Dual Degree Project Entitled

Deep Neural Networks for Medical Image Translation, Segmentation, and Analysis: A Comprehensive Study

Submitted by

Rishav Kumar Rajak

(2018CH70302)

Department of Chemical Engineering

Indian Institute of Technology Delhi, New Delhi - 110016

Course Instructor: Prof. Anurag Singh Rathore

Date	Supervisor Signature
08/06/2023	

DECLARATION AND APPROVAL OF SUPERVISOR

This report was written with the help of my supervisor. I took care of the formatting according to the requirements that were supplied. No other study, website, or article sections have been reproduced.

Rishav Kumar Rajak

Rishav Kumar Rajak

Entry Number: 2018CH70302

Date: 08/06/2023

Report entitled "**Deep Neural Networks for Medical Image Translation, Segmentation, and Analysis: A Comprehensive Study**" is approved.

Dr Anurag Singh Rathore

Department of Chemical Engineering

Indian Institute of Technology Delhi

Date: 08/06/2023

ACKNOWLEDGMENT

At first, I would like to express my sincere gratitude to my project guide, Dr Anurag Singh Rathore, for his timely support, backing, and constant encouragement. I am absolutely delighted to present this as my Master's Thesis here at IIT Delhi. I have learned a great deal from this project, and I would like to acknowledge the contribution of Mr Sanjeet Patil (PhD) and Ms Keerthiveena (Post Doc) of AI/ML, Bioprocessing and Bioseparations Lab in my learning curve. The discussions, presentations, general inquisitive environment, and a relish for novelty have pushed me to perform better in all aspects. I would also like to express my gratitude to the respected professors in my evaluation panel for their constructive criticism and valuable ideas. Finally, my heartfelt acknowledgement goes out to all our professors in the Department of Chemical Engineering, IIT Delhi, who have taught me to explore further and comprehend the realms of Chemical Engineering. My tenure at IIT Delhi has made me better prepared to face all kinds of challenges in life ahead.

Rishav Kumar Rajak

ABSTRACT

Accurately segmenting pathological tissue from magnetic resonance imaging can be very useful from a medical point of view. Manual demarcation is tedious, repetitive, and sensitive to minute observer differences; therefore, medical image segmentation is still tricky. It is essential for many clinical applications, such as radiotherapy planning, surgical planning, and diagnosis. Traditional methods for multi-organ segmentation are based on handcrafted features, which are often not robust to variations in image appearance. Deep learning methods, such as convolutional neural networks (CNNs), have shown promising results for multi-organ segmentation.

However, there can be multiple modalities for a medical image like X-Rays, MRI, CT, Ultrasound and PET Scans. Within each modality, there can be various sub-modalities; in MRI, there are T1-weighted, hrT2, T1ce-weighted, FLAIR etc. Each modality provides a viable quantification of the target areas. Some of the tissues can be visible in T1 but not in T2. So, it would become hard to segment the barely visible tissues. In order to tackle this, one must translate the image from one modality to another and apply segmentation. Medical image translation can improve the quality of images, make them more accessible, and use them for a broader range of applications, such as creating synthetic images for training and testing medical image analysis algorithms. This research has dealt with Heart and Brain scans and only considered Magnetic Resonance Imaging (MRI).

This research proposes both image-to-image translations as well as segmentation tasks. We have developed and used various state-of-the-art models for the segmentation part, including ensembling of UNET, nnU-NET, conditional GAN and UNETR. For the domain adaptation part, we have incorporated and developed pix2pix, and CycleGAN. The final part of the study proposes the development of the current state-of-the-art generative model, the Diffusion Network. The developed model has been evaluated on Myocardial Pathological Segmentation 2020 (MyoPS), Brain Tumor Segmentation 2021(BraTS), and cross-modality Domain Adaptation 2022 (crossMoDA) datasets.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

In recent years, medical imaging has emerged as a vital tool in diagnosing and treating various diseases, especially those affecting the cardiovascular system and the brain. Medical image analysis is pivotal in extracting meaningful information from these images, enabling clinicians to make accurate diagnoses and develop personalized treatment strategies. With the advent of deep learning techniques, particularly convolutional neural networks (CNNs), the field of medical image analysis has witnessed significant advancements. These advancements have paved the way for accurate and efficient multi-organ image segmentation, revolutionizing the way we interpret medical images.

Cardiovascular diseases (CVDs)[1] and brain diseases[2], including strokes and tumours, are among the leading causes of mortality and morbidity worldwide. Early and accurate detection of these diseases is crucial for timely intervention and improved patient outcomes. Medical imaging modalities such as magnetic resonance imaging (MRI), computed tomography (CT), positron emission tomography (PET), and ultrasound provide detailed structural and functional information about the cardiovascular system and the brain. These modalities offer diverse perspectives, capturing various aspects of disease pathology. For instance, cardiac MRI provides high-resolution images of the Heart, revealing intricate details of its structures and functions. At the same time, CT angiography enables the assessment of blood vessels for potential blockages.

Similarly, brain MRI allows visualizing anatomical abnormalities, and functional MRI (fMRI) helps study brain activity. These imaging modalities collectively contribute to comprehensive disease evaluation, making multi-organ image segmentation a vital task. This study will focus on Magnetic Resonance Imaging (MRI). MR Images have different modalities, including native (T1) and post-contrast T1-weighted (T1Gd), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (T2-FLAIR).

The rise of deep learning techniques has significantly transformed the field of medical image analysis. CNNs, in particular, have exhibited remarkable performance in a wide range of computer vision tasks, including image classification, object detection, and segmentation. CNNs leverage hierarchical feature learning, automatically extracting relevant features from raw image data. This capability has proven invaluable in medical image analysis, allowing for accurate and efficient segmentation of various organs and tissues.

In the context of multi-organ image segmentation, CNNs have shown immense promise. Traditional segmentation approaches, which relied on handcrafted features and heuristic algorithms, often faced challenges in accurately delineating complex structures and dealing with inter-patient variability. DCNN-based segmentation methods[3] overcome these limitations by learning directly from the data, capturing the inherent variations and complex relationships in medical images. By training on large, annotated datasets, CNNs can generalize well and adapt to new data, making them highly suitable for multi-organ image segmentation.

Despite the successes of CNN-based segmentation, deploying models trained on one dataset to a different domain (e.g., different medical centres, scanner types, or patient populations) often leads to performance degradation due to domain shift. Domain adaptation [4] or Image Translation techniques have emerged as a solution to address this challenge, enabling models to generalize across different domains. Domain adaptation in the context of medical images refers to the process of adapting a machine learning model trained on a source domain (e.g., data from one medical centre or scanner type) to perform well on a target domain (e.g., data from a different medical centre or scanner type). The goal is to minimize the negative impact of domain shift, which refers to the differences in image characteristics, patient populations, acquisition protocols, and other factors between the source and target domains.

Domain adaptation techniques are crucial in addressing the challenges of deploying models trained on a specific dataset to new, unseen data. In medical image analysis, domain adaptation is particularly relevant due to the diversity and heterogeneity of medical imaging data across different institutions and imaging devices.

Several methods have been proposed for domain adaptation in medical images, aiming to bridge the gap between source and target domains while preserving the model's segmentation

performance. These methods can be broadly categorized into feature- and image-level adaptation[5].

In this study, we report various deep-learning frameworks for semantic segmentation. These paired and unpaired domain adaptation techniques in medical images achieved state-of-the-art performance on the respective datasets on which they were trained. Datasets play a vital role in DNN frameworks during training and inference. Hence, these datasets were collected from licensed organizations and medical institutes. In this study, we will first work on the Myocardial pathology segmentation (MyoPS 2020) challenge[6]. In this task, we have successfully reproduced the paper Myocardial Edema and Scar Segmentation Using a Coarse-to-Fine Framework with Weighted Ensemble[7]. This paper stood first in the challenge. We have entirely segmented the Myocardium, myocardial edema and scar of the heart structure with a mean dice score of 0.67 and 0.73 in LV myocardial scar and the union of scar and edema on the test set, respectively.

After this work, we have switched to a more challenging task, Cross-Modality Domain Adaptation for Medical Image Segmentation, crossMoDA 2022[8]. In this task, first, we have to perform unsupervised domain adaptation from contrast-enhanced T1 to high-resolution T2 brain MR images. Finally, the segmentation task aims to segment two critical brain structures (tumour and cochlea) involved in the follow-up and treatment planning of vestibular schwannoma (VS). The segmentation of these two structures is required for radiosurgery, a standard VS treatment. Contrast-enhanced T1 (ceT1) MR imaging is frequently used for patient diagnosis and monitoring. Non-contrast imaging sequences, such as high-resolution T2 (hrT2) imaging, are gaining popularity as they reduce risks connected with gadolinium-containing contrast agents. Additionally, hrT2 imaging is 10 times more economical than ceT1 imaging, enhancing patient safety.

The translational network in the above task wasn't performing well due to some limitations in the architecture. So, we need to move towards a more accurate and robust translation network. Diffusion Model[9] works very well in this domain. Diffusion models are a class of generative models that learn the data distribution by iteratively transforming a given initial distribution to approximate the target distribution. Unlike traditional generative models that generate samples directly from a fixed distribution, diffusion models iteratively update the

distribution to match the target distribution progressively. Diffusion models have gained attention recently due to their ability to generate high-quality samples and provide efficient likelihood estimation. Later in the study, we developed a conditioned Diffusion Model for paired image translation of FLAIR to T1CE Brain MR scans. The performance of the proposed scheme was evaluated on BraTS 2020 dataset[10] and compared with some of the state-of-the-art translational models.

CHAPTER 2

Review of Literature

Traditional CNN methods have significantly advanced the field of medical image analysis, enabling accurate and automated analysis of medical images. These methods leverage the power of deep learning and convolutional neural networks to learn hierarchical features directly from the image data. CNN architectures, such as AlexNet, VGGNet, GoogLeNet, and ResNet, have been widely adopted in medical image analysis tasks. For instance, AlexNet, introduced by (Krizhevsky et al.) [11], has shown remarkable success in image classification tasks, including medical image classification. Similarly, VGGNet[12] is known for its depth and ability to capture rich spatial information. This architecture has been successfully applied in medical image analysis, demonstrating accurate classification and segmentation results. GoogleNet, proposed by (Szegedy et at.)[13], introduced the concept of Inception modules, allowing the network to capture diverse spatial information. This architecture has been applied to medical image analysis, particularly for tasks where capturing local and global information is crucial.

Furthermore, ResNet[14] addressed the problem of vanishing gradients by introducing skip connections, enabling the training of very deep networks. ResNet architectures have been widely utilized in medical image analysis tasks, including brain tumour segmentation and retinal disease classification. These traditional CNN methods have paved the way for more advanced deep-learning techniques in medical image analysis, and they continue to be foundational tool for researchers and practitioners in the field.

2.1 Myocardial Pathological Segmentation

Deep learning (DL)-based algorithms have recently demonstrated satisfactory performance in medical image segmentation. Convolution neural networks (CNNs), in particular, have the broadest applicability in image segmentation[15]. There are several established strategies for segmenting myocardial pathologies. As an illustration, threshold approaches were used by

(Karim et al.)[16] and (Sandford et al.)[17] to segment scars based on the intensity differential between healthy and diseased Myocardium. To segment edema, (Gao et al.)[18] merged a morphological filtering technique with threshold segmentation. A continuous max-flow optimization problem for edema segmentation was developed by (Moccia et al.)[19], considering the shortcomings of thresholding techniques. In addition, pathologic segmentation also made use of graph cuts [20]. By integrating CNNs with graph cuts, (Li et al.)[21] presented a Learn GC framework for scar quantification from LGE CMR data. Fully convolutional networks were introduced by (Moccia et al.)[19] to analyze cardiac scarring objectively. Many proposed medical picture segmentation frameworks have been based on U-Net since its inception[22]. To accomplish multi-task learning of joint left atrial segmentation and scar quantification on LGE CMR images, (Li et al.)[23] presented a multi-branch U-Net. (Zabihollahy et al.)[24] carried out automatic scar segmentation by using a cascaded multi-planar U-Net.

The top performer of the MyoPS 2020 challenge (Zhai et al.)[7] used an ensembling of coarse to fine segmentation strategy that contains two segmentation neural networks. (Martin-Isla et al.)[25] stacked BCDU Networks[26] first to detect the region of interest and then perform ROI-based segmentation. (Zhang et al.)[27] proposed an encoder-decoder-based bi-directional feature pyramid network (BiFPN) to perform segmentation. (Ankenbrand et al.)[28] applied various data augmentation and ensembling techniques to predict the segmentation mask. (Zhang et al.)[29] combined anatomical structure segmentation network (ASSN) and pathological region segmentation network (PRSN). (Jiang et al.)[30] performed dynamic and attention resampling techniques for their fusion U-Net model. Table 1 shows a comparative study of the top 8 architectures presented in the challenge.

Team	Architecture	Batch Size	Patch Size	Preprocessing	Loss Function	Optimizer	Learning Rate	Dice Score (Scar)	Dice (Edema)
UESTC	U-Net	1	160 x 160	crop	CE and Dice Loss	SGD	6e-3 (Decay)	0.708 ± 0.191	0.731 ± 0.109
UBA	U-Net, BCDU-Net	8	256 x 256	crop, IN into [0,1], HE	Weighted BCE and Dice Loss	Adam	1.00E-04	0.701 ± 0.189	0.698 ± 0.129
NPU	Encoder (Efficient Net), Decoder (BiFPN)	64	288 x 288	crop, z-score	CE, Boundary and Dice Loss	Adam	1e-4 (Decay)	0.681 ± 0.240	0.709 ± 0.122
USTB	Dual Attention U-Net	8	256 x 256	crop, z-score, CLAHE, EM	Dice Loss	SGD	1e-3 (Decay)	0.668 ± 0.255	0.688 ± 0.148
UHW	U-Nets (resnet34 backbone)	12	256 x 256	crop, CLAHE	CE and Focal Loss	Adam	1e-3 (Decay)	0.652 ± 0.195	0.665 ± 0.137
FZU	Channel attention-based CNN	16	128 x 128	crop	Dice Loss	Adam	1.00E-03	0.627 ± 0.215	0.686 ± 0.123
NJUST	2D nnU-Net	6	112 x 112	crop, z-score	CE and Dice Loss	SGD	1.00E-03	0.658 ± 0.241	0.599 ± 0.200
CQUPTI	U-Net and a dense connected path	6	256 x 256	crop, z-score	Weighted CE and Dice loss	Adam	1e-4 (Decay)	0.637 ± 0.227	0.656 ± 0.138

Table 1: Summary of the benchmarked algorithms. BiFPN: Bi-directional Feature Pyramid Network; EM: equalization matching; HE: histogram equalization; IN: intensity normalization; CLAHE: contrast limited adaptive histogram equalization; CE: cross entropy; BCE: binary cross entropy; SGD: Stochastic Gradient Descent.

