UNIVERSITY OF LEEDS

Deep generative model for synthesising and analysing cardiac magnetic resonance images

Nina Cheng

University of Leeds

School of Computer Science

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

May, 2025

Intellectual Property Statement

The candidate confirms that the submitted work is entirely her own, except for contributions to co-authored publications cited therein. References to the work of others have been appropriately acknowledged in the thesis. The specific contributions of the candidate and co-authors are detailed below.

I am the main author of all the publications listed below. I was responsible for the experiment design and manuscript writing, including analysis and discussion of results. In addition, I developed the code, performed data processing and performed experiments. Contributions of other co-authors include assisting with experiment design, providing data annotations, participating in discussion of results, and reviewing the manuscript.

- **Chapter 4:**
    1. Cheng N, Bonazzola R, Ravikumar N, et al. "A generative framework for predicting myocardial strain from cine-cardiac magnetic resonance imaging." *Annual Conference on Medical Image Understanding and Analysis*, 2022: 482-493.
    2. Cheng N, Zakeri A, Ravikumar N, Frangi AF. "Automated Myocardial Strain Quantification via Synthesised Tagging-MRI: A Sparse Multi-Channel Variational Autoencoder Approach." *under review.*
- **Chapter 5:**
    1. Cheng N, Liu Z, Deo Y, et al. "Synthesising 3D Cardiac CINE-MR Images and Corresponding Segmentation Masks using a Latent Diffusion Model." *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2024: 1-5..
- **Chapter 6:**

1. Cheng N, Zakeri A, Ravikumar N, Frangi AF. "Conditional 4D spatio-temporal latent diffusion generative model for cine CMR imaging synthesis." *under review.*

Acknowledgements

First of all, I must express my deep respect and gratitude to my supervisors Dr. Nishant Ravikumar, Dr. Zeike Taylor, Dr. Avan Suinesiaputra and Professor Alex Frangi. During my research journey, they not only guided me with their professional knowledge, but also strengthened my research beliefs with their unparalleled patience and support. Their unique insights and encouragement are important motivations for me to keep moving forward.

I would also like to thank every member of the laboratory, Ning Bi, Rodrigo Bonazzola, Cynthia Maldonado Garcia, Yash Deo, Haoran Dou, Fengming Lin, Kun Wu, Xiang Chen, Qiongyao Liu and Nurbanu Aksoy, whose daily collaboration and support make my research work smoother. In addition, I would like to thank Kattia Hoz de Vila Eduardo and Alejandro Lebrero for providing technical support. Their help played an indispensable role in the development and conduct of my experiments and work.

I would also like to express my special thanks to my friends in the academic community, especially Zhengji Liu and Jianqing Zheng, with whom I have participated. Their insights and experiences inspired me a lot and added more perspectives to my research.

I can't express my gratitude enough to my family, Ms. Juhua Liu and Mr. Zhiping Cheng. They not only gave me firm support in pursuing my academic dreams, but also provided me with a safe haven of love and encouragement. They have always given me firm support and encouragement, constantly inspiring me to broaden my horizons, inspire creativity, and face challenges bravely. I aspire to future achievements that will make them proud. I would also like to thank my friends. Their understanding and support are the source of strength for me to overcome difficulties and keep moving forward.

Finally, I would like to express my gratitude to all those who directly or indirectly contributed to the completion of this thesis. I cherish meeting and growing together with you during this journey, and I sincerely thank you all for your support and trust.

Abstract

Cardiovascular disease (CVDs) is still the main disease causing many deaths around the world. According to the World Heart Federation's 2023 World Heart Report, approximately 20.5 million deaths in 2021 were attributed to CVDs, accounting for nearly one-third of global fatalities. Over the past few decades, deep learning algorithms have increasingly been applied in magnetic resonance imaging (MRI) in the medical field, and in particular, have become central to the diagnosis and prediction of CVDs. However, the dynamic motion of the heart and its complex and changeable anatomy pose many challenges to the interpretation of cardiac magnetic resonance (CMR) data. Traditional manual analysis methods are time-consuming and provide variable results. At the same time, generative models have advanced medical image analysis, especially for downstream cardiac image analysis tasks. The aim is to use these synthetic images as viable alternatives to real data in deep learning model training, providing cutting-edge solutions in data segmentation, registration, and strain analysis.

This thesis systematically investigated several probabilistic generative models applied specifically to cardiac image analysis, including multi-channel variational autoencoders (VAEs), generative adversarial networks (GANs), and latent diffusion models (LDMs), using cine CMR and tagging CMR images as primary subjects. Cine CMR provides high-resolution dynamic sequences to assess cardiac morphology and myocardial function throughout the cardiac cycle. Tagging CMR enables the quantification of myocardial deformation by encoding spatial modulation patterns into the myocardium. The efficacy of these models is validated through multiple metrics and downstream tasks such as cardiac segmentation and myocardial strain analysis. Initially, we comprehensively reviewed existing deep learning-based image generation techniques in medical image synthesis. Subsequently, we introduced a sparse multi-channel VAE to learn the joint latent representation of cine and tagging CMR images. The proposed model can generate tagging CMR from cine CMR alone, thereby enabling myocardial strain estimation straight from cine CMR images. This represents a novel approach within cardiac imaging research and could potentially replace the conventional clinical use of tagging image sequences as a basis for myocardial motion and

strain analysis. Furthermore, we introduced an innovative framework employing latent denoising diffusion implicit models (DDIM) to synthesise full-spatial cine CMR images. We investigated whether these synthetic images can serve as viable substitutes for real data in downstream cardiac image analysis tasks. Building upon this, we present a novel spatial-temporal generative model that leverages latent DDIM conditioned on demographic and clinical factors, capable of synthesising realistic 4D cardiac cine CMR image sequences.

Overall, the methodologies presented in this research demonstrate potential for innovation and practical applications. The method introduced here may potentially revolutionize traditional clinical diagnosis and intervention methods, and introduce new perspectives on applying deep learning models in medical imaging. These models show promising performance in the generative field, not only promising insights into cardiac conditions, but also advancing the development of personalized medical diagnosis and prediction solutions in the field of cardiology.

Contents

List of Figures

List of Tables

Abbreviations

ED	End-diastolic	ES	end-systolic
EDV	end-diastolic volume	ESV	end- systolic volume
HF	heart failure	SV	Stroke Volume
EF	Ejection Fraction	CO	Cardiac output
HR	Heart Rate	VM	Ventricular mass
STE	Speckle Tracking Echocardiography	CMR	Cardiac magnetic resonance imaging
REF	Regional Ejection Fraction	WMSI	Wall Motion Score Index
WT	Wall thickness	PW	posterior wall
IVS	interventricular septum	LVH	left ventricular hypertrophy
LS	longitudinal strain	RS	radial strain
CS	circumferential strain	SSFP	Steady-State Free Precession
RF	Radio frequency	TR	Repetition Time
TE	Echo Time	SNR	signal-to-noise ratio
T1WI	T1 Weighted Image	FSE	Fast Spin-Echo
T2WI	T2 Weighted Image	SS-SE	Single-Shot Spin-Echo
FGE	fast gradient echo	ECG	electrocardiography
SPAMM	Spatial Modulation of Magnetization	CSPAMM	Complementary SPAMM
HARP	harmonic phase	DENSE	displacement encoding with stimulated echoes
SENC	strain encoding	RBMs	restricted Boltzmann machine
SBNs	sigmoid belief networks	DBNs	deep belief networks
SVMs	support vector machine	CNNs	convolutional neural networks
VAEs	variational autoencoders	GANs	generative adversarial networks
WGAN	Wasserstein GAN	JS	Jensen-Shannon
LSGAN	Least Squares GAN	PGGAN	Progressive GAN
AEs	autoencoders	CVAE	Conditional VAE
DVAE	Denoising VAE	HVAE	Hierarchical VAE

VQ-VAE	Vector Quantized VAE	DDPM	Denoising Diffusion Probabilistic Models
LDMs	Latent Diffusion Models	MSE	mean square error
KL divergence	Kullback-Leibler divergence	ELBO	Evidence Lower Bound
NLL	negative log-likelihood	NLP	natural language processing
Medical VDM	Medical Variation Diffusion Model	LV	left ventricl
RV	right ventricle	FCNs	fully convolutional networks
RNN	recurrent neural network	FT	feature tracking
DXA	dual-energy X-ray absorptiometry	DW	Depthwise
PW	Pointwise	ROI	region of interest
RMSE	Root Mean Square Error	SSIM	Structural Similarity Index Measure
PSNR	Peak Signal-to-Noise Ratio	MAE	Mean Absolute Error
FOV	field of view	HD	Hausdorff distance
DSC	dice similarity coefficient	SAX	short axis

Chapter 1

Introduction

CVDs remains the leading cause of death worldwide. In recent years, the application of deep learning in medical MRI has promoted the diagnosis and prediction of CVDs. However, the complex anatomical structure and dynamic motion characteristics of the heart have brought many challenges to the data analysis of CMR imaging. Traditional manual analysis methods are time-consuming and have large variability in results, while deep learning technology, especially generative models, provide a new breakthrough for medical image analysis. In this chapter, we first introduce the background and motivation of this research, the contributions of this thesis, and the structure of this thesis.

1.1 Background and motivation

CVDs remains a leading factor in global mortality [5][6]. According to the 2023 World Heart Report released by the World Heart Federation (WHF) [7], 20.5 million people passed away due to CVDs in 2021. This fact reminds us of the critical importance of exploring and researching heart information. Over the past few decades, the application of deep learning algorithms to MRI has become crucial for diagnosing and predicting of CVDs [8][9]. However, the dynamic motion of the heart and the complex and varied anatomy of the heart pose many challenges in the interpretation of CMR data. Traditional manual analysis methods are time-consuming and provide varying results.

Meanwhile, deep learning technology, particularly generative models, has significantly advanced medical image analysis, especially for downstream cardiac tasks, and has become a powerful tool in the field. [10]. These synthetic images can be used as viable surrogates for real data in deep learning model training to assist clinical diagnosis and treatment planning, potentially revolutionizing the field of medical images and enabling more accurate and personalized medical care, including synthetic data used to train or create specific understandable generative models [11][12]. It also provides cutting-edge solutions in areas such as data segmentation, registration, and strain analysis [13][14].

However, despite the potential, the complexity of cardiac anatomy and motion makes the interpretability of the generated models challenging. The complexity of cardiac dynamics and anatomical changes requires models that can robustly capture these complexities while maintaining interpretability for clinicians. Traditional methods typically require large volumes of high-quality well-labeled data for training. In cardiac

imaging, there are difficulties in obtaining datasets due to variability in image acquisition protocols and patient populations. Furthermore, models trained on limited or biased datasets may not generalize well across different clinical settings. Furthermore, image generation models like generative adversarial networks (GANs) [15] often face issues such as mode collapse (where the generator produces limited output diversity) and training instability.

The overall goal of this thesis is to propose and develop novel generative model framework based on deep learning to achieve more accurate and robust synthetic cardiac images for cardiac motion modeling and analysis of spatio-temporal MRI data. The authors aimed to use generative models to generate synthetic data and improve diagnostic accuracy and enhance cardiac image analysis.

1.2 Contributions

This thesis aims to explore cardiac image based analysis, using generative models that are reliable, accurate and can be used to assist clinical diagnosis and model training, with advanced deep learning-based methods. To accomplish this task, we study cardiac image synthesis between different sequences for myocardial strain calculation; simultaneous generation of cardiac images and their corresponding masks using latent diffusion models; and spatio-temporal cardiac image generation incorporating clinical demographic information conditions. Has the following main contributions:

- A generative model for predicting myocardial strain using cine CMR: This thesis explores joint latent representations between different cardiac sequences (cine CMR and tagging CMR), using only cine CMR to synthesise tagging CMR and estimate myocardial strain. This framework provides a new perspective for traditional clinical collection of cine images to estimate myocardial motion and strain. It is not limited to image synthesis and can be extended to generate other channel sequence information.

- Efficient, high-quality simultaneous synthesis of cine CMR images and their corresponding biventricular segmentation masks: This thesis proposes a novel pipeline for generating synthetic full-spatial cine CMR images via latent denoising diffusion implicit models (DDIM), synthetic images can be used as the viable surrogate for real dataset in deep learning model training for downstream cardiac

image analysis tasks, and can complement real patient datasets and help reduce the burden of manually annotating images.

- Full spatio-temporal 4D cine CMR images clinical demographic information conditioned generation model: This thesis proposes a conditional latent diffusion generation model to generate full spatio-temporal 4D cardiac images, which can capture full spatio-temporal cardiac motion and anatomical changes. Not only can it be used for the generation of health datasets, but it can also be used to incorporate disease types and condition-specific cardiac atlases. The generated images also can be used for downstream task analysis and deep learning model training data enhancement.

Overall, this thesis improves the accuracy and reliability of generative models for cardiac image motion and anatomy analysis, and helps improve the robustness of clinical diagnosis, enabling more accurate and personalized medical care. In addition, synthetic images can be used to expand and replace the training dataset of medical image analysis models, which can better capture and learn the diversity of data and improve its generalization ability and accuracy.

1.3 Thesis structure

Chapter 2 explains the clinical background of this study and briefly introduces the relevant medical imaging concepts. First, the clinical background describes the anatomical structure and motion mechanism of the cardiac, as well as the functional indicators used to assess cardiac health, with a emphasis on myocardial strain. Next, the medical imaging background section is introduced, focusing on CMR imaging technology, including its theoretical basis, physical principles, and various sequencing technologies used in the image acquisition process with focus on cine MRI and tagging MRI.

Chapter 3 presents a literature review from the perspective of technical and methodological principles. **Section 3.1** introduces generative models and explains their theoretical foundations, with emphasis on variational autoencoders (VAEs) and diffusion models. **Section 3.2** reviews the literature on generative model-based medical image analysis and diagnosis techniques, mainly exploring machine learning and deep learning methods, their applications in real-world technologies and cutting-edge research. **Section 3.3** introduces the datasets used in this paper.

Chapter 4 explores the application of generative models in synthesising cardiac tagging MRI images and estimating myocardial strain. This thesis introduces a sparse multi-channel variational autoencoder (smcVAE) model for jointly learning the latent representations of cine CMR and tagging CMR. By taking cardiac cine MRI images as input, the smcVAE model is able to effectively synthesise cardiac tagging MRI images, which can then be used to quantify myocardial strain. This method effectively overcomes the challenges of limited clinical tagging MRI and achieves reliable and effective myocardial strain analysis using only conventional sequence.

Chapter 5 explores the synthesis of full-spatial cine CMR images and corresponding segmentation masks. This paper proposes a latent DDIM. First, an encoder is used to map the input into the low-dimensional latent space, and a diffusion process is performed in the latent space. After back-diffusion to obtain the reconstructed vector, the decoder is decoded to generate synthetic data that can be used as a viable alternative to real data in deep learning model training. Multiple evaluation methods verify that the model can effectively and efficiently generate 3D cine cardiac images with corresponding segmentation masks.

Chapter 6 extends the basic latent diffusion model and proposes a full spatio-temporal 4D conditional latent diffusion generative model to explore the relationship between clinical factors and demographic information with cardiac imaging anatomy. Non-imaging factors are used as conditions for the generative model to explore their relationship and influence on the cardiac anatomical structure. The model can generate full spatial and temporal 4D cardiac image sequences, showing great potential in real data supplementation and downstream task model training data enhancement.

Chapter 7 concludes this thesis, discusses current limitations, and explores potential future research directions.

Chapter 2

Clinical Background and Medical Imaging

This chapter provides a overview of the clinical foundations and imaging techniques essential for understanding cardiac function and disease diagnosis. Beginning with an exploration of cardiac anatomy and physiological principles, we discuss key functional indices used in clinical assessments, such as ejection fraction (EF), stroke volume (SV), and myocardial strain, highlighting their importance in evaluating cardiac health. The chapter then delves into the role of CMR, emphasizing its advantages over other imaging modalities and detailing the various MRI sequences used in clinical and research settings. Special attention is given to cine MRI and tagging MRI, which are crucial for capturing dynamic cardiac motion and assessing myocardial mechanics. This foundational knowledge sets the stage for the subsequent discussions on generative models for cardiac image synthesis and analysis.

2.1 Cardiac anatomy and clinical principles

The heart is an vital organ in the human circulatory system [16], responsible for pumping blood throughout the body. It ensures that tissues and organs receive the oxygen and nutrients they need while removing metabolic waste. As shown in Figure 2.1, the heart consists of four chambers: left atria, right atria and left ventricles, right ventricles. The right atrium and ventricle collect deoxygenated blood from the systemic veins and direct it to the lungs. Meanwhile, the left atrium and ventricle take in oxygenated blood from the lungs and circulate it through systemic blood vessels, delivering it throughout the body.

The cardiac cycle encompasses the complete process of heart contraction and relaxation. It includes atrial contraction, which directs blood into the ventricles, followed by ventricular contraction, consisting of the isovolumetric contraction and ejection phases. During ejection, the aortic and pulmonary valves open, allowing blood to flow into the aorta and pulmonary artery, respectively. Ventricular diastole includes the isovolumetric relaxation period, the rapid filling period, and the slow filling period. During diastole, the ventricular pressure drops, the tricuspid and mitral valves (atrioventricular valves) open, and blood pumped into the ventricles, initially rapidly, then gradually slowing. The Wigger diagram in Figure 2.2 shows the fluctuations in atrial pressure and volume, as well as in the ventricles and arteries throughout the cardiac cycle.

End-diastolic (ED) and end-systolic (ES) are two key stages in the entire cardiac cycle. ED refers to the state of the ventricles at the end of ventricular relaxation in a

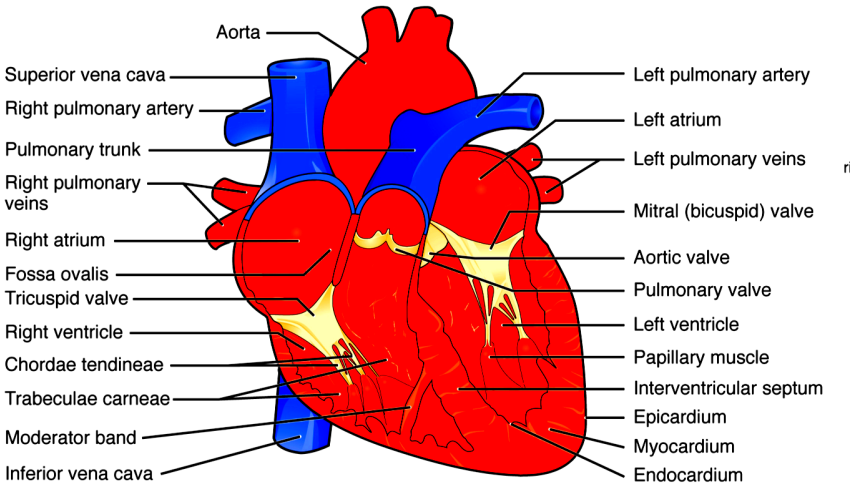

Figure 2.1: **Schematic diagram of the internal anatomy of the heart.** The four chambers of the heart: the left and right ventricles, and the left and right atria, along with the major blood vessels and heart valves. The image is from DeSaix *et al.* [1].

cardiac cycle. At this time, the ventricles are filled with blood and reach their maximum capacity. The ventricular volume is called the end-diastolic volume (EDV). ES refers to the state of the ventricles at the end of contraction during the cardiac cycle. At this time, the amount of blood in the ventricles is the least and the ventricles are in the maximum contraction state. The ventricular volume at this time is called the end-systolic volume (ESV). These parameters and measurements are clinically important for the evaluation of cardiac function and various heart diseases, especially heart failure (HF) and myocardial disease.

2.2 Cardiac functional index

Cardiac function assessment indices [17] are multiple comprehensive indices used in cardiology to assess the overall performance of the cardiac. Clinically, evaluating the overall function of the heart helps diagnose and monitor various cardiovascular diseases. In clinical practice, assessing the overall function of the heart can not only help identify potential pathological changes, but also provide a basis for the formulation of personalized treatment plans. This section mainly introduces the main indices used to

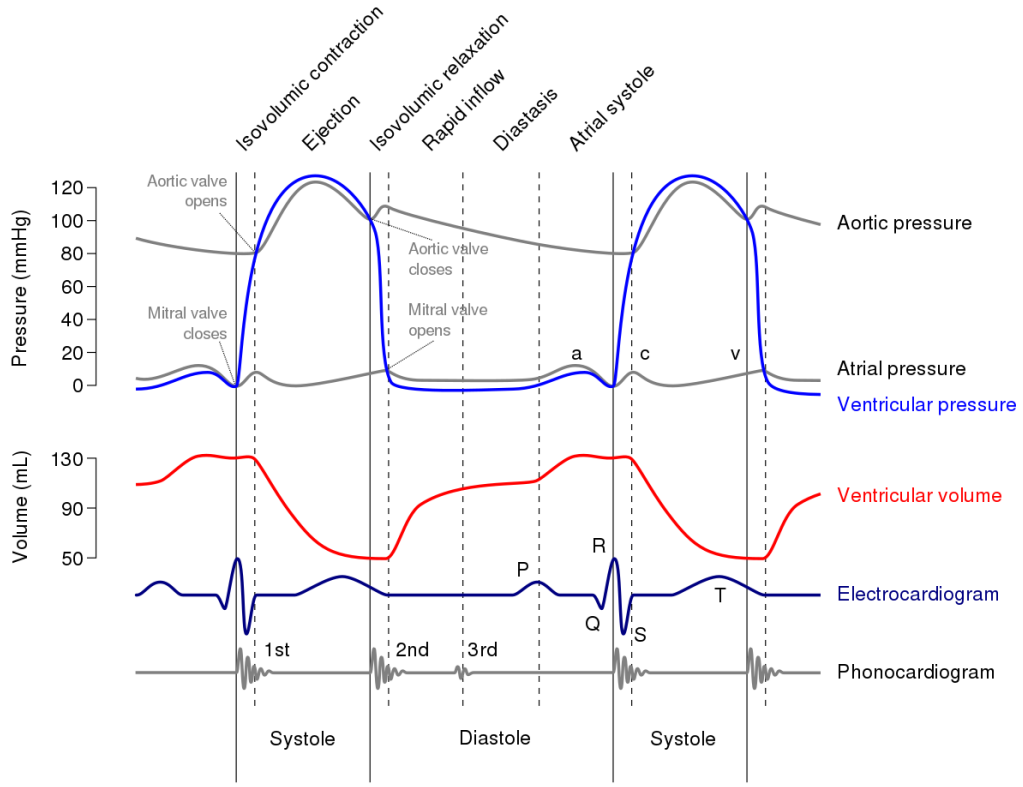

Figure 2.2: **The Wiggers diagram [2] shows the cardiac cycle.** The dotted lines represent different periods and stages of contraction and diastole. The solid traces show the variations in aortic, atrial, and ventricular pressures, and ventricular volume throughout the cardiac cycle, along with the electrocardiogram and phonocardiogram.

assess cardiac function and their physiological significance.

Common cardiac function assessment indicators include EDV, which represents the ventricle's maximum volume at the end of diastole and reflects its filling state. A higher EDV may indicate ventricular dilatation, while a lower EDV may indicate insufficient ventricular filling. ESV represents the minimum ventricular volume at the end of systole, with a normal range of 50-100 mL. A higher ESV may indicate decreased myocardial contractility, while a lower ESV may indicate increased myocardial contractility. SV represents the volume of blood ejected by the LV during each heartbeat, which can be obtained by $SV = EDV - ESV$. Changes in SV can reflect the contractility of the myocardium and the filling state of the ventricle, an increased SV may indicate increased myocardial contractility, while a decreased SV may indicate decreased myocardial contractility or insufficient ventricular filling. EF is a key indicator for evaluating LV function, reflecting the efficiency of the heart pumping blood each time it contracts. The normal EF typically ranges from 50% to 70%, a lower EF usually indicates the presence of HF or cardiomyopathy. The calculation formula is: $EF = (\frac{SV}{EDV}) \times 100\%$. Cardiac output (CO) indicates the total amount of blood pumped by the heart and is an important indicator of its overall pumping function. The normal range is generally 4-8 L/min. Changes in CO can indicate changes in heart function, such as a decrease in cardiac output may indicate HF. The calculation formula is as follows: CO = Heart Rate(HR) $\times$ SV. Ventricular mass (VM) reflects the weight of ventricular muscle, higher ventricular mass indicates ventricular hypertrophy, which may be related to hypertension, cardiomyopathy, etc. In clinical, evaluating VM helps to determine the structural changes and pathological development of the heart.

In addition to these global cardiac function assessment indicators, some regional functional indices are used to assess the function of specific regions or segments of the heart in cardiology. These indicators help detect local myocardial dysfunction, such as myocardial infarction, local ischemia, etc. Figure 2.3 shows the cardiac segment diagram proposed by Cerqueira *et al.* [18], which divides the heart into 17 segments based on the basal segment, mid-segment, apical segment and apex, which is the current general model.

Commonly used regional function indices include regional strain, which is measured by Speckle Tracking Echocardiography (STE) [19] or cardiac magnetic resonance imaging (CMR) [20]. Regional strain can quantify the contraction and relaxation function

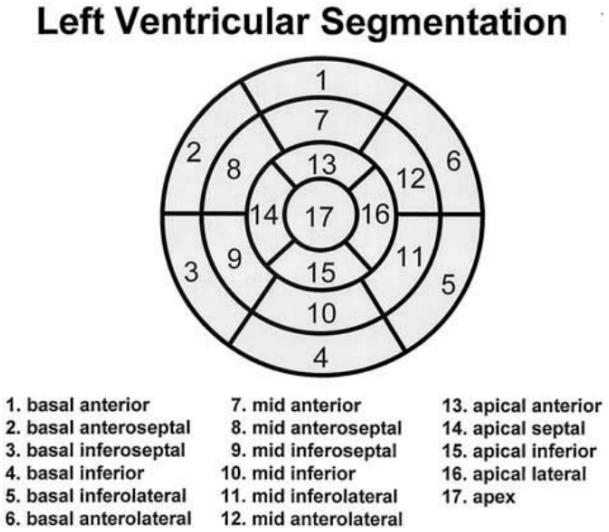

Figure 2.3: **Schematic diagram of the nomenclature of myocardial segments and cardiac tomography.** The heart is divided into 17 segments based on the basal segment, middle segment, apical segment and apex, and each group includes 4 to 8 subsegments, including the anterior, inferior, septal and lateral directions.

of the myocardium and identify myocardial dysfunction at an early stage. Regional Ejection Fraction (REF) is the EF of a specific myocardial segment, which can be used to assess the contractile function of the local myocardium and help identify local myocardial dysfunction. The Wall Motion Score Index (WMSI) calculates the score of the myocardial segment according to its motion status (1-4 points), and the total score is divided by the number of segments assessed. 1 to 4 points represent normal movement, hypokinesia (mild reduction in movement), akinesia (loss of movement), and paradoxical movement (abnormal direction of movement). WMSI provides a quantitative assessment of regional myocardial function, with higher values indicating more severe myocardial damage.

Wall thickness (WT) refers to the thickness of the myocardial wall, which is usually measured in different segments of the heart, including the thickness of the posterior wall (PW) and interventricular septum (IVS). Imaging examinations often measure the thickness during diastole and systole. Increased wall thickness, such as left ventricular hypertrophy (LVH), is common in hypertension, hypertrophic cardiomyopathy, etc. Reduced wall thickness may indicate myocardial disease or myocardial atrophy.

Myocardial strain [21] is also an indicator for evaluating myocardial function, which we will focus on here. Myocardial strain provides more detailed information on the dynamic changes of myocardial deformation than traditional cardiac function evaluation indicators (such as EF). Strain analysis can detect myocardial dysfunction at an early stage before other evaluation indicators, which is helpful for early diagnosis and interventional management of various heart diseases, especially in the early stages of cardiomyopathy and myocardial ischemia. Strain represents the change in the length of the myocardium during contraction and relaxation, usually expressed as a percentage. According to the different directions of myocardial fibers, strain can be divided into longitudinal strain (LS), circumferential strain (CS) and radial strain (RS). Strain is expressed as the percentage change in myocardial length per unit length. The calculation formula is given by the Green-Lagrange strain [22] tensor:

$$\mathbf{E}(t) = \frac{1}{2}\left(\nabla \mathbf{l}(t) + \nabla \mathbf{l}(t)^T + \nabla \mathbf{l}(t)^T \nabla \mathbf{l}(t)\right) \tag{2.1}$$

where $\mathbf{l}(t)$ represents the displacement of the myocardium from the ED phase to the systolic phase. RS and CS are the diagonal components of the tensor $\mathbf{E}$ calculated in cylindrical coordinates.

Figure 2.4 shows strain in three different directions and its changes during cardiac diastole and contraction. LS represents the deformation from base to apex, which is mainly controlled by myocardial fibers in the endocardial layer and is usually used to evaluate the overall function of the LV. Normal LS values are -18% to -22%. Reduced LS indicates decreased myocardial contractility and could be an early indicator of myocardial ischemia, cardiomyopathy, or other heart disease. RS refers to radial myocardial deformation toward the center of the LV cavity, reflecting changes in myocardial thickness and is mainly controlled by the middle myocardial fibers. Normal RS values are 30% to 50%. Reduced RS indicates myocardial hypertrophy or HF. CS is the circumferential deformation of the myocardium around the left ventricle, which is mainly controlled by the myocardial fibers in the epicardial layer. Normal CS values are -20% to -30%. Reduced CS also indicates decreased myocardial contractility.

Strain imaging techniques include STE and CMR. STE uses the natural speckle pattern in ultrasound images to track myocardial motion. It does not rely on Doppler signals [23] and is therefore less affected by angle dependence. CMR is a high-resolution cardiac MRI that can measure myocardial strain very accurately. Notably, myocardial

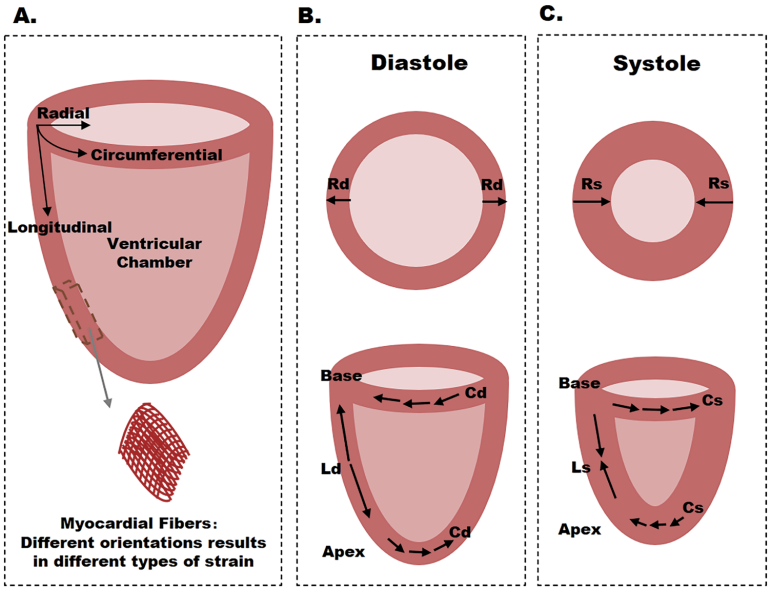

Figure 2.4: **Schematic diagram of myocardial strain.** (a) Long axis view shows myocardial strain in three different directions: radial strain, circumferential strain, and longitudinal strain. (b) Schematic diagram of the changes of the three strains in diastole on short axis and long axis views. (c) Schematic diagram of the changes of the three strains in systole on short axis and long axis views. Cs: systolic circumferential strain, Ld: diastolic longitudinal strain, Ls: systolic longitudinal strain, Rd: diastolic radial strain, Rs: systolic radial strain. This figure is adapted from Zhang *et al.* [3].

tagging MRI [24] is considered the reference standard for measuring myocardial strain and validating other strain measurement techniques. Tagging MRI uses the characteristics of MRI to generate periodic magnetic markers in the myocardium. These markers usually appear in the form of lines or grids and deform with the movement of the myocardium. By analysing the deformation of the markers, the myocardial strain can be accurately calculated. Strain imaging [25] is very effective in assessing the area and extent of myocardial ischemia and myocardial infarction, and can help monitor the progression of the disease and the effectiveness of treatment in patients with HF.

2.3 Cardiac MRI

MRI [26] is a non-invasive imaging technique that utilizes magnetic fields and radiofrequency waves to produce detailed images of the body's internal structures. CMR combines anatomical and functional imaging to provide fine-detailed images of the myocardium and other cardiac anatomy, and can perform qualitative and quantitative assessments of cardiac structure and function. Compared with other imaging techniques, MRI has better contrast for soft tissue than CT [27] and echocardiography [28], and can more clearly distinguish myocardium, fat, blood, and fibrotic tissue without ionizing radiation exposure. MRI includes a variety of imaging sequences that can be used to evaluate a variety of pathological and physiological characteristics of the heart. This section focuses on the various imaging sequences of cardiac MRI, with a particular focus on cine MRI [29] and tagging MRI. These imaging techniques provide rich anatomical and functional information for the heart, are used clinically to evaluate various pathological conditions of the heart, and can be used for early diagnosis and management of cardiovascular diseases.

2.3.1 Steady-State Free Precession (SSFP) sequence

The SSFP sequence [30] uses balanced gradient echo imaging technology to quickly acquire magnetic resonance signals under steady-state conditions and generate high-contrast images. Radio frequency (RF) pulses rapidly and repeatedly excite the SSFP sequence in short intervals. After each pulse excitation, a gradient magnetic field is applied to generate echo signals. In one imaging cycle, the net effect of all gradient pulses is zero, so the transverse magnetization vector maintains a consistent steady state

between multiple excitations. When the steady state is reached after repeated multiple excitations, the length and direction of the magnetization vector remain unchanged, producing a consistent and high-contrast signal.

The SSFP sequence has high contrast and can clearly distinguish the myocardium, blood, and other tissues. Because Repetition Time (TR) and Echo Time (TE) are very short, the SSFP sequence can obtain high-resolution images in a short time and is suitable for dynamic cardiac imaging. And the signal acquired in the steady state has a high signal-to-noise ratio (SNR) and superior image quality. However, the SSFP sequence is sensitive to magnetic field inhomogeneity and requires a long acquisition time for high-resolution imaging of the complete cardiac cycle. Generally, the SSFP sequence is crucial for comprehensive evaluation of cardiac anatomy and function due to its superior temporal and spatial resolution and excellent soft tissue contrast.

The SSFP sequence is extremely widely used in cardiac MRI, and its high temporal and spatial resolution as well as superior soft tissue contrast rendering it essential for assessing cardiac anatomy and function.

2.3.2 T1 Weighted Image (T1WI)

T1-weighted imaging uses differences in longitudinal relaxation times (T1 relaxation times) of different tissues to highlight these differences through specific pulse sequences, thereby obtaining images with high contrast. Commonly used pulse sequences include Inversion Recovery and Fast Spin-Echo (FSE) sequences. T1 relaxation time is shorter in tissues with high fat and protein content, and the image appears as high signal (light), while the T1 relaxation time is longer in liquid tissues such as water and blood, and the image appears as low signal (dark). T1WI is often used in clinical practice to assess myocardial thickness and identify scar and fibrosis areas.

2.3.3 T2 Weighted Image (T2WI)

T2-weighted imaging uses the differences in transverse relaxation times (T2 relaxation times) of different tissues to highlight these differences through specific pulse sequences, thereby obtaining images with high contrast. Commonly used pulse sequences include FSE and Single-Shot Spin-Echo (SS-SE) sequences. In the image performance, water, blood, cerebrospinal fluid, inflammation and edema areas appear as high signals, while fat, high-protein tissue and hemosiderin deposition areas appear as low signals. T2WI is

often used in clinical practice to evaluate lesions such as myocardial edema, myocardial inflammation, and cardiac tumors.

