Imperial College London

MSc Individual Project

Imperial College London

Department of Computing

Cardiac Shape Analysis with Nouveau Variational Autoencoder

Author:
Freddy Jiang

Supervisors:
Prof. Elsa Angelini
Prof. Loïc Le Folgoc

Second Marker:
Dr Ahmed Fetit

Submitted in partial fulfilment of the requirements for the MSc degree in Computing (Artificial Intelligence and Machine Learning) of Imperial College London

September 8, 2024

Abstract

Cardiovascular diseases (CVDs) cause over 20 million deaths annually, with a third occuring prematurely in people under the age of 70. However, CVDs are largely preventable with early detection and intervention. Over recent years, there has been rapid progression in the development of automated techniques for cardiac magnetic resonance imaging (MRI) analysis. Accurate delineation of cardiac components is crucial to assist in anomaly detection and diagnosis, and shape analysis is an essential prerequisite.

The emergence of deep learning has introduced powerful frameworks capable of automating the process of learning compact shape representations. Variational autoencoders (VAEs) are a class of generative models that excel at learning efficient low-dimensional representations of complex data. In particular, the Nouveau VAE (NVAE) is a deep hierarchical VAE that is the state-of-the-art among its class in encoding fine-grained details in high-resolution images.

In this dissertation, we examine how the NVAE framework can be applied to cardiac shape analysis. We propose configurations that can learn from clinically annotated segmentation masks to efficiently encode cardiac anatomic shapes, with significantly improved performance over existing VAE models (up to 0.108 Dice increase for reconstructed masks and 22.0% anatomical validity increase in synthetic masks when used as a generative model, the latter of which ensures the generated shapes conform to realistic cardiac anatomy). Furthermore, we propose a novel metric, the Fréchet ResNet Distance with SimCLR (FRDS), which improves over the Fréchet Inception Distance in measuring the similarity between synthetic and real cardiac segmentation masks. We demonstrate that the learned NVAE encodings can be used in downstream tasks by using them as an anatomical constraint to improve the segmentation performance of a U-Net model (5.3% anatomical validity increase). We find these encodings to generalise well when applied to unseen data, without the need for further training.

Acknowledgments

I would like to thank my supervisors, Prof. Elsa Angelini and Prof. Loïc Le Folgoc, for their invaluable advice and guidance throughout the project. They have dedicated a significant amount of time, even going outside of their working hours to support me, for which I am most grateful. This project would certainly not have been possible without their help.

I would also like to thank my second marker, Dr Ahmed Fetit, for his valuable feedback during the interim meeting.

Finally, I would like to recognise the support of my family and friends. Their presence has been a stellar source of motivation and encouragement throughout this degree.

Contents

Chapter 1

Introduction

1.1 Motivation

Cardiovascular diseases (CVDs) are diseases of the heart and blood vessels. Causing 20.5 million deaths in 2021 and accounting for a third of all deaths globally, CVDs have been the leading cause of death for decades[1]. Over 80% of CVDs are due to heart attacks and strokes, with a third of these deaths occuring prematurely in people under the age of 70[2].

Up to 80% of premature heart attacks and strokes are preventable with treatment[1]. There has been significant efforts in developing techniques and tools to identify patients at risk of CVDs. In particular, it is crucial to monitor the shape of the heart. Magnetic resonance imaging (MRI) is the gold standard for obtaining scans of the heart and surrounding structures. Accurate delineation of heart components, especially the left and right ventricles (lower heart chambers) and the myocardium (surrounding muscle tissue), is an important prerequisite to detect abnormalities and provide reliable diagnosis.

Advancements in machine learning frameworks have enabled the task of segmenting cardiac scans to be semi-automated in clinical practice, from what traditionally involved fully manual delineation. However, the lack of accuracy, robustness and interpretability of these models poses a considerable barrier. As such, experts are still required to provide a significant amount of manual correction and supervision, which is a time-consuming process and introduces intra- and inter-observer variability.

Within the field of generative AI, variational autoencoders[3] (VAEs) excel at learning compact representations of data, and have been used to learn underlying patterns of cardiac shapes. They can provide population-level visualisations and transparency for classification models for diagnosing CVDs[4, 5], which are otherwise notorious for being black-box models. The learned representations can also be injected into, or used alongside, segmentation models to improve output segmentations[6, 7, 8, 9].

As a generative model, VAEs are capable of producing synthetic data that shares the same style and distribution as the training data. Historically, these synthetic data have been inferior to other frameworks like generative adversarial networks (GANs)[10], flow-based methods[11] and autoregressive models[12]. However, research efforts have lead to hierarchical VAE frameworks with strong performance and stability. In particular, Nouveau VAE (NVAE)[13] is the state-of-the-art among its class and has shown powerful capabilities in both generative tasks

and learning compact representations that encode fine-grained details for high-resolution natural image datasets.

In this dissertation, we aim to apply NVAE to cardiac shape analysis and present a comprehensive evaluation of its performance and utility, including its adaptability to various downstream tasks by incorporating it within popular segmentation models.

1.2 Objectives

Our primary objective is to evaluate the performance of Nouveau VAE in cardiac shape encoding. Specifically, the model is trained to perform 2 tasks: (1) it can take in cardiac segmentation masks annotated by clinical experts and output reconstructions with minimal depreciation of quality, and (2) it can generate realistic synthetic masks. A model that excels at both tasks is considered to have learned a meaningful, compact representation of the anatomical shapes. We propose 2 hierarchical architectures adapted from the NVAE framework that greatly improves the quality of reconstructions and generated masks over baseline VAE models. Furthermore, we propose a novel metric, Fréchet ResNet Distance with SimCLR (FRDS) for measuring the similarity between synthetic and real masks.

Our secondary objectives involve investigating the potential of using the learned representations from NVAE to improve downstream tasks. In particular, we extend an existing segmentation framework by using the learned representations as a regulariser. We find that this results in the model producing more anatomically valid segmentations. We investigate the robustness of this addition by evaluating the model on in-distribution data, as well as data from a previously unseen dataset with different acquisition protocols. We find success in applying the framework in a domain adaptation and few-shot learning setting.

1.3 Ethical Considerations

This dissertation is purely research-oriented. We do not warrant nor take liability of any misuse of methods and/or findings associated with this work.

Experimental data is sourced from the Automated Cardiac Diagnosis Challenge (ACDC) dataset[14] and the Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image Segmentation Challenge (M&Ms) dataset[15, 16]. Our research involves preprocessing of previously collected sensitive data that has been completely anonymised and rendered unidentifiable. We use only the data that is publicly available. As such, our work is exempt from ethical approval.

NVAE is open source and can be used in accordance with the NVIDIA Source Code License[1]. Section 3.3 of the license states that the work can be used for research or evaluation purposes only.

Our findings are a result of experiments that involve heavy computation via prolonged training of over 1,000 models[2] and 860 hours of GPU time (Tesla A40 48GB). With a maximum power consumption of 300W, the total energy consumption is estimated to be 258kWh and CO2 emis-

[1] https://github.com/NVlabs/NVAE/blob/master/LICENSE
[2] Including hyperparameter tuning.

sions to be 52.9kg[3]. In comparison, the average electric car emits 1,278kg of CO2 emissions annually.

To limit environmental impact, we make diligent effort in using systems optimised for AI training, such as CUDA and Slurm Workload Manager, as well as writing efficient code.

1.4 Summary

Our experiments are conducted with the PyTorch Lightning framework. As a research paper is expected to be produced from this work, our codebase will only be made publicly available upon completion.

This dissertation is structured as follows. Chapter 2 provides a literature review on VAEs, generative models and their applications in cardiac shape encoding. Chapter 3 analyses the experimental datasets and describes the preprocessing pipeline. Chapter 4 describes the experiments conducted, including our proposed methods and implementation details. In Chapter 5, we introduce FRDS, a metric for evaluating quality of synthetic cardiac shapes. Chapter 6 reports experimental results and presents a quantitative and qualitative evaluation of proposed methods. Finally, Chapter 7 revisits the objectives by summarising our contributions, provides concluding remarks and proposes future directions of this project.

[3] Assuming an emission factor of 0.20493kgCO2e/kWh [17].

Chapter 2

Background

2.1 Generative AI

Generative artificial intelligence (generative AI) encompasses a class of algorithms that can output realistic synthetic data. Advancements in deep neural networks have led to the development of models capable of generating high quality data across domains such as text[18], images[12] and audio[19]. In this section, we overview foundational concepts and models of generative AI, with a focus on the variational autoencoder[20].

2.1.1 Autoencoder

An autoencoder[21] is an artificial neural network that aims to learn an efficient representation of an unlabelled set of data by passing the data through a low-dimensional space. This architecture is a powerful tool for feature extraction by dimensionality reduction.

More formally, a high-dimensional input $x \in \mathbb{R}^D$ is passed through an encoder f to produce a low-dimensional latent variable $z \in \mathbb{R}^d$ where $d \ll D$. z is then passed through a decoder g to output a reconstruction $\hat{x} \in \mathbb{R}^D$ (Figure 2.1). The network is trained to minimise the reconstruc-

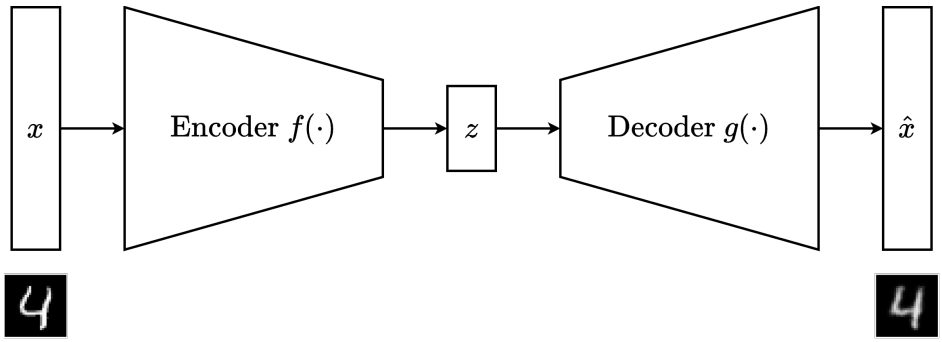


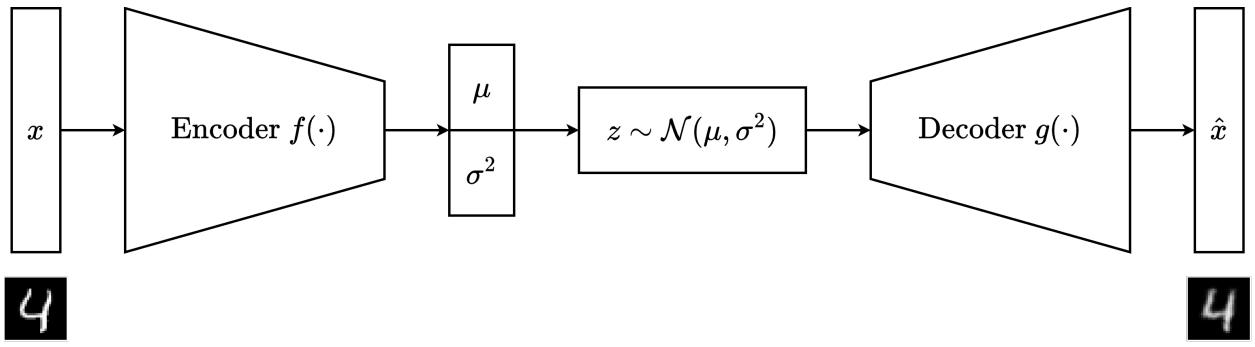
Figure 2.1: The autoencoder architecture. The encoder f maps a high-dimensional input x to a low-dimensional representation z, then the decoder g maps z back to the original space. This diagram uses MNIST[22] as an example dataset. The encoder-decoder architecture is chosen appropriately for the data domain, such as using convolutional layers to process images. Due to the low-dimensional bottleneck, the reconstruction may lose some quality (lossy compression).

tion loss, that is, the difference between x and $\hat{x}$. The low-dimensional bottleneck forces the network to extract the most important features of the input data. Applications of autoencoders include dimensionality reduction[23], denoising[24] and anomaly detection[25].

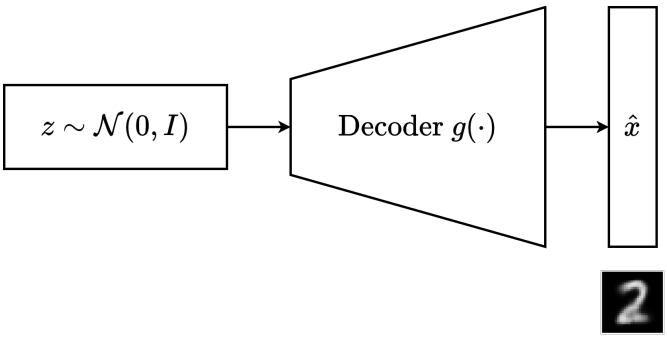
2.1.2 Variational Autoencoder

A variational autoencoder[20] (VAE) is a probablistic extension of an autoencoder. In short, the encoder outputs a distribution $q_\phi(z \mid x)$ over the latent space instead of a fixed point. A sample $z \sim q_\phi(z \mid x)$ is drawn from this distribution and passed through the decoder to output $\hat{x}$ (Figure 2.2a). This variability allows the network to be used for generative tasks (Figure 2.2b). The ability to produce synthetic data allows VAEs to be applied in anomaly correction[6], improving speech recognition systems[26] and options pricing[27].

More formally, a VAE is a directed graphical model that assumes Markovian properties to learn the data distribution as $p_\theta(x,z) = p_\theta(x|z)p_\theta(z)$, where we assume there exists an unobserved factor z that affects data generation and it follows a prior distribution $p_\theta(z)$ (Figure 2.3). Then, the marginal likelihood is $p_\theta(x) = \int_z p_\theta(x|z)p_\theta(z)dz$. We are interested in maximising $\log p_\theta(x)$. However, integrating over the entire latent space is intractable, hence the true pos-



(a) Reconstruction / training process. The encoder f maps a high-dimensional input x to a simple distribution over a low-dimensional space, often a Gaussian. A latent variable z is sampled from this space, then the decoder g maps z back to the original space.



(b) Generative process. Assume a standard Gaussian prior. A sample z is drawn from the prior, then the decoder g maps z to a synthetic data point. The generated image resembles the digit 2.

Figure 2.2: The variational autoencoder architecture. The figures use MNIST[22] as an example dataset. Similar to an autoencoder, the reconstruction may lose quality due to the low-dimensional bottleneck. The generated image also exhibits a lower quality texture.

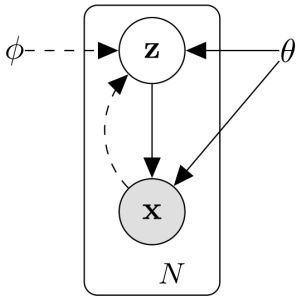


Figure 2.3: The directed graphical model of a variational autoencoder. Solid lines correspond to the decoder process modelled by parameters θ, dashed lines correspond to the encoder process modelled by ϕ. [20, p. 2]

terior $p_\theta(z|x) = \frac{p_\theta(x|z)p_\theta(z)}{p_\theta(x)}$ is also intractable. Instead, we introduce the encoder $q_\phi(z|x)$ to approximate the true posterior. We refer to $p_\theta(x|z)$ as the decoder. In literature, the encoder and decoder are sometimes referred to as the recognition model and generative model [20].

In practice, we typically assume a standard Gaussian prior $p_\theta(z) = \mathcal{N}(z; 0, I)$[1]. Let $q_\phi(z|x) = \mathcal{N}(z; \mu_\phi(x), \text{diag}(\sigma^2_\phi(x)))$. Then, the marginal log-likelihood can be shown to be bound by the evidence lower bound $\mathcal{L}(x, \theta, \phi)$ (ELBO) (2.1).

$$\log p_\theta(x) \geq \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - D_{KL}\left[q_\phi(z|x)||p_\theta(z)\right] =: \mathcal{L}(\theta, \phi; x) \tag{2.1}$$

The ELBO has two terms: a log-likelihood reconstruction term and a KL divergence term that acts as a regulariser. Using this reformulation, we maximise the ELBO during training. $D_{KL}\left[q_\phi(z|x)||p_\theta(z)\right]$ can be computed analytically for Gaussian and Bernoulli priors. $\mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right]$ can be approximated using the reparameterisation trick and MC sampling to obtain a differentiable expression for backpropagation (2.2).

$$\begin{aligned} \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] &= \frac{1}{M}\sum_{m=1}^{M} \log p_\theta\left(x_m \mid T_\phi(x_m, \epsilon_m)\right) \\ T_\phi(x, \epsilon) &:= \mu_\phi(x) + \sigma_\phi(x) \odot \epsilon \\ x_1, \ldots x_M &\sim \{x_n\}^M \\ \epsilon_1, \ldots \epsilon_M &\sim \mathcal{N}(0, I) \end{aligned} \tag{2.2}$$

A restriction of the VAE is that during inference, the sampled z completely determines the generated data point. Assuming a well-trained VAE, this data point can resemble a realistic sample from any class within the train set. Frameworks have been proposed to introduce more control over the type of data generated, such as Conditional VAE (CVAE) [28], which conditions the generation on an additional input.

2.1.3 Disentanglement

In the context of representation learning, the data distribution for a complex task often has many underlying explanatory factors of variation. Disentanglement is the process of separating these factors into distinct latent variables [29].

[1] Some works use a Bernoulli prior or a uniform prior.

We focus on disentanglement for VAEs. A fully disentangled latent space is one where each dimension corresponds to one feature, so changing the value along that dimension (and fixing other dimensions) would cause changes to that feature only. By default, VAEs do not enforce disentanglement. However, it is desirable in some use cases, as it allows for direct control over the factors of variation during generation.

β-VAE[30] is a framework that extends the VAE by adding a hyperparameter β that weights the KL divergence term in the ELBO(2.3).

$$\mathscr{L}(\theta,\phi,\beta;x) \coloneqq \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - \beta D_{KL}\left[q_\phi(z|x)||p_\theta(z)\right] \tag{2.3}$$

$\beta > 1$ encourages disentanglement while maintaining a lower bound on the marginal likelihood. To realise the former, the KL divergence term can be decomposed as (2.4) [12].

$$\begin{aligned}\mathbb{E}_{p_{\text{data}}(x)}[D_{KL}(q_\phi(z|x)||p_\theta(z))] &= I(x;z) + D_{KL}(q_\phi(z)||p_\theta(z)) \\ q_\theta(z) &= \mathbb{E}_{p_{\text{data}}(x)}[q_\phi(z|x)]\end{aligned} \tag{2.4}$$

Penalizing $D_{KL}(q_\phi(z)||p_\theta(z))$ pushes the aggregate posterior $q_\phi(z)$ towards the factorial prior $p_\theta(z)$. This encourages independence between the dimensions of z. Minimising this term also improves generation quality, since a perfect recognition model $q_\phi(z) = p_\theta(z)$ yields synthetic data indistuingishable from real data. However, penalizing the mutual information $I(x;z)$[2] reduces information about x in z. Therefore, β-VAE introduces a trade-off between disentanglement and reconstruction quality.

There has been efforts to address this drawback of β-VAE, and we focus on two of them: the β-TCVAE algorithm[31] and the FactorVAE framework[32].

Both β-TCVAE and FactorVAE build on top of β-VAE by introducing a loss function that encourages a factorised latent distribution (2.5).

$$\begin{aligned}\mathscr{L}(\theta,\phi,\alpha,\beta,\gamma;x) \coloneqq{}& \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - \alpha I(x;z) \\ &+ \beta D_{KL}\left(q_\phi(z)||\prod_j q_\phi(z_j)\right) - \gamma \sum_j D_{KL}\left(q_\phi(z_j)||p_\theta(z_j)\right)\end{aligned} \tag{2.5}$$

Crucially, (2.5) introduces a total correlation (TC) term (weighted by β), which enforces statistically independent factors in the latent distribution. However, TC is analytically intractable. β-TCVAE proposes approximating TC with minibatch sampling, for which we derive below.

$$\begin{aligned}D_{KL}\left(q_\phi(z)||\prod_j q_\phi(z_j)\right) &= \int_z q_\phi(z)\log\frac{q_\phi(z)}{\prod_j q_\phi(z_j)}dz \\ &\approx \frac{1}{N}\sum_{n=1}^{N}\log\frac{q_\phi\left(z^{(n)}\right)}{\prod_j q_\phi\left(z_j^{(n)}\right)} \quad z^{(n)} \sim q_\phi(z) \\ &= \frac{1}{N}\sum_{n=1}^{N}\left(\log q_\phi\left(z^{(n)}\right) - \log\prod_j q_\phi\left(z_j^{(n)}\right)\right)\end{aligned}$$

On the other hand, FactorVAE proposes using a discriminator to approximate the density ratio in the TC term. The discriminator is trained jointly with the VAE, similar to the adversarial training process in GANs[10] (Section 2.1.5).

[2] Equivalent to $D_{KL}[q_\phi(x,z)||q_\phi(z)p_\theta(x)]$.

2.1.4 InfoVAE

In Section 2.1.3, we discussed the shortcomings of β-VAE, in particular, how penalizing the mutual information term can lower reconstruction quality. On the other hand, penalizing $D_{KL}(q_\phi(z)||p_\theta(z))$ can improve both disentanglement and generation quality. InfoVAE[33] is a framework that extends β-VAE by leverging this term to improve reconstruction quality while maintaining good generations.

InfoVAE uses the loss function presented in (2.6). This extends the β-VAE loss (2.3) by introducing an additional regulariser term: the KL divergence between the prior and the aggregate posterior, weighted by a parameter γ.

$$\mathscr{L}(\theta,\phi,\beta,\gamma;x) := \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - \beta D_{KL}\left[q_\phi(z|x)||p_\theta(z)\right] - \gamma D_{KL}\left[q_\phi(z)||p_\theta(z)\right] \tag{2.6}$$

When $p_\theta(x|z)$ is a complex distribution, the authors suggest using $\beta = 0$. This is equivalent to removing the mutual information penalty in the ELBO, as seen by (2.4). $D_{KL}\left[q_\phi(z)||p_\theta(z)\right]$ is not analytically tractable and the authors suggest using an auxiliary discriminator to approximate it.

2.1.5 Various Works

In this paper, we focus on the VAE and its derivatives, in particular, the Nouveau VAE (NVAE) framework[13] (Section 2.2.2). Our work shares similarities with other generative frameworks, so we describe relevant works below.

A generative adversarial network (GAN)[10] is a generative framework that aims to fit the data distribution $p_{\text{data}}(x)$ directly as $p_\theta(x)$ without an explicit latent space. Instead, it introduces a discriminator D. D is a binary classifier that aims to distinguish between real data $x_r \sim p_{\text{data}}(x)$ and synthetic data $x_g \sim p_\theta(x)$ produced by the generator G. G and D are trained simultaneously in a 2-player min-max game. As such, unlike VAEs, training a GAN does not involve calculating a log-likelihood. The aim is for G to generate data indistinguishable from real data. GANs are capable of generating high fidelity data, and as a result have been applied in super resolution tasks. However, they are more difficult to train and can suffer from mode collapse[3]. Furthermore, the lack of an explicit latent space makes GANs less suitable for shape encoding tasks[4] compared to VAEs.

Normalising flow[11] is a method that transforms a random variable z with a simple distribution into a random variable x with a complex distribution using a sequence of invertible transformations $x = f_K \circ \ldots \circ f_1(z)$. After applying f_i at each step, the resulting distribution is normalised. By enforcing f to be invertible, we can compute the probability density of x in terms of z, which allows the log-likelihood to be evaluated. Hence, it is possible to perform exact posterior inference[5], in contrast to VAEs which only allow approximate inference. As example, normalising flow can be used as a generative framework by letting $p(z)$ to be a standard Gaussian prior and $p(x)$ to be the observable data distribution.

[3] The outputs of the generator captures limited diversity of the data distribution.

[4] The process of representing shapes compactly, usually as a low-dimensional vector. An example is an image dataset, where learning the shape of key objects within each image allows the model to locate and segment them.

[5] Given x, find the exact value of z.

Normalising flow can be used to extend VAEs by applying a sequence of invertible transformations to the latent variable z, then passing the transformed z into the decoder. This no longer restricts the approximate posterior to a simple distribution (e.g. Gaussian) and can allow for better inference.

VAEs, GANs and normalising flows are non-autoregressive. An autoregressive generative model generates data sequentially, where the next token[6] is conditioned on previously generated tokens. As a result, autoregressive models can produce higher quality generations than non-autoregressive models, but they are significantly slower at inference time[34, 35, 36].

2.1.6 Evaluating Generation Quality

Evaluating synthetic data can be done empirically, but this succumbs to intra- and inter-observer variability, as well as being time-consuming. The current standard for quantitative evaluation of synthetic image data is the Fréchet Inception Distance (FID)[37], which builds on top of the Inception Score (IS)[38]. We give a brief overview of both metrics.

The algorithm for computing IS involves using a discriminator to classify the synthetic data. The discriminator is usually a pretrained Inception-v3 network. Let p_g be the distribution of synthetic data and $p_d(y|x)$ be the probability that an image x is classified as y by the discriminator. IS is defined by (2.7). For maximum IS, the predictions should be uniformly distributed across the classes and the entropy of the label distribution as predicted by the discriminator should be minimised. This means the generated images are diverse and distinct.

$$IS(p_g, p_d) = \exp\left(\mathbb{E}_{x\sim p_g}\left[D_{KL}\left(p_d(\cdot|x)||\int p_d(\cdot|x)\,p_g(x)\,dx\right)\right]\right) \tag{2.7}$$

FID improves over IS by comparing the synthetic data distribution to the real data distribution. It involves using a pretrained Inception-v3 network f with its classification head removed. Given the real dataset X and the synthetic dataset X', compute the embeddings $f(X)$ and $f(X')$, then approximate them as Gaussian distributions $\mathcal{N}(\mu,\Sigma)$ and $\mathcal{N}(\mu',\Sigma')$. FID is defined as $d_F(\mathcal{N}(\mu,\Sigma),\mathcal{N}(\mu',\Sigma'))^2$, where d_F is the Fréchet distance[39] as described by (2.8).

$$d_F(\mu,\nu) = \left(\inf_{\gamma\in\Gamma(\mu,\nu)}\int_{\mathbb{R}^n\times\mathbb{R}^n}\|x-y\|^2 d\gamma(x,y)\right)^{\frac{1}{2}} \tag{2.8}$$

Compared to IS, FID puts more emphasis on diversity of the synthetic data than quality of individual generations[40]. Note that FID is a distance metric, so lower values correspond to better generation quality.

There has been works to adapt FID to other domains, such as Fréchet Audio Distance (FAD)[41] for evaluating synthetic audio data and Fréchet Video Distance (FVD)[42] for evaluating synthetic video data.

[6] In the context of autoregressive models, a token refers to a building block of data. For example, in the context of generating image data, a token could be a pixel.

2.2 Hierarchical Models

In Section 2.1.2, we present the single-layer VAE (Figure 2.2a) and illustrate how the reconstruction can lose quality due to the low-dimensional bottleneck. Similarly, in Figure 2.2b, the generated image is blurry. This smooth, blurry texture is a limitation of single-layer VAEs, due to the L2 term in the objective function.

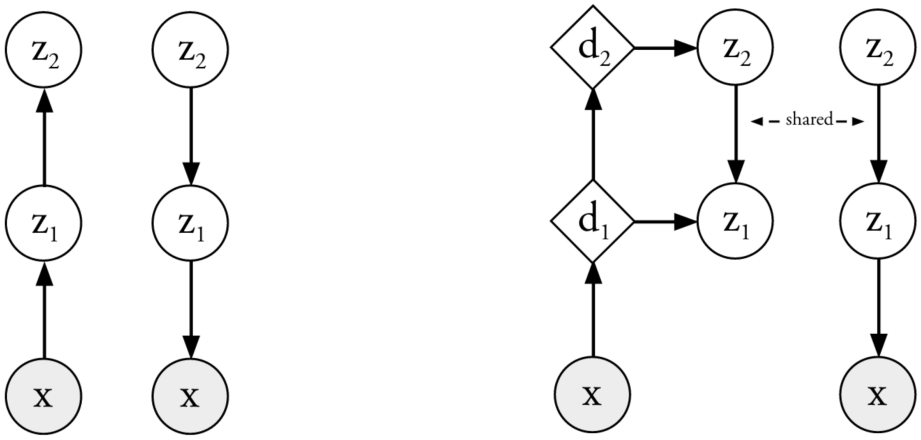
There has been works that attempt to use multiple layers of stochastic latent variables in the VAE model (Figure 2.4a). The theoretical concept is for the variables in the topmost layers to have the smallest dimensionality and learn long-range correlations to capture smooth, global features, while the variables in lower layers have higher dimensionality and build on top of previous layers by adding finer details. However, the VAE model uses the mean field approximation: it works under the assumption that the posterior can be factorised over the latent variables. This is a restrictive assumption and the VAE model degrades in performance after only 2 layers [43].

In this section, we discuss the Ladder VAE [43] and Nouveau VAE [13], two hierarchical VAE frameworks that can be used to train highly expressive generative models scalable to many latent layers.

2.2.1 Ladder VAE

Ladder VAE (LVAE) [43] is a hierarchical VAE framework that introduces a new encoder architecture which combines the approximate likelihood with the decoder model. Combined with stability techniques such as batch normalisation and a deterministic warm-up period for the KL divergence term, LVAE can learn a more distributed representation and achieve better generation quality than the VAE.

