Addressing Data Scarcity in Cardiac MRI Segmentation with Anatomy-Compliant Image-to-Image Translation via Latent Diffusion Models

Nuno Miguel Ferreira Capitão
Master in Engineering Physics
Department of Physics and Astronomy
Faculty of Sciences and Faculty of Engineering of the University of Porto
2024

Addressing Data Scarcity in Cardiac MRI Segmentation with Anatomy-Compliant Image-to-Image Translation via Latent Diffusion Models

Nuno Miguel Ferreira Capitão

Dissertation carried out as part of the Master in Engineering Physics

Department of Physics and Astronomy

2024



Supervisor
Prof. Dr. João V. Lopes, Auxiliar Professor, Department of Physics and Astronomy, Faculty of Sciences of University of Porto

Co-supervisor
Prof. Dr. Qian Tao, Assistant Professor, Department of Imaging Physics, Delft University of Technology

" A computer would deserve to be called intelligent if it could deceive human into believing that it was human. "

Alan Turing

Acknowledgements

I would like to express my gratitude to my host supervisor at TU Delft, Prof. Dr. Qian Tao, to whom I will always be thankful for providing me with the opportunity to begin my research journey, as well as for the immense support and guidance throughout the process. A deep and heartfelt thanks to Yidong Zhao, one of the smartest and most supportive people I have ever met, for always making time to help me, sharing his knowledge, and teaching me how to rigorously follow best research practices. To Yi, my football chat pal, for always being available to share his insights when I needed them. I would also like to acknowledge Nathan for generously sharing his work and ideas, which played an important role in this project. To everyone in the Tao group: it was truly an honor to work with such incredibly intelligent, kind, and supportive people. This has been an extraordinary experience since day one, and none of my successes would have been possible without everyone involved. My appreciation also goes to everyone at the ImPhys department for providing me with the resources necessary to make this project possible.

I also want to extend my heartfelt thanks to Prof. Dr. João Viana Lopes for agreeing to supervise my thesis and for his unwavering support throughout the year. I have long admired his work ethic and high standards, which inspire us as students to excel in our careers. On a personal note, I am deeply thankful to my parents and my brother, who have always sacrificed so much to ensure I had all the tools to succeed academically and for uplifting me during hard times.

Finally, to all my friends from the Student Hotel days, thank you for the unforgettable memories that made Delft such a special place and my experience unforgettable. To João Folhadela, my study pal throughout all our years at university, thank you for keeping me sane and pushing through the numerous challenges with me. Last but not least, to all my work colleagues at Fibersail, thank you for your support and for providing the flexibility that allowed me to pursue my academic goals abroad.

UNIVERSIDADE DO PORTO

Abstract

Faculdade de Ciências da Universidade do Porto
Departamento de Física e Astronomia

MSc. Engineering Physics

Addressing Data Scarcity in Cardiac MRI Segmentation with Anatomy-Compliant Image-to-Image Translation via Latent Diffusion Models

by Nuno Miguel Ferreira CAPITÃO

Learning-based solutions have become increasingly prominent in Cardiac Magnetic Resonance Imaging (CMRI) for automated segmentation tasks, providing radiologists with tools to streamline labor-intensive and error-prone manual segmentation. However, data scarcity presents a significant challenge to achieving optimal model performance due to the limited availability of high-quality data. One potential solution is to synthesize medical images using powerful generative models for style-transfer tasks. Nevertheless, ensuring high fidelity in tissue-specific texturization requires conditioning priors, which are often unavailable or consist of partially labeled annotations, leading to generated images with incoherent or hallucinated anatomy.

CMR images can be reproduced using Bloch simulations on digital phantoms, which provide precise numerical representations of anatomical structures. While these simulations offer controllable and readily available synthetic images, their textures still appear markedly unrealistic compared to in-vivo data, limiting their practical use. To address this issue, one strategy involves aligning the anatomical masks in phantoms with labeled in-vivo tissues to enable style transfer through semantic image synthesis. However, obtaining fully labeled annotations for all organs is impractical, rendering this approach infeasible.

In this work, we propose an innovative medical image generation method based on state-of-the-art latent diffusion models (LDM) to overcome the limitations associated with in-vivo data availability via style transfer techniques. To address the challenge of anatomical compliance, we harness edge information, which is general and easy to compute

within the full FoV. Furthermore, we introduce a novel framework for semantic image synthesis by incorporating Spatially Adaptive Normalization (SPADE) into LDMs — a technique predominantly used with Generative Adversarial Networks (GANs) but largely underexplored in diffusion-based models. Our SPADE-LDM variants leverage both edge information and anatomical masks as dual conditioning and do not require extra annotations to achieve anatomical compliance.

We conducted experiments using Canny, Sobel, and Laplacian operators to evaluate how varying complexities in edge maps, when used as conditioning priors, impact image generation performance and style transfer tasks. Our method was evaluated on the ACDC and MICCAI MSCMR 2019 Segmentation Challenge datasets and benchmarked against GAN-based models. Moreover, we applied our style transfer techniques to MRXCAT in-silico images and introduced a first-stage pipeline to generate pseudo-labels for in-vivo data, extending annotations to the lung region.

Results demonstrate that incorporating edge-based conditioning leads to high-quality, anatomically compliant medical image generation and strongly complements partial semantics. We achieved superior model performance using Sobel edges, reaching an FID score of 25.5 on the ACDC dataset, while relying entirely on edge conditioning. In contrast, models conditioned on Canny edges proved most effective for style transfer tasks. We seamlessly translated various imaging modalities into a single, well-represented modality, enabling accurate segmentation without requiring models trained on multi-sequence acquisitions. Segmentation models trained exclusively on our style-transferred in-silico images achieved highly competitive Dice Scores on MSCMR bSSFP sequences: 0.794 ± 0.055 (LV), 0.758 ± 0.081 (RV), and 0.713 ± 0.088 (Myo). This highlights our method as a promising tool for data augmentation, bridging the gap between physically simulated data and in-vivo CMRI.

Keywords: Latent Diffusion Models, Edge Detection, Deep Learning, Cardiac Magnetic Resonance Imaging, Image-to-Image Translation, Style Transfer, Automated Segmentation, Artificial Intelligence, Digital Phantoms, Numerical Simulations, Medical Image Synthesis.

UNIVERSIDADE DO PORTO

Resumo

Faculdade de Ciências da Universidade do Porto
Departamento de Física e Astronomia

Mestrado Integrado em Engenharia Física

Contornar a Escassez de Dados na Segmentação Cardíaca em RM Preservando a Estrutura Anatómica na Tradução de Imagens com *Latent Diffusion Models*

por Nuno Miguel Ferreira CAPITÃO

Modelos de inteligência artificial têm-se tornado cada vez mais comuns em imagens de Ressonância Magnética Cardíaca (RMC), de forma a equipar radiologistas com ferramentas que simplificam tarefas de segmentação manual, que, por norma, consomem muito tempo e são suscetíveis a erros. No entanto, dada a escassez de dados disponíveis no domínio, treinar modelos eficazes torna-se um desafio difícil de contornar. Uma possível solução passa pela síntese de imagens médicas utilizando modelos generativos para transferir contrastes entre imagens. Contudo, garantir a veracidade das texturas específicas de cada tecido requer o uso de *priors*, que, na sua maioria, não existem ou limitam-se a algumas anotações de radiologistas, resultando em imagens geradas com anatomia incoerente e irrealista.

Imagens de RMC podem ser reproduzidas através de simulações das equações de *Bloch* em fantomas digitais, que consistem em representações numéricas da anatomia humana. Embora estas simulações possam gerar imagens sintéticas de forma imediata, as texturas e os contrastes diferem significativamente daquilo que se observa habitualmente em *in-vivo*. Para ultrapassar esta limitação, uma estratégia envolve alinhar as anotações feitas por radiologistas com as representações dos fantomas, de modo a estabelecer uma correlação direta entre semânticas de cada tecido e, assim, transferir diretamente as texturas presentes em dados reais. Contudo, obter anotações para todos os órgãos é impraticável.

Neste trabalho, propomos um método inovador para gerar imagens médicas, baseado no estado da arte em modelos de difusão, *Latent Diffusion Models* (LDM), para contornar

algumas das limitações associadas à disponibilidade de dados *in-vivo*, através de transferências de estilos entre imagens. Para preservar a anatomia representada, utilizamos a informação dos contornos de cada estrutura, obtida através de *edge detection*, uma vez que esta pode ser facilmente calculada em todo o plano da imagem e é transversal aos diferentes contrastes. Além disso, propomos uma nova arquitetura para LDMs, com o objetivo de sintetizar imagens baseadas em semântica, utilizando a técnica *Spatially Adaptive Normalization (SPADE)* — que tem sido consistentemente utilizada em *Generative Adversarial Networks (GANs)*, mas que continua pouco explorada em modelos de difusão. Os nossos modelos *SPADE-LDM* são capazes de combinar informação das delineações dos órgãos, mapeados por *edge detection*, com segmentações parciais de radiologistas, de forma a gerar imagens com anatomia coerente sem requerer um número extenso de anotações.

Durante as nossas experiências, analisámos os operadores *Canny*, *Sobel* e *Laplacian* como métodos de *edge detection*, com o objetivo de avaliar o impacto dos seus diferentes níveis de complexidade quando usados como *priors* em tarefas distintas: geração de imagens e tradução entre modalidades. O nosso método foi avaliado nas bases de dados públicas *ACDC* e no desafio de segmentação *MICCAI MSCMR 2019*, comparando-o com modelos baseados em *GANs*. Além disso, aplicámos os nossos modelos para transferir contrastes *in-vivo* para imagens sintéticas (*in-silico*) simuladas com o programa *MRXCAT*, e introduzimos ainda um modelo preliminar no nosso método, capaz de gerar *pseudo*-anotações para a região pulmonar.

Os nossos resultados demonstram que o uso de mapas de *edge* como *priors* resulta em imagens médicas de elevada qualidade, representando adequadamente toda a anatomia no plano da imagem, funcionando assim como um complemento à falta de anotações de radiologistas. Os nossos modelos obtiveram o melhor desempenho com o método *Sobel*, alcançando um valor de 25.5 na métrica de *FID* nos dados *ACDC*, utilizando apenas *edge detection*. Por outro lado, mapas pelo método de *Canny* provaram ser mais eficazes na transferência de contrastes entre imagens. Com os nossos modelos, traduzimos vários tipos de modalidades de RMC para uma única modalidade com boa representatividade, permitindo treinar modelos de segmentação sem a necessidade de dispor de uma base de dados com vários tipos de aquisições. Os nossos modelos de segmentação, treinados exclusivamente nas imagens *in-silico* traduzidas, obtiveram resultados notáveis nas sequências *bSSFP* dos dados *MSCMR*: 0.794 ± 0.055 (ventrículo esquerdo), 0.758 ± 0.081 (ventrículo direito) e 0.713 ± 0.088 (miocárdio), medidos pela métrica *Dice Score*.

Tal evidencia o nosso método como uma ferramenta promissora para aproximar dados completamente simulados de imagens *in-vivo* de RMC.

Palavras-Chave: *Latent Diffusion Models*, *Edge Detection*, *Deep Learning*, Ressonância Magnética Cardíaca, Tradução de Imagens, Transferência de Estilo, Segmentação Automática, Inteligência Artificial, Fantomas Digitais, Simulações Numéricas, Síntese de Imagens Médicas.

Contents

Acknowledgements iii

Abstract v

Resumo vii

Contents xi

List of Figures xiii

List of Tables xv

Glossary xvii

1 Introduction 1
- 1.1 Cardiac Magnetic Resonance Imaging . . . 1
- 1.2 Automated CMRI Segmentation . . . 2
- 1.3 In-Silico Bloch-Simulated Images . . . 3
- 1.4 Generative Models for CMR Image Synthesis . . . 4
- 1.5 Project Outline . . . 5

2 Theoretical Background 7
- 2.1 Magnetic Resonance Imaging Physics . . . 7
  - 2.1.1 Longitudinal (Spin-Lattice) Relaxation . . . 11
  - 2.1.2 Late Gadolinium Enhancement . . . 13
  - 2.1.3 Transverse (Spin – Spin) Relaxation . . . 13
  - 2.1.4 Balanced Steady-State Free Precession . . . 15
  - 2.1.5 Cine Imaging . . . 16
- 2.2 Deep Generative Models . . . 17
  - 2.2.1 Historical Trends . . . 19
  - 2.2.2 Generative Adversarial Networks . . . 20
  - 2.2.3 Principles of Non-Equilibrium Thermodynamics in Diffusion Models 23
  - 2.2.4 Denoising Diffusion Probabilistic Models . . . 24
  - 2.2.5 Denoising Diffusion Implicit Models . . . 29
  - 2.2.6 Latent Diffusion Models . . . 33
  - 2.2.7 Spatially-Adaptive Normalization . . . 35

3 Methodology **37**
3.1 Anatomy-Compliant Medical Image Synthesis . . . 37
3.1.1 Image Edge Detection Operators . . . 38
3.1.2 Edge-Informed Spatially-Adaptive Latent Diffusion Models . . . 39
3.2 Image-to-Image Translation via Style-Transfer Techniques . . . 42
3.2.1 Automated Segmentation with nnU-Net . . . 43
3.2.2 MRXCAT Numerical CMR Simulations . . . 44
3.2.3 Pseudo-Labeled Lung Semantic Annotations . . . 45

4 Results **47**
4.1 Image Generation Performance . . . 47
4.1.1 Semantic Image Synthesis . . . 47
4.1.2 Edge-Informed Image Synthesis . . . 49
4.2 Downstream Segmentation . . . 51
4.2.1 Multi-Sequence Acquisitions . . . 51
4.2.2 In-Silico Images . . . 55

5 Discussion **59**

6 Conclusions **63**

A MIDL 2024 Conference: Short Paper Submission **65**

B Training Details **71**
B.1 SPADE-GAN . . . 71
B.2 VQ-AE . . . 73
B.3 LDM . . . 74

Bibliography **75**

List of Figures

2.1 Hydrogen nuclei energy levels under $\mathbf{B}_0$. 8
2.2 Hydrogen protons precessing around the z-axis and shifting towards the xy-plane under the influence of $\mathbf{B}_1$. 9
2.3 MRI spatial encoding with superimposed magnetic field gradients. 10
2.4 Longitudinal magnetization recovery (T_1). 12
2.5 Transverse magnetization recovery (T_2). 14
2.6 Cine 2D bSSFP imaging. 17
2.7 GAN Framework. 22
2.8 DDPM forward and reverse processes. 24
2.9 Accelerated sampling with DDIM. 31
2.10 Illustration of the LDM framework. 34
2.11 SPADE normalization mechanism. 36

3.1 Denoising preprocessing for Canny and Laplacian edges. 39
3.2 SPADE U-Net architecture. 40
3.3 LDM-based image synthesis with dual conditioning (e_x, y) 41
3.4 Automatic configuration in nnU-Net. 43
3.5 First-stage nnU-Net for pseudo-labeled lung annotations. 46

4.1 Qualitative results using semantic image synthesis ($e_x = \emptyset, y$). 48
4.2 Qualitative results using edge conditioning ($e_x, y = \emptyset$). 49
4.3 Qualitative results using dual conditioning (e_x, y). 50
4.4 Averaged spectral energy distribution. 51
4.5 nnU-Net qualitative results on translated T_1-weighted LGE imaging. 52
4.6 nnU-Net qualitative results on translated T_2-mapping. 53
4.7 nnU-Net qualitative results on translated T_1-mapping. 53
4.8 nnU-Net qualitative results on translated T_2-weighted imaging. 54
4.9 Bloch-simulated in-silico CMR images. 55
4.10 Style-transferred in-silico images. 56
4.11 Lung region comparison: 3-labeled vs. 4-labeled Canny-SPADE-LDM. 57

B.1 VAE-based SPADE-GAN architecture. 72

List of Tables

3.1 Simulation parameters for XCAT digital phantoms in cine CMR imaging. . 44
3.2 Tissue-specific parameters for Bloch simulations with MRXCAT. 45

4.1 Image generation performance on ACDC dataset. 48
4.2 Image generation performance on ACDC in a centrally cropped 128×128 area. 50
4.3 Segmentation performance on style-transferred T_1-weighted LGE and T_2-weighted sequences. 52
4.4 Segmentation performance on ACDC and MICCAI MSCMR bSSFP sequences evaluated by a nnU-Net trained on translated in-silico images. 56

B.1 Hyperparameter settings for VAE-based SPADE-GAN training. 71
B.2 Hyperparameter settings for VQ-AE training. 74
B.3 Hyperparameter settings for LDM training. 74

Glossary

AIS	Annealed Importance Sampling
AI	Artificial Intelligence
AdaIN	Adaptive Instance Normalization
ACDC	Automated Cardiac Diagnosis Challenge
AE	Autoencoder
BN	Batch Normalization
BP	Backpropagation
bSSFP	Balanced Steady State Free Precession
CAD	Computer Aided Diagnosis
CMRI	Cardiac Magnetic Resonance Imaging
CNN	Convolutional Neural Network
CVD	Cardiovascular Disease
DBN	Deep Belief Network
DDIM	Denoising Diffusion Implicit Models
DDPM	Denoising Diffusion Probabilistic Models
DL	Deep Learning
ECG	Electrocardiogram
ED	End Diastole
ELBO	Evidence Lower Bound
ES	End Systole
FID	Fréchet Inception Distance

FoV	Field of View
GAN	Generative Adversarial Networks
Gd	Gadolinium
GBCA	Gadolinium based Contrast Agent
GN	Group Normalization
IFT	Inverse Fourier Transform
IFFT	Inverse Fast Fourier Transform
IN	Instance Normalization
KL	Kullback-Leibler
LDM	Latent Diffusion Models
LGE	Late Gadolinium Enhancement
LV	Left Ventricle
MCMC	Markov Chain Monte Carlo
MICCAI	Medical Image Computing and Computer Assisted Intervention
MIDL	Medical Imaging with Deep Learning
MLE	Maximum Likelihood Estimation
MRI	Magnetic Resonance Imaging
MRXCAT	Magnetic Resonance Extended Cardiac Torso
Myo	Myocardium
MSCMR	Multi Sequence Cardiac Magnetic Resonance Imaging
NMSE	Normalized Mean Squared Error
NMR	Nuclear Magnetic Resonance
RBM	Restricted Boltzmann Machine
RF	Radiofrequency
ROI	Region of Interest
RV	Right Ventricle
SiLUs	Sigmoid Linear Units

SNR	Signal-to-Noise Ratio
SPADE	Spatially-Adaptive
SSIM	Similarity Index Measure
TE	Echo Time
TR	Repetition Time
VAE	Variational Autoencoders
VLB	Variational Lower Bound
VQ	Vector Quantized
XCAT	Extended Cardiac Torso

Chapter 1

Introduction

Medical imaging holds an indispensable role in modern healthcare, supporting routine clinical practice across all medical subspecialties. With rapid advancements in artificial intelligence (AI) and its prominence in cutting-edge applications across various fields, developing AI solutions for computer-aided diagnosis (CAD) in medical imaging has become a focal point of ongoing research [1]. These efforts aim to provide physicians with data-driven tools to automate labor-intensive tasks and streamline different stages within clinical workflows.

1.1 Cardiac Magnetic Resonance Imaging

Cardiac Magnetic Resonance Imaging (CMRI) is an advanced imaging technique used to evaluate and diagnose heart conditions. Grounded in the fundamental principles of Nuclear Magnetic Resonance (NMR), it produces detailed images of the heart tissues and surrounding structures without exposure to ionizing radiation. CMRI uses tissue-specific characterization techniques to render comprehensive anatomical, functional, and physiological information.

Balanced Steady-State Free Precession (bSSFP) is a widely used sequence in CMRI due to its ability to produce high signal-to-noise ratio (SNR) images with short acquisition times and excellent contrast between the blood pool and myocardium. It's particularly suitable for evaluating ventricular function and wall motion through cine dynamic imaging, which reconstructs the different phases of the cardiac cycle [2]. T_1-weighted MRI, on the other hand, is the preferred method for identifying myocardial scar tissue and

assessing overall myocardial health. It excels at detecting and quantifying changes in tissue composition [3], such as fibrosis and fatty infiltration, as both conditions exhibit prolonged T_1 relaxation times compared to healthy myocardium. These changes often occur alongside scar tissue formation following a myocardial infarction, indicating damaged areas. Advanced techniques such as contrast-enhanced Late Gadolinium Enhancement (LGE) are commonly used alongside T_1-weighted imaging to pinpoint myocardial scarring or increased extracellular space, providing a clear view of irreversible injury. LGE is crucial for diagnosing various ischemic and non-ischemic cardiomyopathies [4], as well as for risk stratification and clinical decision-making. Alternatively, T_2-weighted CMRI is particularly effective at identifying areas with high water content, such as edema or inflammation [5], as fluids typically exhibit long T_2 relaxation times and appear brighter. As a result, T_2-weighted imaging is most suitable for detecting conditions characterized by increased blood flow and fluid accumulation, such as acute injuries (e.g., heart attacks or trauma), and for distinguishing cardiac tumors based on their water content.

Each CMRI modality offers unique advantages and can be oriented toward specific diagnostic purposes, providing clinicians with a detailed perspective that facilitates accurate diagnosis, treatment planning, and patient monitoring.

1.2 Automated CMRI Segmentation

In standard clinical practice, a radiologist examines each image to manually segment and annotate the Region of Interest (ROI). This process involves classifying pixels within the heart and delineating distinct regions into anatomically meaningful areas. In CMRI, this usually means outlining the contours of the left ventricle (LV), right ventricle (RV), and myocardium (Myo). Precise segmentation highlights important physical, geometrical, and pathological details of tissues, enabling structural processing techniques to create 3D reconstructions and examine cardiac motion. Such information is essential for the diagnosis and treatment of cardiovascular diseases (CVDs), which are known to pose major health risks to the general population [6]. For instance, comparing segmentations between end-systole (ES) and end-diastole (ED) provides a method to estimate various cardiac metrics, such as ventricular volume and ejection fraction. These metrics describe the heart's pumping efficiency and overall function, enabling a thorough assessment of heart performance. Importantly, segmenting the myocardium reveals the mass of the heart muscle, which is a key indicator of cardiac health. Increased myocardial mass is

often linked to conditions such as hypertrophy, which can result from high blood pressure or other cardiac stresses.

Although manual segmentation benefits from expert knowledge, it is also a highly time-consuming and technically arduous task, prone to human error and inter-observer variability. To streamline this process, Deep Learning (DL) approaches, known for their exceptional performance in computer vision tasks [7], have become increasingly prominent in medical imaging. These DL techniques enable precise and efficient automated cardiac image segmentation, allowing experts to concentrate on more complex and meaningful tasks. Nevertheless, data scarcity poses a major challenge to achieving optimal model performance, as such models rely heavily on high-quality labeled data, which is frequently unavailable due to privacy concerns, rare pathologies, and a lack of annotation expertise [8].

1.3 In-Silico Bloch-Simulated Images

Medical images can be closely reproduced using numerical simulations (in-silico) on digital phantoms representing anatomical structures. In the context of MRI, model tissue properties can be derived from Bloch equations [9] to simulate the signal generated by tissues in response to variations in the magnetic field.

Hybrid methods integrate both analytical and voxel-based descriptions. Analytical descriptions employ mathematical tools like non-uniform rational B-splines or polygon meshes [10] to model tissue structures and simulate imaging experiments based on known sequence parameters. These purely analytical descriptions also allow the image domain to be converted into a well-defined k-space discrete Fourier transform, avoiding truncation errors [11]. Meanwhile, voxel-based descriptions, typically obtained from in-vivo data, provide realistic voxelized anatomical representations that incorporate the required motion for CMRI [12]. Although voxel-based models may not generalize to population statistics or pathological cases, the flexibility of analytical phantoms mitigates this limitation by incorporating anatomical variations. The underlying MRI physics are typically simulated using three-dimensional coil sensitivities derived from the Biot-Savart law, which involves discretizing the integral to calculate the magnetic field at each voxel position. Coil elements are arranged in one or more circular arrays around the phantom to replicate typical cardiac coil configurations around the chest. Tissue and sequence operators assign realistic MR signal intensities to the organs based on properties such as T_1,

T_2, proton density, and contrast agent concentration, according to predefined sequence timing parameters. A key benefit of generating synthetic images from phantoms is that these images include full-body anatomical masks.

A well-established in-silico phantom simulation tool for CMRI is the Magnetic Resonance Extended Cardiac Torso (MRXCAT) framework [13]. MRXCAT generates synthetic CMR images from XCAT digital phantoms [14], which provide detailed anatomical and physiological models of the thorax and abdomen, recreating realistic cardiac function. While phantoms serve as a useful means of augmenting and potentially replacing in-vivo datasets [15] with controlled, repeatable, and fully characterized simulations, a common drawback is that they often lack the complexity and realistic texture of actual MRI scans, making them insufficient when used in isolation.

1.4 Generative Models for CMR Image Synthesis

Generative AI presents a promising solution to the data scarcity problem [16] by synthesizing realistic images that improve the generalization of DL models in segmentation tasks. By approximating the underlying distribution of in-vivo data, generative models can be used as a tool for style-transfer based on tissue-specific characteristics found in the training set.

Preserving anatomical integrity in medical image synthesis is crucial for achieving realistic and coherent images. To ensure a high degree of fidelity within the ROI, leveraging semantic information from expert annotation labels provides models with a deeper understanding of the contextual information in the data. Meanwhile, semantic maps are often unavailable or only cover specific organs (i.e. heart) [17], leading to spurious and hallucinated generation of surrounding organs in the thorax and abdomen (i.e. lung, liver). Previous medical image synthesis studies predominantly employ the Spatially-Adaptive (SPADE) [18] framework, which translates a semantic map to an image using Generative Adversarial Networks (GANs) [19]. However, GAN-based generation often lacks diversity in sampling, struggles with mode coverage, and is unstable in the learning process, resulting in degraded quality [20, 21]. Recent attention has turned to diffusion-based generators, which have demonstrated superior performance to GANs in terms of generation quality, offering flexible approaches with strong mode coverage [22–24]. Diffusion models have gained prominence in recent literature, with significant efforts to integrate them into

novel medical imaging solutions [25], though their widespread application is hindered by computational strains.

1.5 Project Outline

In this work, we introduce a novel method for CMR image synthesis. Our focus is on generating realistic CMR images using style-transfer techniques to address the issue of data scarcity in medical imaging. Ultimately, this approach aims to improve the performance and robustness of learning-based models in automated segmentation tasks. We highlight the following contributions:

1. We propose an innovative framework based on emerging state-of-the-art, diffusion-based generative models. While traditional diffusion-based approaches are limited by the high computational demands of sequential Markov Chain sampling and direct image generation in image space [22], we sidestep this challenge by employing efficient Latent Diffusion Models (LDMs) [24] with Denoising Diffusion Implicit Models (DDIMs) sampling [26]. Our method operates within an auto-encoded (AE) compressed image space, enabling our models to run on a single GPU without any computational overhead. This ensures that our work is easily reproducible.

2. Built upon the original LDM implementation, we introduce several changes to the existing framework. We reworked the conditioning mechanism with a more sophisticated encoder, where features are concatenated at a similar abstract level, thereby improving the image generation process. Additionally, for improved semantic synthesis, we incorporate the SPADE normalization technique into the decoder process of the U-Net [27, 28] in a multi-resolution manner. SPADE has been extensively used in GAN-based models, but its application in diffusion-based models remains largely unexplored.

3. To address the challenge of anatomical compliance, we use edge information to delineate anatomical structures and condition the full field of view (FoV). Edge information is a highly underexplored topic in current literature and has only been applied to a small extent to guide image generation with GANs [29, 30]. To our knowledge, its use has not been explored in diffusion models, which promise superior quality compared to GANs. We condition our LDM-based generative models

using either edge maps alone or paired edges with anatomical masks for the SPADE-LDM variants, using the latter as dual anatomy conditioning to complement partial semantics.

4. We leverage our generative models for various automated segmentation tasks using style-transfer techniques. With our edge-conditioned generators, we perform image-to-image translation, delivering accurate segmentation without relying on models trained on extensive multi-sequence datasets.

5. We synthesize in-silico CMR images with the MRXCAT framework and address their shortcomings by transferring characteristic in-vivo tissue textures. Following the work from [31], we employ a first-stage segmentation model, initially trained on XCAT phantoms, to extend our in-vivo data annotations to the lung region and mitigate spurious hallucinated areas. Our approach refines in-silico images to closely resemble realistic CMRI, providing a reliable solution for data augmentation in training segmentation models. While some studies have attempted one-to-one tissue characterization by matching designated in-vivo semantic labels [17], they often fall short due to the low fidelity of remaining organs lacking annotations.

Chapter 2

Theoretical Background

2.1 Magnetic Resonance Imaging Physics

The underlying physics principles of MRI are based on the interaction between an external magnetic field and the intrinsic angular momentum, or spin (s), of a nucleus. Nuclei with an odd atomic number or odd atomic mass have unpaired spins, resulting in a magnetic moment ($\boldsymbol{\mu}$) that interacts with applied magnetic fields. Human tissues are primarily composed of water and fat, both abundant in hydrogen. Each hydrogen atom has a nucleus with a single proton (^{1}H), which serves as the primary source of MRI signals in medical imaging. While each proton has a spin of equal magnitude (s = $^1/_2$), their random orientations cause the net magnetization vector (**M**) to be zero, as the magnetic moments from the individual protons cancel each other out. When a tissue is placed within a strong magnetic field $\mathbf{B}_0 = B_0\hat{z}$ generated by an MRI system, the torque induced by this interaction causes the hydrogen nuclei to precess around $\mathbf{B}_0$ at a frequency ω. The precession rate, known as the Larmor frequency (ω_L), is proportional to the strength of the magnetic field, as described by the Larmor equation:

$$\omega_L = \gamma B_0, \tag{2.1}$$

where γ is the gyromagnetic ratio of the nucleus. This interaction also leads to two distinct energy states for the spins: parallel (lower energy) and anti-parallel (higher energy) directions relative to the field $\mathbf{B}_0$, as depicted in Figure 2.1. The energy gap ΔU between these two states is described by the Zeeman effect:

$$\Delta U = \gamma \hbar B_0, \tag{2.2}$$

where $\hbar$ is the reduced Planck constant. The parallel state, where the nuclei's magnetic moments align with the magnetic field B_0, represents a lower energy configuration since it minimizes the system's potential energy. Therefore, according to thermodynamic principles and the Boltzmann distribution [32], hydrogen nuclei are slightly more likely to occupy these lower energy states, resulting in a non-zero longitudinal magnetization vector **M** (Figure 2.1.b).