2.3.4 cine MRI

Cardiac cine MRI has the characteristics of no ionizing radiation, arbitrary orientation imaging, good soft tissue contrast, and it can also show the rhythmic contraction and relaxation process of the cardiac in the form of continuous frames, also often used for the observation of clinical ventricles and myocardial structures and functional evaluation. The acquisition imaging sequences currently used in clinical practice are mainly fast gradient echo (FGE) sequences or SSFP sequences. During the scanning process, electrocardiography (ECG) gating [31] is required, and the patient need holding their breath multiple times to reduce the impact of heart beats and respiratory movements. The process triggers image acquisition at the R wave of the ECG signal [32] to ensure that each frame of the image is acquired at the same phase of the cardiac cycle. The image at each time point is called a frame, and cardiac cine MRI usually contains 15-50 frames of images, encompassing the full cardiac cycle. After repeated acquisition of multiple time points at multiple slice positions, a $4D(3D+t)$ data structure is formed, and the slice thickness is usually between 6-8 mm.

Cardiac cine MRI offers superior temporal and spatial resolution, enabling it to clearly capture the heart's rapid movements throughout the entire cardiac cycle. It provides detailed cardiac anatomy and records heart movements during contraction and relaxation. Cardiac cine MRI is a powerful tool for evaluating cardiology, it is the most widely used imaging sequence in clinical and scientific research, and has important value [33]. It can accurately assess cardiac and valvular function and quantitatively measure cardiac chamber size and myocardial morphology, which are essential for the early detection and treatment of cardiovascular disease.

2.3.5 tagging MRI

Cardiac cine MRI is currently the most commonly used imaging technique for clinical acquisition, allowing for the observation of global myocardial motion and quantitative measurement of myocardial motion and cardiac structure. However, many cardiovascular diseases may occur before any significant changes in traditional cardiac function indicators occur, and the physiological state of the myocardium may have already

changed. Compared to a variety of other cardiac imaging methods, such as echocardiography, CT, radionuclide single photon emission CT (SPECT) [34] and positron emission tomography (PET) [35], as well as older projection methods of cardiac angiography [36].

Although these methods can assess the tomographic and even 3D motion of the endocardial and epicardial borders of the heart, and the heart valves. However, none of these methods can track the motion of discrete material points within the myocardium. Because the heart passes through any imaging plane and rotates in it during the cardiac cycle, there is a complex relationship between the apparent motion of the endocardial and epicardial borders in tomographic imaging and the contraction of the myocardial wall and the resulting intramyocardial motion and deformation. Therefore, prior to the advent of MRI myocardial tagging, there were no reliable non-invasive methods to assess true myocardial contraction and relaxation. The practical impact of these limitations has been to restrict the understanding of normal human myocardial contraction, as well as the understanding of normal cardiac physiology, and the assessment of myocardial mechanical function, as the relationship between ventricular dynamics and true intramural contraction is altered in conditions such as hypertrophic states and ischemic heart disease.

Cardiac tagging MRI [37] is a specialized imaging technique that allows assessment of regional myocardial function, measuring intramyocardial motion parameters such as strain [38], strain rate [39], torsion [40], and rotation [41]. These parameters allow observation of mechanical behavior of the myocardium that may not be captured in traditional global cardiac function measurements. The principle of cardiac tagging MRI is the application of a magnetization preparation pulse prior to cine imaging. This pulse alters the direction and magnitude of the myocardial magnetization vector under the influence of a spatially varying magnetic field gradient. As a result, the myocardium is ”tagged” with a spatially defined pattern of markers that are visible in the resulting MR images. These visual markers, which usually appear as periodic lines or grids, act as tracers that move as the myocardium deforms during the cardiac cycle. By tracking the motion of these markers, regional myocardial motion and deformation can be quantified, providing a direct measure of myocardial function.

In common techniques for generating these markers, such as spatial magnetization modulation (SPAMM) [42] and complementary SPAMM (CSPAMM) [43], a series of

RF pulses are applied to the myocardium. These RF pulses create a spatially localized, periodic pattern of magnetization within the myocardium. In the case of SPAMM, the pattern consists of regularly spaced lines, while CSPAMM improves on SPAMM by enhancing the contrast and persistence of the markers, allowing for more reliable tracking. Marked areas appear darker in MRI images due to saturation of the magnetization in these areas, while surrounding tissue is unaffected.

Other techniques, such as harmonic phase (HARP) [44], strain encoding (SENC) [45], and stimulated echo displacement encoding (DENSE) [46], are used to further improve the quality of the markers and the accuracy of regional motion measurements. For example, HARP can detect small changes in myocardial deformation, making it particularly useful for studying strain and torsion. SENC and DENSE offer improved spatial resolution and SNR, as well as the ability to capture 3D volumetric motion. These techniques also reduce the need for long acquisitions and expand the range of anatomy that can be imaged.

Figure 2.5 illustrates the process of tagging MRI data acquisition, which involves perturbations of the myocardial magnetization to create visible markers. These markers move with the tissue, providing a dynamic representation of myocardial motion during the cardiac cycle. The original concept of myocardial tissue tagging was first proposed by [47] and has evolved over time to provide increasingly accurate and high-resolution measurements of myocardial function.

The tagging MRI sequence typically involves two phases: a tagging preparation phase and an imaging phase. During the tagging preparation phase, RF pulses are applied perpendicular to the imaging plane, which alters the longitudinal magnetization of the myocardium in the specific slice that intersects the RF pulse. The surrounding tissue is unaffected, resulting in a unique spatial pattern of the tagged region. During the imaging phase, the tagged region appears darker than untagged tissue due to saturation of its magnetization. These markers are intrinsically linked to the tissue and move with it, allowing detailed tracking of myocardial deformation throughout the cardiac cycle. By combining tagging with cine MRI, myocardial tagging patterns can be continuously monitored from ED to ES of the cardiac cycle, providing real-time observation of regional cardiac mechanics.

Figure 2.5 illustrates the process of labeled MRI data acquisition, which involves perturbations to the magnetization of the myocardium to create visible markers. These

markers move with the tissue, providing a dynamic representation of myocardial motion during the cardiac cycle. The original concept of myocardial tissue labeling was first proposed by Zerhouni *et al.* [47] and has evolved over time to provide increasingly accurate and high-resolution measurements of myocardial function.

A tagging MRI sequence typically involves two phases: a tagging preparation phase and an imaging phase. During the tagging preparation phase, RF pulses are applied perpendicular to the imaging plane, which alters the longitudinal magnetization of the myocardium in the specific slice that intersects the RF pulse. The surrounding tissue is unaffected, resulting in a unique spatial pattern of the tagged region. During the imaging phase, the tagged region appears darker than untagged tissue due to saturation of its magnetization. These markers are intrinsically linked to the tissue and move with it, allowing detailed tracking of myocardial deformation throughout the cardiac cycle. By combining tagging with cine MRI, myocardial tagging patterns can be continuously monitored from ED to ES of the cardiac cycle, providing real-time insights into regional cardiac mechanics.

Although tagging MRI can provide important information about myocardial function, the additional RF pulses required to create the tagging limit its clinical use. The need for multiple RF pulses increases acquisition time and reduces the efficiency of the imaging process, which is particularly problematic in clinical settings where rapid imaging is often required. Despite these challenges, tagging MRI remains an important sequence for studying myocardial motion, and current research is focused on improving its efficiency and applicability in clinical practice.

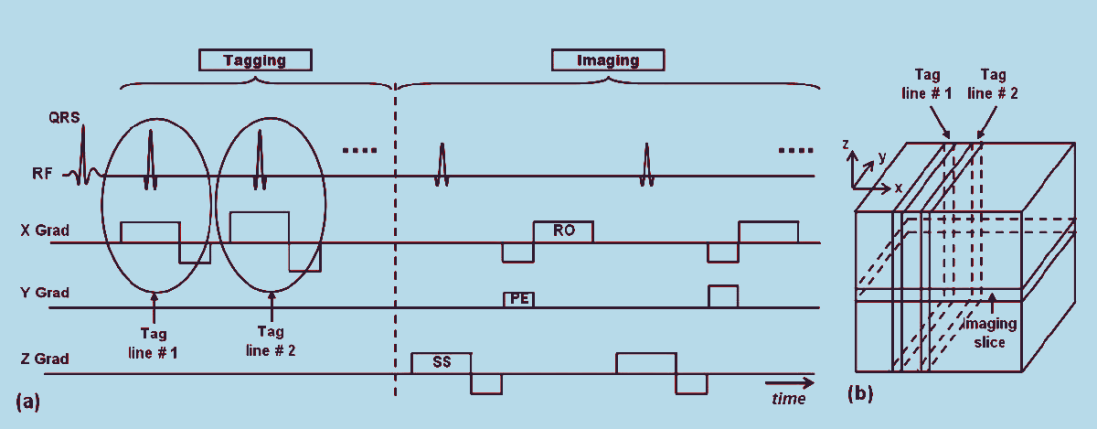

Figure 2.5: **Schematic diagram of tagging data acquisition.** (a) shows the two stages of the acquisition process: tagging preparation and imaging. Each tagged plane requires a slice-selective RF pulse during the tagging stage, immediately followed by the imaging sequence. (b) the relationship between tagging planes and imaging slices. This figure is adapted from Ibrahim *et al.* [4].

Chapter 3

Literature Review and Algorithm Theory

This chapter reviews the existing literature and theoretical foundations related to deep learning and generative models in medical imaging. It first provides an overview of generative models, including autoencoder (AE), variational autoencoder (VAE), generative adversarial network (GAN), and diffusion model, and introduces their algorithmic foundations, applications in medical image analysis, advantages, and limitations. This chapter further explores some recent studies in cardiac image synthesis, segmentation, registration and dynamic analysis, and introduces their applications in cardiac imaging. In addition, the dataset used in this thesis are introduced.

3.1 Deep learning and generative models

Deep generative models [48], a key category of deep learning, have garnered growing attention and significant research interest. In recent years, many different deep generative models have been proposed, establishing them as a leading research area in artificial intelligence [49][50]. Generative models model data distribution through probability density functions, which helps to understand the generation mechanism of complex data and learn high-level feature expressions. Therefore, they are widely used in tasks such as image generation and restoration, data augmentation, medical diagnosis, and virtual reality.

Boltzmann machines [51] and restricted Boltzmann machines (RBMs) [52] were the earliest generative models. RBMs have been employed in dimensionality reduction, classification, and feature learning, and have also proven effective as the initial stage of deep neural networks [53]. However, RBMs face significant challenges, such as difficulties in training due to the vanishing gradient problem and the need for extensive computational resources. Then in the 1990s, Helmholtz machines [54] and sigmoid belief networks (SBNs) appeared [55]. Helmholtz machines introduced a framework for unsupervised learning of latent variable models, but their complex training algorithms and computational inefficiency limited their practical applications. SBNs, while innovative, struggled with issues related to slow convergence and difficulty in scaling to larger datasets. In 2006, Hinton introduced deep belief networks (DBNs) [56], among the earliest non-convolutional models to be effectively trained within deep architectures. DBNs demonstrated impressive performance, outperforming kernelized support vector machines (SVMs) on the MNIST dataset. Despite their early success, DBNs have mostly lost favor in recent years, largely due to the rise of more efficient and scal-

able models like convolutional neural networks (CNNs) and VAEs. DBNs also suffer from challenges such as slow training times and difficulties in fine-tuning deep layers. While DBNs and other early generative models have been overshadowed by more advanced techniques, their contributions to the development of deep learning should not be underestimated. They laid the groundwork for the evolution of more sophisticated models and highlighted the potential of deep architectures, despite their limitations and the emergence of more effective algorithms. In summary, although these model have their own limitations, such as training difficulties, computational inefficiency, and scaling issues, their historical significance in the field of deep learning remains pivotal. These early models provided essential insights and techniques that have influenced the design and optimization of modern generative models.

Goodfellow *et al.* [57] proposed GAN, marking an important progress in generative models. Since then, numerous advancements in GANs have emerged, such as Wasserstein GAN (WGAN) [58], which enhances training stability by using the Earth-Mover distance rather than the Jensen-Shannon (JS) divergence to quantify the difference between generated and real data distributions. However, WGAN is sensitive to the choice of its hyperparameters, and its training can be unstable if not properly tuned. Least Squares GAN (LSGAN) [59], which improves image quality and accelerates convergence by minimizing the least squares error between generated images and real images instead of the traditional cross entropy loss. Despite these improvements, LSGAN may still suffer from mode collapse, causing the generator to produce a narrow range of outputs. Progressive GAN (PGGAN) [60], which improves the quality of generated images by gradually increasing the network depth of the generator and discriminator to gradually generate higher resolution images. This method, however, requires considerable computational resources and training time, making it less practical for real-time or resource-constrained applications. StyleGAN [61] introduces style transfer technology, which generates more diverse and higher quality images by controlling different layers of generated images. Although StyleGAN delivers state-of-the-art image quality, it is computationally intensive and requires extensive hyperparameter tuning to achieve optimal performance. CycleGAN [62] achieves image-to-image translation tasks by introducing cycle consistency loss, eliminating the need for paired datasets. Although CycleGAN is highly effective for tasks where paired data is scarce, it may struggle with generating high-fidelity images in cases where the domain shift between source and

target images is significant.

These improvements improve the performance and stability of GANs, and make important contributions to the quality and diversity of generated images. However, each method also introduces its own set of challenges, such as increased computational requirements, sensitivity to hyperparameters, and potential for mode collapse.

At the same time, the expansion of AE into generative models has also begun to attract widespread attention and in-depth research. β-VAE [63] introduces a hyperparameter β to balance the trade-off between image reconstruction and latent variable distribution, enhancing decoupling and interpretability. However, the introduction of β can lead to challenges in balancing reconstruction quality and disentanglement, potentially resulting in poorer reconstructions if not tuned properly. Conditional VAE (CVAE) [64] introduces conditional variables into VAE, enabling the model to generate specific types of images or data based on conditional information, thereby improving the controllability and diversity of generated data. Despite these benefits, CVAE's performance heavily relies on the availability and quality of the conditional information, which may not always be accessible or easy to define. Denoising VAE (DVAE) [65] introduces the idea of denoising autoencoder, allowing the model to learn data representations in the presence of noise, thus enhancing the robustness of generated images. However, while DVAE enhances robustness, it can sometimes struggle with retaining fine details in the generated images due to the added noise during training. Hierarchical VAE (HVAE) [66] introduces a hierarchical latent variable structure, which enables the model to capture the complex characteristics of data at different levels, thus enhancing the quality of generated images and the model's expressiveness. The complexity of HVAE, however, can lead to increased computational costs and training difficulties, making it challenging to scale and optimize. Vector Quantized VAE (VQ-VAE) [67] introduces vector quantization technology, which discretizes the continuous latent space and combines it with convolutional neural networks to achieve the generation of high-quality images and sequence data. VQ-VAE avoids the complexity of KL divergence and can capture the discrete structural characteristics of data. Nevertheless, the discretization process can introduce quantization errors, which may affect the overall quality of the generated data.

In summary, while these advancements in VAE variants enhance the capabilities of generative models, each method has its limitations. Balancing reconstruction qual-

ity and interpretability, managing conditional dependencies, maintaining robustness without losing detail, handling computational complexity, and mitigating quantization errors are ongoing challenges that require further research and optimization.

In 2015, Jascha *et al.* proposed the diffusion model [68] as a generative model, which greatly improved the generation effect. Since then, the diffusion model has also become one of the hot topics in the field of generative model research. Denoising Diffusion Probabilistic Models (DDPM) [69] applies the diffusion process to the data generation task, generates data by gradually adding and removing noise to learn the data distribution, and improves the quality and diversity of generated data. Improved DDPM [70] improves DDPM, improving the quality and sampling efficiency of generated images through better noise scheduling and loss function design. Latent Diffusion Models (LDM) [71] greatly reduces the computational cost while maintaining high-quality data generation by applying the diffusion process in latent space rather than data space. Guided Diffusion Models [72] introduces guidance signals (such as labels or other conditions) to guide the diffusion process, thereby achieving conditional generation and improving the controllability and diversity of generated data.

Although these improvements have greatly improved the generation effect, each method still has some shortcomings. DDPM could improves the quality and diversity of generated data, but its sampling process is slow and requires more computing resources. While the improved DDPM [70] partially addresses this issue, it still encounters bottlenecks when generating extremely high-resolution images. LDM lowers computational costs by applying the diffusion process in latent space, but this method relies heavily on model pre-training and the choice of latent space, which may limit its effect in some application scenarios. Guided Diffusion Models achieve conditional generation by introducing guidance signals, but it is still a challenge to design effective guidance signals and model structures when high-precision control of the generation results is required. In addition, although these methods perform well in various generation tasks, they still face difficulties in dealing with ultra-large-scale datasets and real-time applications.

In general, although diffusion models have made significant progress in generation tasks, further research is still needed to address their limitations and explore more efficient and accurate generation methods.

In this section, several widely used generative model principles are introduced in detail.

3.1.1 Autoencoders (AEs)

In deep learning, AEs are unsupervised neural network models used for representation learning [73]. The typical AEs are composed of the encoder and the decoder. The encoder transforms the input x to the hidden representation z, while the decoder reconstructs z back into the original data space, producing the output $\widehat{x}$. The encoding and decoding processes can be mathematically represented as:

$$z = f_\theta(x) \tag{3.1}$$

$$\widehat{x} = g_\phi(z) \tag{3.2}$$

Therefore, AEs are often used for data denoising, dimensionality reduction, and feature extraction. The training goal of the AEs is to reduce the difference between the input x and its reconstruction. Common loss functions include mean square error (MSE):

$$L(x, \widehat{x}) = ||x - \widehat{x}||^2 = ||x - g_\phi(f_\theta(x))||^2 \tag{3.3}$$

Specifically, we hope to minimize the following loss function by adjusting the parameters θ and ϕ of the encoder and decoder:

$$\mathcal{L}\theta, \phi) = \frac{1}{n}\sum_{i=1}^{n} ||x_i - g_\phi(f_\theta(x_i))||^2 \tag{3.4}$$

where n represents the number of training samples, and x_i denotes the i-th sample.

To enhance the model's generalization ability, regularizing the AEs is often required. Common regularization methods include adding sparsity constraints, such as $L1$ regularization terms:

$$\mathcal{L}\theta, \phi) = \frac{1}{n}\sum_{i=1}^{n} ||x_i - g_\phi(f_\theta(x_i))||^2 + \lambda||z||_1 \tag{3.5}$$

λ is the regularization parameter.

Some regularization methods introduce noise to the input data and train the AEs to denoise it.

$$\tilde{x} = x + n \tag{3.6}$$

Here, n represents the noise vector, then loss function becomes:

$$\mathcal{L}(\theta, \phi) = \frac{1}{n} \sum_{i=1}^{n} ||x_i - g_\phi(f_\theta(\tilde{x_i}))||^2 \tag{3.7}$$

Due to its potential in compression and feature extraction, AEs have been extensively applied across various areas of medical image analysis. From data preprocessing to high-level disease diagnosis and anomaly detection, AEs provide powerful tools and methods for the medical imaging field. Jan *et al.* [74] introduced various applications of AEs in medical image analysis, including image denoising, reconstruction, modality conversion, classification, segmentation, and anomaly detection. They also emphasized the flexibility and adaptability of these models in dealing with complex medical image data. Amina *et al.* [75] explored a medical image compression method based on deep convolutional AEs. By improving the accuracy of image compression and reconstruction, the method showed advantages in a variety of medical image applications. Bing *et al.* [76] introduced a multimodal collaborative learning method based on AEs, which demonstrated the power of AEs when dealing with multimodal medical image data.

However, AEs are deterministic, with training primarily aimed at minimizing reconstruction errors and encoding data into latent representations in a fixed manner, which restricts the model's generative capabilities.

As an extension of the AEs, the VAEs integrate deep learning with Bayesian statistical inference [77], aiming to learn a robust probabilistic model that explains the data under analysis. This is described in detail in the next subsection.

3.1.2 Variational autoencoders (VAEs)

The objective of the VAEs is to learn a generative model by mapping input data to a latent space and then reconstructing it. Unlike traditional AEs, VAEs encode input data as a probabilistic distribution in the latent space, rather than a fixed code. This enables VAEs to generate new samples by drawing from the latent space distribution.

Encoder

In the VAEs, the encoder maps the input x to the parameters of a probability distribution in latent space, specifically producing two vectors: the mean vector μ and the logarithmic variance vector $log\sigma^2$, as follows:

$$\mu = f_\mu(x; \theta) \tag{3.8}$$

$$\log \sigma^2 = f_{\log \sigma^2}(x; \theta) \tag{3.9}$$

where f_μ and $f_{log\sigma^2}$ are neural networks parameterized by θ.

Latent Variable Sampling

Once the mean and variance vectors are obtained, a point z can be drawn from this distribution. The reparameterization trick [78] is introduced to make this process differentiable, allowing backpropagation during training. The sampled latent variable z is subsequently calculated as:

$$z = \mu + \sigma \odot \epsilon, \epsilon \sim \mathcal{N}(0, I) \tag{3.10}$$

$\odot$ represents element-wise multiplication, and ϵ is drawn from a standard normal distribution.

In VAEs, the latent variable z is assumed to follow a standard normal prior distribution $p(z)$. This prior allows the model to cover diverse data distributions and generate varied samples through learning. Additionally, the probability density function of the standard normal distribution is simple, making it easy to compute and derive.

Posterior Distribution

The posterior distribution $q_\phi(z|x)$ represents the latent variable z given observation data x. According to the Bayesian theorem, the posterior distribution is expressed as:

$$q_\phi(z|x) \approx p(z|x) = \frac{p(x|z)p(z)}{p(x)} \tag{3.11}$$

Here, $p(x|z)$ is the likelihood function, which represents the probability of observing x conditioned on the latent variable z. $p(z)$ is the prior distribution of z, and $p(x)$ is the marginal likelihood, or evidence, representing the total probability of data x.

Variational Inference

In actual calculations, it is difficult to directly solve the posterior distribution because it is usually not feasible to calculate the marginal likelihood $p(x)$. To overcome this problem, VAEs use the variational inference method to approximate the true posterior distribution $p(z|x)$ through a parameterized distribution $q_\phi(z|x)$. The goal of variational inference is to approximate $q_\phi(z|x)$ that is as close as possible to the true posterior distribution $p(z|x)$. Specifically, VAEs achieve this goal by reducing the Kullback-Leibler (KL) divergence:

$$D_{KL}(q_\phi(z|x)||p(z|x)) \tag{3.12}$$

Since it is not feasible to directly minimize this KL divergence, VAEs introduce the Evidence Lower Bound (ELBO) for model optimization:

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{q_\phi(z|x)}[logp_\theta(x|z)] - D_{KL}(q_\phi(z|x)||p(z)) \tag{3.13}$$

The optimization goal becomes to maximize ELBO, thereby indirectly minimizing the KL divergence.

Decoder

The decoder $p_\theta(x|z)$ transforms the latent variable z back to the input data space, producing the reconstructed data $\widehat{x}$.

In VAEs, the loss function plays a critical role and consists of two components: reconstruction loss and regularization loss, as illustrated in equation 3.13. The term $\mathbb{E}_{q_\phi(z|x)}[logp_\theta(x|z)]$ represents the reconstruction loss, which is the expected log probability of generating the input x from the latent variable z. The regularization loss, $D_{KL}(q_\phi(z|x)||p(z))$, captures the KL divergence between the posterior $q_\phi(z|x)$ and the prior $p(z)$. The reconstruction loss measures how closely the generated data matches the original, typically quantified by the negative log-likelihood (NLL):

$$\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] = \frac{1}{N}\sum_{i=1}^{N} \log p_\theta(x^{(i)}|z^{(i)}) \tag{3.14}$$

where N is the batch size, $x^{(i)}$ represents the i-th input data sample, and $z^{(i)}$ is the latent variable drawn from the posterior distribution $q_\phi(z|x)$.

The regularization loss term $D_{KL}(q_\phi(z|x)||p(z))$, measures the KL divergence between the posterior and prior distributions. In VAEs, the prior $p(z)$ is typically assumed to follow a standard normal distribution $\mathcal{N}(0, I)$, while the posterior is modeled as a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$. The KL divergence can be expressed as:

$$D_{KL}(q_\phi(z|x)||p(z)) = \frac{1}{2}\sum_{j=1}^{d}\left(1 + \log\left(\sigma_j^2\right) - \mu_j^2 - \sigma_j^2\right) \tag{3.15}$$

Here d represents the dimensionality of the latent variable z, with μ_j and σ_j^2 denoting the mean and variance of the j-th dimension, respectively.

By combining these two loss terms, the VAEs loss function (the objective to be minimized) is expressed as:

$$\mathcal{L}(\theta, \phi) = \frac{1}{N} \sum_{i=1}^{N} \Big(- \log p_\theta(x^{(i)}|z^{(i)}) \Big) + \frac{1}{2} \sum_{j=1}^{d} \Big(1 + \log\Big(\sigma_j^2\Big) - \mu_j^2 - \sigma_j^2 \Big) \tag{3.16}$$

VAEs have found widespread application in medical image analysis. Attri-VAE[79] creates interpretable medical image representations by disentangling different attributes in the latent space. It uses attribute-based regularization to encode specific attributes along specified dimensions of the latent space, thereby enhancing the interpretability and usefulness of the generated images for tasks such as volume estimation and generating attention maps for specific features in medical images. Mahmoud *et al.*[80] used VAE to generate synthetic eye tracking data that can be used to enhance limited datasets in medical research. This approach demonstrated the potential of VAE to generate reasonable outputs from small datasets and can be used as a data augmentation mechanism to improve performance on classification tasks. Qingyu *et al.*[81] applied the VAE model to brain aging analysis, aiming to learn the latent space of neuroimaging data and perform supervised regression. The disentanglement in the latent representation can intuitively explain the structural developmental patterns of the brain, which makes it a powerful tool for analysing complex medical images and understanding the latent patterns associated with aging.

VAEs have multiple advantages in generating data and learning latent representations, but they also have some limitations. First, because VAEs use a Gaussian distribution to model the latent space, the resulting image is the average of several possible outputs, resulting in a lack of clarity and detail, resulting in generated images often appearing blurry. And due to the need to strike a balance between reconstruction loss and KL divergence, training VAEs can be unstable and may result in poor performance in reconstructing input data or generating realistic samples from the latent space.

3.1.3 Generative Adversarial Networks (GANs)

GANs address the limitations of VAEs and are renowned for producing clearer and more realistic images. GANs do not impose a specific distribution on the latent space. Instead, the generator learns to map simple distributions (such as uniform distributions

or Gaussian distributions) to complex data distributions, which is able to capture more complex structures in the data.

GAN was proposed by Goodfellow *et al.* [57]. It generates data through a confrontation process between two neural networks. GAN consists of two networks: the generator and the discriminator. The generator aims to create realistic data, while the discriminator's role is to differentiate between real and generated data. Through continuous confrontation, the two networks improve each other's performance, ultimately enabling the generator to produce samples nearly indistinguishable from real data.

The goal of GAN training is to shape the generator's distribution so that it closely matches the real data distribution. Let G represent the generator and D represent the discriminator. The generator G takes a random noise vector z (typically sampled from a simple distribution, such as Gaussian or uniform) and generates data $G(z)$. The discriminator D receives a data sample (either real data x or generated data $G(z)$) and outputs a probability $D(x)$ or $D(G(z))$, indicating how likely the sample is real.

The discriminator D aims to maximize the probability of correctly distinguishing between real and generated data. The objective function is:

$$\max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))] \tag{3.17}$$

Here, $\mathbb{E}_{x\sim p_{\text{data}}(x)}[.]$ denotes the expectation over the true data distribution $p_{data}(x)$, while $\mathbb{E}_{z\sim p_z(z)}[.]$ denotes the expectation over the noise distribution $p_z(z)$. The generator G aims to minimize the discriminator's accuracy in recognizing generated data, thereby maximizing $D(G(z))$:

$$\min_G V(D,G) = \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))] \tag{3.18}$$

A more commonly used objective for the generator is to maximize $log(D(G(z))$ rather than as it provides better gradient behavior during the initial stages of training. The final objective function can be expressed as:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))] \tag{3.19}$$

The objective functions for both the discriminator and generator follow the form of cross-entropy loss. The discriminator aims to minimize the cross-entropy of classification errors, while the generator seeks to maximize the cross-entropy, encouraging the generated data to be identified as real.

The GANs optimization process can be regarded as minimizing the JS divergence between the real and generated data distributions. Specifically, the combined objective function of the discriminator and generator is expressed as:

$$V(D,G) = -\log 4 + 2 \cdot \mathrm{JSD}(p_{\mathrm{data}} \| p_g) \tag{3.20}$$

where p_{data} denotes the true data distribution, p_g represents the generator's distribution, and $\mathrm{JSD}(p_{\mathrm{data}} \| p_g)$ represents the JS divergence.

GANs are widely used to generate high-quality medical images to enhance the diversity of training datasets. Maayan *et al.* [82] used GAN-based synthetic images for data augmentation to improve the performance of CNNs in liver lesion classification. Zhang *et al.* [83] proposed SkrGAN for sketch rendering generation of medical images, aiming to augment datasets and enhance model training. This approach demonstrated improved performance in medical image segmentation. Jin *et al.* [84] used 3D conditional GAN to simulate real lung nodules on CT to expand the dataset and achieve enhanced segmentation of lung images. GANs have also made significant progress in medical image segmentation. Rezaei *et al.* [85] employed conditional adversarial training for brain tumor semantic segmentation, achieving promising results. Liu *et al.* [86] applied conditional GAN for automatic cartilage and meniscus segmentation of knee MRI, demonstrating the potential of GANs in complex medical image segmentation tasks.

Although GAN has achieved remarkable success in many fields, it also has some significant shortcomings and challenges. GAN training is often unstable, and adversarial learning between the generator and the discriminator can result in mode collapse, where the generator produces variations of a limited subset of samples rather than covering the full data distribution, reducing the diversity of generated outputs. Therefore, the training process requires careful adjustment of the hyperparameters and selection of the architecture.e architecture. Currently, the evaluation of GAN is usually subjective and lacks a unified objective evaluation standard. Therefore, assessing both the quality and diversity of the generated images remains a challenge.

3.1.4 Diffusion models

The diffusion models are recently widely used generative models that generate high-quality samples by simulating data distribution to gradually "diffuse" from noise. Some

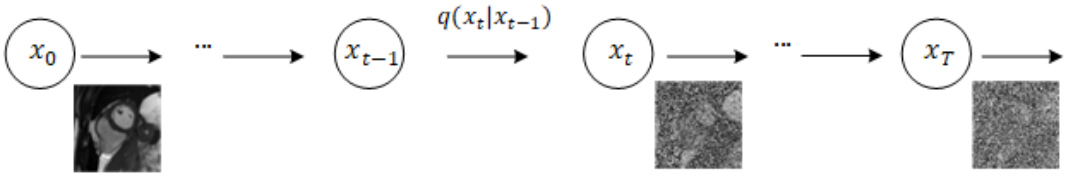

Figure 3.1: **Schematic diagram of the Markov chain in the forward diffusion process.**

studies have shown that diffusion models beat GANs in image synthesis and can achieve better sample quality than state-of-the-art GANs [72][87]. It includes two stages: the forward diffusion process and the reverse diffusion process. In the forward diffusion process, noise is incrementally added to the real data until it becomes pure Gaussian noise. This process can be expressed as a series of Markov chains, in which Gaussian noise is introduced at each step to obtain an approximate posterior $q(x_{1:T}|x_o)$. The forward diffusion process is formulated as:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \tag{3.21}$$

where β_t is a small positive value, indicating the noise intensity introduced at each step. Figure 3.1 shows the embodiment of the Markov chain in the cardiac image data during the process.

The reverse diffusion process gradually removes noise by learning the reverse process and recovers from pure noise to the original data. The training goal is to train a parameterized model that approximates the reverse process. The reverse diffusion process is expressed as:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)) \tag{3.22}$$

Here, μ_θ and Σ_θ are parameters need to be learned.

Figure 3.2 shows the directed graphical model of the reverse diffusion process. The time-dependent parameters of the Gaussian transitions are learnable, and the distribution for each reverse diffusion step depends only on the previous or next time step.

The training aims to find the inverse Markov transitions that maximize the likelihood of the data, which is equivalent to minimizing the variational upper bound of the NLL.

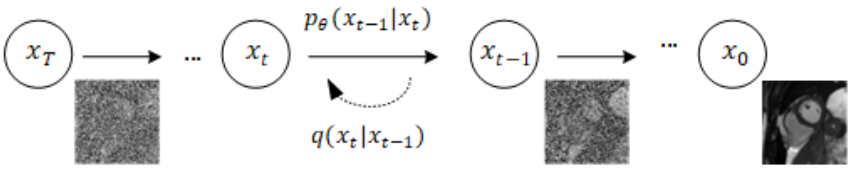

Figure 3.2: **Schematic diagram of the directed graphical model of the reverse diffusion process.**

$$\mathbb{E}[-\log p_\theta(x_0)] \leq \mathbb{E}_q[-\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}] =: L_{vlb} \tag{3.23}$$

KL divergence, an asymmetric statistical distance metric, measures the difference between a probability distribution P and a reference distribution Q. The transition distribution in the Markov chain is a Gaussian distribution, and the variational lower bound L_{VLB} is rewritten using the KL divergence:

$$L_{vlb} = L_0 + L_1 + ... + L_{T-1} + L_T \tag{3.24}$$

where $L_0 = -\log p_\theta(x_0|x_1)$, $L_{t-1} = D_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t))$, $L_T = D_{KL}(q(x_T|x_o)||p(x_T)))$.

In the forward process, the variance schedule needs to be defined, which is often set to a time-dependent constant, such as a linear schedule from $\beta_1 = 10^{-4}$ to $\beta_T = 0.2$.

The overall training objective function is:

$$L_{\text{simple}}(\theta) := \mathbb{E}_{t,\mathbf{x}_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\|^2\right] \tag{3.25}$$

Here, ϵ is Gaussian noise, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. ϵ_θ is a noise estimator parameterized by the model. α_t represents the weight coefficient at time step t, while $\sqrt{1-\bar{\alpha}_t}$ adjusts the noise proportion at the same time step.

The training and sampling algorithm for the diffusion models could be concisely expressed as:

As the emerging generative models, the diffusion models have achieved remarkable success in many fields in recent years [70][88]. In terms of image synthesis, the diffusion models can generate high-quality and diverse images [89][71]. It is also used in image restoration [90][91], such as denoising and filling in missing parts, and image super-resolution [92][93], converting low-resolution images to high-resolution images. The diffusion models can also be applied to video generation [94][95], where high-quality

Algorithm 1 Training

repeat
  $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
  $t \sim \text{Uniform}(\{1, \ldots, T\})$
  $\epsilon \sim \mathcal{N}(0, \mathbf{I})$
  Take gradient descent step on $\nabla_\theta \|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\|^2$
until converged

Algorithm 2 Sampling

$\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$
for $t = T, \ldots, 1$ **do**
  $z \sim \mathcal{N}(0, \mathbf{I})$ if $t > 1$, else $z = 0$
  $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)\right) + \sigma_t z$
end for
return $\mathbf{x}_0$

video content can be generated by generating continuous frames from noise. This application has broad application prospects in fields such as filmmaking, animation generation, and virtual reality. In the domain of natural language processing (NLP), the diffusion models are used for text generation tasks, including text completion and dialogue generation [96][97]. By learning the distribution of text data, the diffusion models can generate coherent, semantically meaningful text.