2.2 Brain Tumor Segmentation Challenge (BraTS)

(Myronenko et al.)[31] proposed a variational encoder decoder-based architecture that shares information between each other and imposes additional constraints within the layers. The model won BraTS 2018 challenge [32] [33]. (Jiang et al.)[34] proposed a two-stage cascade U-Net-type architecture which won BraTS 2019 challenge[33]. (Wang et al.)[35] demonstrated a novel Modality-Pairing Architecture for Brain Tumor Segmentation. The authors devised parallel branching to exploit the feature map of various modalities provided in the BraTS 2020 dataset[36] [32] [37]. (Cirillo et al.)[38] proposed a 3D volume-to-volume Generative Adversarial Network and showed promising results in BraTS 2020 challenge. Figure 1 shows the performance of the model mentioned above.

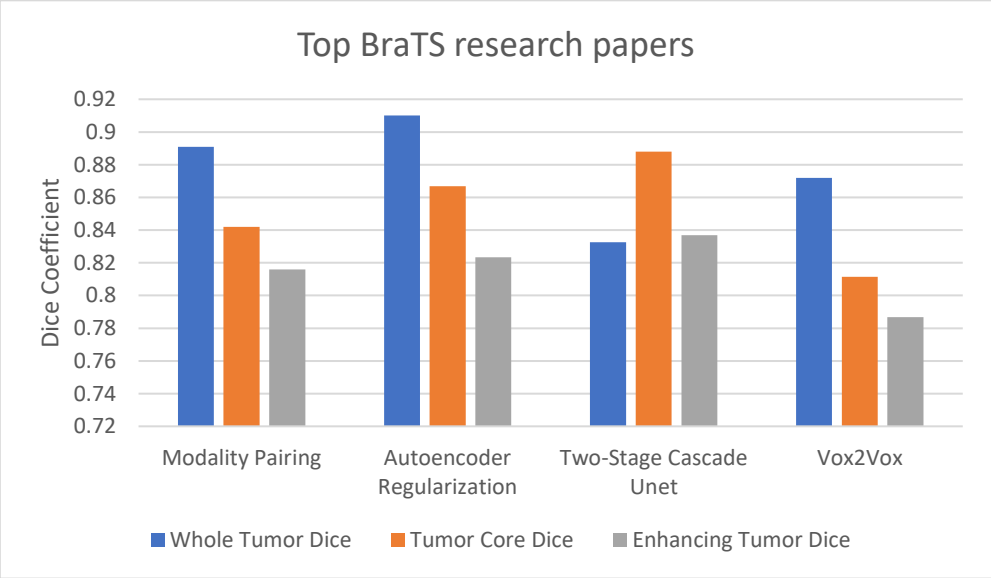


Figure 1: Performance of various architectures presented in the BraTS challenge

2.3 Cross-modality domain adaptation techniques for vestibular schwannoma and cochlea segmentation

(Shin et al.)[39] proposed a target-aware domain adaptation technique using CycleGAN[40] to convert ceT1 to pseudo-hrT2 scans. Then a shared autoencoder was utilized to perform VS and cochleas segmentation. (Dong et al.)[41] developed a pixel alignment and self-training (PAST) network to translate ceT1 scans to hrT2 scans using NiceGAN[42]. After domain shift, segmentation is performed via nnU-Net[43]. (Choi et al.)[44] presented an out-of-the-box CNN framework where CUT[45], patch-wise contrastive learning and adversarial learning was utilized for domain adaptation and nnU-Net for 3D segmentation. (Liu et al.)[39] proposed an image-level domain divergence minimization scheme. Ensembling of CUT, 2D and 3D CycleGAN was used for unpaired image translation and Mean Teacher method [46] to fine-tune the segmentation network. Moving towards other benchmark models, (Yao et al.)[46] proposed a disentangled style GAN framework for style transfer and modified 3D ResUNet[47] with attention modules for segmentation (proposed a regularized translation approach where the images are normalized to MNI space, then modified CycleGAN was utilized for unsupervised domain adaptation from ceT1 to hrT2. After the domain shift, vanilla 3D U-Net architecture was constructed as a segmentation module. A comparison of various proposed methods is made in Table 2.

Team Name	Segmentation Architecture	Translation Network	Cropping	Data Augmentation	Loss Function	Optimizer	Batch Size	Dice Score (VS)	Dice Score (Cochlea)
Samoyed	3D and 2D nnU-Net	CycleGAN with segmentation decoder	Fixed	nnU-Net Aug.	Dice, CE Loss	SGD	2	0.87	0.857
PKU_BIALAB	3D nnU-Net	NiceGAN (2D)	Fixed	nnU-Net Aug.	Dice, CE Loss	SGD	4	0.884	0.806
jwc-rad	3D nnU-Net	CUT (2D)	Fixed	nnU-Net Aug., VS: Int. Aug.	Dice, CE Loss	SGD	2	0.862	0.831
MIP	2.5D U-Net (Attention)	CycleGAN (2D+3D), CUT (3D)	Manual ROI + Rigid Registration	Int. Shift, Contrast, AE	Dice, CE Loss	Adam	1	0.837	0.832
PremiLab	DAR-U-Net	Content-Style GAN (3D)	x	Affine and Elastic Deformation	Dice, Focal Loss	Adabelief	2	0.833	0.807
Epione-Liryc	3D U-Net	CycleGAN with Pair Loss	MNI registration label-based ROI	Flipping, rotation, Int. Noise	Dice Loss	Adam	1	0.853	0.778
MedICl	2.5D U-Net, 3D CNN	CycleGAN (2D)	MNI registration label-based ROI	Affine, Elastic Deformation and Int. Aug.	Dice Loss	Adam	2	0.84	0.752
DBMI_pitt	3D U-Net (Attention)	CUT (3D)	x	Int. Shift, Resizing, Affine Deformation	Attention Dice	Adam	4	0.501	0.816

Table 2: Summary of the benchmarked algorithms. EM: equalization matching; HE: histogram equalization; Int: intensity; Aug.: Augmentation; AE: Affine Deformation; CE: cross entropy; BCE: binary cross entropy; MNI: Montreal Neurological Institute.

2.4 Diffusion Model

Wake-sleep algorithm[48] popularised the concept of testing generative probabilistic and inference models against one another. With a few exceptions, this strategy mainly remained undiscovered for almost 20 years[49]. Recently, effort in developing this concept has exploded. Variational learning and inference methods were created[50], allowing a flexible generative model and posterior distribution over latent variables to be trained against one another.

For image-to-image issues, including unpaired translation[40], unsupervised cross-domain generation[51], multi-domain translation[52], and few-shot translation[53], (GAN-based solutions have also been presented in the past years.

However, current GAN models can fail to translate pictures holistically with constant structural and textural regularity. Concerning image generation, audio synthesis, image super-resolution, unpaired image-to-image translation, and image editing, diffusion models[9] have recently been developed and have produced impressive results[54]. On top of these recent developments, our conditional diffusion models demonstrate adaptability to various image-to-image translation challenges.

Most inpainting and other linear inverse problem diffusion models[55] have modified unconditional models for usage in conditional tasks. (Saharia et al.)[56] proposed a unified conditional diffusion framework that performs image translation, including inpainting, colourization, JPEG restoration and uncropping.

(Sasaki et al.)[57] developed a novel image-to-image translation architecture that used a denoising diffusion probabilistic model[54] without the use of adversarial training.

CHAPTER 3

Methodology and Modelling

3.1 Dataset and Preprocessing

3.1.1 MyoPS 2020

MyoPS 2020 challenge[6] provides three sequences (each subject) of MRI scans[58], i.e., T2 CMR, Balanced Steady-State Free Precession (bSSFP) CMR, and Late Gadolinium Enhancement (LGE) CMR of 45 patients. The ground truth consists of Left Ventricle Blood Pool, Right Ventricle Blood Pool, Normal Myocardium, Edema, and Scar classes. The scans are from male patients with acute Myocardial Infarction and with an average age and weight of 56 and 74 respectively. The challenge provided the dataset in 3D volumes (NIFTI) and divided as 25 pairs of training set and 20 pairs for the validation set. Figure 2 depicts the modalities and segmentation classes in the gold label.

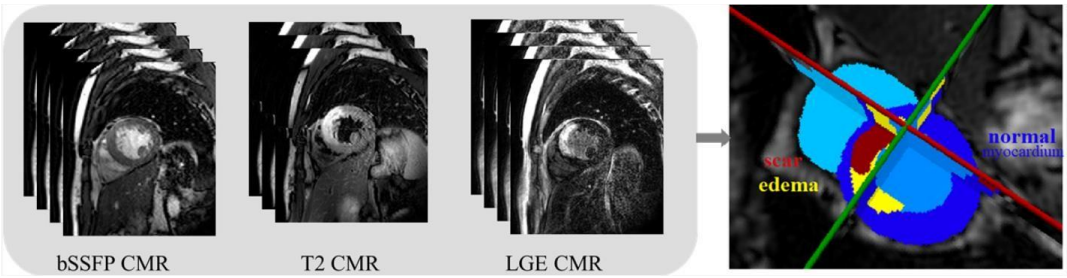


Figure 2: bSSFP, T2, and LGE CMR modalities, along with the segmentation masks [6]

Data preprocessing is necessary before feeding into the model to achieve better results. Hence, in our implementation, the images are cropped with a cardiac bounding box (our region of interest) with a margin of 30 voxels (Figure 3). A voxel is a 3D unit of an image with one value, as in the case of a pixel in a 2D image. The ground truth is also cropped similarly to have consistency with the modalities.

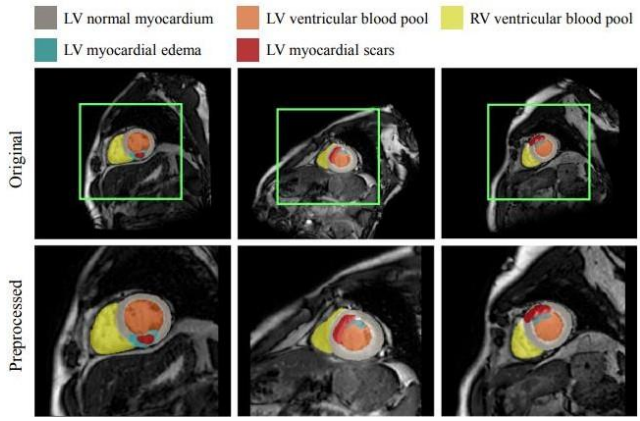


Figure 3: Original vs Preprocessed Image

3.1.2 MICCAI crossMoDA 2022

MICCAI CrossMoDA 2022 released the dataset[59] containing data from two sources, London and Tilburg. London data contains 105 labeled ceT1 images and 105 unpaired hrT2 images, while Tilburg data contains 105 labeled ceT1 images and 105 unpaired hrT2 images. ceT1 images have in-plane matrix of 512x512 with 120 channels while hrT2 images have in-plane matrix of 384x384 or 448x448 with varying channels between 30 to 80. Images and segmentation masks were distributed as compressed NIfTI files (.nii.gz). Figure 4 illustrates the types of images.

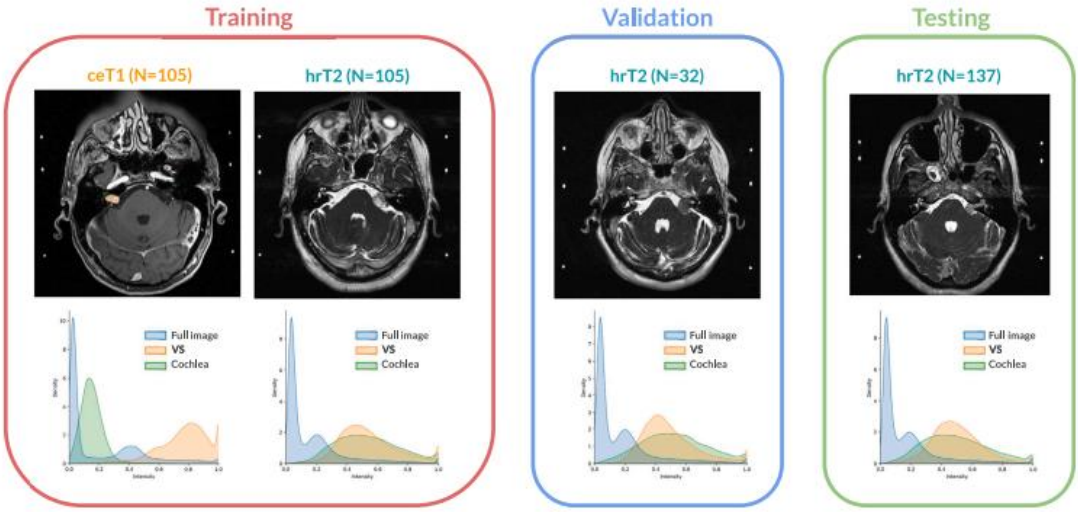


Figure 4: ceT1 and hrT2 scans with Intensity distribution for training, validation and testing set are depicted[39]

Images were of different shapes and had a varied number of channels. Most of the area in the in-plane matrix was useless as it didn't contain VS and cochlea near the periphery. We performed centre-cropping to reduce computational cost and avoid such irrelevant information to make the images in-plane resolution 256x256. Further, the out-of-the-plane distance between each slice was different, and it is necessary to have the same planar distance for combined training; to solve this, we changed the target spacing to [0.5, 0.5, 1.5]. Additionally, the pixel values were significant for the images; this would have caused slow training; therefore, we normalized the pixel values from [-1,1]. Moreover, in the ceT1 scans of London data, 120 slices were there, but the labels were present in almost 20 slices. The rest slices were useless, so we picked ±10 slices from the beginning of labels and end of labels slices.