Figure 2.4b presents the LVAE architecture. The decoder is the same as a VAE model. The encoder recursively corrects the generative distribution conditioned on a data-dependent approximate likelihood. In the downward pass, the decoder computes the approximate posterior



(a) 2-layer VAE (b) 2-layer Ladder VAE

Figure 2.4: Directed graphical models for a multi-layer VAE and a Ladder VAE. Circles denote stochastic variables while diamonds denote deterministic variables. LVAE introduces a new encoder architecture and uses the same decoder as a VAE. [43, p. 2]

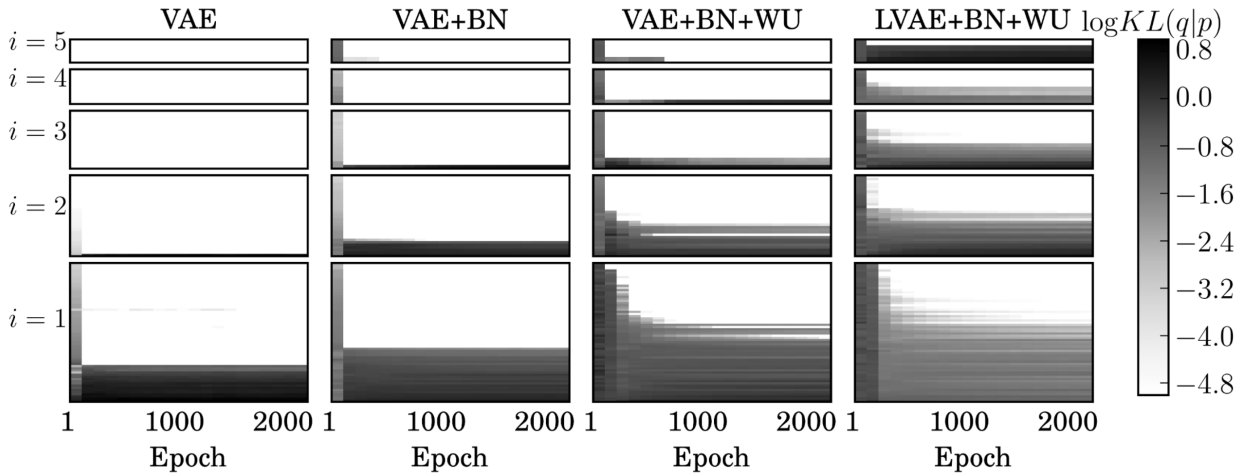


Figure 2.5: $\log D_{KL}$ for each latent unit in a 5-layer LVAE. Higher values correspond to darker shades of grey. Warm-up (WU) helps preserve more active units in the early epochs, some of which gets regularised and inactive later on. Even with batchnorm and warm-up, VAE struggles to keep units active in the topmost layers and succumbs to posterior collapse. LVAE is able to learn a more distributed representation. [43, p. 6]

and the generative distribution. This is opposed to a multi-layer VAE, where the encoder and decoder do not explicitly share information within each layer.

Compared to a multi-layer VAE, it is easier for LVAE to model the explaining away effect. This refers to the phenomenon where latent variables can become statistically dependent on each other: during inference, when a latent variable becomes active, it reduces the need for other latent variables to be active, causing them to collapse.

Batch normalisation (batchnorm) [44] is a method that normalises a batch of values via centering and scaling. In multi-layer VAE and LVAE, batchnorm helps with training the topmost layers to capture meaningful features. The warm-up period is motivated by the ELBO (2.1) containing a KL regularisation term. Since it forces the approximate posterior towards the prior, some latent variables become inactive. To prevent early posterior collapse, a deterministic warm-up period is introduced by extending (2.3) to increase β linearly from 0 to 1 during the first n epochs. Figure 2.5 presents the effect of batchnorm and warm-up period.

The authors test LVAE on low-resolution greyscale datasets only (MNIST [22], OMNIGLOT [45], NORB [46]), with the largest being 32×32. In Section 2.2.2, we discuss how Nouveau VAE improves upon LVAE in architectural design and stability, allowing it to tackle 256×256 colour image generation tasks.

2.2.2 Nouveau VAE

Nouveau VAE (NVAE) [13] is a hierarchical VAE framework that shares similar principles with LVAE, with additional improvements to architectural design and training stability. NVAE is the first variational autoencoder successfully applied to large-resolution[7] image generation tasks, such as CelebA HQ [47] resized to 256×256 and FFHQ [48] resized to 256×256 (both datasets consist of human face images). It also achieves state-of-the-art performance among

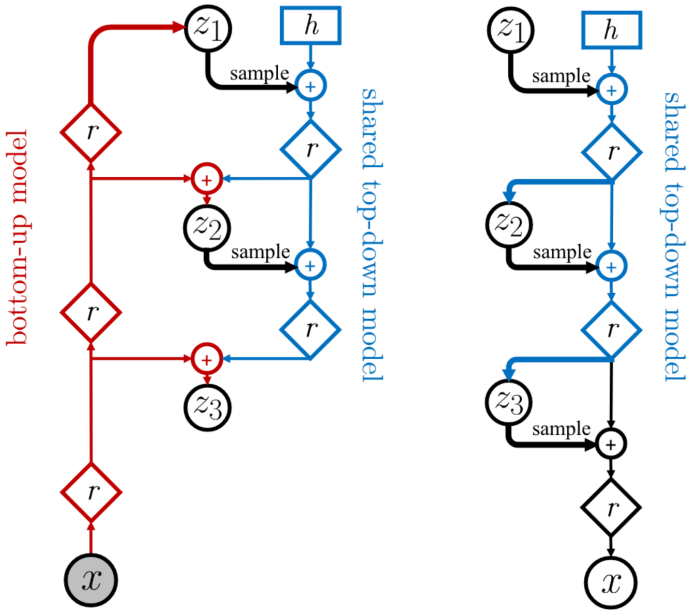
[7] 256×256

non-autoregressive likelihood models on other datasets such as MNIST [22], CIFAR-10 [49] and CelebA 64 [50, 51].

Figure 2.6 presents the NVAE architecture. Like LVAE, it consists of a shared encoder-decoder tower. It follows a similar principle with other hierarchical models: the topmost layer has a small latent dimensionality, and the spatial dimensions increase as we move down the layers. This allows global features to be captured higher up and finer details to be added in lower layers. NVAE uses batchnorm and warm-up period for training stability, but it also incorporates techniques like residual parameterisation for information flow and spectral regularisation (SR) [52, 53].

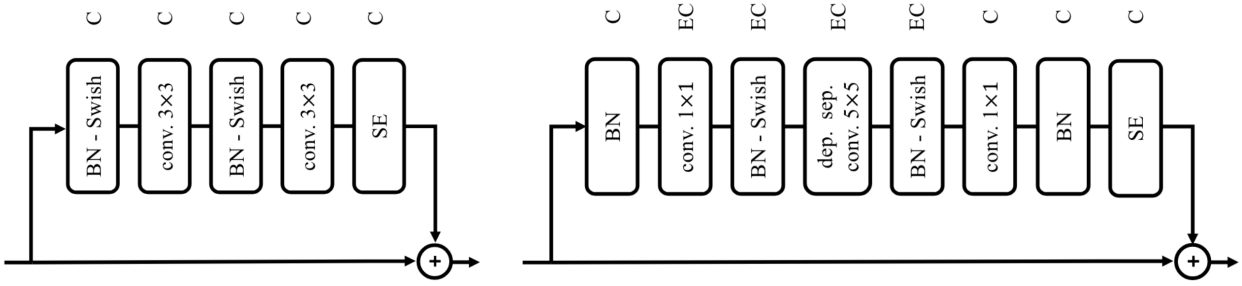
Residual parameterisation of approximate posterior is proposed to improve training stability, especially when there are many hierarchical groups. The residual cells (Figure 2.7) are designed with large kernel sizes to increase receptive field and capture global correlations, while using depthwise separable convolutions to limit the number of parameters (compared to normal convolutions). Since depthwise convolutions operate on individual channels and thus have limited capacity, they are applied after temporarily increasing the number of channels with a 1×1 convolution.

Mathematically, the residual distribution parameterises $q(z|x)$ relative to $p(z)$ so the approximate posterior moves depending on how the prior moves. For a Normal prior



(a) Encoder (b) Decoder

Figure 2.6: Encoder and decoder of a 3-layer NVAE. r denotes a block of residual cells (see Figure 2.7), + denotes a conbination cell that merges the layer-wise prior and posterior, and h is a learnable parameter. Note that each layer can have multiple groups, for example, if the topmost layer has 4 groups, we sample 4 latent variables from z_1 and use 4 combination cells to merge them sequentially. [13, p. 3]



(a) Encoder residual cell

(b) Decoder residual cell

Figure 2.7: Residual cells for NVAE. The encoder residual cell consists of batchnorm, swish activation and convolutions without changing number of channels. The decoder residual cell further applies depthwise convolution at a projected number of channels. Both cells end with a squeeze-and-excitation channel-wise gating layer, which improves performance. [13, p. 4]

$p(z^i) = \mathcal{N}(\mu_i, \sigma_i)$, the KL term in the objective becomes (2.9)[8].

$$D_{KL}\left[q(z^i|x)||p(z^i)\right] = \frac{1}{2}\left(\frac{\Delta\mu_i^2}{\sigma_i^2} + \Delta\sigma_i^2 - \log\Delta\sigma_i^2 - 1\right) \tag{2.9}$$

Since (2.9) is unbounded, the authors propose controlling the Lipschitz constant of the network using spectral regularisation. This is done by adding $\lambda\sum_i s^{(i)}$ to the objective, where $s^{(i)}$ is the largest singular value of the weight matrix in layer i and λ is a smoothness hyperparameter.

An area of concern is that NVAE has very high memory requirements. This is due to (1) the increased channels for depthwise convolution, and (2) overparameterisation in the model and latent space. The authors resort to using mixed-precision and gradient checkpointing [54, 55] to reduce GPU memory. As an example, the model uses 5 groups of $20 \times 4 \times 4$ and 10 groups of $20 \times 8 \times 8$ latent variables for MNIST, which amounts to 14,400 latent parameters for a 28×28 image, greatly exceeding its 784 input dimensions. NVAE uses regularisation techniques to turn off parameters during training. Less groups and smaller latent dimensions can be explored, which reduces memory load and may also improve interpretability and generation stability.

2.3 Cardiac Anatomy and Imaging Modalities

Cardiac anatomy refers to the structure of the (human) heart. The study of cardiac anatomy is important as irregularities in the heart can lead to life-threatening conditions, such as heart failure. In this section, we provide a brief overview of relevant terminology and imaging modalities.

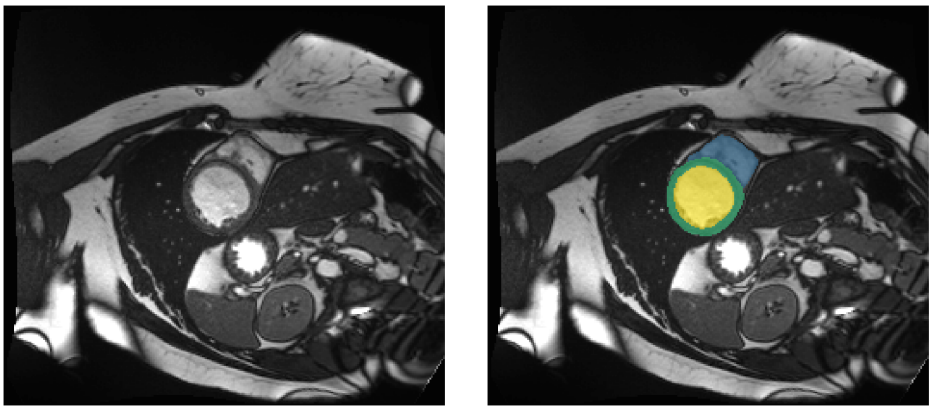
The heart is divided into 4 chambers: deoxygenated blood enters the right atrium, then the right ventricle. It gets pumped to the lungs for oxygenation, then enters the left atrium and left ventricle. The left ventricle pumps the oxygenated blood which circulates throughout the body. The 4 chambers are surrounded by myocardium, the muscle tissue of the heart. Irregularities include dilated cardiomyopathy (enlarged left ventricle) which can lead to weakened

[8] More formally, the Normal prior is defined as $p(z_l^i|z_{<l}) = \mathcal{N}(\mu_i(z_{<l}), \sigma_i(z_{<l}))$ where l is the latent layer index and i is the layer group index. The prior is dependent on previous layers and we have omitted the dependency notations for simplicity.

myocardium and blood clots, hypertrophic cardiomyopathy (thickened myocardium) which can cause heart failure, and abnormal right ventricle which can cause liver congestion.

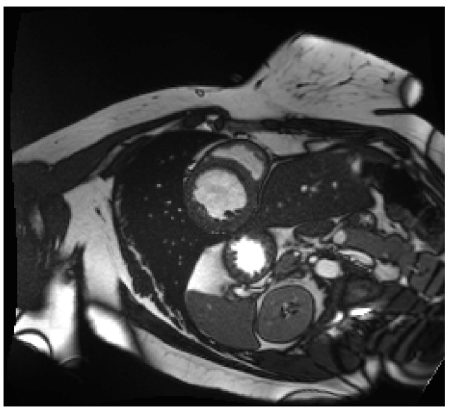
Magnetic resonance imaging (MRI) is a medical imaging technique that uses magnetic fields and radio waves to produce images of internal body structures. MRI and computed tomography (CT) scans are widely used in cardiac imaging. Unlike CT, MRI does not involve ionising radiation (X-rays), making it a safer procedure. However, MRI is more time-consuming - a full cardiac MRI scan takes 30-60 minutes.

A single MRI scan is divided into multiple 2D slices, with each slice representing a cross-section at a particular depth. The slices are stacked to form a 3D volume. A voxel is a 3D pixel, representing a unit of volume. In general, an MRI scan can obtain images in the axial, coronal or sagittal planes, depending on the slice orientation. These planes are relative to the body. However, for cardiac imaging, planes relative to the heart are used: short-axis, long-axis and 4-chamber views. In particular, the short-axis view gives the best cross-section of the left and right ventricles. In addition, the heart is an active muscle and its shape changes over a cycle. Cardiac cine-MRI involves repeated imaging of the heart over a short period of time, from which

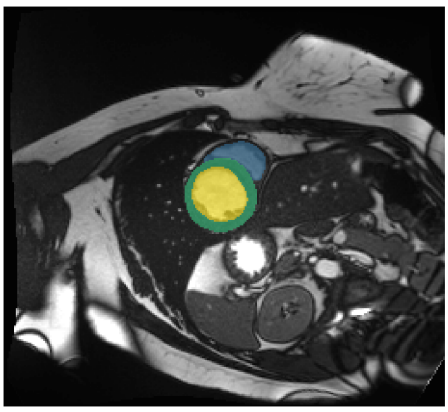


(a) End-diastole (ED)

(b) ED ground truth



(c) End-systole (ES)



(d) ES ground truth

Figure 2.8: A 2D slice of a cardiac cine-MRI scan, presented at both ED and ES phases. Acquired in the short-axis view from a healthy patient. The ground truth segmentations (overlayed with opacity) label the left ventricle (yellow), right ventricle (blue) and myocardium (green). Sourced from the ACDC dataset [14].

two phases can be extracted: end-systole (ES) and end-diastole (ED). ES refers to the heart after contraction, while ED refers to the heart after relaxation. Figure 2.8 presents a preview of a short-axis slice.

2.4 Shape Encoding

Shape encoding is the process of representing shapes compactly. In the field of cardiac imaging, shape encoding is used to encode the segmentation mask of the heart as a low-dimensional vector. An accurate segmentation of the heart provides crucial information about its structure and potential irregularities. A model that can encode these masks acts as a feature extractor, and can be applied in downstream tasks. In this section, we overview 4 works that use VAEs to encode medical segmentation masks for different applications.

2.4.1 Explainable Anatomical Shape Analysis

In 2019, Biffi et al. proposed a deep learning pipeline for interpretable shape analysis of cardiac and brain segmentations [5]. The pipeline takes a previously segmented mask and encodes it using a Ladder VAE (Section 2.2.1). The top-level latent space is chosen to be 2D, and the latent representation is fed into a small MLP classifier to predict or diagnose a condition. Figure 2.9

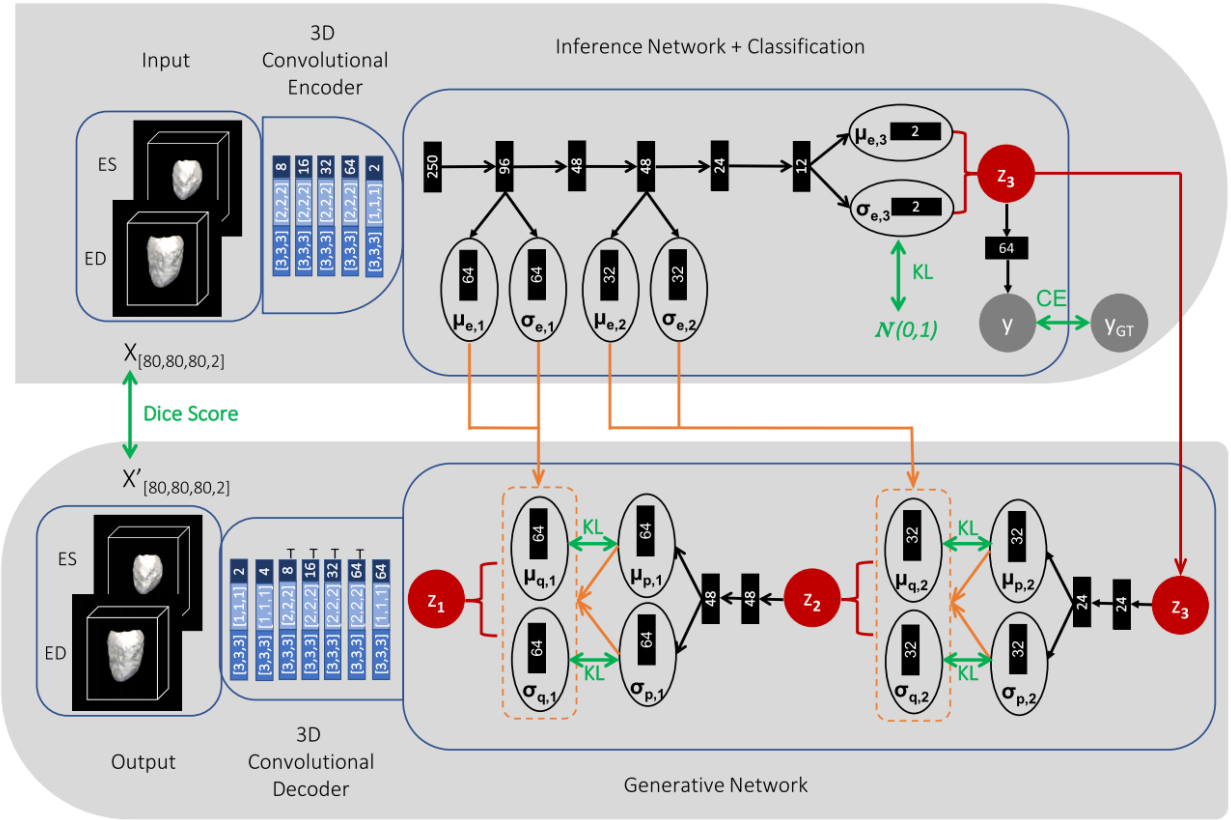


Figure 2.9: Pipeline for interpretable shape analysis involves passing the end-systole and end-diastole 3D segmented images through a convolutional encoder which compresses them each into a 250-dimensional embedding. This is then fed into the LVAE+MLP architecture. Top diagram shows the encoder and bottom diagram shows the decoder. [5, p. 6]

presents the pipeline in more detail. Hence, the latent space, which acts as the classification feature space, can be directly visualised (1) due to the generative properties of the LVAE which gives variability, and (2) without needing additional dimensionality reduction as it is 2D.

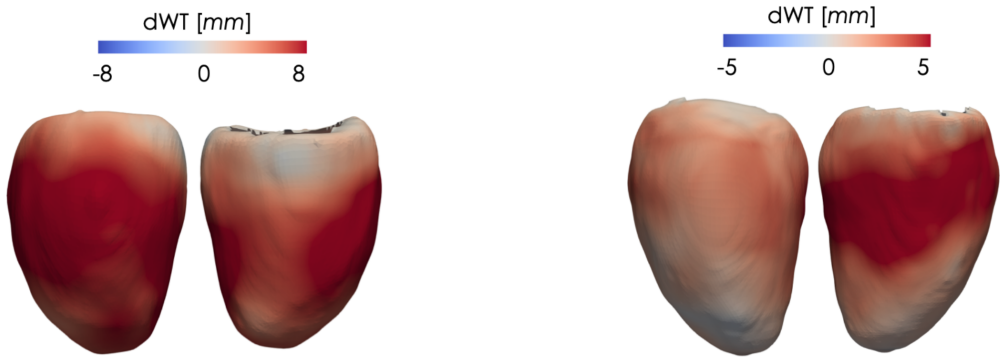
The LVAE and MLP are trained end-to-end using a weighted loss function presented by (2.10), where DSC_{ES} and DSC_{ED} are the Dice coefficients for the end-systole and end-diastole phases to measure reconstruction quality. CE is the cross-entropy loss for the MLP classifier, L is the number of hierarchical layers (3 in the paper) and γ is the linear warm-up for LVAE training. α_i are hyperparameters that weight the KL terms.

$$\mathscr{L}(\theta,\phi,\alpha;x) = DSC_{ES} + DSC_{ED} + \gamma\left(\sum_{l=0}^{L} \alpha_l D_{KL}\left[q_\phi(z_l|x_l)||p_\theta(z_l)\right] + \beta CE\right) \tag{2.10}$$

Using a small MLP limits classifier capacity, which forces the top-level latent space to capture the most discriminative features for diagnosis.

Figure 2.10 presents the learned top-level latent space for cardiac application. 2 distinct clusters form, corresponding to the healthy and HCM classes. Figure 2.11 presents the average shapes of the lateral and septal walls for the healthy and HCM classes.

Figure 2.10: Top-level latent space for cardiac application, corresponding to dimensions 1 and 2 in the plots. Distinct clusters form for the train and test sets, while weaker clusters form for the additional dataset (ACDC[14]). HVol refers to the healthy class.[5, p. 7]



(a) ES (b) ED

Figure 2.11: Wall thickness (dWT) point-wise differences between healthy and HCM average shapes. The average shapes are generated by the LVAE decoder[5, p. 8]

Overall, the pipeline allows for the classification tasks to be more transparent while maintaining high performance: 100% sensitivity and specificity on the in-distribution test set for cardiac diagnosis of healthy and HCM patients, 78% sensitivity and 90% specificity for brain diagnosis. Additionally, the model trained for cardiac application achieves 100% sensitivity and 80% specificity on an additional dataset[9] without retraining. The pipeline produces improved results in reconstruction quality over the previous work[4], which used a non-hierarchical VAE+MLP pipeline.

2.4.2 Cardiac Segmentation with Strong Anatomical Guarantees

In 2020, Painchaud et al. proposed a VAE-based pipeline for automatic correction of cardiac segmentations[6]. The method involves training a VAE that takes in a segmentation mask and identifies anatomically invalid shapes[10], then warps them to the nearest valid shape. At inference time, the trained VAE is appended to an existing segmentation model to ensure the inference pipeline produces anatomically valid segmentations.

Figure 2.12 presents the train and inference pipeline. The authors use a constrained VAE[30], with the constraint being a 1-neuron linear network $y_{\theta_c}(\cdot)$ trained simultaneously with

[9] Automated Cardiac Diagnosis Challenge, MICCAI 2017[14]

[10] The authors define a list of anatomically invalid configurations. For example, for a short-axis segmentation, this includes hole(s) in the left ventricle, right ventricle or myocardium.

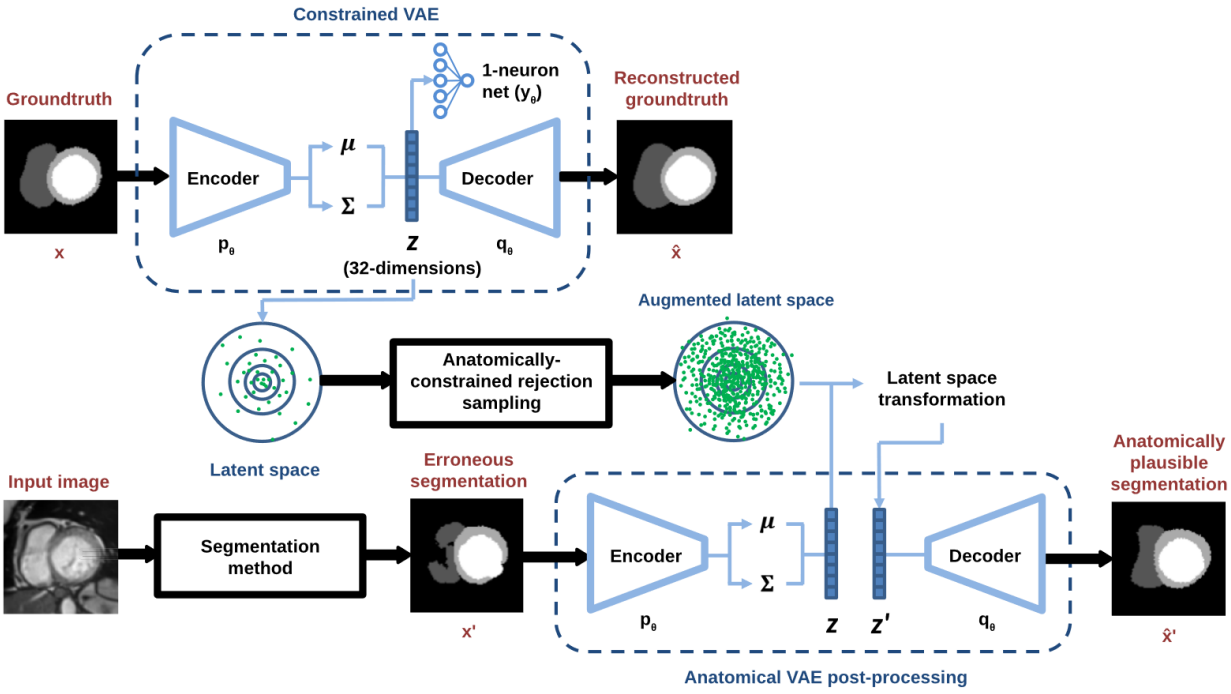


Figure 2.12: Top diagram presents training pipeline that trains a VAE to learn latent representations of gold standard anatomically valid segmentations. The latent space is augmented with more valid representations by linearly interpolating between valid vectors and performing small translations. This latent space is used to correct anatomically invalid segmentations at inference time via warping to the nearest valid shape (bottom diagram). The segmentation method refers to any existing segmentation method. [6, p. 3]

the encoder-decoder. This forces the learned latent space to be more linear and less convoluted, so close points in the latent space correspond to similar reconstructed shapes, which is important for the correction step. The loss function (2.11) is made up of the ELBO and the L2 loss of the constraint.

$$\mathscr{L}(\theta,\phi;x) = \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - D_{KL}\left[q_\phi(z|x)||p_\theta(z)\right] + \|y_{\theta_c}(z) - t\|^2 \tag{2.11}$$

Overall, the pipeline can correct anatomically invalid segmentations without detrimenting performance of various segmentation models (as measured by Dice coefficient, Hausdorff distance and ejection fraction of right and left ventricles).

The fact that the proposed pipeline does not improve segmentation performance while correcting inaccuracies is surprising. A possible explanation is that some quality is lost during the reconstruction stage, which is made up for by the corrections. Furthermore, the framework only uses linear interpolation between valid latent encodings to generate synthetic samples for warping. It does not make use of generating samples from the assumed prior only, and hence the augmented latent space is potentially limited. Using a more expressive model such as a hierarchical VAE could improve reconstruction and generation quality to address both shortcomings.

2.4.3 Probabilistic U-Net

In 2018, Kohl et al. proposed the Probabilistic U-Net[7], a segmentation framework that can provide multiple hypotheses for a single image. In practice, computer vision tasks are often ambiguous. For example, given a CT scan the region of a lesion that is cancerous is subject to inter-observer variability. A model that outputs multiple segmentation maps can provide more information for a clinician to recommend further action than traditional deterministic models. Furthermore, the Probabilistic U-Net can model the joint probability of all pixels in a segmentation. This improves upon models that can only provide pixel-wise probabilities, which ignores

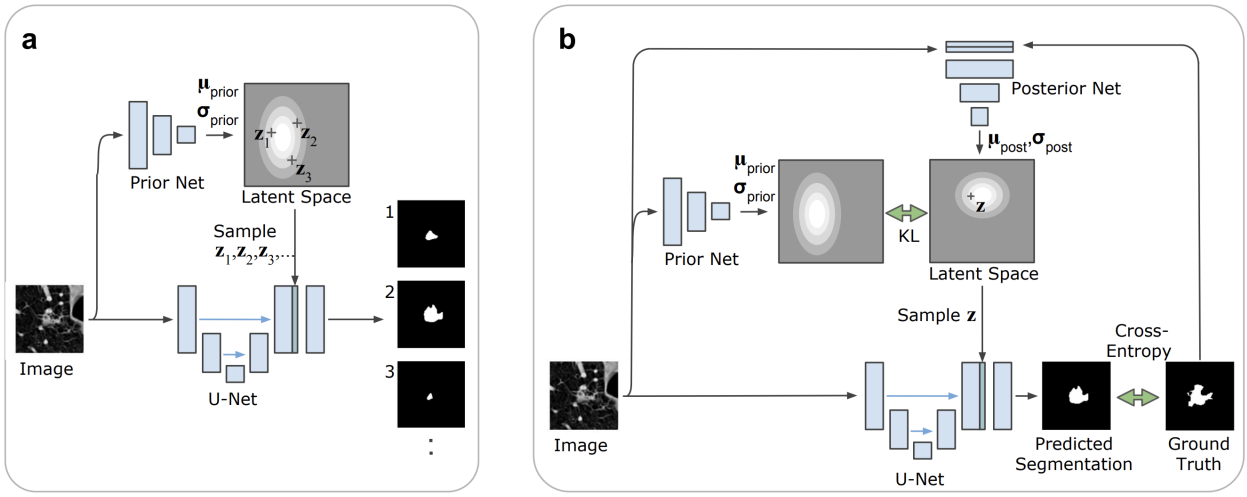


Figure 2.13: The Probabilistic U-Net framework, depicting (a) the sampling process, and (b) the train process. For each execution, a single z is sampled to predict a mask. Blue blocks denote feature maps. Green arrows denote loss. During training, a posterior net is used to recognise segmentation variants and map them to a position (distribution) in the latent space.[7, p. 2]

co-variance between pixels. The framework achieves state-of-the-art performance among existing probabilistic methods in lung lesion segmentation as well as natural image segmentation tasks.