In medical image analysis and generation, diffusion models also show great potential and application value. Rguibi *et al.* [98] proposed a Medical Variation Diffusion Model (Medical VDM) that generates smooth medical images that could retain important features (such as edges) through a variational diffusion model, thereby improving the accuracy and reliability of medical images. These generated images are useful for medical education, training, and to assist clinical diagnosis and treatment planning. Pinaya *et al.* [99] used LDM to generate 3D brain MRI images by mapping brain images to latent representations and generating synthetic images from noise, the model conditioned on covariates such as age, gender, and brain structure volume. The results showed that conditional variables can effectively control data generation, and synthetic data can supplement datasets and become a promising alternative. In addition, the diffusion model has demonstrated outstanding performance in medical image restoration,

super-resolution, and multimodal generation [100][101].

Although diffusion models have shown great potential, there are still some significant defects. Since multiple iterations are required to gradually denoise and generate images, the training and sampling process of diffusion models typically require significant computational resources and time. In addition, the architecture and training process of diffusion models are relatively complex, involving multiple steps and parameter tuning. This makes the implementation and maintenance of the model difficult. Especially in clinical applications, a high level of technical support and prior knowledge are required. Diffusion models rely heavily on large-scale and high-quality datasets. In the medical field, it is challenging to obtain sufficient labeled data, especially for data on rare diseases or specific medical conditions.

3.2 Literature review on cardiac image analysis

Cardiac image analysis includes a series of key tasks that collectively aim to quantitatively analyse cardiac images, elucidate the morphology and function of cardiac physiology, and perform dynamic image analysis [102]. This includes, but is not limited to, image acquisition and preprocessing, such as generating images for image augmentation [103][104]. Segmentation of the image, segmenting the cardiac and its different parts [105][106]. Image registration, aligning cardiac images from different time points or different imaging modalities for dynamic analysis or multimodal fusion [107][108][109]. Cardiac morphology and function analysis, quantifying the morphological characteristics of cardiac structures, such as cardiac wall thickness, chamber and ventricle volume, and evaluating cardiac function, including systolic function (such as EF), diastolic function, myocardial motion [110][111]. Dynamic image analysis, analysing dynamic image sequences to evaluate the motion and changes of the cardiac within a cardiac cycle [112][113]. Then calculate dynamic parameters, such as the change curve of cardiac ventricular volume and myocardial strain during the cardiac cycle [114][115].

In recent years, deep learning and artificial intelligence technologies have been applied to automated cardiac image analysis, feature extraction, and disease diagnosis and prediction, providing effective reference and assistance for clinical diagnosis, treatment planning, and surgical assistance [116][117][118]. The following chapters will explore the literature review of the main cardiac image analysis, including image synthesis,

segmentation, registration, dynamic analysis, and the application of deep learning.

3.2.1 Cardiac image synthesis

The generation of cardiac images generates high-quality synthetic cardiac images by learning the features of a large number of real cardiac images, which are used to assist medical research and clinical applications [119][120]. Generated images are of great significance in training data augmentation, model verification, and medical diagnosis, treatment, and research. Medical imaging datasets are usually limited, and synthetic images can be used for dataset augmentation, which is very important for deep learning model training and can improve the generalization and robustness of the model. Synthesised cardiac images can reduce the reliance on expensive and invasive imaging technologies (such as MRI and CT), reduce medical costs, and can be used for various clinical trials and studies without involving actual patients to explore new treatments and diagnostic techniques. It helps to accelerate the progress of medical research and promote new medical discoveries [121][122].

Compared with traditional image synthesis methods, generative models using deep learning-based methods have demonstrated excellent performance in synthetic image generation. These models are primarily based on architectures like GAN, VAE, and diffusion models. Amirrajab *et al.* [123] combined VAE with GAN, first synthesising cardiac labels through VAE, and then implementing the label-to-image translation task through label-conditional GAN. This study quantitatively evaluated the usability of generated virtual subjects for training cardiac MRI segmentation models in data augmentation, proving that data augmentation improves model generalization and robustness to multi-center data. Skandarani *et al.* [105] also combined VAE and GAN to generate highly realistic MRI and its pixel-accurate ground truth, which can be used for cardiac segmentation of cine MR images. VAE is used to learn the potential representation of the cardiac shape, while GAN generates MRI images that match the given anatomical map. Chartsias *et al.* [124] used CycleGAN to demonstrate the potential of cross-modal synthesis, synthesising MRI images from CT, and used synthetic data to train the segmentation network to obtain better results, proving the practicality of synthetic data. Zhang *et al.* [125] proposed a lesion-focused diffusion model, which simplified the model learning process and enhanced the controllability of the synthetic output by redesigning the diffusion learning objective to concentrate on the lesion area,

while retaining the background information during forward diffusion. Verification of the cardiac lesion segmentation dataset proved that synthetic data can effectively enhance the existing model. Abbasi *et al.* [120] proposed a hybrid controllable method for generating anatomically relevant 3D+t-labeled CMR images, and the generated images performed well in data augmentation and style transfer. Pan *et al.* [126] proposed a Swin-transformer-based DDPM, trained on datasets including chest X-rays, cardiac MRI, pelvic CT, and abdominal CT, and demonstrated the effectiveness of synthetic datasets through classification tasks, demonstrating the potential of this synthetic framework to synthesise high-quality medical images.

However, generative models usually use or fuse complex model structures, and the training process is computationally intensive, especially when processing high-resolution 3D images, which demands significant computational resources and time. The results of combined label research rely on accurate anatomical image labels, which may require a lot of prior knowledge and manual intervention in practical applications. When generating time series images, there may be a problem of temporal consistency, that is, the continuity and consistency of the generated images in the time series are insufficient.

3.2.2 Cardiac image segmentation

Cardiac image segmentation involves precisely isolating and identifying various anatomical structures, such as the left ventricle (LV), right ventricle (RV), and myocardium, in cardiac images. It is one of the core tasks based on medical image analysis, which is achieved through image processing and semantic segmentation, a basic task of computer vision. Cardiac image segmentation can be applied to 2D and 3D images, and has important applications in processing static and dynamic cardiac images. Accurate cardiac image segmentation is crucial for downstream cardiac analysis tasks, such as 3D shape reconstruction and estimation of cardiac clinical indicators. Segmenting the ventricles and atria allows for quantitative measurement of their volumes and enables evaluation of cardiac functions, such as myocardial mass, wall thickness, ventricular volumes, and EF. Monitoring of these clinical indicators can detect and diagnose cardiac lesions earlier and more accurately.

Advancements in deep learning technology have greatly enhanced the accuracy and efficiency of cardiac image segmentation, providing a solid technical foundation for per-

sonalized medicine. Numerous studies have extensively reviewed deep learning-based methods in medical image segmentation. Chen *et al.* [127] and Petitjean *et al.* [128] detailed the wide application of deep learning methods in cardiac image segmentation. Tran *et al.* [129] first applied fully convolutional networks (FCNs) to short axis (SAX) CMR image segmentation, proposing an end-to-end model that outperformed traditional methods in both speed and accuracy, demonstrating strong competitiveness. Subsequently, a number of FCN-based studies emerged to further improve segmentation performance. The focus of the research is on refining network architectures to improve feature extraction, such as the dense U-net developed by Khened *et al.* [130], and on refining loss functions (e.g., weighted cross entropy, weighted Dice loss, deep supervision loss, and focal loss) to enhance segmentation accuracy [131][132][133]. Due to the low through-plane resolution and motion artifacts in CMR scans, these studies have concentrated on 2D networks over 3D networks. However, 2D networks for cardiac segmentation cannot exploit inter-slice dependencies, making it challenging to accurately segment the heart on difficult slices, such as apical and basal slices. To address this, many studies have introduced additional contextual information to guide 2D FCNs, including shape priors learned from labels or multi-view images [134][135], and spatial information from neighboring slices [136]. Techniques such as recurrent units (RNNs) and multi-slice (2.5D) networks have been used to assist segmentation [137][138]. In addition, these networks also utilize information from multiple time frames throughout the cardiac cycle to enhance the spatial and temporal consistency of segmentation outcomes [139][140]. Recent research has focused on achieving anatomically accurate and robust segmentation in various challenging CMR scenarios. To further improve the performance of FCN-based ventricular segmentation, multi-task learning has also been explored [141][142][143]. By performing related auxiliary tasks (such as motion estimation, cardiac function estimation, ventricular size classification, and image reconstruction) to regularize the training process, multi-task learning encourages the network to extract features useful for each task, thereby improving learning efficiency and prediction accuracy.

Although these methods have made significant progress, some challenges still exist. The shortcomings of existing methods include unstable performance when processing images with large anatomical variability and pathological changes, and insufficient spatial and temporal consistency between slices. In addition, although multi-task learning

methods can improve feature extraction capabilities, they may also lead to complex network structures, increase training difficulty and computational costs. Therefore, developing more efficient network structures, exploring more advanced context information fusion methods, and combining the latest artificial intelligence technologies may bring breakthroughs to cardiac MR image segmentation. These improvements are expected to improve the accuracy and reliability of clinical applications, offering stronger support for diagnosing and treating cardiovascular diseases.

3.2.3 Cardiac image registration

Advancements in computing power, along with increased algorithmic capabilities and complexity, have led to significant progress in the field of image registration. This technology is widely used in various clinical scenarios, including disease diagnosis and monitoring, image-guided treatment, and postoperative evaluation. As the spatial resolution of medical images often varies, image registration is also widely used as a tool for data preprocessing in order to perform subsequent tasks such as object detection, segmentation, or classification. The performance of these subsequent tasks depends largely on the quality of the image registration algorithm. Image registration algorithms affect the effectiveness of subsequent processing by aligning images to a common coordinate system, unifying their size and resolution.

In the field of biomedical research, cardiac image registration has always received widespread attention [144][107]. It aligns and matches cardiac images acquired at different times or in different modalities to facilitate subsequent medical image analysis. It includes associating clinical features of images of different cardiac modalities, respiratory motion correction, facilitating the cardiac segmentation process, supplementary information for image fusion, and image guidance for clinical intervention treatment. Numerous studies have extensively reviewed deep learning-based approaches for medical image registration. Chen *et al.* [145] and Khalil *et al.* [146] detailed the wide application of deep learning methods in cardiac image registration.

In cardiac image registration, the LV has received the most attention [147][148][149]. This is because the geometry of the LV is relatively simple and the myocardial wall is thick, making automatic segmentation more feasible. In addition, LV function/dysfunction is associated with most cardiovascular diseases. Rohe *et al.* [150] used CNN to predict image registration parameters, focusing on deformable image registration through

shape matching. They performed well in the registration task of cardiac images and outperformed optimization-based algorithms. De Vos *et al.* [151] first proposed a CNN-based unsupervised deformable image registration method for registering 2D cine CMR images. The model does not require labeled data and achieves end-to-end training by optimizing the similarity metric between image pairs, reducing the cost and complexity of data annotation. Balakrishnan *et al.* [152] introduced VoxelMorph, a fast, unsupervised learning-based framework for paired medical image registration. It defines registration as a parameterized function that maps a pair of input images to a deformation field so that these images are aligned. VoxelMorph has been applied in multiple medical image registration tasks [153][154][155], including brain MRI, cardiac MRI, etc.

After VoxelMorph, deep learning-based cardiac image registration has advanced and has been applied to the analysis of various cardiac sequences. Upendra *et al.* [156] proposed a hybrid CNN and Vision Transformer (ViT) model for deformable image registration of 3D cine CMR images to achieve consistent cardiac motion estimation. The registration results were shown to be superior to the VoxelMorph CNN model and traditional non-rigid image registration algorithms. Chen *et al.* [157] proposed a joint motion estimation and segmentation method for undersampled CMR images, which combines the VoxelMorph framework and improves the accuracy of registration and segmentation through joint optimization. The model is able to predict results close to fully sampled data without the usual image reconstruction stage. Lu *et al.* [154] proposed a bidirectional registration CNN for cardiac motion tracking. The bidirectional recurrent neural network (RNN) models temporal relationships and can automatically learn spatio-temporal information from multiple images with fewer parameters.

Although deep learning-based image registration models perform well in many aspects, their deterministic framework limits their capacity to produce synthetic motions. Therefore, probabilistic models [158] have been proposed and applied to medical image registration problems. Dalca *et al.* [159] proposed a probabilistic generative model based on unsupervised learning, modeling the potential velocity field as a multivariate Gaussian distribution and regularizing it using a standard Gaussian prior. This framework not only achieves competitive registration results but also provides differential homeomorphism guarantees. Krebs *et al.* [160] proposed a model for learning probabilistic motion models from a series of images for spatio-temporal registration, which uses a latent motion matrix to encode motion in a low-dimensional space, thereby achieving

motion simulation and interpolation. The Gaussian process in the low-dimensional latent space captures temporal dependencies but increases the complexity of the model, and also does not specify a pixel-by-pixel explicit probability distribution of deformation.

Another advantage of the probabilistic view over other learning-based approaches is the analytical uncertainty estimates. However, they are difficult to evaluate in high-dimensional complex models. In summary, although the field of cardiac image registration has made significant progress through deep learning-based methods, it still faces great challenges. These challenges include the need to more effectively handle temporal dependencies, data dependencies, better modeling of pixel-level uncertainties, and cross-modality registration in the presence of large differences in imaging features.

3.2.4 Cardiac image dynamic analysis

The strength and direction of myocardial contraction largely control the heartbeat. When the heart is abnormal, it often leads to an abnormal heartbeat [161]. Therefore, tracking myocardial motion, accurately capturing the details of myocardial movement, and quantitatively describing the degree of myocardial tissue deformation is of great significance for evaluating cardiac function and diagnosing cardiovascular diseases. Myocardial deformation assessment follows the basic principles of most deformation imaging techniques: identifying specific patterns or features in the image and finding the best correspondence between these patterns or features over time in subsequent image frames to achieve tracking. By repeating this process throughout the time series, the deformation of local tissues can be estimated.

Myocardial strain analysis is a common method of measuring cardiac dynamic analysis. It derives the displacement gradient field from the dense displacement field and obtains the strain tensor that reflects the details of myocardial deformation. According to the geometric characteristics of the LV, the strain tensor can be divided into radial, circumferential and longitudinal strain components. As a routine sequence in clinical data acquisition, strain analysis in cine CMR is an important tool for evaluating myocardial deformation. Feature tracking (FT) is an important method for strain analysis in cine MRI. It tracks natural myocardial characteristics throughout the cardiac cycle, providing strain measurements in all directions (longitudinal, circumferential, and radial). FT-based strain analysis can detect subtle changes in myocardial function

earlier than traditional methods such as EF.

In recent years, many studies have attempted to calculate myocardial strain from cine CMR images [162][163]. Onishi *et al.* [164] used an FT method to evaluate the longitudinal changes in cardiac function and myocardial strain values in a myocardial disease model and verified the practicality of myocardial strain analysis using cine CMR. Truong *et al.* [165] used CMR FT to calculate the normal range of LA strain and strain rate and compared them with 2D STE.

Tagging CMR marks specific areas of the myocardium in MRI as markers during contraction. Tagging CMR-based strain analysis provides a pioneering approach to describing myocardial markers and is regarded as the reference standard for strain assessment [166]. Tracking marker deformation enables direct evaluation of myocardial strain and demonstrates great reproducibility. Myocardial tagging technology has been used in many clinical and research applications. Valet *et al.* [167] compared the changes in tag contrast over time in tagging images acquired at different field intensities and analysed the images for systolic and diastolic strain measurements. Studies have shown that high-field imaging has significant benefits in terms of the durability of myocardial tagging throughout the cardiac cycle. Ennis *et al.* [168] used tagging MRI to compare regional LV function during systole and diastole in subjects with familial hypertrophic cardiomyopathy (FHC) and normal subjects, demonstrated a reduction in early diastolic strain rate across all regions in the FHC group, highlighting regional differences in systolic and diastolic dysfunction in FHC patients. Denney *et al.* [169] proposed an unsupervised tag extraction and cardiac strain reconstruction algorithm to quantify the 3D myocardial strain on tagging CMR images. The results demonstrated that the algorithm is capable of measuring disease-induced wall motion abnormalities and has the potential to overcome the limitations of routine clinical use of tagging CMR. Herrezuelo *et al.* [170] introduced the application of variational methods in cardiac motion estimation and proposed a new method for motion estimation of tagging CMR sequences based on variational techniques. Validation results in synthetic and real sequences demonstrated the accuracy of the algorithm for motion estimation.

There are also some strain assessment techniques for variations of tagging, such as DENSE and SENC. DENSE [46] was developed in the late 1990s, and provides superior spatial resolution and relatively straightforward post-processing. SENC, originally developed by Osman *et al.* [171], enables faster acquisition, more efficient post-processing,

and demonstrates excellent reproducibility. However, both techniques face limitations, including the need for additional acquisition, low SNR, lack of standardization, and intensive post-processing, which restrict their use as primary research tools.

In recent years, many studies have applied deep learning to cardiac strain analysis. Ye *et al.* [172] introduced an unsupervised deep learning-based motion tracking model for tagging CMR images. They used a bidirectional generative diffeomorphic registration neural network to compute the motion field across consecutive time frames. They further estimated the Lagrangian motion field from the reference frame to other frames through a differentiable combination layer. The results show that this method outperforms traditional motion tracking methods in tag tracking accuracy and inference efficiency. Morales *et al.* [173] proposed an efficient, fully autonomous deep learning pipeline integrating segmentation and motion estimation CNNs to derive volume metrics and strain values from cine CMR. This end-to-end learning pipeline fully automates the analysis of cine MRI data for quantitative characterization of cardiac mechanics in healthy and cardiovascular disease subjects. Ferdian *et al.* [174] introduced a fully automated deep learning framework for estimating myocardial strain from SAX tagging CMR. This approach facilitates unbiased strain assessment within a high-efficiency workflow and demonstrates comparable effectiveness in distinguishing disease-related injuries.

In summary, although the field of cardiac dynamic analysis has made significant progress through machine learning and deep learning-based methods, major challenges still exist. These challenges include the need to more effectively handle data dependencies, enhance model generalization, and standardization issues.

3.3 Datasets

UK Biobank [175] is a prospective cohort study that integrates a large sample size with extensive phenotypic and genotypic data to enhance the prevention, diagnosis and treatment of diseases in middle-aged and elderly people, particularly cardiovascular conditions like heart disease and stroke. Its goal is to improve public health by deeply analysing these data to find effective prevention, diagnosis and treatment methods. From 2006 to 2010, UK Biobank recruited 500,000 subjects aged 40 to 69 from 22 assessment centers across England, Wales, and Scotland. At these centers, subjects were given electronically signed consent forms, responded to touchscreen and verbal

interview questions, and questions about sociodemographic, lifestyle, environmental, and health-related factors, underwent a series of physical measurements, and given blood, urine, and saliva samples. With recruitment in full swing, the content of the assessment visit has been further enhanced to include eye measurements, heel bone ultrasound, electrocardiogram tests, pulse wave velocity and hearing tests for most participants.

The collection of baseline data covers multiple aspects directly related to cardiovascular health, including self-reported data on medications, health conditions, cardiovascular disease family history, arterial stiffness, blood pressure, cardiorespiratory fitness, as well as body size and fat. While these data do not reflect the general population and thus unsuitable for estimating disease prevalence or incidence, the large sample size enables reliable detection of relationships between most baseline characteristics and health outcomes.

UK Biobank data have unprecedented depth and breadth, providing a valuable opportunity to study questions concerning cardiovascular health outcomes. These data include MRI, dual-energy X-ray absorptiometry (DXA), and CT scans of the brain, heart, bones, and abdomen.

Focusing on CMR data, Petersen *et al.* [176] manually analysed CMRs of 5,065 consecutive UK Biobank participants. The researchers manually segmented all slices of each 3D CMR scan at ED and ES. Annotation and evaluation were performed by two core laboratories in London and Oxford. In collaboration with CISTIB, the research team accessed approximately 5,000 CMR samples and manually outlined the LV and RV anatomical structures.

50 SAX stacks were collected under b-SSFP during each patient's cardiac cycle, each containing 8 to 13 slices with the matrix size of 208×162. Ground truth annotations of the LV endocardium, epicardium, and RV were available at both ED and ES, with ED being time frame 0 and ES varying between frames 21 and 26, depending on the patient. For the tagging sequence, 20 SAX stacks were collected, covering the entire cardiac cycle, each containing apical, middle, and basal slices with a matrix size of 200×256. These detailed imaging data provide researchers with valuable resources to help them gain a deeper understanding of changes in cardiac structure and function, thereby advancing the study and management of cardiovascular diseases.

Chapter 4

Automated Myocardial Strain Quantification via Synthesised Tagging-MRI: A Sparse Multi-Channel Variational Autoencoder Approach

Myocardial strain is an important indicator for evaluating cardiac function, which quantitatively characterizes myocardial motion throughout the cardiac cycle. Notably, myocardial strain can detect abnormal changes in myocardial strain in patients even when EF and/or other ventricular volume indices remain within normal or healthy ranges. This allows us to detect cardiac dysfunction early based on quantitative analysis of abnormal cardiac motion patterns. Tagging CMR is regarded as the reference standard for quantifying myocardial strain, and its effectiveness has been demonstrated in different patient populations. However, the clinical application of tagging CMR has been limited by the lack of automated and powerful myocardial strain computational analysis tools and the prolonged image acquisition time compared to cine CMR alone. In this chapter, we employ a sparse multi-channel variational autoencoder (smcVAE) to jointly learn the latent representations of cine CMR images and tagging CMR images. During inference, tagging CMR images can be synthesised from cine CMR images as input, enabling the quantification of myocardial strain from these synthesised images. The results show that our framework can effectively synthesise tagging CMR from cine CMR, and providing valid estimates of myocardial strain.

4.1 Introduction

Strain [39] measures the degree of deformation or change in shape experienced by an object when subjected to external forces or loads. It quantifies how an object's shape or size has been altered due to applied forces. Myocardial strain [21][25] refers to the measurement of deformation in the myocardium (muscular tissue in the heart) as a result of the contraction and relaxation of the cardiac across the cardiac cycle. It provides quantitative information about the motion of the heart muscle, allowing for a more comprehensive evaluation of cardiac function. Cardiac motion from ED (maximum dilation) to ES (maximum compression) during the cardiac cycle leads to alterations in myocardial strain along different directions. Under healthy/normal conditions, myocardial strain is within a specific range and the changes are coordinated and orderly. Previous studies have found that myocardial longitudinal, radial, and circumferential strains decrease with age, accompanied by relative thickening of the ventricular wall and increase in volume mass [177]. Regional changes to myocardial strain have been used as a biomarker to identify cardiac dysfunction resulting from several cardiac diseases (e.g. ischemic heart disease, hypertrophic cardiomyopathy, etc.) [178, 179].

CMR [180] has emerged as the standard for noninvasive assessment of cardiac function, owing to its large field of view, high tissue resolution, absence of radiation, and excellent repeatability. By combining various technologies, MRI can conduct a 'one-stop' inspection of both the structure and function of the heart. At present, various MRI quantitative imaging techniques are developing rapidly, offering valuable insights for the diagnosis and characterization of cardiovascular diseases [181][182]. Cine CMR [183] uses cardiac segment acquisition technology or real-time imaging technology to continuously acquire images of multiple phases at the same level within one cardiac cycle, and is a MRI technology that shows the rhythmic contraction and relaxation of the heart. Tagging CMR [24] utilizes tissue magnetization as a tissue property. A locally magnetically saturated grid of dark-lined tissue markers, called tag, is induced onto the myocardium by applying radio frequency pulses in an orthogonal plane. These marker lines deform with the myocardial tissue during systole, allowing the 'marker' grid can be used to track and evaluate the displacement of the myocardial tissue, serving as the basis for calculating cardiac strain.

Myocardial tagging is considered the reference for quantifying local myocardial motion and strain, allowing for quantitative assessment of myocardial deformation with good reproducibility within and between patient groups. Previous tag-tracking methods can be broadly categorized into three groups: (i) methods that detect and track tag lines in images, such as Findtags [184] and InTag (Creatis, Lyon, France), etc. [185] [186]; (ii) methods based on optical flow which estimate the motion of objects within images by evaluating spatio-temporal variations of image intensity [187][139][188]; and (iii) methods based on HARP analysis [189, 190], which calculate the spatial phase of each pixel in the marked pattern. The phase can be used to calculate the deformation by tracking the cardiac cycle points or by calculating the difference between the spatial frequency of the tagged image area and the undeformed frequency. However, the widespread clinical application of tagging MRI is limited by several factors, including additional sequences will extend acquisition time, reliance on labor-intensive manual or not fully automatic post processing algorithms to estimate strain, and insufficient validation. Therefore, tagging CMR has not been used as widely as other methods, such as cine CMR in the clinical setting due to the absence of rapid and efficient analytical methods.

Motivated by the application of deep learning-based generative models in image

synthesis [191] [192], this study proposes an automatic framework for synthesising tagging CMR from cine CMR and quantifying myocardial strain from the former. This will help to avoid the increased scan time associated with acquiring additional sequences, enabling the quantification of strain directly from cine CMR. The image synthesis problem is probabilistically formulated as the learning of a joint latent representation for two types of image sequences from a subject. We utilize a smcVAE to learn a shared latent space that captures the distinct information channels represented by tagging and cine images for each subject. The latent representations obtained by the smcVAE are then able to generate tagging CMR images for new or unseen objects during inference, simply by taking their cine CMR images as input. We use the fully automatic cardiac motion tag tracking network trained by Ferdian *et al.* [174] to perform tag tracking on tagging images and estimate myocardial strain. To verify the performance of the model, the generated cine images are segmented and compared against the segmentation results from the original cine.

Our method is evaluated on the public dataset from the UK Biobank via five-fold cross-validation, and experimental results using a variety of metrics show that, tagging CMR images synthesised from cine CMR images can be used to quantify myocardial strain, and strains calculated on the tagging images are in good agreement with previous studies [172][193].

Section 4.2 details the framework for image synthesis and strain analysis. In Section 4.3, we present the dataset, image acquisition protocol, and preprocessing procedures. And then present the experiments and results that validate the effectiveness of the model. Section 4.4 provides an in-depth discussion of the methods and results. Finally, Section 4.5 is the conclusion of this chapter.

4.2 Methodology

Figure 4.1 shows an overview of the framework's workflow. In the training process, the smcVAE model is trained on cine and tagging images, the learned latent space enabling the reconstruction of channel information for each image sequence. In the testing process, the smcVAE model is utilized to synthesise tagging from test cine, and conversely, to generate cine from test tagging for model validation. During the strain analysis process, cardiac motion in the synthetic tagging CMR images is estimated using a combined RNN and CNN model, which was also validated on the UKBB dataset.

Predicted landmarks are extracted for strain analysis, allowing for the calculation of both radial and circumferential strains. For synthetic cine CMR images, the trained model is employed for segmentation and motion estimation.

4.2.1 Sparse Multi-channel Variational Autoencoder (smcVAE)

Our method leverages a multi-variate approach derived from smcVAE [194], which can projects observations from diverse sources into a unified latent space for comprehensive data analysis and enables the generation of new observations by sampling from these learned latent representations. The smcVAE framework simultaneously trains two encoder-decoder pairs, with cine CMR and tagging CMR image sequences serving as input channels for each pair. A shared latent space, jointly learned by both encoder-decoder pairs, facilitates the integration of these distinct modalities into a coherent latent representation. This network architecture, depicted in Figure 4.2, comprises an encoder for each channel with seven convolutional layers and a fully connected layer, while the decoder reconstructs the cine CMR and tagging CMR images using seven transposed convolutional layers and Tanh activation. Notably, each convolutional layer employs depthwise separable convolutions [195] instead of conventional convolutions, incorporating both Depthwise (DW) and Pointwise (PW) convolutions. This design, akin to conventional convolution in structure, reduces the computational complexity and the number of parameters. In DW convolution, each channel is processed by an individual convolution kernel, unlike conventional convolution where each kernel processes the entire input. Meanwhile, the PW convolution effectively captures spatial features across channels, similar to traditional convolution operations.

To introduce smcVAE, first, we introduce the multi-channel variational autoencoder (mcVAE) [196]. Suppose the mcVAE has c channels of input data, each observation channel x_c is a $d-$dimensional vector, and s is an $l-$dimensional vector across all channels x_c. During the generation process, $x_c \sim p(x_c \mid s, \theta_c)$, $c = 1, ..., C$, which is the likelihood distribution of observations conditioned on the latent variable. The inference process derives the posterior $p(s \mid x_c, \phi_c)$ from the joint latent space s from each channel to generate observations and provide information about the latent variable distribution. Variational inference is employed to approximate the posterior distribution $q(s \mid x_c, \phi_c)$, under the assumption that each data channel is conditionally independent of the others. Because each channel provides a different approximation, KL divergence

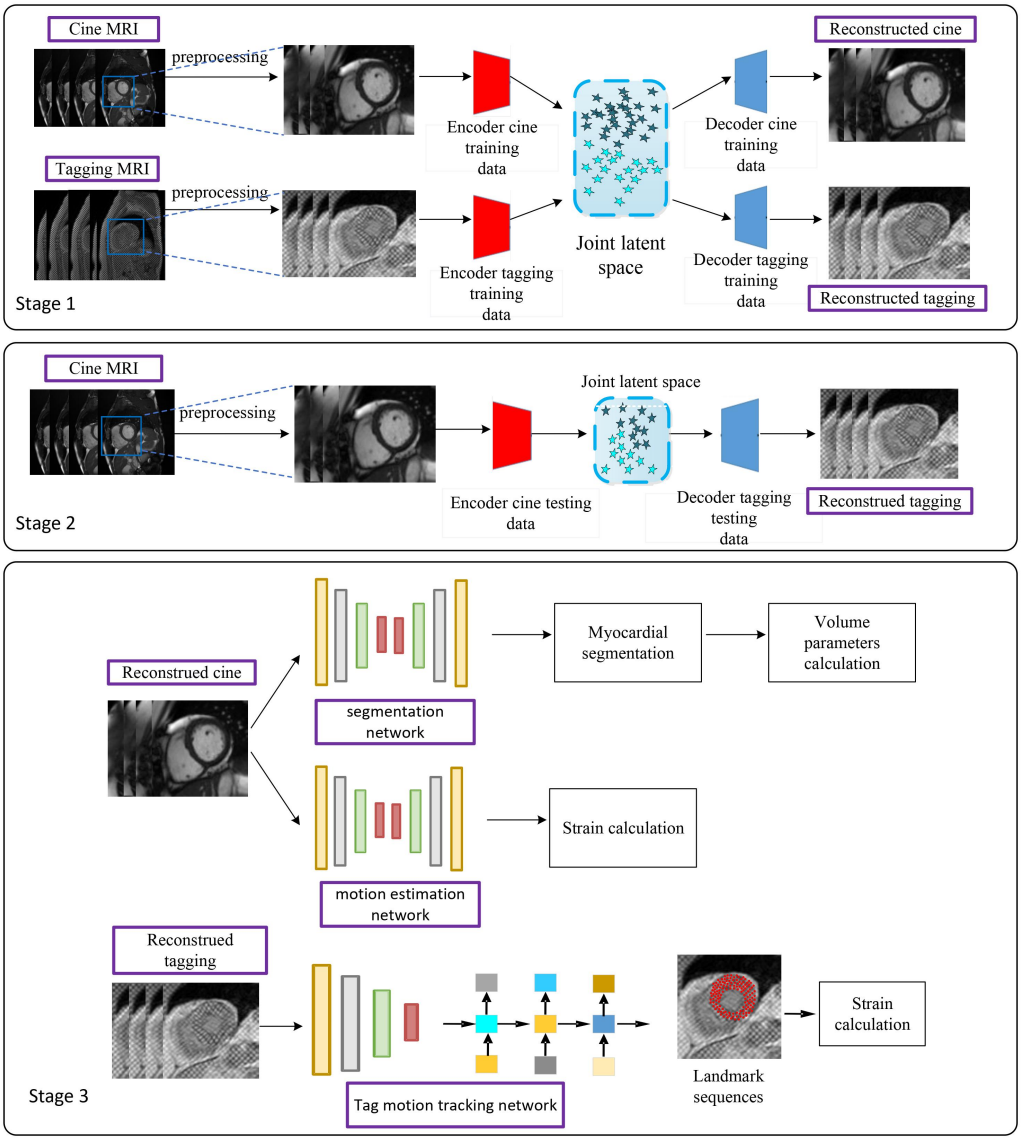

Figure 4.1: **Schematic illustration of our proposed smcVAE framework.**

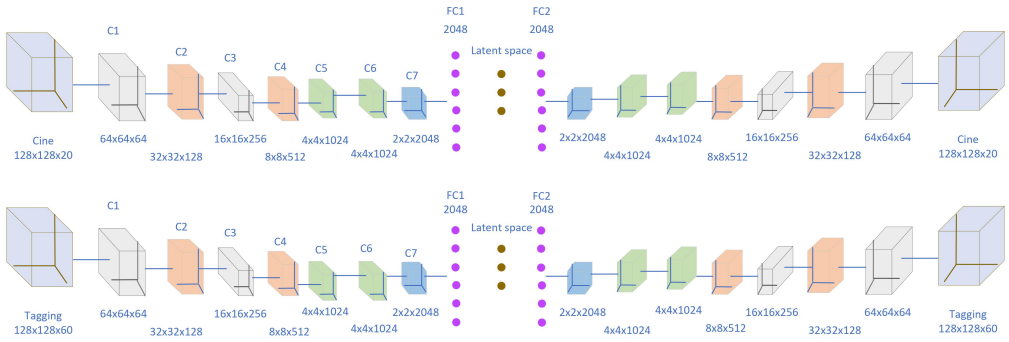

Figure 4.2: **Schematic illustration of the network structure of the smcVAE model.**

constraint [197] is imposed to enforce that each posterior distribution closely aligns with the target prior distribution $p(s)$. The optimization problem can then be expressed as follows:

$$\mathcal{L}(\theta, \varphi, x) = \mathbb{E}_c\Big[\mathbb{E}_{q(s|x_c,\phi_c)} \sum_{i=1}^{C} \ln p(x_i|s,\theta_i) - \mathcal{D}_{KL}(q(s|x_c,\phi_c)\|p(s))\Big], \tag{4.1}$$

$\mathbb{E}_c$ represents the average computed across all channels. After the mcVAE has learned a common latent space, it can reconstruct all other channels from the latent representation by cross-channel decoding, or multiple channels by only encoding information from one channel.

In smcVAE, the difference from mcVAE is that sparsity constraints are imposed on s to automatically infer the dimensions of latent variables. Specifically, choose a dropout posterior for the latent code of s, define the approximate posterior probability as $q(s \mid x_c, \phi_c)$, and parameterize them as $q(s \mid x_c, \phi_c) = \mathcal{N}(\mu_c; \mathrm{diag}((\sqrt{\alpha} \odot \mu_c)^2)$, where $\mu_c = \phi_c x_c$. The sparse s in smcVAE facilitates the computation of the optimization, thereby enhancing the model's interpretability by reducing the complexity of the relationships that need to be considered.

4.2.2 Myocardial Strain Estimation for Tagging

To calculate the strain for tagging CMR, it is necessary first to perform tag tracking for cardiac MRI to monitor cardiac motion within the images. In this work, we reimplemented a fully automatic method based on the approach detailed in [174] for tag

tracking on cardiac tagging MRI. This approach utilizes a combination of CNN and RNN for detecting and tracking myocardial landmarks throughout each image time sequence. Strains are then calculated according to the movements of these landmarks. Strain is subsequently calculated based on the movement of these landmarks, using the Green-Lagrangian strain formula as detailed in equation 4.2. This landmark tracking network (combined with CNN and RNN) is trained and validated on dataset from the UK BioBank. Within this framework, the CNN extracts spatial features, while the RNN integrates the temporal relationship between frames, with both components trained end-to-end.

The strain is calculated using:

$$\epsilon(t) = \frac{1}{2}\Big(\frac{L_t^2 - L_0^2}{L_0^2}\Big) \tag{4.2}$$

where L_t denotes in the given frame t, the segment length, and L_0 is the length at the beginning.

