

University of Nottingham
UK | CHINA | MALAYSIA

Improvement of Model Performance and Generalisability of Deep Convolutional Neural Networks for Medical Image Segmentation

Submitted March 2022, in partial fulfillment of the conditions for the award of the degree **Doctor of Philosophy.**

Mina Jafari
14300229

Supervised by
Xin Chen
Jonathan M. Garibaldi
Susan Francis

School of Computer Science
University of Nottingham

I hereby declare that this dissertation is all my own work, except as indicated in the text:

Signature _____

Date _____ / _____ / _____

To my Mom and my sisters.

Abstract

Automatic and reliable medical image segmentation to assist clinicians in disease diagnosis and treatment planning is highly demanded in clinical practice. Recently, by the advent of deep learning, specifically Deep Convolutional Neural Networks (DCNNs), superior performance is achieved in automatic medical image segmentation. Nowadays DCNNs are known as prosperous visual models and the backbone of image segmentation methods. Despite the attention received from the community in recent years and the success of deep learning based approaches, many challenges remain to be solved, such as the case of unbalanced number of class samples, challenging shapes, model complexity, scarcity of annotated data, multi-modality segmentation, etc.

In this thesis, some of these challenges are addressed in the area of semantic segmentation by proposing novel model architectures, loss functions and image processing techniques to improve the model performance and generalisability of DCNN based methods. In summary, the following problems are tackled. (I) To improve the performance on segmenting small and difficult object regions, a dynamically weighted cross-entropy loss function is proposed to train DCNN models to focus more on the poorly predicted image regions. (II) To improve the segmentation accuracy without increasing the number of model parameters, an efficient DCNN network architecture is proposed by adopting the ideas from residual network and dense network. (III) To achieve multi-modality image segmentation, where image labels of only one modality (source domain) are available for model training and the image labels for the other modality (target domain) are not available, a novel end-to-end DCNN is proposed. In this method, a new loss term is also added to the GAN to minimise the feature distances of domains in the latent space, so that images from different domains are close to each other in the latent space. (IV) To improve the multi-modality image segmentation in terms of performance and model complexity, an extension of the previously proposed method (i.e. III) is proposed. In this method, a new approach is applied to images of both source and target domains for minimising the intensity discrepancy. Subsequently, a dual task encoder-decoder network including image segmentation and reconstruction is utilised to effectively adapt a segmentation network to the unlabelled target domain. Additionally, a shape constraint is imposed by leveraging adversarial learning. Finally, the model is able to segment images of both source and target domains.

All the above developed methods were evaluated using several locally collected and public datasets that cover different organs (e.g. kidney, heart, brain, etc.), different dimensions (2D and 3D) and different imaging modalities (e.g. computed tomography, magnetic resonance imaging, etc.). The proposed methods were comprehensively compared with the state-of-the-art methods in each of the corresponding scenarios using various evaluation metrics (e.g. dice coefficient, surface distance, precision, recall, number of model parameters,

memory usage, training time, inference time, etc.). The experimental results reveal that the proposed methods are effective, and they achieved the state-of-the-art performance. Limitations of the proposed methods and future work are discussed at the end of the thesis.

Acknowledgements

First of all, I would like to express my deepest gratitude to my supervisor, Dr. Xin Chen for giving me the opportunity to work in this exciting project and for his guidance and support during the last four years. I appreciate the time he has spent mentoring me during this time, and bringing hope when I had lost mine. I am also thankful that he has been patient with all my questions. I would also like to thank my second and third supervisors Prof. Jonathan Garibaldi and Prof. Susan Francis for all their guidance on my research. Thanks to all my colleagues, specially Ruizhe Li for all his helps.

Contents

Abstract i

Acknowledgements iii

1 Introduction 1

1.1 Introduction 1

1.2 Technical Challenges 5

1.2.1 Unbalanced Class Samples 6

1.2.2 Model Complexity 6

1.2.3 2D vs 3D 7

1.2.4 Insufficient Training Data 7

1.2.5 Domain Shift 8

1.2.6 Difficulty in Shape Analysis 8

1.3 Contributions 9

1.3.1 FU-Net: Multi-class Image Segmentation using Feedback Weighted U-Net 9

1.3.2 DRU-Net: An Efficient Deep Convolutional Neural Network for Medical Image Segmentation 10

1.3.3 CIDA: Cross-modality Image Segmentation via Domain Adaptation of Unpaired Images 11

1.3.4 LMISA: A Lightweight Multi-modality Image Segmentation Network via Domain Adaptation using Gradient Magnitude and Shape Constraint 12

1.4 Publications 13

1.5 Thesis Structure 14

2 Literature Review **16**

2.1 Classical Methods 17

2.1.1 Threshold based 17

2.1.2 Edge based 18

2.1.3 Region based 19

2.1.4 Energy based 20

2.1.5 Partial Differential Equation (PDE) based 22

2.2 Learning based Methods 23

2.2.1 Deep Learning 23

2.2.2 Domain Adaptation Methods 27

2.2.3 Shape Preservation Method 30

2.3 Performance Metrics 32

2.3.1 Dice 32

2.3.2 Precision and Recall 33

2.3.3 Jaccard 33

2.3.4 Average Symmetric Surface Distance 34

2.3.5 Hausdorff Distance 34

2.3.6 Statistical Tests 34

2.4 Conclusions 35

3 FU-Net: Multi-class Image Segmentation using Feedback Weighted U-Net **36**

3.1 Introduction 36

3.2 Methodology 38

3.2.1 Network Architecture 38

3.2.2 Batch Normalisation and Residual Network 39

3.2.3 Weighted Cross-Entropy Cost Function 40

3.3 Method Evaluation 42

3.3.1 Datasets and Preprocessing 43

3.3.2 Experimental Results 44

3.4 Conclusions and Discussion 48

4 DRU-net: An Efficient Deep Convolutional Neural Network for Medical Image Segmentation 50

4.1 Introduction 50

4.2 Methodology 51

4.3 Method Evaluation 54

4.3.1 Datasets and Preprocessing 54

4.3.2 Experimental Results 56

4.4 Conclusions and Discussion 61

5 CIDA: Cross-modality Image Segmentation via Domain Adaptation of Unpaired Images 63

5.1 Introduction 63

5.2 Methodology 65

5.2.1 GAN based Unsupervised Feature Learning 66

5.2.2 Cross-modality Image Segmentation 68

5.2.3 Network Architecture 68

5.3 Method Evaluation 69

5.3.1 Datasets, Preprocessing and Parameter Settings 69

5.3.2 Experimental Results 70

5.4 Conclusions and Discussion 73

6 LMISA: A Lightweight Multi-modality Image Segmentation Network via Domain Adaptation using Gradient Magnitude and Shape Constraint 74

6.1 Introduction 74

6.2 Methodology 76

6.2.1 Multi-resolution Locally normalised Gradient Magnitude (MLNGM) 77

6.2.2 Image Segmentation Network 80

6.2.3 GAN based Unsupervised Feature Learning 81

6.2.4 Model Training 85

6.3 Method Evaluation 86

6.3.1 Datasets, Preprocessing and Parameter Settings 86

6.3.2 Experimental Results 90
6.4 Conclusions and Discussion 102
7 Conclusion **104**
7.1 Main Contributions and Limitations 104
7.1.1 FU-Net: Multi-class Image Segmentation using Feedback Weighted U-Net 105
7.1.2 DRU-net: An Efficient Deep Convolutional Neural Network for Medical Image Segmentation 106
7.1.3 CIDA: Cross-modality Image Segmentation via Domain Adaptation of Unpaired Images 107
7.1.4 LMISA: A Lightweight Multi-modality Image Segmentation Network via Domain Adaptation using Gradient Magnitude and Shape Constraint 108
7.2 Future Works 109
Bibliography **110**
Appendices **124**
A List of Abbreviations **124**

List of Tables

2.1 Segmentation methods. 17

3.1 Comparison of U-net, BRU-net and FU-net on skin data. The mean dice coefficient values, precision, and recall index are reported. 45

3.2 Comparison of U-net, BRU-net and FU-net on Retinal data. The mean dice coefficient values, precision, and recall index are reported. 45

3.3 Comparison of different methods using different number of training samples. The mean dice coefficient (Dice) ± standard deviation (Std) and P values of paired t-tests are reported. Numbers in bold indicate the best method that performed statistically ($P < 0.01$) better than other methods. 46

4.1 Comparison of U-net, RU-net, DU-net, AttnU-net and DRU-net on ISIC 2018. The mean dice coefficient values, precision, recall, and Jaccard index are reported. 57

4.2 Comparison of U-net, RU-net, DU-net and DRU-net on different sizes of training data using brain MRI dataset. The mean dice coefficient values with Wilcoxon signed rank test results and training time are reported. Based on Wilcoxon signed rank test: $*$ indicates result of DRU-net is different from U-net with statistical significance ($p < 0.01$); $\urcorner$ indicates result of DRU-net is different from RU-net with statistical significance ($p < 0.01$); $\circ$ indicates result of DRU-net is different from DU-net with statistical significance ($p < 0.01$). 58

4.3 Segmentation results of kidney dataset trained on 1.5T, 3T, CKD, and AKI modalities based on different methods. The results are reported based on Dice, Precision, Recall, ans ASSD. Mean $\pm$ standard deviation values for all testing images are reported. 60

4.4 Segmentation results of kidney dataset based on different methods using 5-fold cross validation. The results are reported based on Dice, Precision, Recall, and average symmetric surface distance (ASSD). Mean $\pm$ standard deviation values for all testing images are reported. 60

5.1 Segmentation results of different methods. The results are reported based on Dice, Jaccard, Precision, and Recall. 71

6.1 Segmentation results of 3D kidney dataset based on different image pre-processing methods. The results are reported based on Dice, Precision, Recall, and average symmetric surface distance (ASSD). Mean $\pm$ standard deviation for all testing images are reported. 92

6.2 Segmentation results of kidney dataset based on different methods using 5-fold cross validation. The results are reported based on Dice, Precision, Recall, and average symmetric surface distance (ASSD). Mean ± standard deviation values for all testing images are reported. Asterisk indicates the best 2D method and the overall best domain adaptation method is highlighted in bold. 96

6.3 Segmentation results of kidney dataset from an unseen target domain (1.5T MRIs of 83 subjects) based on different methods. The results are reported based on Dice, Precision, Recall, and average symmetric surface distance (ASSD). Mean ± standard deviation values for all testing images are reported. Asterisk indicates the best 2D method and the overall best domain adaptation method is highlighted in bold. 99

6.4 Based on the segmentation of kidney dataset, the number of parameters, memory usage, training time and inference time of different methods are reported. 99

6.5 Segmentation results of cardiac dataset based on different methods. The results are reported based on Dice and the average symmetric surface distance (ASSD). Asterisk indicates the best 2D method and the overall best domain adaptation method is highlighted in bold. Note: N/A means that the ASSD cannot be calculated due to no prediction. 101

List of Figures

1.1 Visual examples of medical image segmentation in different datasets. a) Skin lesion data [1], b) Retinal data [2], c) CT scan of kidney, d) MRI scan of kidney. As can be seen a label is assigned to every pixel in each image such that pixels with the same label share certain characteristics. 3

1.2 Visual examples of two different modalities of MRI kidney images. a) T1 and b) T2 weighted MRI images of healthy subjects [3] 5

1.3 Examples of unbalanced class samples on brain data as it is seen Substantia Nigra (white label) in comparison with Midbrain (gray label) and backfround (black label) is very small, and hard to detect [4]. 6

2.1 Thresholding segmentation approach on Retinal data (Drive dataset)[5] 18

2.2 Seeded region growing segmentation method [6] 20

2.3 The overall scheme of graph cut segmentation method [7] 21

2.4 The architecture of CNN [8] 24

2.5 The architecture of FCN [9]. 25

2.6 The architecture of DeepLab [10] 25

2.7 The architecture of SegNet [11] 26

2.8 The architecture of U-net [12] 27

2.9 The architecture of PnP-AdaNet[13]. As it is shown, this method is complex and includes a lot of learning modules. 28

2.10 The architecture of DAPNet[14] 28

2.11 The architecture of SIFA[15] 30

3.1 (a) Layer of the original U-net. (b) Layer by adding batch normalisation (BN) and residual block (RB). 39

3.2 Plot of the mapping function of Eq. (3.4) with different values of β 42

3.3 Examples of Skin datsets, the images and the ground-truths [1] . 43

3.4 (a) The original image. (b) The ground truth. (c) (d) (e) Segmentation results by U-net/BRU-net/FU-net respectively using 50 training samples. (f) (g) (h) Segmentation results by U-net/BRU-net/FU-net respectively using 100 training samples. (i) (j) (k) Segmentation results by U-net/BRU-net/FU-net respectively using 200 training samples. 47