3.1.3 RSNA-ASNR-MICCAI BraTS 2021

BraTS 2021[10] [32] [33] training dataset consists of 1251 MR volumes with the dimensions 240 by 240 by 155. To assess tumour heterogeneity, MRI is necessary. T1 weighted sequence (T1), T1-weighted contrast-enhanced sequence employing gadolinium contrast agents (T1Gd) (T1CE), T2 weighted sequence (T2), and Fluid attenuated inversion recovery (FLAIR) sequence are the four MRI sequences that are typically used to detect gliomas. The Enhancing Tumor (ET), which corresponds to an area of relative hyper-intensity in the T1CE relative to the T1 sequence, the Non-Enhancing Tumor (NET), the Necrotic Tumor (NCR), which are both hypointense in T1-Gd relative to T1, and the Peritumoral Edema (ED), which is hyper-intense in FLAIR sequence, are the four distinct tumour sub-regions that can be distinguished from these sequences. Modalities are shown in Figure 5.

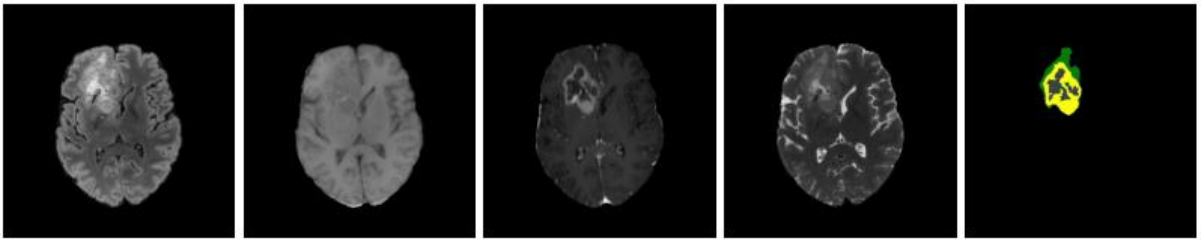


Figure 5: FLAIR, T1, T1CE, T2 and Ground Truth (Left to Right)[10]

BraTS dataset was used for translation network using diffusion model. Usually, translation occurs in pairwise fashion and since the dataset was also paired, we chose two out of four modalities. In our case, we picked FLAIR and T1CE because these datasets were highly contrastive. We slice the 3D MR images axially. Because tumors are uncommon in the upper or lower parts of the brain, we omit the bottom 80 slices and the top 26 slices. We cropped the MRIs to 224 x 224 dimension. The scans contain pixel values in varied range with 10^3 magnitudes. Hence, the images need to be normalized in the range of [-1,1] to make optimizers converge quickly[60]. After preprocessing, we randomly selected 1000 slices of (FLAIR, T1CE) pairs for training and 100 slices for testing.

3.2 Model Architectures

3.2.1 Coarse-to-fine network architecture

This architecture is trained on MyoPS 2020 dataset to segment LV, RV, Myocardial Wall, Edema and Scar. The architecture consists of two parts: Coarse and Fine. In both networks, U-Net[22] is used in its naïve form and modified form respectively. Figure 6 depicts U-Net as a "U" shaped network. The design is symmetrical and is divided into two parts: the expanding path, which is made up of up-sampling layers to extract features and reshape into the original proportions and the contracting path, which shrinks the picture to extract features.

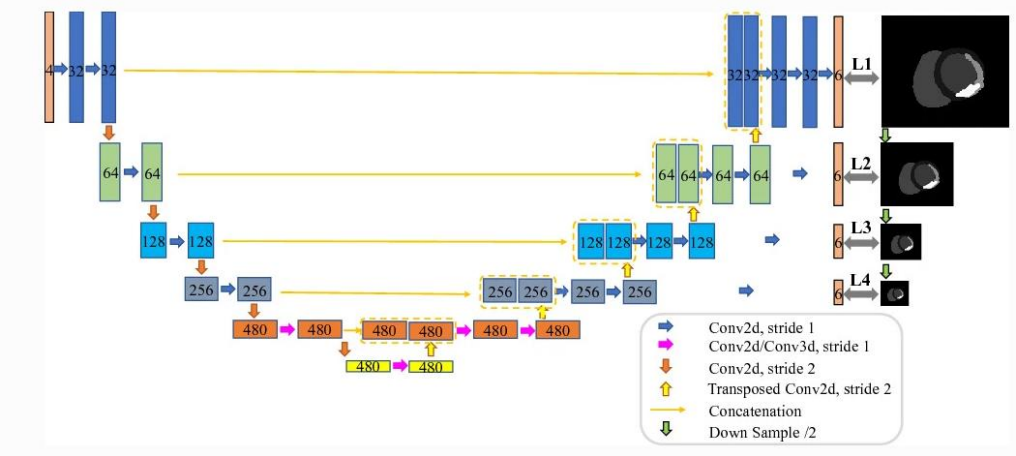


Figure 6: U-Net Architecture[22]

Coarse Network: LV, RV and myocardial segmentation (three classes) are performed in the coarse framework. The segmentation output from this network contains our region of interest, i.e., edema and scar regions. Then in the fine network, myocardial previous location information was used to get a more detailed segmentation of edema and scar. All three channels of CMR images were concatenated and trained to a U-Net model to segment the background, LV, RV, and Myocardium.

Fine Network: The modalities were cropped in accordance with the output of the coarse segmentation network. Then, the cropped images were stacked with the cropped images and a 4-channel input was formed to feed to the fine segmentation network. Here, nnU-Net[43] was utilized. nnU-Net is one of the current state-of-the-art modified U-Net models to perform segmentation for the brain, pancreas tumours, etc. A 2.5D U-Net model can employ inter-slice features and is more efficient than a 3D U-Net model.

Figure 7 depicts the network structure. The image shows that CMR 3 channel images are cropped and fed into the coarse network for myocardial localization. Then the output is concatenated with the cropped CMR images and fed into the fine model to get the predictions.

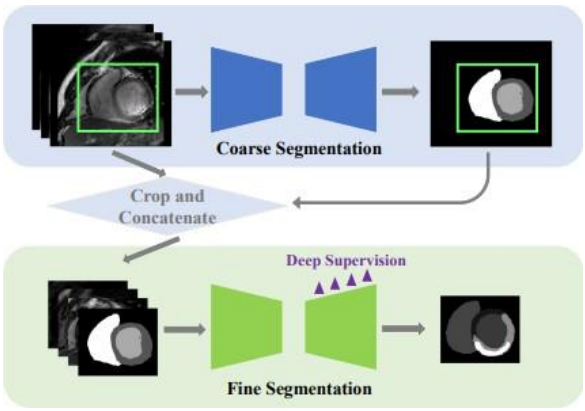


Figure 7: Coarse to Fine Segmentation Network[7]

In the postprocessing step, the weighted ensemble method was developed. The process assigns weights to each class's prediction from different networks and adds them. The class weights concept[61] was introduced to make the model more robust to the minority class. A

weighted ensemble method of 2D and 2.5D outputs from the fine network is exploited to predict each class better. In this segmentation challenge, scar and edema were not correctly localized and visible. Hence, more weights were given to them, as depicted in Figure 8. Implementation and training strategy were replicated from [7].

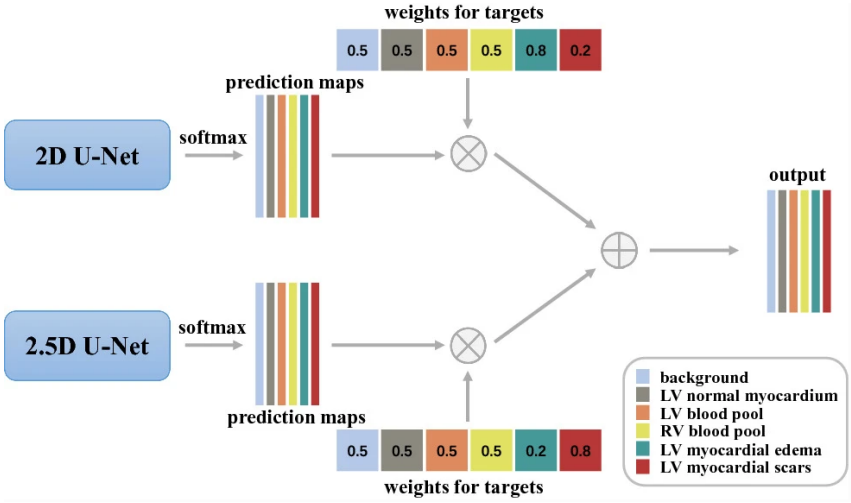


Figure 8: Ensemble Network incorporating class weights[7]

3.2.2 Conditioned GANs for segmentation

Conditioned Generative Adversarial Networks (cGANs)[62] are a type of GAN that can be used for image segmentation tasks. In cGANs, the generator and discriminator networks are conditioned on additional information, such as a segmentation mask or a class label.

For image segmentation tasks, the generator network takes a noise vector and a segmentation mask as input and generates a synthetic image corresponding to the input segmentation mask. The discriminator network takes as input a real or synthetic image and a corresponding segmentation mask and tries to distinguish between the real and synthetic images.

One example of a cGAN for image segmentation is the Pix2Pix model, introduced by (Isola et al.)[63]. In this model, the generator network is a U-Net architecture that takes as input a segmentation mask and generates a corresponding image. The discriminator network is a Patch GAN architecture that takes as input a real or synthetic image and a corresponding segmentation mask and outputs a probability map indicating whether the input is real or synthetic.

The cGAN architecture typically consists of two main components for image segmentation tasks: a generator and a discriminator.

Generator: The generator inputs a random noise vector and a given segmentation mask. The segmentation mask provides conditional information, which helps the generator create more realistic and contextually relevant images. The generator aims to produce synthetic images that resemble the real images corresponding to the given segmentation masks.

Discriminator: The discriminator's role is to distinguish between real images (paired with their corresponding segmentation masks) and synthetic images generated by the generator (also paired with their related segmentation masks). The discriminator is trained to classify the input pairs as real or fake, while the generator is trained to create images that can fool the discriminator.

During training, the generator and discriminator play a minimax game, where the generator tries to create images that the discriminator cannot distinguish from real images, and the discriminator tries to improve its ability to differentiate between real and generated images. The architecture is shown in Figure 9. This adversarial process continues until the generator produces realistic images that match the given segmentation masks, and the discriminator can no longer accurately classify the images as real or fake. Implementation details are provided in Appendix.

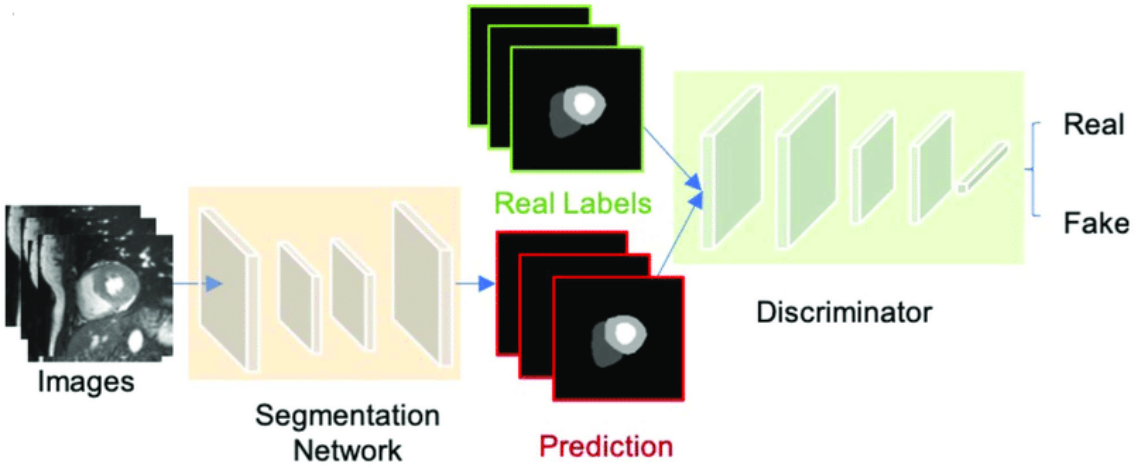


Figure 9: GAN as a segmentation network

3.2.3 CrossMoDA 2022 Challenge Architecture

3.2.3.1 CycleGAN for Unpaired Image-to-Image Translation

The dataset provided to us was unpaired; let source T1 domain images be represented as $x^s = \{x_i^s\}_{i=1}^{N_s}$, and its label be $y^s = \{y_i^s\}_{i=1}^{N_s}$, similarly for target T2 scans, $x^t = \{x_i^t\}_{i=1}^{N_t}$, here N_s = N_t =105. The translated output from CycleGAN can be represented by $\chi^t = \{\chi_i^t\}_{i=1}^{N_s}$. The translated or pseudo hrT2 scans contain the features of ceT1 images, but the intensity and distribution of hrT2 scans would be further used to train the segmentation network. A generator and a discriminator are the two neural networks that make up generative adversarial models (GANs). Two GANs make up a CycleGAN, giving it two generators and two discriminators. One generator transforms ceT1 into hrT2, and the other transforms hrT2 into ceT1, given two sets of unpaired pictures, ceT1, and hrT2, as an example. Discriminators are used throughout the training phase to determine whether generated images are authentic or false. A generator in CycleGAN receives additional feedback from the other generator. This feedback ensures that a picture produced by a generator is cycle consistent, which means that using two generators in succession should provide a similar image. Using their respective discriminators' input, this approach can help generators improve. The CycleGAN network was trained on three losses, Cycle-consistency loss, Adversarial loss, and Identity loss.

$$CycleGAN^{T1\to T2}: x^s \to \chi^t$$

3.2.3.1.1 Adversarial Loss

Mapping $G: S \to T$ and its discriminator D^T, Adversarial loss is incorporated:

$$L_{adv} = \ln\left(D^T(x^t)\right) + \ln\left(1 - D^T\left(C^{S\to T}(x^s)\right)\right) + \ln\left(D^S(x^S)\right) + \ln\left(1 - D^S\left(C^{T\to S}(x^t)\right)\right)$$

3.2.3.1.2 Cycle Consistency Loss

To maintain cycle consistency, L1 loss is imposed between the Source and translated from target to source scans,

$$L_{cycle} = \left\| x^s - C^{T\to S}\left(C^{S\to T}(x^s)\right)\right\|_1 + \left\| x^t - C^{S\to T}\left(C^{T\to S}(x^t)\right)\right\|_1$$

3.2.3.1.3 Identity Loss

This loss takes care of the identity nature of the network; if the ceT1 image is passed to the ceT1 generator, the translated image would be identical to the input image, i.e., not altered, and vice versa.

$$L_{identity} = \left\| x^{\mathrm{t}} - C^{S \to T}(x^t) \right\|_1 + \| x^s - C^{T \to S}(x^s) \|_1$$

The final objective function becomes:

$$\begin{aligned} L(G^{S \to T}, G^{T \to S}, D^{\mathrm{S}}, D^T) \\ = \lambda_1 L_{cycle}(\mathrm{C}^{\mathrm{S} \to T}, \mathrm{C}^{T \to S}) + \lambda_2 L_{adv}(\mathrm{C}^{\mathrm{S} \to T}, \mathrm{C}^{T \to \mathrm{S}}, D^{\mathrm{S}}, D^T) \\ + \lambda_3 L_{identity}(C^{S \to T}, C^{T \to S}) \end{aligned}$$

where G corresponds to Generator, D denotes the discriminator, and C means colour composition and λ_i denotes weighing factor for ith loss. Implementation Details are provided in Appendix.