To further validate the effectiveness of the smcVAE model, we segment the generated cine images using a fully automated deep learning workflow trained by Morales *et al.* [173], which can perform segmentation and motion estimation of the cine images. The segmentation network is used to generate cardiac tissue labels, while the motion estimation network is used to estimate myocardial motion. The networks all consist of encoder-decoder architecture composed of convolutional layers, batch normalization layers [198], and PReLU activations [199] with residual connections [200] to enhance performance.

4.3 Experiments and Results

4.3.1 Dataset

The dataset is sourced from the UK Biobank, a large long-term research dataset of biological samples in the UK. There are two modality sequences: cine CMR and tagging CMR (Access Application No. 11350). Details of data examinations are given in [201]. Cine CMR and tagging CMR were acquired using a 1.5T clinical wide-aperture MRI. Tagging CMR was obtained in basal, middle, and apical three SAX slices with the following acquisition parameters: repetition time = 41.05 ms; echo time = 3.9 ms; flip angle = 12; FOV (field of view) = $350mm \times 241mm$; voxel size = $1.4 \times 1.4 \times 1.0mm$;

tag grid spacing = 6 mm; trigger time = 41 ms; and approximately 20 reconstructed frames. Cine CMR images were acquired with the following parameters: repetition time = 31.56 ms; echo time = 1.1 ms; flip angle= 55; FOV = $370mm \times 296mm$; slice thickness = 8 mm, pixel size = $1.83 \times 1.83mm^2$. All participants gave written informed consent.

4.3.2 Data Preprocessing

Preprocessing of cine CMR images was carried out through a series of steps, including ROI clipping, interpolation, resampling, and intensity normalization. Initially, a CNN model [136] was employed to automatically delineate the ROI surrounding the heart. The image stacks were then resampled to 20 slices. Each slice was subsequently resized to 128 × 128 pixels. And the image intensities were normalized from 0 to 1.

For preprocessing of tagging, the selection of ROIs involves transforming the ROI coordinates derived from cine CMR images. After this, the tagging images undergo preprocessing steps similar to those applied to cine images, including spatial resampling and intensity normalization. Figure 4.3 illustrates the coordinate transformation procedure. A crucial aspect of this process is that the 3D patient coordinates remain consistent between the two imaging sequences, ensuring accurate ROI determination in the tagging images. This coordinate transformation was systematically applied to all subjects in the experiment.

4.3.3 Experimental Setting

In our experiments, we train a smcVAE generative model to jointly learn from cine CMR and tagging CMR. This model is capable of generating tagging CMR images directly from cine CMR and allows to estimate of myocardial strain, including both circumferential and radial strain, during the inference stage. From the public dataset from the UKBB, 535 subjects (10700 slice pairs) were used, containing two sequences (cine CMR and tagging CMR), in experiments, 60% of the dataset is selected for training, 20% for validation, and 20% for testing. Before model input, the images were resized to a 128×128 matrix, with each CMR sequence fixed at 20 frames per slice. For tagging CMR, including 60 slices have all slice images from apical, middle and basal. The generative model neural network was developed using Pytorch 1.7.0 [202] and Python, and trained on NVIDIA Tesla M60 with 16 GB RAM. The Adam optimizer

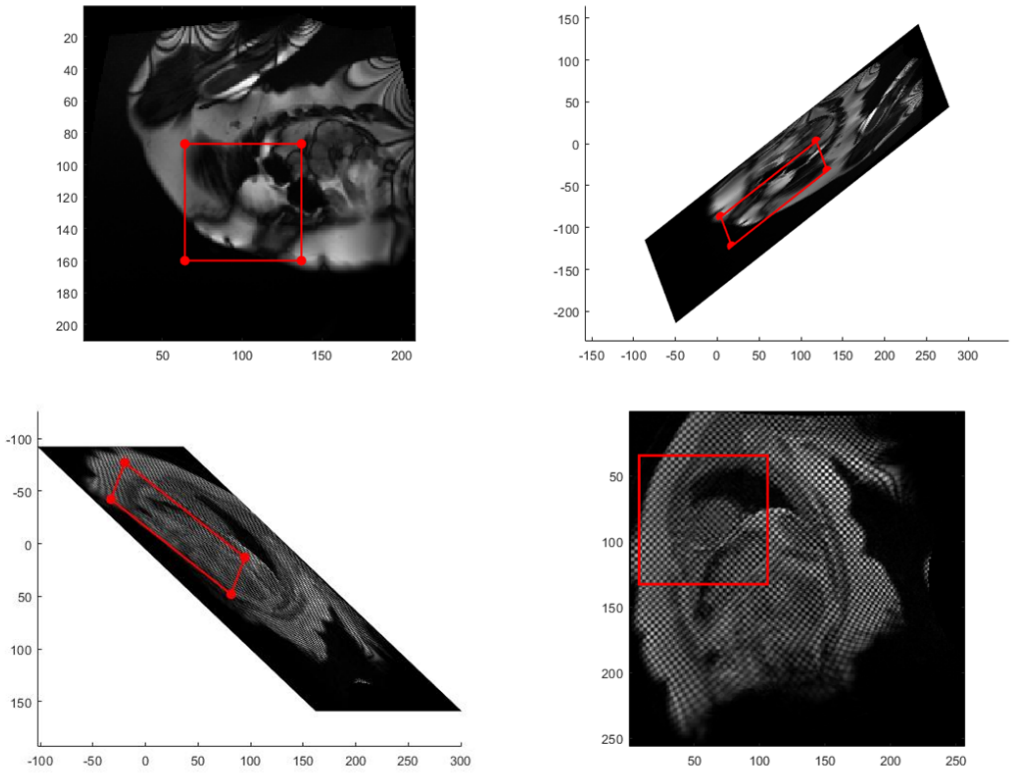

Figure 4.3: **Schematic illustration of the coordinate transformation process.** Top left: 2D cine image ROI; Top right: 3D cine image ROI; Bottom left: 3D tagging image ROI; Bottom right: 2D tagging image ROI.

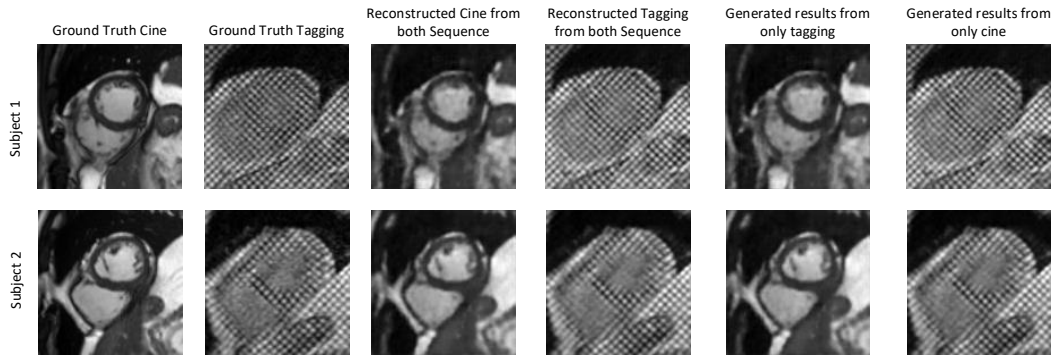

Figure 4.4: Examples comparing generated and original CMR images from different subjects, including both input cine and tagging images to reconstruct cine and tagging images, and test results for synthesising tagging images using only cine images and synthesising cine images using only tagging images.

was employed for training, with a learning rate of 5×10^{-4}, and a weight decay of 10^{-5}. The final outputs of the testing process of this framework are reconstructed cine and tagging CMR images.

4.3.4 Qualitative Evaluations

The results of our framework are presented in Figure 4.4, demonstrating its ability to successfully synthesise images that closely align with the ground truth tagging CMR. The synthesised results exhibit strong consistency with the real images and are visually appealing. However, when using only cine to reconstruct tagging or only tagging to reconstruct cine, the results tend to be relatively blurred. In contrast, when both cine and tagging images are used simultaneously to reconstruct each other, the outputs are more accurate, although some edges of the structures still appear slightly blurred and deformed.

4.3.5 Quantitative Evaluations

Synthetic images should exhibit realistic textures and maintain structural consistency with their corresponding real images. The evaluation of the results is not limited to qualitative evaluation, but quantitative evaluation is also required. For quantitative evaluation, we employ widely recognized image similarity metrics: Root Mean Square

Error (RMSE) [203], Peak Signal-to-Noise Ratio (PSNR) [204], Structural Similarity Index Measure (SSIM) [205], and Mean Absolute Error (MAE) [206]. RMSE quantifies the difference between variables, serving as an objective evaluation metric based on pixel error. When measuring the deviation from the reference image, a lower RMSE indicates higher quality in the synthesised image.

$$\text{RMSE} = \sqrt{\frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} (f^{'}(i,j) - f(i,j))^2} \tag{4.3}$$

where $f^{'}(i,j)$ and $f(i,j)$ represent the gray values of the synthesised image and original image, M and N denote the pixel's number along the image's length and width, respectively.

MAE is the mean absolute error, which is also one of the indicators to measure the image quality. It is calculated by taking the sum of the absolute intensity differences between the evaluation image and the original image and dividing it by the total number of pixels. A lower MAE value indicates a smaller deviation from the original image, reflecting better performance of the generative model. The formula is as follows:

$$\text{MAE} = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} |f(i,j) - f'(i,j)| \tag{4.4}$$

SSIM evaluates image similarity by comparing brightness, contrast, and anatomical structures between images. The SSIM value ranges from 0 to 1, the higher values indicating greater similarity between the images.

$$\text{SSIM} = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \tag{4.5}$$

Here, μ_x and μ_y represent the mean values of images x and y, while σ_x^2 and σ_y^2 denote the variances of x and y. Indicates the deviation/fluctuation range of the image brightness from the gray average value, so it can be used to describe the strength of the contrast. If the variance is large, the image contrast is high. σ_{xy} represents the covariance of x and y, which captures structural differences between the images. If the covariance is relatively small, it means that the structural difference between the two pictures is small. And c_1 and c_2 are two constants to avoid division by zero.

PSNR measures the ratio of a signal's peak energy to the average energy of the noise and is a widely used metric for assessing image reconstruction quality. A higher

PSNR value indicates superior reconstruction quality.

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}} \right) \tag{4.6}$$

MAX_I^2 represents the maximum possible pixel value of the image, which is typically 255 for uint8 data. MSE refers to the mean square error.

$$\text{MSE} = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} \left(f(i,j) - f'(i,j) \right)^2 \tag{4.7}$$

Table 4.1 lists the quantitative evaluation results, including a numerical comparison of the test dataset outcomes for synthesised cine and tagging generated from both two sequence inputs, as well as reconstructed tagging generated using only cine inputs and reconstructed cine generated using only tagging inputs. As expected, the results indicate that generation performance with single-channel input is not as strong as with two-channel input. And the result of generating cine from tagging is the worst, which may be because the rich grayscale changes of cine images cannot be fully inferred from the tagging texture information. In addition, the generation quality of the three regions of apical, middle and basal is different, but overall the SSIM and PSNR of the middle region are the highest, which may indicate that the image information in the middle region is more stable and easy to reconstruct, while the apical and basal regions may increase the difficulty of model prediction due to changes in cardiac motion or anatomical structure. The results show that in the single-channel generation task, the best results of generating image SSIM, PSNR, RMSE and MAE are 0.86 ± 0.04, 28.13 ± 4.85, 0.10 ± 0.01, and 154.86 ± 27.43 respectively.

4.3.6 Motion Tag Tracking

The tag tracking method for tagging images utilizes a neural network model based on RNN and CNN, trained on the UK BioBank dataset [174]. This approach has demonstrated accuracy in tracking myocardial motion in tagging MRI. The network outputs the landmarks at each time point for each slice. Figure 4.5 shows examples of generated tagging images in ED and ES for tag detection and tracking in the two different subjects.

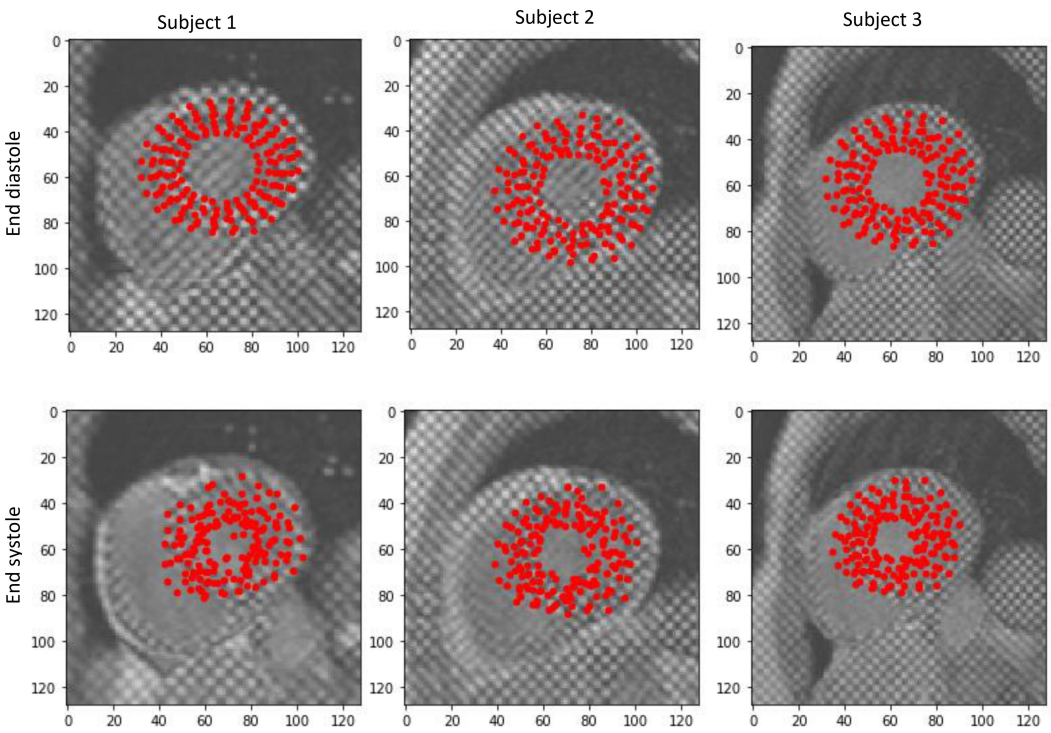

Figure 4.5: Examples of tag tracking estimated during ED (top row) and ES (bottom row) in three different subjects.

Table 4.1: Quantitative numerical comparison of the results generated by a test process. (Boldface denotes best performance).

Reconstruction/generation scenarios	Region	SSIM	PSNR	RMSE	MAE
Reconstruction tagging	Apical	$\mathbf{0.86 \pm 0.04}$	26.99 ± 2.70	$\mathbf{0.10 \pm 0.01}$	$\mathbf{158.61 \pm 44.61}$
	Mid	0.84 ± 0.21	$\mathbf{28.13 \pm 4.85}$	0.14 ± 0.01	159.24 ± 39.54
	Basal	0.82 ± 0.04	26.14 ± 2.27	0.11 ± 0.03	159.73 ± 65.15
Reconstruction cine	Apical	0.78 ± 0.05	23.94 ± 1.79	$\mathbf{0.18 \pm 0.07}$	$\mathbf{154.86 \pm 27.43}$
	Mid	0.69 ± 0.27	23.91 ± 6.68	0.21 ± 0.04	166.95 ± 35.91
	Basal	$\mathbf{0.79 \pm 0.03}$	$\mathbf{24.41 \pm 2.46}$	0.18 ± 0.08	159.46 ± 29.65
Only cine to tagging	Apical	0.72 ± 0.24	24.62 ± 2.53	$\mathbf{0.14 \pm 0.07}$	$\mathbf{160.56 \pm 39.54}$
	Mid	$\mathbf{0.86 \pm 0.05}$	24.57 ± 5.84	0.17 ± 0.02	165.77 ± 35.65
	Basal	0.79 ± 0.03	$\mathbf{26.62 \pm 2.46}$	0.16 ± 0.03	163.85 ± 37.56
Only tagging to cine	Apical	$\mathbf{0.72 \pm 0.18}$	$\mathbf{23.76 \pm 4.56}$	$\mathbf{0.27 \pm 0.08}$	$\mathbf{155.67 \pm 45.73}$
	Mid	0.65 ± 0.29	20.44 ± 7.51	0.31 ± 0.10	170.60 ± 38.65
	Basal	0.70 ± 0.21	21.23 ± 6.36	0.28 ± 0.07	161.67 ± 47.15

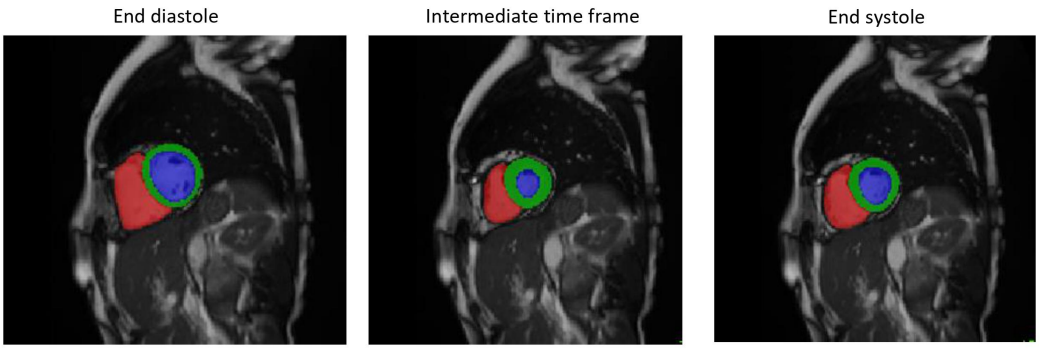

Figure 4.6: Example visualization of anatomical region segmentation results in cine cardiac images. Segmented areas include Myo., LV, and RV.

4.3.7 Myocardial Segmentation for cine CMR

Myocardial segmentation for cine is performed using a fully automated deep learning workflow [173], which incorporates CNNs for both segmentation and motion estimation. The segmentation network for cardiac generates tissue labels RV, LVM, and LV, and the cardiac motion estimation network generates myocardial displacement. Figure 4.6 shows an example of myocardial segmentation results for the original sample, from left to right is at ED, intermediate time frame between ED and ES, ES, respectively.

The performance of the validation of the model reconstructed cine CMR uses Hausdorff distance (HD) and dice similarity coefficient (DSC) evaluation indicators. HD and DSC are commonly used to evaluate medical image segmentation. We measure the

similarity between the segmentation results from the original cine CMR images and the synthesised counterpart data. HD is a measure that describes the similarity between two point sets. The formula for HD is as follows:

$$HD(O,G) = \max\left\{\max_{S_O \in S(O)} d(S_O, S(G)), \max_{S_G \in S(G)} d(S_G, S(O))\right\} \tag{4.8}$$

where O represents the original CMR image segmentation result, while G represents the generated CMR image segmentation result, and S_O and S_G are the elements in the two sets respectively. d represents the Euclidean distance.

The formula of DSC is as follows:

$$Dice(O,G) = \frac{2|O \cap G|}{|O| + |G|} \tag{4.9}$$

The Dice score ranges between 0 and 1. Our results show that the DSC score of the reconstructed cine CMR segmentation result is 0.825 ± 0.0196, and the HD is 6.411 ± 1.196.

Figure 4.7 shows examples of cine CMR original images and generated image segmentation results. The segmentation of the generated images exhibits some missing areas compared to that of the original images, highlighting regions of image blur that are challenging to discern with the naked eye.

4.3.8 Strain Analysis

All statistical analyses were performed using the open-source Python library SciPy Statistics [207]. Tagging CMR strains were calculated separately for the basal, middle, and apical slices, providing results specific to each slice. The error between the original and generated images is expressed as the mean difference ± standard deviation. To quantify agreement, Bland-Altman analysis [208] was employed, plotting the difference between the means of the two measurements. Additionally, violin plots were used to visualize the distributional differences in strain values. Table 4.2 presents the results of the differences between strain values calculated from the tagging generated from cine and those derived from the original tagging images for all subjects in the test set, including both circumferential strain E_c and radial strains E_r. The results show that the strain difference between generated tagging and original tagging in all regions (basal, middle and apical) is between about 0.01 and 0.02, and is within the reported standard deviation. This shows a high consistency between the two, which highlights that the results

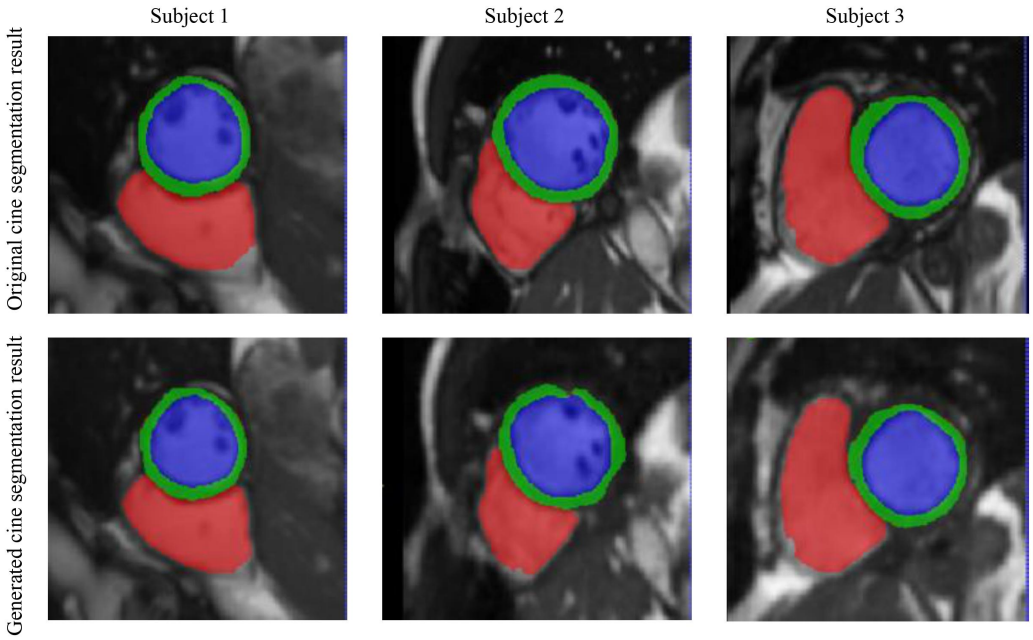

Figure 4.7: Examples of cine CMR original images segmentation results and generated image segmentation results. Segmented areas include Myo.(green area), LV (blue area), and RV (red area).

of generated tagging are very close to the strain results from original tagging. The average deviation of strain results is small and statistically insignificant, which indicates that the differences introduced by generated tagging are negligible in real practice, so the tagging generated based on cine images can reliably estimate strain values. Therefore, the generated tagging can be considered as a viable alternative, especially when the original tagging data is not available or cannot be obtained. However, the results also show that the circumferential strain E_c values of generated tagging are slightly underestimated compared with original tagging in all regions. This suggests that the generated tagging may slightly bias the circumferential strain towards smaller negative values. The radial strain E_r value of the generated tagging is slightly overestimated compared to the original tagging. This indicates that there is a slight upward bias in E_r. Figure 4.8 displays the Bland-Altman plot, comparing the difference between strain values from tagging generated solely from cine images and the original tagging strain values. The plot indicates that the average differences for E_c and E_r are close to zero, with the majority of cases falling in the 95% limits of agreement. Although some outliers with large errors are observed, overall, there is a strong agreement between the strain calculated from tagging images generated from cine and those from the original tagging dataset. At the apical slice, the average difference in circumferential strain is the smallest, and the average difference between the radial strain on the basal slice is the smallest. 95% confidence range of the two measuring methods differences are all within $-0.1 \sim 0.1$.

The violin plot in Figure 4.9 illustrates the distribution and comparison of strain values across different slices between the original and generated tagging images. The horizontal median (represented by the white dot within the violin plot) and the interquartile range (indicated by the black bar in the center) are similar across all tasks, indicating a consistent overall data distribution. The similar widths of the violin plots for the original and generated further reflect close approximation of the strain value distributions between the two sets of observations.

Figure 4.10 presents the circumferential and radial strain calculations, along with error bands, for both the original tagging and the cine-generated tagging throughout the entire dynamic cardiac cycle across all time frames. The curve shows the group's average result (real line) and the standard deviation (shadow areas). Beyond the observation of consistency, we can find that when tag labels have faded after the diastolic

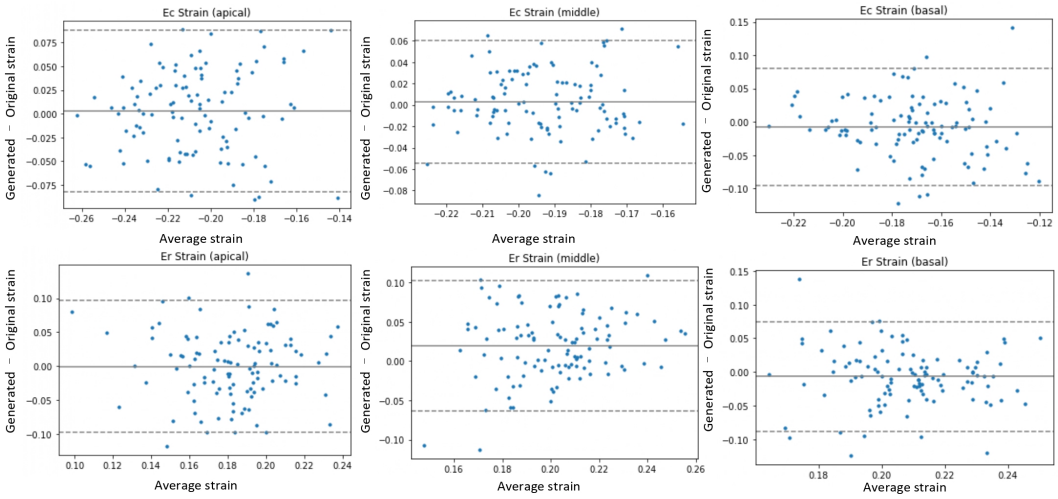

Figure 4.8: Bland-Altman plot of ES LV strain. The strain values obtained from the tagging images generated from the cine images are compared to the strain values obtained from the original tagging images. The first row shows the circumferential strain for three different SAX slices; the second row shows the radial strain. Solid lines represent mean differences; dashed lines represent 95% limits of agreement (mean difference $\pm$ 1.96 $\times$ standard deviation of differences).

Table 4.2: Quantitative numerical comparison of the results generated on the test process, comparing the strain values obtained from the tagging images generated from the cine images and the strain values obtained from the original tagging images, including circumferential strain E_c and radial strains E_r on the three slices (basal, middle and apical).

Region	Strain	Original tagging	Generated tagging
Basal	E_c	-0.2057 ± 0.030	-0.1897 ± 0.029
	E_r	0.2043 ± 0.026	0.2112 ± 0.028
Middle	E_c	-0.2057 ± 0.031	-0.1936 ± 0.024
	E_r	0.1898 ± 0.033	0.1926 ± 0.032
Apical	E_c	-0.2116 ± 0.031	-0.2017 ± 0.030
	E_r	0.1716 ± 0.037	0.1744 ± 0.035

period, the errors of the two often increase at the end of the cine sequence.

To further assess the validity of our findings, we conducted a comparison with the reference values reported by Ferdian *et al.* [174], using identical subjects from the UKBB. We employ Bland-Altman analysis to measure consistency. It is worth noting that the estimates of the corresponding strain values of the two adopt the same method. This approach integrates the detection and tracking of myocardial landmarks via an RNN and CNN, followed by strain calculation based on the motion of these landmarks. Given that the reference values exclusively include circumferential strain, we focused our comparison on the consistency of circumferential strain across three slices (apical, middle, and basal). The results are shown in Figure 4.11, showing that most data points are within the 95% consistency limit. This indicates that our results are close to the reference values in [174] and consistent with the range of myocardial circumferential strain, and the results are of good consistency. It can be observed that the mean difference (bias) in the three SAX slices (apical, middle, and basal) is close to zero, indicating that the systematic bias of the two in different SAX slices is very small. For data points with higher or lower strain, the distribution of the difference does not show an obvious trend, indicating that the consistency of the two methods is not significantly affected by the strain value. The results prove the feasibility of the cine-

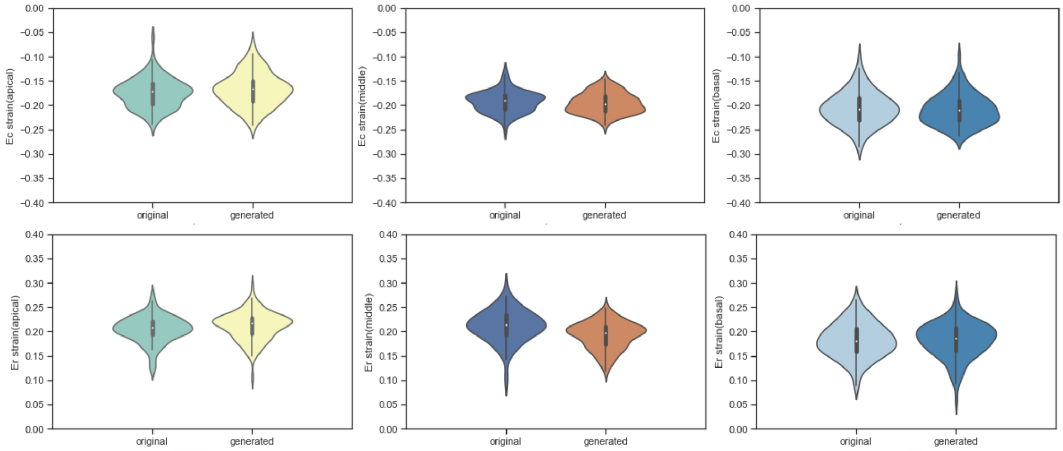

Figure 4.9: Comparative violin plots displaying the difference in distribution between the strain values obtained from the tagging images generated from the cine images and those obtained from the original tagging images.

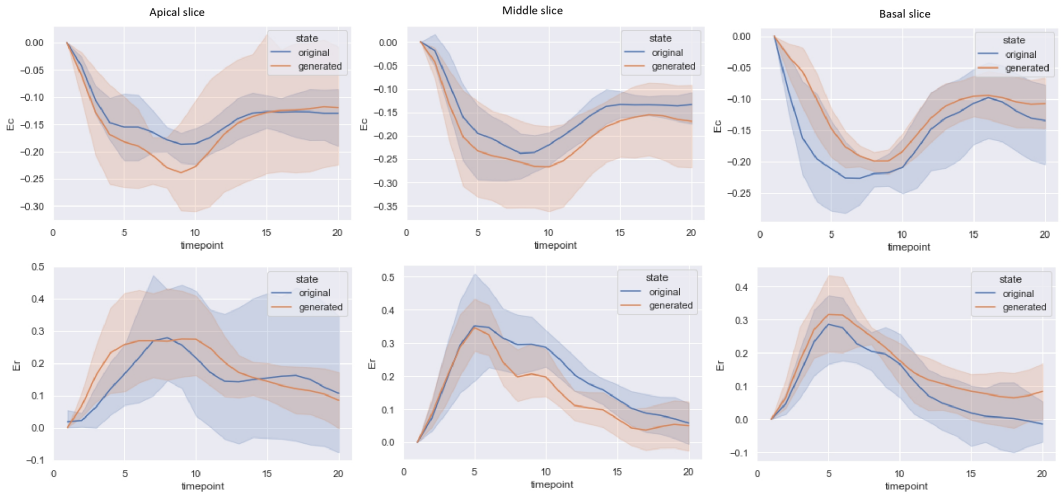

Figure 4.10: Circumferential strain E_c (top row) and radial strain E_r (bottom row) of the original and generated tagging images from the cine CMR images in the test dataset are presented across time with error bands for all time frames in the apical, middle, and basal slices.

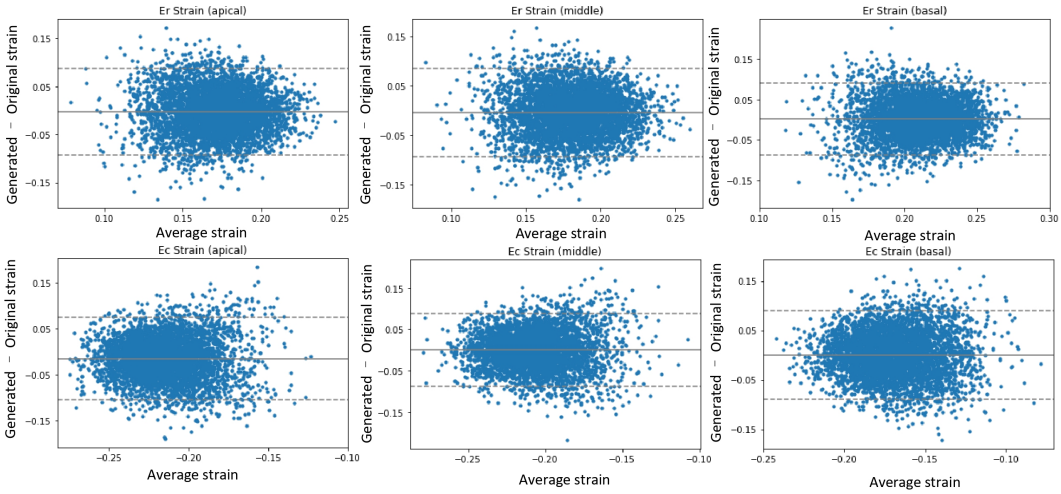

Figure 4.11: Bland-Altman plot of ES LV circumferential strains. Circumferential strain values obtained from tagging images generated from cine images for the same subjects were compared to those returned reference from the UK BioBank. Three different SAX slices are shown from left to right; Solid lines represent mean differences; dashed lines represent 95% limits of agreement (mean ± 1.96 × standard deviation).

generated tagging strain estimation method and provide a stable measurement basis for subsequent research, indicating that this method can be used as an alternative reference.

4.3.9 Ablation study

In this study, we performed an ablation analysis of the proposed methods, focusing on investigating latent space dimensions, because it is an important parameter in our smcVAE model. We have changed the size of the latent space D based on a set of values $\{128, 256, 512, 1024, 2048\}$, and compared their impact on the image quality on the test set. The results, presented in Table 4.3, include both the average and standard deviation for each metric. As D increases, there is a noticeable enhancement in the quality of the generated tagging results, attributable to the richer data representation. Table 4.3 shows that for PSNR and SSIM metrics, $D = 1024$ yielded the highest results, while for RMSE, $D = 2048$ achieved the best performance. This suggests that $D = 1024$

is the preferable choice here.

Latent Dimension (D)	**PSNR (dB)**	**SSIM**	**RMSE**
128d	23.96 ± 4.5	0.71 ± 0.18	0.17 ± 0.09
256d	25.24 ± 4.7	0.74 ± 0.16	0.17 ± 0.08
512d	25.17 ± 4.6	0.76 ± 0.12	0.16 ± 0.06
1024d	**26.62 ± 4.6**	**0.79 ± 0.11**	0.16 ± 0.06
2048d	25.94 ± 4.4	0.77 ± 0.13	**0.16 ± 0.04**

Table 4.3: Ablation study results on the effects of the latent space dimension D, with comparisons in terms of PSNR, SSIM, and RMSE (mean ± standard deviation).

In addition, we have assessed the influence of convolutional layers on the performance of the smcVAE model. By varying the number of convolutional layers, we explored the contribution of network depth to model performance, helping to identify the optimal architecture while also gaining insight into how convolutional layer complexity affects different evaluation metrics. Different numbers of convolutional layers L, $\{5, 6, 7, 8\}$, were considered and their impact is investigated on the model performance on the test data. The outcomes, depicted in Table 4.4, include the mean and standard deviation for each configuration. The results indicate that the number of convolutional layers L could affects the quality of the generated tagging results. As shown in Table 4.4, for PSNR and SSIM metrics, $L = 7$ yields the highest results, while for RMSE, $L = 5$ achieves the best performance, suggesting $L = 7$ as the optimal choice in our model.