The framework extends a U-Net[56] by introducing a Conditional VAE that learns a low-dimensional latent space to encode the segmentation variants (Figure 2.13). The hierarchical space is interleaved with the U-Net decoder. At inference time, random samples in the latent space levels are injected into the U-Net to produce various segmentation hypotheses.

In 2019, Kohl et al. extended the Probabilistic U-Net to the Hierarchical Probabilistic U-Net[8], which uses a CVAE with a hierarchical latent space. This results in improved fidelity, particularly in modelling fine details of lesion shapes, allowing the model to perform well in more complex tasks like instance segmentation.

2.4.4 Anatomically Constrained Neural Networks for Cardiac Imaging

In 2017, Oktay et al. proposed the Anatomically Constrained Neural Network (ACNN) for multi-modal cardiac imaging[9]. Previous existing segmentation methods involve training a model at a pixel-wise level where the objective (cross-entropy) does not involve shape and structure information. ACNN introduces a shape regularisation term to the objective, which allows shape information to be injected into a neural network, encouraging it to learn global anatomical features.

One application involves integrating a convolutional autoencoder f with a segmentation model ϕ (Figure 2.14). During training, ϕ outputs a predicted mask $\hat{y}$ given a scan, and f learns a low-dimensional latent representation of $\hat{y}$, as well as the ground truth y.

The loss formulation is presented in (2.12). The shape loss (scaled by hyperparameter λ_1) is the distance between the latent encodings of $\hat{y}$ y. x is the scan, and w denotes convolution

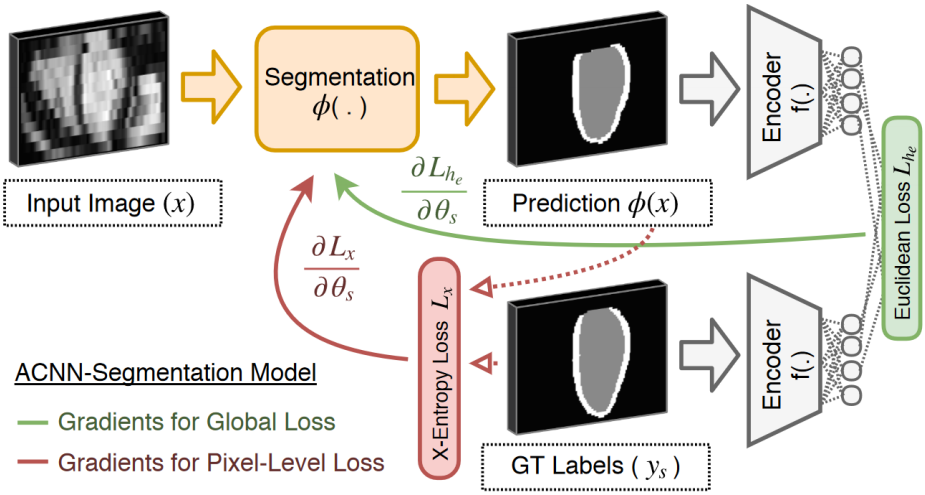


Figure 2.14: Anatomically Constrained Neural Network (ACNN): Training pipeline for image segmentation tasks. The encoder f is a convolutional autoencoder network that is trained end-to-end with the segmentation model ϕ.[9, p. 4]

weights and is a weight-decay term.

$$\mathscr{L}(\theta, \lambda_1, \lambda_2; x, y) := CE(y, \hat{y}) + \lambda_1 \| f_\theta(\hat{y}) - f_\theta(y) \|_2^2 + \frac{\lambda_2}{2} \| w \|^2 \tag{2.12}$$

The autoencoder acts as a regularisation model to constrain the segmentation model towards learning global anatomical features, allowing it to be more robust against imaging artefacts, noise and slice misalignment. ACNN outperforms previous existing models in segmenting the endocardium and myocardium in 2D cardiac MRI scans, as well as the left ventricle cavity in 3D-US cardiac image sequences. Overall, ACNN demonstrates the benefits of incorporating prior shape information into mainstream neural network architectures. As the work uses a single-layer autoencoder with limited capacity, it remains to explore the full potential of using prior information from more expressive models like hierarchical VAEs.

2.5 Concluding Remarks

We have overviewed a selection of works on generative models, with a focus on the development of VAE frameworks due to their ability to learn an explicit compact representation of complex data. Automated learning of such representations is an active area of research in the medical imaging domain, and we have reviewed and criticised key works that leverage shape encoding for cardiac imaging tasks, including improving interpretability of classifying pathologies [5], automatic correction of segmentation masks [6], probabilistic segmentation [7, 8] and using the learned representations to enforce anatomical constraints [9].

The reviewed works demonstrate the importance of shape encoding for cardiac imaging. A shared limitation is the use of single-layer VAEs, which has limited expressive power. Indeed, Kohl et al. presents improved results in using hierarchical VAEs for probabilistic segmentation [8].

In particular, the Nouveau VAE framework [13] has shown promising results in producing both high fidelity reconstructions and generations, suggesting it is capable of learning stronger latent representations more suitable for downstream tasks. An open challenge, therefore, is to explore the potential of NVAE when applied to cardiac imaging.

Chapter 3

Data Analysis

We will primarily be using the Automated Cardiac Diagnosis Challenge (ACDC) dataset[14]. The challenge was introduced during the MICCAI 2017 conference and aims to promote research in developing automated methods for segmenting and classifying cardiac MRI scans.

We will also be using the Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image Segmentation Challenge (M&Ms) dataset[15, 16]. We use this dataset for domain adaptation experiments by transferring learnings from the ACDC dataset. The M&Ms challenge was introduced during the MICCAI 2020 conference and provides cardiac imaging data from multiple clinical centres and acquisition protocols to promote research in developing robust, generalisable models across different clinical settings.

3.1 Data Overview

The ACDC dataset consists of 150 cardiac cine-MRI scans in the short-axis orientation, obtained over a 6-year period at the University Hospital of Dijon (France). A manual segmentation of the left ventricle (LV), right ventricle (RV) and myocardium (MYO), jointly annotated by 2 medical experts, is provided for each scan. The M&Ms dataset includes 320 publicly available short-axis cine-MRI scans obtained from 5 centres in Spain and Germany, with segmentation labels for the LV, RV and MYO, manually annotated by an expert from the centre of origin. In addition, the M&Ms dataset provides 25 unlabelled scans, which we will not use. It also provides 30 additional scans from a 6th centre in Canada that is not publicly available.

For both datasets, each scan includes an end-diastolic (ED) and end-systolic (ES) frame. Each frame consists of a stack of 2D slices. The number of slices differs per scan, averaging around 10 per frame. We have previously presented a slice preview (Figure 2.8). We will use the manually annotated masks as the ground truth (GT) for our experiments.

3.1.1 Study Population

ACDC - The study population consists of 150 patients categorised into 5 pathologies. Each patient contributes 1 MRI scan. The study population has been partitioned into a train set and a test set. The train set has 100 patients, with 20 patients from each class. The test set has 50 patients, with 10 patients from each class.

M&Ms - The study population consists of 320 patients categorised into 9 pathologies. Each patient contributes 1 MRI scan. The study population has been partitioned into a train set, validation set and test set, with varying number of patients from each centre and class.

Appendix A provides details on class and centre distribution.

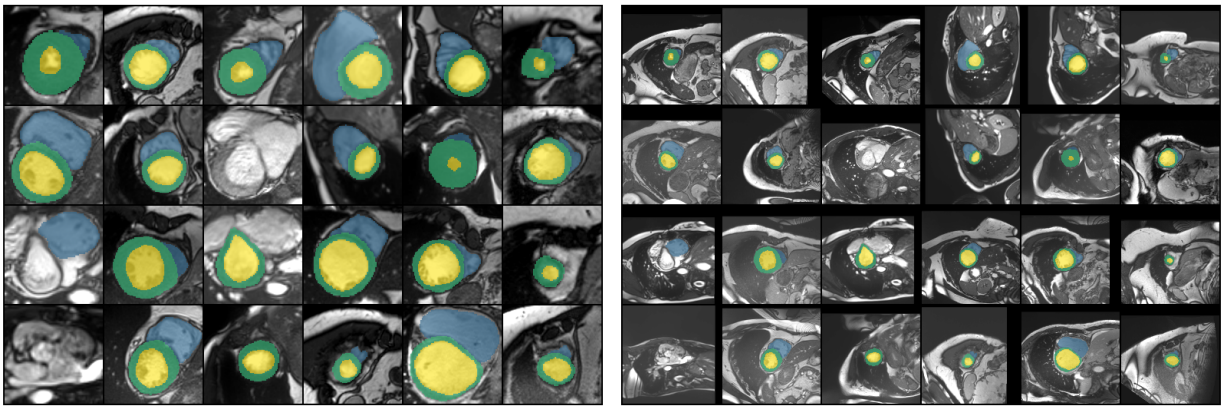
3.2 Preprocessing

3.2.1 Pipeline

ACDC - We perform preprocessing to produce a dataset where each data point is an MRI slice with its corresponding GT segmentation mask. Width, height and voxel spacing of each slice are consistent across the same frame and patient, but differs between patients. We standardise the dimension to 128×128 pixels with the following algorithm.

1. **Crop and pad:** For each frame, we compute the coordinates of the smallest bounding box such that for all masks, all information (LV, RV, MYO) is contained within it. Let this box have dimensions $w \times h$. Without loss of generality, assume $w \geq h$. We crop each slice and mask to $(w+4) \times (w+4)$ around the centre, with 4 pixels of padding on each side.
2. **Resize:** Each cropped slice and mask are resized to 128×128 with nearest-neighbour interpolation.
3. **Rescale intensity:** We perform min-max scaling on each slice by clipping to the $[1, 99]$ intensity percentiles, then linearly mapping to the range $[0, 1]$. This helps resolve pixel-level artefacts introduced during acquisition.
4. **One-hot encode mask:** At this stage, each mask has shape 128×128 with 4 unique values: 0 (background), 1 (RV), 2 (MYO) and 3 (LV). We one-hot encode each mask to $4 \times 128 \times 128$.

Since the bounding box is computed per frame and not per slice, voxel spacing is the same for all slices within the same patient, but may differ between patients.



(a) Preprocessed data (b) Preprocessed data without crop-and-pad

Figure 3.1: Preprocessed slices and masks from the ACDC dataset. The GT mask (overlayed with opacity) segments the slice into the LV (yellow), RV (blue), MYO (green).

Figure 3.1 presents a preview of the preprocessed data points. It also presents the slices with the crop-and-pad stage omitted[1] to illustrate data sparsity in the unprocessed dataset. For the latter, the masks consist of 95.3% background across the train and test set, reduced to 65.9% with crop-and-pad. Filtering part of the background is a standard preprocessing step to encourage models to focus on the structures of interest: LV, RV and MYO. Furthermore, we do not want the background to dominate the reconstruction loss during model training.

M&Ms - We apply the same preprocessing pipeline as ACDC. Volumes from the M&Ms dataset may contain an empty slice-wise data point, where both the MRI slice and the mask are empty. We filter such occurrences from all partitions.

More previews of ACDC and centre-specific M&Ms data can be found in Appendix E.1.

3.2.2 Data Partitioning

ACDC - The pipeline yields 2,978 data points: 1,902 from the train set and 1,076 from the test set. We further partition the train set into a smaller train set and a validation set with a 9:1 split (1,711 and 191 data points respectively).

M&Ms - The pipeline yields 6,880 data points: 3,089 from the train set, 754 from the validation set and 3,037 from the test set.

3.2.3 Segmentation Class Distribution

Each segmentation mask has 4 classes: RV, MYO, LV and background. Across the ACDC dataset, 34.1% of the pixels are non-background: 11.2% RV, 11.9% MYO and 11.0% LV.

Figure 3.2 presents a heatmap of the class distribution.

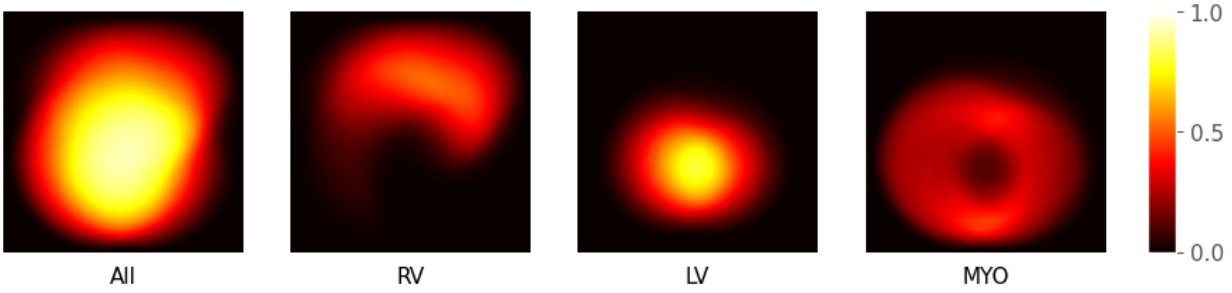


Figure 3.2: Heatmap of segmentation class distribution across the ACDC dataset. The colour of each pixel denotes the percentage that it is occupied by the class.

Across the M&Ms dataset, 27.3% of the pixels are non-background: 8.58% RV, 9.50% MYO and 9.19% LV.

3.2.4 Other Remarks

Out of the 1,711 data points in the ACDC train set, 55 have empty masks[2]. Indeed, these can be observed in Figure 3.1. This is due to the clinical protocol to not segment ambiguous slices.

[1] But with original aspect ratio retained by cropping to min(width, height).
[2] Masks with only background.

These are often basal and apical slices. The basal slice is the topmost slice representing the base of the heart[3], and the apical slice is the last slice with visible ventricular cavity. As such, the dataset imposes a challenge for segmentation models to recognise and not segment these slices.

The disparity in class distribution between ACDC and M&Ms is due to a significantly higher prevalence of empty masks in M&Ms: 1,591 out of 6,880 data points (21.8%). The class distribution among non-empty masks in M&Ms is similar: 35.3% of the pixels are non-background, with 11.0% RV, 12.4% MYO and 11.9% LV.

The heatmap in Figure 3.2 reveals slight misalignment. In particular, the RV is angled towards the top-right corner. On top of this, initial experimentation suggests that some VAE models struggle with generating diverse synthetic masks. To address both issues, we tried adding alignment registration by rotating the masks in the preprocessing pipeline, and rotation augmentation at train time. However, neither additions improved (nor worsened) overall performance, so we decided to omit them.

3.3 Cardiac Shape Anatomical Validity

We discuss a systematic approach to determine whether a cardiac segmentation mask is considered anatomically valid. From this, we can compute the percentage of anatomically valid masks in a set, or % AV. This gives us (1) a metric for evaluating generation quality of generative models, and (2) a metric, along with Dice coefficient, for evaluating quality of segmentation models, among other use cases.

3.3.1 Methodology

We use a fixed set of criteria designed for short-axis cardiac masks from Painchaud et al.[6]. A mask is considered anatomically invalid if it has any of the following: (1) hole(s) in the LV, RV or MYO, (2) hole(s) between LV and MYO, (3) hole(s) between RV and MYO, (4) more than 1 LV, RV or MYO, (5) RV is disconnected from MYO, (6) LV touches RV, (7) LV touches background.

3.3.2 Validity of Dataset

ACDC - Out of 2,978 masks in the dataset, 18 masks are anatomically invalid. 100% of the invalid masks are present in the original, unprocessed dataset and not caused by resolution downsampling in the preprocessing pipeline. 17 masks are invalid due to having an extra RV that is a few pixels wide (Figure 3.3a). The other invalid mask is a basal slice (Figure 3.3c).

All segmentation masks are gold standard annotations by medical experts. The criteria provides a strong guideline but does not guarantee correct classification. We choose to preserve all masks and use the percentage of valid masks in the dataset (99.4%) as a gold standard for generative models to aim for.

M&Ms - Only 69.3% of the masks are valid. This percentage becomes 69.1% for the unprocessed dataset, so resolution downsampling in the preprocessing pipeline does not contribute

[3]The Society for Cardiovascular Magnetic Resonance (SCMR) classifies the top slice with >50% myocardium around the blood during ED phase as the basal slice[57].

(a) (b) (c)

Figure 3.3: Examples of anatomically invalid masks from ACDC: (a) a mask with an extra RV, (b) a zoomed-in view of the extra RV, and (c) a mask of a basal slice; it is invalid as the LV touches the background.

to a lower % AV. Table 3.1 presents a summary and Figure 3.4 presents an example. The invalid masks are due to pixel-level violations of the criteria, rather than global corruptions.

	Centre					
	1	**2**	**3**	**4**	**5**	**All**
% AV	86.3	55.6	53.7	76.4	63.2	69.3

Table 3.1: % anatomical validity of the M&Ms dataset across each centre.

(a) (b)

Figure 3.4: An anatomically invalid mask from M&Ms: (a) a mask with holes between the RV and MYO, and (b) a zoomed-in view.

Chapter 4

Experiments

This chapter is organised as follows. Firstly, we describe our setup for cardiac shape encoding with VAE frameworks. This consists of single-layer VAE models and Nouveau VAE. The performance of the former is used as a baseline benchmark. Secondly, we present a downstream application of the learned latent representations by using them to formulate a shape regularisation term to train an anatomically constrained cardiac segmentation model. We use the ACDC dataset for the experiments listed above. Finally, we apply the same segmentation framework with data from a different source to investigate whether the learned representations can be used for domain adaptation and few-shot learning. For this task, we use the M&Ms dataset.

4.1 Baseline Models for Shape Encoding

We use 3 approaches to train baseline models, all of which use the same single-layer VAE architecture. The first approach uses β-VAE loss. The second approach uses InfoVAE loss, and uses an auxiliary discriminator to approximate $D_{KL}\left[q_\phi(z)||p_\theta(z)\right]$. The third approach also uses InfoVAE loss, but uses minibatch sampling to approximate $D_{KL}\left[q_\phi(z)||p_\theta(z)\right]$.

4.1.1 Architecture Overview

The single-layer VAE architecture is adapted from the framework proposed by Painchaud et al.[6]. The overarching principle of the architecture is that of a standard encoder-decoder network (Figure 2.2a, Figure 2.3). During training, each model takes a one-hot encoded cardiac segmentation mask as input and outputs its reconstruction.

Figure 4.1 presents the architecture. The encoder consists of 4 blocks, followed by an FC layer that outputs the mean and log variance. Each block consists of two 3×3 convolutional layer with ELU activation. The first convolutional layer has a stride of 2 and doubles the number of feature maps. The decoder has a similar but reversed structure.

4.1.2 Implementation Details

β-VAE is a framework that builds on top of VAE by extending the ELBO with a KL weight term. The objective is presented in (2.3). Its drawbacks are discussed in Section 2.1.3, which motivates

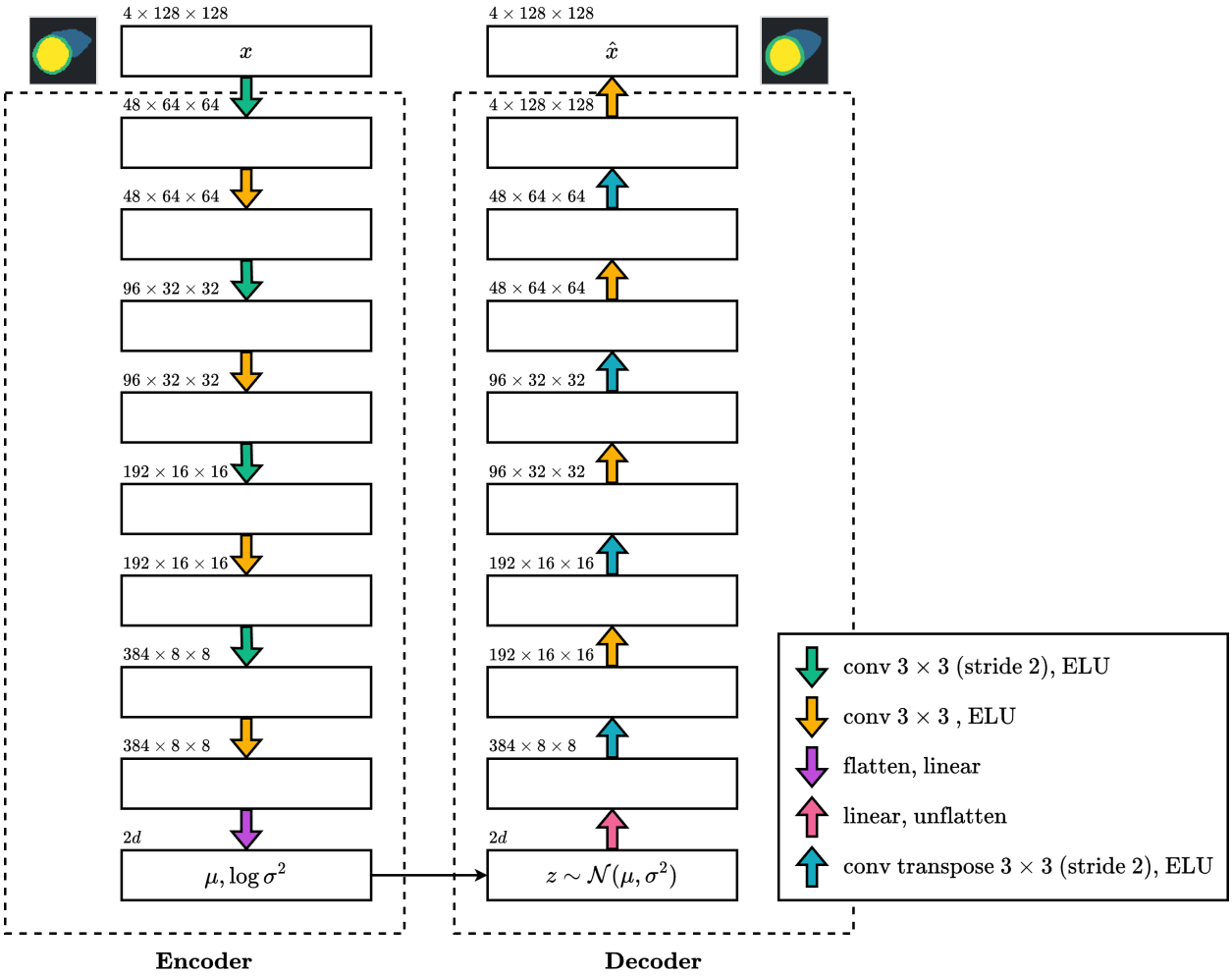


Figure 4.1: VAE architecture for baseline models. A white box represents a feature map (or input/output tensor), with its shape denoted on the top-left corner. A coloured arrow denotes an operation or sequence of operations. d is the dimensionality of the latent space. It is a tunable hyperparameter.

the InfoVAE framework (Section 2.1.4). Its objective is presented in (2.6). We implement this in two ways: by approximating $D_{KL}\left[q_\phi(z)||p_\theta(z)\right]$ with an auxiliary discriminator, and with minibatch sampling. The former is proposed by the authors of InfoVAE [33]. We derive the latter below; it is similar to how TC is estimated in β-TCVAE (Section 2.1.3).

$$\begin{aligned} D_{KL}\left[q_\phi(z)||p_\theta(z)\right] &= \int_z q_\phi(z)\log\frac{q_\phi(z)}{p_\theta(z)}dz \\ &\approx \frac{1}{N}\sum_{n=1}^{N}\log\frac{q_\phi\left(z^{(n)}\right)}{p_\theta\left(z^{(n)}\right)} \quad z^{(n)}\sim q_\phi(z) \\ &= \frac{1}{N}\sum_{n=1}^{N}\left(\log q_\phi\left(z^{(n)}\right)-\log p_\theta\left(z^{(n)}\right)\right) \end{aligned}$$

We refer to the discriminator approach as **InfoVAE-D** and the minibatch approach as **InfoVAE-M**. For InfoVAE-D, we design a 3-layer MLP discriminator with hidden dimension 8 and leaky

Fixed		**Tunable**	
Hyperparameter	**Value**	**Hyperparameter**	**Value**
Batch size	16	Latent dimension	2 – 64
Optimiser	Adam	β-VAE: β	0.01 – 10,000
Learning rate	6×10^{-5}	InfoVAE: β	0 – 5
Weight decay	10^{-2}	InfoVAE: γ	1 – 10,000
InfoVAE-D Discriminator			
Optimiser	Adam		
Learning rate	6×10^{-5}		
Weight decay	10^{-2}		

Table 4.1: Hyperparameter configurations for training baseline models. Only a subset of hyperparameters are tuned due to time constraints. Untuned parameters have not been systematically searched for optimal settings and are marked as "Fixed"; tuned parameters are marked as "Tunable" with the range of searched values. InfoVAE refers to both InfoVAE-D and InfoVAE-M.

ReLU activation (slope 0.2).

Generation process: Producing synthetic masks at inference time requires no additional data. We sample a latent vector from the prior: $z \sim \mathcal{N}(0, I)$, then pass it through the decoder to obtain the synthetic mask (Figure 2.2b).

At inference time, the output $\hat{x}$ is discretised by applying the argmax function to each pixel to obtain a non-probabilistic reconstruction (or generation).

Apart from the difference in loss formulation, the training procedure of the 3 approaches is the same. Table 4.1 presents the configurations. We train for 50 epochs and take the model checkpoint with the lowest validation loss.

4.2 Nouveau VAE for Shape Encoding

4.2.1 Architecture Overview

We design two architectures for cardiac shape encoding, which we refer to as **Default-N** and **LatentSkip-N**. Both are adapted from the original work[13].

Figure 4.2 presents the Default-N architecture, which takes in a one-hot encoded cardiac segmentation mask and outputs its reconstruction. Each architecture consists of a stem, encoder, decoder and conditional coder. The stem is a 3×3 convolutional layer that maps the 4-channel input to a projected space. This helps prevent an initial bottleneck that may cause early information loss. The conditional coder maps the projected space back to the original space. The Default-N encoder-decoder consists of a series of preprocess/postprocess blocks, followed by a shared 3-layer tower with 3 latent spaces. The LatentSkip-N encoder-decoder consists of a 5-layer tower, with every other layer having an explicit latent space. The number of feature channels double at each layer, including preprocess layers. See Section 4.2.2 for details.

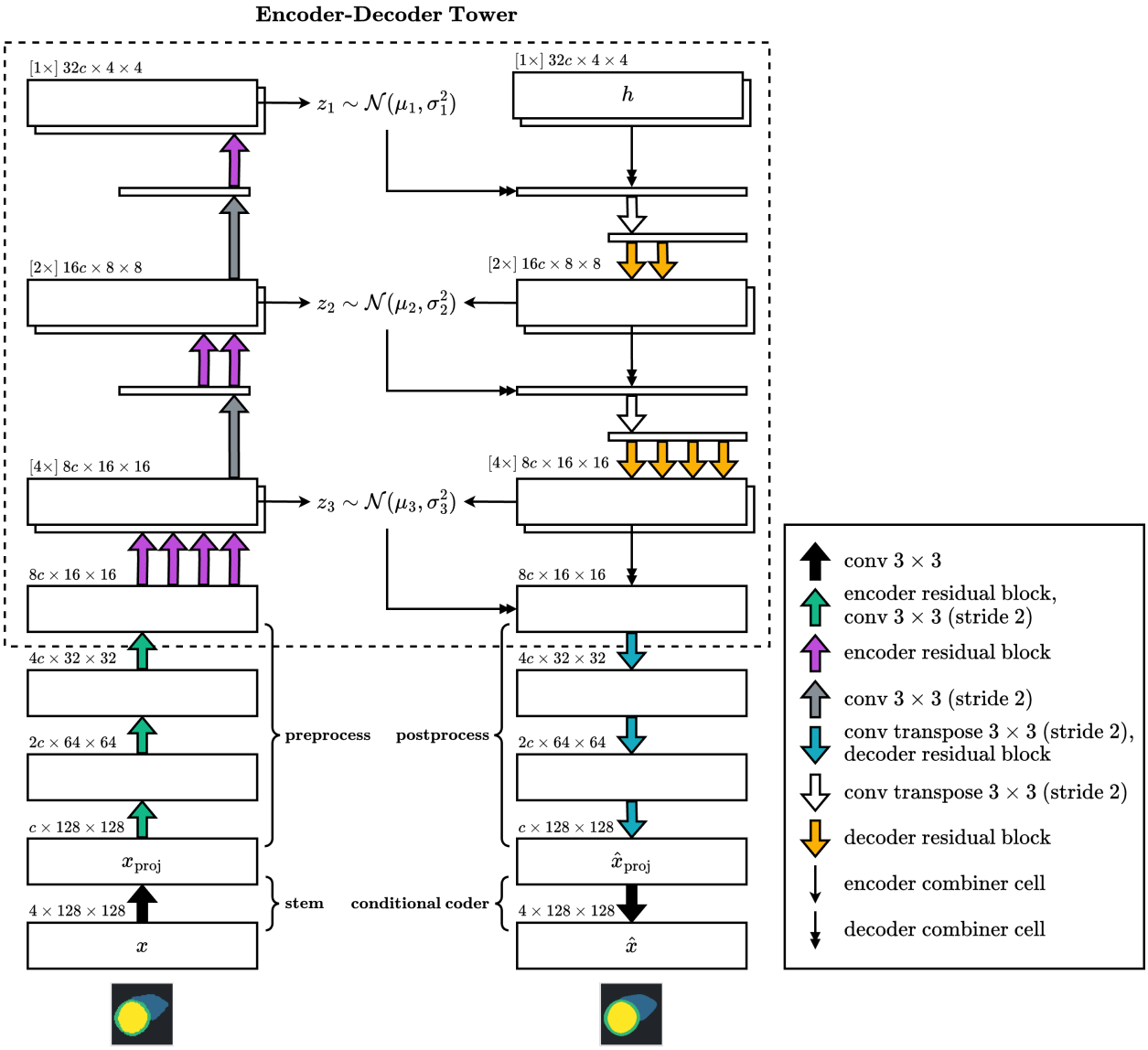


Figure 4.2: NVAE Default-N architecture. A white box represents a feature map (or input/output tensor), with its shape denoted on the top-left corner. A coloured arrow denotes an operation or sequence of operations. c is the projected channels. h is a learnable parameter. The encoder and decoder combiner cells are designed to take in 2 feature maps and combine them together, using a 1×1 convolutional layer to match the number of channels. LatentSkip-N has a similar structure, except with 1 preprocess/postprocess block and a 5-layer tower with 3 latent spaces.