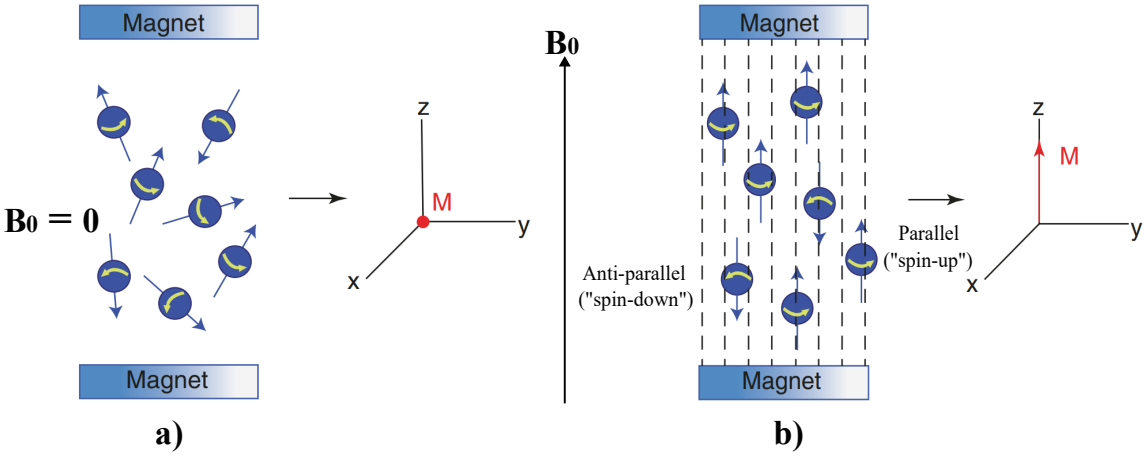

FIGURE 2.1: **a)** In the absence of an external magnetic field, spins are randomly oriented. **b)** When subjected to a strong magnetic field $\mathbf{B}_0 = B_0\hat{z}$, the hydrogen nuclei will align in one of two energy states: parallel or anti-parallel relative to $\mathbf{B}_0$. The parallel energy state is energetically more favorable, resulting in a majority of the spins occupying this configuration and generating a non-zero magnetization vector (**M**). Image adapted from [33].

The magnitude of $|\mathbf{M}|$ is exceedingly small compared to the field $\mathbf{B}_0$ as the spins precess in an incoherent, out-of-phase manner, rendering $|\mathbf{M}|$ undetectable. Instead, MRI signal detection relies on energy transfer principles described by NMR in the radiofrequency (RF) range of the electromagnetic spectrum. When the transmitting coils emit RF pulses, denoted by the field $\mathbf{B}_1$, oscillating at the Larmor frequency ω_L, protons begin to precess in-phase and absorb energy, forcing them to transition to the higher energy state through resonance. The $\mathbf{B}_1$ field is applied perpendicular to the main magnetic field $\mathbf{B}_0$, causing the net magnetization vector **M** to rotate out of equilibrium. This rotation shifts **M** from the longitudinal plane to the transverse xy-plane, decomposing it into two orthogonal vectors: a reduced longitudinal component $\mathbf{M_z}$ and an increased transverse component $\mathbf{M_{xy}}$ as the protons reorient towards the transverse plane. The duration and amplitude of the RF pulse determine the flip angle (α), which is the angle by which **M** is

rotated from the z-axis. This angle can be approximated by

$$\alpha = \gamma\tau \left|\mathbf{B}_1\right|, \tag{2.3}$$

where τ and $|\mathbf{B}_1|$ represent the duration and amplitude of the RF pulse, respectively. A

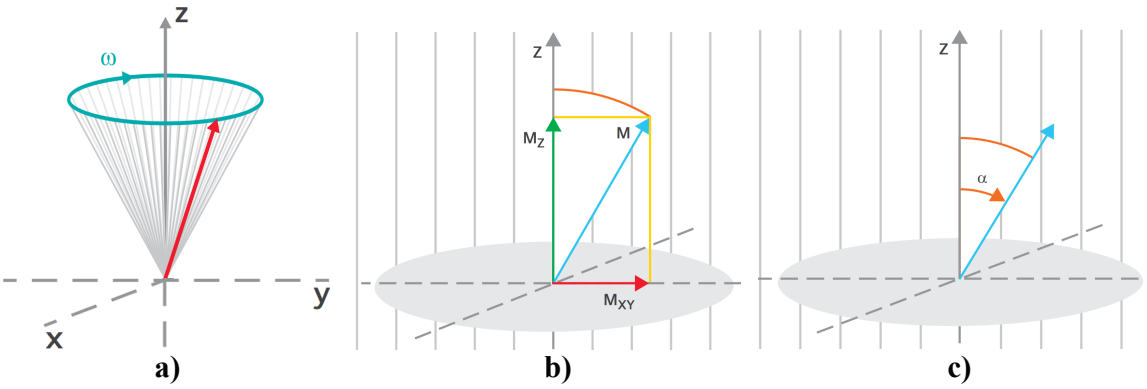

FIGURE 2.2: **a)** Hydrogen proton precessing around $\mathbf{B}_0 = B_0\hat{z}$. **b)** The two orthogonal components of $\mathbf{M}$: the longitudinal component $\mathbf{M_z}$ and the transverse component $\mathbf{M_{xy}}$. **c)** When a strong magnetic field $\mathbf{B}_1$ is applied perpendicular to $\mathbf{B}_0$, $\mathbf{M}$ is rotated out of equilibrium by an angle α towards the transverse xy-plane. Image adapted from [34].

90° pulse completely aligns $\mathbf{M}$ with the transverse plane.

When the $\mathbf{B}_1$ field is turned off, $\mathbf{M}$ realigns with the longitudinal axis as the high-energy ^{1}H protons dissipate their absorbed RF energy and return to their low-energy state. This process, known as relaxation, consists of the continuous decay of the transverse component $\mathbf{M_{xy}}$ until it fully ceases to exist. During relaxation, the changing magnetic field generated by the precessing $\mathbf{M_{xy}}$ induces a voltage in RF receiver coils, placed perpendicular to the transverse plane, known as Free Induction Decay, based on Faraday's law. The induced voltage constitutes the MR signal that is measured by the MRI system to construct an image. Relaxation manifests in two distinct ways: spin-lattice or T_1 relaxation, which involves the recovery of the longitudinal magnetization component $\mathbf{M_z}$; and transverse, spin-spin or T_2 relaxation, which involves the decay of the transverse magnetization component $\mathbf{M_{xy}}$. The relative impact of T_1 and T_2 relaxation times on signal intensity is influenced by the pulse sequence parameters, particularly the repetition time (T_R, i.e., the time between pulses) and the echo time (T_E, i.e., the time between the pulse and signal detection).

The spatial position of the nuclei contributing to the MRI signal is encoded by superimposing weaker magnetic field gradients onto the main magnetic field $\mathbf{B}_0$. The MRI system uses three pairs of gradient coils to create linear variations in the magnetic field

along the x, y, and z axes, known as G_x, G_y, and G_z. Specifically, G_z is used for slice selection, while G_x is applied for frequency encoding to determine the position along the x-axis, and G_y is applied for phase shift encoding to determine the position along the y-axis. These gradients can be combined to form any desired gradient field, enabling MRI to generate images in any orientation. Variations in the magnetic field result in distinct ω_L for every proton at each location, mapping their exact positions during signal detection, as described by:

$$\omega_L(x, y, z) = \gamma(B_0 + G_x \cdot x + G_y \cdot y + + G_z \cdot z). \tag{2.4}$$

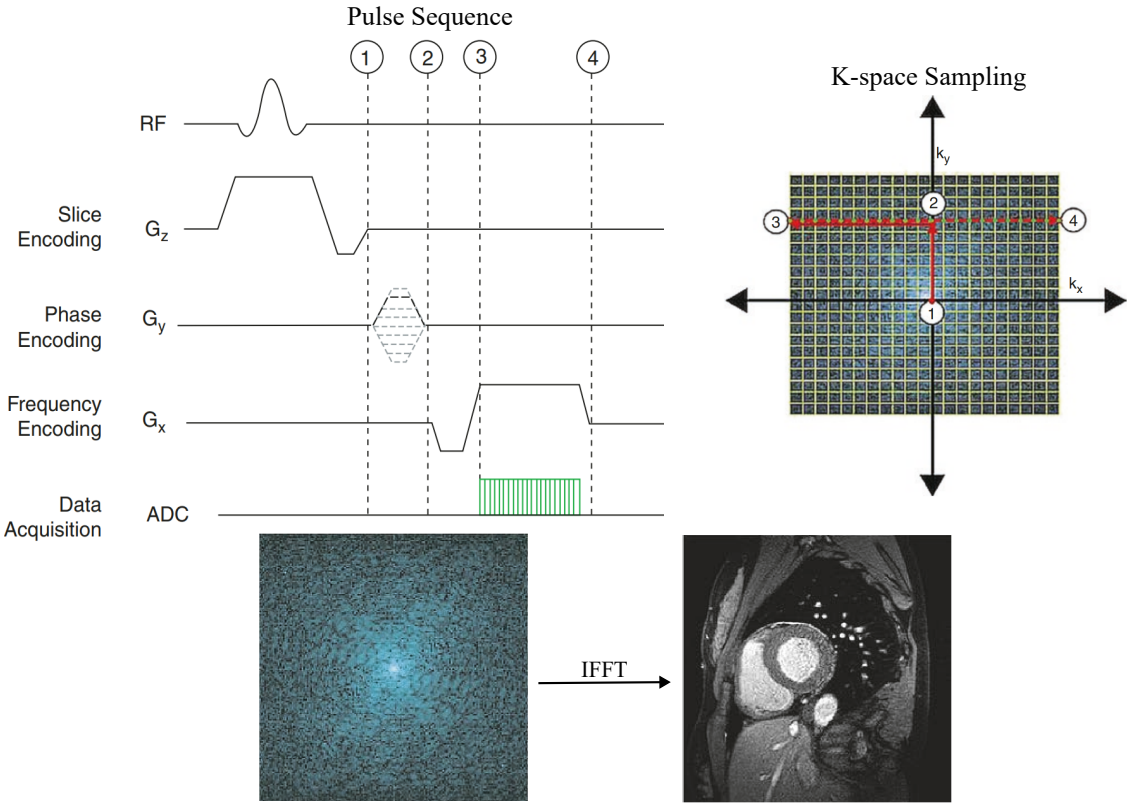

FIGURE 2.3: **Top:** Spatial encoding with superimposed magnetic field gradients. Step 1 involves applying G_z for slice selection along the z-axis. Step 2 is phase encoding, where G_y is used to determine the position along the y-axis, depending on the strength and duration of G_y. In Step 3, G_x is applied for frequency encoding to determine the position along the x-axis. Finally, in Step 4, all k_x points at the encoded k_y are acquired as a k-space line. **Bottom:** Most signal energy is concentrated within the lower spatial frequencies in the center of k-space. Once k-space is filled, it is transformed to the image space domain through the Inverse Fast Fourier Transform (IFFT), an efficient tool for IFT. Image adapted from [35]

During a scan, each encoded signal is conceptually equivalent to performing a Fourier transform, as these signals exist in the frequency domain, commonly known as k-space in MRI, representing image's spatial frequencies. The time-varying signals are digitized and

stored in a *k*-space matrix, with each measurement corresponding to a point on a Cartesian grid. This process is repeated until the *k*-space grid is filled. Low spatial frequency data fill the center of *k*-space, providing information about general shapes and contrast, while high spatial frequency data are stored in the outer regions, depicting fine details. Typically, in MR images, most signal energy is concentrated in the lower spatial frequencies since the imaging scene is dominated by large, high-contrast objects. Once *k*-space is fully sampled, the computer system applies an inverse Fourier transform (IFT) to decode the spatial positions based on the variations in frequency and phase. The resulting image space reveals the spatial anatomical relationships between tissues, reflecting their magnetization behavior based on the signal intensity readouts. The contrast within each voxel is influenced by factors such as proton density and their respective T_1 and/or T_2 relaxation time constants.

Sequential *k*-space sampling significantly slows down the data collection process and is the primary reason why MRI is slower compared to other imaging modalities. This also limits a patient's ability to remain still throughout the procedure [36], potentially causing image artifacts. Moreover, MRI digitizes analog RF signals, and thus adheres to the Nyquist-Shannon theorem, requiring enough *k*-space coverage to avoid aliasing artifacts and ideal spatial resolution. Higher spatial resolutions demand a larger number of samples and a smaller voxel size, which in turn requires longer scanning times and leads to a lower SNR and temporal resolution. These trade-offs must be appropriately selected and balanced for each specific imaging task.

2.1.1 Longitudinal (Spin-Lattice) Relaxation

Longitudinal or T_1 relaxation, also known as spin-lattice relaxation, refers to the recovery of the longitudinal magnetization component $\mathbf{M_z}$ towards its original equilibrium value. This recovery depends on the net transfer of thermal energy between the spins and their surrounding atomic environment, referred to as the "lattice". T_1 is defined as the spin-lattice relaxation time required for $\mathbf{M_z}$ to recover 63% of its equilibrium value. This recovery follows an exponential behavior, which can be derived from Bloch equation [9]:

$$M_z(t) = M_0 \left(1 - e^{-t/T_1}\right), \tag{2.5}$$

where t is the time following the excitation pulse and M_0 is the magnetization at equilibrium.

Energy is transferred to the lattice through molecular motion at the precessing frequency, characterized by the tissue-specific exponential time constant T_1. Efficient energy transfer leads to shorter T_1 relaxation times, whereas less efficient mechanisms result in longer periods. Energy transfer is most efficient when the frequencies of molecular motion align with the Larmor frequency. Large molecules, such as those found in fats with long and complex hydrocarbon chains, exhibit slower motions that can more easily match ω_L, resulting in shorter T_1 relaxation times. In contrast, free water molecules are highly mobile and do not align well with ω_L, leading to longer T_1 relaxation periods.

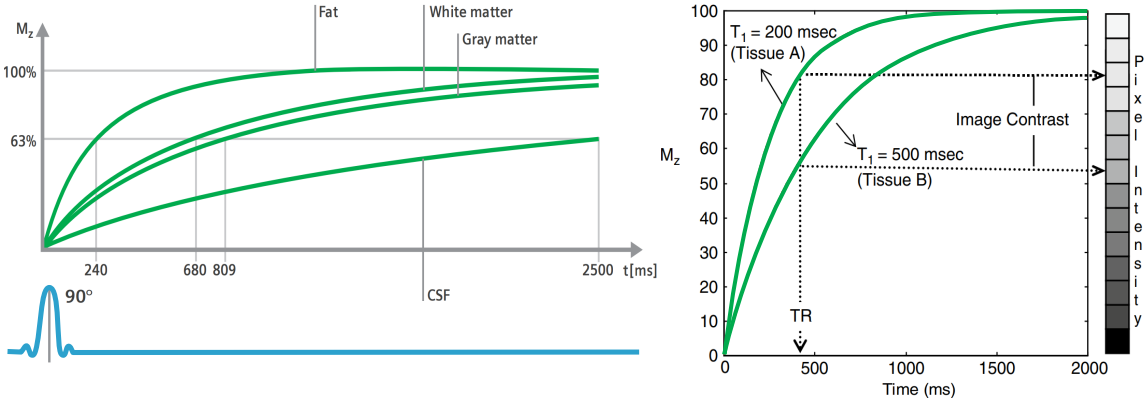

FIGURE 2.4: **Left**: Longitudinal magnetization recovery curves following a 90° excitation pulse. Each curve corresponds to a specific tissue type, with their respective T_1 relaxation constants denoted on the horizontal axis. **Right**: Grayscale contrast differences in T_1-weighted imaging due to variations in T_1 relaxation times between tissues. Image adapted from [34] and [35].

The purpose of T_1-weighted preparation is to evaluate potential T_1 differences within and between tissues. In a pulse sequence designed to generate T_1-weighted images, tissues with shorter T_1 values will induce a stronger signal than those with longer T_1. A shorter T_R results in less time for magnetization recovery between excitations, causing regions with shorter T_1 to appear relatively brighter. The recovery of $\mathbf{M_z}$ back to equilibrium will vary based on the T_1 distribution in specific tissues. For instance, the T_1 of myocardial tissue can be altered by the presence of heterogeneous components such as healthy myocardium, fat, fibrotic tissue, or excess water from edema.

T_1-weighted imaging is only qualitative and does not provide exact T_1 values. Alternatively, T_1-mapping quantifies signal intensity using multiple readouts at distinct inversion times (i.e., intervals within the recovery period after a 180° inversion pulse) or saturation recovery times (i.e., intervals within the recovery period after a 90° saturation

pulse), depending on the imaging technique. These intensities are then ordered and fitted voxel-wise to the exponential recovery model in Equation 2.5, rendering T_1-mapping more suitable for identifying and characterizing subtle changes in tissue composition.

2.1.2 Late Gadolinium Enhancement

Paramagnetic MRI contrast agents enhance image quality by shortening tissue T_1 and T_2 relaxation times through increased local magnetic field strength and field inhomogeneities. In CMRI, gadolinium-based contrast agents (GBCAs) are widely used due to their strong paramagnetic properties. Gadolinium (Gd^{3+}) ions contain seven unpaired electrons, which induce strong magnetic moments in the surrounding protons. This significantly shortens T_1 relaxation times and thus increases signal intensities in areas where the contrast agent is administered. Since Gd^{3+} is highly toxic, it is typically chelated with organic ligands to form GBCAs, which prevent toxic metals from interacting with the body and make it safe for clinical imaging.

Late-phase imaging after GBCA administration ($\sim$ 10–30 minutes) can highlight areas of abnormal contrast agent accumulation in expanded extracellular spaces. This accumulation is typically seen in damaged tissue, which has increased interstitial space due to cell death and subsequent fibrosis (i.e., excess fibrous connective tissue leading to scarring) or edema. The enlarged extracellular space leads to increased accumulation of GBCA in these regions, causing fibrotic scars to appear substantially brighter than healthy heart tissue. This technique is extensively used in T_1-weighted LGE imaging to visualize and quantify infarcted myocardium.

2.1.3 Transverse (Spin – Spin) Relaxation

As outlined in Section 2.1, following the application of RF pulses, protons become coherent and precess in-phase. During relaxation, small variations in the magnetic field cause the spins to precess at different rates (Equation 2.2), driving them out of phase. This dephasing is partly due to ”spin-spin” interactions, where the magnetic fields of nearby protons alter the local magnetic field experienced by each proton. This effect gradually disrupts the coherence of protons along the transverse xy-plane, causing $\mathbf{M_{xy}}$ to decrease over time. The decay of $\mathbf{M_{xy}}$ is characterized by a T_2 relaxation time, also known as spin-spin relaxation. According to Bloch equation [9], once the excitation RF pulse is turned

off, $\mathbf{M_{xy}}$ decays to zero in an exponential process described by:

$$M_{xy}(t) = M_0 e^{-t/T_2}, \tag{2.6}$$

where T_2 is the time constant for $\mathbf{M_{xy}}$ to decay to 37% of its initial value. This process is inherently faster than longitudinal T_1 relaxation, as the signal decays much sooner than the system returns to equilibrium. Spin-spin relaxation, as seen with spin-lattice relaxation, is inefficient in highly mobile protons and exhibits longer T_2 relaxation times in free water and tissues with high water content. Regions characterized by longer T_2 values, such as slow-moving fluids or edematous tissues, appear brighter because they decay more slowly, as illustrated in Figure 2.5. In contrast, fast-flowing fluids appear darker due to phase shifts of moving spins, leading to rapid signal loss.

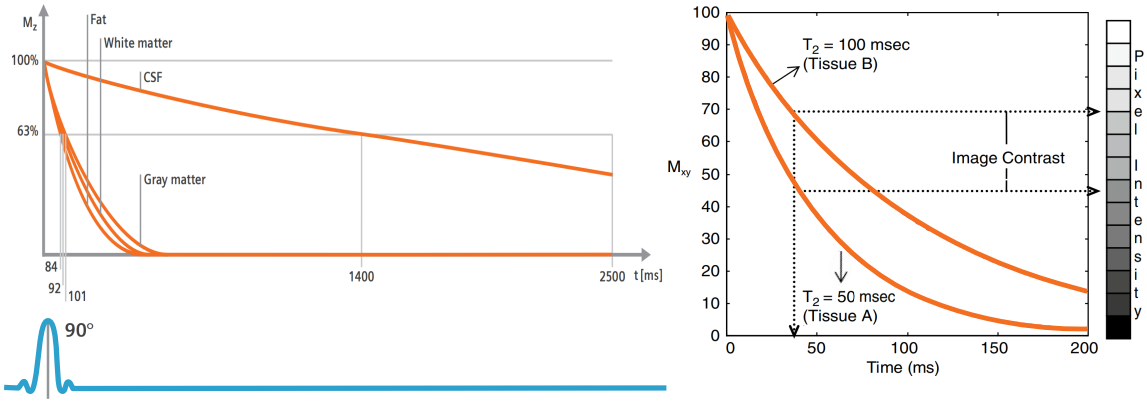

FIGURE 2.5: **Left**: Transverse magnetization recovery curves following a 90° excitation pulse. Each curve corresponds to a specific tissue type, with their respective T_2 relaxation constants at 37% of their equilibrium value denoted on the horizontal axis. **Right**: Grayscale contrast differences in T_2-weighted imaging due to variations in T_2 relaxation times between tissues. Image adapted from [34] and [35].

The dependence on transverse T_2 relaxation is more complex than that of longitudinal T_1. While T_1 is characterized by a time constant associated with the energy exchange between the spin system and its surroundings, T_2 is influenced by both intrinsic spin-spin interactions and static inhomogeneities in the local magnetic field. These inhomogeneities result from variations in the applied field or differences in magnetic susceptibility between adjacent tissues, causing local spin dephasing. The "off-resonance" spins contribute to an extrinsic relaxation, characterized by a shorter apparent time constant, T_2^*, which accentuates spin incoherence and further decreases $\mathbf{M_{xy}}$. Nevertheless, the effects contributing to apparent T_2^* relaxation are reversible and can be controlled through appropriate imaging schemes. T_2-weighted imaging readouts typically isolate intrinsic

T_2 relaxations using spin-echo sequences. These sequences begin with an initial excitation 90° RF pulse, followed by one or more 180° refocusing pulses. The 180° inversion pulse mitigates the dephasing effects from magnetic field inhomogeneities by reversing the direction of phase accumulation. Dephased spins that had moved ahead start to lag behind, while the slower precessing spins that had lagged behind start to catch up. This process restores coherence in the transverse magnetization, producing a strong "echo" signal, with its decay primarily governed by the intrinsic T_2 relaxation time. Note that T_2^*-weighted imaging also holds value for certain diagnostic tasks, particularly for detecting iron overload in cases of hemochromatosis, given its high sensitivity to magnetic field inhomogeneities caused by iron deposits [37, 38].

Similar to T_1-mapping, T_2-mapping quantifies T_2 tissue relaxation times on a voxel-by-voxel basis. This process involves a series of spin echo sequences with 180° RF refocusing pulses, yielding distinct T_E values. The signal intensity readouts at each T_E are then ordered and fitted to the model in Equation 2.6 to determine T_2.

2.1.4 Balanced Steady-State Free Precession

To address the limitations posed by the inevitable loss of signal intensity caused by spins' incoherence and the consequent decrease in M_{xy}, a widely used imaging scheme in CMRI is the Balanced Steady-State Free Precession (bSSFP) sequence. This sequence applies RF excitation pulses with a very short T_R, combined with precise gradient refocusing to maintain steady-state magnetization and minimize off-resonance effects and artifacts.

In bSSFP, symmetric gradient waveforms create a zero net gradient over each T_R period by pairing each gradient pulse with an opposing pulse within the same T_R. This symmetry preserves phase coherence and eliminates cumulative dephasing, resulting in strong and consistent transverse magnetization. The short T_R, compared to T_1 and T_2 relaxation times, minimizes signal decay. Transverse magnetization from the previous pulse adds up to the magnetization generated by the subsequent pulse. This sequence of RF pulses continues until a steady state is reached in both the M_z and M_{xy} components, maximizing signal intensity through constructive interference. The magnetization in bSSFP can be described by the following equation [39]:

$$M_{ss} = \rho_0 \frac{\sin\alpha}{1 + \cos\alpha + (1 - \cos\alpha)\left(\frac{T_1}{T_2}\right)} \tag{2.7}$$

where ρ_0 is the proton density. Equation 2.7 is valid under two assumptions: $T_R \ll T_1, T_2$, which is the case for most biological tissues at 1.5 T with a T_R of $\sim$ 2–3 ms [40], and $T_E = T_R/2$, where the readout gradient is centered around the echo for optimal signal strength.

The steady-state condition in bSSFP offers high SNR and excellent contrast for specific tissues, making it particularly valuable for CMRI, especially for tissues with high fluid content and elevated T_2/T_1 ratios. The T_2/T_1 ratio of blood is approximately three times higher than that of the myocardium [41], causing blood to appear significantly brighter than myocardium in images using the bSSFP scheme. This high-contrast nature of bSSFP allows for precise delineation of the endocardial border between the myocardium and blood, making it the ideal choice for cardiac function assessment.

2.1.5 Cine Imaging

CMRI is able to synchronize image data acquisition with the cardiac cycle, producing high-quality "cine" movies that capture the heart's activity through its distinct phases. Sample acquisitions are timed to an electrocardiogram (ECG)'s R-R interval (i.e., the period between consecutive R waves), which indicate ventricular contractions (systole) followed by relaxation (diastole). This timing targets specific time frames to highlight cardiac phase-dependent features and minimize imaging artifacts from cardiac motion. Data acquisition occurs with a relatively short temporal resolution ($\sim$ 20–50 ms) per frame, depending on the subject's heart rate. This leads to the reconstruction of multiple image frames, each representing different phases of the cardiac cycle, as shown in Figure 2.6.

Cine imaging requires very short T_R values and thus is typically performed with an bSSFP scheme. For each cine frame, k-space data is obtained by repeatedly sampling over consecutive cycles. The resulting cine sequence allows for detailed assessment of heart function, wall motion, and blood flow.

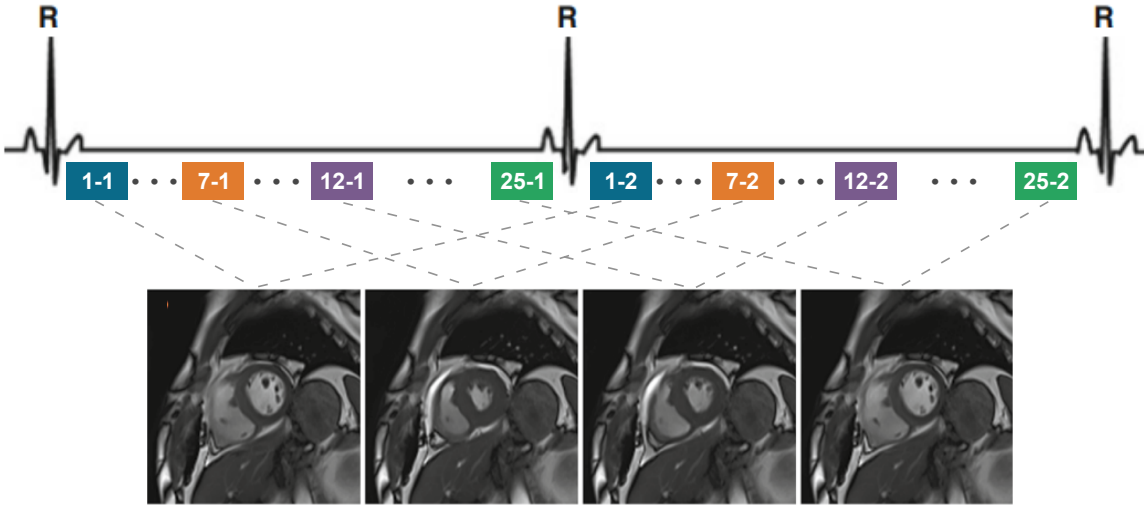

FIGURE 2.6: Cine 2D bSSFP imaging: 25 frames are collected during each R-R interval and over multiple cardiac cycles. Only a fraction of the k-space lines is collected in each interval. In this example, the complete k-space data for each of the cine frames is collected from only two heartbeats for demonstration purposes. Image adapted from [35].

2.2 Deep Generative Models

The principle behind image generation is to extract the underlying probability distribution of a given dataset to synthesize new, similar samples on demand. This involves modeling the true probability distribution $p(x)$ of a set of independent and identically distributed (i.i.d.) samples $\{x_i\}_{i=1}^{n}$ over the image space, defined in $\mathbb{R}^{H\times W\times C}$. Here, H and W represent the height and width of the images, respectively, while C denotes the number of channels.