Convolutional Layers (L)	**PSNR (dB)**	**SSIM**	**RMSE**
5	25.56 ± 4.5	0.71 ± 0.15	**0.16 ± 0.04**
6	24.16 ± 4.8	0.69 ± 0.14	0.18 ± 0.07
7	**26.62 ± 4.6**	**0.79 ± 0.11**	0.17 ± 0.06
8	26.21 ± 4.6	0.76 ± 0.11	0.17 ± 0.07

Table 4.4: Ablation study results on the effects of the convolutional layers L, with comparisons in terms of PSNR, SSIM, and RMSE (mean ± standard deviation).

4.4 Discussions

Strain estimation from tagging images of cardiovascular MRI has been a challenging problem. In this study, we introduce a deep learning framework for estimating regional myocardial strain through cine-to-tagging CMR image synthesis. The testing results prove that the method is feasible. Our findings indicate that the proposed framework could generate high-quality tagging images for accurately estimating local myocardial strain, and the reverse is also true. This method holds potential as a reference and evaluation tool in practical applications, particularly when conventional cardiac imaging sequences are used for strain analysis.

Qualitative evaluation verification has obtained visually satisfactory results, while quantitative evaluations reveal that multi-channel input outperforms single-channel input. This suggests that the joint input model of original channel information is better than single-channel missing reconstruction. Even if no original channel information is provided to the model, our model can recover the data's joint trajectory learned in the latent space to reconstruct the missing channel, which cannot be achieved in single-channel models.

As far as we are aware, this study is the first to synthesise tagging images from cine CMR and apply them to strain analysis. The smcVAE model can be extended to generate other more channel sequence information, not just cine and tagging images. Furthermore, the study is not limited to image synthesis. It demonstrates the value of synthetic tagging CMR images for quantifying myocardial strain. The model was trained and validated using data from the UK BioBank database, which offers a homogeneous imaging protocol and primarily consists of healthy participants. This is a limitation in this current work, as additional heterogeneous datasets would enhance the robustness of the model, for example, adding data from different imaging protocols. Another limitation of our current work is the fixed time frames in the tagging channel ($n = 20$). After obtaining the generated results, the models perform 2D myocardial motion and tag tracking, with myocardial deformation constrained to a 2D plane. However, accurately simulating cardiac motion necessitates considering the three-dimensional cardiac structure and the alignment of cardiac fibers to fully capture the heart's true motion. This limitation stems from the availability of clinically acquired imaging data, which primarily consists of 2D acquisitions of tagging CMR with only three slices. For future work, we plan to extend this research by integrating three

different slices of tagging images to enable 3D analysis. This will involve transitioning from a 2D convolutional network structure to a 3D convolutional network to account for more accurate and intricate cardiac motion. Extending this work to 3D may present challenges, such as heightened computational demands involved in processing 3D data and implementing 3D networks.

In this work, we have proven that the method is capable of generating tagging CMR solely from cine CMR, enabling myocardial motion tracking and strain calculations. Different from the existing tagging image strain estimation method, the method we proposed effectively solved the problem of extensive clinical applications that need to be obtained by obtaining additional sequences. Currently, myocardial strain estimation primarily focuses on calculating mean global strain values. In the future, we aim to extend our method to estimate myocardial segmental strain by assigning segmental labels to each landmark, following the standardized guidelines of the American Heart Association [209].

The smcVAE used in the model has great potential for joint analysis of the combined analysis of the heterogeneous data. It is not limited to image data of different sequences, but also can be used in clinical population data. This is also our future research direction. Additionally, the smcVAE model in this work was trained using general baseline parameters. Future investigations will explore the potential for performance enhancement through finer-grained parameter tuning strategies.

4.5 Conclusion

In this work, we introduced a deep learning framework designed to jointly learn from cine CMR and tagging CMR data, enabling the generation of tagging CMR images solely from cine CMR sequences and the subsequent estimation of myocardial strain. The purpose of this work is to use only conventional clinical acquisition of cine image sequences to estimate myocardial motion and estimate the radial and circumferential strain for the whole cardiac frames in the heart, the strain results should be similar to the strain results of the quantification of the original tagging CMR. Our experimental results confirm the effectiveness of this approach. This study represents a pioneering concept in the field, opening a new avenue for research in myocardial strain calculation.

Chapter 5

Synthesising 3D cine CMR images and corresponding segmentation masks using a latent diffusion model

In this chapter, we propose a novel pipeline for the generation of synthetic full spatial cine CMR images via a latent Denoising DDIM. These synthetic images can be used as viable alternatives to real data in deep learning model training for downstream cardiac image analysis tasks such as cardiac segmentation. To demonstrate the effectiveness of this approach, we generated synthetic CMR images along with their corresponding segmentation masks. We evaluated model performance using a variety of methods, including generated image fidelity, diversity and calculated the volumes of the generated segmentation masks and compare it with the real segmentation masks. The proposed pipeline has the potential to be widely applied to other tasks in various medical imaging modalities. Effective and efficient generation of 3D cine cardiac images with corresponding segmentation masks can supplement real patient datasets and help reduce the burden of manually annotating images.

5.1 Introduction

Cine CMR [110] is currently considered the gold standard for assessing cardiac function. However, the widespread use of this imaging technology is limited by the high operational costs of the image acquisition process. In order to fully realize the utility of cine CMR in the clinic, it is critical to address challenges related to data collection and patient factors. In recent years, deep learning has shown great potential in medical image analysis [210]. However, deep learning also faces some challenges, especially the need for large datasets during model training. Specifically, the effective training of deep learning models relies on rich and diverse datasets, these datasets need to contain different physiological or pathological conditions of various patients to ensure that the models have wide applicability and high accuracy. In addition, the high quality of data and the accuracy of annotations are also key factors in model performance. Therefore, in order to fully utilize deep learning technology to improve the clinical utility of cine CMR, many studies have focused on data collection, model optimization, and other aspects.

Previous solutions to the limited availability of annotated training data by performing data augmentation using spatial and/or intensity transformations, and using generative models (e.g., GAN [211] and VAE [212]) to create synthetic data alongside their annotations. Data augmentation techniques generate diverse training samples by

applying various transformations to existing images, such as rotation, translation, scaling, and adjusting contrast. These transformations can effectively increase the amount and diversity of training data, thereby improving the generalization ability of the model. GAN and VAE are two major generative models used to generate synthetic data. GAN can generate high-quality images, but they often suffer from instability issues during training, which can lead to inconsistent quality of generated synthetic data. In contrast, VAE generates synthetic data by mapping data into a continuous latent space and reconstructing images from it. Although VAE is relatively stable during training, its main drawback is that it cannot capture complex multimodal data distributions. This often results in blurry generated images that lack fine-grained anatomical details visible in real medical images. This ambiguity limits the application of VAE in generating high-fidelity medical images.

Recently, diffusion models [72] have overcome many of the shortcomings of previous generative models and have been proposed as the state-of-the-art approach for generating synthetic data. Diffusion models learn a noise inversion process and then recover images from randomly sampled noise. This approach has been applied in the medical field, including brain neuroimaging [213] and histopathology [214]. Although diffusion models have shown great potential in medical image generation, research on medical images is still relatively limited. In this paper, we propose a latent DDIM [215] for synthesising 3D cine CMR and corresponding segmentation masks. In our opinion, this is an innovative approach which has not been explored in previous studies. The advantage of DDIM is that it can generate high-quality 3D cine CMR images and generate corresponding segmentation masks at the same time. These synthesised images and segmentation masks can be used for model training for various downstream tasks, such as cardiac structure analysis and functional assessment. By using DDIM, we can expand the training dataset and solve the problem of limited annotated data, and improve the performance and generalization ability of deep learning models.

5.2 Methodology

Figure 5.1 shows a schematic diagram of the overall approach, which includes a pretrained VAE and a DDIM. The core idea of the approach is to compress high-dimensional 3D cine CMR and its corresponding segmentation mask data into a low-dimensional

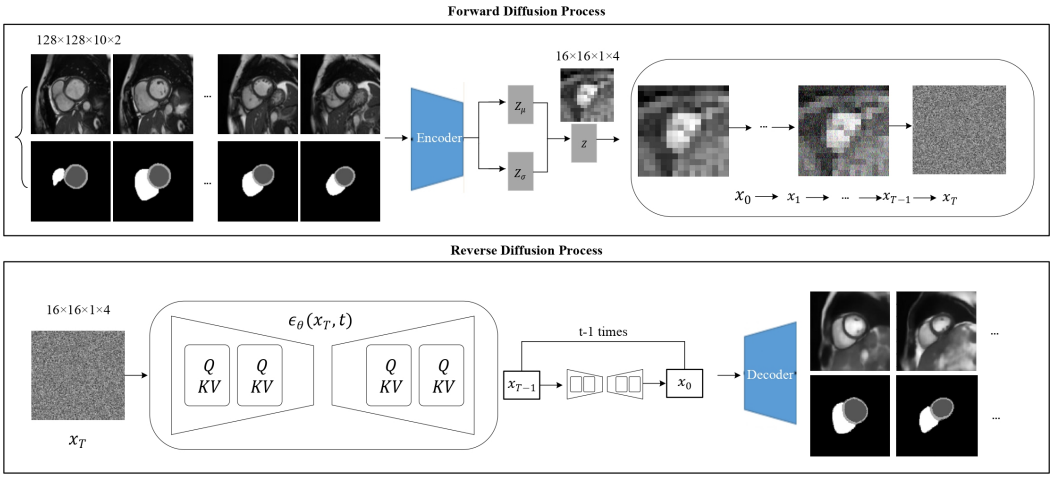

Figure 5.1: **Schematic diagram of the latent diffusion framework.** z_0: latent features of the VAE, z_T: standard Gaussian, t: time step, ϵ_θ: noise added to observed data.

latent space, and to capture and generate high-level semantic information in the latent space. First, the VAE is pre-trained to learn a low-dimensional latent embedding of the cine CMR images and their corresponding segmentation masks. The encoder of the VAE maps the high-dimensional 3D cine CMR images and segmentation masks into a low-dimensional latent space, thereby achieving data compression and feature extraction. The decoder can reconstruct the original high-dimensional data from the latent space. This process can effectively represent complex 3D image data, and preserve important anatomical and functional information. After completing the steps of training the VAE, the DDIM captures high-level semantic information by modeling the data distribution in the latent space. The working principle of DDIM is to start with random noise and gradually reduce the noise through multiple steps of iteration to recover the samples in the latent space. This process is similar to the back-diffusion process and aims to generate realistic low-dimensional latent representations. These latent representations are then converted back to high-dimensional 3D cine CMR images and corresponding segmentation masks through the VAE decoder.

5.2.1 Latent diffusion model (LDM)

The LDM consists of two parts: the first part is a VAE, which compresses the high-dimensional data of 3D cine CMR images and their corresponding segmentation masks into a low-dimensional latent space. The second part is a DDIM, which uses a 3D deep neural network to learn the back-diffusion process. In the first part VAE, the encoder is responsible for mapping the high-dimensional 3D cine CMR images and segmentation masks into a low-dimensional latent space. This process effectively compresses the data, allowing the complex 3D image data to be represented in a lower dimension and retaining important structural information. The decoder of VAE is responsible for reconstructing the high-dimensional original data from the low-dimensional latent space, thereby achieving data compression and restoration. The second part DDIM, focuses on generating high-quality image data and its corresponding segmentation masks through the back-diffusion process. The forward diffusion process is a process of gradually adding Gaussian random noise until the original data is completely transformed into pure Gaussian noise. The back-diffusion process is a process of gradually reducing noise, and recovering the original image and segmentation mask from the noise. DDIM uses a 3D deep neural network in this process, gradually reducing noise through multi-step iterations, and finally generating realistic image data. Unlike traditional diffusion models, a notable feature of DDIM is that its diffusion process does not have to strictly follow the Markov chain [216]. This means that DDIM can improve the efficiency of image generation by reducing the sampling steps and can flexibly adjust the execution time step of the model. This improvement speeds up the image generation process and improves the quality and detail fidelity of the generated images.

As the first part of the model, VAE is responsible for perceptual compression and adopts a network structure of 3D convolution and 3D attention layers. During training, the loss function of VAE includes the reconstruction L1 loss of the image channel, the sparse cross entropy loss of the segmentation mask channel, and the Kullback-Leibler (KL) regularization loss. The L1 loss is used to measure the difference between the original image and the reconstructed image, ensuring that VAE can accurately reconstruct the input 3D cine CMR image. The sparse cross entropy loss is used to evaluate the reconstruction quality of the segmentation mask. This part of the loss can better handle the sparse label problem in the classification task and ensure the accuracy of the segmentation mask. The KL loss is used to minimize the difference between the

latent space and the standard normal distribution. This part of the loss helps VAE learn a continuous and structured latent space, making the generation process more stable and controllable. The latent space dimension used is $16 \times 16 \times 1 \times 4$. This choice takes into account the compactness and information content of the latent space, ensuring that key features are not lost while compressing the data. The hyperparameters for VAE training are set as follows: epochs = 300, KL weight is $1e^{-4}$, batch size is 2, and learning rate is $1e^{-4}$. During training, we select the best model based on the minimum validation loss on the validation set.

DDIM

The diffusion process gradually transforms the original data into pure Gaussian noise by gradually adding noise. This process can be described by the following formula:

$$q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t} x_{t-1}, (1 - \alpha_t)\mathbf{I}) \tag{5.1}$$

where x_t represents the data at the t step, α_t is a scaling factor less than 1, $\mathbf{I}$ is the unit matrix.

The reverse diffusion process starts from the noise and restores the original data by gradually removing noise. The traditional diffusion model uses Markov chain for denoising, while DDIM improves efficiency by simplifying this process. The reverse process can be described by the following formula:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 \mathbf{I}) \tag{5.2}$$

where $\mu_\theta(x_t, t)$ is the predicted mean based on the current data and time step, and σ_t is the standard deviation of the noise.

DDIM proposes a non-Markov chain denoising method to generate samples through a deterministic denoising process. The formula is as follows:

$$x_{t-1} = \sqrt{\alpha_{t-1}} \left(\frac{x_t - \sqrt{1 - \alpha_t} \epsilon_\theta(x_t, t)}{\sqrt{\alpha_t}} \right) + \sqrt{1 - \alpha_{t-1}} \epsilon_\theta(x_t, t) \tag{5.3}$$

where $\epsilon_\theta(x_t, t)$ is the output of the noise prediction network, which represents the noise component in the current data x_t at the time step t.

The latent space representation of VAE (once trained) is used as the input for training the DDIM. The noise prediction network in the DDIM uses a 3D U-net with

self-attention mechanism and is trained by minimising a reconstruction L1 loss, evaluated as,

$$\mathcal{L}_{DDIM} := \mathbb{E}_{z_0,t,\epsilon\sim\mathcal{N}(0,1)}[\| \epsilon - \epsilon_\theta(z_t, t) \|_1] \tag{5.4}$$

where z_t is the result of applying the forward diffusion process to the latent space obtained from the pre-trained VAE. The forward diffusion process gradually transforms the initial latent representation into pure Gaussian noise by gradually adding noise. In the backward diffusion process, the goal is to start with pure Gaussian noise and gradually denoise it to recover the initial latent representation z_0. The output of the backward diffusion process z_0 is generated by a 3D U-net noise prediction network, which predicts the noise present in the current data and denoises it accordingly. The denoised latent representation z_0 forms the input of the VAE decoder to reconstruct the original high-dimensional data. The bottom of Figure 5.1 shows this process, where the prediction network and the decoder are jointly trained to simulate the backward diffusion process. The backward diffusion process is trained with batch size of 32, diffusion time step of 1000, learning rate of $1e^{-5}$, beta of the ADAM optimizer of (0.9, 0.99), and total of 10,000 training iterations.

The VAE decoder converts the denoised latent representation z_0 back into a high-dimensional 3D image and segmentation mask, thereby achieving high-fidelity reconstruction of the original data. After training, the back-diffusion process is sampled to generate the new latent representation. The latent representation is processed by the VAE decoder to generate the synthetic 3D cine CMR image and its corresponding segmentation mask. This process generates high-quality image data, and provides accurate segmentation information at the same time, which helps model training and performance improvement of various downstream tasks.

5.3 Experiments and Results

5.3.1 Dataset and Data Preprocessing

The experimental data comes from 927 SAX cine CMR image sequences from the UK Biobank, including 731 paired images and segmentation masks for training, and 196 segmentation masks for testing. The segmentation masks are based on the method of [217], with automatic segmentation at ED and ES, and label annotations for LV, LV

Myo, and RV. These SAX image stacks usually consist of a variable number of image slices arranged along the axial direction.

To ensure data consistency and efficient model training, all training images were preprocessed. First, the region of interest (ROI) was selected on the images to ensure that each image contained the key parts of the heart. Then, the images were cropped to the same size, i.e., $128 \times 128 \times 10$. This fixed size standardized the data and simplified the subsequent model training process. To make the intensity distribution of the images consistent, the images were intensity normalized and adjusted to the range [0, 1]. During training, the input data is of size $128 \times 128 \times 10 \times 2$, which contains images and corresponding segmentation masks. These data are stacked in the channel dimension so that the model can process both image and segmentation information simultaneously.

5.3.2 Experimental Setting

Using the trained LDM, we first generated 1000 3D cine CMR images and their corresponding segmentation masks. To fully evaluate the performance of the generated images and segmentation masks, we used both quantitative and qualitative methods and compared them with real data from the UK Biobank. To quantitatively evaluate and benchmark our proposed LDM, we compared it with two baseline generative methods: 3D VAE and 3D least squares generative adversarial network (3D LSGAN) [218].

Specifically, for the 3D VAE, we implemented our model architecture following the guidelines provided by [219], adapting it explicitly for handling 3D cardiac cine CMR data. The encoder and decoder utilized 3D convolutional and transposed convolutional layers, respectively. The loss function was composed of the reconstruction loss (mean squared error) and a KL-divergence term. The hyperparameters (such as latent dimension size, learning rate, and batch size) were selected through cross-validation and matched those used for training our proposed LDM to ensure fair comparison.

For the 3D LSGAN, we adapted and extended an existing publicly available implementation from a third-party source, specifically the PyTorch implementation released by Xudong Mao et al. on GitHub (`https://github.com/xudonmao/LSGAN`). This codebase was originally designed for 2D image generation tasks, and we explicitly modified it to support 3D convolutional operations suitable for cine CMR imaging. This

included adjustments to convolutional layers, batch normalization layers, and optimization routines to ensure compatibility and effective training on our 3D cardiac cine datasets.

5.3.3 Qualitative results

Figure 5.2 shows examples of real and generated images for different slices of different real and synthetic objects, and their corresponding segmentation masks, respectively. It can be observed that the synthetic images generated by our model achieve a good level of anatomical fidelity. These synthetic images have good quality and visually realistic appearance, and their corresponding segmentation masks also represent the structures of interest.

In addition, Figure 5.2 also shows the diversity of generated images and segmentation masks in terms of ventricular shape and appearance. This diversity shows that the model can generate a single type of anatomical structure and also cover the differences between different individuals. The generated synthetic images are consistent with the real images in terms of grayscale level, texture details, and edge sharpness. By comparison, it can be found that the myocardium and cardiac chamber structures in the generated images are visually almost the same as the real images. This proves the accuracy and fidelity of the model in data generation and also demonstrates its potential application value in medical image analysis.

5.3.4 Quantitative results

In addition to visual inspection, we also performed quantitative evaluation of the generated synthetic data. We evaluated the performance of VAE and DDIM in our method separately to fully understand the performance of the model in generating high-quality images and segmentation masks. For the VAE, we separately calculated the SSIM and PSNR for the reconstructed images of the test set, as well as the Dice coefficient for the corresponding segmentation masks. These metrics are evaluated to measure the impact of image quality degradation caused by VAE, as this will affect the generation performance of DDIM. By calculating these metrics, the SSIM of the VAE reconstructed image is 0.78 ± 0.03, the PSNR is 23.94 ± 2.76, and the Dice coefficient of the corresponding segmentation mask is 0.91 ± 0.01.

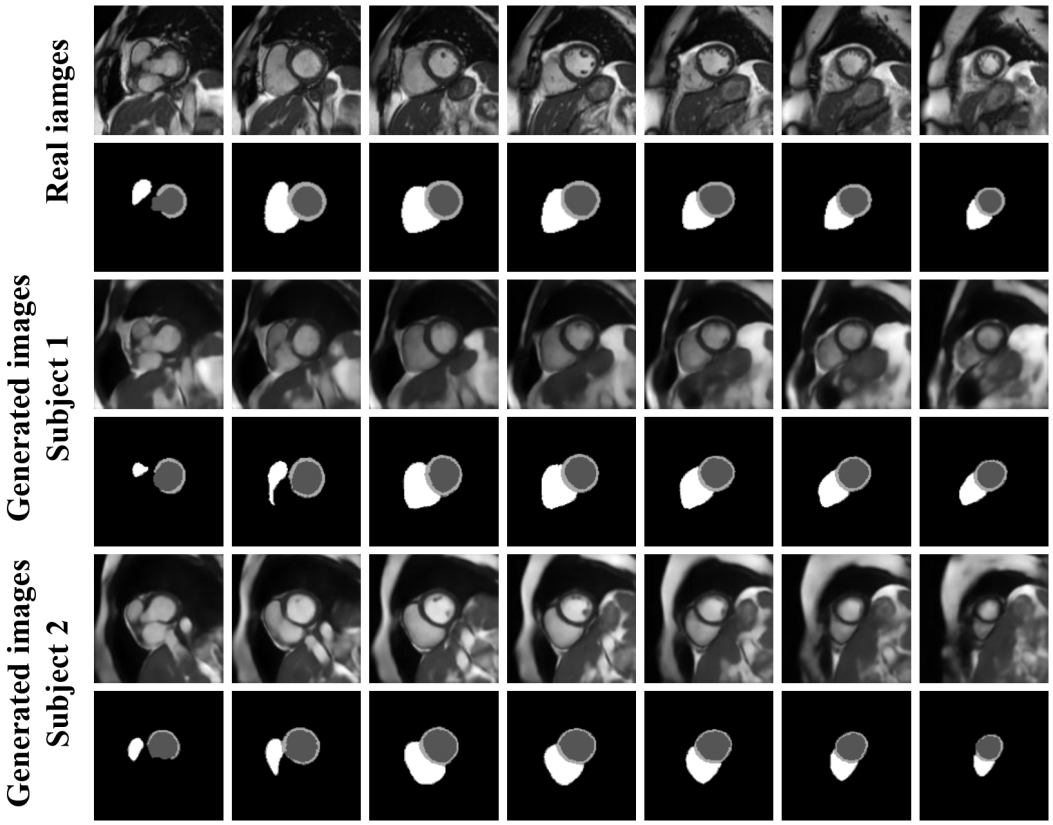

Figure 5.2: Comparison of real images and generated results, examples of full spatial images of different subjects and corresponding segmentation masks.

Table 5.1: Results of quantitative evaluation of FID, FRD, IP, IR, MS-SSIM and 4-G-R SSIM on real test data, comparison on synthetic data generated using 3D VAE, 3D LSGAN and our model.

Method	FID↓	FRD↓	IP ↑	IR↑	MS-SSIM↓	4-G-R SSIM↓
3D VAE	32.74	3.21	0.76	0.74	0.78	0.81
3D LSGAN	48.53	3.47	0.51	0.43	0.91	0.88
Ours	28.37	2.92	0.86	0.76	0.67	0.37

$$\text{Dice} = \frac{2|A \cap B|}{|A| + |B|} \tag{5.5}$$

To comprehensively evaluate the overall generative performance of the DDIM model, we calculate several metrics: Fréchet Inception Distance (FID) scores [220], Fréchet ResNet Distance (FRD) scores [221], Improved Precision (IP), Improved Recall (IR) [222], multiscale structural similarity metric (MS-SSIM) [223] and Four-Grid-Recursive Structural Similarity Index metric (4-G-R SSIM). Quantitative evaluation of these metrics provides a comprehensive understanding of the model's generative performance. The performance scores of the models are shown in Table 5.1.

The FID score measures the similarity between the generated image and the real image in the feature space, using the pre-trained Inception-V3 [224] as the feature extractor. The lower FID score indicates the higher perceived image quality generated.

$$\text{FID}(X, Y) = \|\mu_X - \mu_Y\|^2 + \text{Tr}(\Sigma_X + \Sigma_Y - 2(\Sigma_X \Sigma_Y)^{1/2}) \tag{5.6}$$

where X and Y denote the feature space representation of the generated image and the real image, respectively. μ_X and μ_Y denote the mean vector of the generated image and the real image in the feature space. Σ_X and Σ_Y represent the covariance matrices of the generated image and the real image in the feature space, respectively.

FRD is similar to FID, but is calculated using the pre-trained ResNet50 as the feature extractor, and the similarity judgment between real images and generated images is more consistent with humans. The lower FRD score indicates that the generated image is more similar to the real image. IP and IR evaluate the quality and coverage of measured image generated samples by forming explicit non-parametric representations of real data and generated data manifolds, IP represents the probability that the generated image falls within the support range of the real image manifold, and IR represents the probability that the real image belongs to the generated image manifold. MS-SSIM

and 4-G-R SSIM are used to evaluate the structural similarity of the generated images at different scales. They are calculated by taking the average of the generation results. The lower the MS-SSIM score, the better the diversity of the generated images.

$$\text{MS-SSIM}(x, y) = \left[\prod_{j=1}^{M} \text{SSIM}_j(x, y) \right]^{1/M} \tag{5.7}$$

Compared with the images generated by the competing methods, the MS-SSIM and 4-G-R SSIM scores of the images generated by our model are lower, indicating that the generated image diversity is better. Additionally, to compare the generated segmentation masks and the real segmentation masks, we use box plots to visualize the volume differences of LV, Myo and RV, as shown in Figure 5.3. The generated data n are 1000, 2000, and 4000.

In addition, to visualize the distribution of synthetic data and real training and test datasets, we used t-distributed Stochastic Neighbor Embedding (t-SNE), which is a method of dimensionality reduction and visualization for projecting high-dimensional data into a low-dimensional space. Figure 5.4 shows the visualization of the dataset distribution after t-SNE.

5.4 Discussions

Data augmentation is an important part of training robust machine learning models and is also widely used in the field of medical imaging. Traditional data augmentation techniques, such as rotation, flipping, and scaling, can increase data diversity through simple image transformations, thereby improving the performance and robustness of the model. However, these methods have limitations when dealing with complex data patterns. In contrast, generative models can generate realistic new data by learning the inherent patterns of the data, thereby effectively improving the generalization ability and robustness of the model.

In the field of medical imaging, the application of generative models has greatly expanded the possibilities of data augmentation. For example, generative models can generate high-quality medical images and their corresponding annotations, substantially reducing the workload and human errors that rely on manual annotation [225][226]. At the same time, these generated data can also be used to enrich real training datasets and further improve the performance of the model in various downstream tasks, such

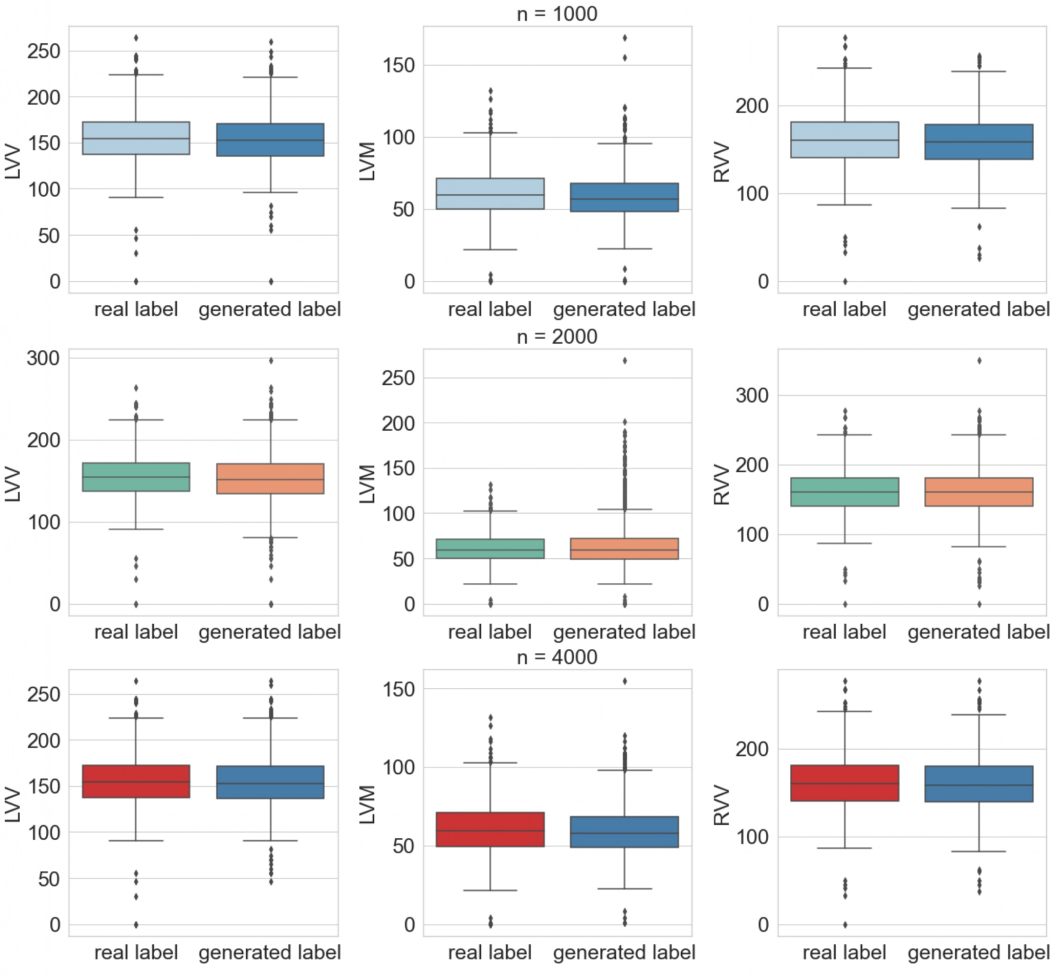

Figure 5.3: Box plot for real and generated segmentation masks of volumes for the LV (LVV), LV Myo (LVM) and RV (RVV) with n = 1000, 2000 and 4000, respectively.

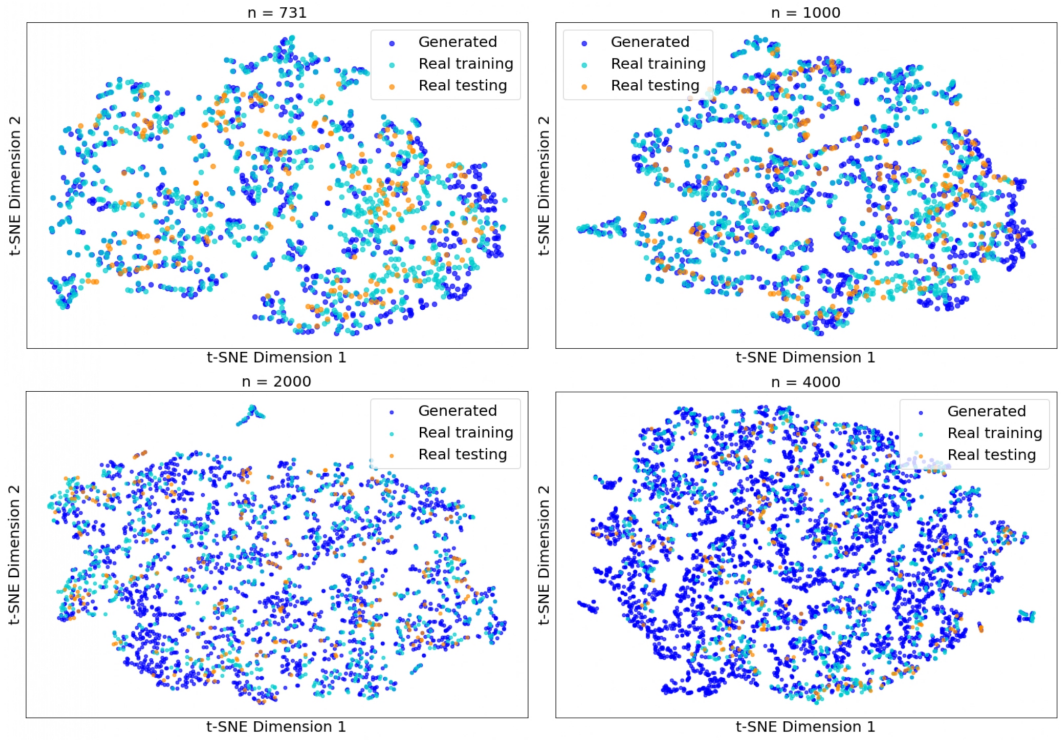

Figure 5.4: t-SNE for the real and generated synthetic images with n = 731, 1000, 2000, and 4000, respectively.

as medical image segmentation and registration.

In this study, we use generative models to simultaneously generate medical images and corresponding cardiac segmentation masks. The generated data can reduce the burden of manual annotation, reduce subjective errors and variability, and also effectively enhance the diversity of training data, thereby improving the performance of the model in cardiac image segmentation and other related tasks [227][228].

Due to the computational complexity of analysing high-dimensional data, we use VAE in the first part of the model to learn low-dimensional representations of the data. This method can effectively reduce the computational complexity by encoding high-dimensional data into a low-dimensional latent space and then decoding it back to high-dimensional data from the latent space. However, this process inevitably leads to the loss of image details and affects the quality of the model generation results.

Recently, some studies have explored the use of LDM to generate medical images, including brain images, echocardiograms, and histopathology images [229]. The LDM can generate higher-quality images by gradually adding noise to the image and learning the denoising process. In this study, we use 3D cine CMR images for experiments. Compared with traditional methods such as 3D VAE and 3D LSGAN, our method achieves improvements in generation performance. Specifically, our method can more accurately capture the details of cardiac images, and the generated images have improvements in visual quality and structural consistency.

Our research shows that the use of advanced generative model technology can effectively make up for the shortcomings of traditional methods in detail preservation and generation quality. Through comparative experiments, we verified the superiority of the proposed method in generating 3D CMR images. This provides a new method for the generation of cardiac medical images, and provides higher quality data support for related downstream tasks (such as cardiac image segmentation and diagnosis), which helps to improve the performance of automated medical image analysis.

5.5 Conclusion

In this study, LDM was used to achieve efficient and high-quality simultaneous generation of cine CMR images and their corresponding biventricular segmentation masks. Our method outperforms traditional GAN or VAE methods in terms of the diversity and fidelity of synthesised cine CMR images and their segmentation masks. LDM can

generate imaging data with high diversity, covering a wide range of cardiac structural variability, which is essential for training more robust and generalizable machine learning models. In addition, our method also has the ability to simultaneously generate images and their corresponding segmentation masks, which could reduces the workload and errors of manual annotation and improves the consistency and accuracy of data annotation. The generated high-quality, annotated CMR image dataset can be widely used in various medical image analysis tasks, including but not limited to cardiac structure segmentation, disease diagnosis, treatment effect evaluation, etc. Overall, this study demonstrates the great potential of LDM in the field of cardiac image generation. By generating high-quality and diverse CMR images and their segmentation masks, our method provides strong data support for medical image analysis, helps improve the accuracy and efficiency of automated analysis, and promotes the development and application of medical imaging technology.