4.2.2 Implementation Details

The majority of our implementation matches the original work[13], which is reviewed in Section 2.2.2. In this section, we focus primarily on specific configurations for cardiac shape encoding, as well as any modifications to the framework. We also cover important mechanisms from the original work in more detail.

Residual block: We adopt the design from the original work[13]. The encoder residual block consists of 2 sequential encoder cells and the decoder residual block consists of 2 sequential decoder cells (Figure 2.7).

Residual parameterisation: Similar to other VAE frameworks, D_{KL} is computed using the reparameterisation trick. However, the shared encoder-decoder network allows z_i to be sampled from a residual distribution that combines the layer-wise prior $p_\theta(z_i|z_{<i})$ and layer-wise variational posterior, the latter of which is derived from the combined feats via the encoder combiner cell. The layer-wise approximate posterior $q_\phi(z_i|x_i, z_{<i})$ is defined as this residual distribution.

$$p_\theta(z_i|z_{<i}) = \mathcal{N}\left(\mu(z_{<i}), \sigma(z_{<i}))\right) \tag{4.1}$$

$$q_\phi(z_i|x_i, z_{<i}) = \mathcal{N}\left(\mu(z_{<i}) + \Delta\mu(z_{<i}, x),\ \sigma(z_{<i})\Delta\sigma(z_{<i}, x)\right) \tag{4.2}$$

At test time, we use deterministic sampling[1] to output the most accurate reconstruction.

Loss formulation: Our baseline results reveal InfoVAE to not improve upon β-VAE[2]. For NVAE, we extend the β-VAE ELBO loss such that each divergence term is scaled separately (see (4.3)). Each scaling factor is annealed from 0 to β_i over a warm-up period, where β_i is a tunable constant. The spectral regularisation term is omitted for clarity; we have implemented it with a static weight term as an optional configuration.

$$\mathcal{L}(\theta, \phi, \beta; x) := \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - \sum_{i=1}^{3} \beta_i \gamma_i \mathbb{E}_{q_\phi(z_{<i}|x)}\left[D_{KL}\left[q_\phi(z_i|x_i, z_{<i})||p_\theta(z_i|z_{<i})\right]\right] \tag{4.3}$$

γ_i is a dynamic balancing coefficient that is active during the warm-up period. It is proportional to the KL term of layer i as well as the size of the latent space. This mechanism is adopted from the original work[13].

Sharing information: The combiner cells play a crucial role in the sharing of information between the encoder and decoder. We mentioned how the encoder combiner cell allows the formation of a residual distribution that combines the prior and approximate posterior. The decoder path involves forming x_1 by combining the learned initial representation h with the latent vector z_1 via the decoder combiner cell. Subsequent layers involve combining x_i with z_{i+1} to form x_{i+1}. This mechanism is analogous to the original work[13].

Generation process: Producing synthetic masks at inference time requires no additional data. It involves traversing the decoder path with the following changes: (1) z_1 is sampled from the top prior $\mathcal{N}(0, I)$, and (2) for each latent layer we take $z \sim \mathcal{N}(\mu_p, \sigma_p^2)$. That is, the distribution is based purely on the prior.

Clamping: We find setting projected channels to a small number to have stable training. However, as noted previously, this may cause an early information bottleneck. To combat this, we propose a mechanism called clamping: at any point, if the feature map channels are less than 16, set it to 16. For example, if the channels in the preprocess layers are $4 \rightarrow 8 \rightarrow 16 \rightarrow 32$, it becomes $16 \rightarrow 16 \rightarrow 16 \rightarrow 32$.

At inference time, the output $\hat{x}$ is discretised by applying the argmax function to each pixel to obtain a non-probabilistic reconstruction or generation.

Table 4.2 presents the configurations for training Default-N and LatentSkip-N. We train for 100 epochs and take the model checkpoint with the lowest validation loss.

[1] $z = \mu_q$

[2] Quantitative results are presented in Table 6.1.

[3] $\beta_1 = 0.9, \beta_2 = 0.999$

Fixed		**Tunable**	
Hyperparameter	**Value**	**Hyperparameter**	**Value**
Warmup epochs	30	β_1	1 – 20
Batch size	8	β_2	1 – 20
Optimiser	Adamax[3], $\epsilon = 10^{-3}$	β_3	1 – 20
Learning rate	10^{-2}		
Weight decay	3×10^{-4}		
Scheduler	Cosine annealing, $\eta_{\min} = 10^{-4}$		
Latent channels	20		
Projected channels	4		
	Default-N	**LatentSkip-N**	
# Tower layers	3	5	
Latent layers index	1, 2, 3	1, 3, 5	
# Groups per layer	1, 2, 4	1, 2, 2, 4, 4	
Spatial dims per latent layer	$4^2, 8^2, 16^2$	$4^2, 16^2, 64^2$	
Preprocess layers	3	1	

Table 4.2: Hyperparameter configurations for training NVAE models. Only a subset of hyperparameters are tuned due to time constraints. The upper half presents shared configuration for both Default-N and LatentSkip-N. Untuned parameters have not been systematically searched for optimal settings and are marked as "Fixed"; tuned parameters are marked as "Tunable" with the range of searched values. Topmost layer is indexed as 1.

4.2.3 Other Considerations

The Default-N and LatentSkip-N architecture perform best out of various designs we worked with. For example, using multiple 3×3 convolutional layers in the preprocess/postprocess stage gives better results than using a single 9×9 convolutional layer. Furthermore, using groups of 1, 2, 4 in the 3-layer tower is more effective than using groups of 2, 4, 8. We find instance normalisation to perform worse than batch normalisation with the equivalent architectural designs. However, alternative methods were not explored thoroughly due to time constraints, in particular long training times.

4.3 Cardiac Segmentation with Shape Loss

We investigate a downstream application by using the learned latent representations of an NVAE model to formulate a shape regularisation term that regulates training a U-Net for cardiac segmentation. The motivation is described by Oktay et al.[9] (Section 2.4.4). To summarise, a U-Net is conventionally trained with cross-entropy loss, which is a pixel-wise loss and thus poor at capturing global shape information. By adding a regularisation term that is based on learned latent representations of GT segmentation masks, we introduce an anatomically constrained U-Net that is more attentive to the cardiac structure as a whole.

4.3.1 Implementation Details

We use the original U-Net architecture proposed by Ronneberger et al. [56], except we also add batch normalisation after each convolution layer for train stability (Figure 4.3).

Loss formulation: Let x be the input scan, y the ground truth segmentation mask and $\hat{y}$ the predicted (probabilistic) mask. The baseline is trained with cross-entropy loss. This is pre-

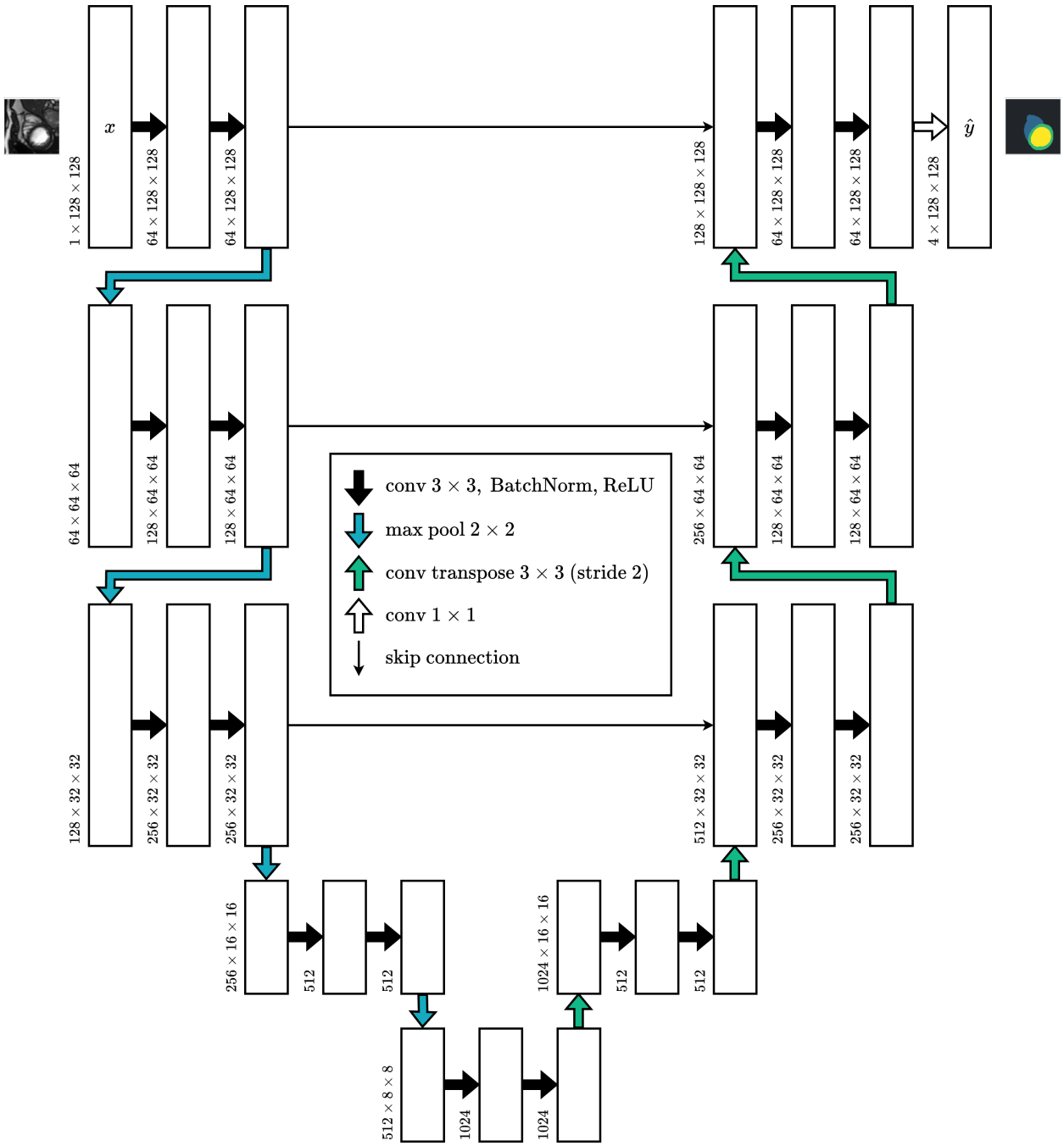


Figure 4.3: U-Net architecture. A white box represents a feature map (or input/output tensor), with its shape denoted in the corner. A coloured arrow denotes an operation or sequence of operations.

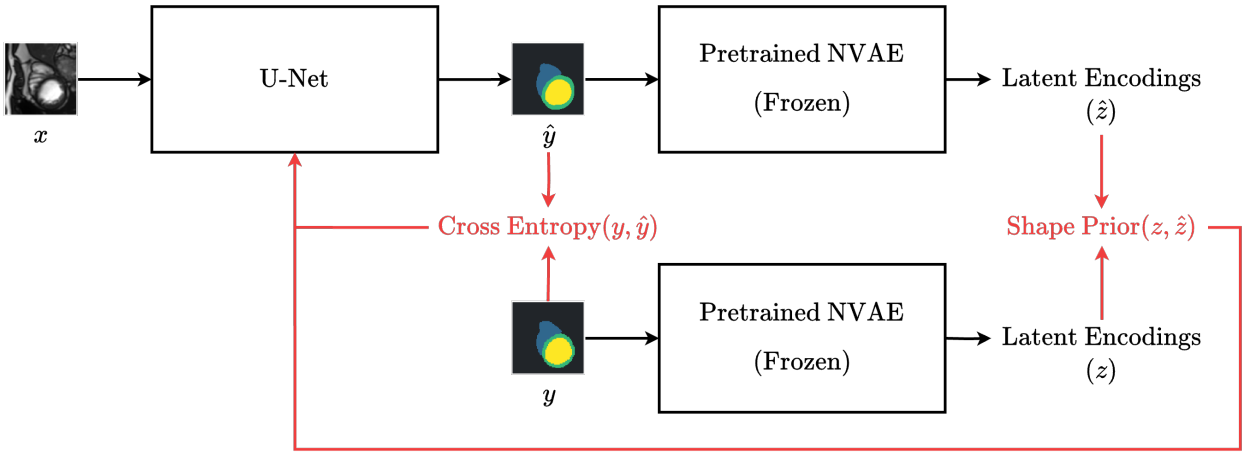


Figure 4.4: ACU-Net pipeline. x is the input scan (slice), y is the GT mask and $\hat{y}$ is the predicted mask. Red arrows denote backpropagation for U-Net training.

sented in (4.4), with width W, height H and number of classes C.

$$\text{CE} = \sum_{w=1}^{W}\sum_{h=1}^{H}\sum_{c=1}^{C}\left(y_{w,h,c}\log\hat{y}_{w,h,c}\right) \tag{4.4}$$

For our anatomically constrained U-Net, which we refer to as ACU-Net, we add a shape loss to the objective. We take a frozen, pretrained NVAE model and obtain the latent representations of y and $\hat{y}$ as z and $\hat{z}$ respectively. Each representation consists of 7 group-level latent vectors[4], which we denote as $z = \left[z_1 \cdots z_7\right]$ and $\hat{z} = \left[\hat{z}_1 \cdots \hat{z}_7\right]$. Each latent vector is defined as the mean of the corresponding residual distribution, $z_i = \mu_{i,q}$[5]. The objective is presented in (4.5) and the pipeline is presented in Figure 4.4.

$$\mathscr{L}(\theta;\theta_N,\phi_N,\alpha,x,y) = \alpha CE + \omega\sum_{i=1}^{7}\left\|z_i - \hat{z}_i\right\|_2^2 \tag{4.5}$$

We use the L2 norm for the shape loss, as we find the NVAE latent representations to have an empirically linear relationship with their corresponding reconstructions[6]. α is a tunable hyperparameter that weights the effect of cross-entropy with respect to the shape loss. ω is a fixed constant that scales the shape loss relative to cross-entropy. In short, cross-entropy is often orders of magnitude larger than the shape loss, so ω is set to a large value to make α easily interpretable. For our chosen pretrained NVAE model, we set $\omega = 126717$ so that $\alpha = 1$ results in equal contribution from both cross-entropy and shape loss. See Appendix B.1 for details.

Data augmentation: We apply random horizontal flip: the scan and mask are reflected across the y-axis with 50% probability. Data augmentation is applied during train time only. We investigated other augmentation techniques like random noise, gamma correction and equalisation, but did not find improved performance.

Table 4.3 presents the configurations. We train for 100 epochs and take the model checkpoint with the lowest validation loss.

[4] Corresponding to the latent groups of NVAE; see Table 4.2.

[5] This process is analogous to NVAE reconstruction at inference time: the same latent vectors are obtained and passed through the decoder tower to yield the reconstruction. See Section 4.2.2 for details.

[6] See Section 6.1.4 in the Evaluation chapter.

Fixed Hyperparameter	Value	Tunable Hyperparameter	Value
Batch size	32	α	0 – 1,000
Optimiser	Adam		
Learning rate	1×10^{-3}		
Weight decay	0		
# Layers	5		

Table 4.3: Hyperparameter configurations for training U-Net models. Only a subset of hyperparameters are tuned due to time constraints. Untuned parameters have not been systematically searched for optimal settings and are marked as "Fixed"; tuned parameters are marked as "Tunable" with the range of searched values.

4.3.2 Shape Loss Requirements

Similar to β-VAE, NVAE has a trade-off between reconstruction and generation quality caused by the strength of its regularisation terms. ACU-Net requires the NVAE model to have strong generative capabilities with solid reconstruction capabilities, for which we provide an explanation as follows.

The shape loss as presented in (4.5) measures the distance between latent representations of the GT mask and the predicted mask. In early stages of training, the predicted mask will have poor quality. Strong generative capabilities ensure that poor quality synthetic masks are very rarely produced, hence the latent representation of the predicted mask is many standard deviations away from the assumed prior. Following similar reasoning, the latent representation of the GT mask is close to the mean of the prior. Therefore, z_i and $\hat{z}_i$ are far apart, and the U-Net model is penalised heavily by the shape loss. Likewise, in later stages of training, the U-Net output is of higher quality and z_i and $\hat{z}_i$ are closer together, reducing the penalty.

In the above explanation, we assume the latent representation is an accurate representation of the corresponding mask. This is only guaranteed if the NVAE model has solid reconstruction capabilities, as the reconstruction is unambiguously defined by the latent representation.

4.4 Domain Adaptation and Few-Shot Learning

In Section 4.3, we introduce the ACU-Net framework with a shape regularisation term. This term is based on the learned latent representations of an NVAE model pretrained on the ACDC dataset. We investigate whether the anatomical information encoded in the latent space can be transferred to a different domain, including an environment where little training data is available. In this section, we apply the ACU-Net framework to the M&Ms dataset.

4.4.1 Domain Adaptation

The experimental pipeline involves taking the ACU-Net model pretrained on the ACDC dataset as configured by Table 4.3 and finetuning it on the M&Ms dataset with (4.5). We perform finetuning with batch size 16 and 50 epochs. The baseline experiment takes the U-Net model pretrained on ACDC and finetunes it on M&Ms with cross-entropy loss.

Centre	# Train Slices	# Validation Slices	% AV (Train & Validation)
1	93	11	100.0
2	81	9	75.6
3	93	11	57.7
4	99	11	74.5
5	116	13	70.5

Table 4.4: M&Ms dataset: Few-shot learning data statistics. Presenting number of data points and % anatomical validity.

4.4.2 Few-Shot Learning

The M&Ms dataset is an aggregate collection of samples from 5 different centres with varying quality and acquisition protocols. A comprehensive breakdown of the dataset is provided in Table A.1b of Appendix A.

We investigate few-shot learning for centre-specific segmentation. We select a small subset from the M&Ms dataset to simulate a small data, domain gap scenario. For each centre, we select 5 subjects from the pre-partitioned train and validation sets that (1) attempt to cover the range of pathologies via stratified sampling, and (2) with the highest % AV. Then, we extract the slice-wise data points and re-partition a smaller train and validation set with a 9:1 split. On average, each subject has 20 slices (10 slices per phase), resulting in approximately 90 data points for training and 10 for validation. Table 4.4 presents the details. Each model is finetuned on 5 subjects from a single centre.

The experimental pipeline involves taking the ACU-Net model pretrained on the ACDC dataset, then performing few-shot finetuning with batch size 16 and 50 epochs. The baseline experiment is taking the U-Net model pretrained on ACDC and few-shot finetuning it with cross-entropy loss.

A concern with the above pipeline is that during finetuning, the model may overfit to the small number of data points. We mitigate this by taking the model checkpoint with the lowest validation loss. However, due to the small validation set, it is difficult to perform early stopping effectively. Therefore, we also explore a parallel set of experiments with the following pipeline: we take the ACDC-pretrained model and immediately use it for centre-specific evaluation. Effectively, we perform zero-shot inference, which guarantees that the features learned from the ACDC dataset do not get depreciated by the small, potentially volatile M&Ms dataset. This is analogous to linear probing[7] when finetuning a pretrained model for classification tasks, which has been shown to give improved results over finetuning the entire model [58, 59].

[7] Freezing a pretrained model and finetuning the linear classifier head only.

Chapter 5

Designing a Robust Metric for Evaluating Quality of Synthetic Cardiac Masks

The Fréchet Inception Distance (Section 2.1.6) is the standard metric for evaluating quality of synthetic images. Previous literature has shown FID to be an effective, reliable metric when applied in the natural image domain. However, we work with cardiac segmentation masks, which differ significantly from natural images. We find FID to be unreliable when used to evaluate synthetic masks. To our knowledge, no work exists that proposes a metric for such domain-specific evaluation: synthetic masks that are not conditioned on subject-level image or scan, and as such, metrics like Dice coefficient (DSC) are inapplicable for measuring generative capability[1].

We propose Fréchet ResNet Distance with SimCLR (FRDS), a novel metric that measures the similarity between synthetic and real cardiac masks. This chapter discusses our work on FRDS, which adapts FID by replacing the Inception-v3 network with a pretrained ResNet model. We use the ACDC dataset for our experiments.

5.1 Methodology

5.1.1 Baseline Metric

FID acts as a baseline for quantitative evaluation of quality of generated cardiac masks. Each data point is a one-hot encoded mask represented as a $4 \times 128 \times 128$ tensor (Section 3.2.1): in particular, it has 4 channels corresponding to the RV, MYO, LV and background. However, FID uses Inception-v3 which is designed for RGB image inputs. We transform the masks into RGB images with the following algorithm.

1. **Remove background:** We remove the background channel from the mask. At this stage, we conveniently have 3 channels: RV corresponds to red, MYO to green and LV to blue.

2. **Rescale values:** The channels are rescaled to have the ImageNet mean and standard deviation[2], as FID uses Inception-v3 pretrained on ImageNet.

[1] Although DSC is useful for measuring reconstruction capability of variational autoencoders.

[2] $\mu = \begin{bmatrix} 0.485 & 0.456 & 0.406 \end{bmatrix}$, $\sigma = \begin{bmatrix} 0.229 & 0.224 & 0.225 \end{bmatrix}$

3. **Resize:** The masks are upsampled to 299×299 pixels with bilinear interpolation as required by Inception-v3.

Computing FID involves comparing the synthetic dataset X' with the real dataset X. We define X to be the test set with 1076 masks. Then, we generate the same number of synthetic masks. We apply the above algorithm to both the real and synthetic masks, then compute FID as described in Section 2.1.6.

5.1.2 SimCLR Pretraining

We hypothesise that the problem which causes FID to be inconsistent for cardiac masks is using a model pretrained on natural images. Certainly, the images in ImageNet differ significantly from segmentation masks in general; the most notable being the discrete colour space and gradient of masks. Therefore, we propose to replace Inception-v3 with a model pretrained on cardiac masks, and keep the rest of the algorithm unchanged.

For pretraining, we use A Simple Framework for Contrastive Learning of Visual Representations (SimCLR) [58, 59], a self-supervised contrastive learning framework for pretraining a convolutional neural network (CNN). The CNN learns representations by comparing similarities between augmented views of the same data point, and contrasting differences between views of different data points. Within self-supervised frameworks for CNNs, SimCLR achieves state-of-the-art performance in various downstream tasks like image classification [59], which testifies its ability to learn robust representations.

Figure 5.1 presents the train workflow; we adopt the same notation from the figure. For downstream tasks, g is discarded and only f is kept. The authors use ResNet-50 [60] as f; we use a smaller ResNet-18 model, as segmentation masks are less complex than natural images. Since ResNet takes RGB images as input, we transform the one-hot encoded masks into RGB

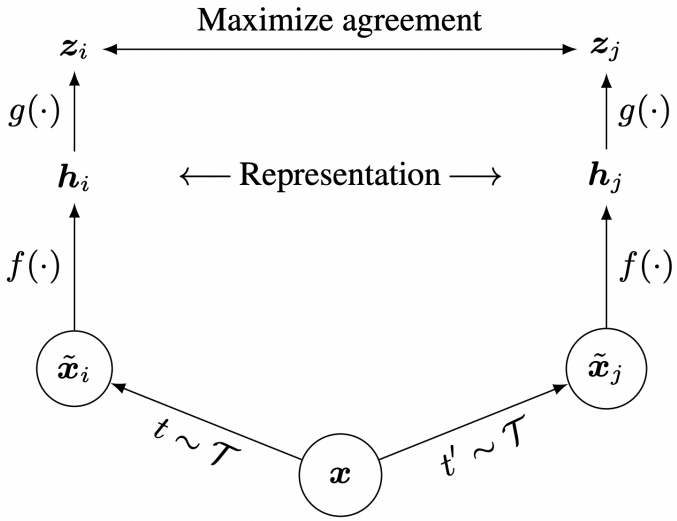


Figure 5.1: SimCLR train workflow. SimCLR is self-supervised so it uses unlabelled data. Instead, within each batch, for each image x it creates a **positive pair**: 2 augmented views $\tilde{x}_i, \tilde{x}_j$. The views are fed into a CNN f without an FC layer to produce embeddings h_i, h_j. During training, the embeddings are projected to a lower dimensional space using a small MLP g, and the entire model tries to match the embeddings of the positive pair z_i, z_j. [58, p. 2]

images with the following algorithm before applying augmentations.

1. **Remove background:** Similar to Section 5.1.1, we remove the background channel from the mask.
2. **Rescale values:** We rescale the values from $[0,1]$ to $[-1,1]$ for train stability.

The choice of data augmentations is crucial for ResNet-18 to learn robust representations. The authors propose a fixed sequence: random crop and resize, random colour distortion, Gaussian blur. In particular, crop and resize forces the model to learn scale invariance (convolutions are already translation equivariant), and colour distortion prevents the model from exploiting the colour space to match positive pairs.

We design a custom augmentation pipeline and describe it as an ordered sequence of affine transformations. The transformations are defined as matrices using homogeneous coordinates.

1. **Random rotation:** Rotate the mask by a randomly chosen angle.

$$R = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \theta \sim \text{Uniform}(-\pi, \pi)$$

   We introduce rotation for the cardiac imaging domain specifically, because structure rotation and alignment are not indicative of heart disease, but rather caused by the cine-MRI procedure. We want to dissuade the model from learning this and instead focus on more indicative features.

2. **Random horizontal flip:** The mask is reflected across the y-axis with 50% probability.

$$H = \begin{cases} \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} & \text{c} < 0.5 \\ I & \text{otherwise} \end{cases} \qquad c \sim \text{Uniform}(0,1)$$

   I is the identity matrix.

3. **Random crop and resize:** Take a random square subset of the image and expand it to the original size (which defines the values of the translation factors t_x, t_y) with nearest neighbour interpolation.

$$C = \begin{bmatrix} s & 0 & t_x \\ 0 & s & t_y \\ 0 & 0 & 1 \end{bmatrix} \quad s \sim \text{Uniform}(0.8, 1)$$

Then, an augmented view is defined as $\tilde{x} = C \cdot H \cdot R \cdot x$.

Figure 5.2 presents the effect of the augmentation sequence. We choose to omit (1) colour distortion, as all masks have the same colours of red, green, blue and black, and (2) Gaussian blur, as we want to penalise generative models that produce smooth, blurry images. Other considerations are discussed within Section 5.2.4.

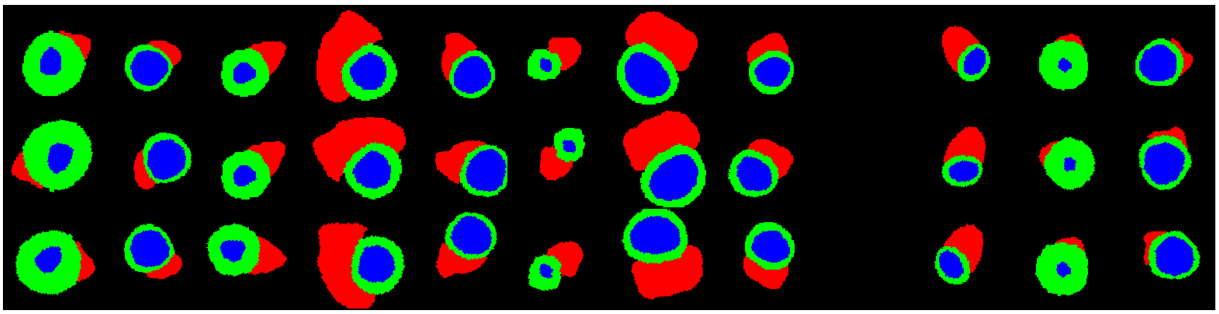


Figure 5.2: Effect of data augmentations for SimCLR pretraining. Each column depicts a mask as an RGB image. Row 1: original. Rows 2, 3: augmented views.

Hyperparameter	**Value**
Epochs	200
Batch size	256
Latent dimension	512
Projected dimension	128
Loss	InfoNCE, $\tau = 0.07$
Optimiser	AdamW
Learning rate	5×10^{-4}
Weight decay	10^{-4}
Scheduler	Cosine annealing, $\eta_{\min} = 10^{-5}$

Table 5.1: Hyperparameter configurations for SimCLR pretraining. The batch size refers to the number of pairs; there are 512 masks in each batch. A large batch size is crucial for learning robust representations.

We adopt the hyperparameters from the SimCLR paper. Table 5.1 presents the full configuration. The authors propose InfoNCE loss, which is based on Noise Contrastive Estimation and uses cosine similarity ((5.1), batch size $2N$).

$$l_{i,j} = -\log \frac{\exp\left(\text{sim}\left(z_i, z_j/\tau\right)\right)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq 1]} \exp\left(\text{sim}\left(z_i, z_k/\tau\right)\right)} \tag{5.1}$$

5.1.3 Computing FRDS

At inference time, we discard the projection head g and use the backbone f, which is a standard ResNet-18 model without the FC layer. f acts as the feature extractor.