3.2.3.2 crossMoDA segmentation networks

The main goal of the crossMoDA challenge is to segment the two critical brain structures, VS and cochlea. To perform segmentation, we have used two state-of-the-art architectures, UNETR and nnU-Net. The architectures are discussed briefly below.

3.2.3.2.1 UNETR Segmentation Architecture

U-Net Transformers[64], also known as UNETR, efficiently capture multi-scale characteristics and learn sequence representations of the input volume using an encoder-based transformer. The encoder and decoder are designed using the well-established "U-shaped" network architecture. The transformer encoder is directly connected to a decoder to determine the final semantic segmentation output using skip connections.

In the UNETR architecture, the authors extracted a sequence represented by z_i ($i \in \{3,6,9,12\}$) of size $\frac{H*W*D}{P^3} * K$ from the transformer and reshape it into a $\frac{H}{P} * \frac{W}{P} * \frac{D}{P} *$ K tensor to replicate

U-Net-like architectures[24] [Figure 10], where features from various encoder resolutions are combined with the decoder. The authors added a de-convolutional layer to the modified feature map to double its resolution at the encoder bottleneck (i.e., the output of the last layer of the transformer). Then, using a deconvolutional layer to upsample the outcome, they concatenated the enlarged feature map with the element map of the previous transformer output (for example, z_9) and fed them into successive 3*3*3 convolutional layers. Finally, SoftMax function is used to get semantic predictions. Implementation Details are provided in Appendix.

3.2.3.2.1.1 Loss Function

The loss function consists of dice and cross-entropy losses, which can be calculated from the equation below. Here, J denotes the different classes, I represents the voxel numbers, $G_{i,j}$ denotes one-hot encoded labels, and $Y_{i,j}$ denotes probability output for class j at voxel i.

$$\mathcal{L}(G,Y) = 1 - \frac{2}{J}\sum_{j=1}^{J}\frac{\sum_{i=1}^{I} G_{i,j}Y_{i,j}}{\sum_{i=1}^{I} G_{i,j}^2 + \sum_{i=1}^{I} Y_{i,j}^2} - \\ -\frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{J} G_{i,j}\log Y_{i,j}.$$

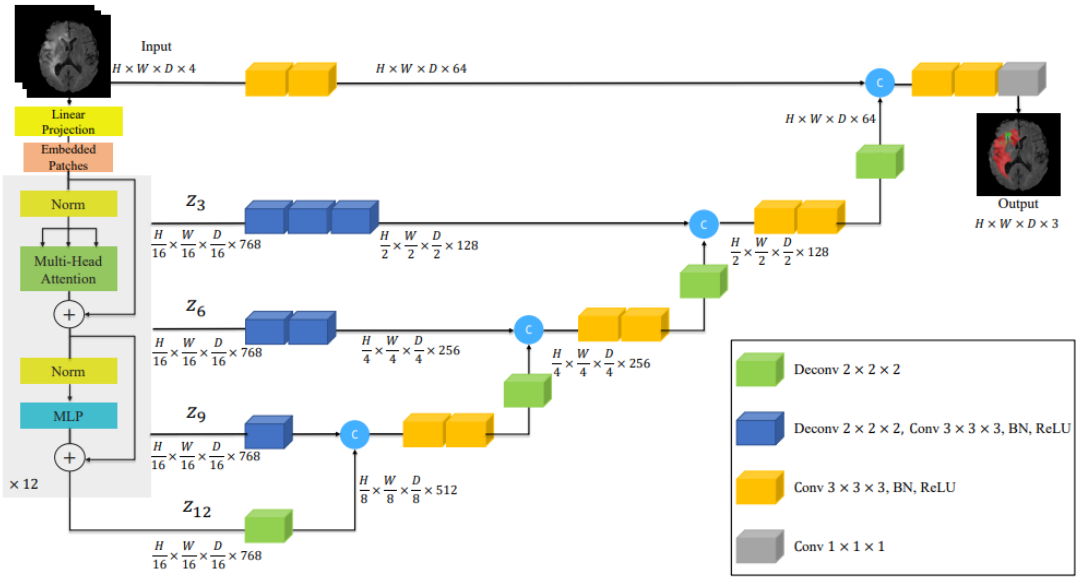


Figure 10: UNetR Architecture [65]

3.2.3.2.2 nnU-Net

nnU-Net[43] has outperformed many state-of-the-art architectures in image segmentation. The "data fingerprint" is a data-dependent hyper-parameter set that the nnU-Net pipeline employs to identify when ingesting training data.

Pipeline fingerprints, shown in Figure 11, are created using the data fingerprint, loss function, optimizer, inferred parameters (batch, patch size, image resampling and normalization) and architecture. Using the best-selected hyper-parameters, the pipeline model creates network training for 2D, 3D- Cascade U-Net, and 3D-full res. The optimal average dice coefficient for the training data is chosen by combining the various network configurations and postprocessing. The predictions for the test data will then be generated using the optimal setup. Similar Implementation was followed as mentioned in [43].

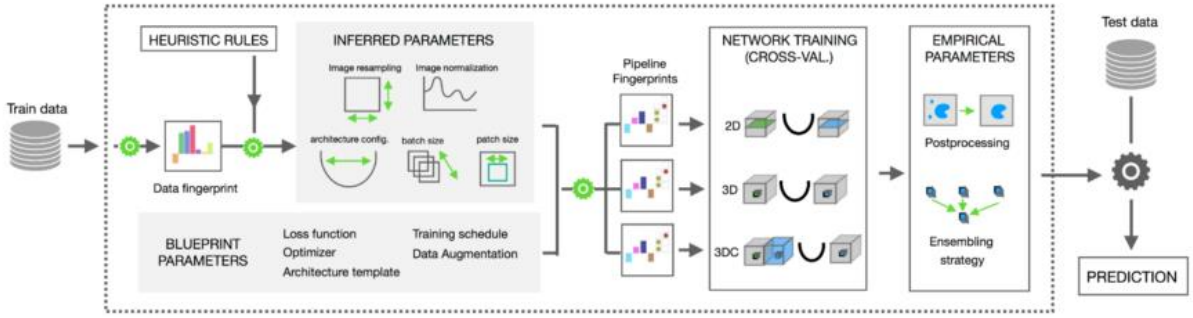


Figure 11: nnU-Net workflow [43]

3.2.4 Diffusion Model for Domain Translation

Diffusion Models**[65]**, the current state of the art for text-to-image generation, whose condition can be changed according to us for unsupervised domain adaptation purposes. Nonequilibrium thermodynamics[9] serves as the basis for diffusion models. Since they are generative models, they may be used to produce data that resembles the data on which they were trained. Diffusion models function adds Gaussian noise to data at each step, and the neural network learns to denoise the data to get the previous form. Diffusion models are relatively novel architecture in the medical imaging domain, which, when exploited, can be

used to give a better translation. In this study, we have successfully developed a conditioned diffusion model by applying the mastered denoising technique to randomly sampled noise. The diffusion model comprises the Forward Diffusion Step (Systematic Addition of Gaussian noise to the image) and Backward Denoising Step (Removing added noise from the images). Let's see these steps one by one.

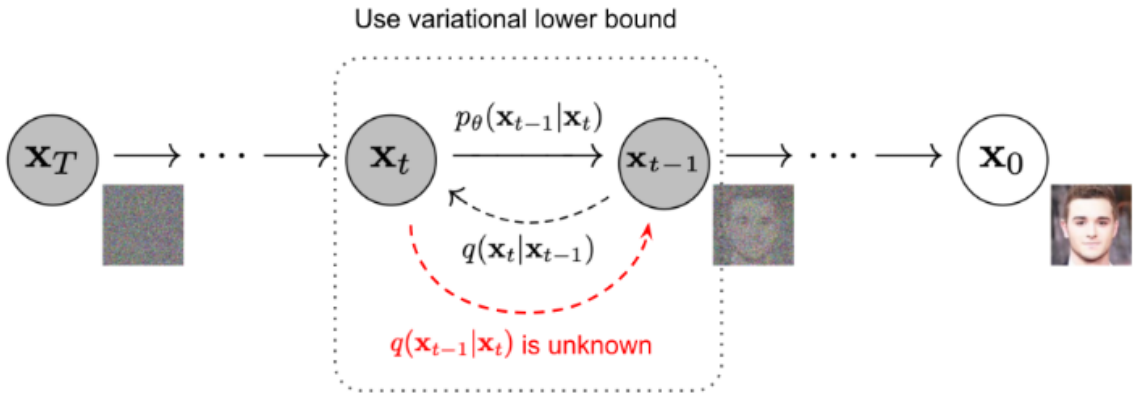


Figure 12: Markov Chain of forward diffusion process[65]

3.2.4.1 Forward Diffusion Step

In the forward diffusion process, we add a small amount of Gaussian noise to a given image x, in T steps, such that it produces noisy images x_1, ……,x_T. The image would gradually become distorted with an increase in step t. The forward step is given by

$$q(x_t|x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_{tI}\right)$$

Where β_1,…,β_T denotes the variances at each timestep, I is the identity matrix, and N denotes normal distribution.

Doing the forward step t times, we can write

$$q(x_t|x_0) = \mathcal{N}\left(x_t; \sqrt{\overline{\alpha_t}}x_0, (1-\overline{\alpha_t})I\right)$$

where $\alpha_t = 1 - \beta_t$ and $\overline{\alpha_t} = \prod_{i=1}^{t} \alpha_i$. To get the image at t timestep, a reparameterization trick should be applied to get x_t as a function of x_0:

$$x_t = \sqrt{\overline{\alpha_t} x_0} + \sqrt{1 - \overline{\alpha_t}\epsilon} \text{ ; where } \epsilon_t \sim \mathcal{N}(0, I)$$

3.2.4.2 Reverse Denoising Process

Since the image is noised now, we need to reverse the above process and sample from $q(x_{t-1}|x_t)$ to get back the true sample. But getting the true image is not possible because the steps are random. In order to denoise, we need to learn a model p_θ that approximates the conditional probability having θ as a parameter.

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

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

x_{t-1} can be predicted from x_t, using the method shown in [65],

$$x_{t-1} = \mathcal{N}\left(x_{t-1}; \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \overline{\alpha_t}}}\epsilon_\theta(x_t, t)\right), \Sigma_\theta(x_t, t)\right)$$

The parameter θ, variance is learnt by U-Net[22]. ϵ_θ is trained by U-Net, such that it predicts the noise in the image and subtract it from x_t during sampling scheme as shown in above equation.

We modified the idea of denoising procedure to perform image translation. While training we stack (concatenate) both the modalities, but the noising would be done in only provided target image. Our model infuses the anatomical feature from one modality during sampling inference, but the intensity and contrastive features are generated from the other modality.

Let (x, y) be the modality pair. We define X= xØy, where Ø denotes concatenate operation. Since the noise is added to the target modality, so over new modified input and prediction becomes:

$$y_t = \sqrt{\overline{\alpha_t} y_0} + \sqrt{1 - \overline{\alpha_t}\epsilon}$$

$$y_{t-1} = \mathcal{N}\left(y_{t-1}; \frac{1}{\sqrt{\alpha_t}}\left(y_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha_t}}}\epsilon_\theta(X_t, t) \right), \Sigma_\theta(y_t, t) \right)$$

The network performs a domain shift from one modality to another. But due to stochastic sampling, the model generates uncertain distribution, which may be dissimilar to the gold label. To overcome this, we performed sampling 5 times for each image. We compared the Fréchet Inception Distance (FID)**[**66] of each output with the original input image and selected the output with minimum FID. Implementation Details are provided in Appendix.

3.2.4.3 Loss Functions

Denoising Diffusion Probabilistic Model[65] uses two types of the loss function, Mean Squared Error and Variation Bound Loss:

3.2.4.3.1 Mean Squared Error (MSE) Loss

In diffusion models, the Mean Squared Error (MSE) loss is commonly used to measure the discrepancy between the reconstructed and original data. It quantifies the average squared difference between the generated samples and the target data. Here's the expression for the MSE loss in diffusion models:

Let's denote the generated samples by the diffusion model as $\hat{\boldsymbol{x}}$, and the target data (original data) as $\boldsymbol{x}$. The MSE loss is calculated as the average of the squared differences between the generated samples and the target data:

$$\mathcal{L}_{\mathcal{MSE}} = \frac{1}{N}\sum_{i=1}^{N}|x_i - \hat{x_i}|^2$$

Where N is the number of samples in the batch, x_i represents the i-th target data sample, and $\hat{x_i}$ represents the corresponding generated sample from the diffusion model.
The MSE loss's goal is to minimize the average squared difference between the generated samples and the target data during the training process. Minimizing this loss encourages the diffusion model to learn to reconstruct the data accurately.

3.2.4.3.2 Variation Bound (VB) Loss

The variation bound loss, including diffusion models, is commonly used in variational inference to approximate the intractable posterior distribution. It provides an upper bound on the negative log-likelihood of the data and is derived using the Kullback-Leibler (KL) divergence[67]. Here's an explanation of the variation bound loss along with its expression:

The variation bound loss is derived from the evidence lower bound (ELBO), which is a lower bound on the log-likelihood of the data. In diffusion models, the ELBO is often used to approximate the negative log-likelihood. The ELBO can be expressed as the sum of two terms: the reconstruction term and the KL divergence term.

1. Reconstruction term measures how well the model can reconstruct the data given a latent variable. In the context of diffusion models, it represents the likelihood term of the diffusion process.
2. KL divergence term quantifies the difference between the approximate posterior distribution and a prior distribution over the latent variables. It accounts for the divergence between the model distribution and the true data distribution in diffusion models.