Generative models revolve around designing a generator $\mathcal{G}$ that transforms random samples drawn from a tractable prior distribution $z \sim p(z)$ into images that closely resemble those from the training set. The vector $z \in \mathbb{R}^d$, commonly referred to as the latent variable, is generally unknown, and $\mathbb{R}^d$ is called the latent space. Formally, this process can be expressed as $\mathcal{G} : \mathbb{R}^d \to \mathbb{R}^{H\times W\times C}$, such that $\mathcal{G}(z) \approx x$, where the observed data x, generated by an associated latent variable z, is described by the joint distribution $p(x, z)$. The latent space dimension d differs from the image space, as natural images are inherently high-dimensional, a problem known as the curse of dimensionality. For instance, a single RGB image with 256×256 px contains $196,608$ dimensions. For this reason, using a latent space $\mathbb{R}^d$ of the same dimensionality as the image space is practically infeasible. Natural images often lie on a lower-dimensional manifold within the high-dimensional

space, where the essential information can be captured in a much lower-dimensional latent space, though the exact dimension of this hidden manifold is unknown. The latent dimension d still needs to be high enough to capture the complex structures and features of image data and disentangle factors of variation.

To obtain the marginal distribution of the sample x, we either can integrate out the latent variable z:

$$p(x) = \int p(x,z)dz = \int p(x|z)p(z)dz, \tag{2.8}$$

or make use of the chain rule of probability:

$$p(x) = \frac{p(x,z)}{p(z|x)}. \tag{2.9}$$

However, both approaches present practical limitations. Integrating over all latent variables z in Equation 2.8 is highly complex, as the number of evaluation points grows exponentially with the number of dimensions, making direct numerical integration infeasible. On the other hand, Equation 2.9 involves an unknown posterior $p(z|x)$ with no closed form, requiring access to a ground truth latent encoder that is typically unavailable. Instead, most practical implementations rely on approximating these integrals and distributions.

Traditional methods like Markov Chain Monte Carlo (MCMC) and variational inference were often used to approximate the posterior $p(z|x)$, but they would yield significant drawbacks. MCMC is computationally prohibitive due to its high rejection rates and the large number of iterations required, especially in high-dimensional spaces. In contrast, variational inference is more efficient and scalable but tends to oversimplify the posterior with a tractable surrogate distribution. Over the years, generative models have transitioned to deep neural networks, leveraging their universal approximation properties to estimate probability distributions [42]. This principle underlies deep generative models, where $\mathcal{G}$ is represented as a feed-forward deep neural network with numerous hidden layers, parameterized by weights $\theta \in \mathbf{R}^{N_\theta}$. Exact inference is often framed as an optimization problem, relying on algorithms to approximate 2.8 or 2.9 through an objective function that ensures $\mathcal{G}_\theta(z;\theta)$ generates samples matching the statistical properties of the training set. By employing DL techniques, these models can extract features at multiple hierarchical levels of abstraction, identifying intricate patterns and dependencies within the data to accurately approximate functions in high-dimensional spaces.

The performance of deep generative models is significantly influenced by several factors, including network architecture, choice of training objectives, regularization methods, training algorithms, and hyperparameters. Selecting the network architecture is a complex task that involves determining the number of layers and their specific functions. This process is usually guided by domain-specific inductive biases and supported by empirical results.

2.2.1 Historical Trends

In the mid-2000s, early deep generative models such as Deep Belief Networks (DBNs) [43] were developed, employing stacked Restricted Boltzmann Machines (RBMs) [44]. Each RBM layer refined features from the previous one, learning hierarchical data representations. Training these networks was memory-intensive and required inefficient, sophisticated algorithms like Contrastive Divergence [45], demanding significant computational resources. These models struggled with vanishing gradients and scalability issues, often leading to poor-quality image generation. DBNs fell out of use in the early 2010s with the rise of more efficient and scalable architectures employing Convolutional Neural Networks (CNNs) [46], which leverage modern hardware and are trained end-to-end using backpropagation [47] (BP).

In 2013, Variational Autoencoders (VAEs) [48] introduced a probabilistic model that optimizes the Evidence Lower Bound (ELBO) using differentiable stochastic sampling, enabling efficient gradient-based optimization through BP. The parameters of a surrogate Gaussian function are fine-tuned using deep learning techniques to closely approximate the true posterior distribution, providing more flexibility and robustness compared to traditional variational inference. VAEs are known as a type of AE because they compress input data to an intermediate bottleneck and then learn to reconstruct it, producing compact and meaningful representations. However, VAEs often produce blurry images due to the trade-off between reconstruction accuracy and latent space regularization, which favors smooth reconstructions over sharp details.

In 2014, the introduction of Generative Adversarial Networks (GANs) [19] left a significant mark on the field of generative modeling, remaining one of the most popular and enduring deep generative models to date. As Turing Award winner Yann LeCun once stated in a Quora post in May 2016, "GANs are the most interesting idea in the last

ten years in machine learning" [49]. GANs improved upon previous models by introducing an adversarial framework that efficiently produces higher quality images. Nevertheless, GANs' adversarial training is often unstable and prone to mode collapse, resulting in a limited diversity of samples. Despite these challenges, the early success and widespread application of GANs over nearly a decade have led to significant engineering efforts aimed at improving their capabilities and overcoming these issues. In 2016, autoregressive models (e.g., PixelRNN [50]) and flow-based models [51] showed some signs of success but were hindered by high computational load and inefficient training. These limitations made them impractical for large-scale, high-quality image synthesis, making GANs the preferred choice.

Since 2020, diffusion-based models [22, 52] have shown remarkable performance and gained significant attention for outperforming GANs in numerous tasks [23], setting new state-of-the-art benchmarks in image generation. Diffusion models are inspired by non-equilibrium thermodynamics [52] and employ a Markovian process that gradually adds random noise to image samples through a series of diffusion steps, ultimately reaching an isotropic Gaussian noise state. Neural networks are then trained to reverse this diffusion process, step by step, refining the noisy images to reconstruct the original, noise-free images. This, in turn, results in highly flexible and tractable generative models that can learn complex data distributions with strong mode coverage. These models are the driving force behind the significant growth in cutting-edge image synthesis applications. Some well-known examples include DALL-E [53], Stable Diffusion [54], and, more recently, Sora [55], which generate fine-grained high-resolution images and video synthesis from text prompts.

2.2.2 Generative Adversarial Networks

The development of Generative Adversarial Networks (GANs) [19] was inspired by game theory and involves a minimax objective, where players aim to minimize the possible maximum loss [56]. GANs consist of two distinct deep neural networks working simultaneously: a generator $\mathcal{G}$ and a discriminator $\mathcal{D}$. The purpose of $\mathcal{G}$ is to produce realistic outputs to deceive $\mathcal{D}$, while $\mathcal{D}$'s role is to distinguish between fakes from $\mathcal{G}$ and real examples, as illustrated in Figure 2.7.

During training, $\mathcal{G}$ progressively improves the quality of its samples, approaching the underlying true data distribution. Meanwhile, $\mathcal{D}$ serves as a proxy by refining its ability to

identify synthetic samples from $\mathcal{G}$. This adversarial framework is highly efficient, avoiding explicit likelihood computations or approximate inference by relying on $\mathcal{D}$'s feedback to $\mathcal{G}$. The optimization ends at a saddle point, representing a minimum for $\mathcal{G}$ and a maximum for $\mathcal{D}$. This condition corresponds to a Nash equilibrium [57], a stable state where neither network can improve without changes from the other. At this stage, $\mathcal{G}$ is considered to have captured the underlying distribution of the training set, enabling GANs to generate images almost statistically indistinguishable from real ones.

The cost functions for $\mathcal{G}$ and $\mathcal{D}$ are formally derived from binary cross-entropy loss [58] and can be expressed as follows:

$$L_{\mathcal{G}} := -\mathbb{E}_{z\sim p_z}\left[\log(1-\mathcal{D}(\mathcal{G}(z)))\right] \tag{2.10}$$

$$L_{\mathcal{D}} := -\mathbb{E}_{x\sim p_{\text{data}}}\left[\log \mathcal{D}(x)\right] - \underbrace{\mathbb{E}_{x\sim p_{\mathcal{G}}(x)}\left[\log(1-\mathcal{D}(x))\right]}_{\mathbb{E}_{z\sim p_z}\left[\log(1-\mathcal{D}(\mathcal{G}(z)))\right]}, \tag{2.11}$$

where $z \in \mathbb{R}^d$ is a random latent vector drawn from a standard normal distribution, $p_z \sim \mathcal{N}(0, \mathbf{I})$, and $x \in \mathbb{R}^{H\times W\times C}$ is a real example from the training set, such that $\mathcal{G} : \mathbb{R}^d \to \mathbb{R}^{H\times W\times C}$. $\mathcal{D}$ operates as a binary classifier to evaluate both real samples x and synthetic samples $\mathcal{G}(z)$, where the output is a single scalar indicating the probability that a given sample belongs to p_{data}. Formally, $\mathcal{D}$ is defined as $\mathcal{D} : \mathbb{R}^{H\times W\times C} \to [0, 1]$. An output of $\mathcal{D} = 1$ denotes that the example is real, whereas a value of $\mathcal{D} = 0$ denotes that the example is fake.

Following a minimax strategy, $\mathcal{D}$'s objective is to maximize the probability of correctly classifying both x and $\mathcal{G}(z)$. In contrast, $\mathcal{G}$ training goal is minimize $\log(1 - \mathcal{D}(\mathcal{G}(z)))$. This creates an adversarial dynamic where $\mathcal{D}$ pushes $\mathcal{D}(\mathcal{G}(z))$ closer to 0, while $\mathcal{G}$ pushes it closer to 1, establishing a zero-sum game. Over time, $\mathcal{G}$ becomes increasingly proficient at producing realistic counterfeits. Since $L_{\mathcal{G}}$ (Equation 2.10) is directly linked to $L_{\mathcal{D}}$ (Equation 2.11), the final minimax optimization problem, reformulated from 2.11, can be expressed by the value function $V(\theta_{\mathcal{G}}, \theta_{\mathcal{D}})$ given by:

$$\min_{\mathcal{G}} \max_{\mathcal{D}} V(\theta_{\mathcal{G}}, \theta_{\mathcal{D}}) := \mathbb{E}_{x\sim p_{\text{data}}}[\log \mathcal{D}(x)] + \mathbb{E}_{z\sim p_z}[\log(1-\mathcal{D}(\mathcal{G}(z)))]. \tag{2.12}$$

Derivations in the original paper [60] demonstrate that this objective is equivalent to minimizing the Jensen-Shannon divergence between the two distributions p_{data} and p_G

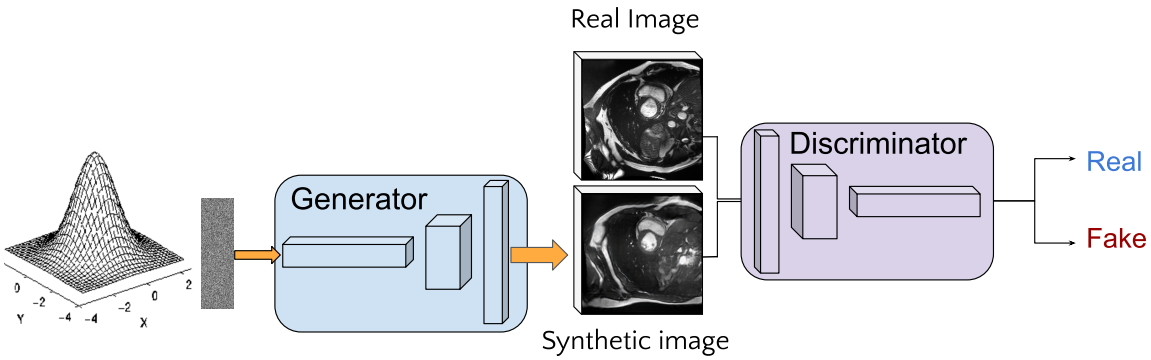

FIGURE 2.7: GANs consist of two neural networks: the generator $\mathcal{G}$, which transforms a random Gaussian vector into synthetic data, and the discriminator $\mathcal{D}$, which evaluates whether the data is real or fake. Image adapted from [59].

[61]. At the Nash equilibrium, $p_{\mathcal{G}} \approx p_{\text{data}}$, meaning $\mathcal{D}$ cannot differentiate between real and synthetic samples, yielding $\mathcal{D}(\mathcal{G}(z)) = 0.5$ for all inputs. This state is important from a game theory perspective, as it implies that neither player finds it advantageous to make changes to their strategy near this equilibrium.

Equation 2.10 is theoretically well-grounded, but it often performs poorly in practice because it may not provide sufficient gradients for $\mathcal{G}$ to learn effectively. In the early stages of training, $\mathcal{G}$ produces low-quality samples that $\mathcal{D}$ can confidently reject, as they differ significantly from the training set. As a result, $\log(1 - \mathcal{D}(\mathcal{G}(z)))$ saturates, causing $\mathcal{G}$'s gradients to vanish. A more reliable and heuristically driven approach involves training $\mathcal{G}$ to maximize $\log \mathcal{D}(\mathcal{G}(z))$ [62], providing stronger gradients early in the learning process. Although this adjustment means that GANs can no longer be framed as a minmax problem, the dynamics remain unchanged since maximizing the log probability is equivalent to minimizing the loss function in Equation 2.10.

GANs face other several challenges in their training process, with one of them being mode collapse. Mode collapse occurs when the synthetic samples become concentrated in a few or even a single mode, failing to capture the full data distribution. As $\mathcal{D}$ becomes highly adept at recognizing certain patterns, $\mathcal{G}$ starts generating convincing samples that easily deceive $\mathcal{D}$. This causes $\mathcal{G}$ to repeatedly produce similar samples because they consistently fool $\mathcal{D}$, neglecting other modes and resulting in a lack of diversity in p_G. Additionally, other issue lies in achieving stability and Nash equilibrium, as the optimization process in GANs differs from conventional methods. Traditional gradient descent navigates a fixed loss landscape to find a minimum, whereas in GANs, each optimization step transforms the landscape, creating a constantly evolving environment. This means gradient descent may never reach an equilibrium, potentially causing GANs to underfit,

as $V(\theta_G, \theta_D)$ might oscillate indefinitely instead of converging to a saddle point. To adequately mitigate these issues, training GANs requires employing several heuristic tricks and careful tuning of the model architecture and hyperparameters [62–64].

2.2.3 Principles of Non-Equilibrium Thermodynamics in Diffusion Models

The foundational concepts behind diffusion models are rooted in physics and probability theory, drawing inspiration from non-equilibrium thermodynamics [52]. Noise is incrementally added to the original samples in a sequential, quasi-static manner, formulated as a Markov chain, while neural networks are trained to reverse each step of the process. Based on the Kolmogorov equations [65], the forward and reverse diffusion processes can be described using the same mathematical framework, facilitating the formulation of algorithms that directly parameterize the reverse process.

The theoretical concept of sequential quasi-static diffusion steps is based on the Jarzynski equality, which relates the free energy difference between two states to the work done during non-equilibrium transitions. Both the Jarzynski equality and sequential Monte Carlo methods form the foundation of Annealed Importance Sampling (AIS) [66], which uses a Markov chain to transition between distributions. In thermodynamics, transitions between equilibrium distributions determine the change in free energy. By using infinitesimal steps, quasi-static transitions between states occur with minimal energy loss and entropy change. Similarly, in diffusion models, the noise schedule and the number of intermediate distributions significantly impact the efficiency and accuracy of sampling.

Reverse diffusion can be interpreted as a form of annealed Langevin dynamics [67, 68], which describes the trajectory of a particle towards low-temperature regions, influenced by both deterministic and stochastic effects. This process involves gradually reducing the entropy of the data distribution, which is linked to decreasing noise levels. In Langevin dynamics, deterministic updates are governed by the gradient of the potential function, guiding the system toward lower energy states or higher-probability regions. Meanwhile, the stochastic term introduces thermal fluctuations, allowing the system to escape potential energy wells, ensuring it explores the state space and reaches an equilibrium distribution. In a similar manner, at each step of the reverse diffusion process, the model applies a learned deterministic update to progressively reduce the noise introduced during the forward process, while simultaneously adding a small amount of noise to preserve the stochastic nature of the process. To the same extent, diffusion models leverage this noise

term to promote robust exploration of the complex data space at each denoising step, thus improving mode coverage and preventing the model from collapsing into local minima.

2.2.4 Denoising Diffusion Probabilistic Models

Leveraging the groundwork laid in [52] related to non-equilibrium thermodynamics, Denoising Diffusion Probabilistic Models (DDPM) [22, 69] have demonstrated the ability of diffusion-based models to excel in image generation tasks, producing diverse, high-quality samples.

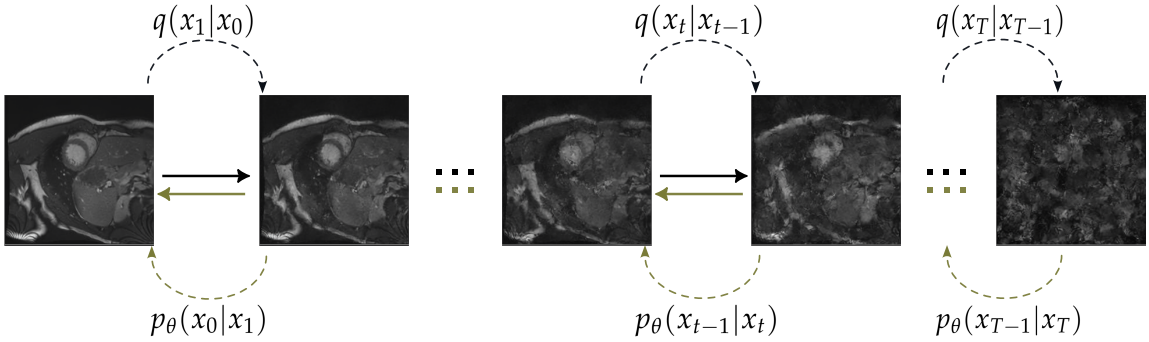

FIGURE 2.8: DDPM: The forward process gradually adds noise to the image, ultimately transforming it into a pure random noise state. The reverse process then learns to remove this noise by modeling the transition kernels from the forward diffusion to restore the original image.

DDPMs consist of two Markov chains: one for forward diffusion and the other for reverse diffusion. The forward chain gradually adds noise to a sample, ultimately transforming it into a state of pure Gaussian noise. In contrast, the reverse chain, parameterized by deep neural networks, learns the transition kernels required to reverse the noising process, reconstructing the original sample. By traversing this reverse chain and employing ancestral sampling, the model learns to transform samples drawn from a Gaussian distribution back to the modes of the underlying true distribution, effectively performing generative modeling. Formally, let x_0 be a sample drawn from the underlying true distribution $q(x_0)$. The forward Markov chain is defined as a sequence of random variables $x_1, x_2, \ldots, x_T$ governed by Gaussian transition kernels $q(x_t \mid x_{t-1})$, which progressively transform $q(x_0)$ into a Gaussian distribution. By applying the chain rule of probability and employing the Markov property, the joint distribution of $x_1, x_2, \ldots, x_T$ conditioned on x_0 can be factorized as follows:

$$q(x_1, x_2, \ldots, x_T|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}). \tag{2.13}$$

Gaussian noise is incrementally added to x_0 over T steps, resulting in a sequence of noisy samples. The step sizes are governed by a variance schedule, denoted as $\beta_t \in (0,1)_{t=1}^{T}$. As the time index t increases, the initial state x_0 gradually loses its distinct features. In this context, β_t serves as the upper bound of the variance, while $\tilde{\beta}_t$ represents the lower bound, expressed as:

$$\tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t. \tag{2.14}$$

By marginalizing the joint distribution in Equation 2.13, an analytical form for $q(x_t|x_0)$ is derived for all $t \in \{0,1,\ldots,T\}$:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}), \tag{2.15}$$

where x_t represents the random variable being modeled by the Gaussian distribution, $\alpha_t := 1-\beta_t$, $\bar{\alpha}_t := \prod_{i=0}^{t}\alpha_i$, and $\mathbf{I}$ represents the identity matrix. When T is large, x_T approaches an isotropic Gaussian distribution. To sample x_t at any arbitrary time step t in closed form, one can sample a Gaussian vector ϵ, where each element is independently drawn from a normal distribution $\epsilon \sim \mathcal{N}(0,\mathbf{I})$, and apply the reparameterization trick from [48] for differentiable stochastic sampling:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \tag{2.16}$$

where ϵ merges $t-1$ Gaussian distributions. Note that as $t \to T$, $\bar{\alpha}_t \approx 0$ and $q(x_T) \approx \mathcal{N}(x_T; 0, \mathbf{I})$. Therefore, if the exact reverse distribution is known, x_T can be sampled from $\mathcal{N}(0,\mathbf{I})$, and the process can be run in reverse to obtain a sample from $q(x_0)$, as illustrated in Figure 2.9. However, estimating $q(x_{t-1}|x_t)$ is infeasible because it requires integrating over all possible prior states $x_{t-1}, x_{t-2}, \ldots, x_0$ to obtain the marginal $q(x_t)$. Moreover, it would also require computing $q(x_0)$ by integrating over the entire data distribution, making it computationally prohibitive. Instead, one can approximate this process using a deep neural network to train a model p_θ that estimates these posteriors, with parameters θ adjusted so that the joint distribution of the reverse Markov chain closely approximates Equation 2.13 from forward diffusion:

$$p_\theta(x_{0:T}) = p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t). \tag{2.17}$$

This reverse process is characterized by a Gaussian prior distribution $p(x_T) = \mathcal{N}(x_T; 0, \mathbf{I})$, and parametric transition kernels $p_\theta(x_{t-1}|x_t)$ given by:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)), \tag{2.18}$$

where $\mu_\theta(x_t, t)$ is parameterized by the DL model, while $\Sigma_\theta(x_t, t) = \sigma_t^2 \mathbf{I}$ remains fixed as time dependent constants. Experiments in [22] suggest that attempting to learn the reverse process's diagonal variance, Σ_θ, results in unstable training and lower-quality samples compared to when variances are fixed. Additionally, ablation studies in [69] reveal that β_t and $\tilde{\beta}_t$ behave similarly across most diffusion steps, with minor deviations near $t = 0$ when the model emphasizes finer details. This indicates that the choice of σ_t have limited influence on sample quality, and $\mu_\theta(x_t, t)$ has much greater impact on the approximate posterior than $\Sigma_\theta(x_t, t)$.

To ensure the reverse posteriors are tractable, x_0 is introduced as a condition in each transition. Although x_{t-1}, by Markov properties, already captures the influence of x_0 up to $t-1$, this step is also important for reducing variance in Equation 2.23. Using Bayes' rule, each transition is then formulated as:

$$q(x_t|x_{t-1}, x_0) = \frac{q(x_{t-1}|x_t, x_0)q(x_t|x_0)}{q(x_{t-1}|x_0)}. \tag{2.19}$$

The training objective is to minimize the Variational Lower Bound (VLB) on the negative log-likelihood, derived from the ELBO used in VAEs [48]:

$$\begin{aligned}
\mathbb{E}_{x\sim q(x)}\left[\log p_\theta(x)\right] &\geq \mathbb{E}_{x_{0:T}\sim q(0:T)}\left[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}\right] && (2.20)\\
&= \mathbb{E}_{x_{0:T}\sim q(0:T)}\left[\log \frac{p_\theta(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)}{\prod_{t=1}^{T} q(x_t|x_{t-1})}\right] && (2.21)\\
&= \dots \\
&= \mathbb{E}_{x_1,x_0\sim q(x)}\left[\log p_\theta(x_0|x_1)\right] - D_{KL}\left(q(x_T|x_0) \parallel p_\theta(x_T)\right) && (2.22)\\
&\quad - \sum_{t=2}^{T} \mathbb{E}_{x_t,x_0\sim q(x)}\left[D_{KL}\left(q(x_{t-1}|x_t, x_0) \parallel p_\theta(x_{t-1}|x_t)\right)\right].
\end{aligned}$$

To delve into each individual term:

1. $\mathbb{E}_{x_1,x_0\sim q(x_0)}\left[\log p_\theta(x_0|x_1)\right]$ serves as a reconstruction term [48] and represents the final step required to transform the noisy samples back to their original state. In [22], this term is modeled using an independent discrete decoder derived from the

Gaussian distribution $\mathcal{N}(x_0; \mu_\theta(x_1, 1), \sigma_1^2\mathbf{I})$, where the transition from x_1 to x_0 is represented by the mean output $\mu_\theta(x_1, \mathbf{I})$.

2. $D_{KL}\left(q(x_T|x_0) \parallel p_\theta(x_T)\right)$ measures how closely the distribution of the final noisy input aligns with the Gaussian prior distribution. Since β_t is fixed and the approximate posterior q has no learnable parameters, this term is a constant and omitted from the loss function.

3. $\sum_{t=2}^{T} \mathbb{E}_{x_t, x_0 \sim q(x_0)} \left[D_{KL}\left(q(x_{t-1}|x_t, x_0) \parallel p_\theta(x_{t-1}|x_t)\right)\right]$ represents the denoising matching criterion. Here, $p_\theta(x_{t-1}|x_t)$ is the model's approximation of the denoising transition step, which aims to match the ground-truth $q(x_{t-1}|x_t, x_0)$. The latter describes how to denoise a noisy image x_t using knowledge of the denoised sample x_0. Minimizing this term ensures that the learned transition kernel closely matches the ground-truth, as measured by KL divergence.

In these derivations, all trainable terms are computed as expectations, which can be estimated via Monte Carlo sampling [70] and optimized using gradient-based methods. The main focus of optimization costs lies in the summation term, which dominates over the reconstruction term. Note that the expectations for each summation term are taken over a single random variable for x_t, due to the conditioning on x_0 in Equation 2.19, which also serves as a Rao-Blackwell estimator [71]. This prevents expectations over the pair x_{t-1}, x_{t+1} at each timestep. Without this approach, the variance of the Monte Carlo estimates would increase, as it is computed by summing up $T-2$ terms, potentially resulting in high variance for large T.

The trainable KL-divergence term is derived by expanding $q(x_{t-1}|x_t, x_0)$ using Bayes' rule:

$$q(x_{t-1}|x_t, x_0) = \frac{q(x_t|x_{t-1}, x_0)q(x_{t-1}|x_0)}{q(x_t|x_0)} \tag{2.23}$$

$$= \frac{\mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathbf{I})\mathcal{N}(x_{t-1}; \sqrt{\bar{\alpha}_{t-1}}x_0, (1-\bar{\alpha}_{t-1})\mathbf{I})}{\mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I})} \tag{2.24}$$

$$= \dots$$

$$\propto \mathcal{N}\left(x_{t-1}; \underbrace{\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})x_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)x_0}{1-\bar{\alpha}_t}}_{\tilde{\mu}_t(x_t, x_0)}, \underbrace{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}}_{\tilde{\beta}_t}\mathbf{I}\right). \tag{2.25}$$

Rearranging Equation 2.15 and substituting x_0 expresses $\tilde{\mu}_t(x_t, x_0)$ as a function of x_t, which can then be approximated by the parameterized mean $\mu_\theta(x_t, t)$. The resulting derivations are given by:

$$\tilde{\mu}_t(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_t \right), \tag{2.26}$$

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right). \tag{2.27}$$

In order to formulate a loss function that minimizes the difference between $\mu_\theta(x_t, t)$ and the ground-truth denoising mean $\tilde{\mu}_t(x_t, t)$, the optimization problem can be expressed by:

$$\arg\min_\theta D_{KL}(q(x_{t-1}|x_t, x_0) \parallel p_\theta(x_{t-1}|x_t)) \tag{2.28}$$

$$= \arg\min_\theta \left[\frac{1}{2||\Sigma_\theta(x_t, t)||^2} \left\| \tilde{\mu}_t(x_t, x_0) - \mu_\theta(\mathbf{x}_t, t) \right\|_2^2 \right] \tag{2.29}$$

$$= \arg\min_\theta \left[\frac{1}{2||\Sigma_\theta(x_t, t)||^2} || \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_t \right) - \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right) ||_2^2 \right] \tag{2.30}$$

$$= \arg\min_\theta \left[\frac{(1-\alpha_t)^2}{2\alpha_t(1-\bar{\alpha}_t)||\Sigma_\theta(x_t, t)||^2} ||\epsilon_t - \epsilon_\theta(x_t, t)||_2^2 \right]. \tag{2.31}$$

Based on Equation 2.31, predicting a noisy sample x_t is equivalent to predicting the noise introduced during the forward diffusion process up to step t. Heuristically, relying on a simplified objective that excludes the weighting term leads to improved performance [22], as it reduces the emphasis on loss terms for smaller values of t, where noise is minimal. By prioritizing more substantial denoising challenges at larger t values, the network can focus on the more difficult tasks that significantly improve its overall performance. The final loss function takes the form:

$$L(\theta) := \mathbb{E}_{t \sim \mathcal{U}[1,T], x_0 \sim q(x_0), \epsilon_t \sim \mathcal{N}(0, \mathbf{I})} \left[||\epsilon_t - \epsilon_\theta(x_t, t)||_2^2 \right] + C, \tag{2.32}$$

where C is a constant not depending on θ, x_t is computed from Equation 2.15, $\mathcal{U}[1, T]$ is a uniform distribution over the set $\{1, 2, \ldots, T\}$, and ϵ_θ is a deep neural network that predict the noise vector ϵ_t given x_t and t. For image-like data, the prevalent architecture used in diffusion models is a time-conditioned Attention U-Net [27, 28], which excels in capturing both local and global context within spatially structured data.

This theoretical framework addresses several limitations of GAN-based approaches.