4.1 a) Residual block. b) Dense net blocks. c) Encoder and decoder blocks in DRU-net. 52

4.2 The overall scheme of DRU-net. 54

4.3 a) Original image. b) Ground truth. c) Segmentation result by U-net. d) Segmentation result by RU-net. e) Segmentation result by DU-net. f) Segmentation result by DRU-net. 57

4.4 Visual results for each comparative methods 1.5T, 3T, CKD, and AKI modalities respectively. First column is original image (Org), second column ground truth (GT), third column shows the results of U-net and last column represents the results of DRU-net 59

5.1 The overall scheme of the proposed method. Generator and discriminator loss applied to encourage the generator to generate synthesized images that has high probabilities of being classified as real by the discriminator. Dissimilarity loss is added to minimise the absolute distances of the latent representations (Z_s and Z_t) of images in the source and target domains. Segmentation loss is require as the task is segmentation. 66

5.2 Visual results of three example images for each comparative methods. 72

5.3 Visualization of the latent feature space of the segmentation network using t-SNE for CT and MRI edge images, without (a) and with (b) the proposed L_{latent} loss. 72

6.1 The overall scheme of LMISA. 1) The images from both domains are input to the image pre-processing method (Multi-resolution Locally normalised Gradient Magnitude (MLNGM)). 2) The output of MLNGM is then input to an encoder-decoder network as a batch of images from both domains. L_G, L_{seg} and L_D are the loss terms to be optimised during model training. . . 76

6.2 Multi-resolution locally normalised gradient magnitude (MLNGM) scheme. 78

6.3 By varying the *thr* value in MLNGM for both CT to MRI and MRI to CT domain adaptation, Dice scores (vertical axis) were compared using LMISA on kidney dataset. 90

6.4 Visualization of the original images (Org), output of MLNGM by *thr = 2*, MLNGM by *thr = 10*, MLNGM by *thr = 14*. The top row is CT and the bottom row is MRI. 90

6.5 Visualization of the original images (Org), canny edge detection (Edges), gradient magnitude (GM), and the proposed method (MLNGM). The top row is CT and the bottom row is MRI. . . 92

6.6 Ablation study of the key components in the proposed method based on Dice score for both CT to MRI and MRI to CT domain adaptation on kidney dataset. 93
6.7 Ablation study of the key components in the proposed method based on Dice score for both CT to MRI and MRI to CT domain adaptation on Cardiac dataset. The average Dice score of four organs are used here. 94
6.8 Visualization of the latent feature space of LMISA with and without *MLNGM* using t-SNE for both CT to MRI and MRI to CT domain adaptation on 3D volumes. 94
6.9 Visual results for each comparative method using CT and MRI as the target domain respectively. The top three rows are CT, and the bottom three rows are MRI images. These images are chosen based on the best, median, and the worst dice score of DRU-net method. 97
6.10 Example slices of 3T MRI and 1.5T MRI, and their corresponding MLNGM output. 99
6.11 Visual results of LMISA-3D using MRI and CT as the target domain respectively. The original image (Org), ground truth annotation (GT) and LMISA-3D prediction are shown for both MRI and CT as the target domain respectively. 102

Chapter 1

Introduction

1.1 Introduction

With the advent of medical imaging equipment, and the possibility of scanning and loading medical images into a computer, automated medical image analysis has captured the attention of many researchers. Initially, medical image analysis was performed with low-level pixel processing approaches such as edge and line detector filters [16][17][18], region growing [6][19], mathematical modelling [20], and expert systems with many if-then-else statements [21]. On the other side, with the rapid development of learning approaches, especially deep learning [22], numerous image analysis methods based on deep learning have been developed in the literature and have made great progress on the designs and performance of medical applications.

Medical images are presented in both 2D and 3D formats, and are captured in various modalities such as CT (Computed Tomography), MRI (Magnetic Resonance Imaging), X-Ray and ultrasound. In this thesis CT and MRI are used for training and evaluation of the methods. CT images are harder to interpret. However, as CT scaning process is faster; then, CT is less sensitive to patient motion during the examination, and less expensive in comparison with MRI. MRI images use magnetic fields and include more details, are clearer, and provide a better picture of organs and soft tissues in comparison with CT

images.

Image segmentation is the process of extracting meaningful information by dividing the image into segments (Figure 1.1). Medical images normally contain a lot of information, however, only some small parts are of interest for a particular clinical application. Using segmentation, unnecessary information can be removed, and elements that are related to the same tissue type or organ can be extracted. Thus, image segmentation is a fundamental and crucial step in many medical image analysis tasks (e.g. computer assisted diagnosis, radiotherapy planning, preparation, assessing before surgical operations, and treatment [23][24]).

There are a lot of segmentation techniques such as: traditional methods including thresholding [5], clustering approaches [25], Gaussian distribution models [26], shape prior based methods including deformable models [27], and machine learning based methods including SVM [28], KNN [29], random forest [30] and deep neural networks [11]. In the next chapter, a proper review is provided on the previous works. Despite the progress in segmentation methods, because of variation in texture, different grades of disease, different imaging techniques, and high input data dimensionality and scarcity of available annotated data, accurate medical image analysis is a challenging task.

The emergence of deep learning, in particular deep convolutional neural networks (DCNNs), makes end-to-end image processing possible, and have rapidly become a noteworthy methodology for analysing medical images [8] [31] [32]. A neural network simulates humans' brain, and is composed of many neurons. Each neuron is connected to other neurons, and is considered as an information-processing unit. The key point about deep learning in comparison to traditional methods is that these approaches instead of using handcrafted features, are able to learn hierarchical feature representations automatically, and share the weights of feature mapping in different layers of the network. In these approaches, the features are mostly designed based on domain-specific knowledge, learned from data, while classical methods require a good feature

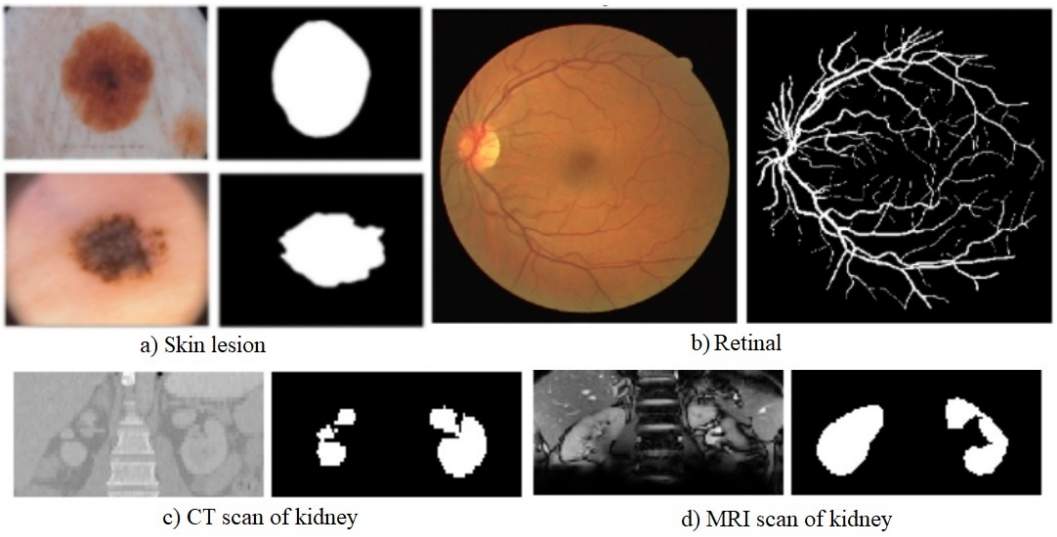


Figure 1.1: *Visual examples of medical image segmentation in different datasets. a) Skin lesion data [1], b) Retinal data [2], c) CT scan of kidney, d) MRI scan of kidney. As can be seen a label is assigned to every pixel in each image such that pixels with the same label share certain characteristics.*

extractor often designed by hand based on expert knowledge [8]. These methods, in the first step, learn hierarchical feature representations by using the inference from simple features; in the next step, building up more complex features from the previous level, gradually.

In terms of structure, a deep learning network comprises of an input layer, several hidden layers, and an output layer. The input image is fed to the input layer. The hidden layers performs feature extraction on the input data to obtain a feature map and prepares feature map for the next layers, these operations normally are convolution, max pooling and activation layers. Although these networks in their basic form were designed for classification task, they could simply be applied to segmentation task by considering every single pixel or voxel in an image, and investigating if the pixel or voxel belongs to the object of interest. The most well-known DCNN architecture for supervised segmentation task is U-net, proposed by Ronneberger et al. [33]. The two main contributions in U-net are the combination of up-sampling and down-sampling layers. Furthermore, there are skip connections between the corresponding convolution and deconvolution layers, which concatenate their feature maps. In the next chapter this model and some other state-of-the-art

methods are explained in detail.

Despite the success achieved by DCNN based segmentation approaches, there are still a large number of open research questions due the requirement of (i) balanced class samples (ii) having sufficient amount of annotated data, and (iii) sampling of the training (source domain) and testing images (target domain) from the same data distribution. However, in medical scenarios, there are hard examples, and image annotations require careful and time-consuming analysis from trained experts. Furthermore, in medical imaging contexts, there are various imaging modalities, scanners and protocols that capture various physical properties and provide different information for disease diagnosis and treatment (Figure 1.2). Therefore, the training and testing instances can come from different underlying data distributions (known as domain shifts). Gradient vanishing is another issue that may happen in deep learning-based methods. During the backpropagation process sometimes, the gradients get smaller and approach to zero. As a result, the weights of the lower layers cannot be updated, and the gradient descent never converges to the optimum. Although there are a number of approaches to cope with the aforementioned limitations, to the best of the author's knowledge, all of them are computationally expensive and cannot be applied on 3D volumes as they consist of a lot of learning modules. Besides, there is still a need of boosting the performance.

Motivated by this, this thesis aims to develop more efficient and effective methods for accurate medical segmentation task by improving a variety of existing DCNN based methods. This thesis begins with investigating the supervised methods (i.e. when enough annotated images are available), their limitations, and proposing some ideas to improve the supervised methods, such as: i) a novel cost function to focus on poorly predicted regions more to cope with hard examples; ii) a powerful supervised segmentation network to boost the performance, but decrease or remain stable the model complexity. The thesis continues by focusing on multi-modality image segmentation when there is

not enough annotated data, and the testing (target) and training (source) are coming from different domains. Hence, the following methods are proposed: iii) an efficient model to segment images from different domains to achieve cross-modality image segmentation of unpaired images; iv) an extension of previous method to improve multi-domain modality segmentation in terms of performance and model complexity.

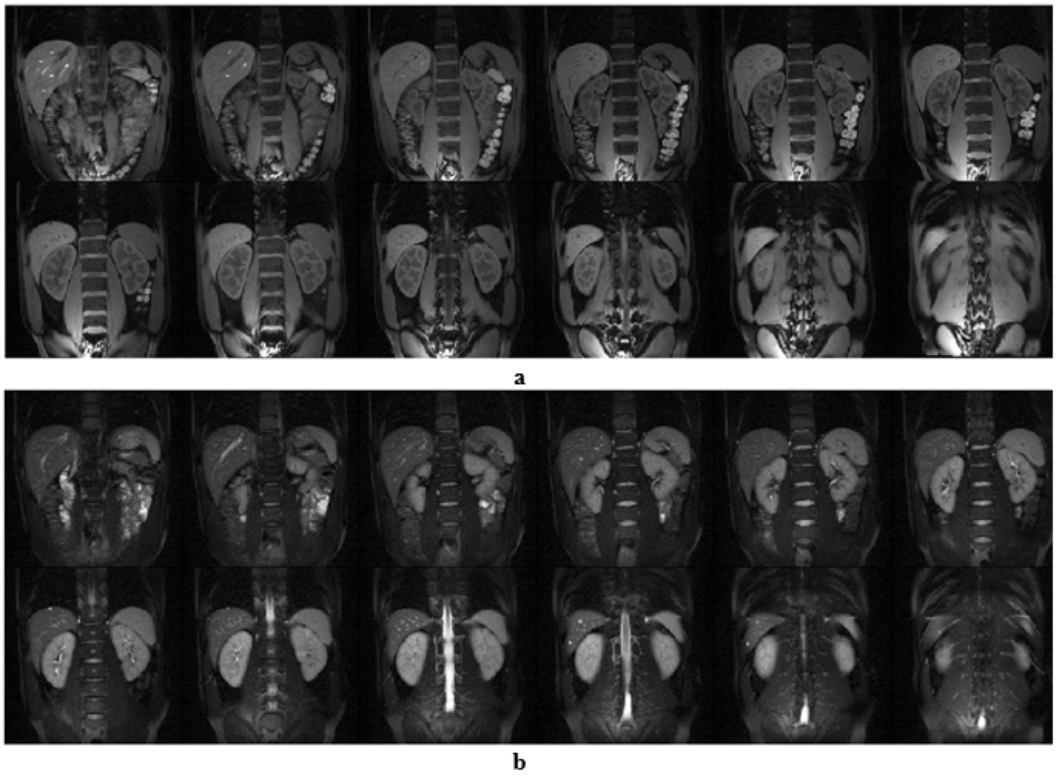


Figure 1.2: *Visual examples of two different modalities of MRI kidney images. a) T1 and b) T2 weighted MRI images of healthy subjects [3]*

1.2 Technical Challenges

A multitude of approaches with various degree of success have been proposed so far to solve the semantic segmentation task. However, despite past efforts, a series of problems still make the current methods unsuitable for some of the applications. In the following, these obstacles are described and the proposed solutions are briefly explained in the next sections.