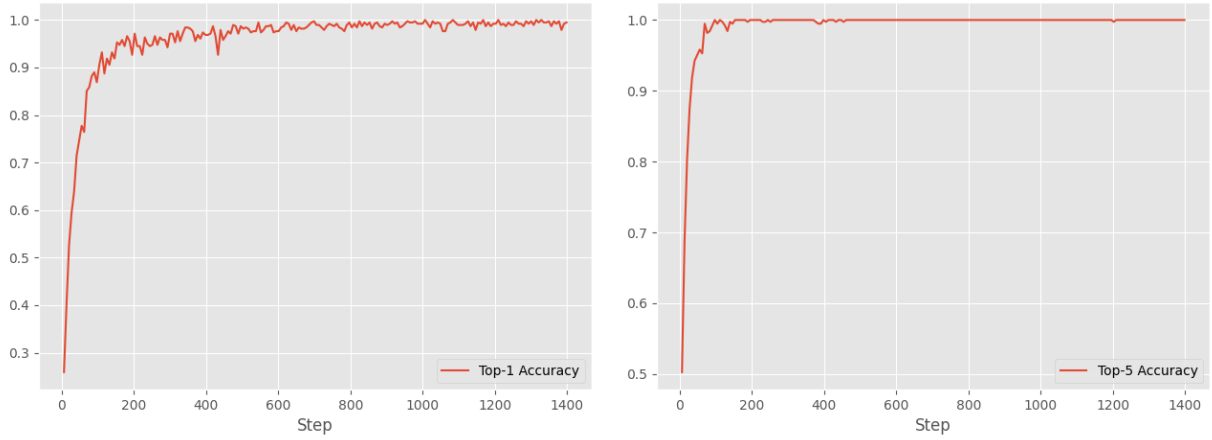
To get the embedding from a one-hot encoded mask, we use the same algorithm as in pretraining (Section 5.1.2) but without applying augmentations: (1) remove background channel from mask, (2) rescale to $[-1, 1]$. Then, we pass the transformed mask through the backbone to get the 512-dimensional embedding.

The rest of the calculation is the same as FID, except that we use 512-dim vectors instead of 2048-dim vectors. See Section 2.1.6 for details.

5.2 Evaluation

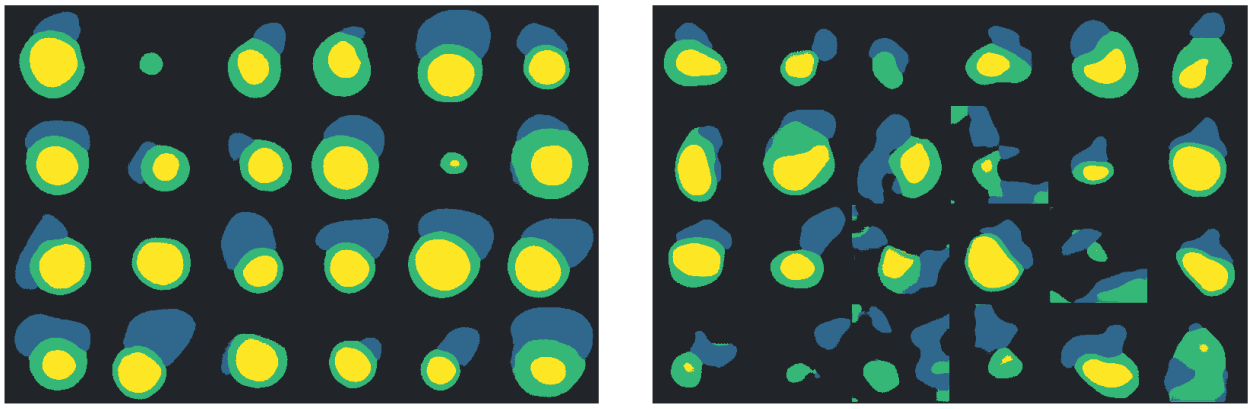
Figure 5.3 presents SimCLR pretraining graphs. In particular, the top-5 accuracy reaches 100% after 97 steps (13 epochs), while the top-1 accuracy continues to increase. Therefore, we use the model checkpoint with the highest top-1 accuracy.

We evalute the performance of FRDS in 3 ways: (1) whether it corrects the inconsistencies of FID in practice, (2) with a custom test suite of various disturbances applied at different levels of intensity to the masks, and (3) by analysing the output embeddings of the pretrained model. Both FRDS and FID are distance metrics, so a lower value indicates better generation quality.



(a) Top-1 accuracy (b) Top-5 accuracy

Figure 5.3: Top-1 and top-5 accuracy graphs for SimCLR pretraining measured with validation set. Metrics are determined by computing the cosine similarity between all data points within each batch, then for each postive pair, whether the similarity is in the top 1 and top 5 respectively.



(a) (b)

Figure 5.4: Example of FID inconsistency. (a) presents generations produced by a well-configured InfoVAE-M model. The model produces anatomically valid masks with a 73.5% rate and achieves 95.08 FRDS and 16.97 FID. (b) presents generations produced by a poorly configured InfoVAE-D model with 16.1% anatomical validity. It has a poor 428.7 FRDS, but achieves a strong FID of 15.50.

5.2.1 Practical Results

Figure 5.4 presents 2 sets of synthetic masks generated by different VAE models. It illustrates the practical inconsistencies of FID and how FRDS gives an evaluation that is more consistent with empirical judgment.

To further illustrate this, we plot anatomical validity against FRDS and FID for a batch of baseline models. These models are obtained during hyperparameter tuning, and thus range from well-configured to poorly configured. Assuming FRDS is a reliable metric, we expect a negative correlation between FRDS and anatomical validity. However, a model can also gener-

(a) Anatomical validity against FRDS. Models with FRDS > 300 are filtered.

(b) Anatomical validity against FID. Models with FID > 25 are filtered.

Figure 5.5: Scatter plots of anatomical validity against FRDS and FID for a batch of baseline models with varying quality.

ate with high validity rate and have high FRDS if (1) the model experiences mode collapse and produces low diversity generations, or (2) the model produces anatomically valid but blurry or impractical shapes. Therefore, we expect the plot to form a shape resembling an upper triangular matrix. Figure 5.5 presents the plots. Indeed, we do observe this phenomenon for FRDS. Meanwhile, the FID plot does not indicate any meaningful correlation with anatomical validity.

However, FRDS is not an ideal metric. We observe that FRDS is poor at penalising synthetic masks that have eroded or chipped edges (Figure 5.6). We hypothesise this may be caused by using crop and resize augmentation in pretraining, which may introduce chipped edges in the augmented views. In Section 5.2.4, we discuss other augmentation pipelines, such as replacing crop and resize with zoom-out. However, these alternative designs do not outperform the current FRDS metric in practical results, test suite results and embedding analysis.

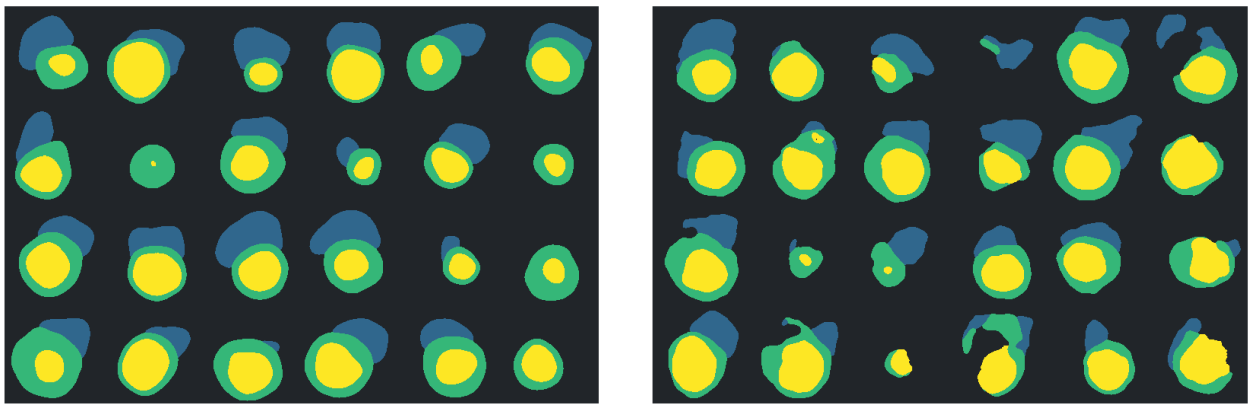
5.2.2 Disturbance Test Suite

To measure the robustness of FRDS, we design a test suite that applies disturbances to the test set at various intensity levels. Then, we measure the FRDS between the undisturbed train set and the disturbed test set[3]. The desired outcome is to observe a positive correlation between FRDS and intensity level.

We apply 4 classes of disturbances: average smoothing, black box crop, elastic deformation, pepper noise. Each class has 4 intensity levels. See Appendix C for details.

Evaluation results are presented in Table 5.2. We observe for both FRDS and FID that (1) the values increases with intensity level, and (2) the values never decrease below the gold standard of 12.9 for FRDS and 4.51 for FID. The only exception is FRDS with elastic deformation at intensity level 1. Overall, this suggests that both metrics are capable of distinguishing between high and low quality masks, including small deteriorations in quality.

[3]Since the train set has 1902 data points and the test set has 1076, we downsample the train set by choosing a fixed random subset of 1076 train data points to compute FRDS.



(a) (b)

Figure 5.6: Example of FRDS inconsistency. (a) presents generations produced by a well-configured NVAE Default-N model. The model produces anatomically valid masks with a 95.9% rate and achieves 43.10 FRDS. (b) presents generations produced by a poorly configured NVAE LatentSkip-N model with 50.6% anatomical validity. However, it achieves a strong 41.39 FRDS.

	FRDS				FID			
Disturbance	**Intensity Level**				**Intensity Level**			
	1	**2**	**3**	**4**	**1**	**2**	**3**	**4**
None	12.9				4.51			
Average smoothing	14.5	15.0	15.5	15.9	7.43	8.21	8.55	8.67
Black box crop	16.1	28.1	73.4	186.0	6.19	21.6	62.5	123.6
Elastic deformation	9.20	23.4	282.1	3881	8.62	14.4	49.0	148.9
Pepper noise	13.3	22.4	267.7	1496	4.78	16.9	31.1	33.0

Table 5.2: FRDS and FID values between train set and disturbed test set for 4 classes of disturbances at increasing intensity levels.

5.2.3 Embedding Analysis

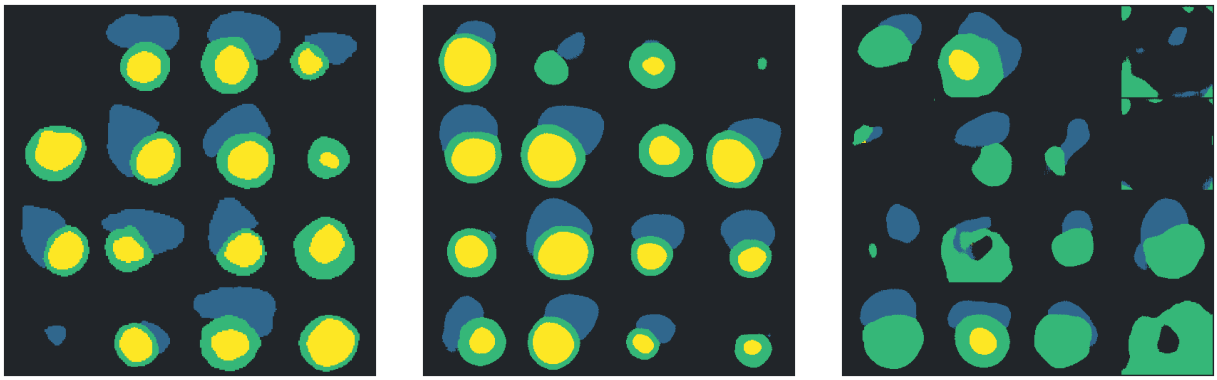
In Figure 5.3, the top-1 and top-5 accuracies are very high which suggest potential overfitting. We investigate the embeddings produced by the model to ensure that there is no presence of overfitting and the model learns meaningful representations. As a baseline, we also present the corresponding embeddings of the pretrained Inception-v3 for FID.

For a set of masks, we compute the 512-dimensional embeddings[4] as described in Section 5.1.3. Then, we apply principal component analysis (PCA) [61] with 2 components, fitted with the train embeddings. This achieves an explained variance of 0.42 for the embeddings used to compute FRDS and 0.68 for the embeddings used to compute FID.

Since FRDS uses the Fréchet distance, the ideal scenario is that embeddings of the train, test and realistic synthetic masks are **not** well separated and interleave with each other, while embeddings of the poorly generated masks are separated from the rest. This would mean that the Fréchet distance is low between the test and realistic synthetic masks, and high between the test and poorly generated masks.

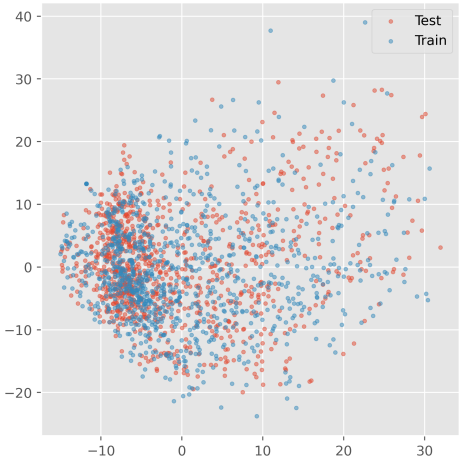
To acquire realistic synthetic masks, we select a baseline VAE model that empirically gener-

[4] 2048-dim embeddings for FID.

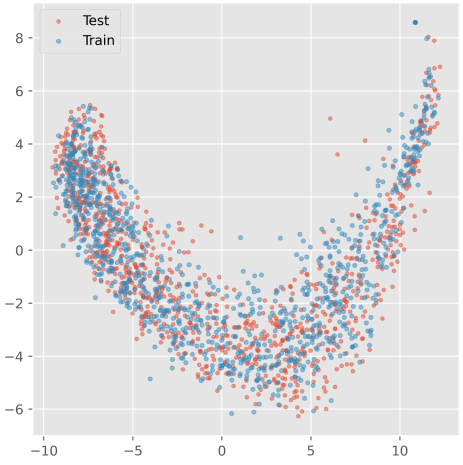


(a) Test set (b) Good generations (c) Bad generations

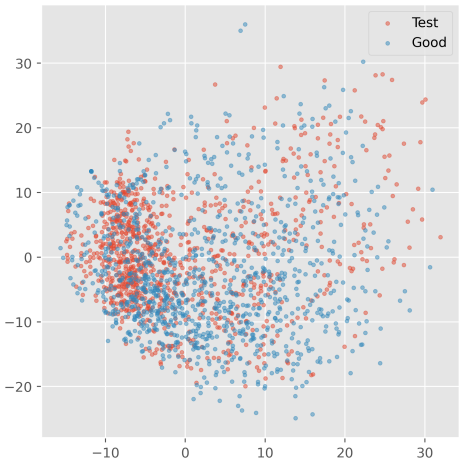
Figure 5.7: Preview of high quality generations from a well-configured VAE model and low quality generations from a poorly configured VAE model, alongside masks from the test set.



(a) FRDS: Test and train sets

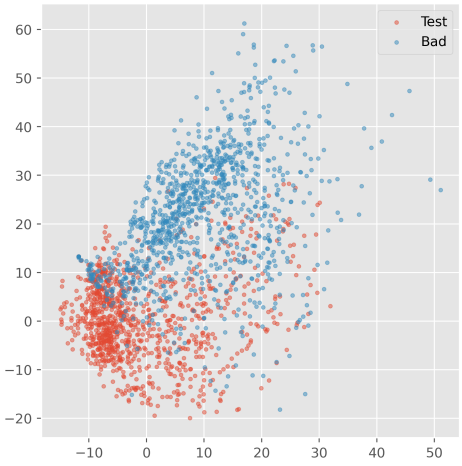


(b) FID: Test and train sets

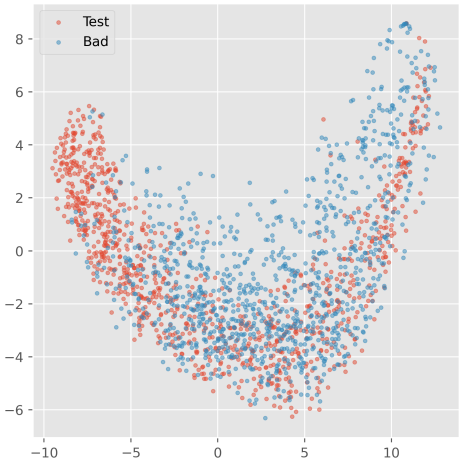


(c) FRDS: Test set and good generations

(d) FID: Test set and good generations



(e) FRDS: Test set and bad generations



(f) FID: Test set and bad generations

Figure 5.8: PCA visualisations of FRDS and FID embeddings compared between various data partitions. x-axis: component 1, y-axis: component 2. Each plot has 1076 data points from each set. Good generations refer to masks generated by a well-configured VAE model, while bad generations refer to masks generated by a poorly configured model. Note that the poorly configured model occasionally generates realistic masks; vice versa.

ates high quality masks. We also select a poorly configured VAE model to acquire low quality masks. Examples of these masks are shown alongside real data in Figure 5.7. Note that the poorly configured model occasionally generates realistic masks.

Figure 5.8 presents the PCA visualisations for the FRDS and FID embeddings on various data partitions. We make several observations. Firstly, the FRDS embeddings of the train and test sets interleave well without forming 2 distinct clusters. This indicates that the model has not overfitted to the train set. Secondly, the FRDS embeddings of the test set and good generations also interleave, but there is a distinct separation from the bad generations. This is also seen in the FID embeddings with less clarity, suggesting the model used for FRDS is more effective in distinguishing between realistic and unrealistic data. Finally, the 2 principle components for FRDS embeddings have a larger variance than that of FID embeddings. For the former, the components of realistic data lie in the range $[-20, 40]$ while for the latter, the range is $[-10, 15]$. Therefore, we expect FRDS values to be relatively larger than FID values, and the absolute values of the two metrics are not directly comparable.

5.2.4 Other Considerations

We hypothesise that the poor ability of FRDS to recognise eroded edges in masks is amplified by the crop and resize augmentation in pretraining. However, we did not find improvement in replacing crop and resize with zoom-out augmentation. We also briefly investigated elastic deformation as a possible augmentation technique, but dropped it as it might affect the learned contours, which are indicative of heart disease.

For the embedding analysis (Section 5.2.3), we attempted to use t-SNE [62] instead of PCA. t-SNE is a non-linear dimensionality reduction technique that has the potential to preserve more information with the same number of components. However, we found the results to be less interpretable than PCA, despite testing various perplexity and iteration values. This could be due to t-SNE's tendency to group data points into clusters, but in some settings we expect the embeddings to interleave.

5.3 Concluding Remarks

We design a novel metric for evaluating quality of generated cardiac masks, which we term Fréchet ResNet Distance with SimCLR (FRDS). We evaluate FRDS in a variety of settings and find it to give more consistent, robust results than FID. We observe that FRDS struggles with recognising eroded masks as poorly generated, and suggest using it in conjunction with another metric that makes up for this weakness, such as the percentage of anatomically valid masks.

Chapter 6

Evaluation

6.1 Shape Encoding

6.1.1 Results

We perform quantitative evaluation as follows. Each model outputs 1076 reconstructed masks (corresponding to ACDC test set) and randomly generates 1076 synthetic masks with the assumed prior. Quality of synthetic masks are measured by FRDS, with X and X' being the test set and synthetic set respectively. % AV (anatomical validity) measures the set of synthetic masks. In particular, while training has been conducted on a per-slice basis, we measure reconstruction quality with the 3D Dice coefficient (DSC) across the 3D volume (see Appendix D.1). We compute DSC by comparing the reconstructed volume segmentation (stack of masks) to the ground truth.

Table 6.1 presents the results of NVAE compared to baseline models. Each row corresponds to the configuration for a framework that gives best overall performance. Since there is a trade-off between reconstruction and generation quality, we present 3 strong NVAE configurations. Our downstream task (Section 4.3) favours high quality, anatomically valid generations, so the best model is bolded: Default-N without clamping nor spectral regularisation. This model significantly improves upon baseline metrics: 0.078 DSC increase for reconstructions, as well as

Baseline Model			**DSC**				**FRDS**	**% AV**
			All	**RV**	**MYO**	**LV**		
β-VAE			0.882	0.893	0.847	0.907	87.6	74.5
InfoVAE-D			0.891	0.901	0.858	0.916	89.7	72.5
InfoVAE-M			0.871	0.875	0.832	0.906	61.2	64.3
NVAE Model	**Clamp**	**SR**						
Default-N	**No**	**No**	**0.969**	**0.976**	**0.953**	**0.978**	**33.2**	**96.5**
Default-N	Yes	Yes	0.989	0.990	0.984	0.993	37.4	85.9
LatentSkip-N	No	No	0.999	0.999	0.998	0.999	83.1	72.0

Table 6.1: Quantitative metrics for NVAE and baseline models on the ACDC dataset.

Baseline Model			# Parameters (M)	Latent Dimension	β	γ
β-VAE			4.6	8	100	
InfoVAE-D			4.6	8	0	100
InfoVAE-M			4.6	8	0	200
NVAE Model	**Clamp**	**SR**		β_1	β_2	β_3
Default-N	**No**	**No**	**2.0**	**10**	**9**	**8**
Default-N	Yes	Yes	2.1	1	1	1
LatentSkip-N	No	No	1.6	1	1	1

Table 6.2: Tuned parameters for the models presented in Table 6.1. For the rest of the configurations, see Table 4.2 (NVAE) and Table 4.1 (baseline).

28.0 FRDS increase and 22.0% anatomical validity increase for generations. The hyperparameter configurations are presented in Table 6.2.

For reconstruction at component level, we observe a trend present in both NVAE and baseline models: LV has the most accurate reconstructions, followed by RV then MYO. This trend exists even for very strong reconstruction models (>0.98 DSC across every component). It is also observed in existing segmentation models[14]. Since NVAE and VAE models are trained with the masks only, we conclude that the difficulty of the components are partly due to the shape of the components themselves, and not caused solely by the scan intensity. We speculate this is due to the torus topology of the MYO, and the high variability of the RV shape.

We further investigate the performance of the best NVAE model by analysing its reconstruction quality per phase and per pathology (Figure 6.1 and Figure 6.2). The model has a slightly higher performance for RV and LV at the ED phase, and for MYO at the ES phase. This is also observed in existing segmentation models[14], suggesting this phenomenon is also caused by

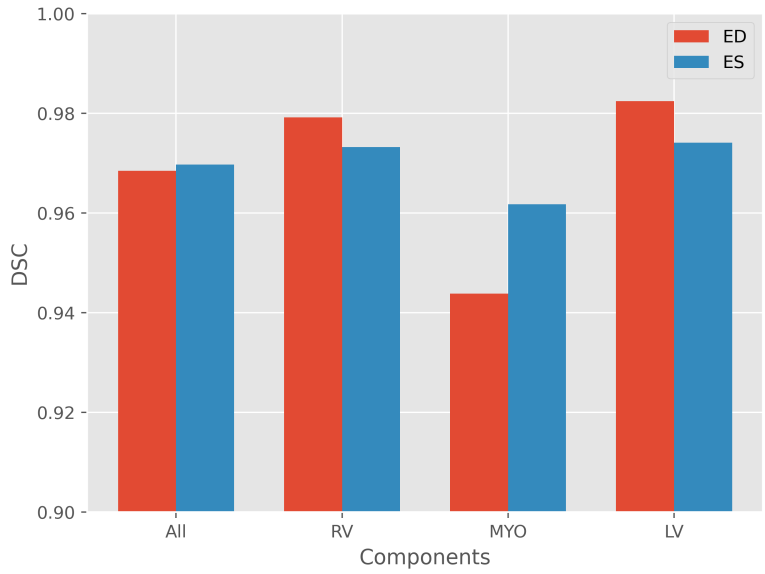


Figure 6.1: Comparison of reconstruction quality of masks at ED and ES phases of top NVAE model.

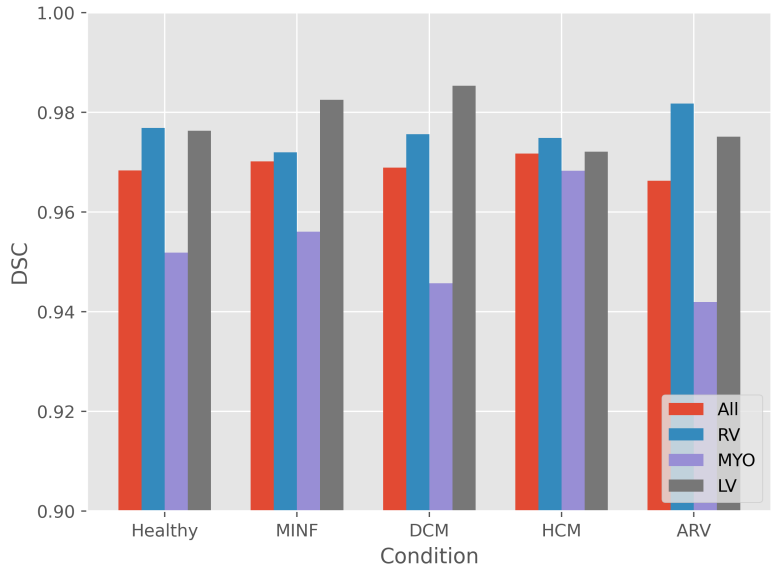


Figure 6.2: Comparison of reconstruction quality of top NVAE model for the different pathologies in the dataset. Abbreviations stand for previous myocardial infarction (MINF), dilated cardiomyopathy (DCM), hypertrophic cardiomyopathy (HCM), abnormal right ventricle (ARV).

the shape of the components. We observe that the model is robust to the various pathologies in the dataset. Interestingly, it has the highest RV performance for patients with abnormal right ventricle, suggesting that some pathologies may be easier to segment. However, the differences are very small ($\pm$ 0.01 DSC).

Unless otherwise specified, subsequent sections will focus on evaluating the top NVAE model.

6.1.2 Reconstruction and Generation Visualisations

Figure 6.3 presents a visual comparison of the original and reconstructed masks produced by NVAE, compared to the β-VAE baseline. The improvement in reconstruction quality is strikingly evident: NVAE is able to reconstruct masks with a very thin margin of error. Figure 6.4 presents a random batch of synthetic masks generated by NVAE and β-VAE. β-VAE (and other baseline models) often produce pixel-wise anatomical inconsistencies, which can be empirically observed when zoomed in. NVAE is able to produce anatomically valid generations without such inconsistencies. Furthermore, NVAE produces more diverse and detailed masks, as can be seen in the RV shapes. Appendix E.2 provides more visualisations.

6.1.3 Temperature

Temperature τ is a hyperparameter at inference time that controls the diversity of generated samples. In particular, it scales the standard deviation of the prior. The generations presented in this chapter are produced with $\tau = 1$. As the models are trained with a standard Gaussian prior, this means the prior is unchanged at inference.

In practice, $\tau < 1$ is often used. This forces the model to sample from higher probability regions of the latent space, which can lead to more stable generations at the cost of some di-

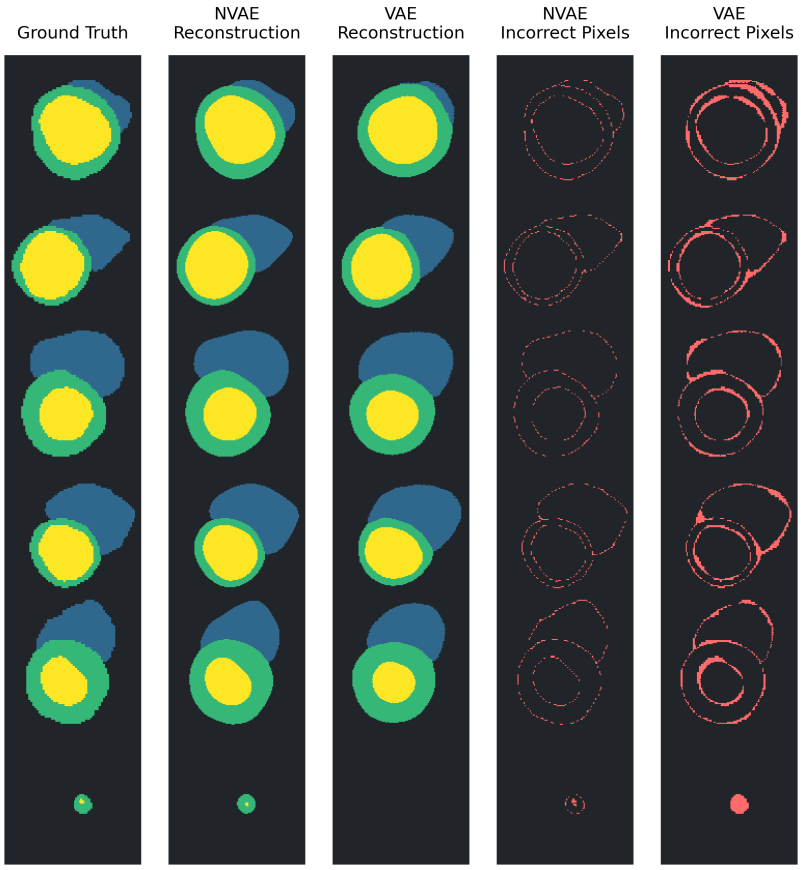
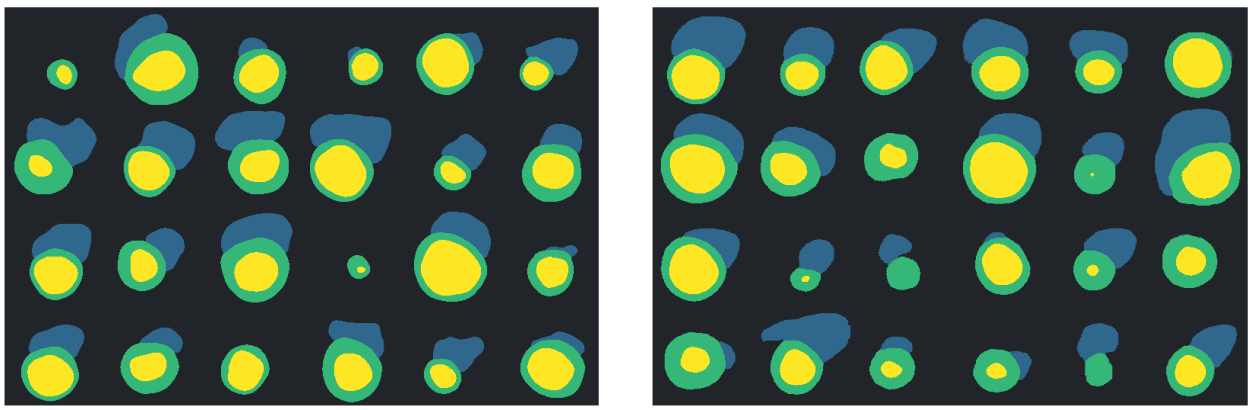


Figure 6.3: Visualisations of reconstructed masks of top NVAE model and β-VAE. Red indicates incorrectly reconstructed pixels.