Chapter 6

Conditional 4D spatio-temporal latent diffusion generative model for cine CMR imaging synthesis

Generating synthetic CMR images that maintain clinical relevance and reflect real subject demographics is essential for advancing medical image analysis and computational modeling. In addition, exploring the relationship between cardiac images and non-imaging clinical factors such as demographic information and diseases is also a key issue in cardiac image analysis. In this chapter, we propose a conditional latent diffusion generative model that aims to generate 4D spatio-temporal cine CMR images by incorporating non-imaging demographic and clinical data such as age, gender, blood pressure, and lifestyle factors as conditional variables. Our model embeds this information into the latent space using a conditional encoder to guide the generation process. We evaluate the model using various structural similarity and clinical measures, and the results show that it is able to generate realistic and diverse cine CMR images. The model achieves high generation performance, and the synthetic images are closely related to the real data distribution.

6.1 Introduction

Cardiac imaging is essential for diagnosing and managing cardiovascular diseases [230][231]. For example, cine CMR imaging can reveal the anatomy of the cardiac and its dynamic contraction and relaxation patterns [232][33]. Generating CMR images is essential for conducting computer simulation experiments and model training for medical image analysis. The generated data should capture sufficient variability while maintaining plausibility and reflect the clinical characteristics and demographics of the subjects observed in real data. In addition to anatomical details, the heart's dynamic temporal motion provides valuable information for clinical diagnosis and treatment management decisions [25][233]. Developing computational tools to link spatio-temporal imaging features with demographic and clinical data is crucial. Therefore, it is crucial to synthesise images suitable for specific goals conditionally. This study aims to enhance our understanding of spatio-temporal cardiac structure and its relationship with demographic and clinical factors through generative models. We develop a conditional LDM for generating cine CMR images. By providing demographic information and clinical factors as condition variables, our model generates 4D spatio-temporal cine CMR images, which are realistic and align with the ground truth distribution.

Recently, great progress has been made in the field of conditional generative models, largely due to the development of deep learning techniques, including conditional

GAN [234], conditional VAE [235], flow-based model [236][237], and conditional diffusion model [238][239]. These deep learning methods have performed well in efficiently approximating the underlying conditional distribution and generating high-quality samples. The application and improvement of these conditional generative models have promoted many developments in different generation tasks. For example, in the image-to-image translation task, conditional generative models can transform one type of image into another type of image. For example, conditional GANs are widely used to transform one type of image into another form of the image [240][241]. In the task of generating medical images, conditional generative models can generate new CT or MRI images from existing image data [242][243]. Additionally, conditional generative models in text-to-image synthesis can generate images from text descriptions, greatly expanding the application scenarios of image generation [244][245].

In medical imaging, there has been a need for research focused on incorporating non-imaging demographic information and clinical variables into the conditional image synthesis process. Wu *et al.* [246] introduced a class conditional GAN model to synthesise enhanced mammography datasets. However, GAN-based models are known to be prone to mode collapse problems, and the generated images may lack sufficient diversity to capture the full range of anatomical variations present in real datasets. Jung *et al.* [247] proposed a conditional cGAN that can synthesise MR images of different stages of Alzheimer's disease (AD) (i.e., normal, mild cognitive impairment, and AD). While GANs are effective in generating disease progression images, it is often difficult to have fine-grained control over the synthesised features, limiting their clinical interpretability. Xia *et al.* [248] proposed a model to synthesise brain MRI based on age and AD disease stage without relying on longitudinal data. However, the lack of longitudinal consistency may result in unrealistic temporal progression patterns. Biffi *et al.* [249] proposed a Ladder VAE (LVAE) generative model for shape analysis, but VAE tends to produce blurrier images due to the inherent trade-off between reconstruction fidelity and latent space regularization. Segmentation and classification results validated the accuracy of the generated images, but the method did not explicitly incorporate clinical variables beyond anatomy. Reynaud *et al.* [250] proposed a Deep ARtificial Twin-Architecture GeNerAtive Networks (D'ARTAGNAN) model to synthesise 3D echocardiographic videos conditioned on LV EF and image. Although this approach leverages both image and clinical data, it is still limited in its ability

to model high-dimensional spatio-temporal dependencies. Campello *et al.* [251] developed a conditional GAN to synthesise cardiac images of different ages. Amirrajab *et al.* [252] introduced a synthetic framework to synthesise CMR images with variable anatomical and imaging features, but the scalability of this approach to real-world datasets with high inter-subject variability remains unclear. Qiao *et al.* [253] introduced a conditional VAE-based generative model to capture the 4D spatio-temporal anatomy of the heart and its interaction with clinical variables, but the images generated by VAE typically have lower spatial resolution and finer anatomical details compared to the diffusion model. While these studies offer valuable insights into conditional medical image synthesis, research generating spatio-temporal cardiac images based on numerous demographic and clinical factors remains limited.

While these studies provide valuable insights into conditional medical image synthesis, they typically rely on GANs or VAEs and thus have inherent limitations in image quality, diversity, and interpretability. Furthermore, existing spatio-temporal cardiac image generation methods remain limited in their ability to integrate numerous demographic and clinical factors in a coherent and high-fidelity manner. To address these challenges, we proposed a conditional LDM that generates realistic cine CMR images based on non-imaging demographic information and clinical variables such as age, gender, height, weight, smoking status, alcohol status, and systolic and diastolic blood pressure. The model adopts a conditional latent diffusion approach to learn the latent representation in the low-dimensional feature space of cardiac images, and adopts a conditional encoder to embed the demographic and clinical condition information into the conditional latent vector, which is input into the model as guidance information during the diffusion process. The proposed model shows satisfactory variety and accuracy in image generation, evaluated through various similarity metrics, structural overlap, surface distance measurements, and clinical parameters such as ventricular volume and mass. The key contributions of this study can be outlined as follows:

- We introduce a 4D spatio-temporal conditional latent diffusion generative model for cine CMR images that simultaneously considers spatial and temporal variations during the cardiac cycle.
- We train the model using both image-based and non-image-based demographic and clinical information, allowing it to generate cardiac image sequences based on various controllable variables.

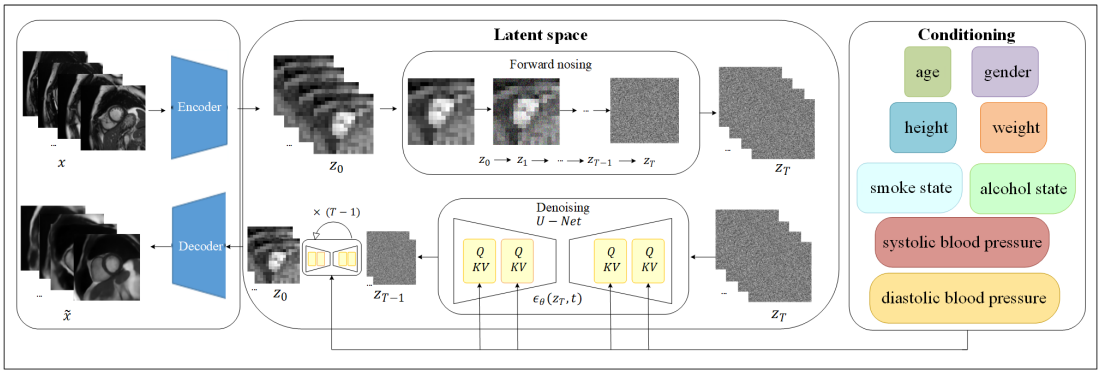

Figure 6.1: **Schematic diagram of the conditional latent diffusion framework.** z_0: latent features of the autoencoder, z_T: standard Gaussian, t: time step, ϵ_θ: noise added to observed data.

- The low-dimensional feature space extracted by the pre-trained autoencoder has a good balance between complexity and detail preservation, and the cross-attention layer introduced in the diffusion model is transformed into a powerful generator with controllable conditions to achieve high-resolution synthesis.
- We show that the model can synthesise highly realistic and varied cardiac MRI sequences that closely align with the ground truth data distribution.

6.2 Methodology

Although diffusion models have demonstrated state-of-the-art performance in image data synthesis and other areas [72][87], generating high-resolution images with these models requires costly function evaluations in pixel space, which is costly to infer and results in huge demands on computing time and computing resources. The LDM model [71] was proposed to address these shortcomings. The LDM performs a diffusion process on a compressed low-dimensional latent space, saving computing resources and speeding up inference while ensuring synthesis quality. The LDM is divided into a compression learning stage and a generation learning stage. The compression learning stage obtains a compressed image space through perceptual learning of the auto-encoding model, which greatly reduces the computational complexity. The proposed conditional LDM uses non-imaging demographic information and clinical data as conditional inputs to generate 4D spatio-temporal CMR images. Figure 6.1 illustrates the overall framework.

6.2.1 LDMs

Diffusion Models are probabilistic frameworks designed to understand and replicate a sample distribution $p(x)$. They achieve this by incrementally removing noise from a variable that is initially normally distributed. This process mirrors training the reverse sequence of a predefined Markov chain of length T. For the task of image generation, the best-performing models [69][254] use a modified variational lower bound on $p(x)$, which closely resembles noise reduction score matching [255]. Essentially, these model functions can be viewed as a sequence of equally weighted denoising autoencoders, denoted as $\epsilon_\theta(x_t, t)$ for $t = 1, 2, ..., T$, here x_t represents a noisy instance of the original input x. The overall target function, which guides the training of these models is simplified as follows:

$$L_{DM} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right], \tag{6.1}$$

where t is uniformly selected from $\{1, ..., T\}$.

The generative model for the latent representation consists of encoder $\mathcal{E}$ and decoder $\mathcal{D}$. The encoder captures an efficient low-dimensional latent space. This space abstracts high-frequency, imperceptible details and is better suited for likelihood-based generative models over high-dimensional pixel space. This allows the model to focus on the most relevant and meaningful aspects of the data, and improving computational efficiency. The overall objective function can be simplified to:

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2\right]. \tag{6.2}$$

The model's neural backbone is implemented using a time-conditional UNet [256]. The time step t is modeled by the sinusoidal embedding method, which maps the time step t to a fixed frequency sine and cosine function space to provide a stable and periodic time representation. The method is defined as follows:

$$\text{TimeEmbedding}(t) = [\sin(\omega t), \cos(\omega t)] \tag{6.3}$$

where ω is a set of predefined frequency parameters, which usually grows exponentially to ensure that different time steps have sufficient distinction in the embedding space. The sinusoidal embedding method is similar to Transformer Positional Encoding [257], which can provide rich temporal information to the model without relying on

additional trainable parameters. This design choice allows the model to leverage temporal information effectively. The forward process is predefined and fixed, the latent variable z_t can be efficiently derived from the encoder $\mathcal{E}$, and samples from the prior distribution $p(z)$, the decoder $\mathcal{D}$ can map these back into the image space in one pass.

6.2.2 Conditioning Mechanisms

Like other generative models [258][235], diffusion models can model conditional distributions in the form of $p(z|y)$. This is achieved using a conditional denoising autoencoder $\epsilon_\theta(z_t, t, y)$, which enables control over the synthesis process via various conditional inputs y, like text descriptions [244], semantic labels [259], or other tasks involving image translation [240]. To enhance the flexibility of the diffusion model as a conditional image generator, its UNet backbone is enhanced by a cross-attention mechanism [260]. The cross-attention mechanism plays a crucial role in integrating conditional information y into the UNet backbone of the diffusion model. It enables the model to selectively focus on relevant aspects of the conditioning signal while generating images. The fundamental idea of cross-attention is to establish relationships between the features extracted from the noisy image z_t and the conditioning information y through learned transformations.

This enhancement is particularly effective for developing attention-based models that handle various input modalities [261]. For preprocessing inputs y from different modalities, introduce a domain-specific encoder τ_θ. This encoder maps y into an intermediate form $\tau_\theta(y) \in \mathbb{R}^{M \times d_\tau}$, which is then passed to the intermediate layers of the UNet through a cross-attention mechanism. The cross attention mechanism is implemented as follows, using the standard scaled dot-product attention formula:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V \tag{6.4}$$

where $Q = W_Q^{(i)} \cdot \varphi_i(z_t)$, represents the feature embeddings from the intermediate layers of the UNet, which capture the current state of the noisy image at time step t. $K = W_K^{(i)} \cdot \tau_\theta(y)$, $V = W_V^{(i)} \cdot \tau_\theta(y)$, K and V are derived from the conditional information y, processed through a domain-specific encoder τ_θ. $\varphi_i(z_t) \in \mathbb{R}^{N \times d_\epsilon^i}$ represents a (flattened) intermediate feature of the UNet implementing ϵ_θ, and the matrices $W_V^{(i)} \in \mathbb{R}^{d \times d_\epsilon^i}$, $W_Q^{(i)} \in \mathbb{R}^{d \times d_\tau}$, and $W_K^{(i)} \in \mathbb{R}^{d \times d_\tau}$ are trainable projection matrices that map features to the appropriate dimensionality for attention computation [260].

The cross-attention mechanism enables the model to integrate conditioning information into the image generation process dynamically. The query matrix Q represents the current state of the image being denoised, while the key matrix K contains relevant information from the conditioning signal y. The attention scores $\frac{QK^T}{\sqrt{d}}$ determine the influence of each conditioning feature on the image, and the weighted sum of values V provides updated features that guide the generation process. This mechanism allows the model to modulate its focus based on different conditioning inputs. For instance, if the conditioning input is text, the attention mechanism aligns textual features with spatial regions in the image to ensure semantic consistency. If the conditioning input is a segmentation mask, it enforces structural coherence, and if the conditioning input is another image, it facilitates style or content transfer.

Using image-conditioning pairs, the conditional LDM via:

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2\right] \tag{6.5}$$

In this process, both the domain-specific encoder τ_θ and the denoising autoencoder ϵ_θ are jointly optimized according to Equation 6.4. This conditioning mechanism is highly flexible, allowing τ_θ to be parameterized by specialized domain models. For instance, when y represents text prompts, τ_θ can be implemented using unmasked transformers [260]. This adaptability ensures that the model can effectively handle various types of input data, enhancing its versatility and performance in different tasks.

The generative model can achieve conditional spatio-temporal sequence generation in the inference phase. The model is conditioned solely on demographic information and clinical variables c, without requiring any image data as input. The sampling process first samples the latent variable $p(z)$ from a simple prior distribution (usually a Gaussian distribution) and connects it with the conditional latent code z_c. Based on the latent variable z and the conditional information c, a diffusion model is used to generate data. The diffusion model gradually transforms the initial noise into the required data through a series of inverse diffusion processes.

6.2.3 Evaluation

To assess the conditional LDM, we employ quantitative evaluation metrics to evaluate the results in addition to visual inspection and utilize clinical metrics to evaluate the similarity of distributions. Additionally, we evaluate the performance of VAE and conditional DDIM separately to gain a comprehensive understanding of the model's

ability to generate high-quality results. For the autoencoder, we first calculate SSIM and PSNR for the reconstructed test set images, as well as the Dice coefficient, Hausdorff distance (HD), and average symmetric surface distance (ASSD) for the corresponding segmentation masks. These metrics are evaluated to measure the impact of image quality degradation caused by the autoencoder, as this will affect the generation performance of the conditional LDM.

To comprehensively assess the generative performance of the conditional LDM, we calculated the following important indicators: including FID score, FRD score, IP and IR, maximum mean discrepancy (MMD), and MS-SSIM and 4-G-R SSIM. Given two sets of sample sets $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_m\}$, the empirical MMD is defined as:

$$\mathrm{MMD}^2(X, Y) = \frac{1}{n(n-1)} \sum_{i \neq j} k(x_i, x_j) + \frac{1}{m(m-1)} \sum_{i \neq j} k(y_i, y_j) - \frac{2}{nm} \sum_{i,j} k(x_i, y_j) \quad (6.6)$$

where k is a kernel function, commonly the Gaussian kernel: $k(x, y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$.

Secondly, we also extracted five imaging phenotypes from the generated images to further evaluate the performance of the model: including LVM, left ventricular and right ventricular end-diastolic volume (LVEDV and RVEDV), left ventricular and right ventricular end-systolic volume (LVESV and RVESV). These phenotype data were compared with real data sharing the same demographic information and clinical variables to evaluate the authenticity and accuracy of the generated data.

Additionally, we examined the correlation between these phenotypes and conditional variables to verify the plausibility and consistency of the generated data under different population characteristics [262]. We calculated the distribution of imaging phenotypes relative to demographic information and clinical conditions, and compared the generated data with real data. Qualitative results were compared using kernel density plots, and quantitative results were compared using KL divergence [263] and Wasserstein distance (WD) [264]. KL divergence is an information theory metric that quantifies the difference between two probability mass functions. Specifically, KL divergence measures the amount of information lost by encoding data using distribution P under the assumption that distribution Q is the actual distribution. KL divergence can be expressed as:

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} \quad (6.7)$$

where $\mathcal{X}$ is the set of all possible events, $P(x)$ and $Q(x)$ represent the probability mass functions of event x under distributions P and Q, respectively. For continuous distributions, KL divergence is calculated as:

$$D_{\mathrm{KL}}(P \parallel Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} \, dx \tag{6.8}$$

WD measures the distance between two probability distributions. It represents the least effort required to convert one distribution into another, where work is quantified by the mass transfer from u to v multiplied by the distance it is moved. The formula for the WD is as follows:

$$W(P, Q) = \inf_{\gamma \sim \Pi(P,Q)} \mathbb{E}_{(u,v)\sim\gamma}[\|u - v\|] \tag{6.9}$$

where $\Pi(P, Q)$ represents the set of all joint distributions with marginal distributions P and Q respectively. γ is one of these joint distributions. $\mathbb{E}$ is the expected value operator. $\|u - v\|$ is the distance between u and v.

6.3 Experiments and Results

6.3.1 Dataset

We performed the experiments using 415 SAX cine CMR sequences from the UK Biobank study. Each scan consists of 50 time frames, covering a complete cardiac cycle, and each frame comprises a stack of 10 SAX image slices along the heart's vertical axis. Therefore, each subject's scan forms a 4D image volume with dimensions of (x, y, z, t), where x and y denote the spatial resolution of each 2D slice, z indicates the number of slices (i.e., spatial depth), and t represents time points across the cardiac cycle. In other words, a single 4D image stack can be interpreted as 50 time frames of 3D image volumes: each composed of multiple 2D slices, providing rich spatio-temporal information for learning.

All training images were cropped to a fixed spatial resolution of 128×128, and voxel intensity values were normalized to the range $[0, 1]$. The entire dataset was randomly split into 199 scans for training, 40 for validation, and 176 for testing. To leverage the full 4D nature of the data, we adopted a 3D CNN architecture, which applies convolutional filters in three spatial dimensions (x, y, z) at each time point. Compared to traditional 2D CNNs, which operate on single image slices, 3D CNNs can capture

volumetric anatomical context within each 3D frame. Moreover, by processing all 50 time points sequentially, the model can learn the dynamic features of the beating heart.

Regarding demographic and clinical data, all participants were healthy volunteers, including 223 females and 192 males, aged between 40-69 years, with weights ranging from 49 to 116 kg and heights from 147 to 195 cm. Smoking State (SS) and Alcohol State (AS) included 0: Never, 1: Previous, and 2: Current. Systolic blood pressure (SBP) ranged from 95 to 191 mm Hg, and diastolic blood pressure (DBP) was between 60-106 mm Hg.

6.3.2 Experimental Setting

The model utilizes PyTorch [202] for its implementation. The autoencoder part transforms high-dimensional data into a reduced latent representation. The encoder features 3D convolutional layer and 3D attention layer network structure, outputting the latent code z_0. The decoder also consists of 3D convolutional and attention layers network structure. Both encoder and decoder utilize convolutional and transposed convolutional layers with a kernel size of 3. The training loss includes the image reconstruction L1 and KL regularization loss used to align the latent space with a standard normal distribution. The latent space dimensions used are $16 \times 16 \times 1 \times 50$. The hyperparameters for network training are: epochs = 300, KL weight is $1e-6$, batch size is 2, and learning rate is $1e-4$. We select the best model based on the minimum validation loss.

After the autoencoder part is trained, it is used as the input for training the conditional DDIM. The conditional mapping network is built using MLP and outputs the latent code z_c for the input condition c. The noise prediction network uses the 3D U-net with a self-attention mechanism and is trained by minimizing the reconstruction L1 loss. z_t is the result of applying the forward diffusion process to the latent space obtained from the pre-trained autoencoder. The backward diffusion process outputs z_0, which is the output from the 3D U-net noise prediction network. The denoised latent representation z_0 constitutes the input to the decoder. The back-diffusion process employs a batch size of 64 during training, a diffusion time step of 1000, a learning rate of $1e-5$, and a beta of (0.9, 0.99) for the ADAM optimizer, for a total of 50,000 training iterations. The latent space generated by the back-diffusion process is reconstructed by the decoder to generate spatio-temporal cine CMR images. After training, the back-diffusion process is sampled to generate the latent representation and used to

create the synthetic images. The model is completed on the NVIDIA A100 DGX.

For comparison, we use several baseline comparison methods. Methods for conditional generation applied in other domains are extended from 2D image synthesis to 4D data modeling. CGAN: A conditional adaptation of Generative Adversarial Networks (GANs) [258]. CVAE: A conditional generative model CVAE [64] where the conditional information is concatenated with the image in both the encoder and decoder. CHeart [265]: A model for conditional generation of 3D-t cardiac anatomy, for comparison, we using cardiac images as input instead of cardiac segmentation masks while maintaining the same model structure.

6.3.3 Qualitative evaluation

Evaluation of generative models is a well-known challenge since ground truth data is often inaccessible. We first qualitatively analyse the obtained results, with examples of real and synthetic images shown in Figure 6.2. We conducted a qualitative evaluation through visual inspection, without external expert consultation. The synthetic images exhibit high anatomical fidelity, preserving key structural features such as the ventricles' shape and positioning. Additionally, they demonstrate high resolution and contrast, with few artifacts, ensuring visual clarity. The overall appearance closely resembles real data, confirming their visual realism. In addition, the generated images are diverse in terms of the heart and appearance of the ventricles.

6.3.4 Quantitative evaluation

In addition to visual inspection, we also quantitatively evaluate the generated synthetic data to assess how close between the generated and real data. After training, the model generates new full spatio-temporal cine CMR synthetic images taking demographic information and clinical variables being the sole input. Given the randomness of LDM generation, multiple cardiac image sequences can be generated for different combinations of input conditions of gender, age, height, weight, smoke and alcohol status, systolic and diastolic blood pressure. We select 50 random samples from the Gaussian distribution and generate 50 synthetic cardiac cine CMR image sequences for that set of input conditions accordingly.

We evaluate the performance of VAE and DDIM in our methods separately. Figure 6.3 shows an example comparison of myocardial segmentation masks for VAE test set

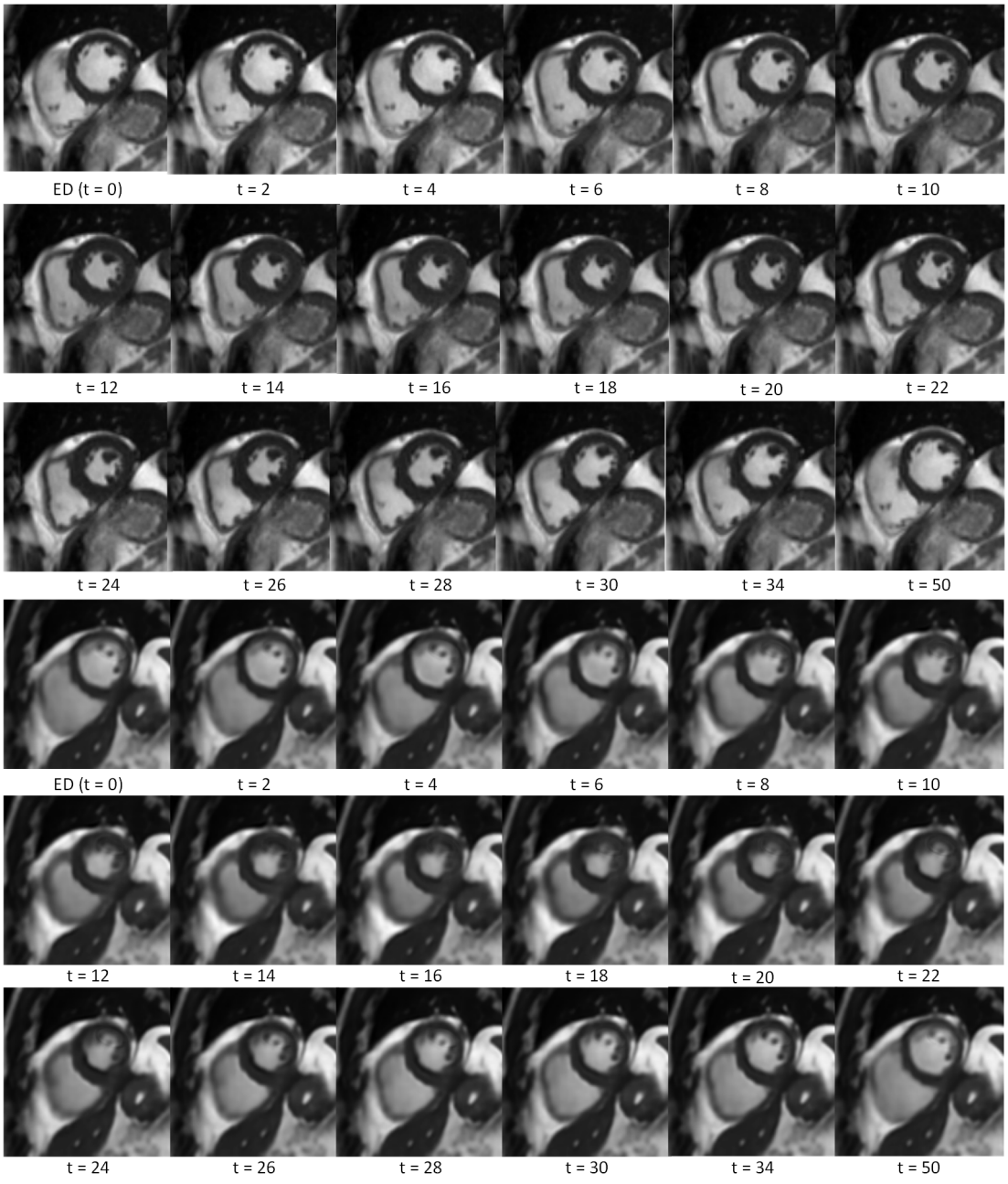

Figure 6.2: **Schematic diagram of examples of image generation.** Some examples include the ED frame at time $t = 0$ to $t = 50$. The first three rows are examples of images from real data, and the last three rows are examples of images from synthetic data.

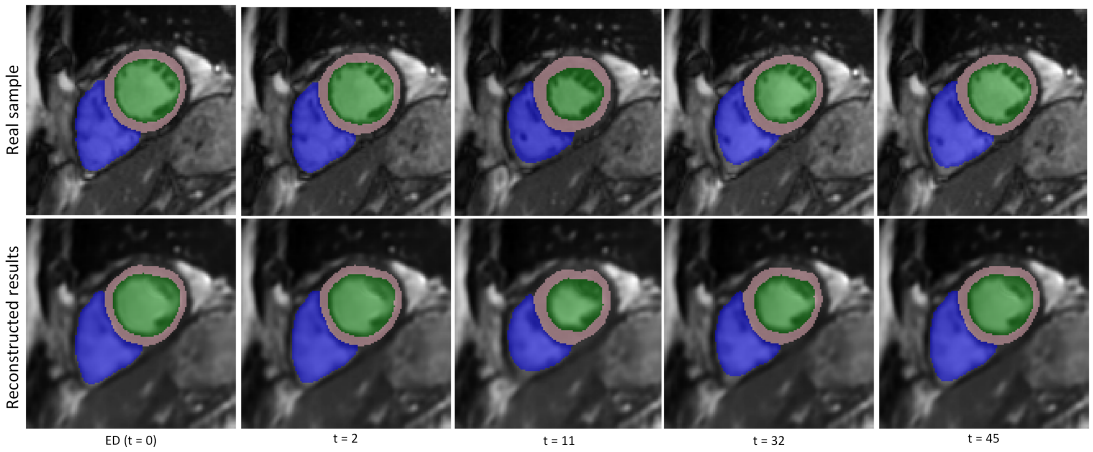

Figure 6.3: **Schematic diagram of comparison of myocardial segmentation masks for test set images and reconstructed images after VAE.**

Table 6.1: Results of quantitative evaluation of the image fidelity and diversity of synthetic data generated using CGAN, CVAE, CHeart, and our model.

Method	FID↓	FRD↓	IP↑	IR↑	MMD↓	MS-SSIM↑	4-G-R SSIM↑
CGAN [258]	43.73 ± 3.24	11.49 ± 0.75	0.79 ± 0.05	0.72 ± 0.03	0.26 ± 0.03	0.56 ± 0.04	0.33 ± 0.03
CVAE[64]	33.60 ± 2.46	7.23 ± 0.54	0.75 ± 0.06	0.74 ± 0.03	0.19 ± 0.04	0.59 ± 0.04	0.36 ± 0.03
CHeart [253]	7.74 ± 0.85	6.64 ± 0.64	0.84 ± 0.03	0.78 ± 0.02	0.23 ± 0.02	0.61 ± 0.03	0.36 ± 0.02
Ours	7.01 ± 0.69	6.17 ± 0.57	0.90 ± 0.03	0.80 ± 0.02	0.18 ± 0.02	0.66 ± 0.03	0.39 ± 0.02

images and reconstructed images. For VAE, we calculate the SSIM, PSNR of the test set and the reconstructed results, as well as the Dice coefficient, HD, and ASSD of the myocardial segmentation masks. The SSIM of VAE is 0.89 ± 0.03, the PSNR is 29.57 ± 1.06, the Dice coefficient of the myocardial segmentation mask is 0.95 ± 0.01, the HD is 8.26 ± 2.63, and the ASSD is 1.69 ± 0.91. The reason for evaluating these indicators is to measure the impact of image quality degradation caused by VAE, which will affect the generation performance of conditional LDM.

The realism of the generated images is measured by the FID score. This method uses the pre-trained Inception-V3 [224] model for feature extraction and is calculated by comparing the feature distribution of the generated and the real image. The lower the FID score, the greater the similarity between the generated and real images in terms of perceptual quality. FRD uses the pre-trained ResNet50 model to extract features to evaluate the similarity between the generated and real results. A lower FRD score indicates a closer alignment between the distributions of the generated and

real results in the feature space, and this evaluation is more consistent with human perception. IP and IR evaluate the quality and coverage of image generation samples by forming an explicit non-parametric representation of the real and the generated data manifold. IP represents the probability that the generated image falls within the support range of the real image manifold, and the higher the value, the better; while IR represents the probability that the real result belongs to the generated result manifold, and more higher the value, the better the coverage of the generated result to the real data. MMD quantifies the distance between sample distributions of two datasets in the Reproducing Kernel Hilbert Space (RKHS). A smaller MMD value indicates greater similarity between the distributions of generated and real images, suggesting superior performance of the generative model. MS-SSIM and 4-G-R SSIM measure image similarity. The generated results are evaluated with the reference images in terms of structure, contrast, and brightness. A higher MS-SSIM score indicates a higher similarity between the generated result and the ground truth, i.e., better quality. Compared with images generated by competing methods, the images generated by our method have higher MS-SSIM and 4-G-R SSIM values, indicating better generated image similarity. These specific results are shown in Table 6.1. The proposed model demonstrates strong image generation accuracy, with average FID, FRD, IP, IR, MMD, MS-SSIM and 4-G-R SSIM of 7.01, 6.17, 0.9, 0.8, 0.18, 0.66 and 0.39 respectively.

Table 6.2 shows the image generation performance comparison between CGAN, CVAE, CHeart and our method. Clinical measurements obtained from each real sample are compared with those obtained from 50 synthetic samples under identical conditions. The results in Table 6.2 show that our model achieves low measurement differences in clinical phenotypes between real samples, with average differences of 28.15 mL, 13.56 mL, 39.45 mL, 19.52 mL and 27.43 g for LVEDV, LVESV, RVEDV, RVESV and LVM, the quantitative indicators demonstrate that the proposed model exhibits fidelity, indicating a high similarity between generated and ground truth samples. Moreover, the model's output aligns well with expected cardiac structures.

We further assess the realism and variety of the generated results in relation to the ground truth through distance evaluation. Tables 6.3 and 6.4 show the KL divergence and WD between the synthetic and real data distributions, respectively. Our method achieves the best KL or WD indicators in most of them. The KL divergence for LVEDV, LVESV, RVEDV, RVESV and LVM are 28.15, 13.56, 39.45, 22.45, 31.76, and the WD

Table 6.2: Comparison of the generation capabilities between CGAN, CVAE, CHeart, and our method. Clinical measurements obtained from each real sample are compared with those from 50 synthetic results under identical conditions.

Method	d_{LVEDV}(mL)	d_{LVESV}(mL)	d_{RVEDV}(mL)	d_{RVESV}(mL)	d_{LVM}(g)
CGAN	38.51 ± 15.87	21.26 ± 9.82	46.82 ± 26.46	25.61 ± 12.51	39.46 ± 25.31
CVAE	36.45 ± 18.15	19.51 ± 8.94	48.52 ± 27.61	23.86 ± 11.75	34.81 ± 27.31
CHeart	29.72 ± 16.61	16.59 ± 7.45	42.73 ± 26.51	22.45 ± 10.56	31.76 ± 21.51
Ours	28.15 ± 19.85	13.56 ± 7.28	39.45 ± 20.57	19.52 ± 10.25	27.43 ± 17.53

values are 29.72, 13.56, 39.45, 19.52, 27.43, respectively.

In addition to quantitative evaluation, we qualitatively assess the distribution of clinical measurements about age and weight for real and synthetic images, including LVEDV, LVESV, RVEDV, RVESV, and LVM, as shown in Figure 6.4 and Figure 6.5. The qualitative evaluation involves visually inspecting kernel density plots, we compare the contour shapes, density patterns, and data concentration regions between real and synthetic distributions. A strong similarity in these density patterns suggests that the synthetic data effectively captures key statistical properties of the real data. The results indicate that the synthetic data distribution, conditioned on age and weight, closely resembles the real distribution. Given the general expectation that individuals with higher weights tend to have larger hearts, the alignment between real and synthetic distributions may seem intuitive. However, the near-perfect match observed in Figures 6.4 and 6.5 raises concerns about potential overfitting. This suggests that our model might be capturing not only the true underlying distribution but also noise or specific patterns from the training data. A more detailed evaluation is necessary to determine whether the generated data generalizes well or merely replicates training characteristics. Future work should investigate regularization techniques or adversarial validation to ensure the synthetic data maintains biological plausibility without excessive mimicry.

In addition, we compared our method with other methods for generating cine CMR images, all of which are dedicated to generating time series medical images. Table 6.5 shows the performance comparison of each method on the corresponding indicators, where the missing values indicate that the original paper did not report the indicator. The comparison shows that our method is better than the comparison methods in SSIM, Dice coefficient and FID, and is comparable to the best method in PSNR, indicating good performance in generating high-quality Cine CMR images. In particular, it shows

Table 6.3: KL divergence between the distribution of synthetic and real data.

Similarity of distributions	LVEDV	LVESV	RVEDV	RVESV	LVM
CGAN	38.51 ± 15.87	21.26 ± 9.82	46.82 ± 26.46	25.61 ± 12.51	39.46 ± 25.31
CVAE	36.45 ± 18.15	19.51 ± 8.94	48.52 ± 27.61	23.86 ± 11.75	34.81 ± 27.31
CHeart	29.72 ± 16.61	16.59 ± 7.45	42.73 ± 26.51	19.52 ± 10.56	27.43 ± 21.51
Ours	28.15 ± 19.85	13.56 ± 7.28	39.45 ± 20.57	22.45 ± 10.25	31.76 ± 17.53

Table 6.4: WD between the distribution of synthetic and real data.