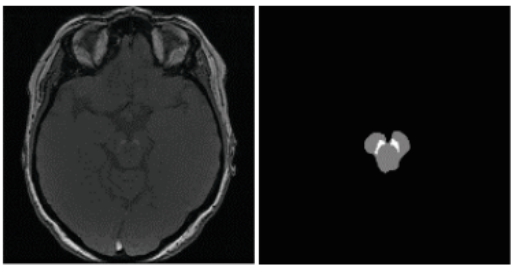


Figure 1.3: *Examples of unbalanced class samples on brain data as it is seen Substantia Nigra (white label) in comparison with Midbrain (gray label) and backfround (black label) is very small, and hard to detect [4].*

1.2.1 Unbalanced Class Samples

The problem of class imbalance arises when the distribution of examples across the classes is biased. The distribution can vary from a slight to a severe imbalance between the minority classes and the majority classes. Imbalanced class is a challenge for most of the machine learning models as these models were developed assuming balanced classes however, in real life it is not common to have properly balanced data. For most multi-class image segmentation problems, also the number of pixels in each class is different from each other which potentially leads to classes with more number of samples dominate the training process, and less accurate predictions for some classes than others. Additionally, some of the image regions are easier to be classified (i.e. higher segmentation accuracy) than others due to more distinct local image characteristics. It would be more efficient if the network can be dynamically adapted to learn from pixel locations with lower predicted accuracies during the training process. Examples of such challenging cases can be found in Figure 1.3.

1.2.2 Model Complexity

Most of the DCNN methods are computationally expensive, requiring one or more GPUs for training. On the other side, by decreasing the model complexity, performance drops. Therefore, most methods achieved better performance by integrating sophisticated network blocks (e.g. dense network[34], attention network[24], etc.), multiple network modules and various aspects of learning,

which require more parameters to be learned. Generally, the implementation of such a complicated framework even in 2D is memory intensive and training prohibitive, and it is almost impossible to work with 3D volumes using a single GPU workstation. Thus, it would be more desirable to reduce the number of model parameters and memory consumption while still achieve a similar or better performance.

1.2.3 2D vs 3D

Most of the prior works assume the input images as 2D slices. As the object orientation changes, there is a difference between the 2D and 3D annotations. Besides, 2D images inherently fail to leverage context from adjacent slices, therefore, this assumption also results in performance drop. The ability to process 3D voxels leverages inter slice information and boosts the performance, however, the computational cost is increased as the number of parameters is increased.

1.2.4 Insufficient Training Data

The performance of deep learning models is dependent on the amount of annotated data for training. However, most of the time, the available data is not fully annotated. The insufficient training data is a common issue in medical imaging, where reliable annotation still required to be prepared by experts manually, thus it is expensive labor, a difficult task, and time-consuming. However, data collection is increasingly inexpensive, and a large amount of medical images is produced everyday that generates the availability of large unlabeled datasets. Learning from unlabeled data without any supervision, normally perform significantly worse than the fully-supervised methods. Thus, a robust and effective method for learning using limited annotations is highly demanding.

1.2.5 Domain Shift

In the supervised setting, a common assumption is that the source (training images) and the target domains (test images) share similar characteristics or sampled from the same data distribution. Therefore, the performance of the target domain depends on the performance of the source domain, and similarity between the distributions of the source and target domains. However, normally there is a difference in data distribution between the source and target datasets that is called domain shift and is a common case scenario in medical applications. Most of the models cannot cope with changes in the distribution, thus the accuracy drops when evaluates on unseen data. Many approaches have been developed to take care of domain shift, which aim to find a common representation to minimise the divergence between the domains, known as domain adaptation approaches. But most of these methods require some ground-truth labels from the target domain to learn the proper adapted model. However, in most case scenarios the target domain is unlabeled. Therefore, it is required to utilise the unsupervised domain adaptation approaches.

1.2.6 Difficulty in Shape Analysis

The usage of shape information in deep learning networks is limited, according to [35]. DCNNs are capable of encoding local shape features including local edge segments, but less capable of preserving the global shape or contour of objects. Some of the reasons for the above challenges can be due to the filters, focus on local intensity changes rather than encoding the shape. Therefore, maintaining contextual information of shape and boundary of the object is still a challenging task and a fundamental topic in computer vision. According to the literature, shape-aware components have been widely used in segmentation methods to constrain a segmentation result to a class of learned shapes. However, a light-weight and more accurate method in terms of segmenting the shape and contour of the object properly is still needed.

1.3 Contributions

This thesis improves current semantic segmentation approaches and proposes a generalised model for multi-modality image segmentation. The proposed approaches focus on addressing the main problems and challenges as mentioned previously. In the following, the objectives and the contributions are described.

1.3.1 FU-Net: Multi-class Image Segmentation using Feedback Weighted U-Net

Most of the semantic segmentation methods face the class imbalance problem during training, i.e. a large number of candidate locations per image need to be evaluated but only a few locations contain objects. This class imbalance causes performance drop and inefficient training process because most locations are easy with no valuable learning clue, but they suppress the training. Hard negative mining [36][37] and reweighting the contributions of pixels [38] are some common ideas to solve the imbalance in class samples. However, these methods are so complex, and cannot make decision dynamically according to the predictions. In this thesis, a dynamically weighted cross-entropy loss function is proposed to address this class imbalance problem by reshaping the standard cross-entropy loss such that it is able to reweight the loss assigned to examples. The weighting scheme is calculated based on the pixel-wise prediction accuracy during the training process. This method assigns higher weights to pixels with lower segmentation accuracies to enable the network to learn more from poorly predicted image regions, and down-weights the well-classified examples to avoid overwhelming the training process by these easy samples. The proposed method is named as feedback weighted U-net (FU-net). The method is evaluated on the segmentation datasets, including a public dataset (ISIC 2018 grand-challenge) for skin lesion segmentation, a retinal public dataset (Drive), and a local brain MRI dataset, where the number of pixels in each class is extremely unbalanced to each other. Experimental results reveal that the pro-

posed FU-net outperforms the performance of the other methods specially the classes with small number of samples. This method is described in detail in chapter 3.

In addition to evaluating the proposed method on medical datasets, FU-net is assessed on images from other field of studies including prognostic and health management of automotive and aerospace and compared with other state-of-the-arts methods [39].

1.3.2 DRU-Net: An Efficient Deep Convolutional Neural Network for Medical Image Segmentation

Residual network (ResNet) [40] and densely connected network (DenseNet) [34] have significantly improved the training efficiency and performance of DCNNs mainly for object classification tasks. However, these methods have complex learning blocks and are time consuming in the train and test process. In addition, still there is a need of boosting performance. Thus, in this thesis, an efficient, yet simple, DCNN architecture is proposed by considering advantages of ResNet and DenseNet networks. The proposed method is integrated into an encoder-decoder DCNN model for medical image segmentation. The method adds additional skip connections compared to ResNet but uses significantly fewer model parameters than DenseNet. The proposed method is evaluated on a public dataset (ISIC 2018 grand-challenge) for skin lesion segmentation, a local brain MRI dataset, and a local Kidney dataset. In comparison with ResNet-based, DenseNet-based and attention network (AttnNet) based methods within the same encoder-decoder network structure, the proposed method achieves significantly higher segmentation accuracy with fewer number of model parameters than DenseNet and AttnNet. This method is presented in chapter 4.

Furthermore, DRU-net is evaluated on images from another field of study, buildings and architectures [41].

1.3.3 CIDA: Cross-modality Image Segmentation via Domain Adaptation of Unpaired Images

In medical image segmentation, supervised machine learning model trained using one image modality (e.g. CT) is often prone to failure when is applied to another image modality (e.g. MRI) even for the same type of organ. This is due to large geometric differences across different subjects and significant intensity variations of different image modalities. There are a number of techniques for addressing domain adaptation, including those based on discrepancy-based methods [42], self-ensembling methods [43], methods built upon adversarial learning and cycle consistency [44] [45], and disentanglement strategies [46] [47]. Although a good progress has been achieved in tackling this problem, these frameworks normally consist of multiple network modules and various aspects of adversarial learning. The implementation of such a complicated framework even in 2D is memory intensive and training prohibitive, and it is almost impossible to work with 3D volumes using a single GPU workstation. In contrast to the aforementioned methods, a novel method to achieve cross-modality image segmentation of unpaired images is proposed, where image labels of only one modality is available for model training. In this method, Canny edge detection is firstly applied to images of both domains for minimising the intensity discrepancy. Subsequently, a generative adversarial network (GAN) is utilised to learn the latent representation of the images of both domains in an unsupervised manner. A new loss term is added to the GAN for minimising the feature distances of both domains in the latent space, so that images from different domains are close to each other in the latent space. The learned encoder in GAN is then used as the encoder of an encoder-decoder segmentation network (i.e. U-net). The segmentation network is then trained using images and their corresponding labels of one domain only. Finally, images from the other domain can be segmented, as they share similar latent feature representations with images of the learned domain. The proposed

method was evaluated using a public CT dataset (KiTS19, MICCAI 2019) and a local MRI dataset for kidney segmentation. The segmentation model is trained only using CT images and able to segment kidneys in MRI images. The proposed method achieved significantly higher performance in comparison with other state-of-the-art methods. The description of this method is presented in chapter 5.

1.3.4 LMISA: A Lightweight Multi-modality Image Segmentation Network via Domain Adaptation using Gradient Magnitude and Shape Constraint

LMISA is an extension to CIDA in terms of accuracy and model complexity. LMISA is a novel end-to-end deep neural network to achieve multi-modality image segmentation. In this method, a multi-resolution locally normalised gradient magnitude approach is firstly applied to images of both domains for minimising the intensity discrepancy. Subsequently, a dual task encoder-decoder network including image segmentation and reconstruction is utilised to effectively adapt a segmentation network to the unlabelled target domain. Additionally, a shape constraint is imposed by leveraging adversarial learning. Finally, images from the target domain are segmented, as the network learns a consistent latent feature representation with shape awareness from both domains. Both 2D and 3D versions of the proposed method are implemented, in which CT and MRI images for kidney and cardiac organ segmentation are utilised. Experimental results reveal that the proposed method achieves significantly higher performance with a much lower model complexity in comparison with other state-of-the-art methods. More importantly, the method is also capable of producing superior segmentation results than other methods and CIDA for images of an another unseen target domain without model retraining. This method is presented in chapter 6.

1.4 Publications

The research presented in this thesis has been published in the following journals and conferences:

- M Jafari, R Li, Y Xing, D Auer, S Francis, J Garibaldi, X Chen, FU-Net: Multi-class Image Segmentation Using Feedback Weighted U-Net, International Conference on Image and Graphics 11902, 2019, 529-537.

- M Jafari, D Auer, S Francis, J Garibaldi, X Chen, DRU-Net: An Efficient Deep Convolutional Neural Network for Medical Image Segmentation, 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), 1144-1148.

- M Jafari, S Francis, C Buchanan, J Garibaldi, X Chen, ASKDCNN: Automatic Segmentation of Kidneys using A Novel Deep Convolutional Neural Network, 3rd International Symposium on Functional Renal Imaging 2019.

- M Jafari, S Francis, J Garibaldi, X Chen, LMISA: A Lightweight Multi-modality Image Segmentation Network via DomainAdaptation using Gradient Magnitude and Shape Constraint, Medical Image Analysis Journal, vol. 81, p. 102536, 2022.

In addition to them, FU-net and DRU-net are applied to another scenarios and published in journals as follows:

- D Rengasamy, M Jafari, B Rothwell, X Chen, GP Figueredo, Deep learning with dynamically weighted loss function for sensor-based prognostics and health management, Sensors 20 (3), 723.

- SM Sajjadian, M Jafari, X Chen, Automatic Architectural Drawing Labelling Using Deep Convolutional Neural Network, Sustainability in Energy and Buildings 2021, 69-78.