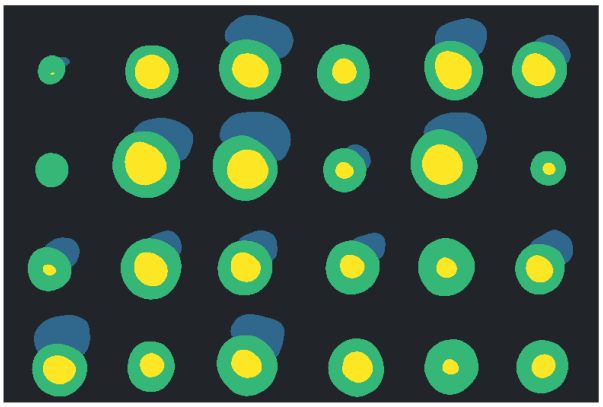
(a) NVAE generations (b) β-VAE generations

Figure 6.4: Visualisations of generated masks of top NVAE model and β-VAE.

versity loss[63]. The authors of NVAE use $\tau \in \{0.5, 0.7\}$ for complex datasets[13]. However, we find that (1) $\tau = 1$ produces high quality, anatomically valid generations, and (2) $\tau < 1$ improves % AV but significantly detriments generation diversity. As such, we stick with $\tau = 1$ at inference. Figure 6.5 presents the results.

τ	FRDS	% AV
1.0	33.2	95.8
0.9	56.3	97.6
0.8	90.6	97.9
0.7	137.6	99.0
0.6	201.4	99.3
0.5	294.1	99.9

(a) Effect of τ on FRDS and % anatomical validity



(b) Preview of generations with $\tau = 0.5$

Figure 6.5: Effect of temperature τ on generation quality of top NVAE model.

6.1.4 Learned Latent Space

Figure 6.6 presents the marginal KL divergence per group. If the KL divergence is 0, then the distributions $q_\phi(z_i|x, z_{<i})$ and $p_\theta(z_i|z_{<i})$ are identical. This indicates posterior collapse, as the posterior does not use the current sample x to encode z_i. The figure shows that none of the latent groups have collapsed, as all marginal KLs remain significantly above 0. This justifies our use of all groups in formulating the shape loss in the downstream segmentation task (Section 4.3).

At inference time, reconstructions are computed with deterministic sampling from the residual distributions, which means the reconstruction of an existing mask is defined unam-

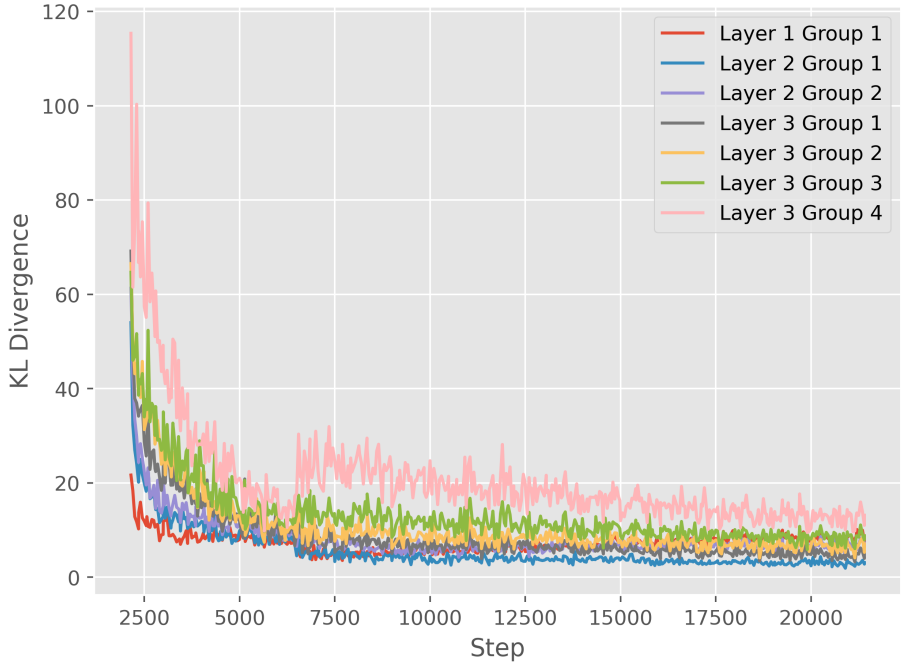


Figure 6.6: Marginal KL divergence per group of top NVAE model during training. The first 10 epochs (2140 steps) are omitted for clarity.

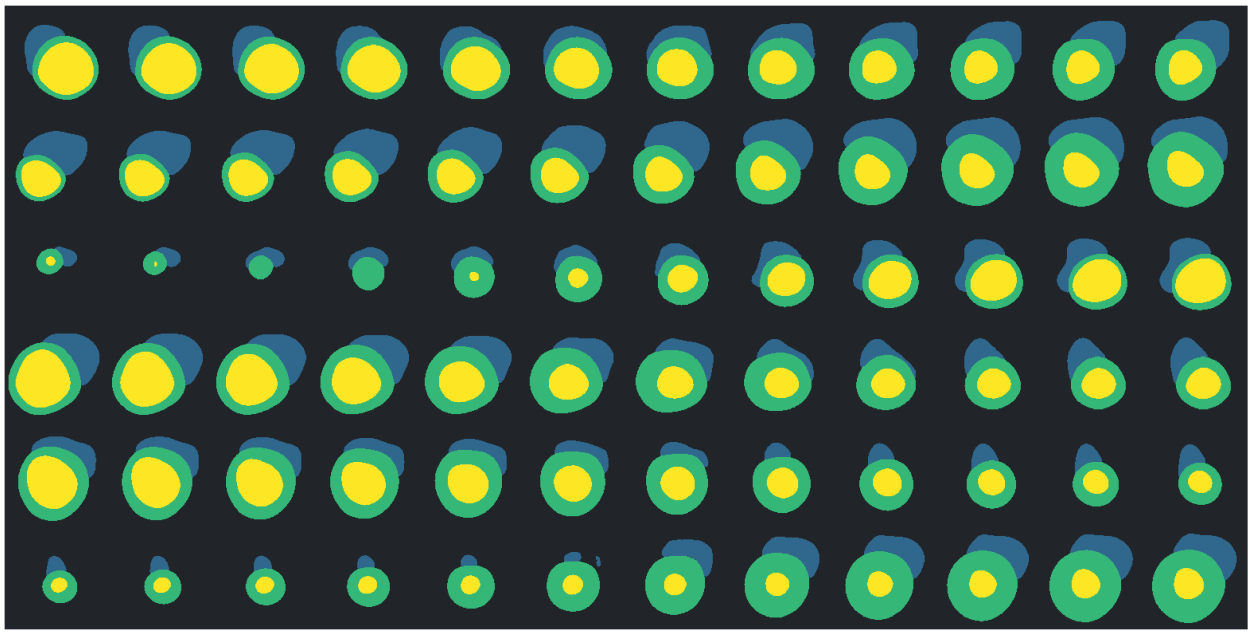


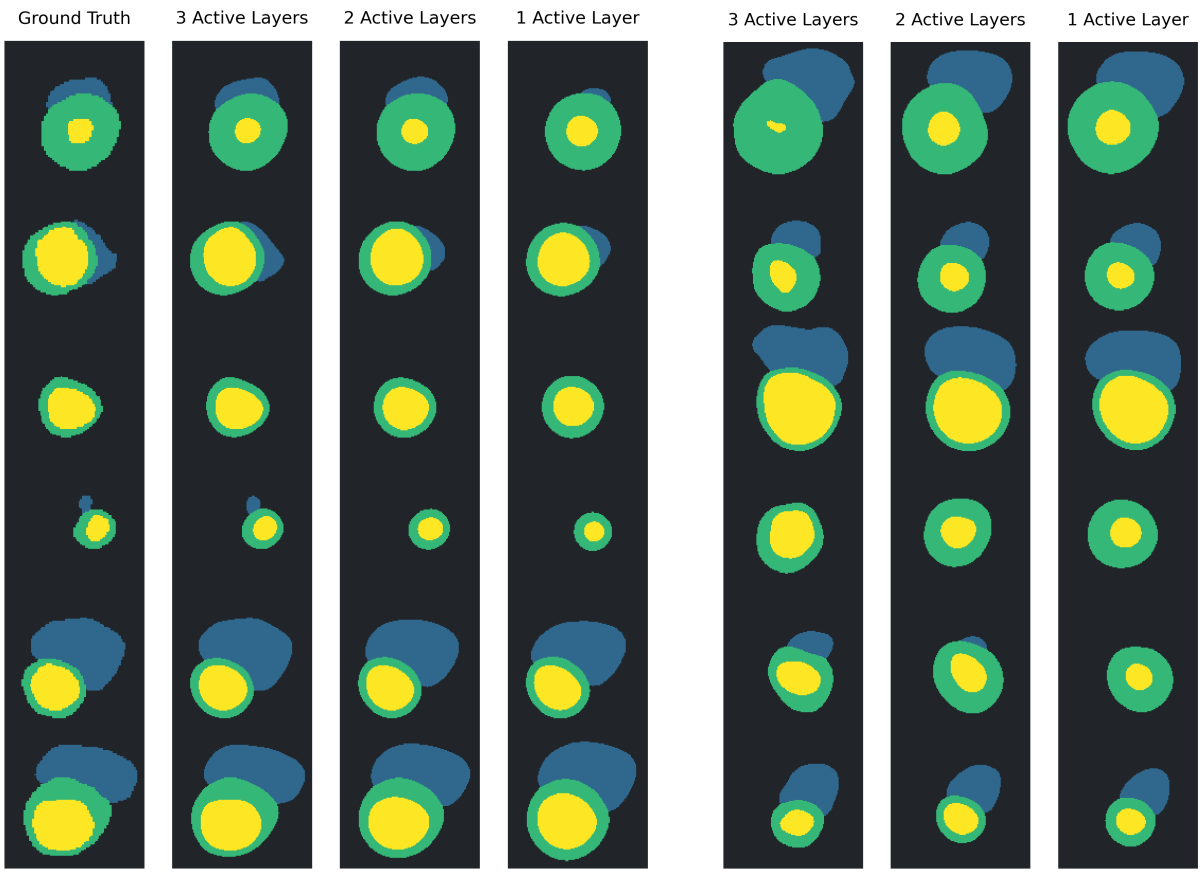
Figure 6.7: Latent traversal of top NVAE model. In each row, the first and last columns are reconstructions of existing masks, and the columns in between are reconstructions of linearly interpolated vectors.

biguously by its latent representation. Therefore, we can perform latent traversal by linearly interpolating between the latent vectors of 2 masks, then obtaining the reconstructions by traversing the decoder path. Figure 6.7 presents the results. Empirically, the shape changes smoothly and gradually as it interpolates between a pair of masks. In some examples, there are more abrupt changes in the RV (middle of last row).

We perform an ablation study to investigate the contribution of each latent layer by turning them off during reconstruction and generation. By default, all 3 layers are active. 2 active layers refers to the topmost and middle layers, and 1 active layer refers to the topmost layer only. If a layer is inactive during reconstruction, the decoder does not draw information from the approximate posterior. That is, z is sampled from $\mathcal{N}(\mu_p, \sigma_p)$. If a layer is inactive during generation, z is set to μ_p instead of being sampled from $\mathcal{N}(\mu_p, \sigma_p)$. Table 6.3 presents the results. The topmost layer encodes smooth, global features, while the other two layers build on this by encoding detailed, fine-grained features to form more refined, intricate shapes. These refined features result in better reconstructions for all cardiac components and generation diversity (FRDS), but also cause slightly lower anatomical validity in generations. This can be observed

# Active Layers	**DSC**				**FRDS**	**% AV**
	All	**RV**	**MYO**	**LV**		
3	0.969	0.976	0.953	0.978	33.2	96.5
2	0.935	0.938	0.915	0.952	41.0	98.6
1	0.873	0.865	0.848	0.907	58.6	99.6

Table 6.3: Effect of turning off latent layers on reconstruction and generation performance of top NVAE model.



(a) Reconstructions (b) Generations

Figure 6.8: Visualising the effect of turning off latent layers on reconstructed masks and generated masks, using top NVAE model. (a) The same original sample is used within each row. (b) The same latent vector is used within each row.

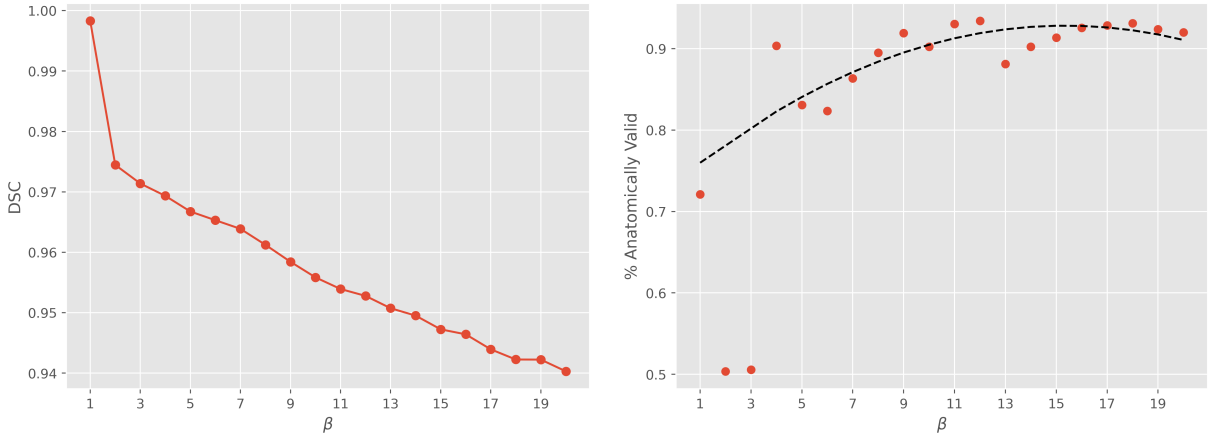
in row 4 of Figure 6.8a, where the model reconstructs the RV with 3 active layers, but fails with 1 or 2 active layers. In row 3 of Figure 6.8b, the model refines the RV shape with 3 active layers. Overall, the model performs best from having all 3 latent layers active. Furthermore, with 1 active layer, the model is still capable of producing outputs with quality comparable to (and exceeding) the single-layer baseline models.

6.1.5 Increasing KL Weight Term

Figure 6.9 presents the effect of increasing β on NVAE performance. We present findings for the LatentSkip-N architecture, but similar trends are observed for the Default-N architecture. Figure 6.10 and Figure 6.11 presents sample visualisations. We observe a larger β leads to smoother outputs. This is expected as β controls regularisation strength. As a result, the generated masks are more stable and anatomically valid, at the cost of some loss in detail. It also causes deterioration in reconstruction quality. This results in a trade-off between reconstruction and generation quality, and the best β is dependent on the downstream application.

This phenomenon is also observed in baseline models[1].

[1] β for β-VAE and γ for InfoVAE.



(a) DSC against β (b) % AV against β

Figure 6.9: Effect of increasing β on reconstruction quality (DSC) and % anatomical validity of generations, for the LatentSkip-N architecture.

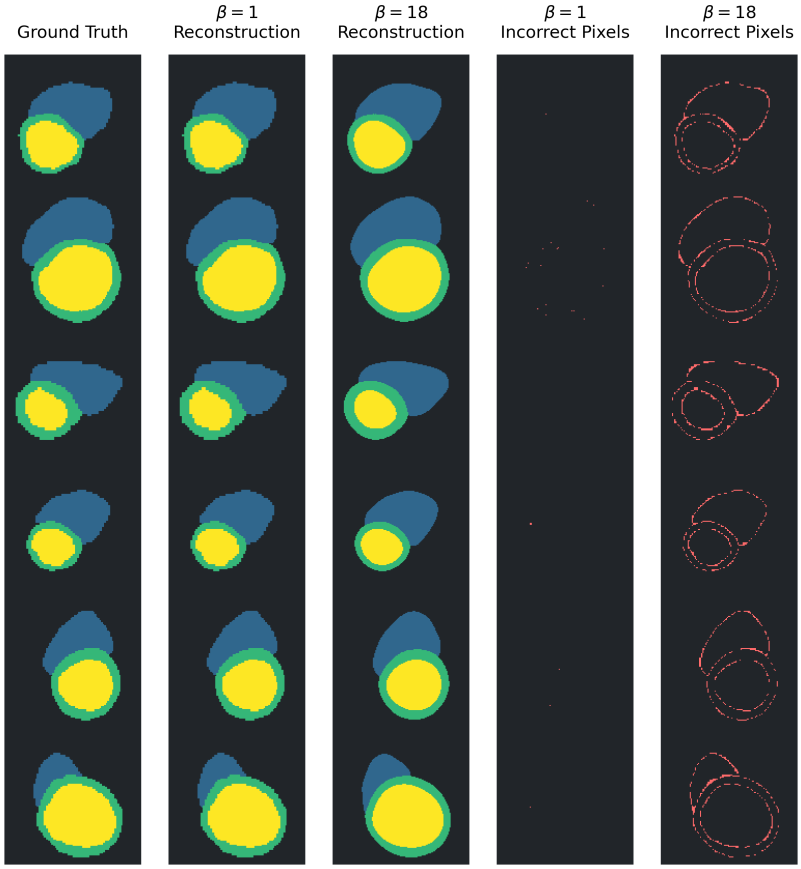
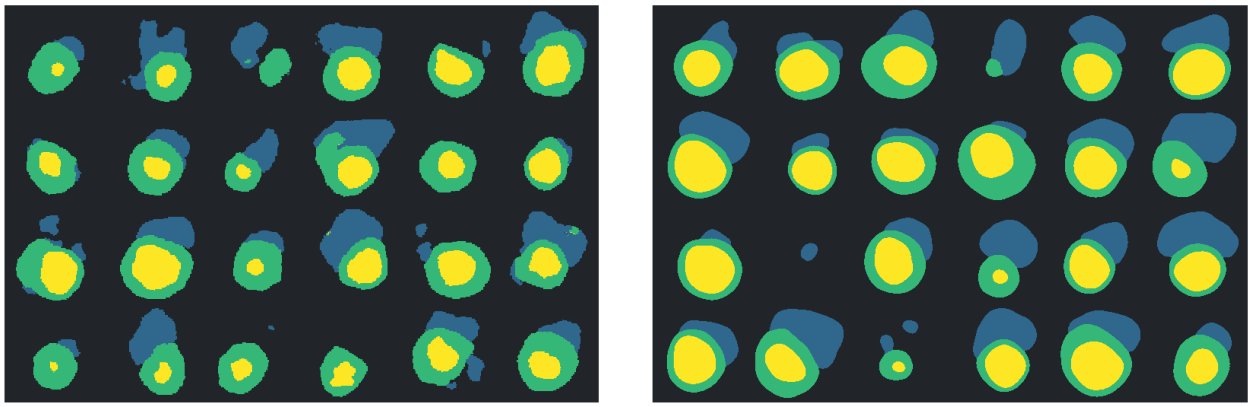


Figure 6.10: Visualisations of reconstructed masks for the LatentSkip-N architecture, with $\beta = 1$ and $\beta = 18$. Red indicates incorrectly reconstructed pixels. Best seen when zoomed in.



(a) $\beta = 1$ (b) $\beta = 18$

Figure 6.11: Visualisations of generated masks for the LatentSkip-N architecture, with $\beta = 1$ and $\beta = 18$. The difference in smoothness is best seen when zoomed in.

6.2 Cardiac Segmentation with Shape Loss

6.2.1 Results

We perform quantitative evaluation with the ACDC test set of 1076 (slice, mask) pairs. % AV measures the set of output segmentations. Similarly, while training has been conducted on a per-slice basis, we measure segmentation quality with the 3D Dice coefficient (DSC) across the 3D volume.

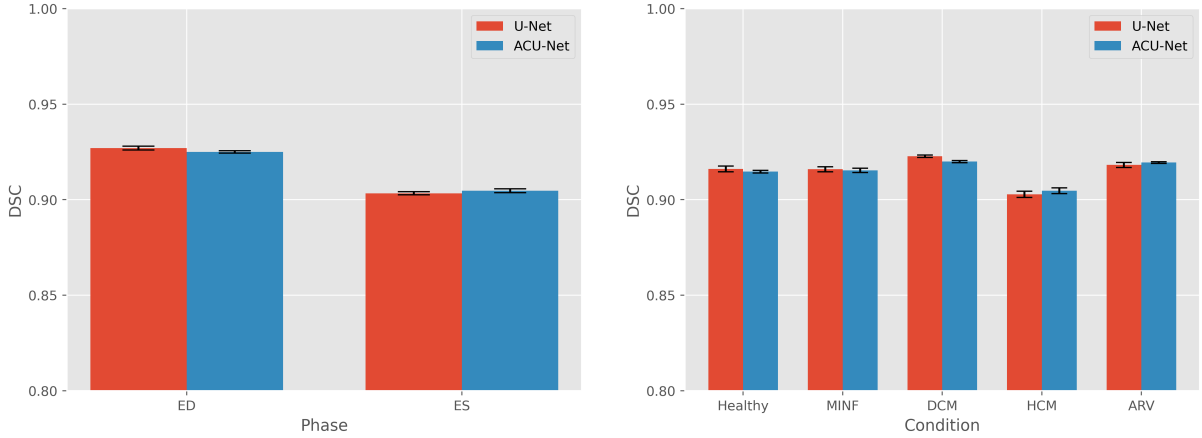
Table 6.4 presents the results of ACU-Net compared to the U-Net baseline. In addition, we include the results of the best 2D U-Net framework from MICCAI 2017[2] by Baumgartner et al. [64], which also uses cross-entropy loss. We find that (1) our baseline results align with Baumgartner et al. (within ±0.006 DSC in all classes), and (2) ACU-Net matches the performance of U-Net in DSC, while achieving a 5.3% increase in anatomical validity ($p < 0.014$[3]). This increase could be attributed to using the NVAE latent vectors as a shape regularisation term, which enforces a more global shape consistency compared to the pixel-wise cross-entropy loss.

We further investigate by analysing the segmentation quality per phase and per pathology. Figure 6.12a and Figure 6.12b) presents the DSC plots. ACU-Net is able to match U-Net per-

[2] The challenge that introduced the ACDC dataset.
[3] p-value as computed by Welch's t-test; see Appendix D.2.

Model	**DSC**				**% AV**
	All	**RV**	**MYO**	**LV**	
Baumgartner et al.		0.908	0.897	0.937	
U-Net	0.915 ± 0.001	0.909 ± 0.002	0.894 ± 0.001	0.943 ± 0.001	84.3 ± 1.37
ACU-Net	**0.915 ± 0.001**	**0.907 ± 0.001**	**0.893 ± 0.001**	**0.944 ± 0.001**	**89.6 ± 0.65**

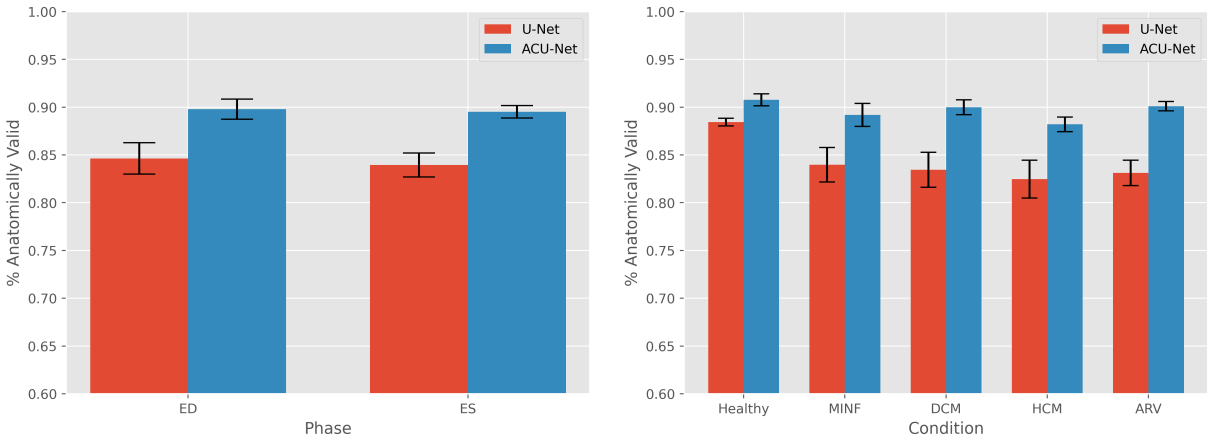
Table 6.4: Quantitative metrics for cardiac segmentation models on the ACDC dataset. ACU-Net is trained with $\alpha = 1$. For U-Net and ACU-Net, 5 models are trained with different seeds (but same configuration and train-validation split), and the mean metrics with standard error are reported.



(a) Quality at ED and ES phases.

(b) Quality for the different pathologies.

Figure 6.12: Comparison of segmentation quality of ACU-Net and U-Net for different phases and pathologies. Presenting mean DSC with standard error bars.



(a) % AV at ED and ES phases.

(b) % AV for the different pathologies.

Figure 6.13: Comparison of anatomical validity of output segmentations of ACU-Net and U-Net models for different phases and pathologies. Presenting mean % AV with standard error bars.

formance across both ED and ES phases, and across all pathologies. Both ACU-Net and U-Net have relatively lower performance for HCM condition; however, NVAE does not have a specific bias towards any pathology (Figure 6.2). This suggests that the scan intensity may contribute to the difficulty of segmenting patients with HCM, although the differences are subtle (±0.03 DSC). Figure 6.13a and Figure 6.13b presents the % AV plots. ACU-Net is able to produce more anatomically valid segmentations across all phases and pathologies compared to U-Net, especially for patients with a diagnosed condition.

6.2.2 Segmentation Visualisations

Figure 6.14 presents a visual comparison of the output segmentations of ACU-Net and the ground truth. In general, the model is able to produce anatomically valid masks that accurately

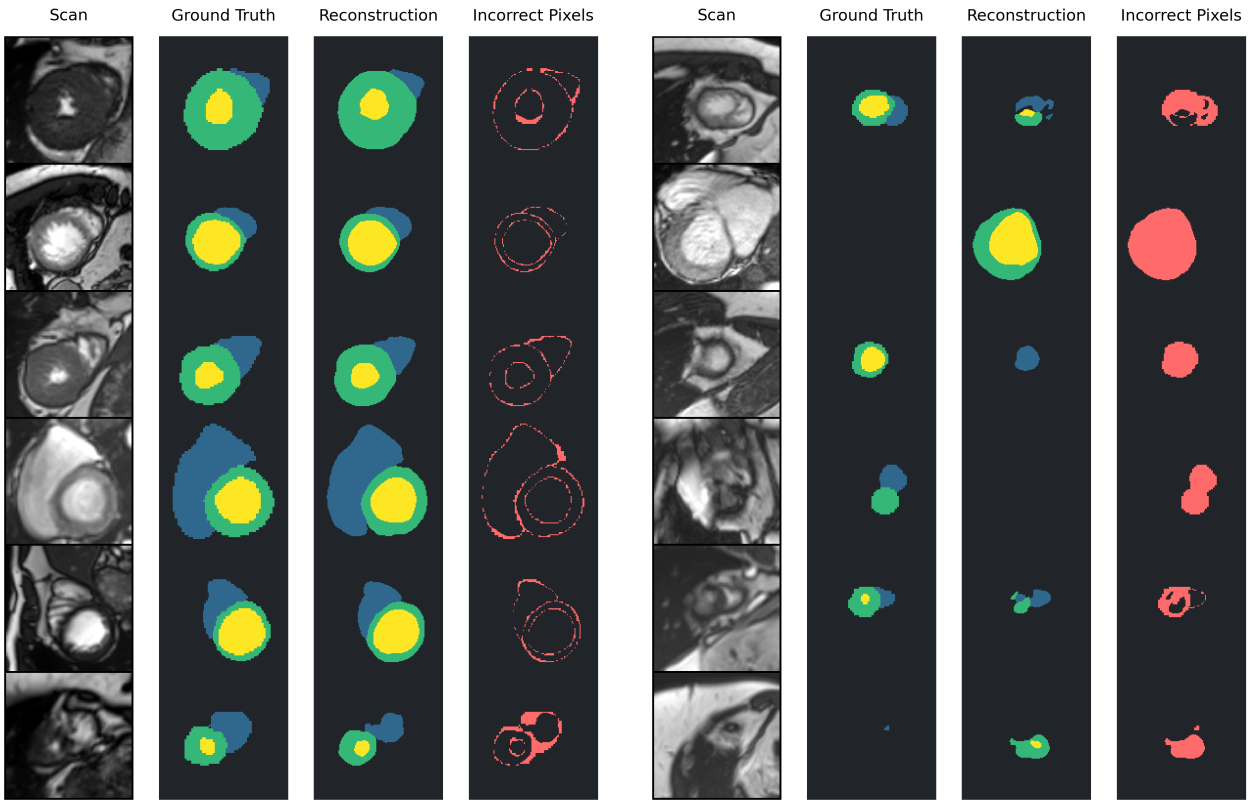


Figure 6.14: Visualisations of output segmentations of ACU-Net on the ACDC dataset. Presenting (a) random samples, and (b) some of the worst samples.

segment the RV, MYO and LV. We also showcase some of the worst output segmentations[4]. A key weakness of ACU-Net, which is also observed in U-Net, is that it struggles to determine whether the slice is a basal or apical slice (or otherwise ambiguous), for which the experts leave the annotated mask empty. However, the model may attempt to segment said slice. Conversely, the model may output an empty mask for a slice that is meant to be segmented. Only 55 out of 1711 train data points and 75 out of 1076 test data points have empty masks, so this is not a critical issue when using the ACDC dataset. The model is also more likely to struggle with segmenting the components when they are smaller.