The variation bound loss is obtained by taking the negative of the ELBO, resulting in an upper bound on the negative log likelihood. The expression for the variation bound loss can

be written as follows:

$$\mathcal{L}_{\mathcal{VB}} = -\text{ELBO} = E_{x\sim p_{\text{dt}}(x)}[\log q\,(x;\theta) - \log p_{\text{data}}\,(x)] + \text{KL}\big(q(x;\theta)|p(x)\big)$$

where $q(x;\theta)$ is the approximate posterior distribution, $p_{data}(x)$ is the true data

distribution, and p(x) is the prior distribution over the latent variables.

The variation bound loss consists of the reconstruction term (the first term) and the KL divergence term (the second term). Minimizing this loss encourages the model to find a balance between accurately reconstructing the data and aligning the approximate posterior with the prior distribution.

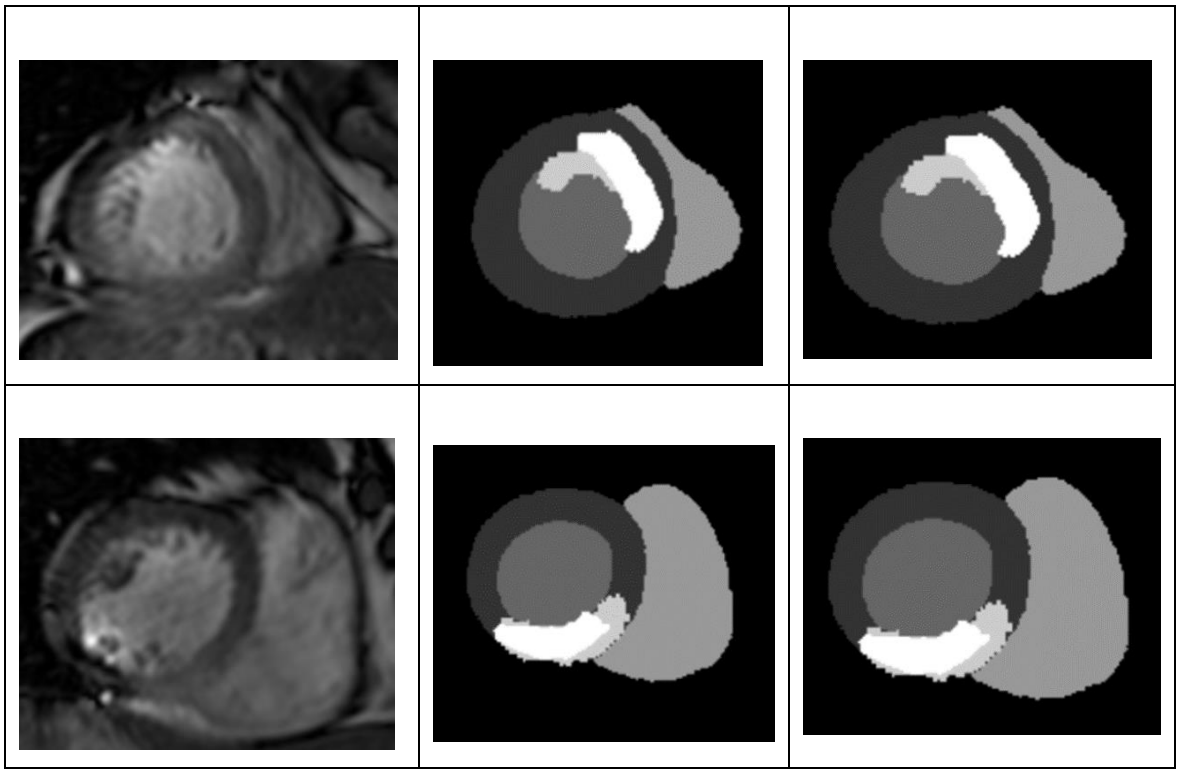
CHAPTER 4

Results and Discussion

4.1 MyoPS Dataset

4.1.1 Coarse-to-fine strategy results

Figure 13 depicts the results predicted by the coarse-to-fine model. The first row is patient 07, and the second is patient 23. The leftmost image denotes the input image, the middle image represents labelled data, and the rightmost is the predicted output. We can say that the model has performed well after seeing the predicted and ground truth image.



. *Figure 13: Patient 07, Patient 23 (Top to bottom) Input Image, Ground Truth, and Predicted Mask (Left to Right)*

Dice Coefficient	2D U-Net	Coarse to Fine Strategy
Edema	0.3880	**0.4124**
Scar	0.5926	**0.6296**
Average	0.4903	**0.521**

Table 3: 2D U-Net vs. Coarse-to-Fine Strategy Comparison

Table 3 compares vanilla 2D U-Net and Coarse to Fine Strategy. This shows the better performance of the proposed network.

The First Part of this study provides a coarse-to-fine framework for segmenting extremely tiny targets like cardiac edema and scars. There are two phases in this network. Because edema and scar are dispersed on the Myocardium, the first segmentation framework predicts five cardiac structures, including edema and scar, to derive the predictions of the region of interest. The fine segmentation framework stacks three sequences of CMR pictures and coarse network output prediction input to get a more accurate mask. The authors also presented a new weighted ensemble approach which provides 2.5D and 2D fine segmentation networks with a particular weight based on the model's performance in each class. The coarse-fine architecture performs admirably on the test set and may be used for other problems involving tiny target segmentation.

4.1.2 cGAN Results

Table 4 depicts conditional GAN models' results when different parameters are selected. cGAN, with a patch size 16*16, gave an average result of 0.74 on the training set. But the results were not good on the validation set. Figure 14 depicts the conditional GAN results for patient 55 and patient 26, respectively (training set). The model with patch size 16*16 shows the best results.

Model	Image Shape	Augmentation	Dice Score
cGAN with 16*16 Patch	256*256	No	Edema-**0.709781** Scars- 0.770868
cGAN with 70*70 Patch	256*256	No	Edema-0.54798 Scars-0.722137
cGAN with 13*13 Patch	208*208	No	Edema- 0.660125 Scars- **0.78237**
cGAN with 13*13 Patch	208*208	Horizontal Flip. Vertical Flip, Rotate (120)	Edema- 0.42723 Scars- 0.61324
Only Generator (Previous Model)	256*256	No	Edema- 0.4124 Scars- 0.6296

Table 4: Comparison between cGAN models with different settings

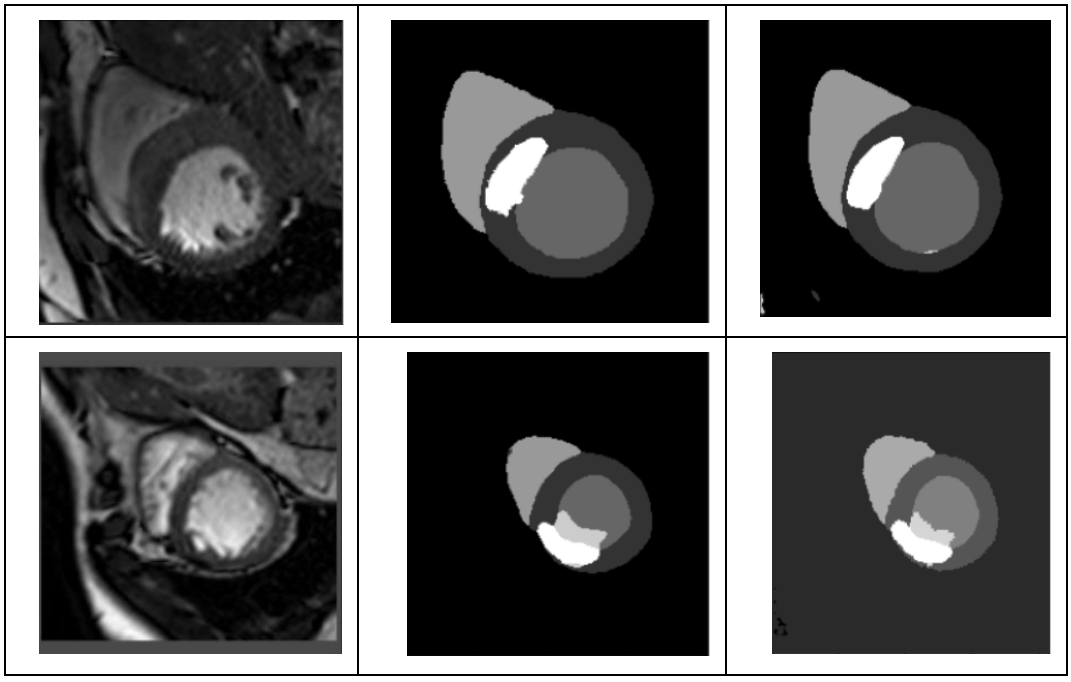


Figure 14: Top Row (Patient 55): ceT1 image, Ground Truth, Predictions (Left to Right); Bottom Row (Patient 26): ceT1 image, Ground Truth, Predictions (Left to Right)

In a cGAN, the generator aims to generate realistic output samples (segmentation masks) that can fool the discriminator while the discriminator tries to distinguish between real and generated samples. If the discriminator converges too quickly and becomes too strong, it may overpower the generator, making it difficult for the generator to produce high-quality segmentations.

In our case, the discriminator converges too quickly [69], which makes it very accurate in distinguishing real and generated samples early in the training process. This can lead to a situation known as "mode collapse," where the generator cannot explore the entire output space and produces limited variations of segmentations.

As a result, the generator might produce overly simplistic segmentations or fail to capture the complexity of the validation dataset. This can lead to poor performance on the validation dataset, as the generator cannot generate diverse and accurate segmentations.

4.2 crossMoDA dataset

4.2.1 Translation results using CycleGAN

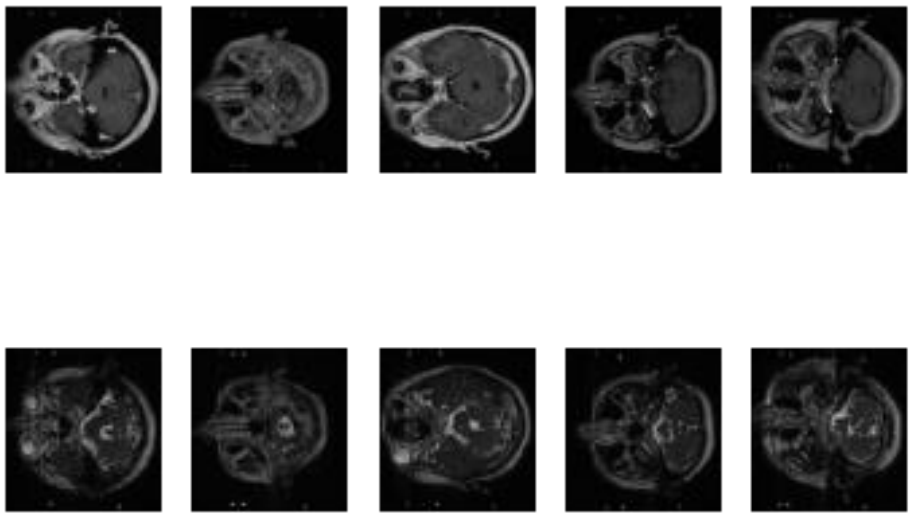


Figure 15: Top row contains the original ceT1 image; the bottom row includes the translated ceT1 to hrT2 image

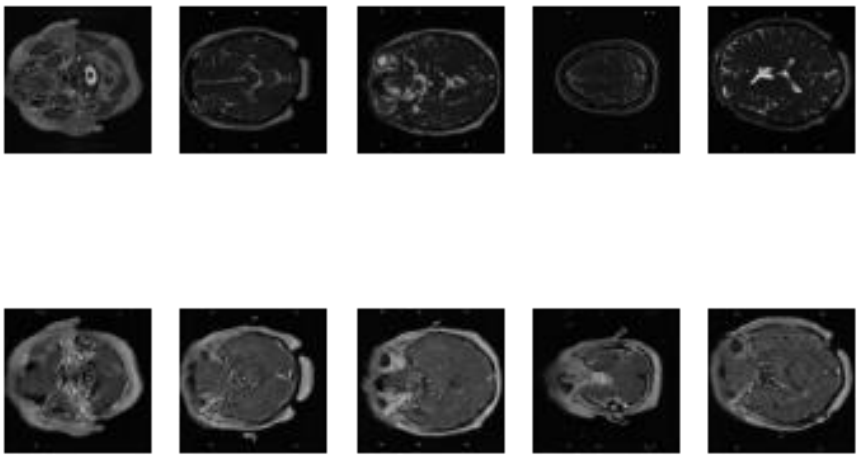


Figure 16: Top row contains the original hrT2 image, and the bottom row includes the translated hrT2 to ceT1 image

After feeding the original images ceT1 to the first generator of CycleGAN, represented by the first row of Figure 15, the model generated the pseudo hrT2 outputs, represented by the bottom row of Figure 15. Similarly, when original hrT2 images were fed to the second generator of CycleGAN, the results were pseudo ceT1, shown in the bottom row of Figure 16. Between the fake and actual scans from the same Source, we can see how consistent overall image contrast qualities remain.

4.2.2 Comparison between UNETR and nnU-Net results (Original Data)

Model	Patch Size	Batch Size	Iterations	Dice Score
UNETR	96*96*16	50	800	**0.8412**
nnU-Net (3D)	128*128*16	2	1000	0.7821

Table 5: Results and parameters comparison between UNETR and nnU-Net

We trained and tested the model on the original ceT1 annotated data for both models, and the results can be represented in Table 5. UNETR outperformed 3D full-res nnU-Net on the validation set. Analyzing the original scans' validation data results, we went ahead with the

UNETR model to train and test on translated output from the CycleGAN. Predictions and ground truths of two of the Tilburg patients can be shown in Figure 17.