1.5 Thesis Structure

The overall structure of this thesis is comprised of seven chapters. The rest of this thesis is organized as follows.

1. Chapter 2 reviews the existing semantic segmentation methods by starting from classical methods such as thresholding, edge and region based approach following by learning based semantic segmentation approaches including supervised, semi-supervised and unsupervised learning methods.

2. Chapter 3 proposes FU-net. This chapter starts with reviewing ROI based and reweighting the contributions of pixels techniques for balancing the foreground and background class samples. Next, a background of the most successful segmentation method (i.e. U-net in 2017) is described. Later the proposed weight map method and finally quantitative and visual results are presented.

3. Chapter 4 represents DRU-net. This chapter begins with a literature review on different network blocks and architectures. Later, the description of the proposed method, and the quantitative and visual results in comparison to other state-of-the-art methods are presented.

4. Chapter 5 proposes CIDA. In this chapter, a review on related works on domain adaption and a background is presented. Later, the description of the proposed method and the segmentation results are reported based on different criteria, and compared with other state-of-the-art methods.

5. Chapter 6 proposes LMISA. This chapter begins with a review on previous domain adaption ideas and their limitations. Next, the proposed method is described in detail. Later, the experimental results are investigated including ablation study, assessing the performance, the model complexity (i.e. number of parameters, memory usage, training and inference time).

6. Chapter 7 draws the a conclusion including a briefly review of the proposed approaches and the summary of the problems that have been addressed in this thesis, and finally, the future works are discussed in this chapter.

Chapter 2

Literature Review

Medical image segmentation plays a vital role for disease diagnosis and treatment. Studying and interpreting the medical images, manually by the radiologist, is a time consuming and sophisticated task, especially by increasing medical imaging modalities, scanners and protocols. Besides, it is not very accurate as it is dependent on the experience of the radiologist. There are several existing techniques which are used for image segmentation automatically, all of these techniques have their own merits and demerits. These techniques can be classified into different categories. A summary of these methods is provided in Table 2.1.

In general, traditional image segmentation methods (e.g. threshold-based segmentation method [5], and region-based image segmentation method [6]) use the knowledge of digital image processing and mathematics to segment the image. The calculation is simple and fast, but the accuracy of the segmentation cannot be guaranteed in terms of contour detail. In contrast, learning-based methods especially deep learning [48][49][40] surpass the accuracy of the traditional segmentation methods. The fully convolutional network [9] is the pioneering work that successfully uses deep learning for image semantic segmentation. There are other outstanding segmentation networks like U-Net [33], DeepLab [10], SegNet [11], and Mask R-CNN [50], etc that have been used in various applications.

In particular in this chapter, some of the methods in table 2.1 are described in details. Firstly, an overview of the classical segmentation methods is provided including thresholding, edge, region, energy, and partial differential equation (PDE) based methods. Then, the supervised deep learning based methods are reviewed. Next, the literature on the problem of multi-modality image segmentation is investigated. Then, the related works are reviewed on shape aware methods. Finally, at the end of this chapter, the metrics for evaluation of the segmentation methods are reviewed.

Table 2.1: Segmentation methods.

Category	**Method**	**Advantages**	**Limitations**
Classical Methods			
Threshold based	Thresholding	Low computational difficulty No need to prior knowledge Fast Easy concept and easy to implement	Hard to define a threshold value Unable to consider spatial details Sensitive to noise Unable to apply on multi-channel images
Edge based	Gradient operator Optimal edge detector Gray-level histogram technique	Accurate on images with good contrast between regions Fast and simple to compute Work well when there is noise immunity	Inaccurate on unclearly defined edges Unable to work when there are too many edges Less immune to noise
Region based	Region growing Region splitting and merging	Simple and more immune to noise Straightforward principle	Time and memory expensive Dependent on the selection of seed region Probability of over or under segmentation
Energy based	Active contour Graph-cut	Accurate on images with large shape varieties Fast and efficient	Sensitive to noise Sensitive to the number of initial settings like number and position of counters
Partial Differential Equation (PDE) based	Level Set Model Mumford Shah Model C-V Model	Stability and irrelevancy with topology Able to solve the problems of corner point producing and curve breaking	Sensitive to noise Sensitive to the choice of its parameters Complex process
Learning-based Methods			
Supervised	Support vector machine Random forest Linear regression CNN-based methods	Greater accuracy in comparison with classical methods Specific about the definition of the classes Task specific A simple process	Require a large set of high-quality labelled data Cannot cluster or classify data by discovering the features
Semi-supervised	Pseudo-labelling methods Graph based methods Generative-based methods	Need a small set of annotated data Require less expert analysis Shrink human labor and improve accuracy A great moderation among the advantages and disadvantages of supervised and unsupervised learning	Less accurate in comarison with supervised learning It can lead to worse performance if an algorithm with the wrong assumption is chosen
Unsupervised	Hard Clustering Fuzzy Clustering Generative-based methods	No need to annotated dataset Can explain the hidden patterns Minimise the error rate by human	It is not task-specific Less accurate results

2.1 Classical Methods

In the following, some of the classical image segmentation methods are reviewed.

2.1.1 Threshold based

Kohler in 1981 [5] proposed a segmentation method to segment each voxel according to its intensity, $I(x)$, using one or more thresholds, T. In its simplest

form, the method performs a binary segmentation using a single threshold as follows:

$$C(x) = \begin{cases} 1, & \text{if } I(x) \geq T. \\ 0, & \text{otherwise.} \end{cases} \tag{2.1}$$

where $I(x)$ is the intensity of the image at position x and $C(x)$ is the predicted class label of the voxel at position x. Lee et al. [51] have compared performance of five thresholding algorithms for image segmentation. Otsu's method (maximum variance) [52], maximum entropy method [53], and k-means clustering [54] are some of the popular methods that are usually used to select the threshold value. This method is easy to implement and efficient. There is no need of prior information or calculation. However, this method can not cope with noisy and blurred edges, it is lack of sensitivity and sharpness, and it is complex for multi-dimensional data. Figure 2.1 represents the overall scheme of thresholding technique on a sample of Retinal data (Drive dataset).

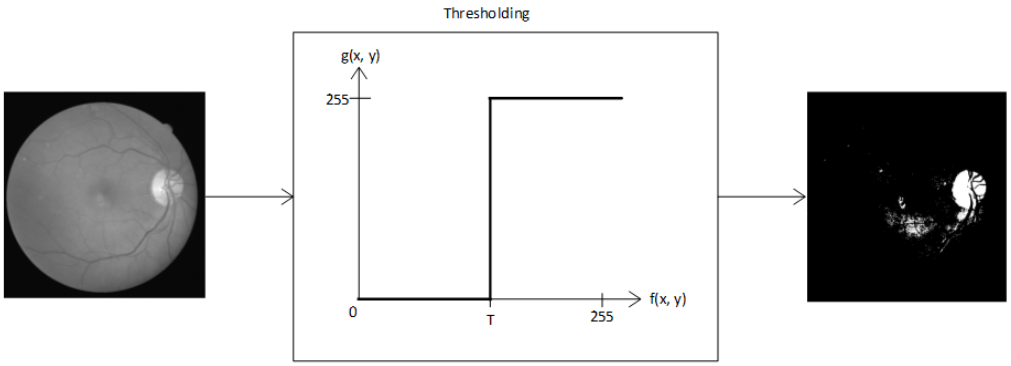


Figure 2.1: *Thresholding segmentation approach on Retinal data (Drive dataset)[5]*

2.1.2 Edge based

Edge based segmentation attempts to segment images by detecting the edges or pixels between different regions that have rapid transition in their intensities. Gray-level histogram and gradient based methods are the two main edge based segmentation methods [55]. Common edge detection operators are Sobel [56], Canny [57], Laplace [58], Laplacian of Gaussian (LOG) [59], etc. In general,

edge based methods are suitable for images that are simple and noise-free.

Edge based techniques may use as the backbone of another segmentation technique. Tony and Meng [60] segment edges as either straight or curved using minimum description length (MDL) approximation using break points. Firstly, in the pre-processing step, junction candidates (related to regions of interest) are detected. These junction features need to be matched to edges using a criterion. For each matched pair, a break point is considered at the closest point to the junction. Obtained breakpoints show the candidate points to form partitions.

2.1.3 Region based

Region based methods segment an image by considering the neighboring pixels within one region that have similar values according to a criterion. The performance of this method is sensitive to the selection of the similarity criterion, and still sensitive to noise. However, region based methods compare to edge based methods are simpler and more immune to noise.

Seeded region growing [6] is one of the commonly used region based segmentation methods. This method begins the process of segmentation with a set of seed pixels inside the object of interest. The seeds are set automatically or manually, then, expand to the neighbouring pixels providing that they satisfy predefined criteria. These criteria compare intensity or colour of the current pixel with the seed. This process continues till no neighbouring pixels satisfy the criteria. Recently, some of the researchers improved this technique [61][62][63][64]. This method performs better in noisy images. However, this method suffers from under-segmenting or over-segmenting. Figure 2.2 represents the scheme of region growing segmentation technique.

Morphological image processing is another region based method that is usually utilised by integrating with other segmentation algorithms. Morphological techniques use a mask to investigate each pixel value. The value of each

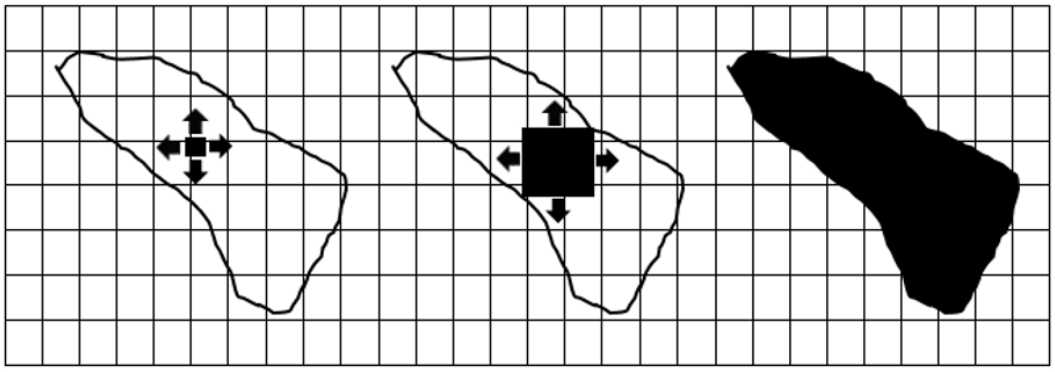


Figure 2.2: *Seeded region growing segmentation method [6]*

pixel is determined by the neighbouring pixels inside the mask. The simplest morphological operations are dilation and erosion. In a binary image, if the current pixel value in the mask is 1, dilation adds all pixels in the mask. If there are any pixels in the mask with value 0, erosion removes the current pixel with value 1 [65][66].

The concept of watershed segmentation [67][68] is based on mathematical morphology and is classified in the group of region based methods. The main idea of this method is to find the watershed lines, and there are also many different algorithms to compute watersheds such as watershed by flooding, watershed by topographic distance, watershed by the drop of water principle, and topological watershed [69][70][71]. This method produces more reliable output, however, needs to specify the seed points and may results in over-segmentation or under-segmentation.

2.1.4 Energy based

Graph cut is another classical segmentation method that uses the concepts of graph. All pixels have additional edges to two special nodes, a source (s) and a sink (t). In this method, a weight is assigned to the edges, so that the weights between similar pixels should be large, and small between different pixels. The segmentation is performed using a minimum graph cut algorithm; there are several algorithms to find the minimum cut [7][72][73]. These algorithms divide the nodes of a graph into two sets as shown in Figure 2.3. This method is fast

and accurate. However, as the size of the graph grows quadratically with the size of the image, this method cannot be used on very large images, besides, it may fail to identify the correct cuts. Therefore, this method cannot segment very sophisticated shapes.