6.2.3 Shape Loss Weight

We find $\alpha = 1$ to perform best for ACU-Net. This suggests that the cross-entropy loss and the shape loss have equal importance.

ACU-Net performs well when both cross-entropy and shape regularisation terms are weighted. We find that when the cross-entropy term is removed ($\alpha = 0$), the model is still able to match U-Net performance in DSC (within ± 0.002), but the output segmentations are only 51.5% anatomically valid. It is difficult to determine the exact cause, but we speculate that the model requires both local (cross-entropy) and global (shape loss) information to train optimally.

[4] < 0.4 DSC when measured per slice.

6.2.4 Domain Adaptation and Few-Shot Learning

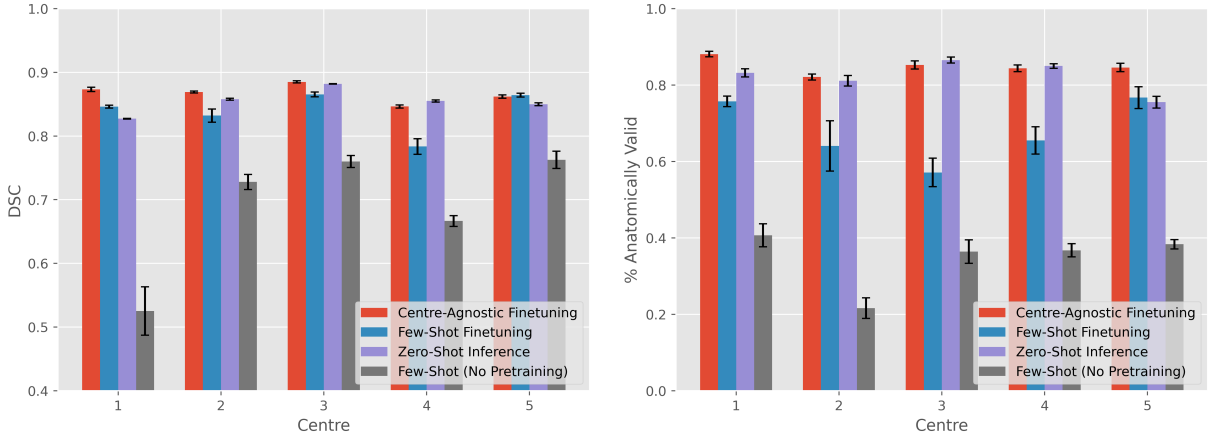
We perform quantitative evaluation for U-Net and ACU-Net trained on the full domain-agnostic M&Ms dataset. Table 6.5 presents the results. Overall, ACU-Net is able to match U-Net performance in DSC, with a slight increase in % AV ($p < 0.05$). In particular, ACU-Net has a 4.9% increase in anatomical validity for Centre 1 ($p < 0.014$). This suggests that the shape loss, which uses the latent vectors pretrained on ACDC, is most effective for Centre 1.

By investigating the pathology distribution of the ACDC and M&Ms datasets (Table A.1), we find there are 4 overlapping conditions: Healthy, HCM, DCM, ARV. There is exactly one centre from M&Ms that have subjects with all 4 conditions in both train and test sets: Centre 1. We speculate that this similarity in subject conditions may contribute to the effectiveness of the shape loss for Centre 1. That said, there are many factors that can affect performance and generalisation, for example, the acquisition protocol differs between each centre and dataset.

We summarise our experiments for domain adaptation as described in Section 4.4: (1) We take the ACDC-pretrained model and finetune with the entire M&Ms dataset, (2) we take the ACDC-pretrained model and finetune with 5 centre-specific subjects (few-shot learning) to produce centre-specific models, and (3) we perform zero-shot inference with the ACDC-pretrained model. To gauge the effectiveness of using the ACDC-pretrained model, we perform a fourth experiment pipeline: we train U-Net and ACU-Net models from scratch with 5 centre-specific subjects. The 4 experimental settings are labelled as (1) Centre-Agnostic Finetuning, (2) Few-Shot Finetuning, (3) Zero-Shot Inference and (4) Few-Shot (No Pretraining). Figure 6.15 presents ACU-Net performance comparisons between them. Centre-Agnosting Finetuning is the only set of experiments that uses all available training data without being constrained to a few-shot setting, and it performs best overall, suggesting that the model has the capacity to learn from all centres. Zero-shot Inference matches Centre-Agnosting Finetuning in DSC and % AV for some centres, and performs slightly worse for others. This suggests that the pretrained ACU-Net model transfers well to the M&Ms dataset. This is further supported by how few-shot

Model	**Centre**	**DSC**				**% AV**
		All	**RV**	**MYO**	**LV**	
U-Net	1	0.870 ± 0.003	0.873 ± 0.004	0.815 ± 0.004	0.923 ± 0.002	83.2 ± 1.26
	2	0.872 ± 0.001	0.877 ± 0.001	0.844 ± 0.001	0.895 ± 0.003	81.9 ± 0.35
	3	0.885 ± 0.002	0.872 ± 0.003	0.871 ± 0.002	0.912 ± 0.001	85.7 ± 0.42
	4	0.849 ± 0.002	0.825 ± 0.002	0.824 ± 0.002	0.898 ± 0.002	82.3 ± 0.25
	5	0.866 ± 0.002	0.871 ± 0.001	0.830 ± 0.003	0.899 ± 0.004	81.6 ± 0.83
	All	0.865 ± 0.002	0.859 ± 0.001	0.835 ± 0.002	0.903 ± 0.002	82.7 ± 0.43
ACU-Net	1	0.873 ± 0.003	0.873 ± 0.006	0.820 ± 0.004	0.926 ± 0.002	88.1 ± 0.72
	2	0.869 ± 0.002	0.876 ± 0.002	0.842 ± 0.002	0.889 ± 0.001	82.1 ± 0.78
	3	0.885 ± 0.002	0.870 ± 0.004	0.872 ± 0.002	0.913 ± 0.002	85.3 ± 1.05
	4	0.846 ± 0.003	0.825 ± 0.002	0.817 ± 0.004	0.897 ± 0.004	84.4 ± 0.87
	5	0.862 ± 0.003	0.872 ± 0.002	0.822 ± 0.004	0.893 ± 0.003	84.6 ± 1.11
	All	0.863 ± 0.002	0.858 ± 0.003	0.831 ± 0.003	0.900 ± 0.002	84.7 ± 0.68

Table 6.5: Quantitative metrics for cardiac segmentation domain adaptation experiments on the M&Ms dataset. Each experiment is repeated 5 times and mean metrics with standard error are reported.



(a) DSC (b) % AV

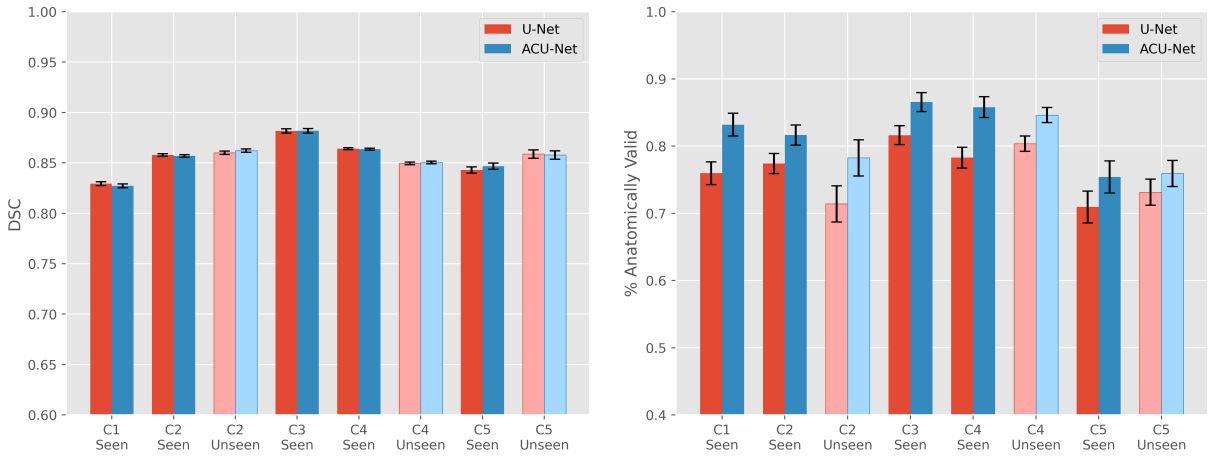
Figure 6.15: Comparison of segmentation quality of ACU-Net for different experimental settings across the centres. Presenting mean DSC and % AV with standard error bars.

learning from scratch performs significantly worse than the finetuned models. Finally, zero-shot inference outperforms few-shot learning, suggesting that finetuning with a small dataset is volatile and degrades the features learned from ACDC pretraining.

Table 6.6 presents the quantitative results of zero-shot inference. Overall, ACU-Net matches U-Net performance in DSC (±0.003 DSC), with an average 5.2% increase in anatomical validity across all centres ($p < 0.03$ for Centres 1, 3, 4). We find that ACU-Net performs similarly across previously seen and unseen pathologies (Figure 6.16).

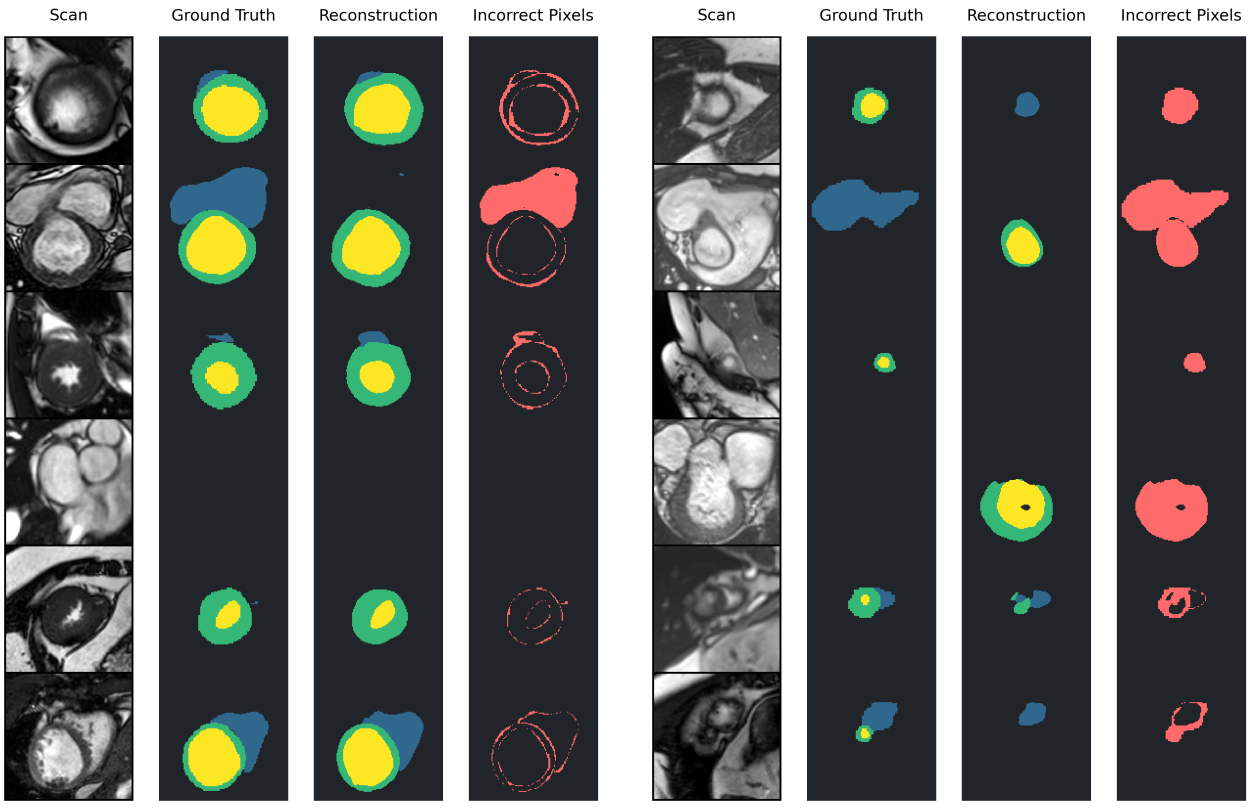
Model	**Centre**	**DSC**				**% AV**
		All	**RV**	**MYO**	**LV**	
U-Net	1	0.829 ± 0.002	0.844 ± 0.003	0.751 ± 0.005	0.894 ± 0.003	76.0 ± 1.67
ACU-Net	1	0.827 ± 0.001	0.837 ± 0.005	0.747 ± 0.004	0.898 ± 0.003	83.2 ± 1.06
U-Net	2	0.858 ± 0.001	0.864 ± 0.002	0.834 ± 0.001	0.876 ± 0.003	76.4 ± 1.64
ACU-Net	2	0.858 ± 0.002	0.864 ± 0.004	0.831 ± 0.001	0.880 ± 0.002	81.1 ± 1.38
U-Net	3	0.882 ± 0.002	0.872 ± 0.005	0.861 ± 0.001	0.912 ± 0.002	81.6 ± 1.43
ACU-Net	3	0.882 ± 0.001	0.874 ± 0.001	0.857 ± 0.001	0.915 ± 0.002	86.5 ± 0.78
U-Net	4	0.855 ± 0.001	0.831 ± 0.003	0.833 ± 0.001	0.899 ± 0.002	79.6 ± 1.18
ACU-Net	4	0.855 ± 0.002	0.831 ± 0.002	0.832 ± 0.001	0.901 ± 0.003	84.9 ± 0.58
U-Net	5	0.847 ± 0.003	0.841 ± 0.005	0.822 ± 0.003	0.879 ± 0.004	71.5 ± 2.24
ACU-Net	5	0.850 ± 0.002	0.839 ± 0.003	0.822 ± 0.004	0.888 ± 0.005	75.5 ± 1.53

Table 6.6: Quantitative metrics for cardiac segmentation zero-shot experiments on the M&Ms dataset. U-Net: Take ACDC-pretrained U-Net model. ACU-Net: Take ACDC-pretrained ACU-Net model. Each experiment is repeated 5 times and mean metrics with standard error are reported.



(a) DSC (b) % AV

Figure 6.16: Comparison of segmentation quality of ACU-Net and U-Net for zero-shot inference. Presenting mean DSC and % AV with standard error bars. C1-C5: Centres 1-5. Seen: using test subset consisting of subjects with conditions seen during ACDC training (Healthy, HCM, DCM, ARV). Unseen: conditions not seen during ACDC.



(a) (b)

Figure 6.17: Visualisations of output segmentations of ACU-Net trained on ACDC, performing zero-shot inference on the M&Ms dataset. Presenting (a) random samples, and (b) some of the worst samples.

Figure 6.17 visualises the output segmentations of ACU-Net performing zero-shot inference on M&Ms, compared to the ground truth. Samples are selected from all 5 centres. The model is capable of producing accurate segmentations despite not having seen the M&Ms dataset. It shares the same weaknesses as seen previously on the ACDC dataset: it can output a segmentation when the GT segmentation is empty, and vice versa. However, 21.8% of the M&Ms data points have empty masks, over 3 times as frequent as the ACDC dataset. On top of that, some GT masks may segment only one or two components (for example, only the RV, or only the RV and MYO), which poses a challenge for the model.

Chapter 7

Conclusion

7.1 Contributions

We revisit the objectives of this dissertation and summarise our contributions.

1. **Evaluate NVAE's capabilities in reconstructing existing segmentation masks and generating synthetic masks, compared to other VAE frameworks.**

   We propose a novel metric for evaluating quality of synthetic masks, the Fréchet ResNet Distance with SimCLR (FRDS). FRDS measures the similarity between synthetic and real cardiac segmentation masks, and is a more robust and consistent metric compared to Fréchet Inception Distance and other existing methods. We evaluate with VAE and NVAE models trained on the ACDC dataset[14] and demonstrate that FRDS aligns well with empirical judgment of synthetic mask quality, but struggles with judging masks with eroded boundaries.

   We design 2 hierarchical architectures based on the NVAE framework for cardiac shape encodin, and perform experiments on the ACDC dataset. The first architecture excels at generating realistic synthetic masks, with 22.0% increase in anatomical validity compared to single-layer VAE models, while offering a competitive reconstruction quality (0.969 DSC[1], compared to 0.891 DSC for single-layer VAE). Furthermore, it achieves a significantly better FRDS of 33.2 (compared to 61.2). The second architecture excels at reconstruction quality, achieving 0.999 DSC.

   Compared to existing NVAE literature[13], our architectures have significantly less parameters and capacity. We demonstrate that a smaller model remains effective in learning compact, robust cardiac shape representations.

2. **Determine the extent to which NVAE's latent representations of cardiac shapes can be used to improve segmentation models.**

   We investigate a downstream task by introducing an anatomical constraint to the U-Net segmentation model[56] in the form of a shape regularisation term, which we refer to as ACU-Net. The shape loss uses the learned latent space of the pretrained, frozen NVAE

[1] Dice coefficient

model to regularise the segmentation model and make it more attentive to the global anatomical shape structure of the heart.

The introduction of a shape loss is not new: the concept is proposed by Oktay et al.[9]. However, we use a pretrained NVAE model while Oktay et al. use an autoencoder trained end-to-end with the segmentation model. Using the ACDC dataset, we demonstrate that the NVAE latent representations are effective in improving segmentation performance, with 5.3% increase in anatomical validity compared to a non-anatomically constrained U-Net while maintaining the same DSC.

3. **Investigate the potential of using NVAE for domain adaptation and cardiac segmentation with few-shot learning.**

   We investigate domain adaptation and few-shot learning by applying the same ACU-Net configuration to the M&Ms dataset[15, 16], which provides scans with different acquisition protocols and vendors.

   We demonstrate that ACU-Net and the NVAE shape regularisation term remains effective when applied to the M&Ms dataset and produces results superior to U-Net (up to 7.2% increase in anatomical validity), despite the latent representations being trained only on the ACDC dataset. With an ACU-Net model pretrained on ACDC, we find zero-shot inference on M&Ms to outperform few-shot learning, with performance almost matching ACU-Net finetuned on the entire M&Ms dataset.

7.2 Future Work

We address the limitations of our work and suggest possible directions for future research.

1. **Datasets**

   A hard bottleneck on model performance is the correctness of ground truth segmentations provided by the ACDC and M&Ms datasets, as our models learn directly from them. A point of concern is the low anatomical validity of the M&Ms dataset at 69.3%, which is mainly caused by pixel-level violations of the anatomical constraints. Indeed, both datasets are acquired from clinical sites and not curated specifically for research applications. We choose to leave the data as is to maintain the integrity of the original datasets.

2. **Nouveau VAE Architecture**

   A key difference between our NVAE architectures for cardiac shape encoding and the configurations proposed by the original authors[13] is that our architectures have less tunable parameters and latent dimensions. This is primarily motivated by time and memory constraints, as well as to avoid overparameterisation on the ACDC and M&Ms datasets, which we consider to be less complex than some of the experimental datasets[47, 48] used in the original paper. Nonetheless, the authors observe that larger models tend to have better performance, and given more time and resources, it would be worth investigating architectural changes such as more layers and groups.

3. **Fréchet ResNet Distance with SimCLR**

   We have acknowledged that FRDS does not align perfectly with empirical judgement of synthetic mask quality, especially for masks with eroded or chipped edges. As such, we

suggest using a complementary metric like % anatomical validity, as well as manual inspection of the results to ensure the absence of corrupt judgment.

4. **ACU-Net**

   We demonstrate the effectiveness of using the pretrained NVAE latent space to form a shape loss for the U-Net segmentation model. Our formulation is a simple prototype and can be extended in several ways, such as weighting each latent group with a separate hyperparameter, or incorporating the standard deviation of the latent residual distributions.

   We provide a U-Net baseline for our ACU-Net experiments. If time permits, it would be interesting to compare how ACU-Net with NVAE shape loss performs against a segmentation model with a shape loss from VAE.

Future work involves addressing the aforementioned limitations, as well as applying the NVAE framework to other downstream applications. For example, replacing LVAE with NVAE for interpretable shape analysis and cardiac diagnosis [5], or using a conditional NVAE with a U-Net for probabilistic segmentation [7, 8]. The NVAE framework can potentially be adapted for 3D cardiac shape analysis by allowing volumetric inputs of a single subject. Finally, our methods can be explored for segmentation of other anatomical structures and medical modalities, such as those concerning brain or lung lesions.

7.3 Final Remarks

In this dissertation, we have demonstrated the effectiveness of the Nouveau VAE framework in learning compact, robust representations of cardiac shapes. Compared to VAE models, NVAE models improve significantly in both reconstruction quality (up to 0.108 DSC increase by achieving 0.999 DSC) and synthetic mask generation (up to 28.0 decrease in FRDS and 22.0% increase in anatomical validity). We propose a novel metric, FRDS, for evaluating synthetic mask quality. Furthermore, we demonstrate the potential of the NVAE latent representations by using them as a regulariser to improve segmentation models (5.3% increase in anatomical validity). We show that these techniques can be used for domain adaptation and zero-shot inference on a previously unseen dataset.

We hope our work contributes to the overarching vision of fully automated, interpretable and accurate segmentation of cardiac scans and diagnosis of cardiac diseases in clinical practice.

Bibliography

[1] Federation WH. World Health Report 2023. World Health Federation; 2023. Accessed: 2024-05-28. Available from: https://world-heart-federation.org/wp-content/uploads/World-Heart-Report-2023.pdf.

[2] Cardiovascular diseases; 2024. Accessed: 2024-05-28. Available from: https://www.who.int/health-topics/cardiovascular-diseases.

[3] An J, Cho S. Variational autoencoder based anomaly detection using reconstruction probability. Special lecture on IE. 2015;2(1):1-18.

[4] Biffi C, Oktay O, Tarroni G, Bai W, De Marvao A, Doumou G, et al. Learning interpretable anatomical features through deep generative models: Application to cardiac remodeling. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11. Springer; 2018. p. 464-71.

[5] Biffi C, Cerrolaza JJ, Tarroni G, Bai W, De Marvao A, Oktay O, et al. Explainable anatomical shape analysis through deep hierarchical generative models. IEEE transactions on medical imaging. 2020;39(6):2088-99.

[6] Painchaud N, Skandarani Y, Judge T, Bernard O, Lalande A, Jodoin PM. Cardiac segmentation with strong anatomical guarantees. IEEE transactions on medical imaging. 2020;39(11):3703-13.

[7] Kohl S, Romera-Paredes B, Meyer C, De Fauw J, Ledsam JR, Maier-Hein K, et al. A probabilistic u-net for segmentation of ambiguous images. Advances in neural information processing systems. 2018;31.

[8] Kohl SA, Romera-Paredes B, Maier-Hein KH, Rezende DJ, Eslami S, Kohli P, et al. A hierarchical probabilistic u-net for modeling multi-scale ambiguities. arXiv preprint arXiv:190513077. 2019.

[9] Oktay O, Ferrante E, Kamnitsas K, Heinrich M, Bai W, Caballero J, et al. Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation. IEEE transactions on medical imaging. 2017;37(2):384-95.

[10] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. Advances in neural information processing systems. 2014;27.

[11] Rezende D, Mohamed S. Variational inference with normalizing flows. In: International conference on machine learning. PMLR; 2015. p. 1530-8.

[12] Makhzani A, Frey BJ. Pixelgan autoencoders. Advances in Neural Information Processing Systems. 2017;30.

[13] Vahdat A, Kautz J. NVAE: A deep hierarchical variational autoencoder. Advances in neural information processing systems. 2020;33:19667-79.

[14] Bernard O, Lalande A, Zotti C, Cervenansky F, Yang X, Heng PA, et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE transactions on medical imaging. 2018;37(11):2514-25.

[15] Campello VM, Gkontra P, Izquierdo C, Martin-Isla C, Sojoudi A, Full PM, et al. Multi-centre, multi-vendor and multi-disease cardiac segmentation: the M&Ms challenge. IEEE Transactions on Medical Imaging. 2021;40(12):3543-54.

[16] Martín-Isla C, Campello VM, Izquierdo C, Kushibar K, Sendra-Balcells C, Gkontra P, et al. Deep learning segmentation of the right ventricle in cardiac MRI: the M&Ms challenge. IEEE Journal of Biomedical and Health Informatics. 2023;27(7):3302-13.

[17] Department for Energy Security & Net Zero. 2024 Government Gas Conversion Factors for company reporting. Department for Energy Security & Net Zero; 2024. This publication is licensed under the terms of the Open Government Licence v3.0. Available from: https://assets.publishing.service.gov.uk/media/66a9fe4ca3c2a28abb50da4a/2024-greenhouse-gas-conversion-factors-methodology.pdf.

[18] Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:230213971. 2023.

[19] Kong Z, Ping W, Huang J, Zhao K, Catanzaro B. Diffwave: A versatile diffusion model for audio synthesis. arXiv preprint arXiv:200909761. 2020.

[20] Kingma DP, Welling M. Auto-encoding variational bayes. arXiv preprint arXiv:13126114. 2013.

[21] Kramer MA. Nonlinear principal component analysis using autoassociative neural networks. AIChE journal. 1991;37(2):233-43.

[22] LeCun Y. The MNIST database of handwritten digits. http://yann lecun com/exdb/mnist/. 1998.

[23] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. science. 2006;313(5786):504-7.

[24] Vincent P, Larochelle H, Bengio Y, Manzagol PA. Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on Machine learning; 2008. p. 1096-103.

[25] Sakurada M, Yairi T. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In: Proceedings of the MLSDA 2014 2nd workshop on machine learning for sensory data analysis; 2014. p. 4-11.

[26] Hsu WN, Zhang Y, Glass J. Unsupervised domain adaptation for robust speech recognition via variational autoencoder-based data augmentation. In: 2017 IEEE automatic speech recognition and understanding workshop (ASRU). IEEE; 2017. p. 16-23.

[27] Bergeron M, Fung N, Hull J, Poulos Z. Variational autoencoders: A hands-off approach to volatility. arXiv preprint arXiv:210203945. 2021.

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

[29] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence. 2013;35(8):1798-828.

[30] Higgins I, Matthey L, Pal A, Burgess CP, Glorot X, Botvinick MM, et al. beta-vae: Learning basic visual concepts with a constrained variational framework. ICLR (Poster). 2017;3.

[31] Chen RT, Li X, Grosse RB, Duvenaud DK. Isolating sources of disentanglement in variational autoencoders. Advances in neural information processing systems. 2018;31.

[32] Kim H, Mnih A. Disentangling by factorising. In: International conference on machine learning. PMLR; 2018. p. 2649-58.

[33] Zhao S, Song J, Ermon S. Infovae: Balancing learning and inference in variational autoencoders. In: Proceedings of the aaai conference on artificial intelligence. vol. 33; 2019. p. 5885-92.

[34] Kim D, Lai CH, Liao WH, Takida Y, Murata N, Uesaka T, et al. PaGoDA: Progressive Growing of a One-Step Generator from a Low-Resolution Diffusion Teacher. arXiv preprint arXiv:240514822. 2024.

[35] Jabri A, Fleet D, Chen T. Scalable adaptive computation for iterative generation. arXiv preprint arXiv:221211972. 2022.

[36] Hang T, Gu S, Li C, Bao J, Chen D, Hu H, et al. Efficient diffusion training via min-snr weighting strategy. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023. p. 7441-51.

[37] Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems. 2017;30.

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

[39] Fréchet M. Sur la distance de deux lois de probabilité. In: Annales de l'ISUP. vol. 6; 1957. p. 183-98.

[40] Ho J, Salimans T. Classifier-free diffusion guidance. arXiv preprint arXiv:220712598. 2022.

[41] Kilgour K, Zuluaga M, Roblek D, Sharifi M. Fr\'echet Audio Distance: A Metric for Evaluating Music Enhancement Algorithms. arXiv preprint arXiv:181208466. 2018.

[42] Unterthiner T, van Steenkiste S, Kurach K, Marinier R, Michalski M, Gelly S. FVD: A new metric for video generation. ICLR 2019 Workshop DeepGenStruct. 2019.

[43] Sønderby CK, Raiko T, Maaløe L, Sønderby SK, Winther O. Ladder variational autoencoders. Advances in neural information processing systems. 2016;29.

[44] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. pmlr; 2015. p. 448-56.

[45] Lake BM, Salakhutdinov R, Tenenbaum JB. Human-level concept learning through probabilistic program induction. Science. 2015;350(6266):1332-8.

[46] LeCun Y, Huang FJ, Bottou L. Learning methods for generic object recognition with invariance to pose and lighting. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.. vol. 2. IEEE; 2004. p. II-104.