Unlike GANs, DDPMs use a straightforward training objective based on variational inference to optimize the ELBO and approximate MLE. This improves mode coverage through likelihood maximization, promoting greater diversity in generated samples. By systematically injecting and removing noise, the model also learns to reconstruct diverse data variations, effectively exploring the data space and avoiding the mode collapse often seen in GANs. Furthermore, the model is easier to train, relying on a single, tractable likelihood loss function, leading to a more stable and predictable learning process without the volatile adversarial dynamics typical of GANs.

The primary bottleneck in diffusion-based models lies in their sequential Markov Chain sampling [22]. Moreover, the intermediate noisy samples, which serve as latent codes, retain the original size of the data, significantly increasing the complexity of generating image-like data. While GANs generate samples in a single forward pass, sampling with diffusion models is particularly slow and computationally intensive, as it requires T iterative steps. As a result, the high computational cost and energy demands of these processes significantly limit their broader use, especially without access to a modern GPU. For instance, generating 50,000 images at a 32×32 resolution using a DDPM takes about 20 hours on an Nvidia 2080 Ti GPU, whereas a GAN can produce the same number of images in under a minute [69].

2.2.5 Denoising Diffusion Implicit Models

Given the challenges posed by the slow and costly sampling process in diffusion-based models, recent research has focused on accelerating this process while preserving sample fidelity [26, 69, 72–74]. Denoising Diffusion Implicit Models (DDIMs) [26], a variant of implicit probabilistic models [75], offer a more efficient alternative to traditional DDPMs. Unlike DDPMs, which rely on a stochastic, Markovian reverse sampling process from a probability distribution, DDIMs employ a deterministic, non-Markovian sampling trajectory. DDIMs retains the forward marginals and training objectives of DDPMs while significantly speeding up sampling without compromising sample quality.

Consider Equations 2.19, 2.25, and 2.27, each now parameterized by a standard deviation σ_t. The revised formulations are presented below:

$$q_\sigma(x_t|x_{t-1}, x_0) = \frac{q_\sigma(x_{t-1}|x_t, x_0)q_\sigma(x_t|x_0)}{q_\sigma(x_{t-1}|x_0)}, \tag{2.33}$$

$$q_\sigma(x_{t-1}|x_t, x_0) = \mathcal{N}\left(x_{t-1}; \sqrt{\bar{\alpha}_{t-1}}x_0 + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\frac{x_t - \sqrt{\bar{\alpha}_t}x_0}{\sqrt{1-\bar{\alpha}_t}}, \sigma_t^2\mathbf{I}\right), \tag{2.34}$$

$$\tilde{\mu}_t(x_t, x_0) = \sqrt{\bar{\alpha}_{t-1}}x_0 + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\frac{x_t - \sqrt{\bar{\alpha}_t}x_0}{\sqrt{1-\bar{\alpha}_t}}. \tag{2.35}$$

Recall that $q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t\mathbf{I})$, where $\tilde{\beta}$ is defined in Equation 2.14, and $\tilde{\beta}_t = \sigma_t^2$. Let $\sigma_t^2 = \eta \cdot \beta_t$, with $\eta \in \mathbb{R}^+$ introduced as a hyperparameter to control the degree of stochasticity in the sampling process. If $\eta = 0$, the process becomes deterministic, entirely eliminating stochastic variability, and the model functions as an implicit probabilistic model [75], generating samples via a fixed procedure from x_T to x_0. Note that DDIMs deviate from the Markovian assumption in DDPMs by allowing x_t in the forward process to depend not only on x_{t-1} but also on the initial state x_0. Importantly, this still preserves the same marginal noise distribution in DDPMs (Equation 2.13):

$$q(x_1, x_2, \ldots, x_T|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}, x_0). \tag{2.36}$$

Alternatively, setting $\eta = 1$ for all t reinstates the Markovian properties of the forward process, reverting the generative model to a standard DDPM. Therefore, DDIMs can be considered a generalization of DDPMs, as they introduce non-Markovian dynamics in the reverse process, enabling more flexible and controlled sample generation.

The generative process $p_\theta(x_{0:T})$ is defined such that each transition kernel $p_\theta(x_{t-1}|x_t)$ leverages information from the posterior $q_\sigma(x_{t-1}|x_t, x_0)$. Starting with $x_0 \sim q(x_0)$ and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, a noisy observation x_t is obtained using Equation 2.16, where the model's output $\epsilon_\theta(x_t, t)$ predicts the noise component ϵ_t. The corresponding denoised observation $f_\theta(x_t, t)$, which estimates x_0, is given by:

$$f_\theta(x_t, t) := \frac{(x_t - \sqrt{1-\alpha_t}\epsilon_\theta(x_t, t))}{\sqrt{\alpha_t}}. \tag{2.37}$$

Subsequently, this result is used compute the approximate posterior, according to the following expression:

$$p_\theta(x_{t-1}|x_t) = \begin{cases} \mathcal{N}(f_\theta(x_1, 1), \sigma_1^2 I) & \text{if } t = 1, \\ q_\sigma(x_{t-1}|x_t, f_\theta(x_t, t)) & \text{otherwise.} \end{cases} \tag{2.38}$$

The Gaussian distribution at $t = 1$ ensures that the generative process covers all steps. Based on these derivations, x_{t-1} can be sampled from x_t via the expression:

$$x_{t-1} = \underbrace{\sqrt{\frac{\bar{\alpha}_{t-1}}{\bar{\alpha}_t}}\left(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t,t)\right)}_{\text{predicted } x_0} + \underbrace{\sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\epsilon_\theta(x_t,t)}_{\text{"direction pointing to" } x_t} + \underbrace{\sigma_t\epsilon}_{\text{random noise}}, \tag{2.39}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is a Gaussian independent of x_t, and $\alpha_0 := 1$.

The Parameters θ are optimized using the variational inference objective outlined below:

$$\begin{aligned} L_\sigma(\theta) := \mathbb{E}_{x_0\sim q_\sigma(x_0)}[\log q_\sigma(x_T|x_0) + \sum_{t=2}^{T}\log q_\sigma(x_{t-1}|x_t,x_0) \\ - \sum_{t=1}^{T}\log p_\theta(x_{t-1},x_t) - \log p_\theta(x_T)], \end{aligned} \tag{2.40}$$

where $q_\sigma(x_{1:T}|x_0)$ and $p_\theta(x_{0:T})$ are factorized from 2.36 and 2.17, respectively. While it may appear necessary to train individual models for each σ value due to their unique variational objectives, an important property is that for all $\sigma > 0$, there exists a constant $C \in \mathbb{R}$ such that $L_\sigma(\theta) = L(\theta) + C$ (Theorem 1 in [26]). Note that in Equation 2.32, the parameters θ associated with the model ϵ_θ are not shared across different timesteps t. This indicates that each timestep t sampled from the uniform distribution $\mathcal{U}[1,T]$ is optimized independently. As a result, the parameters remain distinct for each timestep, and the optimal solution for $L_\sigma(\theta)$ is equivalent to that of $L(\theta)$.

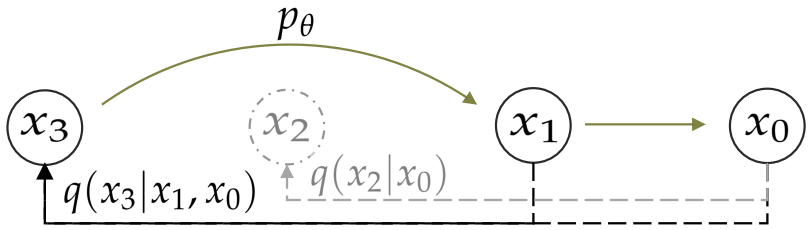

FIGURE 2.9: Accelerated sampling with DDIM using a subset of noisy latents x_{τ_1}, x_{τ_3}. The skipped latent x_{τ_2} is implicitly handled by the deterministic nature of DDIM, which adjusts the reverse process to predict x_{τ_3} directly from x_{τ_1} while preserving the noise schedule, thus reducing the number of steps. Image adapted from [26].

Different choices of σ lead to distinct generative processes, all while using the same model, ϵ_θ, without retraining for each new setting. Given that the denoising objective L is independent of the specific forward procedure as long as the posterior $q_\sigma(x_t|x_0)$ remains constant, one can exploit this property to reduce the number of diffusion steps T. By focusing only on a subset of latents $\{x_{\tau_1}, \ldots, x_{\tau_S}\}$ rather than tracing the entire chain from

$t = 1$ to T, the process becomes more efficient. Here, τ represents an increasing subsequence of time indices $\tau_1 < \tau_2 < \ldots < \tau_S$ chosen from the set $1, 2, \ldots, T$ with $S < T$, where the marginals for this subset are defined as $q(x_{\tau_i}|x_0) = \mathcal{N}(\sqrt{\alpha_{\tau_i}}x_0, (1 - \alpha_{\tau_i})\mathbf{I})$. Intuitively, DDIMs require fewer sampling steps compared to stochastic methods, as it eliminates randomness from the process and each step more precisely navigates towards low-temperature regions. Consider s and t as two points within this accelerated sequence, where $s < t$. The update mechanism in DDIM for these points is described by the posterior:

$$q_{\sigma,s<t}(x_s|x_t, x_0) = \mathcal{N}\left(x_s; \sqrt{\frac{\bar{\alpha}_s}{\bar{\alpha}_t}}\left(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t)\right) + \sqrt{1 - \bar{\alpha}_s - \sigma_t^2}\epsilon_\theta(x_t, t), \sigma_t^2 I\right). \tag{2.41}$$

This allows models to be trained with an arbitrary number of forward steps, while only selecting a subset of these steps to be used during the generation phase. In other words, the training procedure remains the same and only the method of generating samples is changed. These modifications are regulated by two parameters: τ, which influences the sampling speed, and σ, which allows interpolation between the deterministic DDIMs and the stochastic sampling method of DDPMs. As a result, by training a single generative model for the Markovian inference process in DDPMs, one inherently trains for multiple non-Markovian forward processes that are parameterized by σ, as described earlier.

Experiments in [26] and [69] demonstrate that DDIMs can produce high-quality samples up to 50 times faster in terms of wall-clock time compared to DDPMs. When models were trained using $T = 1000$ diffusion steps, DDIM $(\eta = 0)$ produced the best quality samples when using a small subset S, whereas DDPMs $(\eta = 1)$ performed much worse with the same small S. Nevertheless, DDPMs always performed better when one could afford to run the full reverse Markov chain (i.e., $S = T = 1000$). In many cases, samples generated with only 20 steps were already very similar to those generated with 1000 steps in terms of high-level features, with only minor differences in details. x_T alone serves as a robust latent encoding of the image, whereas minor details that affect sample quality are encoded in the parameters, as longer sample trajectories result in better quality samples but have no substantial influence on high-level features.

2.2.6 Latent Diffusion Models

Although DDIMs provide an efficient alternative for sampling via deterministic, non-Markovian trajectories, diffusion-based models still encounter significant computational bottlenecks when processing image-like data, as the noisy latents retain the same size as the original image, always operating directly in image space. Moreover, as likelihood-based methods, diffusion models tend to use excessive computational resources to model imperceptible details due to their mode-covering behavior. As a result, training and evaluating these models require resource-intensive evaluations and gradient computations in high-dimensional spaces, leading to substantial computational overhead in the absence of modern GPUs.

As outlined in Section 2.2, natural image data exist on a hidden lower-dimensional manifold within the high-dimensional image space. Therefore, training diffusion models directly on this manifold can substantially reduce computational load. These manifolds can be implicitly defined via an AE, which learns the underlying structure through a reconstruction loss. While most bits of an image contribute to perceptual details, the essential semantic and conceptual content is preserved, even with significant compression. This concept underlies LDMs [24], a sophisticated variant of diffusion models designed to operate in a low-dimensional encoded latent space. By first reducing pixel-level redundancy with an AE and then manipulating or generating semantic concepts within the encoded latent space, LDMs are able generate high-quality samples with improved computational efficiency and scalability. This approach overcomes many limitations associated with traditional DDPMs, making LDMs a versatile tool for various applications [76–78].

Rather than jointly training an AE and a diffusion model, LDMs address each component separately. For perceptual compression, an encoder $\mathcal{E}$ first maps a high-dimensional input image $x \in \mathbb{R}^{H\times W\times C}$ to a compressed latent representation z, formally expressed as $z = \mathcal{E}(x) : \mathbb{R}^{H\times W\times C} \to \mathbb{R}^{h\times w\times C}$. The downsampling factor f is defined as $f = \frac{H}{h} = \frac{W}{w} = 2^m$, where $m \in \mathbb{N}$. The resulting efficient, low-dimensional latent space captures essential image features while abstracting away high-frequency, imperceptible details. To reconstruct the compressed image, a decoder $\mathcal{D}$ takes the latent representation z and maps it back to the original image space, given by $\tilde{x} = \mathcal{D}(z) : \mathbb{R}^{h\times w\times C} \to \mathbb{R}^{H\times W\times C}$. The flexibility in the compression level requires careful consideration, given that the inductive biases in diffusion models, inherited from the U-Net architecture, scale better with higher spatial

dimensionality due to their convolutional backbone. In the subsequent stage, diffusion modeling is performed within this encoded latent space produced by the first-stage perceptual compression, enabling semantic compression by capturing and learning semantic and conceptual structures within z.

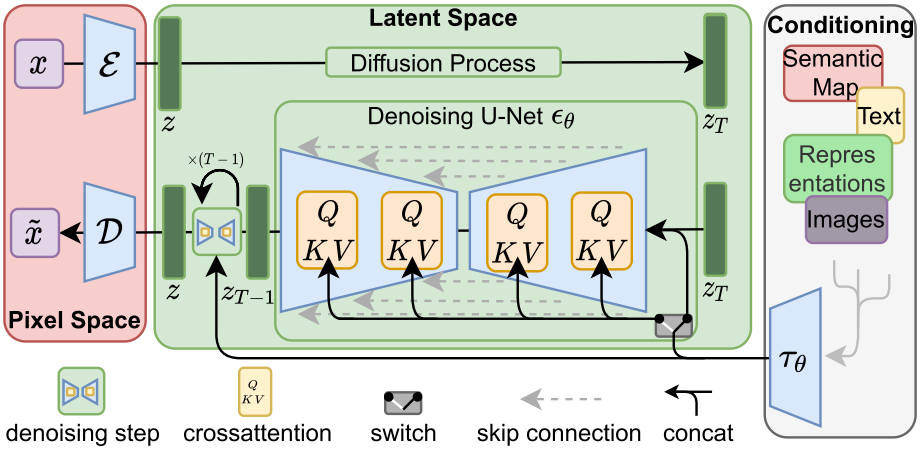

FIGURE 2.10: LDM: Images are first encoded into a compressed latent space using a pre-trained AE. The diffusion process then gradually adds noise to the latent representations, and a neural network with a U-Net-based architecture is trained to reverse this process, similar to DDPM. After denoising, the latent representations are decoded back into the original image space. Image taken from [24].

Note that the AE only needs to be trained once and can then serve as a pre-trained model, with its latent space serving as the basis for training multiple LDMs. In the second stage, aside from the encoding and decoding steps handled by the AE, the process follows the same procedures as in DDPMs or DDIMs but operates within the latent space instead of the original image space, as illustrated in Figure 2.10. During forward diffusion, noise is progressively added over t steps to the encoded latents z_0, derived from the initial image samples $\mathcal{E}(x_0)$, resulting in a series of noisy latent variables z_t. The model then performs reverse diffusion to recover z_0, which is subsequently transformed back into the original image space using $\mathcal{D}(z_0)$.

To integrate controlled image synthesis into LDMs with conditioning priors y, the proposed framework introduces an encoder τ_θ that projects y into intermediate representations. These representations can be mapped into the U-Net either via cross-attention mechanisms or direct concatenation, where $\tau_\theta(y) : \mathbb{R}^{H\times W\times C} \to \mathbb{R}^{h\times w\times C}$. The latter approach has demonstrated superior results for semantic image synthesis [24], with τ_θ consisting of a spatial rescaling setup that encodes y through direct multi-stage interpolation rather than using parameterized layers. The loss function for conditioned LDMs is given

by:

$$L_{LDM}(\theta) := \mathbb{E}_{t\sim\mathcal{U}[1,T],z_0\sim q(z_0),y,\epsilon_t\sim\mathcal{N}(0,\mathbf{I})}\left[||\epsilon_t - \epsilon_\theta(z_t, t, \tau_\theta(y))||_2^2\right] + C. \tag{2.42}$$

where the training objective, similar to Equation 2.32, is to minimize the difference between the estimated ϵ_t and ϵ_θ.

2.2.7 Spatially-Adaptive Normalization

Segmentation maps are visual tools used to delineate and categorize various regions or objects within an image. In the context of CMRI, these maps represent the anatomical structures of the heart, with each segment assigned a unique label, as outlined in Section 1.2. Leveraging segmentation maps as conditioning priors provide detailed, context-aware, and semantically meaningful information, allowing models to effectively distinguish between the different regions within the ROI. This ensures that the generated images remain coherent, with specific textures and styles designated to each tissue label, thus minimizing ambiguities in image interpretation. In generative modeling, this process is known as semantic image synthesis [18, 79, 80].

Integrating semantic layouts directly has proven suboptimal due to the effects of normalization layers like Batch Normalization (BN) [81], Instance Normalization (IN) [82] or Group Normalization (GN) [83]. While these layers are essential for stabilizing and accelerating training by maintaining consistent mean and variance in the inputs, they also inadvertently strip away the semantic information. This limitation is relevant to any architecture that employs convolution operations followed by normalization directly on semantic masks (e.g., Pix2PixHD [84]). Conditional normalization layers address the issue by adjusting the mean and variance of layer activations to match those of a reference. A well-known example is Adaptive Instance Normalization (AdaIN) [85], which has proven successful in style transfer tasks [86]. However, AdaIN aligns feature maps uniformly across spatial coordinates, which results in the loss of important local details and structures. To preserve spatial integrity in semantic information, [18] introduces SPADE Normalization, a technique using spatially-varying affine transformations to modulate feature maps at each pixel location.

Consider a segmentation mask $m \in \mathbb{R}^{H\times W\times 1}$, where each entry represents a pixel's semantic label from a set of $\mathbb{N}$ distinct classes. Using a single-channel integer map can lead to semantic ambiguity in convolution operations due to similar integer values at nearby locations. To prevent this, m is transformed into a multi-channel format using

one-hot encoding, such that $m : \mathbb{R}^{H \times W \times 1} \rightarrow \mathbb{R}^{H \times W \times \mathbb{N}}$, where each channel corresponds to a specific semantic label represented in binary form. Let $h' \in \mathbb{R}^{\mathcal{H} \times \mathcal{W} \times \mathcal{C}}$ denote the activation of the i-th layer of a deep neural network. Using a BN layer for normalization, the SPADE operations for a batch of samples with size N are defined by

$$\mu_c^i = \frac{1}{N\mathcal{H}^i\mathcal{W}^i} \sum_{n,y,x} h_{n,c,y,x}^i \tag{2.43}$$

$$\sigma_c^i = \sqrt{\frac{1}{N\mathcal{H}^i\mathcal{W}^i} \sum_{n,y,x} \left(\left(h_{n,c,y,x}^i \right)^2 - \left(\mu_c^i \right)^2 \right)} \tag{2.44}$$

$$h_{n,c,y,x}^{\prime i} = \gamma_{c,y,x}^i \left(m' \right) \frac{h_{n,c,y,x}^i - \mu_c^i}{\sigma_c^i} + \beta_{c,y,x}^i \left(m' \right), \tag{2.45}$$

where $n \in N$, $c \in \mathcal{C}^i$, $y \in \mathcal{H}^i$, $x \in \mathcal{W}^i$, and $m' \in \mathbb{R}^{\mathcal{H} \times \mathcal{W} \times \mathbb{N}}$ is the projection of the mask m onto an embedding space with nearest neighbors interpolation to match the layer's features resolution. The variables μ_c^i and σ_c^i represent the mean and standard deviation of the activations in channel c, respectively, while $\gamma_{c,y,x}^i \left(m' \right)$ and $\beta_{c,y,x}^i \left(m' \right)$ are the modulation parameters obtained from a two-layer convolution network, as illustrated in Figure 2.11. Note that these parameters are represented as tensors and vary according to the spatial coordinates (y, x) in m'.

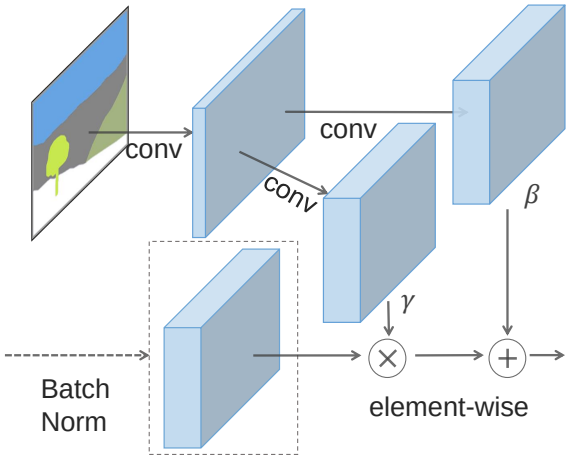

FIGURE 2.11: SPADE normalizes pixel-wise features with affine transformations, modulated by parameters γ and β derived from a semantic mask via a two-layered convolutional network. Image source: [18].

Chapter 3

Methodology

3.1 Anatomy-Compliant Medical Image Synthesis

We propose several LDM-based generators $\mathcal{G}$ [24], described in detail in Section 3.1, designed to translate anatomical structures into realistic anatomical images for style-transfer tasks. To complement or fully replace partial semantics y, we leverage edge maps e_x as conditioning priors to delineate anatomical features and impose a fully anatomical constraint. For edge detection, we integrate Canny [87], Sobel [88], and Laplacian [89] operators and assess how the varying complexities of these edge maps influence image generation and style-transfer performance.

Our models were trained and evaluated on the Automated Cardiac Diagnosis Challenge (ACDC) dataset [90], which includes cine bSSFP CMRIs from 100 patients designated for training and 50 patients allocated for testing. Each MRI scan is paired with segmentation maps that delineate LV, RV, and Myo regions. We processed each axial slice independently, resulting in a training set of 1,882 images and a testing set of 1,076 images. To prevent skewed signal intensity distributions and ensure that image details remain discernible, we applied min-max normalization based on the 1st and 99th percentiles. In the preprocessing stage, each image was resized to 256×256 px using bicubic interpolation [91], and the in-plane resolution was standardized to 1.56 mm^2/px to align with the median resolution across all axial slices. During training, we applied random affine transformations, including rotation by 90° and mirroring, to ensure rotational and reflectional invariance.

We benchmarked image generation performance against two SPADE-GAN variants introduced in [18]: one variant uses a random vector as input for the generator, while the

other employs an encoded latent vector in a VAE setup. Training details are provided in Appendix B.1. Our experiments were conducted using three distinct conditioning priors: partially labeled annotations via semantic image synthesis, edge guidance alone, and a combined method leveraging both edge maps and partial annotations. These ablations allowed us to assess the individual contributions of edges and semantics to image performance. We chose not to condition exclusively on edges for SPADE-GAN variants, as it diverges from the model's sole purpose of semantic image synthesis. Nevertheless, given that Canny edges are binary maps, we also integrated them into SPADE-GANs by concatenating them with the one-hot label encodings from semantic layouts. This approach is not feasible with Sobel and Laplacian edges, as converting them to binary form requires a magnitude threshold, which discards most edge information. For quantitative evaluation, our results are measured by Fréchet Inception Distance (FID) [92, 93], Structural Similarity Index Measure (SSIM) [94], and Normalized Mean Squared Error (NMSE) [95]. Furthermore, we analyzed the energy distributions within the Fourier space representations.

Our preliminary framework employing Canny edges was published in the short papers session of the Medical Imaging with Deep Learning (MIDL) 2024 Conference [96]. The full paper is available online [97] or in Appendix A.

3.1.1 Image Edge Detection Operators

Canny, Sobel, and Laplacian edge operators each have unique characteristics and require different preprocessing techniques. Both Canny and Laplacian edge detection methods are particularly sensitive to noise due to their gradient-based operators, which can amplify minor intensity variations, resulting in spurious edges that do not accurately reflect anatomical structures. Canny edges are represented as binary maps with thin, well-defined edges derived from first-order derivatives. Instead of relying on intensity magnitudes, the Canny operator classifies each pixel as either an edge or not. Meanwhile, the Laplacian operator detects edges by identifying zero-crossings in the second derivative of image intensity. The strength of these edges is determined by the magnitude of the Laplacian at each pixel. Larger values indicate stronger edges and appear as brighter pixels, representing more abrupt changes in the rate of intensity variation. Since the Laplacian is a second-order derivative, it is more sensitive to noise, amplifying high-frequency components.

To reduce the influence of noise in Canny and Laplacian edges, we applied Gaussian smoothing using a 5×5 kernel [98]. This technique averages pixel values with a weighted distribution, where pixels closer to the center receive higher weights according to the Gaussian bell curve. As the distance from the center increases, the weights decrease exponentially. The smoothed images then undergo Non-Local Means Denoising [99], where each pixel value is replaced with a weighted average of similar patches across the image rather than just neighboring ones—using a 5×5 patch size with low filter strength (5) for Canny, and a 7×7 patch size with higher filter strength (10) for Laplacian. These noise filters effectively preserved edge information outlining relevant anatomical structures while suppressing artifacts, though with a slight loss of detail, as illustrated in Figure 3.1.

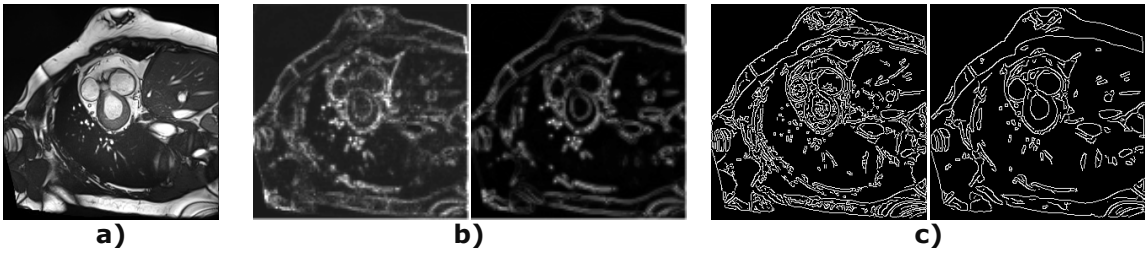

FIGURE 3.1: Denoising preprocessing for Canny and Laplacian edges: **a)** Ground-truth image; **b)** Laplacian edge maps; **c)** Canny edge maps. In **b)** and **c)**, (**Left**) shows edge maps from the original ground-truth, while (**Right**) shows maps after Gaussian smoothing and non-local means denoising.

The Sobel operator uses convolution kernels as first-order derivative approximators to compute gradients in both the horizontal and vertical directions. The gradient magnitudes in Sobel edges indicate the rate of increase or decrease in intensity between neighboring pixels. We found noise filters unnecessary with the Sobel operator, as its edges are less sensitive to noise due to the inherent smoothing effects of the convolution operations, which focus on local neighborhoods.

3.1.2 Edge-Informed Spatially-Adaptive Latent Diffusion Models

Following the framework described in Section 2.2.6, we employ a two-stage image synthesis process to perform diffusion modeling within an encoded latent space. First, we train an AE with Vector-Quantized (VQ) regularization [100, 101] using a downsampling factor of $f = 4$, resulting in latent representations $z \in \mathbb{R}^{64\times64\times C}$. More details are provided in Appendix B.2. For diffusion modeling, we set the number of Markov chain steps

to $T = 1000$ with a linear variance schedule and leverage efficient DDIM sampling, performing only 200 forward passes to significantly improve training speed. During inference, we run the full 1000-step Markov chain to ensure superior image generation quality. Our models are efficiently handled by a single 24 GB NVIDIA RTX 3090 GPU, bypassing the computational overhead often associated with diffusion-based models. Training hyperparameters are detailed in Appendix B.3.