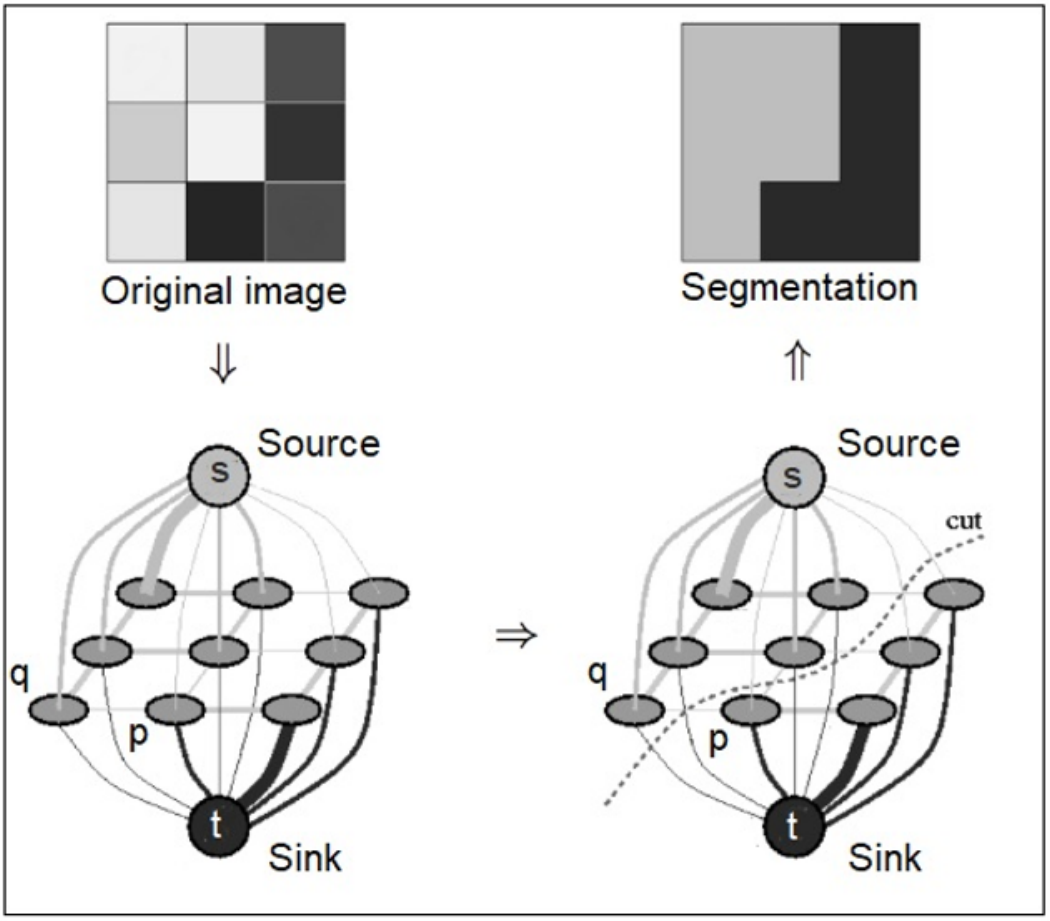


Figure 2.3: *The overall scheme of graph cut segmentation method [7]*

In Markov Random Field (MRF) segmentation [74] each pixel in the image is considered as a node in a graph. Each node has an edge to another node (neighbour pixels) and has a probability distribution that shows the probability of the pixel regarding to each class. Based on Markov approach, the probability related to the closest neighbours should be maximised.

Active contours [75][76] uses energy forces to segment an image. The contours are defined as $v(s) = [x(s), y(s)]$ where $x(s)$ and $y(s)$ are the coordinates for part s of the contour. The cost function is composed of an internal E_{int}

and external cost E_{ext} as follows:

$$\int_0^1 E_{int}(v, s) + E_{ext}(v(s))\, ds \tag{2.2}$$

The internal cost term is smoothness term from adjacent control points, and is used to govern deformable changes. The external loss term is driven by the image and it iteratively deforms the contour to be fitted to the object's boundary. The disadvantage of this method is that it cannot separate weak boundaries, fuzzy boundaries, and discontinuous object boundaries.

2.1.5 Partial Differential Equation (PDE) based

The basic idea for PDE-based methods is to deform a curve, surface or image to a PDE with boundary conditions. Then, the solution of the equation is considered as the segmentation results. These methods can also extend the framework from 2-D to 3-D or even higher dimensions. Level set method is one of the common methods based on PDE [77][78] that performs segmentation similar to active contours. Curve propagation [79] is another common technique in this category that performs image segmentation through the propagation planar curves/surfaces. In general, PDE based methods are robust, accurate, and efficient. However, they may over-segment as the curve may pass through object boundaries.

As a summary, classical approaches are engineered algorithms that rely on theory and mathematical models. The calculation is simple and the segmentation speed is fast, but the accuracy of the segmentation cannot be guaranteed and fail to detect objects under varying geometry and size. Therefore, the classical methods and handcrafted based methods belong to learning-based models are prone to failure because of the potential extracted features. While learning-based methods deliver significantly better accuracy than traditional methods as computers learn the features from data and can deal with the shape and size variations. In the next section, the learning-based methods are

reviewed.

2.2 Learning based Methods

Learning based methods typically requires a training process, where initially a human-designed feature descriptor (e.g. SIFT [80] etc.) is applied to represent local image characteristics. Subsequently, the extracted features are used to train a classification model for pixel-level classification to achieve image segmentation. Since 2012, based on the idea of deep convolutional neural network (DCNN) proposed by LeCun [8], the performance of computer vision systems have been significantly improved in a wide range of applications.

In this section, deep learning based image segmentation methods are reviewed as in this thesis all the proposed methods are based on deep learning based methods. At the end of the section, domain adaption based methods and shape aware techniques are reviewed.

2.2.1 Deep Learning

Since 2012, based on the idea of convolutional neural network (CNN) proposed by LeCun et al. [81] and followed by a technological breakthrough that allows deeper neural networks to be trained [48], DCNNs have demonstrated remarkable capabilities in performing classification, segmentation, object detection, and other image processing tasks [82][49]. Briefly, the CNN-based methods recognise objects based on a multi-scale feature representation obtained by applying many convolutional filters and non-linear activation functions at different image scales. The parameters of the convolutional filters are automatically learned during the training process through iterative back propagation of the errors between the predicted outputs and the ground truth images. This enables an automatic feature learning and representation, which is the key advantage against classical machine learning methods that are based on manually designed features. A DCNN is designed to better utilise spatial

information by taking 2D or 3D images as input. Moreover, a DCNN exploits three mechanisms, local receptive field, weights sharing, and subsampling. The architecture of a DCNN is comprised as a set of layers after each other, such as convolutional layers and pooling layers. An example of DCNN architecture is shown in Figure 2.4.

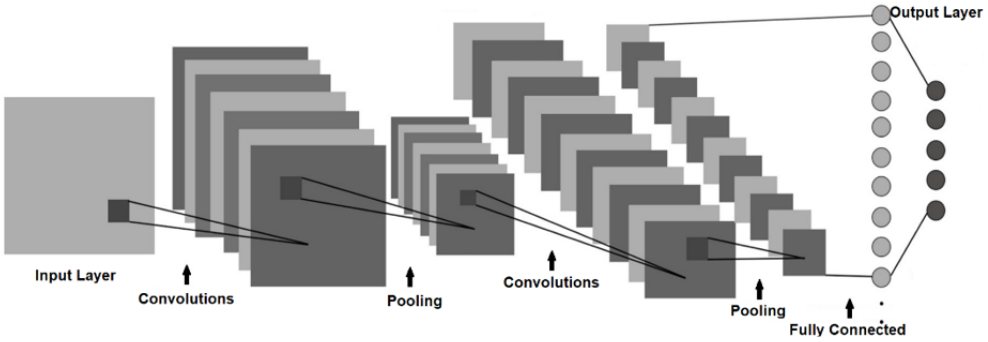


Figure 2.4: *The architecture of CNN [8]*

In earlier approaches, image segmentation was treated as a pixel-wise classification problem [83], and DCNN classification models were trained in a patch-based manner. These methods required millions of image patches for training and suffered from low computational efficiency in both training and testing stages. A popular image segmentation model is based on an encoder-decoder structure that is able to be trained more efficiently and require less training samples than the patch-based models. The encoder extracts image features from the input image, and the features are mapped to the low-resolution version. In the decoder, transpose convolution layer (deconvolution) is applied to upsample the feature map to the original image resolution.

Fully convolutional network (FCN) [9] is an encoder-decoder based architecture. The input image to the FCN can be in any size, and the produced output is in the same size of input. FCN includes pixel-by-pixel classification on the upsampled feature map to complete the segmentation task. FCN avoids the use of dense layers, thus, less parameters are required and the networks are faster to train. The architecture of a FCN is shown in Figure 2.5.

Although FCN achieves a superior performance in segmentation task, the

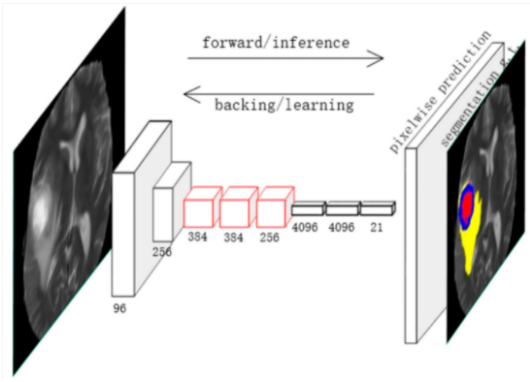


Figure 2.5: *The architecture of FCN [9].*

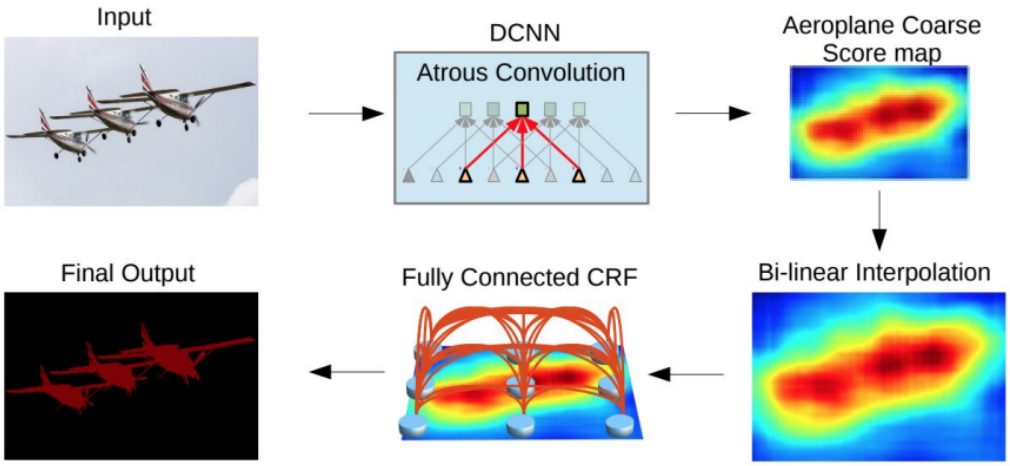


Figure 2.6: *The architecture of DeepLab [10]*

results of upsampling are not precise enough in terms of covering the details of the image because of continuous usage of down sampling. Furthermore, spatial consistency between pixels needs to be considered to have a more accurate segmentation. In DeepLab proposed by Chen et al. [84], Atrous convolutions are used to control the resolution and enlarge the field of view of the filter by remaining the number of parameters and computational time. Atrous convolution is a convolution with up-sampled filters that insert holes ('trous' in French) between nonzero filter taps. The backbone of DeepLab is the VGG-16 network. Beside, DeepLab uses conditional random field (CRF) [85] to improve the accuracy of segmentation boundaries at the end. The architecture of DeepLab is shown in Figure 2.6. An improvement on DeepLab is presented in [10]. The new version of DeepLab is more flexible by proposing different

sampling rates on an input image to easily segment the same objects with different sizes.

SegNet [11] is another encoder-decoder based structure for semantic segmentation. The novelty of SegNet is in the decoder path to map the low resolution encoder feature maps to full resolution feature map using non-linear upsampling. SegNet is an efficient structure as it only stores the max-pooling indices of the feature maps of the corresponding encoder layers and uses them in the decoder network. The overal scheme of SegNet is presented in Figure 2.7.

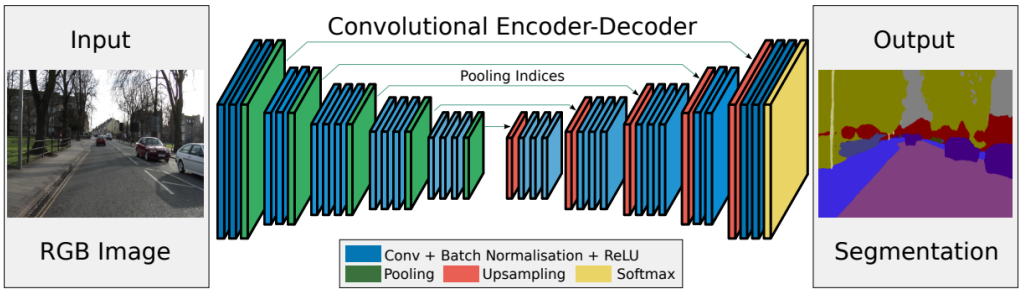


Figure 2.7: *The architecture of SegNet [11]*

One of the latest state-of-the-art methods known as U-net [33] is based on an end-to-end DCNN architecture. U-net consists of a contracting (down-sampling) path to extract image features in multiple image scales and an expansive (up-sampling) path to produce a high resolution pixel-wise segmented image. Figure 2.8 shows the structure of U-net. Following on this pioneer work, several improvements and modifications have been proposed. For instance, Drozdzal et al. [86] added short skip connections in addition to the long skip connections in the U-net to improve training efficiency and segmentation accuracy. Several other researches extended the 2D U-net to address 3D volume segmentation problems [87][88]. In Res-UNet [89] and H-DenseUNet (hybrid densely connected UNet) [90] each sub-module of U-Net is replaced with a residual connection and dense connection respectively. Ibtehaz et al. [91] modified U-net by using MultiRes block and Res path instead of two convolutional layers and the skip connections. The results on public medical image

data sets showed that the proposed architecture achieved a high accuracy. Oktay et al. [92] proposed Attention U-Net by adding the attention mechanism to U-Net structure. This method requires a large amount of parameters and calculation as there is a need of an attention module for the output of each encoder layer before connecting to the corresponding decoder layer.

