J. De Fauw, J. R. Ledsam, K. Maier-Hein, S. Eslami, D. Jimenez Rezende, and O. Ronneberger, "A probabilistic U-Net for segmentation of ambiguous images," *Advances in neural information processing systems*, vol. 31, 2018.

[209] A. Fry, T. J. Littlejohns, C. Sudlow, N. Doherty, L. Adamska, T. Sprosen, R. Collins, and N. E. Allen, "Comparison of sociodemographic and health-related characteristics of uk biobank participants with those of the general population," *American journal of epidemiology*, vol. 186, no. 9, pp. 1026–1034, 2017.

[210] 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.

[211] A. Zakeri, A. Hokmabadi, N. Bi, I. Wijesinghe, M. G. Nix, S. E. Petersen, A. F. Frangi, Z. A. Taylor, and A. Gooya, "Dragnet: Learning-based deformable registration for realistic cardiac mr sequence generation from a single frame," *Medical Image Analysis*, vol. 83, p. 102678, 2023.

[212] J. J. Yu, A. W. Harley, and K. G. Derpanis, "Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness," in *European Conference on Computer Vision*. Springer, 2016, pp. 3–10.

[213] 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*. Springer, 2021, pp. 46–55.

[214] 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.

[215] 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.

[216] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "An unsupervised learning model for deformable medical image registration," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9252–9260.

[217] Y. Huang, S. Ahmad, J. Fan, D. Shen, and P.-T. Yap, "Difficulty-aware hierarchical convolutional neural networks for deformable registration of brain mr images," *Medical image analysis*, vol. 67, p. 101817, 2021.

[218] 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*. Springer, 2021, pp. 46–55.

[219] X. Jia, A. Thorley, W. Chen, H. Qiu, L. Shen, I. B. Styles, H. J. Chang, A. Leonardis, A. De Marvao, D. P. Oâ€™Regan *et al.*, "Learning a model-driven variational network for deformable image registration," *IEEE Transactions on Medical Imaging*, vol. 41, no. 1, pp. 199–212, 2021.

[220] A. Hoopes, M. Hoffmann, B. Fischl, J. Guttag, and A. V. Dalca, "Hypermorph: Amortized hyperparameter learning for image registration," in *International Conference on Information Processing in Medical Imaging*. Springer, 2021, pp. 3–17.

[221] T. C. Mok and A. Chung, "Conditional deformable image registration with convolutional neural network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 35–45.

[222] D. Ha, A. Dai, and Q. V. Le, "Hypernetworks," *arXiv preprint arXiv:1609.09106*, 2016.

[223] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1501–1510.

[224] T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn, "Gradient surgery for multi-task learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5824–5836, 2020.

[225] L. Mansilla, R. Echeveste, D. H. Milone, and E. Ferrante, "Domain generalization via gradient surgery," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6630–6638.

[226] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," *Advances in neural information processing systems*, vol. 28, 2015.

[227] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[228] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *ICLR*, 2015.

[229] T. C. Mok and A. Chung, "Large deformation diffeomorphic image registration with laplacian pyramid networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 211–221.

[230] 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.

[231] C. Qin, W. Bai, J. Schlemper, S. E. Petersen, S. K. Piechnik, S. Neubauer, and D. Rueckert, "Joint learning of motion estimation and segmentation for cardiac mr image sequences," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 472–480.

[232] 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*. Springer, 2019, pp. 420–429.

[233] T. Estienne, M. Vakalopoulou, S. Christodoulidis, E. Battistela, M. Lerousseau, A. Carre, G. Klausner, R. Sun, C. Robert, S. Mougiakakou *et al.*, "U-resnet: Ultimate coupling of registration and segmentation with deep nets," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2019, pp. 310–319.

[234] 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.

[235] 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.

[236] A. Yezzi, L. Zollei, and T. Kapur, "A variational framework for joint segmentation and registration," in *Proceedings IEEE Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA 2001)*. IEEE, 2001, pp. 44–51.

[237] K. M. Pohl, J. Fisher, W. E. L. Grimson, R. Kikinis, and W. M. Wells, "A bayesian model for joint segmentation and registration," *NeuroImage*, vol. 31, no. 1, pp. 228–239, 2006.

[238] D. Mahapatra and Y. Sun, "Joint registration and segmentation of dynamic cardiac perfusion images using mrfs," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2010, pp. 493–501.

[239] D. Grzech, M. F. Azampour, B. Glocker, J. Schnabel, N. Navab, B. Kainz, and L. Le Folgoc, "A variational bayesian method for similarity learning in non-rigid image registration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 119–128.

[240] 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 biobankâ€™s cardiovascular magnetic resonance protocol," *Journal of cardiovascular magnetic resonance*, vol. 18, no. 1, pp. 1–7, 2015.

[241] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[242] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[243] H. Wang and A. A. Amini, "Cardiac motion and deformation recovery from mri: a review," *IEEE transactions on medical imaging*, vol. 31, no. 2, pp. 487–503, 2011.

[244] P. Lu, W. Bai, D. Rueckert, and J. A. Noble, "Modelling cardiac motion via spatio-temporal graph convolutional networks to boost the diagnosis of heart conditions," in *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges: 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers 11*. Springer, 2021, pp. 56–65.

[245] K. Punithakumar, I. B. Ayed, A. Islam, A. Goela, I. G. Ross, J. Chong, and S. Li, "Regional heart motion abnormality detection: An information theoretic approach," *Medical image analysis*, vol. 17, no. 3, pp. 311–324, 2013.

[246] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "Animating arbitrary objects via deep motion transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2377–2386.

[247] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," *Advances in neural information processing systems*, vol. 28, 2015.