Similarity of distributions	LVEDV	LVESV	RVEDV	RVESV	LVM
CGAN	38.51 ± 15.87	21.26 ± 9.82	46.82 ± 26.46	25.61 ± 12.51	39.46 ± 25.31
CVAE	36.45 ± 18.15	19.51 ± 8.94	48.52 ± 27.61	23.86 ± 11.75	34.81 ± 27.31
CHeart	28.15 ± 16.61	16.59 ± 7.45	42.73 ± 26.51	22.45 ± 10.56	31.76 ± 21.51
Ours	29.72 ± 19.85	13.56 ± 7.28	39.45 ± 20.57	19.52 ± 10.25	27.43 ± 17.53

greater advantages in myocardial segmentation accuracy (Dice coefficient) and overall image quality (FID).

Figure 6.6 shows the visualization of the latent space of the generated image sequence after dimensionality reduction by t-distributed stochastic neighbor embedding (t-SNE) [266], which shows that the generative model can capture the temporal dynamic changes in the cardiac sequence. In each time frame, the volume of the heart LV first decreases and then increases, and the thickness of the myocardium first increases and then decreases, which is consistent with the contraction and relaxation patterns of the real heart.

Table 6.5: Comparison of different methods for generating cine CMR images. Missing values indicate that the corresponding metric was not reported in the original paper.

Method	**SSIM** ↑	**PSNR** ↑	**Dice** ↑	**FID** ↓
TexDC [267]	-	-	-	54.82
SADM [268]	0.851	28.992	-	-
Diffusion Deformable Model [269]	-	30.725	0.802	-
Ours	0.89 ± 0.03	29.57 ± 1.06	0.95 ± 0.01	7.01 ± 0.69

6.4 Discussions

The model proposed in this work is based on the conditional LDM, which integrates conditional branches to control and simulate the effect of various demographic information and clinical variables on the synthesis process. The qualitative and quantitative

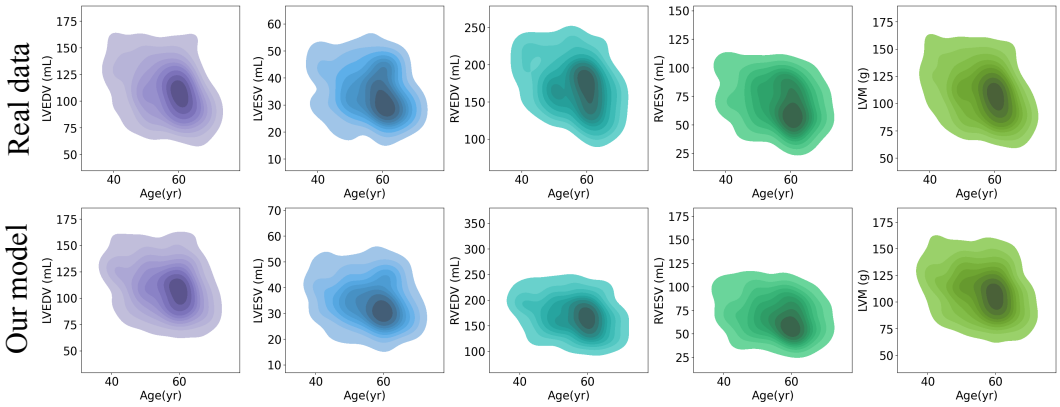

Figure 6.4: **Schematic diagram of comparative distribution of clinical measurements for real and synthetic data.** Kernel density plots of imaging phenotypes versus age are shown separately. In each subplot, the x-axis represents age while the y-axis represents the measurement of the imaging phenotype. Darker regions indicate higher data concentration, whereas lighter regions signify sparser data.

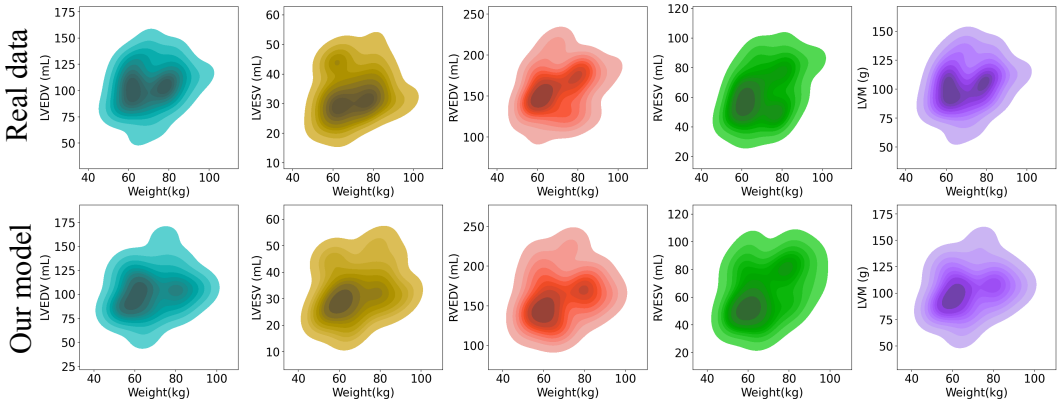

Figure 6.5: **Schematic diagram of comparative distribution of clinical measurements for real and synthetic data.** Kernel density plots of imaging phenotypes versus weight are shown separately. In each subplot, the x-axis represents weight and the y-axis represents the measurement of the imaging phenotype. Darker regions represent areas where data are more concentrated. Lighter areas represent areas where data are more sparse.

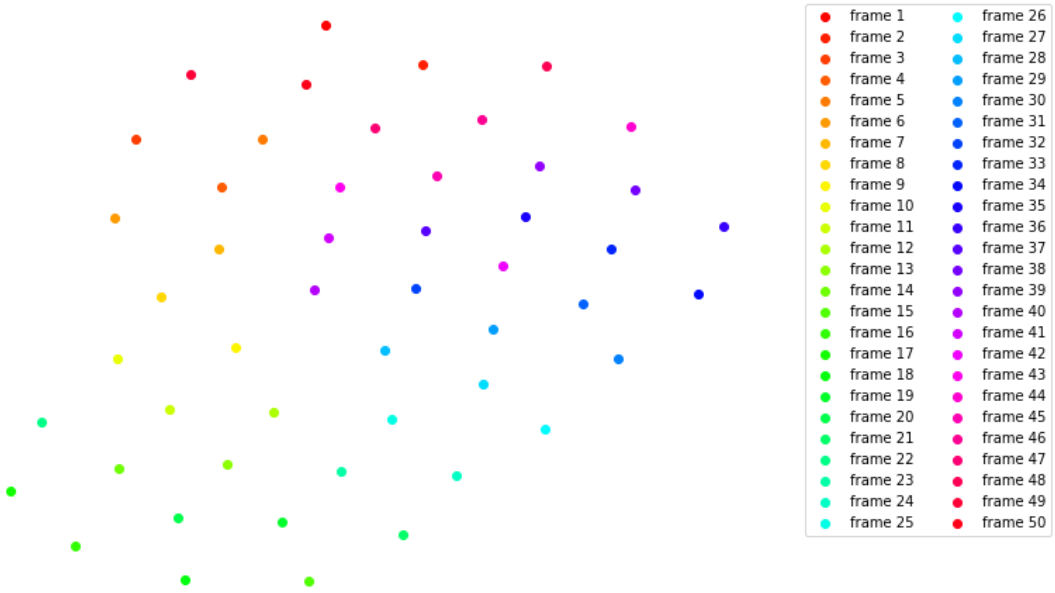

Figure 6.6: **Schematic diagram of the T-distributed Stochastic Neighbor Embedding (t-SNE) visualization representing the latent space of the synthesised full sequence of cine CMR images.** Each dot represents an individual time frame, with color indicating the sequence index. The image decoded from the latent code by one subject is visualized.

results of the experiment show that the model has good generation performance. The control of the conditional variables can show the influence of demographic information and clinical factors on the shape and anatomical changes of the heart. The evaluation results of several clinical measurement indicators (left and right ventricular volume and mass) show that the generated sample distribution aligns closely with the actual samples in both qualitative and quantitative terms, and has diversity. Although the generation results of the model perform well, the similarity between the distributions of generated and real samples needs to be further improved, and the relationship between the anatomical shape and motion of the cardiac and the conditional information needs to be further explored.

The trained cardiac image generation model can be used for potential downstream tasks, such as augmented data for model training to enhance the ability of medical diagnosis and research, privacy-preserving data generation and discovery of complex patterns and changes in images. The conditional generation model means that the model can learn cardiac patterns for specific conditions (such as gender, age, height and weight) and evaluate deviations in a personalized way, thereby providing more accurate diagnosis and support. The model can deeply understand features and identify potential risk factors after learning complex patterns and changes in different demographic information and clinical factors. In addition, the trained model can produce substantial synthetic data for various applications. Synthetic data can enhance data augmentation strategies, thereby improving the performance of machine learning models [270][271]. Additionally, the models can generate synthetic data to promote fairness in predictive models [272], or act as virtual simulation tools for in-silico experiments [273]. The diversity and authenticity of generated data can address data scarcity issues, particularly when access to real data is limited or challenging. These synthetic data can also be used for privacy protection research [274], providing safe and available alternative data, thereby promoting the development of medical research and clinical applications. Further, generative models can also be used for personalized medicine and precision medicine. By generating personalized models based on the specific characteristics of patients, doctors can better formulate treatment plans, predict treatment effects, and accurately manage patients' health. This capability will help advance the development of individualized treatment plans and drive the medical field toward a more personalized and precise direction in the future.

This work has some limitations. First, we preprocessed the input image size to $128 \times 128 \times 10$ and used 50 time frames, which makes the computational cost and complexity of the 4D data model high. Future research directions can focus on reducing the dimensionality and computational complexity. In addition, our training data consists of healthy subjects from the UK Biobank. In the future, it can be extended to longitudinal clinical subjects with cardiovascular disease and multi-center datasets to enhance the variety of the generated models.

6.5 Conclusion

In this study, we develop a conditional generative model based on latent diffusion that can generate spatio-temporal cine cardiac images with the input of controllable demographic information and clinical factors as conditions. The results demonstrate that the generated 4D CMR images are highly realistic and diversity, and can synthesise realistic cardiac anatomical structures and motion changes. This study lays the foundation for generative modeling research in cardiac MRI imaging, which can be further extended to include different disease types or focus on anatomical structure representation and mesh form in the future. In addition, this work could be utilized in a range of downstream applications, including data augmentation, construction of cardiac images under specific conditions, and downstream tasks such as synthetic image cardiac segmentation and registration.

Chapter 7

Conclusion and Future Work

This chapter summarizes the main results of this paper in detail and demonstrates the positive progress in applying deep learning technology to cardiac image generation and analysis. In addition, we discuss some inherent limitations of existing methods and propose possible directions for improvement. These improvement directions could help to overcome the shortcomings of current methods, and also provide new ideas for future research to further enhance and optimize the techniques and methods proposed in this thesis.

7.1 Conclusion

Motivated by the application of generative models in the field of medical imaging to obtain synthetic images, this research is dedicated to exploring the generation and analysis of multiple generative models on CMR of multiple sequence modalities. The synthesis and analysis of the cardiac through deep learning techniques are explored, and three follow-up works are proposed.

Firstly, we introduce a smcVAE model to learn the joint latent representation of multiple sequential CMR images, and use only cine CMR to synthesise tagging CMR and estimate myocardial strain. This model does not require additional clinical sequence acquisition, and obtains reliable myocardial strain estimation. Furthermore, the model can be extended to generate demographic and clinical information from other channels, not just the generation of images domain.

Furthermore, focusing on cine CMR, we introduce a latent denoising diffusion implicit generative model to synthesise full-spatial cine CMR images and corresponding biventricular segmentation masks. The synthesised images can be used as the viable substitute for real datasets in deep learning model training and downstream tasks such as cardiac image registration or segmentation, and the synthesised cardiac anatomical images can help alleviate the burden of manually annotated images.

Additionally, the relationship between non-imaging factors such as clinical and demographic information and imaging is explored through a conditional latent diffusion generative model. The full spatio-temporal 4D cine CMR images generation model can describe the 4D spatio-temporal images of the cardiac and its interactions with non-imaging demographic and clinical factors. It can be used to generate healthy datasets, and also to generate cardiac images that combine disease types and condition-specific. This is of great value in creating comprehensive datasets with limited abnormal

samples, and the proposed model also shows great potential for personalized modeling and pathological development evaluation.

These contributions explore multiple deep learning generative models, which together improve the accuracy and reliability of cardiac multi-sequence image generation, and are of great significance in the applications of image enhancement, data augmentation, automatic annotation, and personalized medicine. Overall, it improves the image processing quality and efficiency, contributes to the robustness of medical image analysis, and provides strong support for personalized medicine and medical research.

7.2 Limitations and Future Research Directions

In this thesis, although our research has achieved promising results, there are still some challenges and limitations that need further attention to promote better cardiac image generation and analysis.

In deep learning-based cardiac image analysis, richer image latent representations and more time-related features should be considered. In Chapter 4, we designed a sparse dual-image channel VAE to learn the joint latent space of cine CMR and tagging CMR images. However, other non-imaging population information or clinical factors were not considered. It would be interesting to expand VAE to more channels. Model architectures that integrate more information may be more improved than dual-channel image models. In addition, most cardiac image analysis only uses CMR images in ED and ES frames, without considering other time frames in the entire cardiac cycle. Incorporating these time frames into analysis and research can achieve more accurate performance.

A more general challenge is the problem of domain adaptation between different datasets, which is one of the common limitations of current deep learning-based methods. During the testing phase, deep learning networks are usually able to perform well on data that is similar to the training data, but their performance may drop when applied to new data that differs greatly from the training data. This challenge is particularly significant in the field of medical image analysis. For example, data from different imaging equipment and multiple centers may differ in appearance, causing the model to perform poorly on these data.

Although the latent diffusion framework proposed in Chapters 5 and 6 shows robustness, its generalizability may be limited when faced with different datasets and various

pathological conditions. This issue is critical because the effectiveness of medical imaging models largely relies on their applicability in various real-world scenarios. Only when they can maintain high performance across different datasets and pathological conditions, the models play their true value in clinical practice.

Another common challenge in medical image analysis models is the limitation of datasets. Current research often relies on small datasets with a limited number of publicly available samples. For example, in cardiovascular disease research, commonly used datasets usually have only about 100 samples. This limitation in data size brings problems. First, small datasets may not fully represent all possible pathological conditions and patient characteristics, which affects the generalization ability of the model and its reliability in practical applications. In addition, the acquisition and annotation of medical image data are expensive, and involve issues such as privacy protection, which makes it unlikely to expand the size of such datasets in the future. This limitation further restricts the range of options researchers have when developing and validating new models.

Another notable limitation is the lack of quality control evaluation of the results of generative models, especially in the context of deep enhancement for medical image analysis. Although the proposed models show satisfactory generation capabilities, their practical utility depends on the quality and reliability of the generated samples in clinical applications. However, there are currently challenges and deficiencies in this regard. Firstly, when generating medical images, generative models may have artifacts, noise, or other unnatural features, which may mislead doctors when making diagnoses. Second, even if the generated images look realistic visually, their clinical validity remains questionable if they do not accurately reflect the real pathological conditions. In addition, generative models may overfit the training data, thus lacking generalizability across different datasets and pathological conditions.

In summary, although deep learning-based image generation and analysis methods have achieved remarkable success in a wide range of computer vision applications, successful and accurate cardiac image synthesis and analysis require special attention to the inherent limitations of deep learning methods and relevant priors of cardiac imaging, such as cardiac motion, cardiac cycle, etc. Taking these factors into account is both necessary and crucial to obtain more efficient and accurate results in the analysis of cardiovascular disease.

Looking ahead, future research can aim to develop model architectures that integrate more information. Such as multi-channel VAE models that integrate more non-imaging demographic information and clinical factors. These models can capture richer potential representations, thereby improving the accuracy and generalization of image generation and analysis. As well as incorporating dynamic feature processing, design algorithms that can handle cardiac cycles and cardiac motion, considering all time frames throughout the cardiac cycle to improve the model's adaptability to dynamic changes and analysis accuracy.

Developing effective domain adaptation methods is also a future research direction, enabling the model to maintain high performance under different datasets and imaging conditions. For example, using techniques such as transfer learning and adaptive learning to improve the applicability of models in different domains.

In terms of data augmentation, generative models and data augmentation techniques are used to augment existing datasets. By generating high-quality synthetic images, the diversity of data can be increased, thus enhancing the model's generalization capability. As well as promoting multi-center collaboration, integrating data resources from different imaging devices and multiple centers to create larger and more diverse datasets. This will help improve the robustness and reliability of the model under different conditions.

In terms of quality control, to implement comprehensive quality control protocols, including quantitative evaluation indicators, clinical validation, cross-dataset validation, and improving model interpretability, to ensure that the generated data is not only diverse but also clinically valid, thereby improving the reliability of the generated models in practical applications.

References

[1] Peter DeSaix, Gordon J Betts, Eddie Johnson, Jody E Johnson, Korol Oksana, Dean H Kruse, Brandon Poe, James A Wise, and Kelly A Young. Anatomy & physiology (openstax), 2013.

[2] Jamie R Mitchell and Jiun-Jr Wang. Expanding application of the wiggers diagram to teach cardiovascular physiology. *Advances in physiology education*, 38(2):170–175, 2014.

[3] Xinyuan Zhang, Ritzia Vinu Alexander, Jie Yuan, and Yichen Ding. Computational analysis of cardiac contractile function. *Current cardiology reports*, 24(12):1983–1994, 2022.

[4] El-Sayed H Ibrahim. Myocardial tagging by cardiovascular magnetic resonance: evolution of techniques–pulse sequences, analysis algorithms, and applications. *Journal of Cardiovascular Magnetic Resonance*, 13(1):36, 2011.

[5] Björn Dahlöf. Cardiovascular disease risk factors: epidemiology and risk assessment. *The American journal of cardiology*, 105(1):3A–9A, 2010.

[6] N Poulter. Global risk of cardiovascular disease. *Heart*, 89(suppl 2):ii2–ii5, 2003.

[7] Joselyn Rwebembera, James Marangou, Julius Chacha Mwita, Ana Olga Mocumbi, Cleonice Mota, Emmy Okello, Bruno Nascimento, Lene Thorup, Andrea Beaton, Joseph Kado, et al. 2023 world heart federation guidelines for the echocardiographic diagnosis of rheumatic heart disease. *Nature Reviews Cardiology*, 21(4):250–263, 2024.

[8] Mahboobeh Jafari, Afshin Shoeibi, Marjane Khodatars, Navid Ghassemi, Parisa Moridian, Roohallah Alizadehsani, Abbas Khosravi, Sai Ho Ling, Niloufar

Delfan, Yu-Dong Zhang, et al. Automated diagnosis of cardiovascular diseases from cardiac magnetic resonance imaging using deep learning models: A review. *Computers in Biology and Medicine*, 160:106998, 2023.

[9] Manuel A Morales, Warren J Manning, and Reza Nezafat. Present and future innovations in ai and cardiac mri. *Radiology*, 310(1):e231269, 2024.

[10] Yanbin Liu, Girish Dwivedi, Farid Boussaid, and Mohammed Bennamoun. 3d brain and heart volume generative models: A survey. *ACM Computing Surveys*, 56(6):1–37, 2024.

[11] Abhishek Thakur and Gopal Kumar Thakur. Developing gans for synthetic medical imaging data: Enhancing training and research. *Int. J. Adv. Multidiscip. Res*, 11(1):70–82, 2024.

[12] Ryo Nishikimi, Masahiro Nakano, Kunio Kashino, and Shingo Tsukada. Variational autoencoder–based neural electrocardiogram synthesis trained by fem-based heart simulator. *Cardiovascular Digital Health Journal*, 5(1):19–28, 2024.

[13] Tanmay Mukherjee, Muhammad Usman, Rana Raza Mehdi, Emilio Mendiola, Jacques Ohayon, Diana Lindquist, Dipan Shah, Sakthivel Sadayappan, Roderic Pettigrew, and Reza Avazmohammadi. In-silico heart model phantom to validate cardiac strain imaging. *bioRxiv*, pages 2024–08, 2024.

[14] Wei Du, Yongkang Huo, Rixin Zhou, Yu Sun, Shiyi Tang, Xuan Zhao, Ying Li, and Gaoyang Li. Consistency label-activated region generating network for weakly supervised medical image segmentation. *Computers in Biology and Medicine*, 173:108380, 2024.

[15] Atif Ahmed Showrov, Md Tarek Aziz, Hadiur Rahman Nabil, Jamin Rahman Jim, Md Mohsin Kabir, MF Mridha, Nobuyoshi Asai, and Jungpil Shin. Generative adversarial networks (gans) in medical imaging: Advancements, applications and challenges. *IEEE Access*, 2024.

[16] Mariachiara Di Cesare, Pablo Perel, Sean Taylor, Chodziwadziwa Kabudula, Honor Bixby, Thomas A Gaziano, Diana Vaca McGhie, Jeremiah Mwangi, Borjana Pervan, Jagat Narula, et al. The heart of the world. *Global heart*, 19(1), 2024.

[17] Israel Mirsky, Andre Pasternac, and R Curtis Ellison. General index for the assessment of cardiac function. *The American journal of cardiology*, 30(5):483–491, 1972.

[18] MD Cerqueira, NJ Weissman, V Dilsizian, AK Jacobs, S Kaul, WK Laskey, DJ Pennell, JA 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*, 105(4):539–542, 2002.

[19] Sergio Mondillo, Maurizio Galderisi, Donato Mele, Matteo Cameli, Vincenzo Schiano Lomoriello, Valerio Zacà, Piercarlo Ballo, Antonello D'Andrea, Denisa Muraru, Mariangela Losi, et al. Speckle-tracking echocardiography: a new technique for assessing myocardial function. *Journal of Ultrasound in Medicine*, 30(1):71–83, 2011.

[20] João AC Lima and Milind Y Desai. Cardiovascular magnetic resonance imaging: current and emerging applications. *Journal of the American College of Cardiology*, 44(6):1164–1171, 2004.

[21] Mihaela Silvia Amzulescu, Mathieu De Craene, Helénè Langet, Agnes Pasquet, David Vancraeynest, Anne-Catherine Pouleur, Jean-Louis Vanoverschelde, and BL Gerber. Myocardial strain imaging: review of general principles, validation, and sources of discrepancies. *European Heart Journal-Cardiovascular Imaging*, 20(6):605–619, 2019.

[22] Robert M Hackett and Robert M Hackett. Strain measures. *Hyperelasticity Primer*, pages 5–8, 2018.

[23] N Grenier, F Basseau, M Rey, and L LaGoarde-Segot. Interpretation of doppler signals. *European radiology*, 11(8):1295–1307, 2001.

[24] Nathaniel Reichek. Mri myocardial tagging. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 10(5):609–616, 1999.

[25] Otto A Smiseth, Hans Torp, Anders Opdahl, Kristina H Haugaa, and Stig Urheim. Myocardial strain imaging: how useful is it in clinical decision making? *European heart journal*, 37(15):1196–1207, 2016.

[26] Donald B Plewes and Walter Kucharczyk. Physics of mri: a primer. *Journal of magnetic resonance imaging*, 35(5):1038–1054, 2012.

[27] Lee W Goldman. Principles of ct and ct technology. *Journal of nuclear medicine technology*, 35(3):115–128, 2007.

[28] Catherine M Otto. *Textbook of clinical echocardiography*. Elsevier Health Sciences, 2013.

[29] Andrew C Larson, Richard D White, Gerhard Laub, Elliot R McVeigh, Debiao Li, and Orlando P Simonetti. Self-gated cardiac cine mri. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 51(1):93–102, 2004.

[30] Oliver Bieri and Klaus Scheffler. Fundamentals of balanced steady state free precession mri. *Journal of Magnetic Resonance Imaging*, 38(1):2–11, 2013.

[31] Leif Sörnmo and Pablo Laguna. Electrocardiogram (ecg) signal processing. *Wiley encyclopedia of biomedical engineering*, 2006.

[32] M Sabarimalai Manikandan and KP Soman. A novel method for detecting r-peaks in electrocardiogram (ecg) signal. *Biomedical signal processing and control*, 7(2):118–128, 2012.

[33] Udo Sechtem, Peter W Pflugfelder, Richard D White, Robert G Gould, William Holt, Martin J Lipton, and Charles B Higgins. Cine mr imaging: potential for the evaluation of cardiovascular function. *American Journal of Roentgenology*, 148(2):239–246, 1987.

[34] Ronald J Jaszczak, R Edward Coleman, and Chun Bin Lim. Spect: Single photon emission computed tomography. *IEEE Transactions on Nuclear Science*, 27(3):1137–1153, 1980.

[35] Dale L Bailey, Michael N Maisey, David W Townsend, and Peter E Valk. *Positron emission tomography*, volume 2. Springer, 2005.

[36] Gopi Kiran Reddy Sirineni, Mannudeep Karanvir Singh Kalra, Krishna Mohan Pottala, Mushabbar Ali Syed, Stefan Tigges, and Aaron Darius Cann. Visualization techniques in computed tomographic coronary angiography. *Current Problems in Diagnostic Radiology*, 35(6):245–257, 2006.

[37] Monda L Shehata, Susan Cheng, Nael F Osman, David A Bluemke, and João AC Lima. Myocardial tissue tagging with cardiovascular magnetic resonance. *Journal of Cardiovascular Magnetic Resonance*, 11(1):55, 2009.

[38] Ha Q Vo, Thomas H Marwick, and Kazuaki Negishi. Mri-derived myocardial strain measures in normal subjects. *JACC: Cardiovascular Imaging*, 11(2 Part 1):196–205, 2018.

[39] George R Sutherland, Giovanni Di Salvo, Piet Claus, Jan D'hooge, and Bart Bijnens. Strain and strain rate imaging: a new clinical approach to quantifying regional myocardial function. *Journal of the American Society of Echocardiography*, 17(7):788–802, 2004.

[40] J Trainini, J Lowenstein, M Beraudo, M Wernicke, A Trainini, VM Llabata, and CF Carreras. Myocardial torsion and cardiac fulcrum. *Morphologie*, 105(348):15–23, 2021.

[41] Satoshi Nakatani. Left ventricular rotation and twist: why should we learn? *Journal of cardiovascular ultrasound*, 19(1):1–6, 2011.

[42] Leon Axel and Lawrence Dougherty. Mr imaging of motion with spatial modulation of magnetization. *Radiology*, 171(3):841–845, 1989.

[43] Moriel NessAiver and Jerry L Prince. Magnitude image cspamm reconstruction (micsr). *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 50(2):331–342, 2003.

[44] Gerald B Folland. *Harmonic analysis in phase space*. Number 122. Princeton university press, 1989.

[45] Mirja Neizel, Dirk Lossnitzer, Grigorios Korosoglou, Tim Schäufele, Antje Lewien, Henning Steen, Hugo A Katus, Nael F Osman, and Evangelos Giannitsis. Strain-encoded (senc) magnetic resonance imaging to evaluate regional heterogeneity of myocardial strain in healthy volunteers: Comparison with conventional

tagging. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 29(1):99–105, 2009.

[46] Anthony H Aletras, Shujun Ding, Robert S Balaban, and Han Wen. Dense: displacement encoding with stimulated echoes in cardiac functional mri. *Journal of magnetic resonance (San Diego, Calif.: 1997)*, 137(1):247, 1999.

[47] Elias A Zerhouni, David M Parish, Walter J Rogers, Andrew Yang, and Edward P Shapiro. Human heart: tagging with mr imaging–a method for noninvasive assessment of myocardial motion. *Radiology*, 169(1):59–63, 1988.

[48] Ruslan Salakhutdinov. Learning deep generative models. *Annual Review of Statistics and Its Application*, 2(1):361–385, 2015.

[49] GM Harshvardhan, Mahendra Kumar Gourisaria, Manjusha Pandey, and Siddharth Swarup Rautaray. A comprehensive survey and analysis of generative models in machine learning. *Computer Science Review*, 38:100285, 2020.

[50] Jakub M Tomczak. Why deep generative modeling? In *Deep Generative Modeling*, pages 1–12. Springer, 2021.

[51] Geoffrey E Hinton. Boltzmann machine. *Scholarpedia*, 2(5):1668, 2007.

[52] Asja Fischer and Christian Igel. An introduction to restricted boltzmann machines. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 17th Iberoamerican Congress, CIARP 2012, Buenos Aires, Argentina, September 3-6, 2012. Proceedings 17*, pages 14–36. Springer, 2012.

[53] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. *Advances in neural information processing systems*, 19, 2006.

[54] Peter Dayan, Geoffrey E Hinton, Radford M Neal, and Richard S Zemel. The helmholtz machine. *Neural computation*, 7(5):889–904, 1995.

[55] Radford M Neal. Connectionist learning of belief networks. *Artificial intelligence*, 56(1):71–113, 1992.

[56] Geoffrey E Hinton. Deep belief networks. *Scholarpedia*, 4(5):5947, 2009.

[57] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[58] Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *ArXiv*, abs/1701.07875, 2017.

[59] Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2813–2821, 2016.

[60] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *ArXiv*, abs/1710.10196, 2017.

[61] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2018.

[62] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251, 2017.

[63] Irina Higgins, Loic Matthey, Arka Pal, Christopher P Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR (Poster)*, 3, 2017.

[64] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

[65] Daniel Im Im, Sungjin Ahn, Roland Memisevic, and Yoshua Bengio. Denoising criterion for variational auto-encoding framework. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[66] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Towards deeper understanding of variational autoencoding models. *arXiv preprint arXiv:1702.08658*, 2017.

[67] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[68] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[69] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[70] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.

[71] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[72] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[73] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[74] Jan Ehrhardt and Matthias Wilms. Autoencoders and variational autoencoders in medical image analysis. In *Biomedical Image Synthesis and Simulation*, pages 129–162. Elsevier, 2022.

[75] Amina Fettah, Rafik Menassel, Abdeljalil Gattal, and Abdelhak Gattal. Convolutional autoencoder-based medical image compression using a novel annotated medical x-ray imaging dataset. *Biomedical Signal Processing and Control*, 94:106238, 2024.

[76] Bing Cao, Zhiwei Bi, Qinghua Hu, Han Zhang, Nannan Wang, Xinbo Gao, and Dinggang Shen. Autoencoder-driven multimodal collaborative learning for med-

ical image synthesis. *International Journal of Computer Vision*, 131(8):1995–2014, 2023.

[77] Gudmund R Iversen. *Bayesian statistical inference*. Number 43. Sage, 1984.

[78] Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. *Advances in neural information processing systems*, 28, 2015.

[79] Irem Cetin, Maialen Stephens, Oscar Camara, and Miguel A González Ballester. Attri-vae: Attribute-based interpretable representations of medical images with variational autoencoders. *Computerized Medical Imaging and Graphics*, 104:102158, 2023.

[80] Mahmoud Elbattah, Colm Loughnane, Jean-Luc Guérin, Romuald Carette, Federica Cilia, and Gilles Dequen. Variational autoencoder for image-based augmentation of eye-tracking data. *Journal of Imaging*, 7(5):83, 2021.

[81] Qingyu Zhao, Ehsan Adeli, Nicolas Honnorat, Tuo Leng, and Kilian M Pohl. Variational autoencoder for regression: Application to brain aging analysis. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pages 823–831. Springer, 2019.

[82] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018.

[83] Tianyang Zhang, Huazhu Fu, Yitian Zhao, Jun Cheng, Mengjie Guo, Zaiwang Gu, Bing Yang, Yuting Xiao, Shenghua Gao, and Jiang Liu. Skrgan: Sketching-rendering unconditional generative adversarial networks for medical image synthesis. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*, pages 777–785. Springer, 2019.

[84] Dakai Jin, Ziyue Xu, Youbao Tang, Adam P Harrison, and Daniel J Mollura. Ct-realistic lung nodule simulation from 3d conditional generative adversarial net-

works for robust lung segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*, pages 732–740. Springer, 2018.

[85] Mina Rezaei, Konstantin Harmuth, Willi Gierke, Thomas Kellermeier, Martin Fischer, Haojin Yang, and Christoph Meinel. A conditional adversarial network for semantic segmentation of brain tumor. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers 3*, pages 241–252. Springer, 2018.

[86] Sibaji Gaj, Mingrui Yang, Kunio Nakamura, and Xiaojuan Li. Automated cartilage and meniscus segmentation of knee mri with conditional generative adversarial networks. *Magnetic resonance in medicine*, 84(1):437–449, 2020.

[87] François Mazé and Faez Ahmed. Diffusion models beat gans on topology optimization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 9108–9116, 2023.

[88] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[89] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022.

[90] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. Diffir: Efficient diffusion model for image restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13095–13105, 2023.

[91] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Refusion: Enabling large-size realistic image restoration with latent-space diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1680–1691, 2023.

[92] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022.

[93] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in Neural Information Processing Systems*, 36, 2024.

[94] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.

[95] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[96] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993, 2021.

[97] Justin Lovelace, Varsha Kishore, Chao Wan, Eliot Shekhtman, and Kilian Q Weinberger. Latent diffusion for language generation. *Advances in Neural Information Processing Systems*, 36, 2024.

[98] Zakaria Rguibi, Abdelmajid Hajami, Dya Zitouni, Amine Elqaraoui, Reda Zourane, and Zayd Bouajaj. Improving medical imaging with medical variation diffusion model: An analysis and evaluation. *Journal of Imaging*, 9(9):171, 2023.

[99] Walter Hugo Lopez Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F. Da Costa, Virginia Fernandez, Parashkev Nachev, Sébastien Ourselin, and Manuel Jorge Cardoso. Brain imaging generation with latent diffusion models. *ArXiv*, abs/2209.07162, 2022.

[100] Zakaria Rguibi, Abdelmajid Hajami, Zitouni Dya, Amine Elqaraoui, Reda Zourane, and Zayd Bouajaj. Improving medical imaging with medical variation diffusion model: An analysis and evaluation. *Journal of Imaging*, 9, 2023.

[101] Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models in medical imaging: A comprehensive survey. *Medical image analysis*, 88:102846, 2022.

[102] Geert Litjens, Francesco Ciompi, Jelmer M Wolterink, Bob D de Vos, Tim Leiner, Jonas Teuwen, and Ivana Išgum. State-of-the-art deep learning in cardiovascular image analysis. *JACC: Cardiovascular imaging*, 12(8 Part 1):1549–1565, 2019.

[103] Fanwei Kong and Shawn C Shadden. A generalizable deep-learning approach for cardiac magnetic resonance image segmentation using image augmentation and attention u-net. 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*, pages 287–296. Springer, 2021.

[104] Didier RPRM Lustermans, Sina Amirrajab, Mitko Veta, Marcel Breeuwer, and Cian M Scannell. Optimized automated cardiac mr scar quantification with gan-based data augmentation. *Computer methods and programs in biomedicine*, 226:107116, 2022.

[105] Youssef Skandarani, Nathan Painchaud, Pierre-Marc Jodoin, and Alain Lalande. On the effectiveness of gan generated cardiac mris for segmentation. *arXiv preprint arXiv:2005.09026*, 2020.

[106] Shusil Dangi, Cristian A Linte, and Ziv Yaniv. A distance map regularized cnn for cardiac cine mr image segmentation. *Medical physics*, 46(12):5637–5651, 2019.

[107] Bob D De Vos, Floris F Berendsen, Max A Viergever, Marius Staring, and Ivana 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: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pages 204–212. Springer, 2017.

[108] Fabian Gigengack, Lars Ruthotto, Martin Burger, Carsten H Wolters, Xiaoyi Jiang, and Klaus P Schafers. Motion correction in dual gated cardiac pet us-

ing mass-preserving image registration. *IEEE transactions on medical imaging*, 31(3):698–712, 2011.

[109] Xishi Huang, John Moore, Gerard Guiraudon, Douglas L Jones, Daniel Bainbridge, Jing Ren, and Terry M Peters. Dynamic 2d ultrasound and 3d ct image registration of the beating heart. *IEEE transactions on medical imaging*, 28(8):1179–1189, 2009.

[110] Rob J van der Geest and Johan HC Reiber. Quantification in cardiac mri. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 10(5):602–608, 1999.