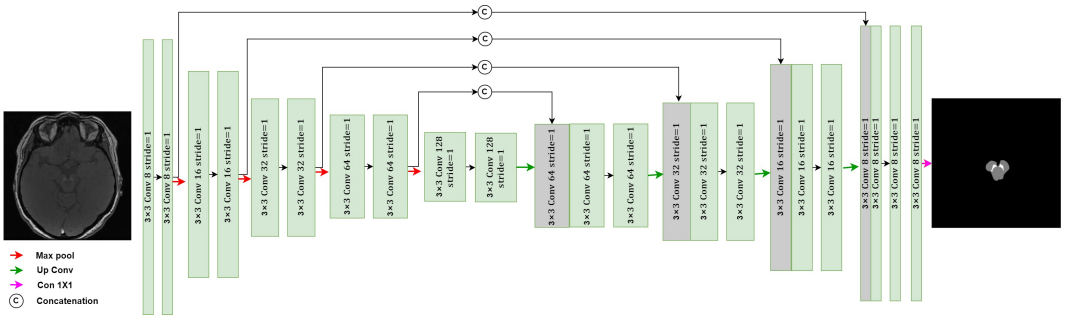


Figure 2.8: *The architecture of U-net [12]*

2.2.2 Domain Adaptation Methods

In addition to having a sufficient amount of annotated data, the success of supervised models requires sampling of the training and testing images come from the same data distribution. However, in medical imaging contexts, there are various imaging modalities, scanners and protocols that capture various physical properties and provide different information for disease diagnosis and treatment. Many approaches have been proposed with different strategies to address the problem of domain shift. In this section, two families of the best performed domain adaptation approaches, i.e. image and feature level adaptation, are reviewed.

The first set of solutions is image-to-image translation based on a cycleGAN foundation [93] by remapping the distribution from the source domain to the target domain or vice versa to tackle the domain shift in image level [94] [95] [45]. These pixel-level approaches tend to reduce the marginal distributions of the two domains as much as possible. However, it does not preserve the semantic consistency, hence additional constraints need to be added. Hoffman *et al.* by proposing semantic consistency loss functions [96],

and Zhang *et al.* by adding shape constraints [44] [94], achieved a better pixel-pixel mapping at a semantic level. Moreover, Jiang *et. al* [97] proposed shape and feature-based losses to preserve tumor features in a segmentation task. Firstly, they used an unsupervised tumor-aware adaptation approach to synthesise target images through adversarial training. Secondly, they applied a combination of synthesised and real images with corresponding labels to a U-net for tumor segmentation.

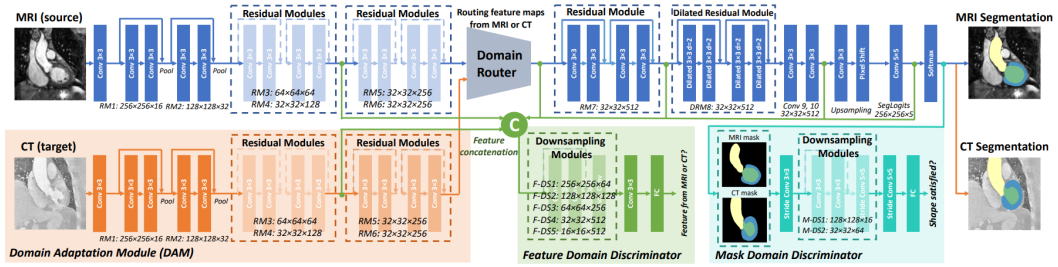


Figure 2.9: *The architecture of PnP-AdaNet[13]. As it is shown, this method is complex and includes a lot of learning modules.*

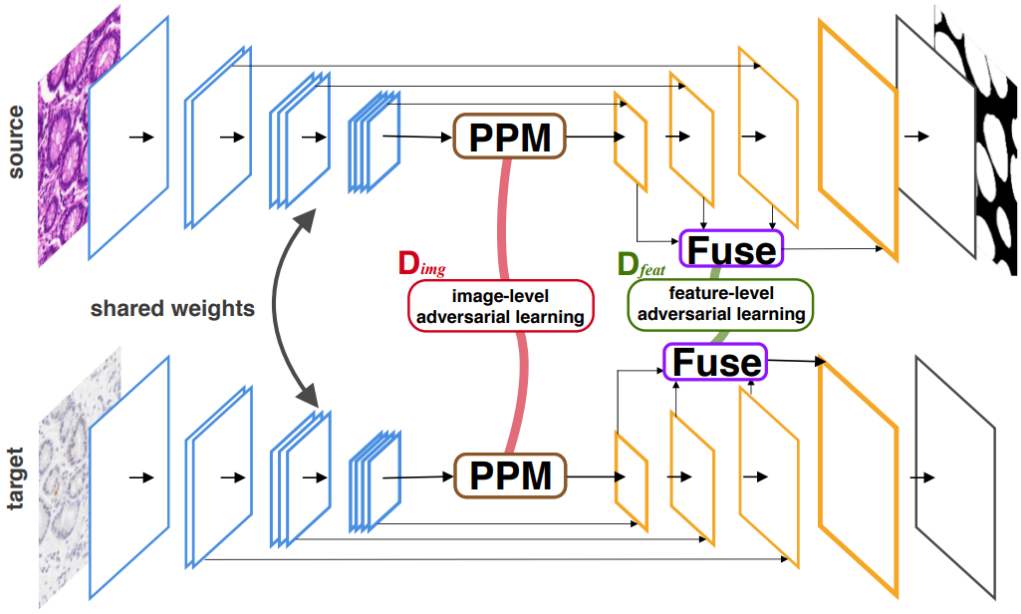


Figure 2.10: *The architecture of DAPNet[14]*

Image level translation from source to target domain is time consuming. Alternatively, learning a unified feature space of the source and target domain is more effective. Feature space alignment learns a domain-invariant feature representation by aligning the latent feature spaces of source and target

domains in an unsupervised manner via adversarial learning [98] [99], cross-domain reconstruction [100], or divergence minimisation [101] [102]. In medical imaging, feature alignment via adversarial domain adaptation has been commonly used for different applications. For instance: in PnP-AdaNet [13], the early layers of a segmentation encoder, which are trained based on the source images, are replaced during target training, but the higher layers are fixed and reused. PnP-AdaNet includes two discriminators and requires three different stages for training, which is time consuming. The overall scheme of PnP-AdaNet is represented in Figure 2.9. Hou *et. al* [14] proposed DAPNet that tackles domain shift using both multi-resolution feature-level and image-level losses instead of only using pixel-level information. The feature level loss minimises the spatial inconsistency of the two domains, and the image level loss attempts to minimise image color and style differences of the two domains. The two components are implemented as domain classifiers with adversarial training. The architecture of DAPNet is shown in Figure 2.10. Wu and Zhuang [103] presented a feature-based cross-modality segmentation approach by proposing a new metric which explicitly computes the domain discrepancy using characteristic functions of distributions. In another paper, Wu and Zhuang [104] proposed to approximate a common probability distribution based on the distributions of the latent features of both domains. This is achieved using two networks based on variational autoencoders (VAEs) and a regularisation loss for this variational approximation. SIFA [15] is an end-to-end method that simultaneously aligns both image and feature levels. Firstly, source images are transferred to the appearance of the target data, using generative adversarial networks with a cycle-consistency loss. Next, a segmentation model is trained using the combination of synthesised and target images. Three different discriminators are embedded in the SIFA architecture. The overall scheme of SIFA is shown in Figure 2.11

Another idea is to preserve semantic feature-level information by disentangling medical images into content space (i.e. shape) and style space (i.e.

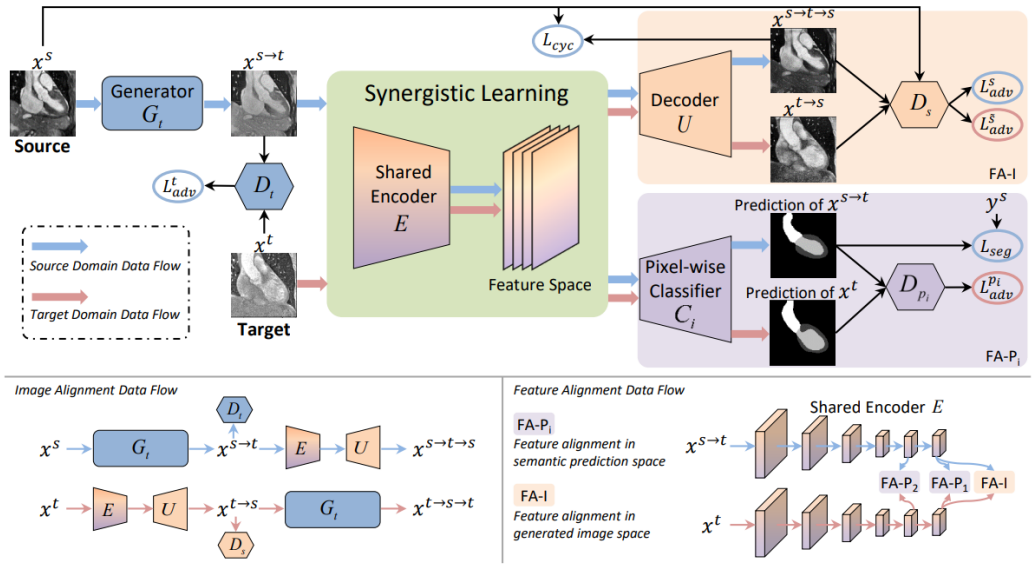


Figure 2.11: *The architecture of SIFA[15]*

appearance) [46] [47][105]. For example, Pei et al. [105] disentangles each domain into the domain-invariant and domain-specific features. To improve the quality of the generated images and segmentation results, self-attention modules, cycle consistency idea, and several discriminators are used resulting in a complex network architecture that is difficult to be extended to a multi-domain situation.

In summary, model complexity is one of the biggest limitations of the aforementioned methods. As these frameworks normally consist of multiple network modules and various loss functions of adversarial learning, applying such a complicated framework to 3D images requires a large amount of memory and tremendous time for training. Parameter tuning for training such networks is not a trivial task. Therefore, the development of a simple, effective and accurate method to address the cross modality image segmentation problem, especially for 3D medical images, is still demanding.

2.2.3 Shape Preservation Method

As it is mentioned before, segmentation methods suffer from relying on a pixel-level prediction [32] and sub-sampling resulting from stacked layers of convo-

lutions and pooling. This may lead to ignoring higher-order properties, such as boundary smoothness and the topological shape/label interactions [106]. Various strategies have been explored to preserve object boundaries such as trainable up-sampling layers using deconvolution operations; however, they only can recover some of the object boundaries. Conditional random field (CRF) [85] is another approach to increase the sharpness of the output; however, this method requires extra computational costs. There are some other works, which incorporate topology priors into segmentation pipeline which are described in the following.

Early works in the context of neural networks on shape analysis regardless their success come at the cost of increased model complexity and computational needs. Such as ShapeBM [107] that has focused on deep Boltzmann Machines (DBMs) with sparse pixel connectivity. However, this model requires setting a large number of parameters and results in model training becoming uncontrollable depending on the size of images. Convolutional deep belief nets [108] are proposed for encoding shape prior information. Variational models [109][110] are applied to discover priors on the shape and structure in label space. Recently by using simple approaches, shape priors are applied, namely adjacency [111] and boundary conditions [112]. Oktay et al. [113] integrated global shape information into a neural network by using PCA based statistical [114] and active shape models [115]. In this framework, a stacked convolutional auto-encoder is utilised to learn a non-linear representation of the underlying structure and enforces network predictions to follow the learnt statistical shape distributions. A similar work is proposed by Avendi et al. in [116] that uses an auto-encoder to infer the shape of the objects, and finally uses deformable model to achieve segmentation of the target objects. Another method for localisation of shapes using a deep network is proposed by Mansoor et al. [106] where the final segmentation is performed using a statistical shape model. The method proposed in [117] uses a CNN model; at first, the network is trained in a supervised manner. In the next step, the network is fine-tuned by using

unlabeled data where the ground truth is generated with the help of a level set-based method. A deep fully convolutional neural network, that is able to predict shapes instead of classifying each pixel separately is proposed by Al Arif et al. [118]. In this model a U-net architecture is modified to generate a signed distance function (SDF) from the input image, and converted to shape parameters, in which the loss is defined.