[47] Karras T, Aila T, Laine S, Lehtinen J. Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:171010196. 2017.

[48] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2019. p. 4401-10.

[49] Krizhevsky A, Hinton G, et al.. Learning multiple layers of features from tiny images. Toronto, ON, Canada; 2009.

[50] Liu Z, Luo P, Wang X, Tang X. Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision; 2015. p. 3730-8.

[51] Larsen ABL, Sønderby SK, Larochelle H, Winther O. Autoencoding beyond pixels using a learned similarity metric. In: International conference on machine learning. PMLR; 2016. p. 1558-66.

[52] Yoshida Y, Miyato T. Spectral norm regularization for improving the generalizability of deep learning. arXiv preprint arXiv:170510941. 2017.

[53] Miyato T, Kataoka T, Koyama M, Yoshida Y. Spectral normalization for generative adversarial networks. arXiv preprint arXiv:180205957. 2018.

[54] Chen T, Xu B, Zhang C, Guestrin C. Training deep nets with sublinear memory cost. arXiv preprint arXiv:160406174. 2016.

[55] Martens J, Sutskever I. Training deep and recurrent networks with hessian-free optimization. In: Neural Networks: Tricks of the Trade: Second Edition. Springer; 2012. p. 479-535.

[56] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer; 2015. p. 234-41.

[57] Schulz-Menger J, Bluemke DA, Bremerich J, Flamm SD, Fogel MA, Friedrich MG, et al. Standardized image interpretation and post processing in cardiovascular magnetic resonance: Society for Cardiovascular Magnetic Resonance (SCMR) board of trustees task force on standardized post processing. Journal of cardiovascular magnetic resonance. 2013;15(1):35.

[58] Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: International conference on machine learning. PMLR; 2020. p. 1597-607.

[59] Chen T, Kornblith S, Swersky K, Norouzi M, Hinton GE. Big self-supervised models are strong semi-supervised learners. Advances in neural information processing systems. 2020;33:22243-55.

[60] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770-8.

[61] Pearson K. LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin philosophical magazine and journal of science. 1901;2(11):559-72.

[62] Van der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of machine learning research. 2008;9(11).

[63] Kingma DP, Dhariwal P. Glow: Generative flow with invertible 1x1 convolutions. Advances in neural information processing systems. 2018;31.

[64] Baumgartner CF, Koch LM, Pollefeys M, Konukoglu E. An exploration of 2D and 3D deep learning techniques for cardiac MR image segmentation. In: Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges: 8th International Workshop, STACOM 2017, Held in Conjunction with MICCAI 2017, Quebec City, Canada, September 10-14, 2017, Revised Selected Papers 8. Springer; 2018. p. 111-9.

[65] Milletari F, Navab N, Ahmadi SA. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). Ieee; 2016. p. 565-71.

Appendix A

Dataset Details

A.1 Patient Conditions

ACDC - The study population consists of 150 patients categorised into 5 pathologies. Details are provided in Table A.1a. The classification rules are based on the measurements of volumes and masses of the LV, RV and MYO, and can be found on the challenge site[1].

M&Ms - The study population of the full dataset consists of 375 patients categorised into 6 centres and 9 pathologies. However, data from the 6th centre in Canada is not publicly available due to legal reasons, comprising 30 patients. Furthermore, scans from 25 patients are unlabelled; a corresponding ground truth segmentation is not provided. In this dissertation, we use only the 320 labelled scans from the 5 centres in Spain and Germany. Details are provided in Table A.1b.

[1] https://www.creatis.insa-lyon.fr/Challenge/acdc/databases.html

Pathology	# Patients Train	Test
Healthy	20	10
MINF	20	10
DCM	20	10
HCM	20	10
ARV	20	10

(a) ACDC

Pathology		Centre 1	2	3	4	5
Healthy	**Train**	15	23	-	0	0
	Validation	1	3	-	2	3
	Test	5	7	-	9	11
HCM	**Train**	23	26	-	0	0
	Validation	2	2	-	2	4
	Test	0	9	-	3	11
DCM	**Train**	29	0	-	0	0
	Validation	1	0	-	0	2
	Test	7	0	-	0	7
HHD	**Train**	0	1	-	0	0
	Validation	0	0	-	4	0
	Test	0	3	-	6	1
ARV	**Train**	8	0	-	0	0
	Validation	0	0	-	0	1
	Test	4	0	-	2	0
AHS	**Train**	0	0	-	0	0
	Validation	0	0	-	1	0
	Test	0	0	-	0	0
IHD	**Train**	0	0	-	0	0
	Validation	0	0	-	0	0
	Test	0	0	-	3	1
LVNC	**Train**	0	0	-	0	0
	Validation	0	0	-	0	0
	Test	0	0	-	0	2
Other	**Train**	0	0	-	0	0
	Validation	0	0	-	1	0
	Test	0	0	-	17	7
Total	**Train**	75	50	25	0	0
	Validation	4	5	5	10	10
	Test	16	19	21	40	40

(b) M&Ms

Table A.1: Study population statistics for ACDC and M&Ms datasets, organised by patient pathology and data centre. Hyphen (-) indicates data not available; although, Centre 3 provides Healthy, HCM and DCM subjects only. Pathologies are previous myocardial infarction (MINF), dilated cardiomyopathy (DCM), hypertrophic cardiomyopathy (HCM), abnormal right ventricle (ARV), hypertensive heart disease (HHD), athlete heart syndrome (AHS), ischemic heart disease (IHD), left ventricule non-compaction (LVNC).

Appendix B

Architecture Details

B.1 ACU-Net Constant

The training objective of the ACU-Net is presented in (B.1).

$$\mathscr{L}(\theta;\theta_N,\phi_N,\alpha,x,y) = \alpha CE + \omega \sum_{i=1}^{7} \left\| z_i - \hat{z}_i \right\|_2^2 \tag{B.1}$$

In our experiments, we set $\omega = 126717$. To motivate this, we train (1) a baseline U-Net model with cross-entropy loss, and (2) an ACU-Net model with the shape loss term only. The objective of the latter is given by (B.2).

$$\mathscr{L}(\theta;\theta_N,\phi_N,\alpha,x,y) = \sum_{i=1}^{7} \left\| z_i - \hat{z}_i \right\|_2^2 \tag{B.2}$$

During inference, both models achieve 0.91 DSC, so they are comparable. We find that over the entire training stage, the train loss of ACU-Net (with shape loss only) is on average 126717 times smaller than the train loss of U-Net (cross-entropy only). Figure B.1 presents the train graphs, with the shape loss scaled to match the cross-entropy loss. Therefore, the goal of ω in (B.1) is to balance the two terms such that the tunable α weight can be easily interpreted: $\alpha = 1$ means that cross-entropy and shape loss have equal relative influence.

Note that the choice of ω is dependent on the pretrained NVAE model.

Figure B.1: Train loss graphs of U-Net trained with cross-entropy loss and ACU-Net trained with shape loss only ($\alpha = 0$). The shape loss is scaled by $\omega = 126717$ to match the cross-entropy loss.

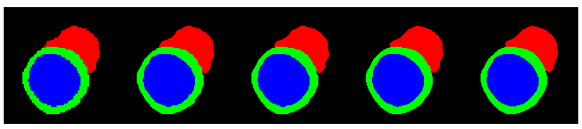
Appendix C

FRDS Disturbance Suite

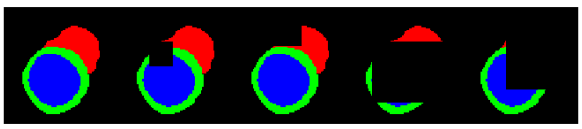
We design a test suite that applies disturbances to segmentation masks at various intensity levels as part of evaluating the robustness of the FRDS metric.

- **Average smoothing:** For $k \in \{3,5,7,9\}$, perform average smoothing with kernel size $k \times k$. Use stride 1 and $\lfloor \frac{k}{2} \rfloor$ padding to retain original dimensions. Then, re-discretise the values. This is equivalent to performing a majority vote within each $k \times k$ window. This tests the metric's ability in penalising blurry generations.
- **Black box crop:** For $c_{\min}, c_{\max} \in \{(0.1,0.3),(0.2,0.5),(0.3,0.7),(0.4,0.9)\}$, choose α, β each within the range $[c_{\min}, c_{\max}]$. Crop a $128\alpha \times 128\beta$ black box randomly within the mask. This tests the metric's ability in penalising incomplete or invalid generations.
- **Elastic deformation:** For $\sigma \in \{8,6,4,2\}$, perform elastic deformation with $\alpha = 300$ and σ, and nearest neighbour interpolation. This tests the metric's sensitivity to unrealistic shape contours.
- **Pepper noise:** For $p \in \{0.0005, 0.005, 0.05, 0.5\}$, each pixel has a p probability of being set to black. This tests the metric's ability to detect fine-grained inaccuracies.

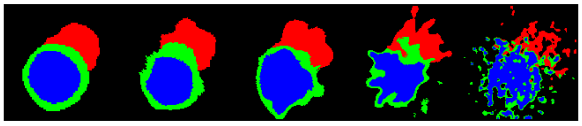
Figure C.1 presents the visual effects of the disturbances.



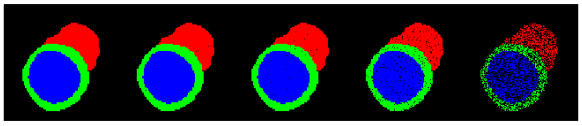
(a) Average smoothing



(b) Black box crop



(c) Elastic deformation



(d) Pepper noise

Figure C.1: Preview of the effects of disturbances on the same mask. Leftmost mask is the original. Then, disturbance level increases from left to right. The effect of average smoothing is empirically subtle unless zoomed in.

Appendix D

Evaluation Metrics

D.1 Dice Coefficient

The Dice coefficient (Dice-Sørensen coefficient, Sørensen-Dice index, Dice similarity coefficient), or DSC for short, is a measure of overlap between two sets of data X and Y. It is formulated as:

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|}$$

In the context of Boolean data, DSC is equivalent to the F1 score: harmonic mean of precision and recall. The score ranges from 0 to 1, where 0 indicates no overlap and 1 indicates perfect overlap.

Similarly, DSC can be computed for 3D volumes[65]. Let A and B be binary segmentation volumes with N voxels. The formulation becomes:

$$DSC = \frac{2\sum_i^N A_i B_i}{\sum_i^N A_i + \sum_i^N B_i}$$

The 2 formulations are identical. For multi-class segmentation, the DSC is computed for each class and the average is taken.

D.2 Welch's t-test

The Welch's t-test is a statistical test that approximates the solution to the Behrens-Fisher problem. In short, it adapts the Student's t-test but does not assume equal variance between two populations.

Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be independent and identically distributed samples from two populations that are normally distributed with unknown means μ_X, μ_Y and unknown variances σ_X^2, σ_Y^2. Furthermore, σ_X^2 and σ_Y^2 are not necessarily equal. The Behrens-Fisher problem is to test the null hypothesis $H_0 : \mu_X = \mu_Y$.

Welch's t-test formulates the test statistic t as:

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{s_{\bar{X}}^2 + s_{\bar{Y}}^2}}$$
$$\text{where } s_{\bar{X}} = \frac{s_X}{\sqrt{n}},\ s_{\bar{Y}} = \frac{s_Y}{\sqrt{m}}$$

Here, $\bar{X}$ and $\bar{Y}$ are the sample means, $s_{\bar{X}}$ and $s_{\bar{Y}}$ are the standard errors, and s_X and s_Y are the sample standard deviations.

The Welch–Satterthwaite equation is used to approximate the degrees of freedom ν. Assuming $n = m$:

$$\nu = \frac{(n-1)\left(s_{\bar{X}}^2 + s_{\bar{Y}}^2\right)^2}{s_{\bar{X}}^4 + s_{\bar{Y}}^4}$$

Then, the t-statistic with ν is used to test the null hypothesis with the Student's t-distribution. A p-value of $p < \alpha$ indicates that there is an α chance that $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are observed if H_0 is true. As example, $p < 0.05$ is often used to reject the null hypothesis with 95% confidence.

Appendix E

Additional Visualisations

E.1 Data Preview

Presenting more previews of preprocessed data from the ACDC and M&Ms datasets. The GT mask (overlayed with opacity) segments the slice into the LV (yellow), RV (blue), MYO (green).

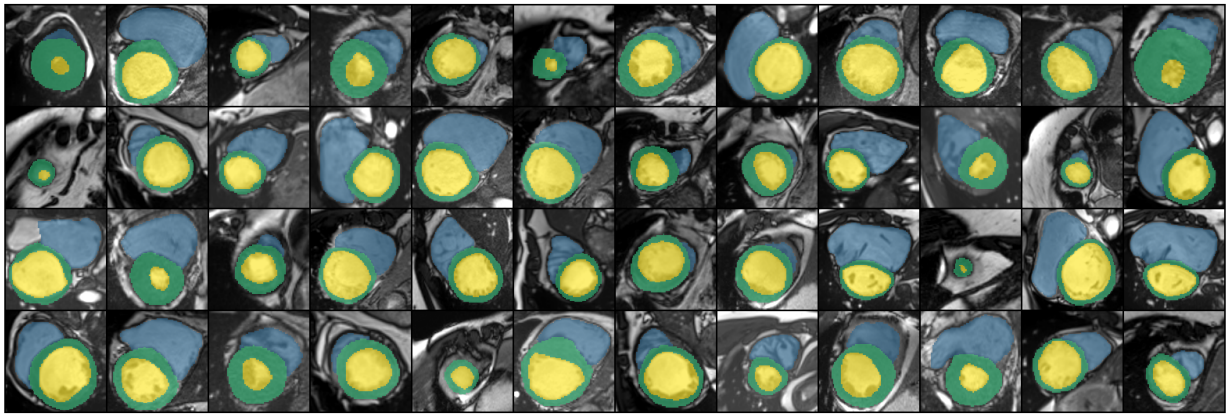


Figure E.1: ACDC dataset.

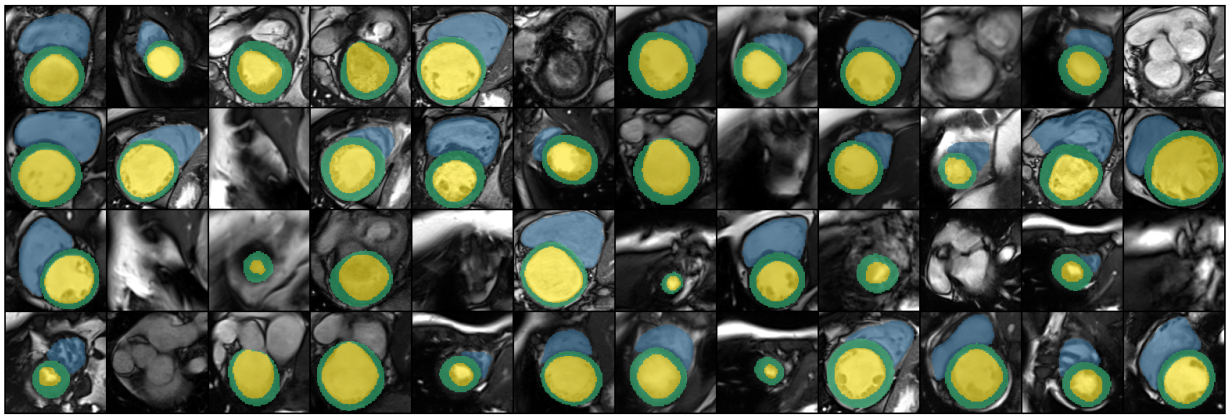


Figure E.2: M&Ms dataset, Centre 1.

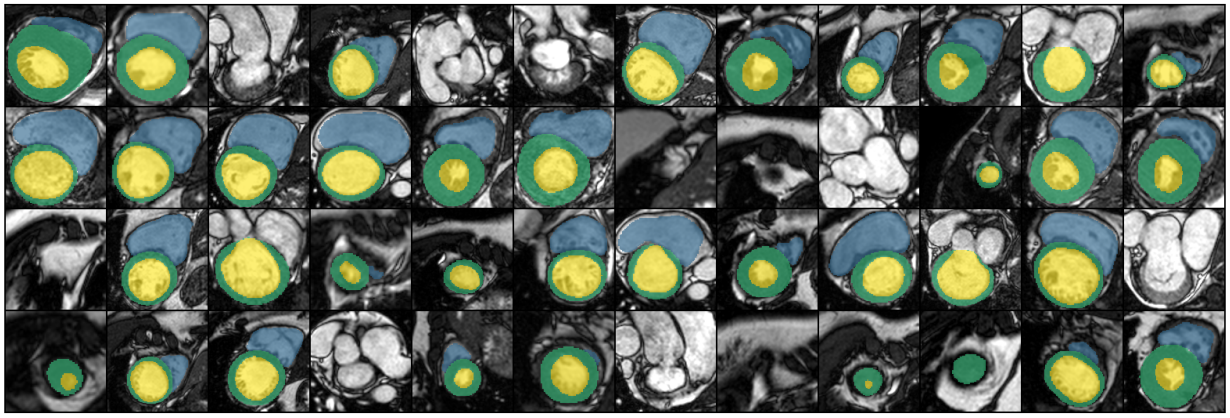


Figure E.3: M&Ms dataset, Centre 2.

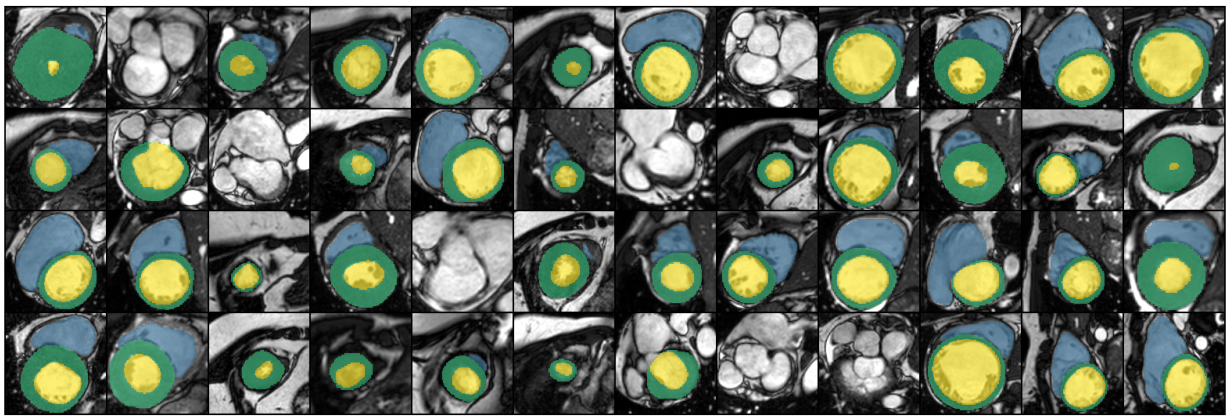


Figure E.4: M&Ms dataset, Centre 3.

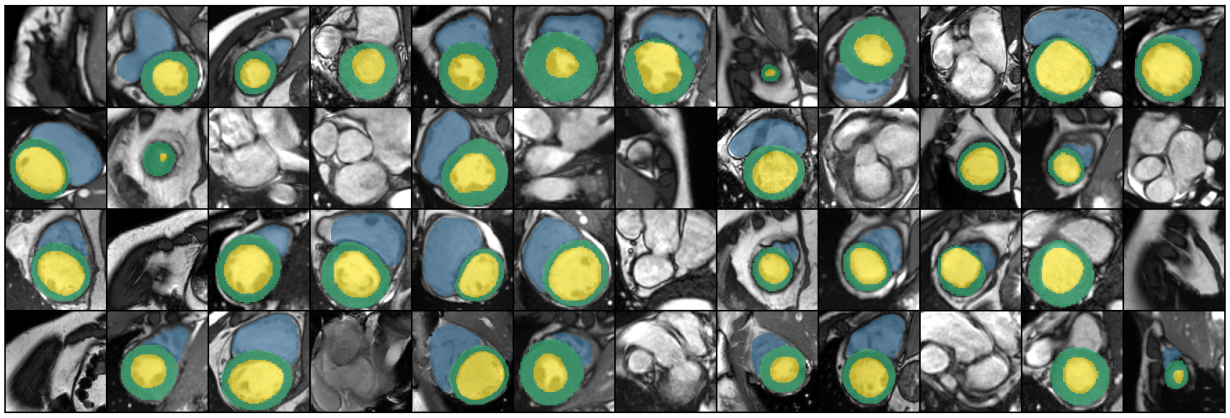


Figure E.5: M&Ms dataset, Centre 4.

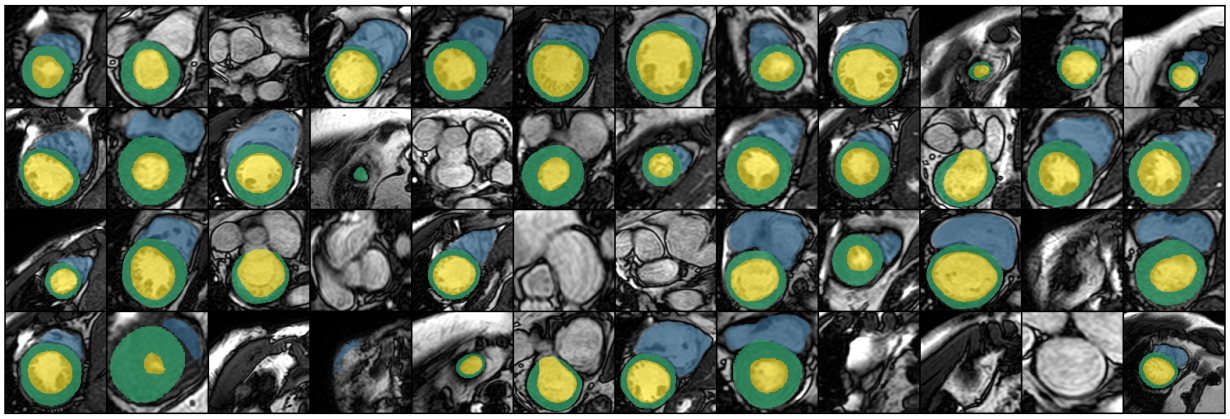


Figure E.6: M&Ms dataset, Centre 5.

E.2 Synthetic Masks

Presenting more previews of generated masks from the best NVAE models.

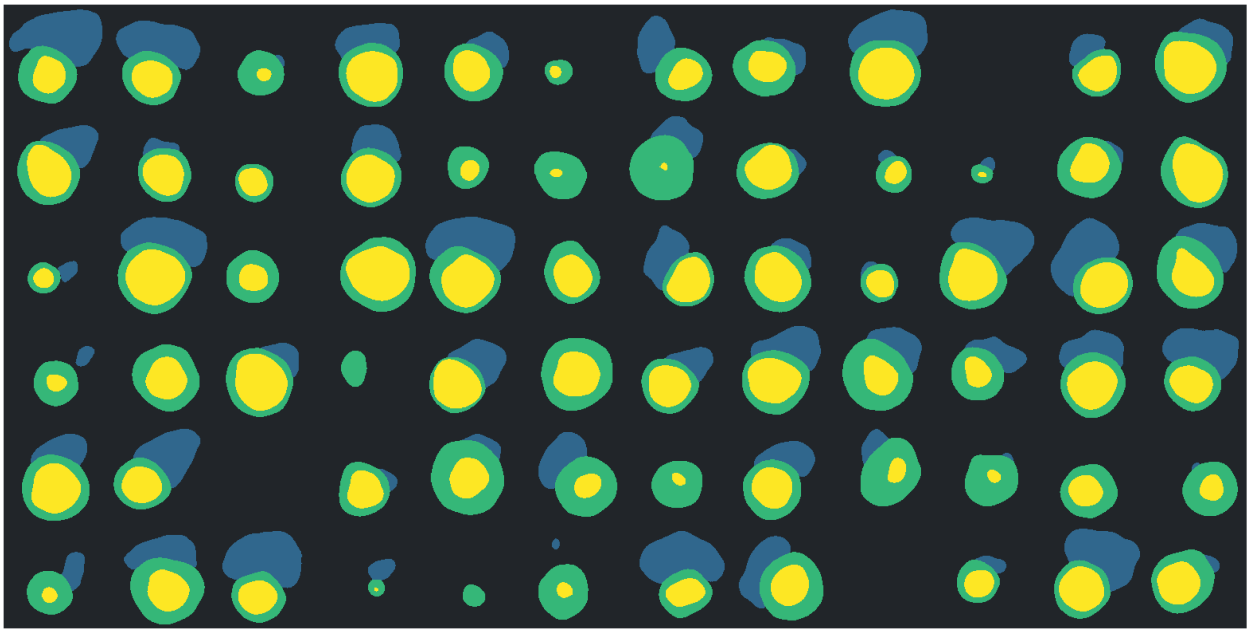


Figure E.7: Default-N

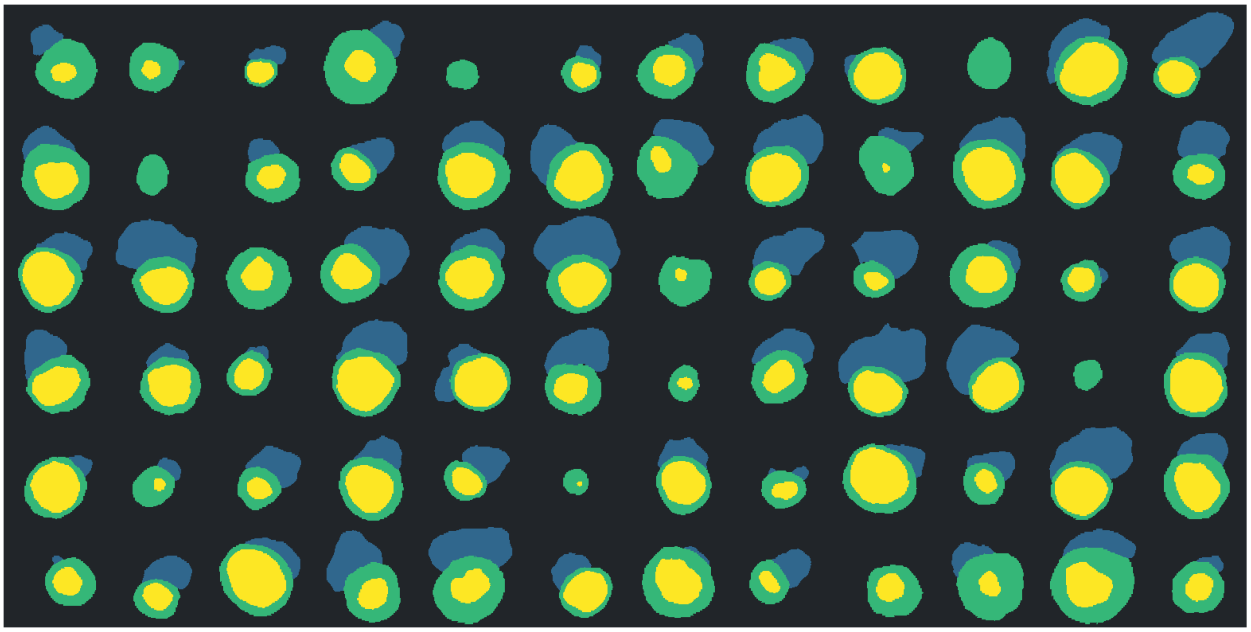


Figure E.8: Default-N with clamping and SR

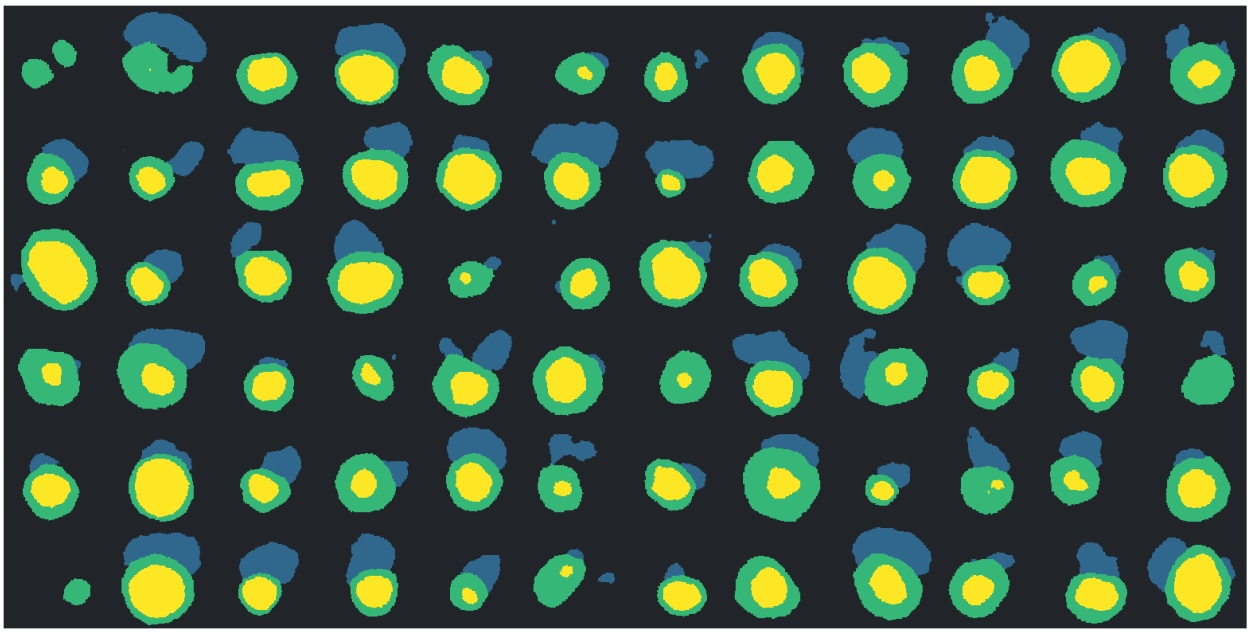


Figure E.9: LatentSkip-N