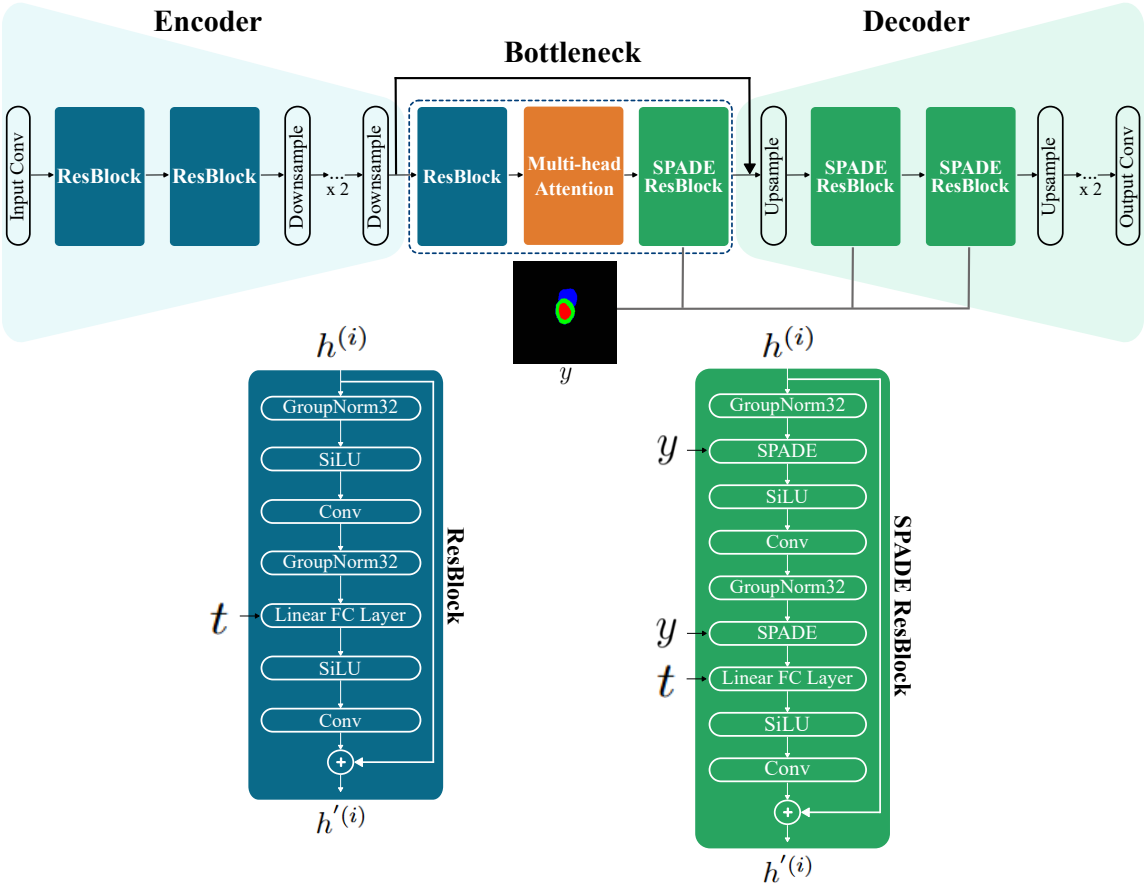

FIGURE 3.2: SPADE U-Net architecture: **Encoder** uses residual blocks with convolution layers to progressively downsample the input image, capturing features at multiple resolutions. **Bottleneck**, at the lowest resolution, captures the most abstract and complex features. **Decoder** upsamples compressed features, merging them with high-resolution details from the encoder via skip connections to restore the original image. SPADE normalization is applied during upsampling to integrate labeled annotations, leveraging semantic information at each spatial resolution.

To incorporate SPADE's semantic image synthesis capabilities into LDMs, we modify the U-Net architecture by replacing normalization layers in the decoder with SPADE normalization, as depicted in Figure 3.2. We also substitute non-linearities and batch normalization within SPADE blocks with Sigmoid Linear Units (SiLUs) [102] and 32-channel

GN [83]. These adjustments are designed to exploit semantic information at multi resolutions and preserve its integrity throughout the generation process. We limit these changes to the decoder to avoid overemphasizing semantic information. When conditioned exclusively on edge map priors $(e_x, y = \emptyset)$, the U-Net retains its original architecture.

The paired anatomy constraints (e_x, y) are concatenated into a two-channel tensor and integrated at each Markov chain reverse diffusion step. This tensor is embedded with the generator $\mathcal{G}$ through a multi-layered convolutional encoder $\tau_\theta(e_x, y) : \mathbb{R}^{256\times256\times2} \rightarrow \mathbb{R}^{64\times64\times C}$, as illustrated in Figure 3.3. For single conditioning, the encoder is defined as $\tau_\theta(e_x = \emptyset \vee y = \emptyset) : \mathbb{R}^{256\times256\times1} \rightarrow \mathbb{R}^{64\times64\times C}$. This approach deviates from the original implementation, which relied on direct multi-stage interpolation. We found this method suboptimal, as it mixes high-level latent features from z with low-level concatenated features. Instead, our encoder extracts features at multiple resolutions—256, 128, and 64—to enable progressive, spatially aware feature extraction while preserving crucial edge information. Moreover, it introduces conditioning parameters that can be concurrently fine-tuned via BP during training, enhancing adaptation to our condition information and improving the network's generalization, rather than solely relying on U-Net.

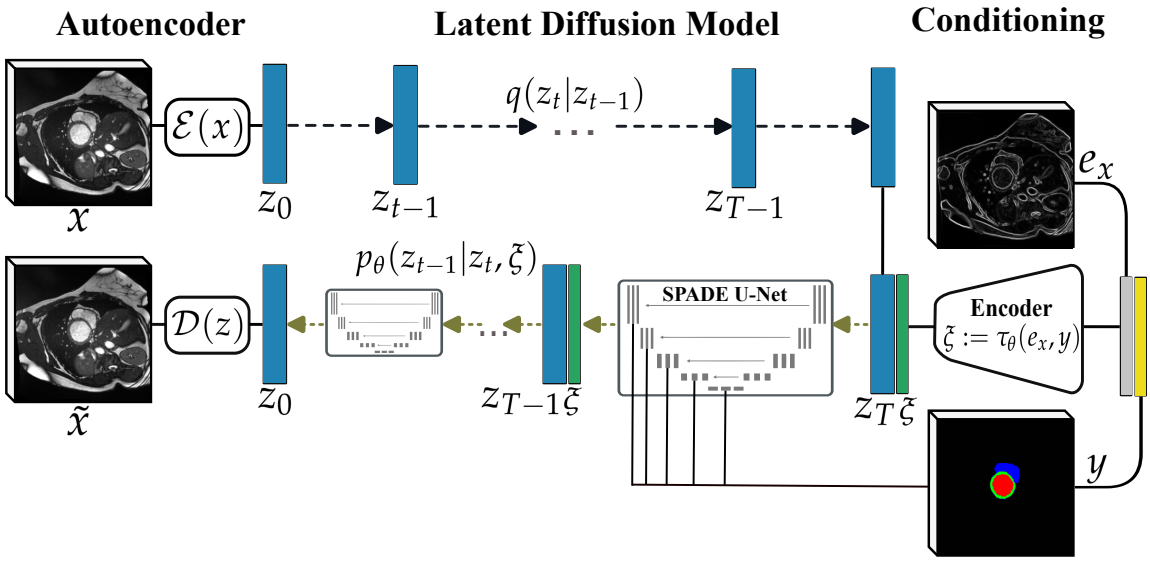

FIGURE 3.3: LDM-based image synthesis with dual conditioning (e_x, y): Edge maps and labeled annotations are encoded into high-level features that match the latent representations, and concatenated at each diffusion step before processing through the SPADE U-Net.

3.2 Image-to-Image Translation via Style-Transfer Techniques

In the context of practical clinical settings, we use our LDM-based generative models, trained on CMRI bSSFP sequences from the ACDC dataset, to perform image-to-image translation across distinct CMRI modalities. We assess whether the style-transferred images accurately replicate tissue-specific textures and enable precise automated segmentation using a model trained exclusively on the well-represented bSSFP sequence. Ultimately, this approach aims to determine whether we can extend the generalization capabilities of single-modality segmentation models to other imaging modalities without requiring multi-sequence acquisitions for training.

For the purpose of this task, we evaluated our models on the MICCAI Multi-Sequence CMR (MSCMR) Segmentation Challenge dataset [103, 104]. To directly translate characteristic bSSFP textures via conditioning priors, we use readily available edge maps (e_x) to impose anatomical constraints on the full FoV. We apply this style-transfer technique to the provided T_1-weighted LGE and T_2-weighted CMR images. The generated style-transferred images are then automatically segmented using a pre-trained model detailed in Section 3.2.1. Our quantitative results, measured by the Dice Score [105], are directly compared with the performance on the original images for reference. Additionally, we extend our analysis to private T_1 and T_2-mapping acquisitions collected at TU Delft Proton Therapy Center [106], but due to the lack of available labeled annotations for these modalities, we limit our evaluation to visual qualitative analysis.

We extend our image-to-image translation to synthetic Bloch-simulated bSSFP CMR images generated using the open-source MRXCAT framework on XCAT digital phantoms, as described in Section 3.2.2. The purpose of this use case is to transfer characteristic in-vivo textures to unrealistic in-silico images to make them more representative of in-vivo CMRI. Importantly, the anatomical masks in the phantoms provide a direct translation of semantically-labeled tissue information from the ACDC dataset, on which our SPADE-LDM variants were trained, to the in-silico images. Given that we were limited to three labeled segmented regions (i.e., LV, RV, and Myo), we complement partial semantics with edge maps as dual conditioning (e_x, y) using the framework outlined in Section 3.1.2. The resulting style-transferred in-silico images are then exclusively used as a training set to train a segmentation model from scratch and evaluate accuracy in segmenting the in-vivo ACDC holdout set and MICCAI MSCMR available bSSFP sequence images. We also

perform ablations on the semantics by comparing performance with models conditioned only on edge maps.

3.2.1 Automated Segmentation with nnU-Net

Our segmentation models are trained using the nnU-Net framework [107], an automated semantic segmentation pipeline built on the U-Net architecture, renowned for its state-of-the-art performance in numerous challenges [108, 109]. The automatic configuration provided by nnU-Net streamlines the entire workflow, including preprocessing, network architecture, training, and post-processing, as shown in Figure 3.4. This eliminates the time-consuming trial-and-error approach to method design.

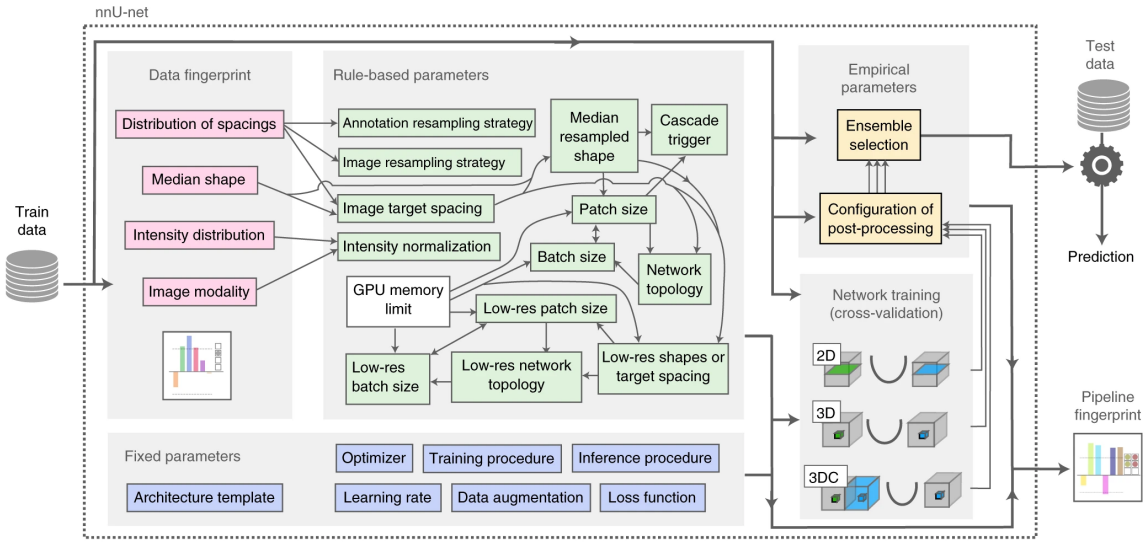

FIGURE 3.4: Automatic configuration in nnU-Net framework using interdependent heuristic rules. Adapted from [107].

For a given training set, the pipeline creates a "dataset fingerprint" that adapts design choices (e.g., network topology, batch size, or intensity normalization) based on interdependent hard-coded heuristic rules derived from the extensive datasets available in the Medical Segmentation Decathlon Challenge [110]. A detailed description of all heuristic rules is provided in the supplementary notes of the original paper [107]. The nnU-Net configuration generates three different options (2D, 3D, and 3D cascade), from which one must be chosen manually. It also sets up a 5-fold cross-validation for validation during training and ensemble processes. Given the anisotropic nature of our data, with varying in-plane resolution and axial slice thickness, we used the 2D U-Net configuration for our datasets. For simplicity, we selected only the best-performing model from the cross-validation and did not perform ensemble.

3.2.2 MRXCAT Numerical CMR Simulations

We generate synthetic breath-hold cine CMR images using MRXCAT [13], a framework for numerical simulations based on 4D XCAT digital phantoms [14, 111]. XCAT phantoms provide anatomical tissue masks that incorporate physiological parameters, such as heart rate and respiratory cycle duration, to simulate cardiac contraction and respiratory motion associated with CMRI. The anatomical data in XCAT is derived from the Visible Human Project by the National Library of Medicine [112]. Bloch simulations assign distinct MR properties to each activity value, reproducing tissue-specific signal behavior within a given imaging scheme.

MRXCAT is distributed in MATLAB source code format and includes digital phantoms derived from 54 individual patients. We use these phantoms to perform numerical simulations, replicating the bSSFP imaging sequence modeled by the magnetization behavior defined in Equation 2.7, under the same underlying assumptions. Since the ACDC dataset provides labeled annotations for the RV region, but it is denoted with a similar activity value to an artery in the XCAT anatomy files, we modified the original implementation to ensure the RV section is distinctly labeled. The other activity values remained unchanged, following the default baseline files.

TABLE 3.1: Simulation parameters for XCAT digital phantoms in cine CMR imaging.

Parameters	**Values**
Number of Cine Frames	30
Number of Slices	20
Time per Frame	0.0333 s
Heart Period	1 s
Array Size	512×512 px
Rotation Angle (x-axis)	115°
Rotation Angle (y-axis)	35°
Rotation Angle (z-axis)	240°
Pixel Size	0.16 cm
Slice Thickness	1.36 cm

The parameter settings used for XCAT phantoms in cine CMR imaging are presented in Table 3.1. The 20 axial slices ensure full heart coverage for every patient, accounting for anatomical variability. Rotation values around the x, y, and z axes are carefully selected to align the phantoms parallel to the short-axis plane of the heart. To introduce variability in the in-silico images, we also applied uniform random fluctuations to the rotations, with values ranging between $-5°$ and $5°$. The MR properties assigned to each tissue

were based on reference values from textbook literature at 1.5T, derived from in vivo measurements [41]. Examples of tissue-specific parameters are provided in Table 3.2. The pulse sequence simulations were configured with $\alpha = 50^\circ$ (Equation 3.2) and T_E ranging from 0 to 20 ms. Each patient had a slightly different T_E value, with T_R set to $T_R = 2T_E$.

TABLE 3.2: Tissue-specific parameters for Bloch simulations with MRXCAT, based on reference values from textbook literature at 1.5T [41].

Tissue	ρ_0 **(%)**	T_1 **Relaxation Times**	T_2 **Relaxation Times**
Blood	95	1200	100
Bone	12	250	20
Fat	70	350	30
Muscle	80	900	50
Liver	90	800	50

The resulting synthetic images undergo post-processing to more closely resemble MRI scans. A low-pass filter with a 1.2 cm circular radius is applied at each pixel location, smoothing sharp edges and fine details to better approximate realistic signal intensities at tissue interfaces. Additionally, Gaussian noise is introduced to closer replicate realistic SNR. MRXCAT determines the appropriate standard deviation based on the largest Myo tissue within the phantoms, which serves as the ROI. The final images are cropped to 256 $\times$ 256 px, centered on the heart region, to ensure consistency with our generative models resolution.

3.2.3 Pseudo-Labeled Lung Semantic Annotations

Building on the approach proposed by [31], we introduce a first-stage nnU-Net segmentation model trained on paired in-silico images and XCAT digital phantoms to generate pseudo-labeled lung annotations for in-vivo data, as illustrated in Figure 3.5. This process allows us to extend partially labeled ACDC semantics to 4-labels, which is our clearest and most direct technique for translating tissue-specific textures into in-silico images.

Although the lung cavities are realistically depicted with dark textures based on assigned $\rho_0 = 0$, our edge conditioning priors may mislead the image generation process by imposing incorrect anatomical constraints on these ambiguous, empty regions, potentially leading to hallucinated anatomy. This holds particularly true with Canny Edges, which is our most simplistic form of anatomical constraint. Since the lung region is adjacent to the heart and close to our ROI, we explore whether incorporating pseudo-labeled

semantics can mitigate spurious regions and ultimately lead to better segmentation results.

We limit this framework to the lung region, as it is the only organ near the ROI that exhibits significant structural and textural similarities to in-vivo images. Since the lung semantics correspond to darker regions, any misplaced pseudo-labels tend to outline the outer areas of the image without compromising its overall integrity.

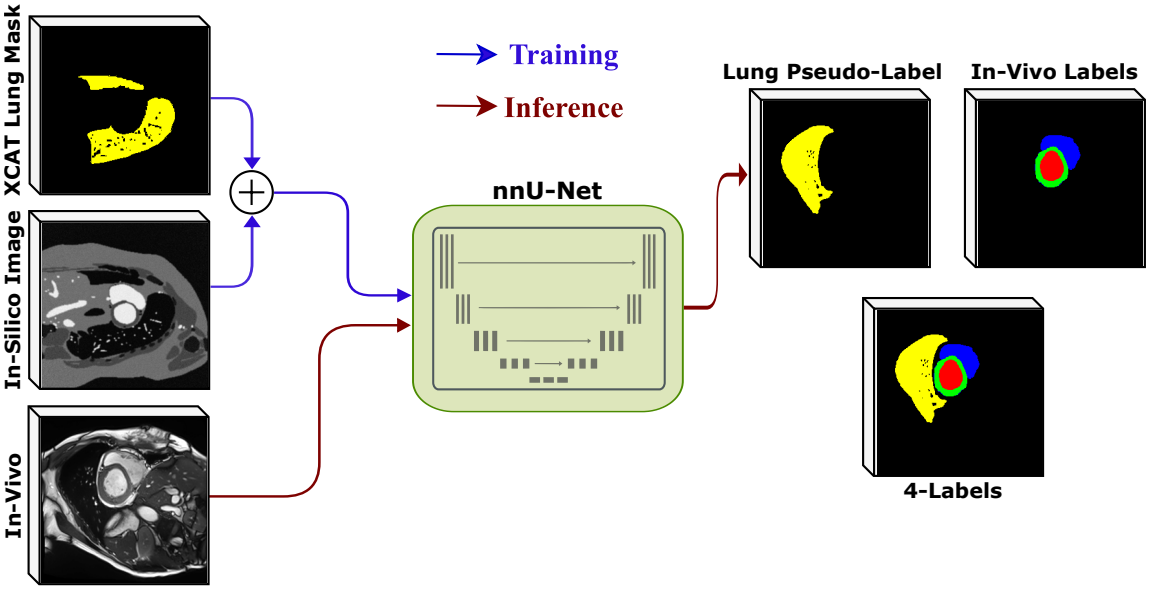

FIGURE 3.5: First-stage nnU-Net trained on paired in-silico images with XCAT anatomical lung masks to generate in-vivo pseudo-annotations for the lung region, extending ACDC semantics to 4-labels. Image adapted from [31].

Chapter 4

Results

4.1 Image Generation Performance

4.1.1 Semantic Image Synthesis

Table 4.1 presents the quantitative evaluation results for image generation performance. Among the models conditioned solely on partial semantics, SPADE-LDM achieved the best FID score (82.6). In terms of SSIM and NMSE, neither of the models showed any significant differences. For the SPADE-GAN variants, employing a VAE setup resulted in an improved FID (182.8) compared to the baseline architecture (149.5). Examples of generated CMR images, along with their Fourier frequency spectra, are included in Figure 4.1. The ROI is consistently well-represented, reflecting the embedded semantic information. However, the surrounding anatomical structures exhibit noticeable deviations from the ground-truth (GT), with some areas displaying unrealistic, hallucinated anatomical features. In comparison, the SPADE-GAN variants provide a more precise depiction of the ROI, while the SPADE-LDM model shows less coherence, particularly along the delineated edges. Despite this, SPADE-LDM is more adept at preserving overall anatomical integrity, as evidenced in the thoracic muscles. Figure 4.4a compares the average spectral energy distribution across the ACDC holdout dataset. SPADE-GAN variants align closely with the GT at low and mid-frequencies but show deviations at higher frequencies. Conversely, SPADE-LDM shows divergence in the mid-frequency range but achieves a closer approximation to the GT at higher frequencies.

TABLE 4.1: Image generation performance on ACDC dataset [90] measured by FID, SSIM and NMSE. Models without edge conditioning are denoted by $e_x = \emptyset$. A lower value is preferable for metrics marked with ↓, while a higher value is better for those marked with ↑.

Model $\mathcal{G}$	FID ↓	SSIM ↑	NMSE ↓
SPADE-GAN	80.5	0.482 ± 0.128	0.021 ± 0.005
SPADE-GAN ($e_x = \emptyset$)	182.8	0.194 ± 0.063	0.024 ± 0.006
SPADE-GAN w/ VAE	74.4	0.491 ± 0.128	0.020 ± 0.005
SPADE-GAN w/ VAE ($e_x = \emptyset$)	149.5	0.193 ± 0.060	0.024 ± 0.006
Canny-LDM	41.7	0.559 ± 0.159	0.019 ± 0.005
Laplacian-LDM	34.9	0.634 ± 0.187	0.017 ± 0.005
Sobel-LDM	**25.5**	0.712 ± 0.219	0.015 ± 0.005
Canny-SPADE-LDM	44.2	0.592 ± 0.136	0.019 ± 0.005
Laplacian-SPADE-LDM	38.8	0.654 ± 0.166	0.016 ± 0.004
Sobel-SPADE-LDM	26.5	$\mathbf{0.763 \pm 0.175}$	$\mathbf{0.014 \pm 0.004}$
SPADE-LDM ($e_x = \emptyset$)	82.6	0.187 ± 0.057	0.024 ± 0.006

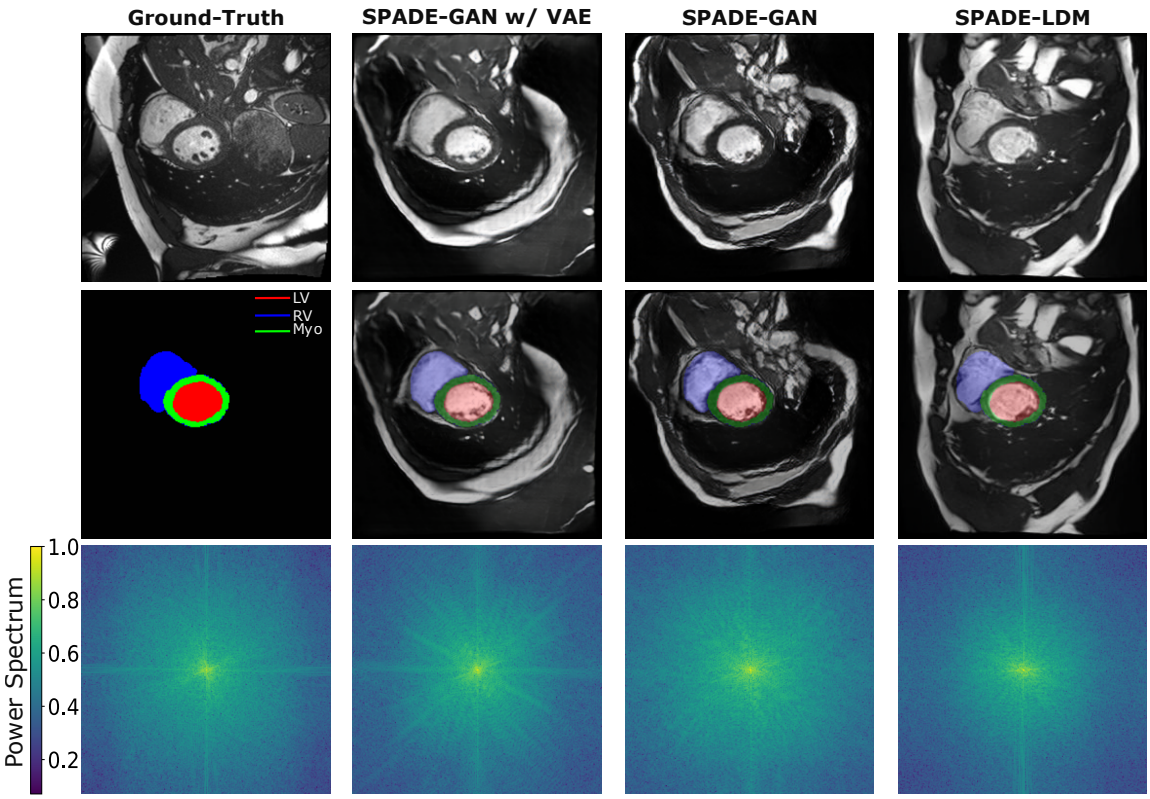

FIGURE 4.1: **Top:** Qualitative image generation comparison between models using only semantic image synthesis ($e_x = \emptyset, y$). **Center:** Generated images overlaid with ground-truth semantic layout. **Bottom:** Fourier spectra of the corresponding images.

4.1.2 Edge-Informed Image Synthesis

In all cases, models leveraging edge maps as conditioning priors, including the SPADE-GAN variants, demonstrated substantial improvements in image performance for all evaluated metrics. Figures 4.3 and 4.2 present the corresponding generated CMR images for SPADE and non-SPADE variants, respectively, as compared to Figure 4.1. Our edge-informed models consistently delineated organ details and anatomical structures within the full FoV that other models failed to capture. Figure 4.4b highlights how the energy distributions closely align with the GT at high frequencies. The best-performing LDM-based models incorporate Sobel gradient operators, achieving FID scores of 25.5 and 26.5 with Sobel-SPADE and Sobel-SPADE-LDM variants, respectively. The latter achieved superior results in terms of SSIM (0.763 ± 0.1) and NMSE (0.014 ± 0.004). Conversely, VAE-based SPADE-GAN performed significantly worse in terms of both FID (74.4) and an SSIM (0.491 ± 0.128). LDM-based models employing Laplacian edges also obtained competitive results, whereas those using Canny edges underperformed.

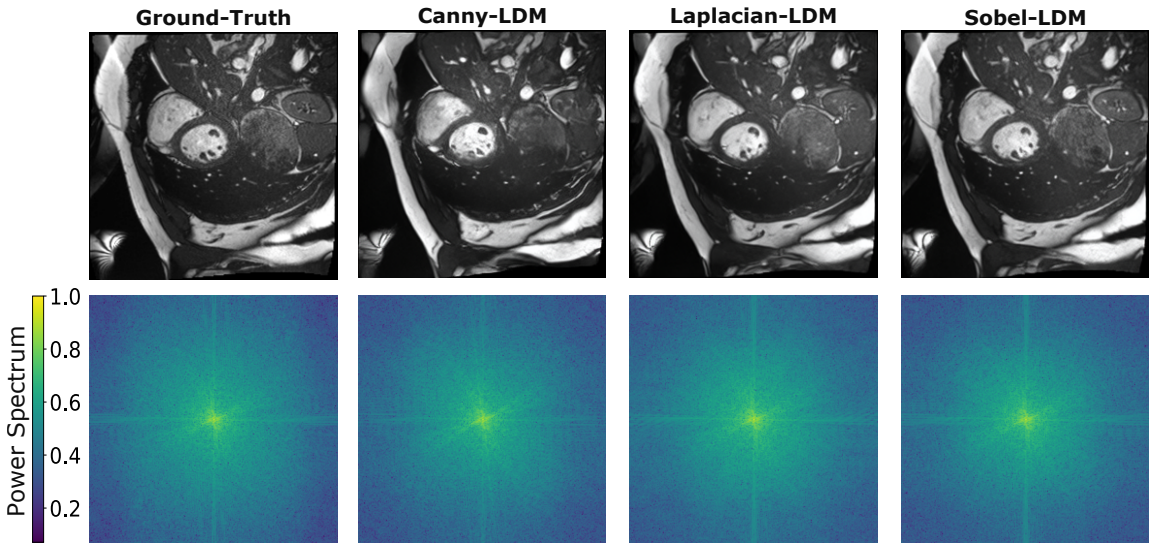

FIGURE 4.2: **Top:** Qualitative image generation comparison between models using only edge conditioning ($e_x, y = \emptyset$). **Bottom:** Fourier spectra of the corresponding images.

Our edge-informed SPADE-LDM variants showed consistent improvements both in SSIM and NMSE, at the expense of increased FID. For a more equitable comparison of image generation performance within the ROI, Table 4.2 presents metrics evaluated on a 128×128 centrally cropped area around the heart, focusing on semantically significant regions. The results within the ROI are consistent with those observed in the full FoV:

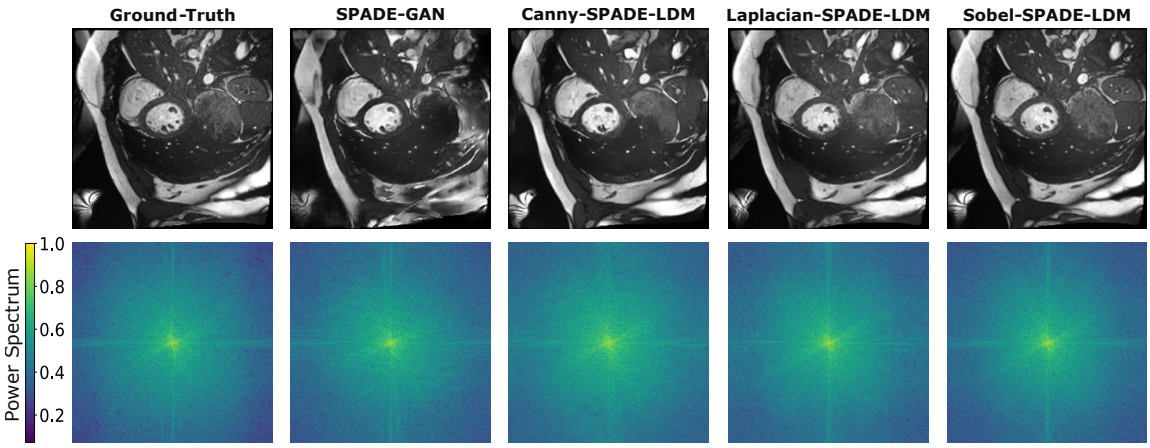

FIGURE 4.3: **Top:** Qualitative image generation comparison between models using edge conditioning and partial semantics as dual conditioning (e_x, y). **Bottom:** Fourier spectra of the corresponding images.