Regardless of the achieved success, almost all of the aforementioned methods suffer from the need of a lot of parameter setting and high computational complexity as most of them utilise a lot of different learning modules.

2.3 Performance Metrics

The performance of the segmentation methods can be evaluated using different metrics. The selection of an appropriate evaluation metric can be helpful to show how well the model performs based on various aspects. In this section, various performance metrics are provided to assess the segmentation performance of the models.

2.3.1 Dice

Dice score is a famous metric in assessing segmentation methods. Dice evaluates the number of pixels correctly labelled and determines the accuracy of the segmentation. Dice is described as follows:

$$Dice = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{2.3}$$

where true positive (TP) presents pixels correctly segmented as foreground, false positive (FP) represents pixels that falsely segmented as foreground and false negative (FN) indicates pixels incorrectly detected as background. The value range of Dice is betwwen 0 and 1. A Dice score close to 1 reveals a high quality segmentation and a score close to 0 shows a low quality segmentation

output.

2.3.2 Precision and Recall

The Precision measure, as expressed in the following equation, indicates the proportion of correctly predicted positive pixels over the total number of ground truth positive pixels. A larger number of FP decreases the Precision value, which results in over-segmentation.

$$Precision = \frac{TP}{TP + FP} \tag{2.4}$$

While Recall shows the proportion of actual positives that was identified correctly by the method. Larger FN results in a lower Recall value, indicating under-segmentation.

$$Recall = \frac{TP}{TP + FN} \tag{2.5}$$

2.3.3 Jaccard

The Intersection-Over-Union (IoU), also known as the Jaccard Index, is another commonly used metric in semantic segmentation. This metric is described as the ratio of the area of the overlap between the predicted segmentation and the ground truth segmentation to the area of union between the predicted segmentation and the ground truth segmentation.

$$Jaccard = \frac{TP}{TP + FP + FN} \tag{2.6}$$

Jaccard is very similar to Dice as can be described also as follows:

$$Jaccard = \frac{Dice}{2 - Dice} \tag{2.7}$$

The difference between Jaccard and Dice is that Jaccard tends to penalize

instances of bad classification more than the Dice score.

2.3.4 Average Symmetric Surface Distance

Average symmetric surface distance (ASSD) is used to quantify the dissimilarity of the segmentation output from the ground truth based on the distance definition. ASSD is the average of all the distances from points of the boundary of segmented output to the boundary of the ground truth. ASSD score of 0 mm indicates a perfect segmentation output.

2.3.5 Hausdorff Distance

Hausdorff distance is another performance metric based on distance to calculate the distance between two point sets. The average of this distance between the ground truth labels and the output of model can be calculated in millimeters or voxels. When the segmentation task involves with complex boundaries and small thin segments such as cerebral vessel segmentation, Hausdorff distance is suggested [119].

2.3.6 Statistical Tests

The paired t-test compares the mean from two groups (individual object, or related units). The t-test assumes that the mean difference between samples is normally distributed. Thus, the measured differences need to be normally distributed in order to verify whether the t-test is valid.

The Wilcoxon signed rank test is another statistical hypothesis test. This technique uses the sum of the signed ranks as the test statistic. Wilcoxon signed rank test in contrast to the dependent samples of t-tests is a non-paracontinuous-level test. The t-test investigates if the average difference between two observations is zero, while the Wilcoxon test investigates if the difference between two observations has a mean signed rank of zero.

In this thesis, dice, recall and precision continuously are used in all chapters

to assess the performance, over-segmentation, under-segmentation of proposed methods and compare with other methods. ASSD is used instead of Hausdorff Distance to evaluate the difference between predicted results and ground-truth based on distance metric as Hausdorff distance is especially recommended for segmentation tasks with complex boundaries and small thin segments. In some of the chapters to have a fair comparison with other state-of-the-arts, the same evaluation metrics as those state-of-the-arts are chosen.

2.4 Conclusions

Image segmentation is a fundamental and crucial step in many image analysis tasks. From classical image segmentation methods (e.g. region growing) to more robust methods (e.g. level-set [120] and graph-cut [121]), various techniques have been proposed to achieve automatic image segmentation in a wide range of clinical problems. However, in most of these methods the extraction of discriminant features from the images is still done manually by human. A logical solution is that computers learn the features automatically as many deep learning algorithms [8], but the classical methods and handcrafted based models are prone to failure because of an inaccurate representation of features. Deep Convolutional neural networks (DCNNs) are known as the most successful type of models for image analysis in deep learning methods. In this chapter, in addition to reviewing the classical approaches, a comprehensive review is provided on state-of-the-art deep learning based methods. Furthermore, the limitations of each method are described. According to the limitations and challenges mentioned, imbalanced number of class samples, domain shift, shape preservation, etc, in the following chapters some solutions are proposed. The proposed methods are compared with other recent methods based on the evaluation metrics introduced in this chapters.

Chapter 3

FU-Net: Multi-class Image Segmentation using Feedback Weighted U-Net

3.1 Introduction

Many deep CNN based methods have been proposed to address image segmentation tasks. In earlier approaches, image segmentation is treated as a pixel-wise classification problem [83]. Deep CNN classification models are trained in a patch-based manner. These methods require millions of image patches for training and suffer from low computational efficiency in both training and testing stages. One of the latest state-of-the-art methods (known as U-net [33]) is based on an end-to-end deep CNN architecture. It is trained more efficiently and requires fewer training samples than the patch-based models. Following on this pioneer work, several improvements and modifications have been proposed. For instance, Drozdzal et al. [86] added short skip connections in addition to the long skip connections in the U-net to improve training efficiency and segmentation accuracy. RU-net and R2U-net, proposed by Alom et al. [122], are based on U-net plus recurrent neural network and U-net plus the combination of recurrent neural network and residual network respectively. A

nested U-net architecture called U-net++ is introduced in [123] that is proposed to replace the direct skip connections from encoder to decoder part by dense skip connections. A chain of multiple U-nets are utilised in LadderNet [124] to improve the flow of information.

Most of the recent methods encounter the problem of class imbalance during training, as usually only a few locations per image contain objects. Therefore, the training could be dominated by a lot of easy examples over a small number of hard examples. Initially, an active learning formulation [125] is proposed to select its target with fewer data samples to learn from rather than random sampling. Another idea is focus-of-attention [36] mechanism to discard background regions of the image, while spending more computation on promising object-like regions. R-CNN [126] and Fast R-CNN [127] extract thousands of RoIs per image that probably includes objects, but only a small number of them are randomly selected and passed to the next stages.

Online hard example mining (OHEM) [128] algorithm can measure contributions of each RoI, reweigh and sample most beneficial ones for training in each iteration. OHEM automatically selects the hard examples by using stochastic gradient descent in which training examples are sampled according to a non-uniform, and non-stationary distribution of the loss value of each example. However, for real-time scenarios, an RoI-free hard example mining strategy is desired to have a higher speed. In an ROI-free model [129] the negatives and positives are sorted using the highest confidence loss. Next, the top ones are selected by considering the ratio 3:1 between the negatives and positives. Focal loss is a loss function to cope with class imbalance without sampling in an one-stage object detection scenario. Focal loss can dynamically contribute less on easy examples during training and focus more on the hard examples by modifying the cross-entropy loss function. There is still a need of a light-weight, non-ROI based and dynamically weighed scheme for segmentation task as most of the methods focus on the problem of object classification, while the application to image segmentation has not been thoroughly investi-

gated.

As the main contribution of this chapter, the U-net method is improved by introducing a dynamically weighted cross-entropy loss function. The weight for each pixel is calculated based on the predicted accuracy in each iteration. The pixel locations with higher prediction accuracies are assigned with lower weights, and vice versa. This enables the network to learn more from poorly predicted image regions. The proposed method is named as feedback weighted U-net (FU-net). The effectiveness of the FU-net is demonstrated using Skin lesion data, Retinal data, and a a challenging brain magnetic resonance image (MRI) dataset with extremely unbalanced classes as well as different numbers of training samples. This method is published in international conference on image and graphics, and is titled as FU-Net: Multi-class Image Segmentation Using Feedback Weighted U-Net [130].

3.2 Methodology

FU-net is designed to tackle the imbalanced issue between foreground and background classes during training. In the following sub-sections, the network architecture and the proposed weighting loss function are described respectively.

3.2.1 Network Architecture

The U-net proposed by Ronnebergeret et al. [33] is based on convolutional neural network, and consists of a contracting path and an expansive path. In the contracting path, each layer consists of two 3×3 convolutions (Conv), and each convolution is followed by a rectified linear unit (ReLU) as illustrated in Figure 3.1(a). The feature map in the next successive layer is a down-sampled version of the output from the previous layer by using a max pooling of stride 2. Due to the down-sampling process, only very abstracted information remained at the end of the contracting path. To capture and rebuild the spatial context,

a decoding path is required. In the expansive (decoding) path, the output feature map in each layer is up-sampled using 2 × 2 up-convolution with halved number of feature channels in the previous layer. Each layer also has two 3 × 3 convolutions, and each followed by a ReLU. Additionally, there are some concatenation operations to combine feature maps from the contracting layers to the corresponding expansive layers. 1 × 1 convolution is used in the final layer to convert the dimension of feature maps to the number of classes. Subsequently, softmax function [22] is applied to map the output value of each pixel to the range of [0, 1]. In the U-net paper [33], the authors proposed a weighted cross entropy loss function for parameter optimisation that is expressed in Eq. (3.1).

$$L_{seg} = -\frac{1}{N_s}\sum_{i=1}^{N_s}\sum_{c=1}^{C} y_s^{i,c} w(x_s) log(\hat{y}_s^{i,c}) \tag{3.1}$$

where c and i repeat over all classes and pixels. $\hat{y}_s^{i,c}$ is the predicted output, and $y_s^{i,c}$ is a one-hot encoding representation of the corresponding ground truth label for the input sample image x_s. $w(x_s)$ is the weight for pixel x_s. In Unet paper [33], the weights are pre-calculated by assigning higher values to challenging boundary pixels based on a distance map. The weights are pre-determined and application dependent.

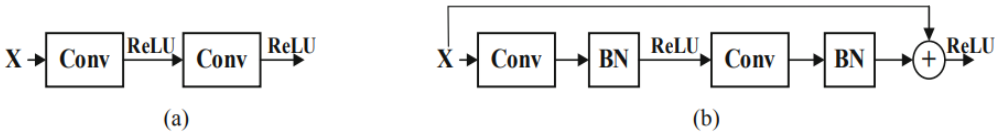


Figure 3.1: *(a) Layer of the original U-net. (b) Layer by adding batch normalisation (BN) and residual block (RB).*

3.2.2 Batch Normalisation and Residual Network

Batch normalisation (BN) [131] has been shown to allow the training of deep networks converging faster with higher learning rates, which is described as follows:

$$x'_n = \frac{x_n - \mu_\beta}{\sqrt{\sigma_\beta^2 + \epsilon}} \tag{3.2}$$

x_n is the intensity values of the nth image in the mini-batch, where $n = 1 \dots M$ and M are the number of images in the mini-batch. μ_β and σ_β are the mean and standard deviation of the mini-batch images. A constant value, ϵ, is used to achieve numerical stability. In the proposed network architecture, BN is applied right after each convolution operation and before the ReLU.

Residual network block (RB) [40] uses some skip short connections that enables flow of information across layers. It allows the parameters to be optimised much easier even in a very deep network. RB comprises multiple convolutional layers in which each layer consists of a residual module f and a skip short connection. By considering x_t as is the input, the output of the $t + 1th$ block is defined as:

$$x_{t+1} = f(x_t) + x_t \tag{3.3}$$

where $f(x_t)$ is a sequence of convolutions, BN, and ReLU. In this chapter, the residual block is employed before the ReLU and after the BN in the second convolution in each layer. Figure 3.1 (b) shows a modified U-net layer with the combined BN and RB. BN and RB are well-known techniques to achieve faster convergence and train deeper networks [88]. More importantly, $w(x_s)$ is automatically calculated in the proposed method. The weights are pixel-wise values which are iteratively updated in each training iteration for each training image. Calculation of the weight is introduced in next sub-section.