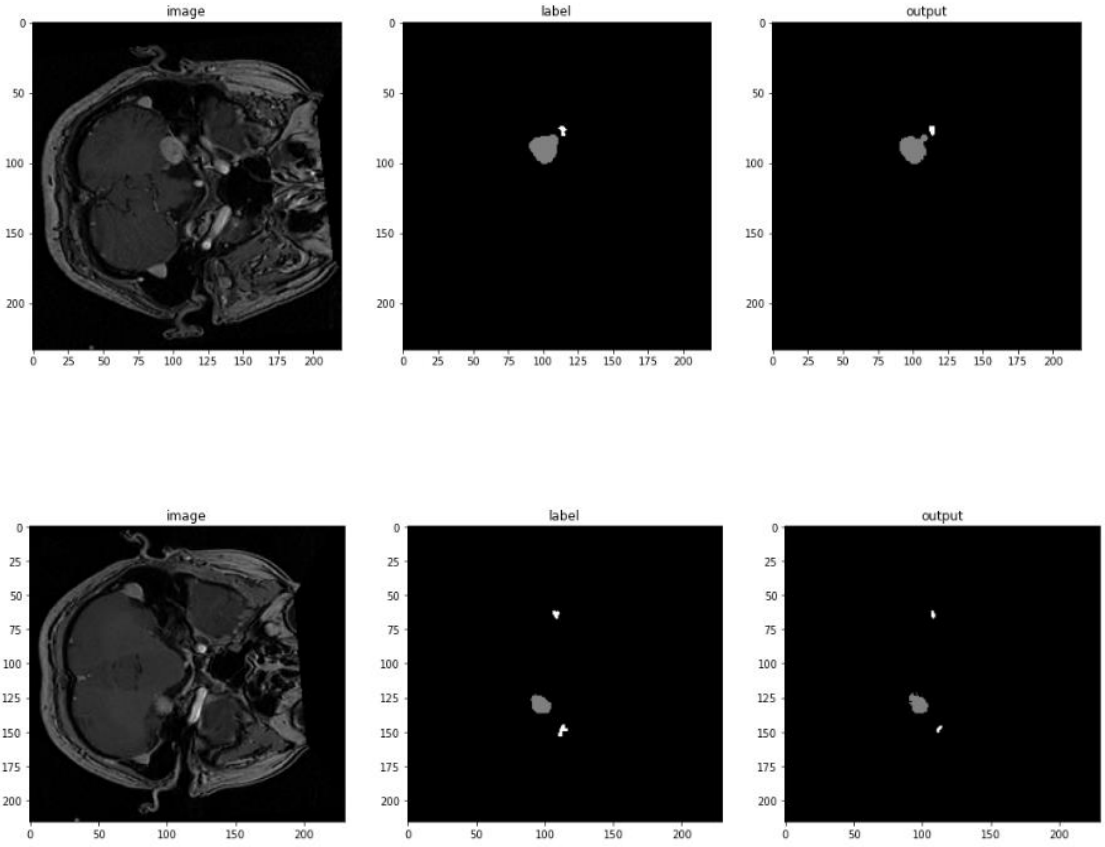
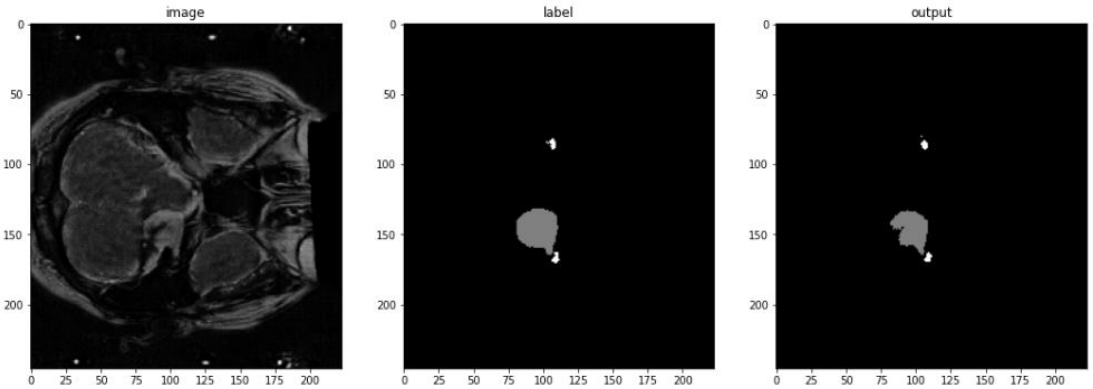


Figure 17: Top Row (Patient 16 Tilburg): ceT1 image, Ground Truth, Predictions (Left to Right); Bottom Row (Patient 92 Tilburg): ceT1 image, Ground Truth, Predictions (Left to Right)

4.2.3 Segmentation results on translated data using UNETR



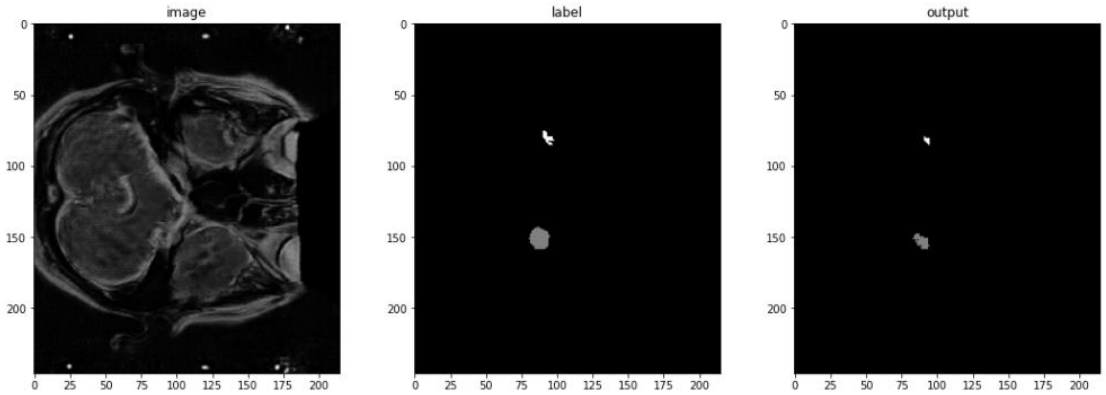


Figure 18: Top Row (Patient 76 Tilburg): pseudo hrT2 image, Ground Truth, Predictions (Left to Right); Bottom Row (Patient 90 Tilburg): pseudo hrT2 image, Ground Truth, Predictions (Left to Right)

Going with the original results of original ceT1 images, as U-Entry performed better than nnU-Net, we have trained the translated image with the UNETR model. The average dice score came out to be 0.75 compared to our original dataset trained UNETR model, 0.865. The main reason for the reduction of 0.115 dice score between the actual and translated segmentation results is improper domain adaptation. Although CycleGAN has performed very well on some datasets but failed to give good results in the CrossMoDA dataset. As shown in Figure 18, the middle image of each row is the actual label, and the rightmost images are the predictions. A deformity in the shape of VS in the translated image (left most images) accounts for false negative rates in the images (i.e., some parts of labels are missing in the prediction). This is the main reason for getting a lower dice score.

The bottleneck part is the translation of ceT1 to hrT2 scans, which, when improved, leads to good results on the segmentation part. GANs have some disadvantages for image generation and translation[68], which are listed below:

1. Mode collapse: One of the significant challenges with GANs is mode collapse. It occurs when the generator fails to capture the target distribution's full diversity and produces limited or repetitive outputs. This can result in generated samples that lack diversity and fail to represent the full complexity of the underlying data distribution.

2. Training instability: GANs can be challenging to train due to their adversarial nature. The training process involves a delicate balance between the generator and

discriminator networks. Achieving convergence and stability during training can be difficult, and GANs are known to be sensitive to hyperparameter choices, network architectures, and optimization settings. In some cases, GANs may suffer from issues like mode dropping, where specific modes of data distribution are not well-represented in the generated samples.

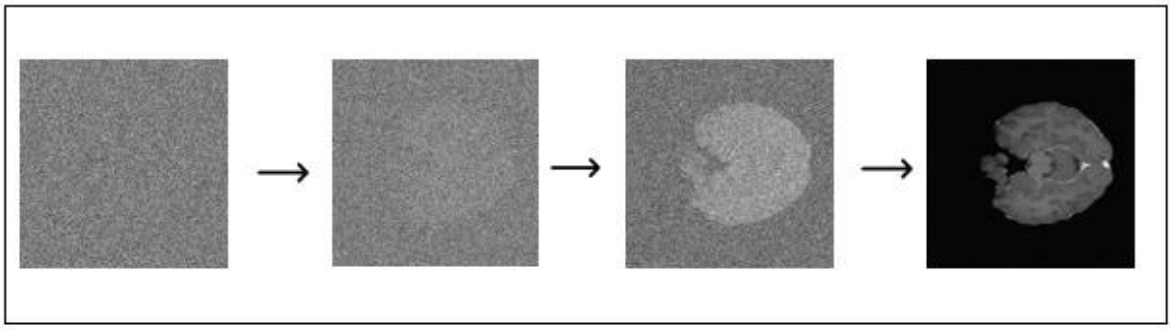
3. Lack of explicit likelihood: Unlike diffusion models, GANs do not directly model the likelihood of the data distribution. Instead, they focus on learning a mapping from noise to data samples through the adversarial training process. The lack of an explicit likelihood estimation can make it challenging to reason about uncertainty or perform tasks requiring access to likelihood information.

4. Difficulty in controlling generation: GANs can struggle with precise control over the generated outputs. While recent advances, such as conditional GANs, have enabled some level of control by conditioning the generator on specific inputs, achieving fine-grained control over attributes or features of the generated samples can be challenging and need to be addressed by some stable architecture or improvements in the current model.

4.3 Image Translation results using Diffusion Models

The model was trained on FLAIR and T1CE pair as discussed above. The noising-denoising process was performed on T1CE modality while anatomical features of FLAIR was infused in the noisy samples, to retain the structure of FLAIR but image intensity, distribution of T1CE. Figure 20 shows the visualization of predictions of 4 cases from diffusion model. In each row, left image is the FLAIR, middle is T1CE and right one is the output predictions. It is quite impressive to see the model has done pretty good job in translation, which can't distinguish with bare eyes. Since the process is stochastic, some of the outputs where not at all good. So we iterated the sampling procedure 5 times for the same image, computed the FID score, and the predictions with the minimum FID score are been selected and shown.

Figure 19 shows the denoising process of case 21, 52nd slice. Diffusion steps was set to 1000, so 1000 intermediaries' samples would have been generated. But for the sake of visualization, we have selected sample from each 100th step i.e., 1000, 900...0. In this figure, the snapshots are taken at 1000, 600, 300 and 0 steps respectively. The steps are in reducing fashion because we are moving in reverse way. The figure clearly shows how noise have been subtracted from the model (denoising process) to get the final refined translated output.



. Figure 19: Denoising process of T1CE image at 1000, 600, 300 and 0 steps respectively

Fréchet Inception distance of each of the case are being computed using pre-trained VGG-16 model [12] and shown in Table 6. Original FLAIR and translated output were given as an input to the model, where it compares the image distribution at latent dimension as transformed by convolution layers. The minimum the value, the more output would be better.

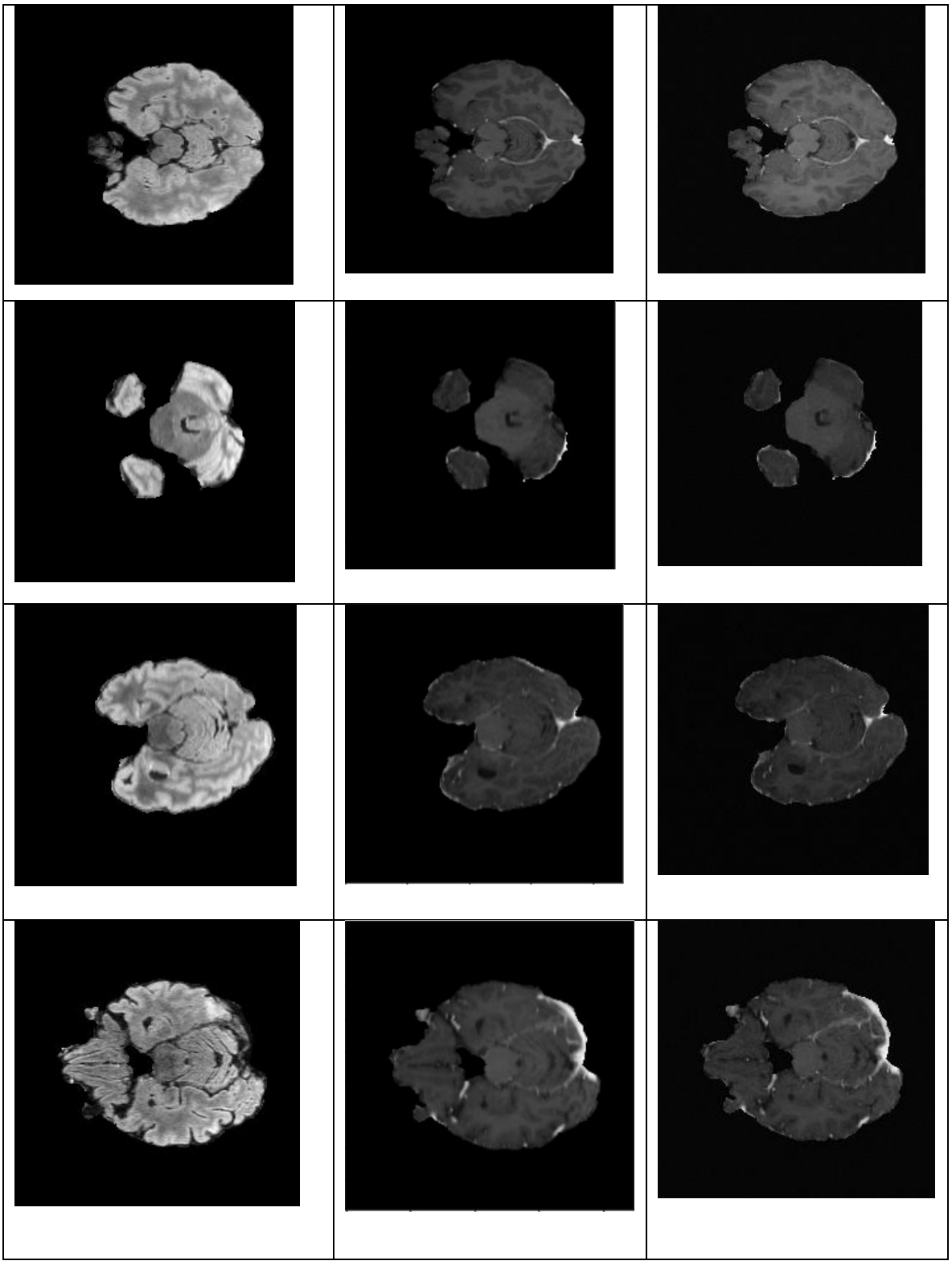


Figure 20: FLAIR, T1CE and Translated Output (Left to Right) (Case 21 slice 52, Case 25 slice 34, Case 25 slice 46, Case 29 slice 46 (Top to Bottom)

Image ID	FID (1e-5)
Case 21, Slice 52	9.89
Case 25, Slice 34	5.13
Case 25, Slice 46	4.12
Case 29, Slice 46	4.16

Table 6: Fréchet Inception Distance between FLAIR and Predictions

Figure 21 shows the Mean Squared Error of the actual sample's vs the reconstructed samples with respect to steps. There is a sudden decrease in loss value at 200 steps, and the loss remains more or less same after that. One of the reasons is because the images were normalized, so the optimizer must find the global minima quickly.