[111] Covadonga Fernández-Golfín, Marta Pachón, Cecilia Corros, Ana Bustos, Beatriz Cabeza, Joaquín Ferreirós, Leopoldo Pérez de Isla, Carlos Macaya, and José Zamorano. Left ventricular trabeculae: quantification in different cardiac diseases and impact on left ventricular morphological and functional parameters assessed with cardiac magnetic resonance. *Journal of cardiovascular medicine*, 10(11):827–833, 2009.

[112] Hanchao Yu, Xiao Chen, Humphrey Shi, Terrence Chen, Thomas S Huang, and Shanhui Sun. Motion pyramid networks for accurate and efficient cardiac motion estimation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23*, pages 436–446. Springer, 2020.

[113] Hanchao Yu, Shanhui Sun, Haichao Yu, Xiao Chen, Honghui Shi, Thomas S Huang, and Terrence Chen. Foal: Fast online adaptive learning for cardiac motion estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4313–4323, 2020.

[114] Brecht Heyde, Szymon Cygan, Hon Fai Choi, Beata Lesniak-Plewinska, Daniel Barbosa, An Elen, Piet Claus, Dirk Loeckx, Krzysztof Kaluzynski, and Jan D'hooge. Regional cardiac motion and strain estimation in three-dimensional echocardiography: A validation study in thick-walled univentricular phantoms. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, 59(4):668–682, 2012.

[115] Ahmed Elnakib, Garth M Beache, Georgy Gimel'farb, and Ayman El-Baz. Intramyocardial strain estimation from cardiac cine mri. *International journal of computer assisted radiology and surgery*, 10:1299–1312, 2015.

[116] Yankun Cao, Zhi Liu, Pengfei Zhang, Yushuo Zheng, Yongsheng Song, and Lizhen Cui. Deep learning methods for cardiovascular image. *Journal of Artificial Intelligence and Systems*, 1(1):96–109, 2019.

[117] Kelvin KL Wong, Giancarlo Fortino, and Derek Abbott. Deep learning-based cardiovascular image diagnosis: a promising challenge. *Future Generation Computer Systems*, 110:802–811, 2020.

[118] Karen Andrea Lara Hernandez, Theresa Rienmüller, Daniela Baumgartner, and Christian Baumgartner. Deep learning in spatiotemporal cardiac imaging: A review of methodologies and clinical usability. *Computers in Biology and Medicine*, 130:104200, 2021.

[119] Michael Loecher, Luigi E Perotti, and Daniel B Ennis. Using synthetic data generation to train a cardiac motion tag tracking neural network. *Medical image analysis*, 74:102223, 2021.

[120] Samaneh Abbasi-Sureshjani, Sina Amirrajab, Cristian Lorenz, Juergen Weese, Josien Pluim, and Marcel Breeuwer. 4d semantic cardiac magnetic resonance image synthesis on xcat anatomical model. In *Medical Imaging with Deep Learning*, pages 6–18. PMLR, 2020.

[121] Daniel Toth, Serkan Cimen, Pascal Ceccaldi, Tanja Kurzendorfer, Kawal Rhode, and Peter Mountney. Training deep networks on domain randomized synthetic x-ray data for cardiac interventions. In *International Conference on Medical Imaging with Deep Learning*, pages 468–482. PMLR, 2019.

[122] S Gurusubramani and B Latha. Enhancing cardiac diagnostics through semantic-driven image synthesis: a hybrid gan approach. *Neural Computing and Applications*, 36(14):8181–8197, 2024.

[123] Sina Amirrajab, Cristian Lorenz, Jürgen Weese, Josien P. W. Pluim, and Marcel M. Breeuwer. Pathology synthesis of 3d consistent cardiac mr images using 2d vaes and gans. *ArXiv*, abs/2209.04223, 2022.

[124] Agisilaos Chartsias, Thomas Joyce, Rohan Dharmakumar, and Sotirios A. Tsaftaris. Adversarial image synthesis for unpaired multi-modal cardiac data. In *SASHIMI@MICCAI*, 2017.

[125] Han-Di Zhang, Jiancheng Yang, Shouhong Wan, and Pascal Fua. Lefusion: Synthesizing myocardial pathology on cardiac mri via lesion-focus diffusion models. *ArXiv*, abs/2403.14066, 2024.

[126] Shaoyan Pan, Tonghe Wang, Richard L. J. Qiu, Marian Axente, Chih-Wei Chang, Junbo Peng, Ashish B Patel, Joseph Shelton, Sagar Patel, Justin Roper, and Xiaofeng Yang. 2d medical image synthesis using transformer-based denoising diffusion probabilistic model. *Physics in Medicine and Biology*, 68, 2023.

[127] Chen Chen, Chen Qin, Huaqi Qiu, Giacomo Tarroni, Jinming Duan, Wenjia Bai, and Daniel Rueckert. Deep learning for cardiac image segmentation: a review. *Frontiers in cardiovascular medicine*, 7:25, 2020.

[128] Caroline Petitjean and Jean-Nicolas Dacher. A review of segmentation methods in short axis cardiac mr images. *Medical image analysis*, 15(2):169–184, 2011.

[129] Phi Vu Tran. A fully convolutional neural network for cardiac segmentation in short-axis mri. *arXiv preprint arXiv:1604.00494*, 2016.

[130] Mahendra Khened, Varghese Alex Kollerathu, and Ganapathy Krishnamurthi. Fully convolutional multi-scale residual densenets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers. *Medical image analysis*, 51:21–45, 2019.

[131] Wufeng Xue, Gary Brahm, Sachin Pandey, Stephanie Leung, and Shuo Li. Full left ventricle quantification via deep multitask relationships learning. *Medical image analysis*, 43:54–65, 2018.

[132] Mingqiang Chen, Lin Fang, and Huafeng 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)*, pages 764–767. IEEE, 2019.

[133] Yeonggul Jang, Yoonmi Hong, Seongmin Ha, Sekeun Kim, and Hyuk-Jae Chang. Automatic segmentation of lv and rv in cardiac mri. In *Statistical Atlases and*

Computational Models of the Heart. ACDC and MMWHS Challenges: 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017, Quebec City, Canada, September 10-14, 2017, Revised Selected Papers 8, pages 161–169. Springer, 2018.

[134] Caizi Li, Qianqian Tong, Xiangyun Liao, Weixin Si, Shu Chen, Qiong Wang, and Zhiyong Yuan. Apcp-net: Aggregated parallel cross-scale pyramid network for cmr segmentation. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 784–788. IEEE, 2019.

[135] Clement Zotti, Zhiming Luo, Alain Lalande, and Pierre-Marc Jodoin. Convolutional neural network with shape prior applied to cardiac mri segmentation. *IEEE journal of biomedical and health informatics*, 23(3):1119–1128, 2018.

[136] Qiao Zheng, Hervé Delingette, Nicolas Duchateau, and Nicholas Ayache. 3-d consistent and robust segmentation of cardiac images by deep learning with spatial propagation. *IEEE transactions on medical imaging*, 37(9):2137–2148, 2018.

[137] Jay Patravali, Shubham Jain, and Sasank Chilamkurthy. 2d-3d fully convolutional neural networks for cardiac mr segmentation. In *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges: 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017, Quebec City, Canada, September 10-14, 2017, Revised Selected Papers 8*, pages 130–139. Springer, 2018.

[138] Xiuquan Du, Susu Yin, Renjun Tang, Yanping Zhang, and Shuo 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*, 7:1–10, 2019.

[139] Wenjun Yan, Yuanyuan Wang, Zeju Li, Rob J Van Der Geest, and Qian Tao. Left ventricle segmentation via optical-flow-net from short-axis cine mri: preserving the temporal coherence of cardiac motion. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11*, pages 613–621. Springer, 2018.

[140] Nicoló Savioli, Miguel Silva Vieira, Pablo Lamata, and Giovanni Montana. Automated segmentation on the entire cardiac cycle using a deep learning work-flow. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 153–158. IEEE, 2018.

[141] Chen Qin, Wenjia Bai, Jo Schlemper, Steffen E Petersen, Stefan K Piechnik, Stefan Neubauer, and Daniel Rueckert. Joint motion estimation and segmentation from undersampled cardiac mr image. In *Machine Learning for Medical Image Reconstruction: First International Workshop, MLMIR 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 1*, pages 55–63. Springer, 2018.

[142] Shusil Dangi, Ziv Yaniv, and Cristian A Linte. Left ventricle segmentation and quantification from cardiac cine mr images via multi-task learning. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 21–31. Springer, 2018.

[143] Liang Zhang, Georgios Vasileios Karanikolas, Mehmet Akçakaya, and Georgios B Giannakis. Fully automatic segmentation of the right ventricle via multi-task deep neural networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6677–6681. IEEE, 2018.

[144] Bob D De Vos, Floris F Berendsen, Max A Viergever, Hessam Sokooti, Marius Staring, and Ivana Išgum. A deep learning framework for unsupervised affine and deformable image registration. *Medical image analysis*, 52:128–143, 2019.

[145] Xiang Chen, Andres Diaz-Pinto, Nishant Ravikumar, and Alejandro F Frangi. Deep learning in medical image registration. *Progress in Biomedical Engineering*, 3(1):012003, 2021.

[146] Azira Khalil, Siew-Cheok Ng, Yih Miin Liew, and Khin Wee Lai. An overview on image registration techniques for cardiac diagnosis and treatment. *Cardiology research and practice*, 2018(1):1437125, 2018.

[147] Xuesong Lu, Rongqian Yang, Qinlan Xie, Shanxing Ou, Yunfei Zha, and Defeng Wang. Nonrigid registration with corresponding points constraint for automatic segmentation of cardiac dsct images. *Biomedical Engineering Online*, 16:1–15, 2017.

[148] Guanyu Yang, Yang Chen, Lijun Tang, Huazhong Shu, and Christine Toumoulin. Automatic left ventricle segmentation based on multiatlas registration in 4d ct images. In *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, pages 413–416. IEEE, 2014.

[149] Jan Ehrhardt, Timo Kepp, Alexander Schmidt-Richberg, and Heinz Handels. Joint multi-object registration and segmentation of left and right cardiac ventricles in 4d cine mri. In *Medical Imaging 2014: Image Processing*, volume 9034, pages 149–156. SPIE, 2014.

[150] Marc-Michel Rohé, Manasi Datar, Tobias Heimann, Maxime Sermesant, and Xavier Pennec. Svf-net: learning deformable image registration using shape matching. In *Medical Image Computing and Computer Assisted Intervention-MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*, pages 266–274. Springer, 2017.

[151] B. D. Vos, Floris F. Berendsen, Max A. Viergever, Marius Staring, and Ivana Igum. End-to-end unsupervised deformable image registration with a convolutional neural network. *ArXiv*, abs/1704.06065, 2017.

[152] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8):1788–1800, 2019.

[153] Roshan Reddy Upendra, Brian Jamison Wentz, Richard Simon, Suzanne M Shontz, and Cristian A Linte. Cnn-based cardiac motion extraction to generate deformable geometric left ventricle myocardial models from cine mri. In *International Conference on Functional Imaging and Modeling of the Heart*, pages 253–263. Springer, 2021.

[154] Jiayi Lu, Renchao Jin, Manyang Wang, Enmin Song, and Guangzhi Ma. A bidirectional registration neural network for cardiac motion tracking using cine mri images. *Computers in Biology and Medicine*, 160:107001, 2023.

[155] Roshan Reddy Upendra, SM Kamrul Hasan, Richard Simon, Brian Jamison Wentz, Suzanne M Shontz, Michael S Sacks, and Cristian A Linte. Motion extraction of the right ventricle from 4d cardiac cine mri using a deep learning-based

deformable registration framework. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3795–3799. IEEE, 2021.

[156] Roshan Reddy Upendra, Richard A. Simon, Suzanne Shontz, and Cristian A. Linte. Deformable image registration using vision transformers for cardiac motion estimation from cine cardiac mri images. In *Functional Imaging and Modeling of the Heart*, 2023.

[157] Chen Qin, Wenjia Bai, Jo Schlemper, Steffen Erhard Petersen, Stefan K. Piechnik, Stefan Neubauer, and Daniel Rueckert. Joint motion estimation and segmentation from undersampled cardiac mr image. *ArXiv*, abs/1908.07623, 2018.

[158] Nick Chater, Joshua B Tenenbaum, and Alan Yuille. Probabilistic models of cognition: Conceptual foundations. *Trends in cognitive sciences*, 10(7):287–291, 2006.

[159] Adrian V. Dalca, Guha Balakrishnan, John V. Guttag, and Mert Rory Sabuncu. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *Medical image analysis*, 57:226–236, 2019.

[160] Julian Krebs, Hervé Delingette, Nicholas Ayache, and Tommaso Mansi. Learning a generative motion model from image sequences based on a latent motion matrix. *IEEE Transactions on Medical Imaging*, 40:1405–1416, 2020.

[161] Verena Schulte-Frohlinde, Yosef Ashkenazy, Ary L Goldberger, Plamen Ch Ivanov, Madalena Costa, Adrian Morley-Davies, H Eugene Stanley, and Leon Glass. Complex patterns of abnormal heartbeats. *Physical review E*, 66(3):031901, 2002.

[162] Sabha Bhatti, Srikanth Vallurupalli, Stephanie Ambach, Adam Magier, Evan Watts, Vien Truong, Abdul Hakeem, and Wojciech Mazur. Myocardial strain pattern in patients with cardiac amyloidosis secondary to multiple myeloma: a cardiac mri feature tracking study. *The International Journal of Cardiovascular Imaging*, 34:27–33, 2018.

[163] Yossi Tsadok, Zvi Friedman, Brian A Haluska, Rainer Hoffmann, and Dan Adam.

Myocardial strain assessment by cine cardiac magnetic resonance imaging using non-rigid registration. *Magnetic Resonance Imaging*, 34(4):381–390, 2016.

[164] Ryutaro Onishi, Junpei Ueda, Seiko Ide, Masahiro Koseki, Yasushi Sakata, and Shigeyoshi Saito. Application of magnetic resonance strain analysis using feature tracking in a myocardial infarction model. *Tomography*, 9(2):871–882, 2023.

[165] Vien T Truong, Cassady Palmer, Sarah Wolking, Brandy Sheets, Michael Young, Tam NM Ngo, Michael Taylor, Sherif F Nagueh, Karolina M Zareba, Subha Raman, et al. Normal left atrial strain and strain rate using cardiac magnetic resonance feature tracking in healthy volunteers. *European Heart Journal-Cardiovascular Imaging*, 21(4):446–453, 2020.

[166] Elias A. Zerhouni, David Michael Parish, Walter J. Rogers, Andrew Yang, and Edward P. Shapiro. Human heart: tagging with mr imaging–a method for non-invasive assessment of myocardial motion. *Radiology*, 169 1:59–63, 1988.

[167] V Uma Valeti, Wookjin Chun, Donald D Potter, Philip A Araoz, Kiaran P McGee, James F Glockner, and Timothy F Christian. Myocardial tagging and strain analysis at 3 tesla: comparison with 1.5 tesla imaging. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 23(4):477–480, 2006.

[168] Daniel B Ennis, Frederick H Epstein, Peter Kellman, Lameh Fananapazir, Elliot R McVeigh, and Andrew E Arai. Assessment of regional systolic and diastolic dysfunction in familial hypertrophic cardiomyopathy using mr tagging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 50(3):638–642, 2003.

[169] Thomas S Denney Jr, Bernhard L Gerber, and Litao Yan. Unsupervised reconstruction of a three-dimensional left ventricular strain from parallel tagged cardiac images. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 49(4):743–754, 2003.

[170] Noemi Carranza-Herrezuelo, Ana Bajo, Filip Sroubek, Cristina Santamarta, Gabriel Cristóbal, Andrés Santos, and María J Ledesma-Carbayo. Motion estimation of tagged cardiac magnetic resonance images using variational techniques. *Computerized Medical Imaging and Graphics*, 34(6):514–522, 2010.

[171] Nael F Osman. Detecting stiff masses using strain-encoded (senc) imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 49(3):605–608, 2003.

[172] Meng Ye, Mikael Kanski, Dong Yang, Qi Chang, Zhennan Yan, Qiaoying Huang, Leon Axel, and Dimitris Metaxas. Deeptag: An unsupervised deep learning method for motion tracking on cardiac tagging magnetic resonance images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7261–7271, 2021.

[173] Manuel A Morales, Maaike Van den Boomen, Christopher Nguyen, Jayashree Kalpathy-Cramer, Bruce R Rosen, Collin M Stultz, David Izquierdo-Garcia, and Ciprian Catana. Deepstrain: a deep learning workflow for the automated characterization of cardiac mechanics. *Frontiers in cardiovascular medicine*, 8:730316, 2021.

[174] Edward Ferdian, Avan Suinesiaputra, Kenneth Fung, Nay Aung, Elena Lukaschuk, Ahmet Barutcu, Edd Maclean, Jose Paiva, Stefan K Piechnik, Stefan 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*, 2(1):e190032, 2020.

[175] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared Oâ€™Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

[176] Steffen E Petersen, Nay Aung, Mihir M Sanghvi, Filip Zemrak, Kenneth Fung, Jose Miguel Paiva, Jane M Francis, Mohammed Y Khanji, Elena Lukaschuk, Aaron 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*, 19(1):18, 2016.

[177] DS Mansell, Vito Domenico Bruno, Eva Sammut, Amedeo Chiribiri, Thomas Johnson, Igor Khaliulin, D Baz Lopez, HS Gill, KH Fraser, M Murphy, et al. Acute regional changes in myocardial strain may predict ventricular remodel-

ling after myocardial infarction in a large animal model. *Scientific Reports*, 11(1):18322, 2021.

[178] Sara Shimoni, Gera Gendelman, Oded Ayzenberg, Nahum Smirin, Peter Lysyansky, Orly Edri, Lisa Deutsch, Avraham Caspi, and Zvi Friedman. Differential effects of coronary artery stenosis on myocardial function: the value of myocardial strain analysis for the detection of coronary artery disease. *Journal of the American Society of Echocardiography*, 24(7):748–757, 2011.

[179] Laurens F Tops, Victoria Delgado, Nina Ajmone Marsan, and Jeroen J Bax. Myocardial strain to detect subtle left ventricular systolic dysfunction. *European journal of heart failure*, 19(3):307–313, 2017.

[180] Prabhakar Shantha Rajiah, Christopher J François, and Tim Leiner. Cardiac mri: state of the art. *Radiology*, 307(3):e223008, 2023.

[181] Andrew B Rosenkrantz, Mishal Mendiratta-Lala, Brian J Bartholmai, Dhakshinamoorthy Ganeshan, Richard G Abramson, Kirsteen R Burton, J Yu John-Paul, Ernest M Scalzetti, Thomas E Yankeelov, Rathan M Subramaniam, et al. Clinical utility of quantitative imaging. *Academic radiology*, 22(1):33–49, 2015.

[182] Nicole Seiberlich, Vikas Gulani, Adrienne Campbell-Washburn, Steven Sourbron, Mariya Ivanova Doneva, Fernando Calamante, and Houchun Harry Hu. *Quantitative magnetic resonance imaging*. Academic Press, 2020.

[183] Aaron D Curtis and Hai-Ling M Cheng. Primer and historical review on rapid cardiac cine mri. *Journal of Magnetic Resonance Imaging*, 55(2):373–388, 2022.

[184] Michael A. Guttman, Jerry L Prince, and Elliot R. McVeigh. Tag and contour detection in tagged mr images of the left ventricle. *IEEE transactions on medical imaging*, 13 1:74–88, 1994.

[185] Mi-Young Jeung, Philippe Germain, Pierre Croisille, Soraya El ghannudi, Catherine Roy, and Afshin Gangi. Myocardial tagging with mr imaging: overview of normal and pathologic findings. *Radiographics*, 32(5):1381–1398, 2012.

[186] Jamal N Khan, Anvesha Singh, Sheraz A Nazir, Prathap Kanagala, Anthony H Gershlick, and Gerry 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*, 84(5):840–848, 2015.

[187] Kevin W Moser, John G Georgiadis, and Richard O Buckius. On the use of optical flow methods with spin-tagging magnetic resonance imaging. *Annals of biomedical engineering*, 29(1):9, 2001.

[188] Sven Loncaric and Zoran Majcenic. Optical flow algorithm for cardiac motion estimation. In *Proceedings of the 22nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (Cat. No. 00CH37143)*, volume 1, pages 415–417. IEEE, 2000.

[189] Ayman M Khalifa, ABM Youssef, and Nael F Osman. Improved harmonic phase (harp) method for motion tracking a tagged cardiac mr images. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pages 4298–4301. IEEE, 2006.

[190] Xiaokai Wang, Maureen L Stone, Jerry L Prince, and Arnold D Gomez. A novel filtering approach for 3d harmonic phase analysis of tagged mri. In *Medical Imaging 2018: Image Processing*, volume 10574, pages 277–284. SPIE, 2018.

[191] Kang Wang, Rui Zhao, and Qiang Ji. A hierarchical generative model for eye image synthesis and eye gaze estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 440–448, 2018.

[192] Youbao Tang, Yuxing Tang, Yingying Zhu, Jing Xiao, and Ronald M Summers. A disentangled generative model for disease decomposition in chest x-rays via normal image synthesis. *Medical Image Analysis*, 67:101839, 2021.

[193] Chen Qin, Shuo Wang, Chen Chen, Huaqi Qiu, Wenjia Bai, and Daniel Rueckert. Biomechanics-informed neural networks for myocardial motion tracking in mri. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*, pages 296–306. Springer, 2020.

[194] Luigi Antelmi, Nicholas Ayache, Philippe Robert, and Marco Lorenzi. Sparse multi-channel variational autoencoder for the joint analysis of heterogeneous data. In *International Conference on Machine Learning*, pages 302–311. PMLR, 2019.

[195] François Chollet. Xception: Deep learning with depthwise separable convolutions. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 1251–1258, 2017.

[196] Hirokazu Kameoka, Li Li, Shota Inoue, and Shoji Makino. Semi-blind source separation with multichannel variational autoencoder. *arXiv preprint arXiv:1808.00892*, 2018.

[197] James M Joyce. Kullback-leibler divergence. In *International encyclopedia of statistical science*, pages 720–722. Springer, 2011.

[198] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

[199] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.

[200] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[201] Steffen E Petersen, Paul M Matthews, Fabian Bamberg, David A Bluemke, Jane M Francis, Matthias G Friedrich, Paul Leeson, Eike Nagel, Sven Plein, Frank 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*, 15(1):1–10, 2013.

[202] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[203] Peter Ndajah, Hisakazu Kikuchi, Masahiro Yukawa, Hidenori Watanabe, and Shogo Muramatsu. An investigation on the quality of denoised images. *International Journal of Circuit, Systems, and Signal Processing*, 5(4):423–434, 2011.

[204] D Poobathy and R Manicka Chezian. Edge detection operators: Peak signal to noise ratio based comparison. *IJ Image, Graphics and Signal Processing*, 10:55–61, 2014.

[205] Richard Dosselmann and Xue Dong Yang. A comprehensive assessment of the structural similarity index. *Signal, Image and Video Processing*, 5:81–91, 2011.

[206] Tianfeng Chai and Roland R Draxler. Root mean square error (rmse) or mean absolute error (mae). *Geoscientific model development discussions*, 7(1):1525–1534, 2014.

[207] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

[208] Nurettin Özgür Doğan. Bland-altman analysis: A paradigm to understand correlation and agreement. *Turkish journal of emergency medicine*, 18(4):139–141, 2018.

[209] American Heart Association Writing Group on Myocardial Segmentation, Registration for Cardiac Imaging:, Manuel D Cerqueira, Neil J Weissman, Vasken Dilsizian, Alice K Jacobs, Sanjiv Kaul, Warren K Laskey, Dudley J Pennell, John A Rumberger, Thomas 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*, 105(4):539–542, 2002.

[210] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248, 2017.

[211] Yuzhen Lu, Dong Chen, Ebenezer Olaniyi, and Yanbo Huang. Generative adversarial networks (gans) for image augmentation in agriculture: A systematic review. *Computers and Electronics in Agriculture*, 200:107208, 2022.

[212] Mohamed El-Kaddoury, Abdelhak Mahmoudi, and Mohammed Majid Himmi. Deep generative models for image generation: A practical comparison between variational autoencoders and generative adversarial networks. In *Mobile, Secure,*

and Programmable Networking: 5th International Conference, MSPN 2019, Mohammedia, Morocco, April 23–24, 2019, Revised Selected Papers 5, pages 1–8. Springer, 2019.

[213] Walter HL Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F Da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Brain imaging generation with latent diffusion models. In *MICCAI Workshop on Deep Generative Models*, pages 117–126. Springer, 2022.

[214] Aman Shrivastava and P Thomas Fletcher. Nasdm: Nuclei-aware semantic histopathology image generation using diffusion models. *arXiv preprint arXiv:2303.11477*, 2023.

[215] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[216] Wai-Ki Ching and Michael K Ng. Markov chains. *Models, algorithms and applications*, 2006.

[217] Yan Xia, Xiang Chen, Nishant Ravikumar, Christopher Kelly, Rahman Attar, Nay Aung, Stefan Neubauer, Steffen E Petersen, and Alejandro F Frangi. Automatic 3d+ t four-chamber cmr quantification of the uk biobank: integrating imaging and non-imaging data priors at scale. *Medical Image Analysis*, 80:102498, 2022.

[218] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

[219] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.

[220] Artem Obukhov and Mikhail Krasnyanskiy. Quality assessment method for gan based on modified metrics inception score and fréchet inception distance. In *Software Engineering Perspectives in Intelligent Systems: Proceedings of 4th Computational Methods in Systems and Software 2020, Vol. 1 4*, pages 102–114. Springer, 2020.

[221] Hao Tang, Wei Wang, Dan Xu, Yan Yan, and Nicu Sebe. Gesturegan for hand gesture-to-gesture translation in the wild. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 774–782, 2018.

[222] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems*, 32, 2019.

[223] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003.

[224] Xiaoling Xia, Cui Xu, and Bing Nan. Inception-v3 for flower classification. In *2017 2nd international conference on image, vision and computing (ICIVC)*, pages 783–787. IEEE, 2017.

[225] Felix Nensa. The future of radiology: The path towards multimodal ai and superdiagnostics. *European Journal of Radiology Artificial Intelligence*, page 100014, 2025.

[226] Hongwei Ding, Qi Tao, and Nana Huang. Bdgan: Boundary and diversity-aware generative adversarial network for imbalanced medical image augmentation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[227] Jiahao Xia, Yutao Hu, Yaolei Qi, Zhenliang Li, Wenqi Shao, Junjun He, Ying Fu, Longjiang Zhang, and Guanyu Yang. Fcas: Fine-grained cardiac image synthesis based on 3d template conditional diffusion model. *arXiv preprint arXiv:2503.09560*, 2025.

[228] Tewodros Weldebirhan Arega, François Legrand, Stéphanie Bricq, and Fabrice Meriaudeau. Using mri-specific data augmentation to enhance the segmentation of right ventricle in multi-disease, multi-center and multi-view cardiac mri. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 250–258. Springer, 2021.

[229] Edward Ferdian, Debbie Zhao, Gonzalo D Maso Talou, Gina M Quill, Malcolm E Legget, Robert N Doughty, Martyn P Nash, and Alistair A Young. Diff · 3:

A latent diffusion model for the generation of synthetic 3d echocardiographic images and corresponding labels. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 129–140. Springer, 2023.

[230] Subhi J Al'Aref, Khalil Anchouche, Gurpreet Singh, Piotr J Slomka, Kranthi K Kolli, Amit Kumar, Mohit Pandey, Gabriel Maliakal, Alexander R Van Rosendael, Ashley N Beecy, et al. Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. *European heart journal*, 40(24):1975–1986, 2019.

[231] Marcelo F Di Carli, Tal Geva, and Ravin Davidoff. The future of cardiovascular imaging. *Circulation*, 133(25):2640–2661, 2016.

[232] Udo Sechtem, Peter Pflugfelder, and Charles B Higgins. Quantification of cardiac function by conventional and cine magnetic resonance imaging. *Cardiovascular and interventional radiology*, 10:365–373, 1987.

[233] Bart Bijnens, M Cikes, C Butakoff, Marta Sitges, and Fatima Crispi. Myocardial motion and deformation: What does it tell us and how does it relate to function? *Fetal diagnosis and therapy*, 32(1-2):5–16, 2012.

[234] Marc Górriz Blanch, Marta Mrak, Alan F Smeaton, and Noel E O'Connor. End-to-end conditional gan-based architectures for image colourisation. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE, 2019.

[235] William Harvey, Saeid Naderiparizi, and Frank Wood. Conditional image generation by conditioning variational auto-encoders. *arXiv preprint arXiv:2102.12037*, 2021.

[236] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (ToG)*, 40(3):1–21, 2021.

[237] Patrick Schlosser, David Munch, and Michael Arens. Investigation on combining 3d convolution of image data and optical flow to generate temporal action pro-

posals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[238] Yuanzhi Zhu, Zhaohai Li, Tianwei Wang, Mengchao He, and Cong Yao. Conditional text image generation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14235–14245, 2023.

[239] Juyeon Ko, Inho Kong, and Hyunwoo J Kim. Stochastic conditional diffusion models for semantic image synthesis. *arXiv preprint arXiv:2402.16506*, 2024.

[240] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[241] Yu Li, Randi Fu, Xiangchao Meng, Wei Jin, and Feng Shao. A sar-to-optical image translation method based on conditional generation adversarial network (cgan). *Ieee Access*, 8:60338–60343, 2020.

[242] Avi Ben-Cohen, Eyal Klang, Stephen P Raskin, Shelly Soffer, Simona Ben-Haim, Eli Konen, Michal Marianne Amitai, and Hayit Greenspan. Cross-modality synthesis from ct to pet using fcn and gan networks for improved automated lesion detection. *Engineering Applications of Artificial Intelligence*, 78:186–194, 2019.

[243] Hoo-Chang Shin, Neil A Tenenholtz, Jameson K Rogers, Christopher G Schwarz, Matthew L Senjem, Jeffrey L Gunter, Katherine P Andriole, and Mark Michalski. Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In *Simulation and Synthesis in Medical Imaging: Third International Workshop, SASHIMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*, pages 1–11. Springer, 2018.

[244] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016.

[245] Yanlong Dong, Ying Zhang, Lin Ma, Zhi Wang, and Jiebo Luo. Unsupervised text-to-image synthesis. *Pattern Recognition*, 110:107573, 2021.

[246] Eric Wu, Kevin Wu, David D. Cox, and William Lotter. Conditional infilling gans for data augmentation in mammogram classification. *ArXiv*, abs/1807.08093, 2018.

[247] Euijin Jung, Miguel Luna, and Sanghyun Park. Conditional generative adversarial network for predicting 3d medical images affected by alzheimer's diseases. In *PRIME@MICCAI*, 2020.

[248] Tian Xia, Agisilaos Chartsias, Chengjia Wang, and Sotirios A. Tsaftaris. Learning to synthesise the ageing brain without longitudinal data. *Medical image analysis*, 73:102169, 2019.

[249] Carlo Biffi, Juan J. Cerrolaza, Giacomo Tarroni, Wenjia Bai, Antonio de Marvao, Ozan Oktay, Christian Ledig, Loic Le Folgoc, Konstantinos Kamnitsas, Georgia Doumou, Jinming Duan, Sanjay K. Prasad, Stuart A. Cook, Declan P. Oâ€™Regan, and Daniel Rueckert. Explainable anatomical shape analysis through deep hierarchical generative models. *IEEE Transactions on Medical Imaging*, 39:2088–2099, 2019.

[250] Hadrien Reynaud, Athanasios Vlontzos, Mischa Dombrowski, Ciarán M. Lee, Arian Beqiri, Paul Leeson, and Bernhard Kainz. D'artagnan: Counterfactual video generation. *ArXiv*, abs/2206.01651, 2022.

[251] Víctor M. Campello, Tian Xia, Xiao Liu, Pedro Sanchez, Carlos Martín-Isla, Steffen E. Petersen, S. Seguí, Sotirios A. Tsaftaris, and Karim Lekadir. Cardiac aging synthesis from cross-sectional data with conditional generative adversarial networks. *Frontiers in Cardiovascular Medicine*, 9, 2022.

[252] Sina Amirrajab, Yasmina Al Khalil, Cristian Lorenz, Jürgen Weese, Josien P. W. Pluim, and Marcel M. Breeuwer. A framework for simulating cardiac mr images with varying anatomy and contrast. *IEEE Transactions on Medical Imaging*, 42:726–738, 2022.

[253] Mengyun Qiao, Shuo Wang, Huaqi Qiu, Antonio de Marvao, Declan P. Oâ€™Regan, Daniel Rueckert, and Wenjia Bai. Cheart: A conditional spatio-temporal generative model for cardiac anatomy. *IEEE transactions on medical imaging*, 43:1259 – 1269, 2023.

[254] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022.

[255] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[256] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[257] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, and Chunhua Shen. Conditional positional encodings for vision transformers. *arXiv preprint arXiv:2102.10882*, 2021.

[258] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[259] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019.

[260] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[261] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021.

[262] Wenjia Bai, Hideaki Suzuki, Jian Huang, Catherine Francis, Shuo Wang, Giacomo Tarroni, Florian Guitton, Nay Aung, Kenneth Fung, Steffen E Petersen, et al. A population-based phenome-wide association study of cardiac and aortic structure and function. *Nature medicine*, 26(10):1654–1662, 2020.

[263] Yuheng Bu, Shaofeng Zou, Yingbin Liang, and Venugopal V Veeravalli. Estimation of kl divergence: Optimal minimax rate. *IEEE Transactions on Information Theory*, 64(4):2648–2674, 2018.

[264] Huidong Liu, GU Xianfeng, and Dimitris Samaras. A two-step computation of the exact gan wasserstein distance. In *International Conference on Machine Learning*, pages 3159–3168. PMLR, 2018.

[265] Mengyun Qiao, Shuo Wang, Huaqi Qiu, Antonio De Marvao, Declan P Oâ€™Regan, Daniel Rueckert, and Wenjia Bai. Cheart: A conditional spatio-temporal generative model for cardiac anatomy. *IEEE transactions on medical imaging*, 2023.

[266] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[267] Cong Liu, Xiaohan Yuan, ZhiPeng Yu, and Yangang Wang. Texdc: Text-driven disease-aware 4d cardiac cine mri images generation. In *Proceedings of the Asian Conference on Computer Vision*, pages 3005–3021, 2024.

[268] Jee Seok Yoon, Chenghao Zhang, Heung-Il Suk, Jia Guo, and Xiaoxiao Li. Sadm: Sequence-aware diffusion model for longitudinal medical image generation. In *International Conference on Information Processing in Medical Imaging*, pages 388–400. Springer, 2023.

[269] Boah Kim and Jong Chul Ye. Diffusion deformable model for 4d temporal medical image generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 539–548. Springer, 2022.

[270] Cherry Khosla and Baljit Singh Saini. Enhancing performance of deep learning models with different data augmentation techniques: A survey. In *2020 International Conference on Intelligent Engineering and Management (ICIEM)*, pages 79–85. IEEE, 2020.

[271] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A Mann. Data augmentation can improve robustness. *Advances in Neural Information Processing Systems*, 34:29935–29948, 2021.

[272] Richard J Chen, Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6):493–497, 2021.

[273] Yan Xia, Nishant Ravikumar, Toni Lassila, and Alejandro F Frangi. Virtual high-resolution mr angiography from non-angiographic multi-contrast mris: synthetic vascular model populations for in-silico trials. *Medical Image Analysis*, 87:102814, 2023.

[274] Zhaozhi Qian, Thomas Callender, Bogdan Cebere, Sam M Janes, Neal Navani, and Mihaela van der Schaar. Synthetic data for privacy-preserving clinical risk prediction. *medRxiv*, pages 2023–05, 2023.