LDM-based models outperform SPADE-GAN variants, with Sobel edge priors demonstrating superior results, while Canny edges are subpar.

TABLE 4.2: Performance comparison of image generation on ACDC in a centrally cropped 128×128 area, focusing on semantically significant regions in the ROI. A lower value is preferable for metrics marked with $\downarrow$, while a higher value is better for those marked with $\uparrow$.

Model $\mathcal{G}$	FID $\downarrow$	SSIM $\uparrow$	NMSE $\downarrow$
SPADE-GAN	95.2	0.448 ± 0.143	0.028 ± 0.014
SPADE-GAN ($e_x = \emptyset$)	146.7	0.287 ± 0.086	0.030 ± 0.015
SPADE-GAN w/ VAE	92.8	0.460 ± 0.142	0.027 ± 0.013
SPADE-GAN w/ VAE ($e_x = \emptyset$)	126.2	0.286 ± 0.085	0.030 ± 0.015
Canny-LDM	57.9	0.501 ± 0.181	0.023 ± 0.007
Laplacian-LDM	55.1	0.556 ± 0.205	0.022 ± 0.009
Sobel-LDM	**46.2**	0.636 ± 0.242	0.022 ± 0.010
Canny-SPADE-LDM	60.5	0.546 ± 0.146	0.023 ± 0.008
Laplacian-SPADE-LDM	52.7	0.601 ± 0.021	0.022 ± 0.010
Sobel-SPADE-LDM	47.6	$\mathbf{0.710 \pm 0.192}$	$\mathbf{0.020 \pm 0.008}$
SPADE-LDM ($e_x = \emptyset$)	95.2	0.180 ± 0.069	0.030 ± 0.016

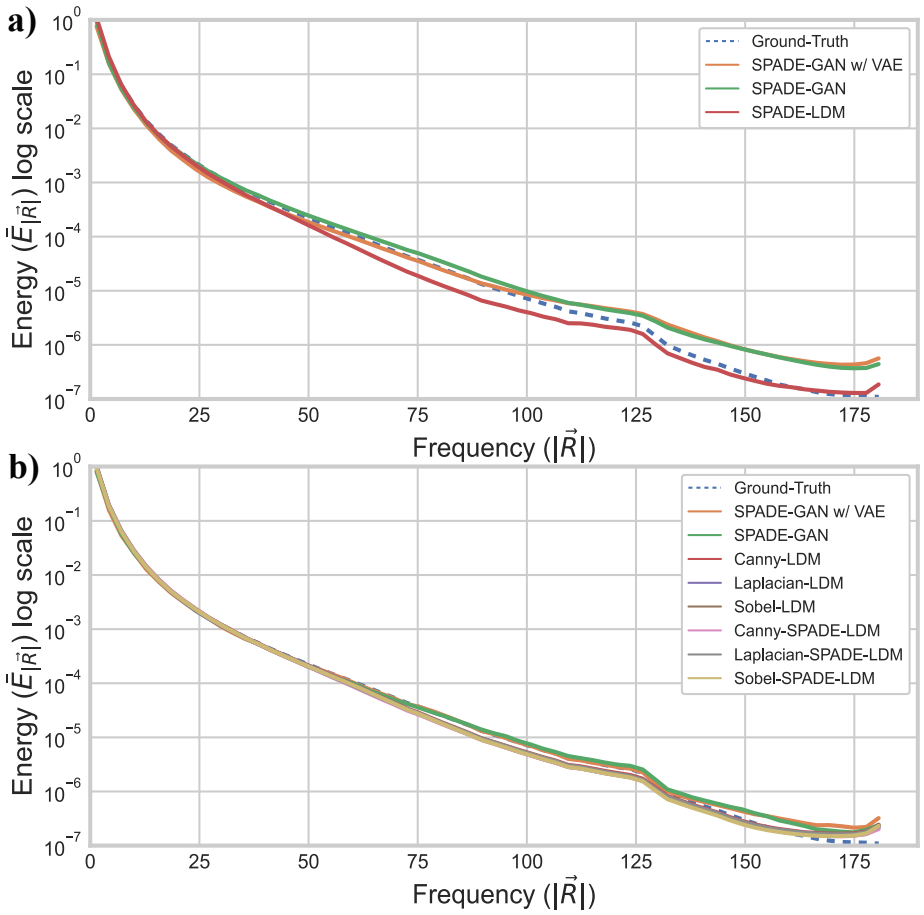

FIGURE 4.4: Averaged spectral energy distribution as a function of radial distance $|\vec{R}| = \sqrt{x^2 + y^2}$ from the 2D grid origin. **a)** Models without edge conditioning ($e_x = \emptyset, y$). **b)** Models with both edge conditioning ($e_x, y = \emptyset$) and dual conditioning (e_x, y).

4.2 Downstream Segmentation

4.2.1 Multi-Sequence Acquisitions

We present the quantitative downstream results of image-to-image translation applied to T_1-weighted LGE and T_2-weighted images in Table 4.3. Style-transferred images significantly improved Dice Scores on T_2-weighted images compared to direct segmentation on the original images. Among the models evaluated, Canny-LDM model achieved overall superior performance, increasing Dice Scores from 0.303 $\pm$ 0.188 (LV), 0.088 $\pm$ 0.112 (RV), and 0.233 $\pm$ 0.214 (Myo), to 0.819 $\pm$ 0.099, 0.692 $\pm$ 0.148, and 0.707 $\pm$ 0.108, respectively. Meanwhile, Laplacian-LDM showed highest improvement in the RV region (0.832 $\pm$ 0.094), whereas Sobel-LDM performed the worst. Figure 4.8 depicts style-transferred T_2-weighted images along with the corresponding nnU-Net segmentation outputs. Visually, in all cases, the style-transfer process seamlessly translated tissue-specific textures

within the ROI that resemble characteristic bSSFP imaging schemes. This is further evidenced by the resulting segmentation maps, which closely align with the GT.

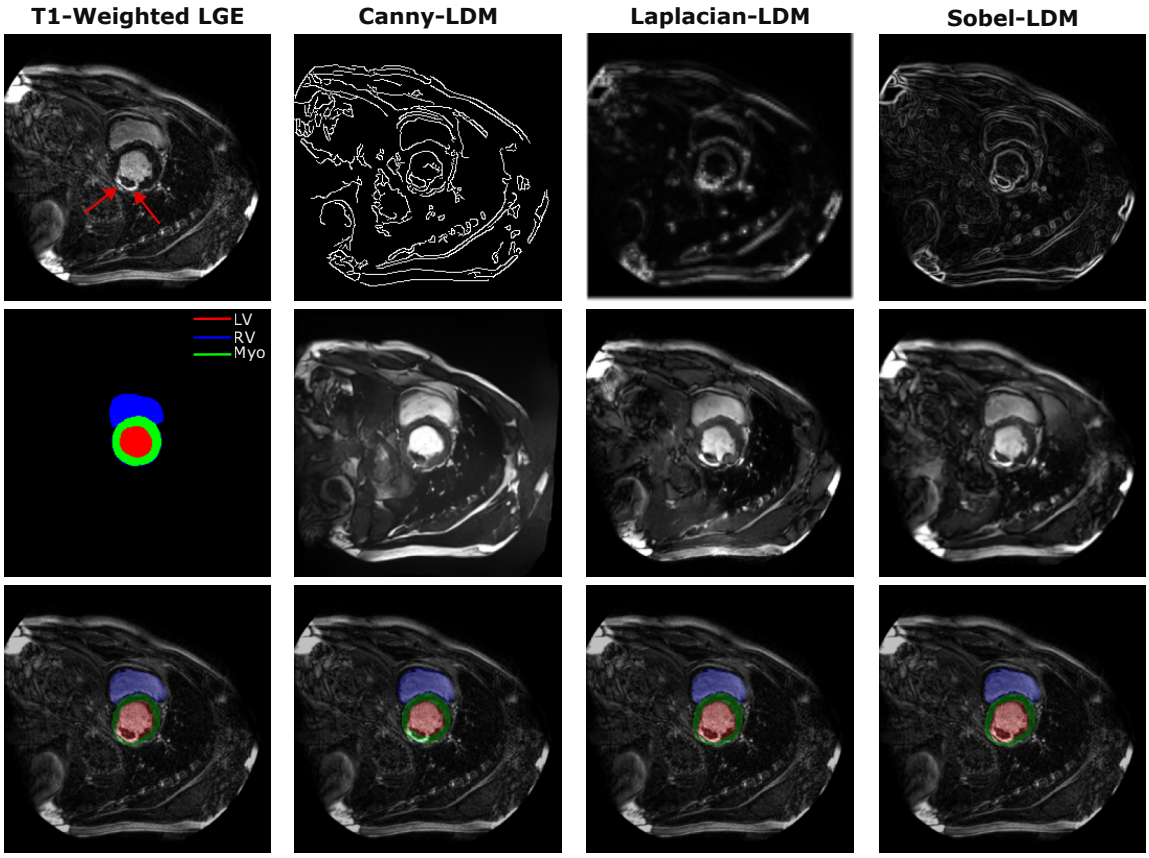

FIGURE 4.5: **Top:** Ground-truth T_1-weighted LGE image and the edge anatomical boundaries detected by the Canny, Laplacian, and Sobel operators. Red arrows indicate expansion of the interstitial space, highlighted by LGE. **Center:** Ground-truth segmentation map and generated style-transferred images. **Bottom:** Images overlaid with the corresponding nnU-Net segmentation output. The image furthest to the left shows the direct segmentation on the original image.

TABLE 4.3: Segmentation performance on style-transferred MICCAI MSCMR T_1-weighted LGE and T_2-weighted sequences, with images translated from edge-conditioned LDM-based models trained on the ACDC bSSFP training set. Results are evaluated using a nnU-Net trained on the same dataset and measured by Dice Score.

	Dice Score Performance					
	T_1-weighted LGE			T_2-weighted		
Model $\mathcal{G}$	LV	RV	Myo	LV	RV	Myo
Canny-LDM	$\mathbf{0.850 \pm 0.095}$	$\mathbf{0.690 \pm 0.167}$	$\mathbf{0.671 \pm 0.108}$	0.819 ± 0.099	$\mathbf{0.692 \pm 0.148}$	$\mathbf{0.707 \pm 0.108}$
Laplacian-LDM	0.849 ± 0.074	0.685 ± 0.150	0.641 ± 0.102	$\mathbf{0.832 \pm 0.094}$	0.538 ± 0.225	0.696 ± 0.124
Sobel-LDM	0.786 ± 0.118	0.626 ± 0.206	0.523 ± 0.151	0.708 ± 0.188	0.436 ± 0.240	0.550 ± 0.202
Original	0.822 ± 0.137	0.679 ± 0.234	0.601 ± 0.172	0.303 ± 0.188	0.088 ± 0.112	0.233 ± 0.214

In comparison, image-to-image translation did not yield similar improvements in T_1-weighted LGE as observed with T_2-weighted images, given that segmentation results on the original images already produced relatively high Dice Scores: 0.822 ± 0.137 (LV), 0.679

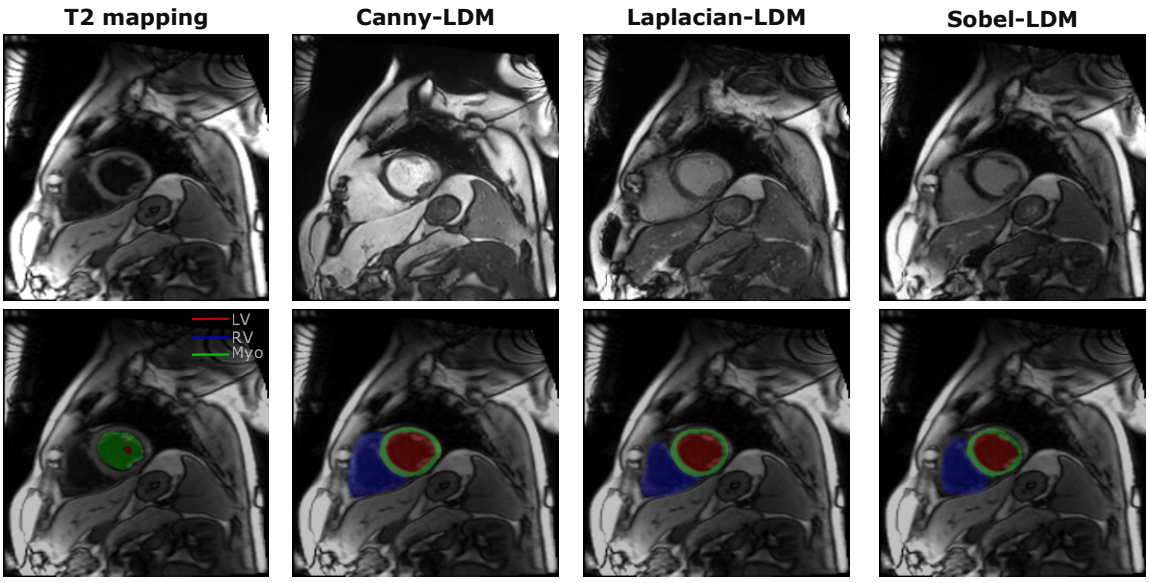

FIGURE 4.6: **Top:** Ground-truth T_2-mapping image and the edge anatomical boundaries detected by the Canny, Laplacian, and Sobel operators. **Bottom:** Images overlaid with the corresponding nnU-Net segmentation output. The image furthest to the left shows the direct segmentation on the original image.

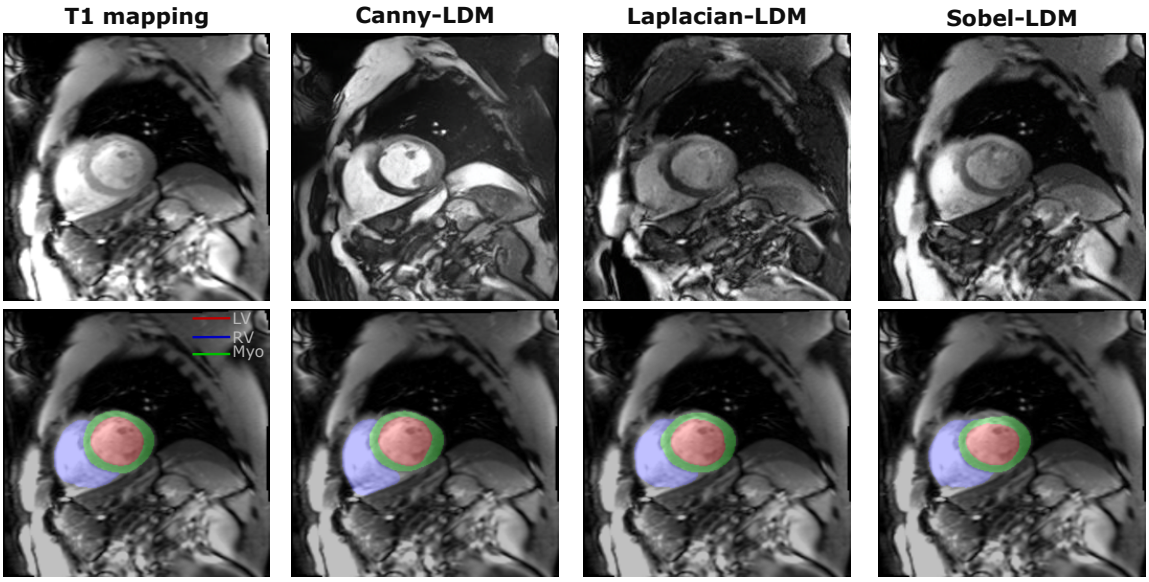

FIGURE 4.7: **Top:** Ground-truth T_1-mapping image and the edge anatomical boundaries detected by the Canny, Laplacian, and Sobel operators. **Bottom:** Images overlaid with the corresponding nnU-Net segmentation output. The image furthest to the left shows the direct segmentation on the original image.

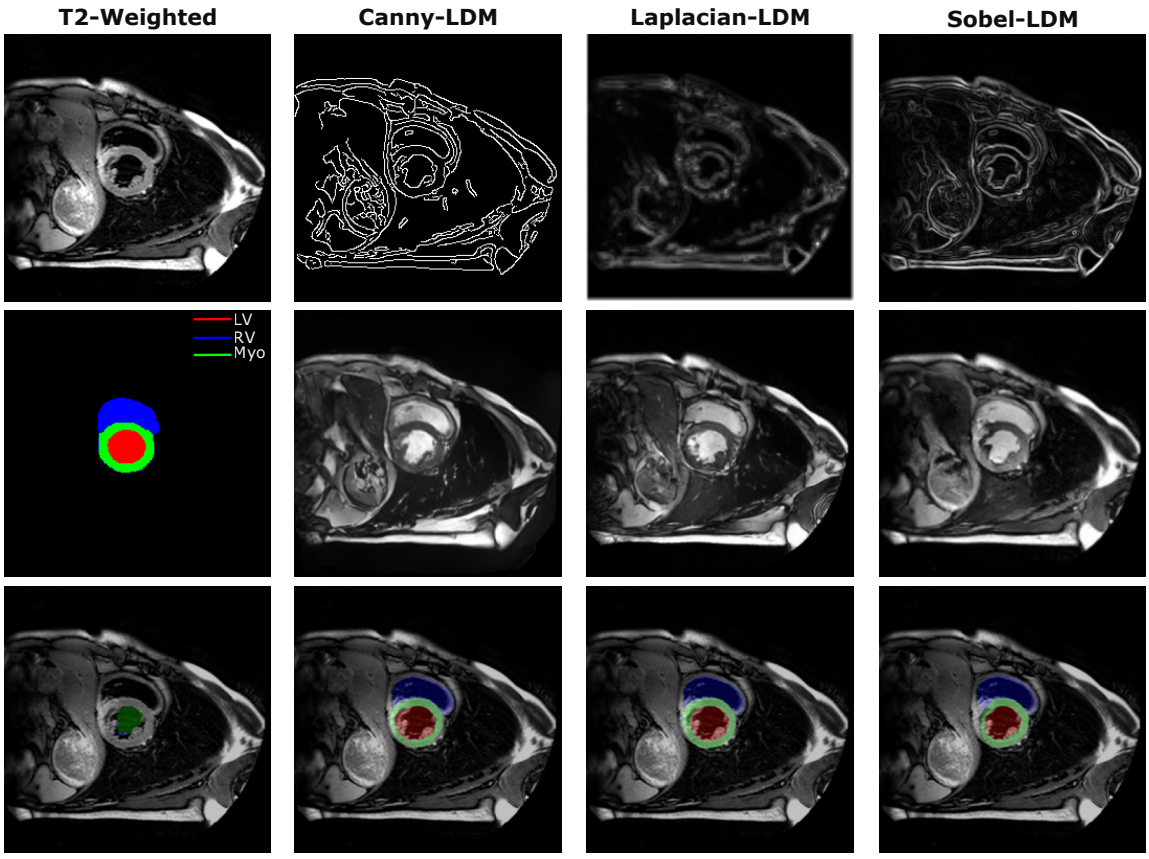

FIGURE 4.8: **Top:** Ground-truth T_2-weighted image and the edge anatomical boundaries detected by the Canny, Laplacian, and Sobel operators. **Center:** Ground-truth segmentation map and generated style-transferred images. **Bottom:** Images overlaid with the corresponding nnU-Net segmentation output. The image furthest to the left shows the direct segmentation on the original image.

$\pm$ 0.234 (RV), and 0.601 $\pm$ 0.172 (Myo). Nevertheless, style-transferred images via Canny-LDM and Laplacian-LDM models were still able to slightly improve performance across all regions. Canny-LDM showed the most notable improvements, increasing Dice Scores to 0.850 $\pm$ 0.095 (LV), 0.690 $\pm$ 0.167 (RV), and 0.671 $\pm$ 0.110 (Myo). On the other hand, Sobel-LDM texturization yielded even lower Dice Scores. Figure 4.5 provides an example of our style-transferred T_1-weighted LGE images. The originally delineated regions align with the GT in the RV, but show deviations in the LV and Myo areas near the damaged tissue highlighted by LGE. These regions are closer to the GT when Canny-LDM translation is applied. Figures 4.6 and 4.7 present style-transferred T_1 and T_2-mapping images, respectively. Although we lack GT references for a quantitative analysis of downstream automated segmentation performance, the T_2-mapping results clearly show improvements, with the delineated heart regions resembling the expected areas. Similar to T_1-weighted

LGE, segmentation results on T_1-mapping are already visually close to the GT, with no noticeable improvements. As seen in the other imaging schemes, Canny-LDM proves to be more suitable for image-to-image translation in both modalities compared to Laplacian-LDM and Sobel-LDM.

4.2.2 In-Silico Images

Some examples of Bloch-simulated CMR images generated using the MRXCAT framework are provided in Figure 4.9, representing various axial slices from a cine frame at ED, paired with their corresponding XCAT phantoms. Figure 4.10 depicts style-transferred

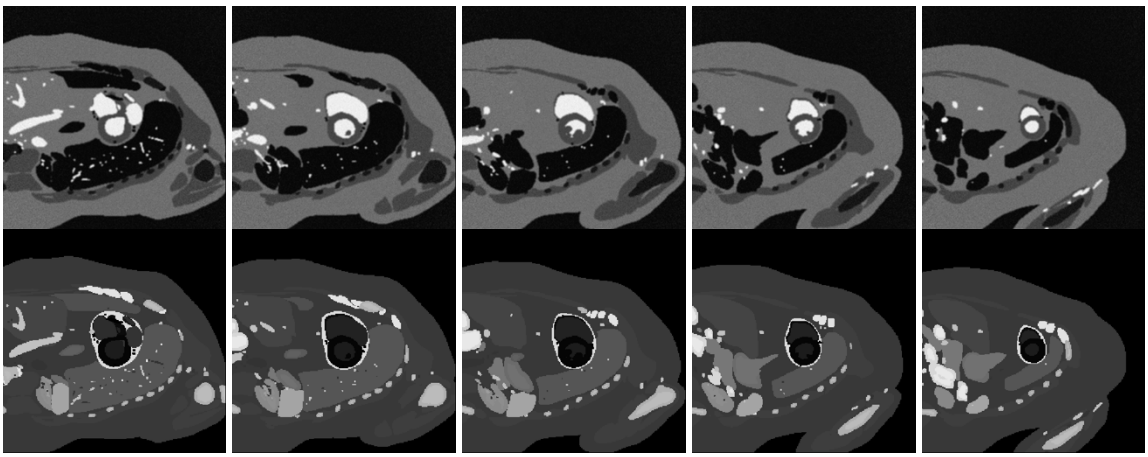

FIGURE 4.9: **Top:** Bloch-simulated in-silico CMR images showing different axial slices from a cine frame at ED. **Bottom:** Corresponding XCAT anatomical masks with activity values representing distinct tissues.

in-silico images after translating in-vivo characteristic textures using our edge-informed SPADE-LDM models. Consistent with other style-transfer tasks, Canny-SPADE-LDM excelled at translating the tissue-specific textures of the bSSFP scheme, including the inherent noise typically observed in in-vivo CMR images. In contrast, the images generated by Sobel-SPADE-LDM remained very similar to the original in-silico versions. Although Laplacian-SPADE-LDM showed improvements over Sobel-SPADE-LDM, it still fell short in rendering realistic textures. The in-vivo downstream segmentation results, obtained from nnU-net models trained exclusively on these synthetic images, are detailed in Table 4.4. Results for the ACDC holdout set were consistently lower compared to the MICCAI MSCMR bSSFP sequences. In all cases, the SPADE variants showed improved Dice Scores. When trained on the original in-silico images, we obtained extremely low Dice Scores on both datasets: 0.002 ± 0.005 (LV), 0.001 ± 0.004 (RV), and 0.024 ± 0.051 (Myo) on MSCMR bSSFP, and 0.009 ± 0.041 (LV), 0.004 ± 0.002 (RV), and 0.051 ± 0.099

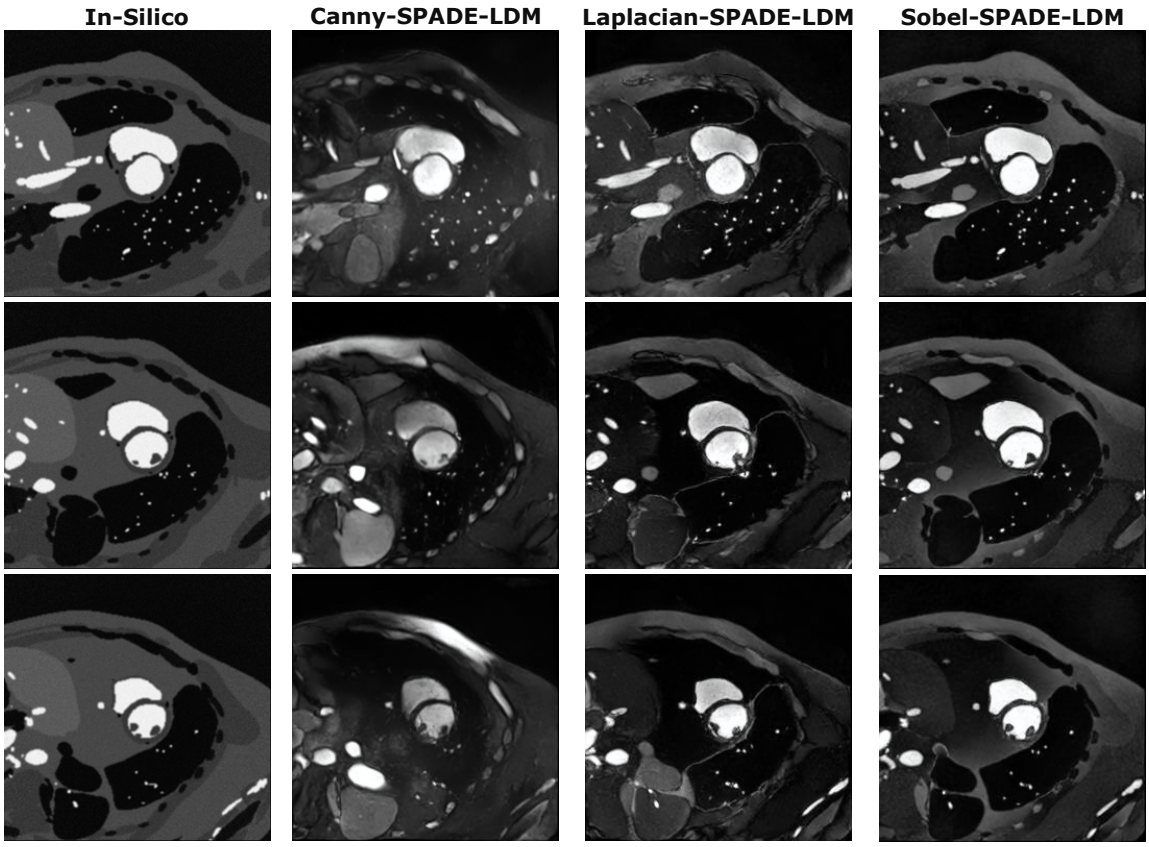

FIGURE 4.10: Style-transferred in-silico images generated via edge-informed SPADE-LDM models.

TABLE 4.4: Segmentation performance on the ACDC holdout set and MICCAI MSCMR bSSFP sequences. Results are evaluated using an nnU-Net exclusively trained on translated in-silico images and measured by Dice Score.

	Dice Score Performance					
	ACDC			MICCAI MSCMR (bSSFP)		
Model $\mathcal{G}$	LV	RV	Myo	LV	RV	Myo
Canny-LDM	0.684 ± 0.180	0.460 ± 0.283	0.581 ± 0.177	0.768 ± 0.105	0.621 ± 0.195	0.600 ± 0.121
Laplacian-LDM	0.124 ± 0.163	0.022 ± 0.053	0.134 ± 0.167	0.413 ± 0.209	0.159 ± 0.154	0.293 ± 0.169
Sobel-LDM	0.072 ± 0.135	0.013 ± 0.043	0.201 ± 0.182	0.096 ± 0.126	0.011 ± 0.050	0.227 ± 0.143
Canny-SPADE-LDM	0.654 ± 0.195	0.490 ± 0.276	0.661 ± 0.198	$\mathbf{0.797 \pm 0.075}$	0.705 ± 0.112	0.657 ± 0.109
Laplacian-SPADE-LDM	0.144 ± 0.164	0.121 ± 0.177	0.105 ± 0.143	0.416 ± 0.215	0.167 ± 0.217	0.327 ± 0.180
Sobel-SPADE-LDM	0.152 ± 0.184	0.034 ± 0.081	0.491 ± 0.231	0.334 ± 0.189	0.040 ± 0.091	0.540 ± 0.177
(4-labels) Canny-SPADE-LDM	$\mathbf{0.712 \pm 0.141}$	$\mathbf{0.508 \pm 0.318}$	$\mathbf{0.693 \pm 0.142}$	0.794 ± 0.055	$\mathbf{0.758 \pm 0.081}$	$\mathbf{0.713 \pm 0.088}$
(4-labels) Laplacian-SPADE-LDM	0.134 ± 0.194	0.091 ± 0.137	0.134 ± 0.194	0.204 ± 0.184	0.049 ± 0.128	0.191 ± 0.159
(4-labels) Sobel-SPADE-LDM	0.105 ± 0.158	0.015 ± 0.042	0.416 ± 0.245	0.261 ± 0.209	0.014 ± 0.053	0.431 ± 0.193
Original In-Silico MRXCAT	0.009 ± 0.041	0.004 ± 0.002	0.051 ± 0.099	0.002 ± 0.005	0.001 ± 0.004	0.024 ± 0.051
ACDC Training Set	0.947 ± 0.031	0.871 ± 0.073	0.893 ± 0.033	0.948 ± 0.017	0.849 ± 0.0366	0.837 ± 0.035

(Myo) on ACDC. These poor results highlight the inherent lack of realism associated with in-silico representations. In line with the qualitative analysis, models trained on style-transferred images with Canny edges achieved superior performance. Both Laplacian and Sobel edge priors also increased the original Dice Scores, but the extent of these improvements was very limited. Our 4-labeled end-to-end framework, which integrates pseudo-labeled lung semantics, further improved the translation performance of Canny-SPADE-LDM, ultimately achieving the highest overall Dice Scores across all segmented regions: 0.794 ± 0.055 (LV), 0.758 ± 0.081 (RV), and 0.713 ± 0.088 (Myo) on MSCMR bSSFP, and 0.712 ± 0.141 (LV), 0.508 ± 0.318 (RV), and 0.683 ± 0.142 (Myo) on ACDC. Figure 4.11 illustrates how this method mitigated hallucinated lung anatomy in images generated by Canny-SPADE-LDM, which appeared as small bright regions in the 3-labeled model. These pseudo-labeled annotations did not improve performance in the Laplacian-SPADE-LDM or Sobel-SPADE-LDM models, where results actually declined.