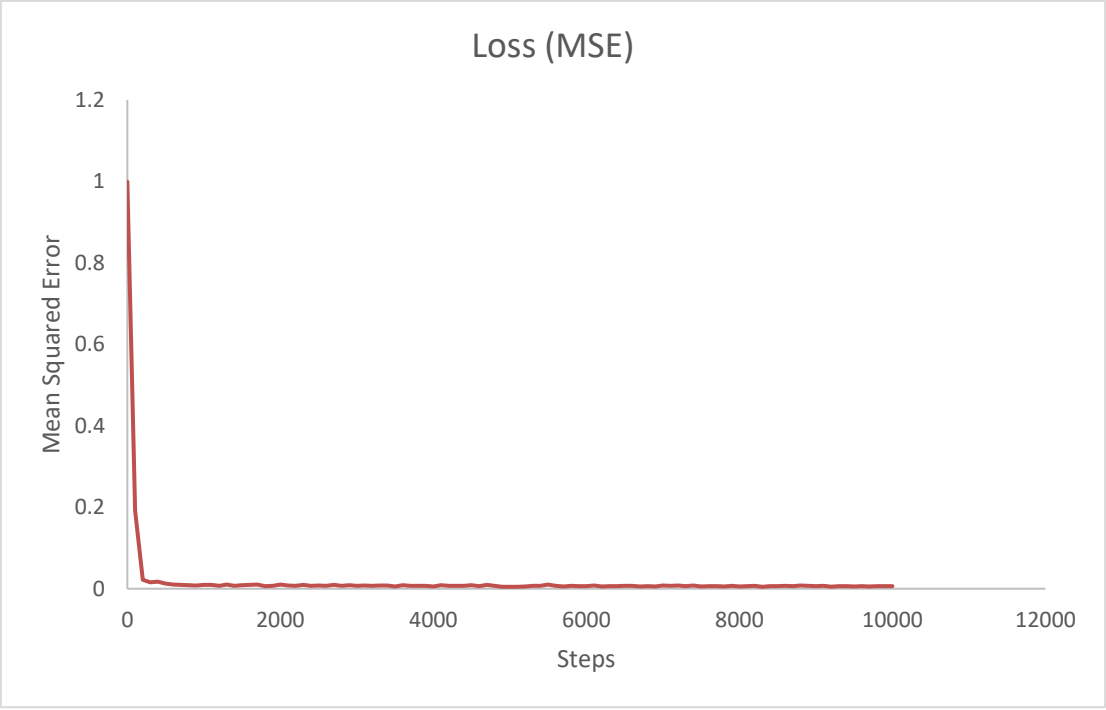


Figure 21: Mean Squared Error of the training data w.r.t steps

Figure 22 depicts the mean variation bound loss between the reconstructed and target image.

It basically measures the KL divergence between the two. VBL has decreased but the values keep on fluctuating. It can be due to small batch size (4). When batch size is closer to 1, the parameters of the model are updated frequently, hence there are variations in predictions as well as loss. But in the long run, it converges.

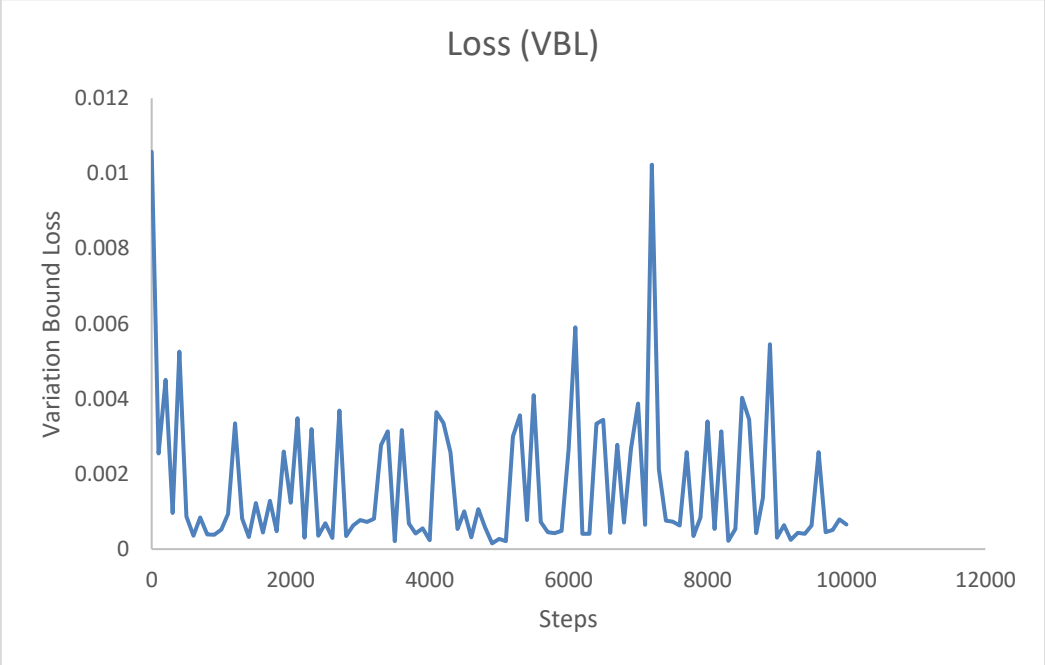


Figure 22: Variation Bound Loss of the training data w.r.t steps

CHAPTER 5

Conclusion and Further Work

In conclusion, multi-organ image segmentation plays a pivotal role in accurately diagnosing and treating cardiovascular and brain diseases. Accurate segmentation of organs and tissues is crucial in medical imaging analysis as it provides valuable insights into the underlying pathologies and helps guide appropriate treatment strategies. The advent of deep learning techniques, particularly CNNs, has revolutionized the field of medical image analysis, enabling more precise and efficient segmentation of organs and tissues than traditional methods.

CNNs have performed exceptionally in various computer vision tasks, including medical image analysis. These networks can automatically learn and extract complex features from medical images, enabling them to capture intricate patterns and structures within the data. By utilizing large datasets and powerful computational resources, CNNs can be trained to recognize and segment-specific organs or tissues with high accuracy. This capability has dramatically improved the efficiency and reliability of medical image analysis, facilitating quicker and more accurate diagnosis and treatment planning.

Integrating diverse imaging modalities has further enhanced the comprehensive evaluation of diseases. In medical imaging, different modalities such as magnetic resonance imaging (MRI), computed tomography (CT), and positron emission tomography (PET) provide unique and complementary information about the anatomy, function, and metabolism of organs. By combining the data from multiple modalities, clinicians and researchers can obtain a more comprehensive understanding of diseases, improving diagnostic accuracy and better treatment outcomes. Multi-organ image segmentation techniques enable extracting and analysing organ-specific information from these multimodal datasets, facilitating a more holistic approach to disease evaluation.

However, one of the challenges in applying CNNs to medical image analysis is the domain shift problem. Medical imaging datasets are often collected from different sources, leading to

imaging protocols, equipment, and patient populations variations. This variability can negatively impact the performance of CNN models, as they may fail to generalize well when applied to unseen data from different domains. To address this issue, domain adaptation techniques are essential to ensure the generalizability of CNN models across other imaging domains. These techniques aim to bridge the gap between the source domain (where the model is trained) and the target domain (where the model is deployed) by adapting the model to the target domain's characteristics.

In the subsequent sections of this thesis, we delve into the methodologies, experiments, and results of multi-organ image segmentation. This research focuses on CNN-based approaches, which have shown great promise in this field. We explore various CNN architectures and their application to different medical imaging tasks, highlighting their strengths and limitations. Additionally, we investigate domain adaptation techniques specifically designed for medical image analysis, aiming to improve the generalizability of CNN models. Our experiments involve training and evaluating the CNN models on diverse datasets, encompassing different imaging modalities and patient populations.

The thesis further explores the application of various segmentation algorithms and translation networks in multi-organ image segmentation. These algorithms and networks, such as nnU-Net, UNETR, conditional GAN, pix2pix, and cycleGAN, offer different approaches and techniques to address the challenges of accurately segmenting organs and tissues from medical images.

One of the segmentation algorithms utilized in this thesis is nnU-Net. nnU-Net is a neural network-based architecture that leverages the power of deep learning for image segmentation. By employing a U-Net architecture consisting of an encoder-decoder framework with skip connections, nnU-Net has shown promising results in segmenting organs and tissues with high precision and recall. The utilization of nnU-Net in this thesis contributes to the exploration of its effectiveness in multi-organ image segmentation tasks.

Another segmentation algorithm explored is UNETR. UNETR is an extension of the U-Net architecture that incorporates the concept of transformers. Transformers are self-attention mechanisms that capture long-range dependencies in the data, making them particularly suitable for capturing contextual information in medical images. By combining the U-Net

architecture with transformers, UNETR has demonstrated improved segmentation performance in various medical imaging tasks. The integration of UNETR in this thesis allows for evaluating its potential in multi-organ image segmentation.

This thesis has also employed conditional generative adversarial networks (GANs) for image-to-image translation tasks. The conditional GAN framework enables the generation of realistic and high-quality synthetic images guided by input conditions. By training the GAN on paired datasets of input and target images, it learns to transform images from one domain to another while preserving important anatomical details and structures. In the context of multi-organ image segmentation, conditional GANs offer a means to generate synthetic images from different imaging modalities or domains, facilitating the development of domain adaptation techniques.

Translation networks such as pix2pix and cycleGAN have been incorporated into this thesis to explore their potential in image translation tasks related to multi-organ image segmentation. Pix2pix is a conditional GAN that has been designed explicitly for image-to-image translation. It learns a mapping between input images and their corresponding output images through adversarial training. CycleGAN, on the other hand, introduces cycle consistency loss, enabling unpaired image translation. It can learn translations between two domains without requiring direct correspondence between individual images. By leveraging these translation networks, this thesis investigates the effectiveness of image translation techniques in addressing domain shift issues and improving the generalizability of segmentation models.

Some limitations with GANs, such as mode collapse, training instability, etc., need to be addressed by improved generative models. In addition to the aforementioned algorithms and networks, The conditional diffusion model, developed as part of this thesis, focuses on translating medical images from one domain to another while preserving important anatomical details and structures. The model generates a sequence of intermediate images conditioned on one modality by utilising diffusion processes. The conditional diffusion model aims to achieve accurate and high-quality image translation through the gradual refinement of these intermediate images.

While the conditional diffusion model has been developed as part of this thesis, it is essential to acknowledge that further enhancements are needed to make it more robust and to improve its translation performance. The thesis identifies areas requiring additional research and development, such as we aim to make it more robust to different modalities. For example, when the dataset of MRI and CT are provided, using class conditional embeddings 4 translations would be done MRI-CT, CT-MRI, MRI-MRI and CT-CT by just passing the name convention and preparation of the dataset. These enhancements aim to strengthen the performance and applicability of the conditional diffusion model for practical medical imaging scenarios.

Overall, this thesis investigates and explores a range of segmentation algorithms, including nnUNET, UNETR, conditional GANs, and translation networks such as pix2pix and CycleGAN. Additionally, it presents the development of a conditional diffusion model, highlighting its potential while acknowledging the need for further improvement. Through thoroughly examining these algorithms and models, the thesis aims to contribute to advancing multi-organ image segmentation techniques, ultimately improving the accuracy and reliability of diagnosis and treatment planning for cardiovascular and brain diseases.

REFERENCES

[1] "Cardiovascular diseases (CVDs)." https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (accessed Jun. 06, 2023).

[2] V. L. Feigin *et al.*, "Global, regional, and national burden of neurological disorders during 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015," *The Lancet Neurology*, vol. 16, no. 11, pp. 877–897, Nov. 2017, doi: 10.1016/S1474-4422(17)30299-5.

[3] X. Liu, Z. Deng, and Y. Yang, "Recent progress in semantic image segmentation," *Artif Intell Rev*, vol. 52, no. 2, pp. 1089–1106, Aug. 2019, doi: 10.1007/s10462-018-9641-3.

[4] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia, "A Brief Review of Domain Adaptation." arXiv, Oct. 07, 2020. doi: 10.48550/arXiv.2010.03978.

[5] C. Chen, Q. Dou, H. Chen, J. Qin, and P.-A. Heng, "Synergistic Image and Feature Adaptation: Towards Cross-Modality Domain Adaptation for Medical Image Segmentation." arXiv, Jun. 18, 2019. doi: 10.48550/arXiv.1901.08211.

[6] J. Qiu *et al.*, "MyoPS-Net: Myocardial pathology segmentation with flexible combination of multi-sequence CMR images," *Medical Image Analysis*, vol. 84, p. 102694, Feb. 2023, doi: 10.1016/j.media.2022.102694.

[7] S. Zhai, R. Gu, W. Lei, and G. Wang, "Myocardial Edema and Scar Segmentation Using a Coarse-to-Fine Framework with Weighted Ensemble," in *Myocardial Pathology Segmentation Combining Multi-Sequence Cardiac Magnetic Resonance Images*, X. Zhuang and L. Li, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 49–59. doi: 10.1007/978-3-030-65651-5_5.

[8] R. Dorent *et al.*, "CrossMoDA 2021 challenge: Benchmark of cross-modality domain adaptation techniques for vestibular schwannoma and cochlea segmentation," *Medical Image Analysis*, vol. 83, p. 102628, Jan. 2023, doi: 10.1016/j.media.2022.102628.

[9] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep Unsupervised Learning using Nonequilibrium Thermodynamics," 2015, doi: 10.48550/ARXIV.1503.03585.

[10] U. Baid *et al.*, "The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification." arXiv, Sep. 12, 2021. Accessed: Jun. 06, 2023. [Online]. Available: http://arxiv.org/abs/2107.02314

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[12] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," 2014, doi: 10.48550/ARXIV.1409.1556.

[13] C. Szegedy *et al.*, "Going Deeper with Convolutions," 2014, doi: 10.48550/ARXIV.1409.4842.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," 2015, doi: 10.48550/ARXIV.1512.03385.