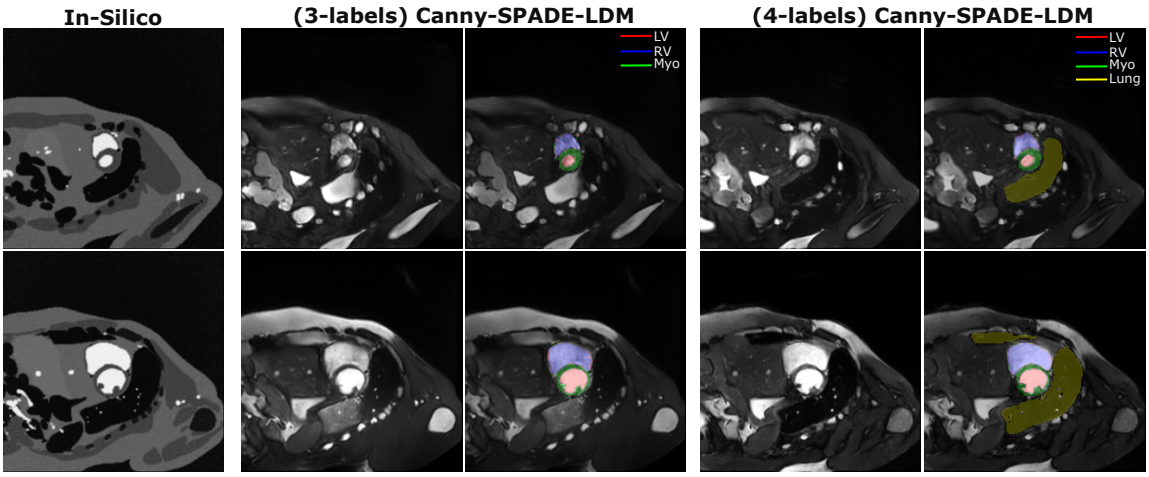

FIGURE 4.11: Comparison of spurious lung regions generated by the baseline 3-labeled model and the 4-labeled Canny-SPADE-LDM model with pseudo annotations.

Chapter 5

Discussion

Based on our results, leveraging edge conditioning priors significantly and consistently improved image generation performance, ensuring anatomical compliance within the full FoV. Our method effectively addressed hallucinated features adjacent to the ROI, resulting in higher fidelity in the generated images compared to approaches that relied exclusively on partial semantics.

In our quantitative analysis, we adhere to standard practices by emphasizing the FID metric, which is widely used to benchmark generative models against existing work. However, it is important to acknowledge the limitations of FID, where it is often considered an ill-posed problem [113]. FID measures the distance between real and generated image distributions in the feature space of a pretrained network, optimized for natural images and trained on the ImageNet dataset. The applicability of FID for medical imaging remains an open topic [114], as medical images possess distinct characteristics that may not be adequately captured by this metric. Therefore, it is important not to underestimate the significance of qualitative analysis, as well as SSIM and NMSE metrics.

Our LDM-based models markedly outperformed SPADE-GAN variants, highlighting the potential of diffusion-based approaches to surpass the predominantly employed GAN-based models in medical image synthesis. However, when comparing the energy distribution in Fourier spectra, low and mid-frequencies in SPADE-GANs were closest to the GT, both with and without edge conditioning. This implies that SPADE-GAN might be more adept at reproducing subtle features that are not easily discernible through qualitative analysis, even though some anatomical structures outside the ROI visibly deviate from the GT. Moreover, when $e_x = \emptyset$, SPADE-GAN tends to outperform SPADE-LDM in regions where semantic information is available, as evidenced by the examples in Figure

4.1 and the notably higher SSIM in Table 4.2. This suggests that SPADE might be better suited for GANs than for diffusion-based models in semantic image synthesis tasks. Nevertheless, LDMs demonstrated superior overall performance, as indicated by their consistently lower FID scores. Additionally, their high-frequency energy distribution better matches the GT, reflecting finer details and improved anatomical representations.

Models conditioned solely on edge map priors exhibited image generation performance similar to their counterparts that also included semantics. This highlights that edge conditioning alone is sufficient to generate high-quality images within the ROI without relying on paired label annotations. However, the performance varies depending on the choice of edge detector operator. Sobel edges yielded the best results, with models producing images that were nearly indistinguishable from the GT. Laplacian edges followed closely with slightly inferior performance, while Canny edges performed subpar. We hypothesize that this performance disparity correlates with the intrinsic characteristics of the edge information provided by each operator. Sobel edges impose a robust anatomical constraint by identifying tissue boundaries and quantifying the strength of transitions based on intensity contrasts. A low gradient magnitude suggests a gradual transition, indicating similar tissue types or a less distinct boundary (e.g., fat vs. muscle). Conversely, a high gradient magnitude often corresponds to a sharp transition between two different tissues or structures, reflecting significant differences in tissue density or composition. The Laplacian operator can provide similar information but relies on second-order derivatives, which requires strong denoising tools to be viable and diminishes much of the rate-of-change intensity information. Such information is absent in Canny edge maps, as they are thin, binary representations that provide only the most basic form of edge-conditioning guidance.

Through image-to-image translation, we were able to achieve accurate automated segmentation on multiple imaging modalities while training a model exclusively on a single imaging scheme. Our style-transfer techniques led to marked improvements in performance on T_2-weighted and T_2-mapping images, with slight improvements on T_1-weighted LGE and T_1-mapping when compared to direct segmentation on the original images. These improvements are more pronounced in T_2 parametrizations, given the dark contrasts within the ROI caused by fast-flowing blood, which we translate into the characteristic bright-blood appearance from bSSFP sequences. In contrast, T_1 parametrization is less sensitive to spin dephasing from rapid motion and naturally produces brighter

contrasts compared to T_2, enabling the nnU-Net to perform well from the outset. Interestingly, Canny edge priors produced the best results, while Sobel edges performed the worst. This contrasts with our findings from image generation experiments. We believe this discrepancy is also attributed to the edge information used for conditioning. Sobel edges may limit flexibility in style-transfer tasks, as gradient magnitudes can vary significantly across different modalities compared to the contrasts seen during training. On the other hand, Canny edges, with their simplicity in binary maps, rely only on similar anatomical structures to enable one-to-one translation.

Models trained on synthetic data derived from Canny-SPADE-LDM style-transferred in-silico images achieved highly competitive Dice Scores on the MICCAI MSCMR dataset, comparable to those trained on in-vivo data. These findings highlight the versatility of our medical imaging synthesis method, which is capable of translating readily available physically simulated data into realistic CMR images. Despite the noteworthy results on the MSCMR bSSFP sequences, the performance gap observed with the ACDC dataset remains unclear and requires further considerations. Preliminary visual analysis suggests that this disparity in performance may be due to differences in the distribution of axial slices in the generated phantoms, which seem to align more closely with the MICCAI MSCMR dataset than with ACDC. Our in vivo pseudo-labeled lung semantic annotations, generated from XCAT digital phantoms, effectively addressed ambiguities in the lung regions of Canny edge maps and reduced the occurrence of hallucinated features. Given the lungs' anatomical proximity to the ROI and their contribution to spatial context, this ultimately resulted in improved segmentation performance. However, we recall the limitations of extending this approach to other organs, owing to the disparity in textures and structures observed in in-silico compared to in-vivo CMR images. Additionally, we emphasize one of the successes of the Canny-LDM models for this task, which is their ability to translate, to some extent, the inherent noise associated with MRI scans. We acknowledge that introducing varying degrees of noise to the generated image data might further improve our results. We leave this for future work.

Chapter 6

Conclusions

In this work, we introduced a novel LDM-based method for medical image synthesis, leveraging both labeled annotations and edge maps for style-transfer tasks. Our experiments demonstrate that our method effectively complements partial semantics and imposes a full anatomical constraint without requiring exhaustive additional annotations. We significantly improved image generation performance and produced more realistic cardiac MRI images with reduced hallucinated anatomical regions. Furthermore, our results highlight the potential of state-of-the-art diffusion-based approaches to outperform commonly employed GAN-based models in medical image synthesis. Our framework is computationally efficient, which is paramount for democratizing diffusion models, ensuring reproducibility, and facilitating future work.

Using image-to-image translation, our method enabled accurate segmentation across different imaging modalities while relying on a model trained on a single imaging scheme. This alleviates challenges posed by data scarcity in learning-based models, which typically require multi-sequence acquisitions in the training set to achieve generalization in automated segmentation tasks. Additionally, we extended our style-transfer techniques to translate readily available in-silico data into realistic CMR images. Our end-to-end framework for downstream segmentation tasks, using style-transferred numerically simulated data, achieved highly competitive Dice scores on the MICCAI MSCMR bSSFP sequences, comparable to models trained on in-vivo data, while relying exclusively on synthetic images. This demonstrates our method as a powerful tool for data augmentation.

Throughout our experiments, we conducted an extensive analysis of various image edge operators. Our findings indicate that Canny edge detection, with its simplicity in binary edge maps, proves to be highly effective in style-transfer tasks but is subpar in

image generation. In contrast, more complex operators like Sobel and Laplacian excel in image generation but are less adaptable due to varying gradient-based contrasts, limiting their ability to translate into the style representations learned during training.

Appendix A

MIDL 2024 Conference: Short Paper Submission

Anatomy-compliant medical image synthesis by latent diffusion models

Nuno Capitão[1] N.M.nunomiguelFerreiraCapitao@tudelft.nl
Yidong Zhao[1] Y.Zhao-8@tudelft.nl
Yi Zhang[1] Y.Zhang-43@tudelft.nl
Nathan Geerts[1] N.Geerts@student.tudelft.nl
João V. Lopes[2] jlopes@fc.up.pt
Qian Tao*[1] Q.Tao@tudelft.nl
[1] *Delft University of Technology, Lorentzweg 1, Delft, The Netherlands*
[2] *Department of Physics and Astronomy, Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal*

Abstract

Data scarcity presents a significant challenge for achieving optimal model performance in medical imaging, due to the limited availability of high-quality data. One potential solution to address this issue is to synthesize medical images using powerful generative models with conditioning prior. However, obtaining full anatomical annotations of all organs for anatomical conditioning is impractical, resulting in synthetic images with incoherent or hallucinated anatomy. In this paper, we propose an innovative medical image generation method based on state-of-the-art latent diffusion models (LDM). To tackle the anatomy compliance challenge, we leverage both the anatomical mask, which is specific to the organ of interest, and the edge information, which is general and easy to compute in the full field of view (FOV), as dual conditioning. Our method does not require extra annotations to achieve anatomy compliance. Our method was evaluated on the ACDC dataset and compared with GAN baselines. Results demonstrate that incorporating edge-based conditioning strongly complements image semantics, leading to high-quality, anatomy-compliant medical image generation.

Keywords: Latent diffusion models (LDM), Medical image synthesis, Edge.

1. Introduction

Learning-based approaches have gained prominence in many medical image processing tasks (Zhou et al., 2021). However, they rely heavily on high-quality labeled data, which is frequently unavailable given the high acquisition costs, rare pathologies, privacy concerns, and lack of annotation expertise (Castro et al., 2020). Generative AI offers a promising solution to the data scarcity problem (Dorjsembe et al., 2024). Previous medical image synthesis studies predominantly leveraged the SPADE (Park et al., 2019) framework which translates the semantic map to an image with Generative Adversarial Networks (GAN) (Goodfellow et al., 2014). Built upon SPADE, (Abbasi-Sureshjani et al., 2020) proposed generating cardiac magnetic resonance (CMR) images given heart segmentation maps. However, GAN-based generation can have degenerated quality (Müller-Franzes et al., 2023; Skandarani et al., 2021) and has recently been outperformed by diffusion-based generation (Ho et al., 2020; Dhariwal and Nichol, 2021; Rombach et al., 2022). Additionally and importantly, it

© 2024 CC-BY 4.0, N. Capitão, Y. Zhao, Y. Zhang, N. Geerts, J.V. Lopes & Q. Tao.

is often overlooked that the available semantic maps usually only cover specific organs (i.e. heart) (Abbasi-Sureshjani et al., 2020), leading to spurious and hallucinated generation of surrounding organs in the thorax and abdomen (i.e. lung, liver).

Image edges can delineate anatomical structures with little computational effort and can be used to guide image generation with GANs (Luo et al., 2021; Yu et al., 2019). Nevertheless, its use has not been explored in state-of-the-art diffusion models which promise superior quality to GANs. In this paper, we propose a novel LDM-based generative model conditioned on both incomplete organ annotations and image edges in the full field of view as dual anatomy conditioning. Our method ensures anatomical compliance in medical image generation, which has been understudied but is important for downstream tasks such as training AI models or educating radiologists (Skandarani et al., 2023).

2. Methods and Materials

We propose an LDM-based conditional generator $\mathcal{G}$ (Rombach et al., 2022) to translate anatomy into realistic images, which performs the diffusion process in an auto-encoded (AE) latent space. The anatomy is modeled by combining the partly labeled tissue map y and the image edge e_x. The edge e of an image x is retrieved by Canny detector (Canny, 2009). The conditions are integrated into $\mathcal{G}$ through a binary boundary map derived from the image gradients, concatenated with a one-hot tensor representation of the semantic labels map, as represented in Figure 1.

We trained our models using the publicly available Automated Cardiac Diagnosis Challenge (ACDC) dataset (Bernard et al., 2018), and compared the proposed model to GAN-based conditional generation with SPADE (Park et al., 2019). To validate the utility of the edge as an additional condition, we assessed both models' performance with and without edge guidance, and evaluate quantitatively using metrics of FID, SSIM, NMSE. Additionally, we compared the Fourier space of generated images by different models.

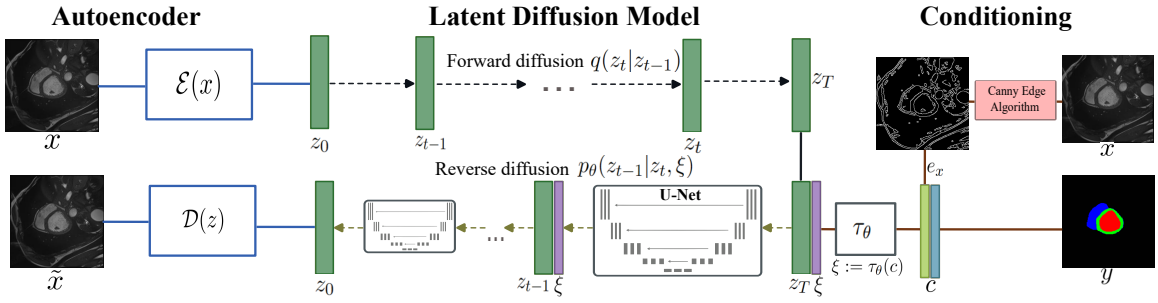

Figure 1: Illustration of the LDM-based edge conditioning pipeline.

3. Results and Conclusion

The quantitative evaluation results are reported in Table 1. The proposed LDM-based generation achieved the best performance in terms of FID (61.9/98.1) compared to the best SPADE variant (101.9/161.4), with or without edge guidance. Our LDM has a lower SSIM

than the SPADE-based model without edge guidance. However, the edge-guided LDM achieved an SSIM of 0.608 which is much higher than the SPADE variant (0.556). Using edge guidance also significantly improved the generation quality for SPADE variants.

Figure 2 presents some examples of generated CMR images and their Fourier frequency spectra. For LDM, we observe that the generated images are closest to the ground truth and correctly delineate organ details that the other models fail to capture. Nevertheless, we note that all edge-informed models show similar improvement. Moreover, we observe that for the edge-informed models, LDM aligns closely with the ground truth in both low and high-frequency ranges, whereas SPADE models only achieve closer alignment in higher frequencies.

Table 1: Image generation performance measured by FID, SSIM and NMSE.

	w/ edge conditioning			w/o edge conditioning		
Method	FID ↓	SSIM ↑	NMSE ↓	FID ↓	SSIM ↑	NMSE ↓
LDM	61.9	0.608 ± 0.123	0.071 ± 0.035	98.1	0.283 ± 0.094	0.136 ± 0.080
SPADE	109.8	0.556 ± 0.119	0.117 ± 0.041	160.3	0.328 ± 0.093	0.132 ± 0.077
SPADE w/ VAE	101.9	0.556 ± 0.128	0.104 ± 0.034	161.4	0.338 ± 0.090	0.145 ± 0.093

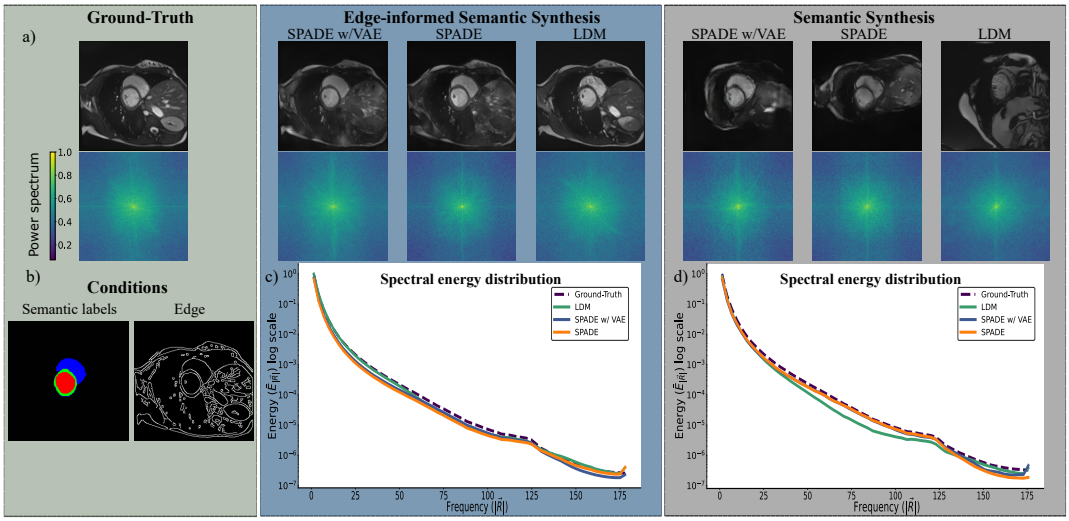

Figure 2: Qualitative results: (a) The ground truth and its Fourier spectrum. (b) Left: Segmentation map. Right: Edge anatomical boundaries by the Canny operator. (c),(d) Comparison of the average model spectra as a function of $|\overrightarrow{\mathbf{R}}| = \sqrt{x^2 + y^2}$, representing a radius distance from the 2D grid origin.

In conclusion, we proposed a novel medical image generation method using LDM with dual conditioning by both semantic labels and Canny edges. Our experiments showed that our method improved image synthesis quality and anatomy compliance, without requiring additional exhaustive annotations. Furthermore, we have shown that LDM outperforms GAN-based models in both the image and frequency space.

References

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 Tal Arbel, Ismail Ben Ayed, Marleen de Bruijne, Maxime Descoteaux, Herve Lombaert, and Christopher Pal, editors, *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, number 121 in Proceedings of Machine Learning Research, pages 6–18. PMLR, February 2020. Third Conference on Medical Imaging with Deep Learning ; Conference date: 06-07-2020 Through 08-07-2020.

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, Gerard Sanroma, Sandy Napel, Steffen Petersen, Georgios Tziritas, Elias Grinias, Mahendra Khened, Varghese Alex Kollerathu, Ganapathy Krishnamurthi, Marc-Michel Rohe, Xavier Pennec, Maxime Sermesant, Fabian Isensee, Paul Jager, Klaus H Maier-Hein, Peter M Full, Ivo Wolf, Sandy Engelhardt, Christian F Baumgartner, Lisa M Koch, Jelmer M Wolterink, Ivana Isgum, Yeonggul Jang, Yoonmi Hong, Jay Patravali, Shubham Jain, Olivier Humbert, and Pierre-Marc Jodoin. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Trans. Med. Imaging*, 37(11):2514–2525, November 2018.

John F. Canny. Canny edge detection. 2009. URL https://api.semanticscholar.org/CorpusID:47899497.

Daniel C. Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):3673, Jul 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17478-w. URL https://doi.org/10.1038/s41467-020-17478-w.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021.

Zolnamar Dorjsembe, Hsing-Kuo Pao, Sodtavilan Odonchimed, and Furen Xiao. Conditional diffusion models for semantic 3d brain mri synthesis, 2024.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

Yanmei Luo, Dong Nie, Bo Zhan, Zhiang Li, Xi Wu, Jiliu Zhou, Yan Wang, and Dinggang Shen. Edge-preserving mri image synthesis via adversarial network with iterative multi-scale fusion. *Neurocomputing*, 452:63–77, 2021. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2021.04.060. URL https://www.sciencedirect.com/science/article/pii/S0925231221006020.

Gustav Müller-Franzes, Jan Moritz Niehues, Firas Khader, Soroosh Tayebi Arasteh, Christoph Haarburger, Christiane Kuhl, Tianci Wang, Tianyu Han, Teresa Nolte, Sven Nebelung, et al. A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. *Scientific Reports*, 13(1):12098, 2023.

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization, 2019.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

Youssef Skandarani, Pierre-Marc Jodoin, and Alain Lalande. Gans for medical image synthesis: An empirical study, 2021.

Youssef Skandarani, Pierre-Marc Jodoin, and Alain Lalande. Gans for medical image synthesis: An empirical study. *Journal of Imaging*, 9(3):69, March 2023. ISSN 2313-433X. doi: 10.3390/jimaging9030069.

Biting Yu, Luping Zhou, Lei Wang, Yinghuan Shi, Jurgen Fripp, and Pierrick Bourgeat. Ea-gans: Edge-aware generative adversarial networks for cross-modality mr image synthesis. *IEEE Transactions on Medical Imaging*, 38(7):1750–1762, 2019. doi: 10.1109/TMI.2019.2895894.

S. Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S. Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L. Prince, Daniel Rueckert, and Ronald M. Summers. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, 109(5): 820–838, 2021. doi: 10.1109/JPROC.2021.3054390.

Appendix B

Training Details

B.1 SPADE-GAN

The original SPADE-GAN work [18] builds upon the Pix2PixHD framework [84] by introducing a generator that disentangles style and semantics. This approach enables controlled image synthesis by integrating SPADE normalization within the generator's upsampling layers. The architecture used in this work is presented in Figure B.1, and all the hyperparameters considered are shown in Table B.1.

TABLE B.1: Hyperparameter settings for VAE-based SPADE-GAN training. The baseline architecture removes all the $\mathcal{E}$ layers and uses a random vector drawn from a Gaussian distribution as the generator's input.

Hyperparameters	**Values**
Channels ($\mathcal{D}$)	64
Number of Layers ($\mathcal{D}$)	4
Channel Multiplier ($\mathcal{D}$)	1-2-4-8
Channels ($\mathcal{G}$)	64
Number of Upsampling Layers ($\mathcal{G}$)	4
Channel Multiplier ($\mathcal{G}$)	16-8-4-2
Depth ($\mathcal{G}$)	2
Scaling Factor λ ($\mathcal{G}$)	5
Channels ($\mathcal{E}$)	16
Number of Layers ($\mathcal{E}$)	5
Channel Multiplier ($\mathcal{E}$)	1-2-4-8-8
Dimension of Latent Vector z	256
Batch size	32
Epochs	100
Learning Rate	2×10^{-4}
Learning Rate Exponential Decay	0.99

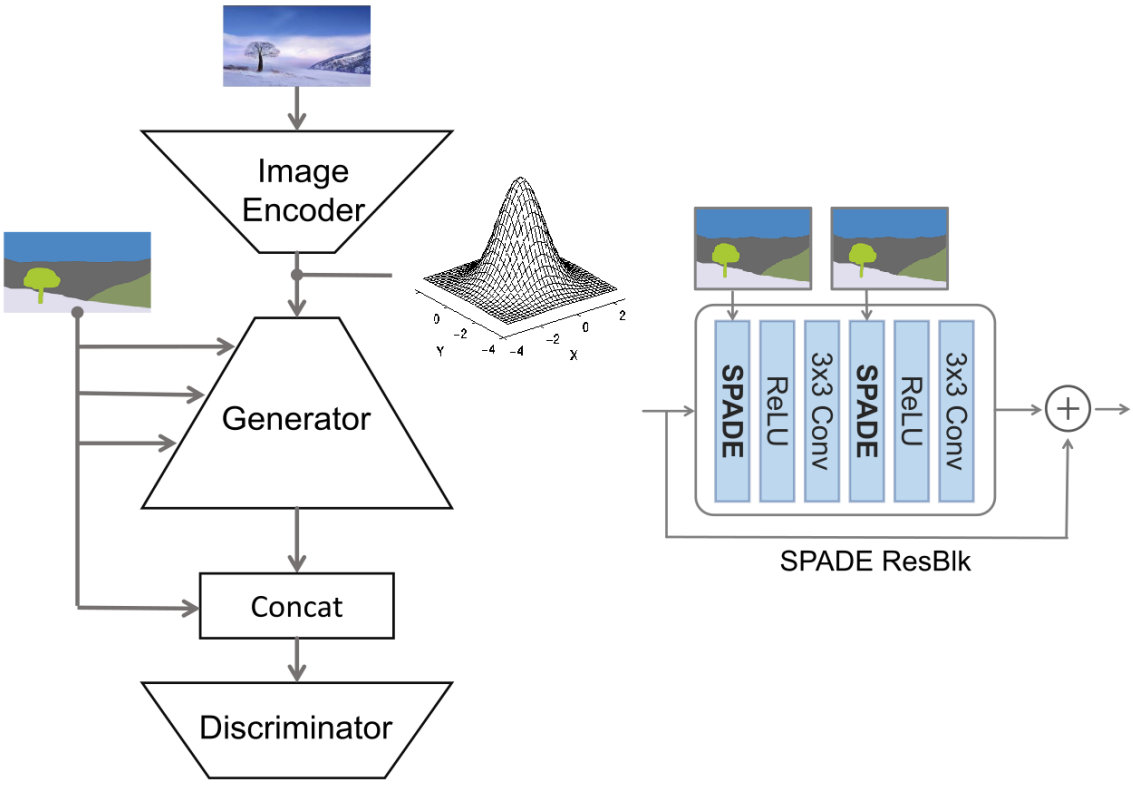

FIGURE B.1: (**Left**) VAE-based SPADE-GAN architecture, where the generator's input is a latent representation encoded from the real image. (**Right**) SPADE Residual Block used within $\mathcal{G}$ layers to integrate semantics. Image adapted from [18].

The model $\mathcal{G}$ is trained with a multi-scale discriminator based on PatchGAN [115]. Instead of evaluating the entire image, the discriminator's final layer uses a convolutional layer to classify each 4×4 patch of the image as real or fake. The multi-scale approach applies distinct discriminators at different resolutions to improve overall performance. We employed two discriminators: one operates at the original resolution, while the other processes a version downsampled by $f = 2$ (i.e., 128×128 px). The final loss is calculated by averaging the losses from both discriminators. The final training objective follows the principles described in Section 2.2.2 but employs a Hinge loss format to impose a stronger penalty for incorrect predictions, mitigating issues such as vanishing gradients and mode collapse. Moreover, an L_1 loss term is included in $\mathcal{G}$ to improve low-frequency details. The loss functions are defined below:

$$L_{\mathcal{G}} := -\mathbb{E}_{z\sim p_z, y\sim p_{\text{data}}}[\mathcal{D}(\mathcal{G}(z,y))] + \lambda \mathbb{E}_{z\sim p_z,(x,y)\sim p_{\text{data}}}[\|\mathcal{G}(z,y) - x\|_1] \tag{B.1}$$

$$L_{\mathcal{D}} := -\mathbb{E}_{(x,y)\sim p_{\text{data}}}[\min(0, -1 + \mathcal{D}(x,y))] - \mathbb{E}_{z\sim p_z, y\sim p_{\text{data}}}[\min(0, -1 - \mathcal{D}(\mathcal{G}(z,y)))] \tag{B.2}$$

where λ is a scaling factor for the L_1 term. The discriminator in Equation B.2 takes the input image concatenated with the semantic layout. In a VAE setup, training also includes a *KLD* loss:

$$L_{KLD} := \mathbb{E}_{x \sim p_{\text{data}}} \left[D_{KL}(q(z|x) \parallel p(z)) \right], \tag{B.3}$$

where the approximate posterior $q(z|x)$ is parameterized by the mean and variance vectors determined via the reparameterization trick in [48], and $p(z)$ is a Gaussian prior.

B.2 VQ-AE

An AE is a type of neural network that compresses data into a latent representation using an encoder ($\mathcal{E}$) and reconstructs it through a decoder ($\mathcal{D}$). For image-like data, AEs efficiently learn the underlying manifold, enabling diffusion modeling in a low-dimensional space by encoding images into the latent space via $\mathcal{E}$ and reconstructing them with $\mathcal{D}$ after the diffusion process.

The AE is trained using a combination of a perceptual loss [116] and a patch-based adversarial objective [117]. The adversarial objective involves a discriminator, D_ψ, which is optimized to differentiate between original images and reconstructions, $\mathcal{D}(\mathcal{E}(x))$. This ensures that the reconstructions remain confined to the image manifold by enforcing local realism and avoiding the blurriness that can result from relying solely on pixel-space losses, such as L_2 or L_1 objectives. The hyperparameters used for VQ-AE training are listed in Table B.2.

To avoid arbitrarily high-variance latent spaces, the latent variable z is regularized to have small variance through an additional loss term, L_{reg}. We use a VQ layer to regularize the latent space by learning a high-dimensional codebook $|\mathcal{Z}|$. The VQ layer maps continuous latent representations to discrete points in a learned codebook, which contains a finite set of embedding vectors, each representing a specific feature of the latent space. During the forward pass, each input latent vector is replaced by the closest embedding vector from the codebook, quantizing the continuous latent vector into a discrete codebook entry. The discretization process introduces regularization by reducing variability in the latent space. The full objective to train the autoencoder model $(\mathcal{E}, \mathcal{D})$ is defined below:

$$L_{AE} := \min_{\mathcal{E},\mathcal{D}} \max_{\psi} \Big(L_{rec}(x, \mathcal{D}(\mathcal{E}(x))) - L_{adv}(\mathcal{D}(\mathcal{E}(x))) + \log D_{\psi}(x) + L_{reg}(x; \mathcal{E}, \mathcal{D}) \Big). \quad \text{(B.4)}$$

TABLE B.2: Hyperparameter settings for VQ-AE training.

Hyperparameters	**Values**
Downsampling Factor (f)	4
Channels	128
Number of Layers	3
Channel Multiplier	1-2-4
Depth	2
Dimension of Codebook $\lvert\mathcal{Z}\rvert$	8192
Dimension of Latent Vector z	$64 \times 64 \times 3$
Batch size	16
Epochs	1000
Learning Rate	1×10^{-5}
Learning Rate Exponential Decay	0.999

B.3 LDM

Table B.3 lists the hyperparameters used to train our LDMs. All models were optimized with Exponential Moving Average (EMA) [118].

TABLE B.3: Hyperparameter settings for LDM training.

Hyperparameters	**Values**
Noise Schedule	linear
Diffusion Steps	1000
Training Diffusion Sampling (DDIM)	200
Channels	128
Number of Layers	3
Channel Multiplier	1-4-8
Depth	2
Channels (τ_θ)	64
Number of Layers (τ_θ)	3
Channel Multiplier (τ_θ)	1-1-2
Batch size	16
Epochs	100
Learning Rate	1×10^{-4}
Learning Rate Exponential Decay	0.95

Bibliography

[1] A. Barragán-Montero, U. Javaid, G. Valdés, D. Nguyen, P. Desbordes, B. Macq, S. Willems, L. Vandewinckele, M. Holmström, F. Löfman, S. Michiels, K. Souris, E. Sterpin, and J. A. Lee, "Artificial intelligence and machine learning for medical imaging: A technology review," *Physica Medica*, vol. 83, pp. 242–256, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1120179721001733 [Cited on page 1.]

[2] R.-M. Menchón-Lara, F. Simmross-Wattenberg, P. Casaseca-de-la Higuera, M. Martín-Fernández, and C. Alberola-López, "Reconstruction techniques for cardiac cine MRI," *Insights Imaging*, vol. 10, no. 1, p. 100, Sep. 2019. [Cited on page 1.]

[3] A. J. Taylor, M. Salerno, R. Dharmakumar, and M. Jerosch-Herold, "T1 mapping," *JACC: Cardiovascular Imaging*, vol. 9, no. 1, pp. 67–81, 2016. [Online]. Available: https://www.jacc.org/doi/abs/10.1016/j.jcmg.2015.11.005 [Cited on page 2.]

[4] G. D. Aquaro, C. De Gori, L. Faggioni, M. L. Parisella, D. Cioni, R. Lencioni, and E. Neri, "Diagnostic and prognostic role of late gadolinium enhancement in cardiomyopathies," *European Heart Journal Supplements*, vol. 25, no. Supplement C, pp. C130–C136, 04 2023. [Online]. Available: https://doi.org/10.1093/eurheartjsupp/suad015 [Cited on page 2.]

[5] Y. Amano, M. Tachi, H. Tani, K. Mizuno, Y. Kobayashi, and S. Kumita, "T2-weighted cardiac magnetic resonance imaging of edema in myocardial diseases," *scientificWorldJournal*, vol. 2012, p. 194069, sep 2012. [Cited on page 2.]

[6] N. Townsend, D. Kazakiewicz, F. Lucy Wright, A. Timmis, R. Huculeci, A. Torbica, C. P. Gale, S. Achenbach, F. Weidinger, and P. Vardas, "Epidemiology of

cardiovascular disease in europe," *Nature Reviews Cardiology*, vol. 19, no. 2, pp. 133–143, Feb 2022. [Online]. Available: https://doi.org/10.1038/s41569-021-00607-3 [Cited on page 2.]

[7] J. Chai, H. Zeng, A. Li, and E. W. Ngai, "Deep learning in computer vision: A critical review of emerging techniques and application scenarios," *Machine Learning with Applications*, vol. 6, p. 100134, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666827021000670 [Cited on page 3.]

[8] D. C. Castro, I. Walker, and B. Glocker, "Causality matters in medical imaging," *Nature Communications*, vol. 11, no. 1, p. 3673, Jul 2020. [Online]. Available: https://doi.org/10.1038/s41467-020-17478-w [Cited on page 3.]

[9] F. Khashami, *Bloch Equation Description with Relaxation Time*. Cham: Springer Nature Switzerland, 2024, pp. 71–89. [Online]. Available: https://doi.org/10.1007/978-3-031-47976-2_4 [Cited on pages 3, 11, and 13.]

[10] W. Paul Segars and B. M. W. Tsui, "MCAT to XCAT: The evolution of 4-D computerized phantoms for imaging research: Computer models that take account of body movements promise to provide evaluation and improvement of medical imaging devices and technology," *Proc. IEEE Inst. Electr. Electron. Eng.*, vol. 97, no. 12, pp. 1954–1968, Dec. 2009. [Cited on page 3.]

[11] T. M. Ngo, G. S. K. Fung, B. M. W. Tsui, E. R. McVeigh, and D. A. Herzka, "Three dimensional digital polyhedral phantom framework with analytical fourier transform and application in cardiac imaging," 2010. [Online]. Available: https://api.semanticscholar.org/CorpusID:41398132 [Cited on page 3.]

[12] S. Aja-Fernández, L. Cordero-Grande, and C. Albewla-López, "A mri phantom for cardiac perfusion simulation," in *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*, 2012, pp. 638–641. [Cited on page 3.]

[13] S. Buoso, T. Joyce, N. Schulthess, and S. Kozerke, "Mrxcat2.0: Synthesis of realistic numerical phantoms by combining left-ventricular shape learning, biophysical simulations and tissue texture generation," *Journal of Cardiovascular Magnetic Resonance*, vol. 25, no. 1, p. 25, 2023. [Online]. Available: https://

www.sciencedirect.com/science/article/pii/S109766472300025X [Cited on pages 4 and 44.]

[14] E. Zurich, "MRXCAT: An open-source framework for MRI simulations," https://biomed.ee.ethz.ch/mrxcat.html, Accessed: 2024-07-05. [Cited on pages 4 and 44.]

[15] C. G. Xanthis, D. Filos, K. Haris, and A. H. Aletras, "Simulator-generated training datasets as an alternative to using patient data for machine learning: An example in myocardial segmentation with mri," *Computer Methods and Programs in Biomedicine*, vol. 198, p. 105817, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260720316503 [Cited on page 4.]

[16] Z. Dorjsembe, H.-K. Pao, S. Odonchimed, and F. Xiao, "Conditional diffusion models for semantic 3d brain mri synthesis," *IEEE Journal of Biomedical and Health Informatics*, p. 1–10, 2024. [Online]. Available: http://dx.doi.org/10.1109/JBHI.2024.3385504 [Cited on page 4.]

[17] S. Abbasi-Sureshjani, S. Amirrajab, C. Lorenz, J. Weese, J. P. W. Pluim, and M. M. Breeuwer, "4d semantic cardiac magnetic resonance image synthesis on xcat anatomical model," in *International Conference on Medical Imaging with Deep Learning*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:211132564 [Cited on pages 4 and 6.]

[18] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," 2019. [Cited on pages 4, 35, 36, 37, 71, and 72.]

[19] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014. [Cited on pages 4, 19, and 20.]

[20] G. Müller-Franzes, J. M. Niehues, F. Khader, S. T. Arasteh, C. Haarburger, C. Kuhl, T. Wang, T. Han, T. Nolte, S. Nebelung, J. N. Kather, and D. Truhn, "A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis," *Sci. Rep.* [Cited on page 4.]

[21] Y. Skandarani, P.-M. Jodoin, and A. Lalande, "Gans for medical image synthesis: An empirical study," 2021. [Cited on page 4.]

[22] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *CoRR*, vol. abs/2006.11239, 2020. [Online]. Available: https://arxiv.org/abs/2006.11239 [Cited on pages 4, 5, 20, 24, 26, 28, and 29.]

[23] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," 2021. [Cited on page 20.]

[24] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2022. [Cited on pages 4, 5, 33, 34, and 37.]

[25] A. Kazerouni, E. K. Aghdam, M. Heidari, R. Azad, M. Fayyaz, I. Hacihaliloglu, and D. Merhof, "Diffusion models in medical imaging: A comprehensive survey," *Medical Image Analysis*, vol. 88, p. 102846, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841523001068 [Cited on page 5.]

[26] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," 2022. [Cited on pages 5, 29, 31, and 32.]

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015. [Online]. Available: https://arxiv.org/abs/1505.04597 [Cited on pages 5 and 28.]

[28] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention u-net: Learning where to look for the pancreas," 2018. [Online]. Available: https://arxiv.org/abs/1804.03999 [Cited on pages 5 and 28.]

[29] Y. Luo, D. Nie, B. Zhan, Z. Li, X. Wu, J. Zhou, Y. Wang, and D. Shen, "Edge-preserving mri image synthesis via adversarial network with iterative multi-scale fusion," *Neurocomputing*, vol. 452, pp. 63–77, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231221006020 [Cited on page 5.]

[30] B. Yu, L. Zhou, L. Wang, Y. Shi, J. Fripp, and P. Bourgeat, "Ea-GANs: Edge-aware generative adversarial networks for cross-modality MR image synthesis," *IEEE Trans. Med. Imaging*, vol. 38, no. 7, pp. 1750–1762, Jul. 2019. [Cited on page 5.]

[31] S. Amirrajab, S. Abbasi-Sureshjani, Y. A. Khalil, C. Lorenz, J. Weese, J. Pluim, and M. Breeuwer, "Xcat-gan for synthesizing 3d consistent labeled cardiac mr images on anatomically variable xcat phantoms," 2020. [Online]. Available: https://arxiv.org/abs/2007.13408 [Cited on pages 6, 45, and 46.]

[32] B. Widom, *The Boltzmann distribution law and statistical thermodynamics*. Cambridge University Press, 2002, p. 1–15. [Cited on page 8.]

[33] R. Y. Kwong, M. Jerosch-Herold, and B. Heydari, Eds., *Cardiovascular Magnetic Resonance Imaging*, 2nd ed., ser. Contemporary Cardiology. Springer-Verlag New York, LLC, 2019. [Online]. Available: https://link.springer.com/book/10.1007/978-1-4939-8841-9 [Cited on page 8.]

[34] S. Healthineers, "Magnets, spins, and resonances: An introduction to the basics of magnetic resonance," 2024, accessed: 2024-07-24. [Online]. Available: https://www.magnetomworld.siemens-healthineers.com/publications/mr-basics [Cited on pages 9, 12, and 14.]

[35] M. A. Syed, S. V. Raman, and O. P. Simonetti, Eds., *Basic Principles of Cardiovascular MRI: Physics and Imaging Techniques*, 1st ed., ser. Medicine, Medicine (R0). Springer International Publishing Switzerland: Springer Cham, 2015, 97 b/w illustrations, 142 illustrations in colour. [Online]. Available: https://doi.org/10.1007/978-3-319-22141-0 [Cited on pages 10, 12, 14, and 17.]

[36] J. C. Ye, "Compressed sensing mri: a review from signal processing perspective," *BMC Biomedical Engineering*, vol. 1, no. 1, p. 8, Mar 2019. [Online]. Available: https://doi.org/10.1186/s42490-019-0006-z [Cited on page 11.]

[37] M. A. Yassin, A. T. Soliman, V. Desanctis, S. Abusamaan, A. Elsotouhy, and N. Aldewik, "Hereditary Hemochromatosis in an Adult Due to H63D Mutation: The Value of Estimating Iron Deposition By MRI T2* and Dissociation Between Serum Ferritin Concentration and Hepatic Iron Overload," *Blood*, vol. 124, no. 21, pp. 4891–4891, 12 2014. [Online]. Available: https://doi.org/10.1182/blood.V124.21.4891.4891 [Cited on page 15.]

[38] N. Martini, A. Meloni, V. Positano, D. D. Latta, P. Keilberg, L. Pistoia, A. Spasiano, T. Casini, A. Barone, A. Massa, A. Ripoli, and F. Cademartiri, "Fully automated regional analysis of myocardial t2* values for iron quantification

using deep learning," *Electronics*, vol. 11, no. 17, 2022. [Online]. Available: https://www.mdpi.com/2079-9292/11/17/2749 [Cited on page 15.]

[39] K. Scheffler and S. Lehnhardt, "Principles and applications of balanced ssfp techniques," *European Radiology*, vol. 13, no. 11, pp. 2409–2418, Nov 2003. [Online]. Available: https://doi.org/10.1007/s00330-003-1957-x [Cited on page 15.]

[40] E. Haacke, R. Brown, M. Thompson, and R. Venkatesan, *Magnetic resonance imaging: physical principles and sequence design*. St. Louis: Mosby, 1999. [Cited on page 16.]

[41] M. A. Bernstein, K. F. King, and X. J. Zhou, "Table of constants and conversion factors," in *Handbook of MRI Pulse Sequences*. Amsterdam: Elsevier, 2004, p. 961. [Cited on pages 16 and 45.]

[42] Y. Lu and J. Lu, "A universal approximation theorem of deep neural networks for expressing probability distributions," 2020. [Online]. Available: https://arxiv.org/abs/2004.08867 [Cited on page 18.]

[43] G. E. Hinton, "Deep belief networks," *Scholarpedia*, vol. 4, p. 5947, 2009. [Online]. Available: https://api.semanticscholar.org/CorpusID:7905652 [Cited on page 19.]

[44] G. Montufar, "Restricted boltzmann machines: Introduction and review," 2018. [Online]. Available: https://arxiv.org/abs/1806.07066 [Cited on page 19.]

[45] I. E. Fellows, "Why (and when and how) contrastive divergence works," 2014. [Online]. Available: https://arxiv.org/abs/1405.0602 [Cited on page 19.]

[46] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, "Handwritten digit recognition with a back-propagation network," in *Advances in Neural Information Processing Systems*, D. Touretzky, Ed., vol. 2. Morgan-Kaufmann, 1989. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/1989/file/53c3bce66e43be4f209556518c2fcb54-Paper.pdf [Cited on page 19.]

[47] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct 1986. [Online]. Available: https://doi.org/10.1038/323533a0 [Cited on page 19.]

[48] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2022. [Online]. Available: https://arxiv.org/abs/1312.6114 [Cited on pages 19, 25, 26, and 73.]

[49] Y. LeCun, "Gans are the most interesting idea in the last ten years in machine learning," 2016, quora post, accessed June 15, 2024. [Online]. Available: https://www.quora.com/What-are-generative-adversarial-networks [Cited on page 20.]

[50] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," 2016. [Online]. Available: https://arxiv.org/abs/1601.06759 [Cited on page 20.]

[51] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real nvp," 2017. [Online]. Available: https://arxiv.org/abs/1605.08803 [Cited on page 20.]

[52] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," 2015. [Cited on pages 20, 23, and 24.]

[53] J. Betker, G. Goh, L. Jing, TimBrooks, J. Wang, L. Li, LongOuyang, JuntangZhuang, JoyceLee, YufeiGuo, WesamManassra, PrafullaDhariwal, CaseyChu, YunxinJiao, and A. Ramesh, "Improving image generation with better captions." [Online]. Available: https://api.semanticscholar.org/CorpusID:264403242 [Cited on page 20.]

[54] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, D. Podell, T. Dockhorn, Z. English, K. Lacey, A. Goodwin, Y. Marek, and R. Rombach, "Scaling rectified flow transformers for high-resolution image synthesis," 2024. [Cited on page 20.]

[55] OpenAI, "Sora: Video synthesis software," 2024, available at https://www.openai.com/sora. [Cited on page 20.]

[56] R. Holzman, "The minimax property in infinite two-person win-lose games," 2023. [Online]. Available: https://arxiv.org/abs/2310.19314 [Cited on page 20.]

[57] C. A. Holt and A. E. Roth, "The nash equilibrium: A perspective," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, pp. 3999 – 4002, 2004. [Online]. Available: https://api.semanticscholar.org/CorpusID:9037168 [Cited on page 21.]

[58] PyTorch, "torch.nn.bceloss," 2024, accessed: 2024-05-22. [Online]. Available: https://pytorch.org/docs/stable/generated/torch.nn.BCELoss.html [Cited on page 21.]

[59] A. Bissoto, E. Valle, and S. Avila, "The six fronts of the generative adversarial networks," 2019. [Online]. Available: https://arxiv.org/abs/1910.13076 [Cited on page 22.]

[60] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org. [Cited on page 21.]

[61] F. Nielsen, "On the jensen–shannon symmetrization of distances relying on abstract means," *Entropy*, vol. 21, no. 5, p. 485, May 2019. [Online]. Available: http://dx.doi.org/10.3390/e21050485 [Cited on page 22.]

[62] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," 2016. [Online]. Available: https://arxiv.org/abs/1606.03498 [Cited on pages 22 and 23.]

[63] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2016.

[64] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," 2017. [Cited on page 23.]

[65] W. Feller, "On the theory of stochastic processes, with particular reference to applications," 1949. [Online]. Available: https://api.semanticscholar.org/CorpusID:121027442 [Cited on page 23.]

[66] R. M. Neal, "Annealed importance sampling," 1998. [Cited on page 23.]

[67] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," 2020. [Cited on page 23.]

[68] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," 2021. [Cited on page 23.]

[69] A. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," *CoRR*, vol. abs/2102.09672, 2021. [Online]. Available: https://arxiv.org/abs/2102.09672 [Cited on pages 24, 26, 29, and 32.]

[70] N. Metropolis and S. Ulam, "The monte carlo method," *J. Am. Stat. Assoc.*, vol. 44, no. 247, pp. 335–341, Sep. 1949. [Cited on page 27.]

[71] C. P. Robert and G. O. Roberts, "Rao-blackwellization in the mcmc era," 2021. [Cited on page 27.]

[72] T. Salimans and J. Ho, "Progressive distillation for fast sampling of diffusion models," 2022. [Cited on page 29.]

[73] H. Zheng, W. Nie, A. Vahdat, K. Azizzadenesheli, and A. Anandkumar, "Fast sampling of diffusion models via operator learning," 2023.

[74] C. Meng, R. Rombach, R. Gao, D. P. Kingma, S. Ermon, J. Ho, and T. Salimans, "On distillation of guided diffusion models," 2023. [Cited on page 29.]

[75] S. Mohamed and B. Lakshminarayanan, "Learning in implicit generative models," 2017. [Cited on pages 29 and 30.]

[76] W. H. L. Pinaya, P.-D. Tudosiu, J. Dafflon, P. F. Da Costa, V. Fernandez, P. Nachev, S. Ourselin, and M. J. Cardoso, "Brain imaging generation with latent diffusion models," in *Deep Generative Models: Second MICCAI Workshop, DGM4MICCAI 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings*. Berlin, Heidelberg: Springer-Verlag, 2022, p. 117–126. [Online]. Available: https://doi.org/10.1007/978-3-031-18576-2_12 [Cited on page 33.]

[77] F. Khader, G. Müller-Franzes, S. Tayebi Arasteh, T. Han, C. Haarburger, M. Schulze-Hagen, P. Schad, S. Engelhardt, B. Baeßler, S. Foersch, J. Stegmaier, C. Kuhl, S. Nebelung, J. N. Kather, and D. Truhn, "Denoising diffusion probabilistic models for 3d medical image generation," *Scientific Reports*, vol. 13, no. 1, p. 7303, May 2023. [Online]. Available: https://doi.org/10.1038/s41598-023-34341-2

[78] H. Vu Quoc, T. Tran Le Phuong, M. Trinh Xuan, and S. Dinh Viet, "Lsegdiff: A latent diffusion model for medical image segmentation," in *Proceedings of the 12th International Symposium on Information and Communication Technology*, ser. SOICT '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 456–462. [Online]. Available: https://doi.org/10.1145/3628797.3629010 [Cited on page 33.]

[79] W. Wang, J. Bao, W. Zhou, D. Chen, D. Chen, L. Yuan, and H. Li, "Semantic image synthesis via diffusion models," 2022. [Online]. Available: https://arxiv.org/abs/2207.00050 [Cited on page 35.]

[80] G. Eskandar, M. Abdelsamad, K. Armanious, and B. Yang, "Usis: Unsupervised semantic image synthesis," 2021. [Online]. Available: https://arxiv.org/abs/2109.14715 [Cited on page 35.]

[81] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015. [Cited on page 35.]

[82] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2017. [Cited on page 35.]

[83] Y. Wu and K. He, "Group normalization," 2018. [Online]. Available: https://arxiv.org/abs/1803.08494 [Cited on pages 35 and 41.]

[84] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," 2018. [Cited on pages 35 and 71.]

[85] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," 2017. [Cited on page 35.]

[86] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," 2020. [Cited on page 35.]

[87] P. Durgadevi, T. Akilan, A. Pradhan, A. M. Shariff, N. Yadav, and P. Uppal, "Canny edge detection techniques for image segmentation," in *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, 2022, pp. 986–989. [Cited on page 37.]

[88] F. Liu, J. Wang, B. Liu, R. Min, G. Yu, F. An, and C. Wang, "Low computation and high efficiency sobel edge detector for robot vision," in *2021 IEEE International Conference on Real-time Computing and Robotics (RCAR)*, 2021, pp. 684–689. [Cited on page 37.]

[89] H. Xu, "An improved edge detector based on the laplacian operator," in *2013 6th International Congress on Image and Signal Processing (CISP)*, vol. 01, 2013, pp. 416–421. [Cited on page 37.]

[90] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. Gonzalez Ballester, G. Sanroma, S. Napel, S. Petersen, G. Tziritas, E. Grinias, M. Khened, V. A. Kollerathu, G. Krishnamurthi, M.-M. Rohe, X. Pennec, M. Sermesant, F. Isensee, P. Jager, K. H. Maier-Hein, P. M. Full, I. Wolf, S. Engelhardt, C. F. Baumgartner, L. M. Koch, J. M. Wolterink, I. Isgum, Y. Jang, Y. Hong, J. Patravali, S. Jain, O. Humbert, and P.-M. Jodoin, "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved?" *IEEE Transactions on Medical Imaging*, vol. 37, no. 11, pp. 2514–2525, Nov 2018. [Online]. Available: https://ieeexplore.ieee.org/Xplorehelp/downloads/license-information/IEEE.html [Cited on pages 37 and 48.]

[91] D. Khaledyan, A. Amirany, K. Jafari, M. H. Moaiyeri, A. Z. Khuzani, and N. Mashhadi, "Low-cost implementation of bilinear and bicubic image interpolation for real-time image super-resolution," 2020. [Online]. Available: https://arxiv.org/abs/2009.09622 [Cited on page 37.]

[92] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," 2018. [Online]. Available: https://arxiv.org/abs/1706.08500 [Cited on page 38.]

[93] M. Seitzer, "pytorch-fid: FID Score for PyTorch," https://github.com/mseitzer/pytorch-fid, August 2020, version 0.3.0. [Cited on page 38.]

[94] I. Bakurov, M. Buzzelli, R. Schettini, M. Castelli, and L. Vanneschi, "Structural similarity index (ssim) revisited: A data-driven approach," *Expert Systems with Applications*, vol. 189, p. 116087, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417421014238 [Cited on page 38.]

[95] A. Poli and M. Cirillo, "On the use of the normalized mean square error in evaluating dispersion model performance," *Atmospheric Environment. Part A. General Topics*, vol. 27, pp. 2427–2434, 10 1993. [Cited on page 38.]

[96] "Medical imaging with deep learning 2024." Paris, France: MIDL, 2024. [Online]. Available: https://2024.midl.io/ [Cited on page 38.]

[97] N. M. F. Capitão, Y. Zhao, Y. Zhang, N. Geerts, J. V. Lopes, and Q. Tao, "Anatomy-compliant medical image synthesis by latent diffusion models," in *Medical Imaging*

with Deep Learning, 2024. [Online]. Available: https://openreview.net/forum?id=m9pHODd2HS [Cited on page 38.]

[98] M. K. Chung, "Gaussian kernel smoothing," 2021. [Online]. Available: https://arxiv.org/abs/2007.09539 [Cited on page 39.]

[99] J.-R. Liao and C. Y. Chan, "Efficient implementation of non-local means image denoising algorithm," in *2019 IEEE 8th Global Conference on Consumer Electronics (GCCE)*, 2019, pp. 566–567. [Cited on page 39.]

[100] H. Wu and M. Flierl, "Vector quantization-based regularization for autoencoders," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, pp. 6380–6387, Apr. 2020. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/6108 [Cited on page 39.]

[101] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," 2018. [Cited on page 39.]

[102] S. Elfwing, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," 2017. [Online]. Available: https://arxiv.org/abs/1702.03118 [Cited on page 40.]

[103] X. Zhuang, "Multivariate mixture model for myocardial segmentation combining multi-source images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 12, pp. 2933–2946, 2019. [Cited on page 42.]

[104] J. Qiu, L. Li, S. Wang, K. Zhang, Y. Chen, S. Yang, and X. Zhuang, "Myops-net: Myocardial pathology segmentation with flexible combination of multi-sequence cmr images," *Medical Image Analysis*, vol. 84, p. 102694, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S136184152200322X [Cited on page 42.]

[105] J. Bertels, T. Eelbode, M. Berman, D. Vandermeulen, F. Maes, R. Bisschops, and M. B. Blaschko, *Optimizing the Dice Score and Jaccard Index for Medical Image Segmentation: Theory and Practice*. Springer International Publishing, 2019, p. 92–100. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-32245-8_11 [Cited on page 42.]

[106] T. Delft, "Proton therapy research," https://www.tudelft.nl/en/proton-therapy-research/, 2024, accessed: 2024-09-07. [Cited on page 42.]

[107] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, Feb 2021. [Online]. Available: https://doi.org/10.1038/s41592-020-01008-z [Cited on page 43.]

[108] "AMOS 2022 Challenge," https://amos22.grand-challenge.org/final-ranking/, accessed: 2024-07-08. [Cited on page 43.]

[109] J. Ma, "Cutting-edge 3d medical image segmentation methods in 2020: Are happy families all alike?" 2021. [Online]. Available: https://arxiv.org/abs/2101.00232 [Cited on page 43.]

[110] A. L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. van Ginneken, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers, P. Bilic, P. F. Christ, R. K. G. Do, M. Gollub, J. Golia-Pernicka, S. H. Heckers, W. R. Jarnagin, M. K. McHugo, S. Napel, E. Vorontsov, L. Maier-Hein, and M. J. Cardoso, "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," 2019. [Online]. Available: https://arxiv.org/abs/1902.09063 [Cited on page 43.]

[111] W. P. Segars, G. Sturgeon, S. Mendonca, J. Grimes, and B. M. W. Tsui, "4D XCAT phantom for multimodality imaging research," *Med. Phys.*, vol. 37, no. 9, pp. 4902–4915, Sep. 2010. [Cited on page 44.]

[112] E. Zurich, "MRXCAT: An open-source framework for MRI simulations," https://biomed.ee.ethz.ch/mrxcat.html, Accessed: 2024-07-05. [Cited on page 44.]

[113] S. Jayasumana, S. Ramalingam, A. Veit, D. Glasner, A. Chakrabarti, and S. Kumar, "Rethinking fid: Towards a better evaluation metric for image generation," 2024. [Online]. Available: https://arxiv.org/abs/2401.09603 [Cited on page 59.]

[114] M. Woodland, A. Castelo, M. A. Taie, J. A. M. Silva, M. Eltaher, F. Mohn, A. Shieh, S. Kundu, J. P. Yung, A. B. Patel, and K. K. Brock, "Feature extraction for generative medical imaging evaluation: New evidence against an evolving trend," 2024. [Online]. Available: https://arxiv.org/abs/2311.13717 [Cited on page 59.]

[115] C. Li and M. Wand, "Precomputed real-time texture synthesis with markovian generative adversarial networks," 2016. [Cited on page 72.]

[116] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," 2018. [Cited on page 73.]

[117] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," 2018. [Cited on page 73.]

[118] S. Li, Z. Liu, J. Tian, G. Wang, Z. Wang, W. Jin, D. Wu, C. Tan, T. Lin, Y. Liu, B. Sun, and S. Z. Li, "Switch ema: A free lunch for better flatness and sharpness," 2024. [Online]. Available: https://arxiv.org/abs/2402.09240 [Cited on page 74.]