[15] N. Baron *et al.*, "Quantification of myocardial edema and necrosis during acute myocardial infarction," in *2008 Computers in Cardiology*, Bologna, Italy: IEEE, Sep. 2008, pp. 781–784. doi: 10.1109/CIC.2008.4749158.

[16] R. Karim *et al.*, "Evaluation of state-of-the-art segmentation algorithms for left ventricle infarct from late Gadolinium enhancement MR images," *Medical Image Analysis*, vol. 30, pp. 95–107, May 2016, doi: 10.1016/j.media.2016.01.004.

[17] V. Sandfort *et al.*, "Automatic high-resolution infarct detection using volumetric multiphase dual-energy CT," *Journal of Cardiovascular Computed Tomography*, vol. 11, no. 4, pp. 288–294, Jul. 2017, doi: 10.1016/j.jcct.2017.04.006.

[18] H. Gao, K. Kadir, A. R. Payne, J. Soraghan, and C. Berry, "Highly automatic quantification of myocardial oedema in patients with acute myocardial infarction using bright blood T2-weighted CMR," *J Cardiovasc Magn Reson*, vol. 15, no. 1, p. 28, Dec. 2013, doi: 10.1186/1532-429X-15-28.

[19] S. Moccia *et al.*, "Development and testing of a deep learning-based strategy for scar segmentation on CMR-LGE images," *Magn Reson Mater Phy*, vol. 32, no. 2, pp. 187–195, Apr. 2019, doi: 10.1007/s10334-018-0718-4.

[20] Y. Lu, Y. Yang, K. A. Connelly, G. A. Wright, and P. E. Radau, "Automated quantification of myocardial infarction using graph cuts on contrast delayed enhanced magnetic resonance images," *Quantitative Imaging in Medicine and Surgery*, vol. 2, no. 2, pp. 816–886, Jun. 2012, doi: 10.3978/j.issn.2223-4292.2012.05.03.

[21] L. Li, V. A. Zimmer, J. A. Schnabel, and X. Zhuang, "Medical Image Analysis on Left Atrial LGE MRI for Atrial Fibrillation Studies: A Review," 2021, doi: 10.48550/ARXIV.2106.09862.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," 2015, doi: 10.48550/ARXIV.1505.04597.

[23] W. Li, L. Wang, and S. Qin, "CMS-UNet: Cardiac Multi-task Segmentation in MRI with a U-Shaped Network," in *Myocardial Pathology Segmentation Combining Multi-Sequence Cardiac Magnetic Resonance Images*, X. Zhuang and L. Li, Eds., in Lecture Notes in Computer Science, vol. 12554. Cham: Springer International Publishing, 2020, pp. 92–101. doi: 10.1007/978-3-030-65651-5_9.

[24] F. Zabihollahy, J. A. White, and E. Ukwatta, "Convolutional neural network-based approach for segmentation of left ventricle myocardial scar from 3D late gadolinium enhancement MR images," *Med. Phys.*, vol. 46, no. 4, pp. 1740–1751, Apr. 2019, doi: 10.1002/mp.13436.

[25] C. Martín-Isla, M. Asadi-Aghbolaghi, P. Gkontra, V. M. Campello, S. Escalera, and K. Lekadir, "Stacked BCDU-Net with Semantic CMR Synthesis: Application to Myocardial Pathology Segmentation Challenge," in *Myocardial Pathology Segmentation Combining Multi-Sequence Cardiac Magnetic Resonance Images*, X. Zhuang and L. Li, Eds., in Lecture Notes in Computer Science, vol. 12554. Cham: Springer International Publishing, 2020, pp. 1–16. doi: 10.1007/978-3-030-65651-5_1.

[26] R. Azad, M. Asadi-Aghbolaghi, M. Fathy, and S. Escalera, "Bi-Directional ConvLSTM U-Net with Densley Connected Convolutions," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, Seoul, Korea (South): IEEE, Oct. 2019, pp. 406–415. doi: 10.1109/ICCVW.2019.00052.

[27] J. Zhang, Y. Xie, Z. Liao, J. Verjans, and Y. Xia, "EfficientSeg: A Simple But Efficient Solution to Myocardial Pathology Segmentation Challenge," in *Myocardial Pathology Segmentation Combining Multi-Sequence Cardiac Magnetic Resonance Images*, X. Zhuang and L. Li, Eds., in Lecture Notes in Computer Science, vol. 12554. Cham: Springer International Publishing, 2020, pp. 17–25. doi: 10.1007/978-3-030-65651-5_2.

[28] M. J. Ankenbrand, D. Lohr, and L. M. Schreiber, "Exploring Ensemble Applications for Multi-sequence Myocardial Pathology Segmentation," in *Myocardial Pathology Segmentation Combining Multi-Sequence Cardiac Magnetic Resonance Images*, X. Zhuang and L. Li, Eds., in Lecture Notes in Computer Science, vol. 12554. Cham: Springer International Publishing, 2020, pp. 60–67. doi: 10.1007/978-3-030-65651-5_6.

[29] Z. Zhang *et al.*, "Multi-Modality Pathology Segmentation Framework: Application to Cardiac Magnetic Resonance Images," 2020, doi: 10.48550/ARXIV.2008.05780.

[30] H. Jiang, C. Wang, A. Chartsias, and S. A. Tsaftaris, "Max-Fusion U-Net for Multimodal Pathology Segmentation with Attention and Dynamic Resampling," in *Myocardial Pathology Segmentation Combining Multi-Sequence Cardiac Magnetic Resonance Images*, X. Zhuang and L. Li, Eds., in Lecture Notes in Computer Science, vol. 12554. Cham: Springer International Publishing, 2020, pp. 68–81. doi: 10.1007/978-3-030-65651-5_7.

[31] A. Myronenko, "3D MRI brain tumor segmentation using autoencoder regularization," 2018, doi: 10.48550/ARXIV.1810.11654.

[32] S. Bakas *et al.*, "Segmentation Labels for the Pre-operative Scans of the TCGA-GBM collection." The Cancer Imaging Archive, 2017. doi: 10.7937/K9/TCIA.2017.KLXWJJ1Q.

[33] B. H. Menze *et al.*, "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," *IEEE Trans. Med. Imaging*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015, doi: 10.1109/TMI.2014.2377694.

[34] Z. Jiang, C. Ding, M. Liu, and D. Tao, "Two-Stage Cascaded U-Net: 1st Place Solution to BraTS Challenge 2019 Segmentation Task," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi and S. Bakas, Eds., in Lecture Notes in Computer Science, vol. 11992. Cham: Springer International Publishing, 2020, pp. 231–241. doi: 10.1007/978-3-030-46640-4_22.

[35] Y. Wang *et al.*, "Modality-Pairing Learning for Brain Tumor Segmentation," 2020, doi: 10.48550/ARXIV.2010.09277.

[36] S. Bakas *et al.*, "Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features," *Sci Data*, vol. 4, no. 1, p. 170117, Sep. 2017, doi: 10.1038/sdata.2017.117.

[37] K. Clark *et al.*, "The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository," *J Digit Imaging*, vol. 26, no. 6, pp. 1045–1057, Dec. 2013, doi: 10.1007/s10278-013-9622-7.

[38] M. D. Cirillo, D. Abramian, and A. Eklund, "Vox2Vox: 3D-GAN for Brain Tumour Segmentation," 2020, doi: 10.48550/ARXIV.2003.13653.

[39] H. Shin, H. Kim, S. Kim, Y. Jun, T. Eo, and D. Hwang, "COSMOS: Cross-Modality Unsupervised Domain Adaptation for 3D Medical Image Segmentation based on Target-aware Domain Translation and Iterative Self-Training," 2022, doi: 10.48550/ARXIV.2203.16557.

[40] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," 2017, doi: 10.48550/ARXIV.1703.10593.

[41] H. Dong, F. Yu, J. Zhao, B. Dong, and L. Zhang, "Unsupervised Domain Adaptation in Semantic Segmentation Based on Pixel Alignment and Self-Training." arXiv, Sep. 29, 2021. Accessed: Jun. 06, 2023. [Online]. Available: http://arxiv.org/abs/2109.14219

[42] R. Chen, W. Huang, B. Huang, F. Sun, and B. Fang, "Reusing Discriminators for Encoding: Towards Unsupervised Image-to-Image Translation," 2020, doi: 10.48550/ARXIV.2003.00273.

[43] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nat Methods*, vol. 18, no. 2, pp. 203–211, Feb. 2021, doi: 10.1038/s41592-020-01008-z.

[44] J. W. Choi, "Using Out-of-the-Box Frameworks for Contrastive Unpaired Image Translation for Vestibular Schwannoma and Cochlea Segmentation: An approach for the crossMoDA Challenge." arXiv, Dec. 07, 2021. Accessed: Jun. 06, 2023. [Online]. Available: http://arxiv.org/abs/2110.01607

[45] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive Learning for Unpaired Image-to-Image Translation." arXiv, Aug. 20, 2020. Accessed: Jun. 06, 2023. [Online]. Available: http://arxiv.org/abs/2007.15651

[46] K. Yao *et al.*, "A Novel 3D Unsupervised Domain Adaptation Framework for Cross-Modality Medical Image Segmentation," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 10, pp. 4976–4986, Oct. 2022, doi: 10.1109/JBHI.2022.3162118.

[47] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 94–114, Apr. 2020, doi: 10.1016/j.isprsjprs.2020.01.013.

[48] G. E. Hinton, P. Dayan, B. J. Frey, and R. M. Neal, "The 'Wake-Sleep' Algorithm for Unsupervised Neural Networks," *Science*, vol. 268, no. 5214, pp. 1158–1161, May 1995, doi: 10.1126/science.7761831.

[49] C. Sminchisescu, A. Kanaujia, and D. Metaxas, "Learning Joint Top-Down and Bottom-up Processes for 3D Visual Inference," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06)*, New York, NY, USA: IEEE, 2006, pp. 1743–1752. doi: 10.1109/CVPR.2006.169.

[50] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," 2013, doi: 10.48550/ARXIV.1312.6114.

[51] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised Cross-Domain Image Generation," 2016, doi: 10.48550/ARXIV.1611.02200.

[52] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation," 2017, doi: 10.48550/ARXIV.1711.09020.

[53] M.-Y. Liu *et al.*, "Few-Shot Unsupervised Image-to-Image Translation," 2019, doi: 10.48550/ARXIV.1905.01723.

[54] A. Nichol and P. Dhariwal, "Improved Denoising Diffusion Probabilistic Models," 2021, doi: 10.48550/ARXIV.2102.09672.

[55] J. Song, C. Meng, and S. Ermon, "Denoising Diffusion Implicit Models," 2020, doi: 10.48550/ARXIV.2010.02502.

[56] C. Saharia *et al.*, "Palette: Image-to-Image Diffusion Models," 2021, doi: 10.48550/ARXIV.2111.05826.

[57] H. Sasaki, C. G. Willcocks, and T. P. Breckon, "UNIT-DDPM: UNpaired Image Translation with Denoising Diffusion Probabilistic Models," 2021, doi: 10.48550/ARXIV.2104.05358.

[58] M. J. Ankenbrand, "Models from: Exploring ensemble applications for multi-sequence myocardial pathology segmentation." Zenodo, Aug. 14, 2020. doi: 10.5281/ZENODO.3985837.

[59] J. Shapey *et al.*, "Segmentation of Vestibular Schwannoma from Magnetic Resonance Imaging: An Open Annotated Dataset and Baseline Algorithm." The Cancer Imaging Archive, 2021. doi: 10.7937/TCIA.9YTJ-5Q73.

[60] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Applied Soft Computing*, vol. 97, p. 105524, Dec. 2020, doi: 10.1016/j.asoc.2019.105524.

[61] W. Huang, G. Song, M. Li, W. Hu, and K. Xie, "Adaptive Weight Optimization for Classification of Imbalanced Data," in *Intelligence Science and Big Data Engineering*, C. Sun, F. Fang, Z.-H. Zhou, W. Yang, and Z.-Y. Liu, Eds., in Lecture Notes in Computer Science, vol. 8261. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 546–553. doi: 10.1007/978-3-642-42057-3_69.

[62] M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," 2014, doi: 10.48550/ARXIV.1411.1784.

[63] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," 2016, doi: 10.48550/ARXIV.1611.07004.

[64] A. Hatamizadeh *et al.*, "UNETR: Transformers for 3D Medical Image Segmentation," 2021, doi: 10.48550/ARXIV.2103.10504.

[65] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," 2020, doi: 10.48550/ARXIV.2006.11239.

[66] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium." arXiv, Jan. 12, 2018. Accessed: Jun. 06, 2023. [Online]. Available: http://arxiv.org/abs/1706.08500

[67] M. Günder, N. Piatkowski, and C. Bauckhage, “Full Kullback-Leibler-Divergence Loss for Hyperparameter-free Label Distribution Learning,” 2022, doi: 10.48550/ARXIV.2209.02055.

[68] P. Dhariwal and A. Nichol, "Diffusion Models Beat GANs on Image Synthesis." arXiv, Jun. 01, 2021. Accessed: Jun. 06, 2023. [Online]. Available: http://arxiv.org/abs/2105.05233

APPENDIX

Training and Implementation Details

UNETR

UNETR Architecture was trained on PyTorch Environment. The system contains 24 GB NVIDIA A5000 chipset with 32 GB RAM. Batch size and optimizer was set to 6 and AdamW. The model was trained for 90000 iterations with a learning rate of 0. 0001.The data was split into 80:20 training and test set respectively with a patch size of 16 x 16 x 16. Random rotations of 90, 180 and 270 degrees, random shift and scale intensity and random flip in axial dimension were incorporated as a data augmentation technique to increase the dataset.

GANs (CycleGAN and conditional GAN)

Both the GAN networks were trained on TensorFlow Environment. The training was performed on Google Colab with Tesla V100 GPU (16 GB) and 24 GB RAM. The training was done for 100 epochs with a batch size of 4. The learning rate was set to 0.0002. Gaussian Distribution was used for weights initialization. Loss tradeoff value, λ was set to 10.

Diffusion Model

Diffusion Model was trained on PyTorch Environment on the system having 24 GB NVIDIA A5000 chipset with 32 GB RAM. Variance Beta Scheduler was chosen as linear with timestep T=1000 steps. The batch size was taken as 4, learning rate as 0.0001, optimizer as Adam and the model was trained for 100000 iteration steps. The architecture was set to use only one attention head at 16 resolution and first layer contains 128 channels. Apart from that all the hyperparameters set are provided in Appendix of [54].