3.2.3 Weighted Cross-Entropy Cost Function

As it is mentioned before, there is an imbalanced class samples that suppress the cross entropy loss, and easy examples could dominate the learning process. In other words, the object class with a larger number of pixels contributes

more to the cross-entropy loss calculation and has larger influence on the gradient values for parameter optimisation. Therefore, a method is proposed for automatically calculating the weight $w(x_s)$ in Eq. (3.1) to dynamically weight easy and hard examples.

Abraham and Khan [132] and Wang et al. [133] have applied dice coefficient loss to address the class imbalance issue. A Weight calculated based on the number of pixels per class has also been proposed [82]. Different from these methods that use fixed weight calculations, the weights are dynamically calculated according to the predication performance in each iteration. The motivation is to increase the contribution from pixels that have larger prediction errors to the loss function calculation. This not only enables the balance of different classes implicitly, but also allows difficult local image regions to be emphasized for model training. A pixel-wise weight map is generated based on the pixel-wise probability values that are produced in each training iteration. The pixel locations with lower prediction accuracies are assigned to higher weights and vice versa. Hence, the network is able to focus on learning from poorly predicted image regions. The feedback weight is a continuous function that maps the input values to the range of [0.01 1], which is expressed as:

$$w(x_s) = e^{-log_{100} \times (\hat{y}_s^{i,c})^{\beta}} \tag{3.4}$$

In Eq. (3.4), larger values of $\hat{y}_s^{i,c}$ indicate higher predicated probability values of the true class, which are assigned to lower weights for calculating the loss function for network backpropagation. Figure 3.2 shows the behaviors of the weighting functions by varying the hyperparameter β in Eq. (3.4). β is experimentally determined in the next section. Note that $log100$ is used to constrain the minimum weight to be 0.01 instead of 0, which prevents the pixels with high prediction accuracies being completely neglected from training.

Note that the same mapping function in Eq. (3.4) is applied to all training images in a mini-batch [134] for parameter optimisation. In each batch, a

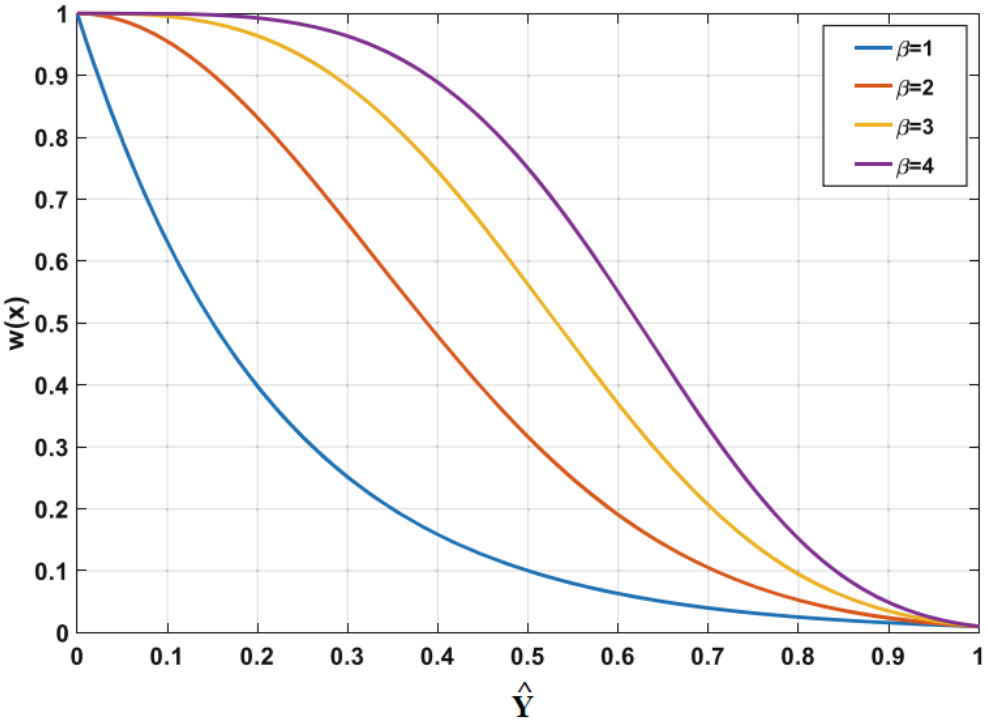


Figure 3.2: *Plot of the mapping function of Eq.* (3.4) *with different values of* β

training image with larger poorly predicted regions contributes more than an image with a higher prediction accuracy as the weights are updated after calculating the loss of a batch. This effectively not only balances the image regions but also balance the 'easy' and 'difficult' training examples. This is particularly beneficial for model training based on a small number of training examples with certain bias. This advantage is demonstrated by varying the size of the training data in the evaluation section.

3.3 Method Evaluation

In this section, the proposed method is evaluated based on three different datasets, i.e. Skin lesion, Retinal, and Brain datasets.

3.3.1 Datasets and Preprocessing

Skin Data

Skin Lesion Analysis Toward Melanoma Detection grand challenge dataset [1] consists of 10015 RGB images of skin lesions with binary annotations. Cases include a representative collection of all important diagnostic categories in the realm of pigmented lesions. Some examples of this dataset are presented in Fig. 3.3

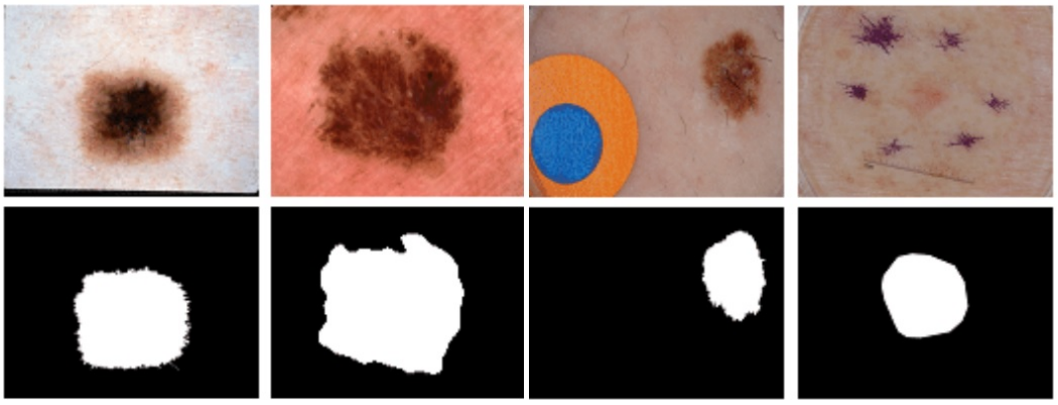


Figure 3.3: *Examples of Skin datsets, the images and the ground-truths [1]*

For the experiments, the images were firstly converted to gray level and resized to 192×256 pixels with a split of 70%-10%-20% for training, validation and testing (same as in [132]).

Retinal Data

The Retinal database [2] contains 40 Retinal images with blood vessel segmentation obtained manually by experts. The images are collected in the Netherlands from a diabetic retinopathy screening program. There is no sign of diabetic retinopathy in 33 of the images. There are signs of mild early diabetic retinopathy in 7 of them. Each image was captured and stored in a RGB color format. A pre-processing step is performed on the images by converting the color images to gray level images, and and resized to 512×512 pixels with a split of 70%-10%-20% for training, validation and testing.

Brain Data

T1-weighted brain MRI for segmentation of Midbrain (MB) and Substantia Nigra (SN). Certain quantitative measurements (e.g. volume) of SN has been found to associate with Parkinson disease [4]. However, it is extremely time consuming to annotate it manually and it is challenging to train a machine learning model for automatic segmentation due to the small size of SN. The T1-weighted brain MRI data were acquired in Nottingham University Hospital and was approved by the local ethics committee for this research. The dataset contains a total of 102 subjects with 30 axial image slices each. Experienced radiologist manually selected 3 or 4 slices that contain both the MB and SN, and annotated the contours of MB and SN. This resulted in a total of 310 2D slices for the segmentation evaluation.

3.3.2 Experimental Results

Skin and Retinal Datasets

First, original U-net, U-net with batch normalisation and residual block (BRU-net), BRUnet with feedback weight (FU-net) are compared with each other on Skin and Retinal datasets in Table 3.1 and Table 3.2 respectively. The dice coefficient, Precision and Recall measures are applied as the evaluation criteria. Note that the separate effects of adding batch normalisation and residual block were not tested, as they were normally used simultaneously to achieve better performance.

The parameters for these Skin and Retinal datasets are as follows. The batch size was 4. The optimiser was Adam [135] with learning rate of 0.001. The number of feature channels in the first layer was 16. No dropout was applied, and the number of epochs was 200. The number of iterations within each epoch for the three experiments was equal to number of training images divided by batch-size. The value of β was experimentally set to 3 using the validation set.

According to Table 3.1 and Table 3.2, FU-net achieved a significantly better results according to wilcoxon signed rank test $P < 0.01$ on both datasets (i.e. the dice score 0.8217 on Skin data and 0.7631 on Retinal data) in compare with other methods.

Table 3.1: Comparison of U-net, BRU-net and FU-net on skin data. The mean dice coefficient values, precision, and recall index are reported.

Method	**Dice**	**Precision**	**Recall**
U-net	0.7414	0.7575	0.7585
BRU-net	0.7980	0.8225	0.8407
FU-net	0.8217	0.8636	0.8408

Table 3.2: Comparison of U-net, BRU-net and FU-net on Retinal data. The mean dice coefficient values, precision, and recall index are reported.

Method	**Dice**	**Precision**	**Recall**
U-net	0.6778	0.6431	0.7349
BRU-net	0.7439	0.7365	0.7530
FU-net	0.7631	0.8610	0.6961

Brain Dataset

To investigate the proposed method more comprehensively, another set of experiments are prepared on brain data when the size of training data varies. Therefore, three experiments for each method are performed as follows: randomly selected 200/100/50 images for training, 10 images for validation and the remaining 100/200/250 images for testing. The parameters for model training are listed as follows. The batch size was 5. The optimiser was Adam [135] with learning rate of 0.001. The number of feature channels in the first layer was 16 and doubled in each of the down-sampled layers. The dropout rate was 0.25, and the number of epochs was 400. The number of iterations within each epoch for the three experiments were 40, 20 and 10 respectively (corresponding to experiments with 200, 100 and 50 training images). The performances is evaluated by varying the hyperparameter β (Eq. (3.4)) from 1 to 4 for the 100 training/200 testing experiment. When $\beta = 3$, it achieved

the best performance. Hence, $\beta = 3$ was used for all the remaining experiments. The main aim of the evaluation is to compare the performances of the proposed improvements rather than achieve an ultimate performance for a particular medical application. Hence, data augmentation was not used.

Table 3.3 lists the numerical results of the mean Dice ± standard deviation (Std) of the three methods by varying the number of training samples. The P values of paired t-test are also reported by comparing U-net with BRU-net and BRU-net with FU-net respectively.

Table 3.3: Comparison of different methods using different number of training samples. The mean dice coefficient (Dice) ± standard deviation (Std) and P values of paired t-tests are reported. Numbers in bold indicate the best method that performed statistically ($P < 0.01$) better than other methods.

Number of training / testing	**Method**	**Mean of Dice ± Std**	
		Midbrain (MB)	**Substantia Nigra (SN)**
200/100	U-net	0.9000 ± 0.03	0.7095 ± 0.17
	BRU-net	0.8775 ± 0.14	0.7164 ± 0.18
	FU-net	0.8929 ± 0.05	$\mathbf{0.7563 \pm 0.15}$
100/200	U-net	0.8584 ± 0.18	0.7022 ± 0.18
	BRU-net	0.8550 ± 0.16	0.7005 ± 0.15
	FU-net	0.8710 ± 0.15	$\mathbf{0.7575 \pm 0.16}$
50/250	U-net	0.8135 ± 0.19	0.4831 ± 0.26
	BRU-net	0.8088 ± 0.15	0.6387 ± 0.17
	FU-net	0.8182 ± 0.20	$\mathbf{0.7087 \pm 0.24}$
	P values		
200/100	U-net/BRU-net	0.0589	0.6086
	BRU-net/FU-net	0.1260	0.0026
100/200	U-net/BRU-net	0.6050	0.8706
	BRU-net/FU-net	0.6062	< 0.0001
50/250	U-net/BRU-net	0.5725	< 0.0001
	BRU-net/FU-net	0.3138	< 0.0001

It is seen from the results in Table 3.3 that all three methods achieved similar segmentation performance (no statistical significance) for the MB segmentation regardless of the number of training samples. However, for the SN class where the number of pixels is much smaller than the MB class and more difficult to be segmented, the proposed FU-net consistently outperformed the U-net and BRU-net methods for all the experiments with statistical sig-

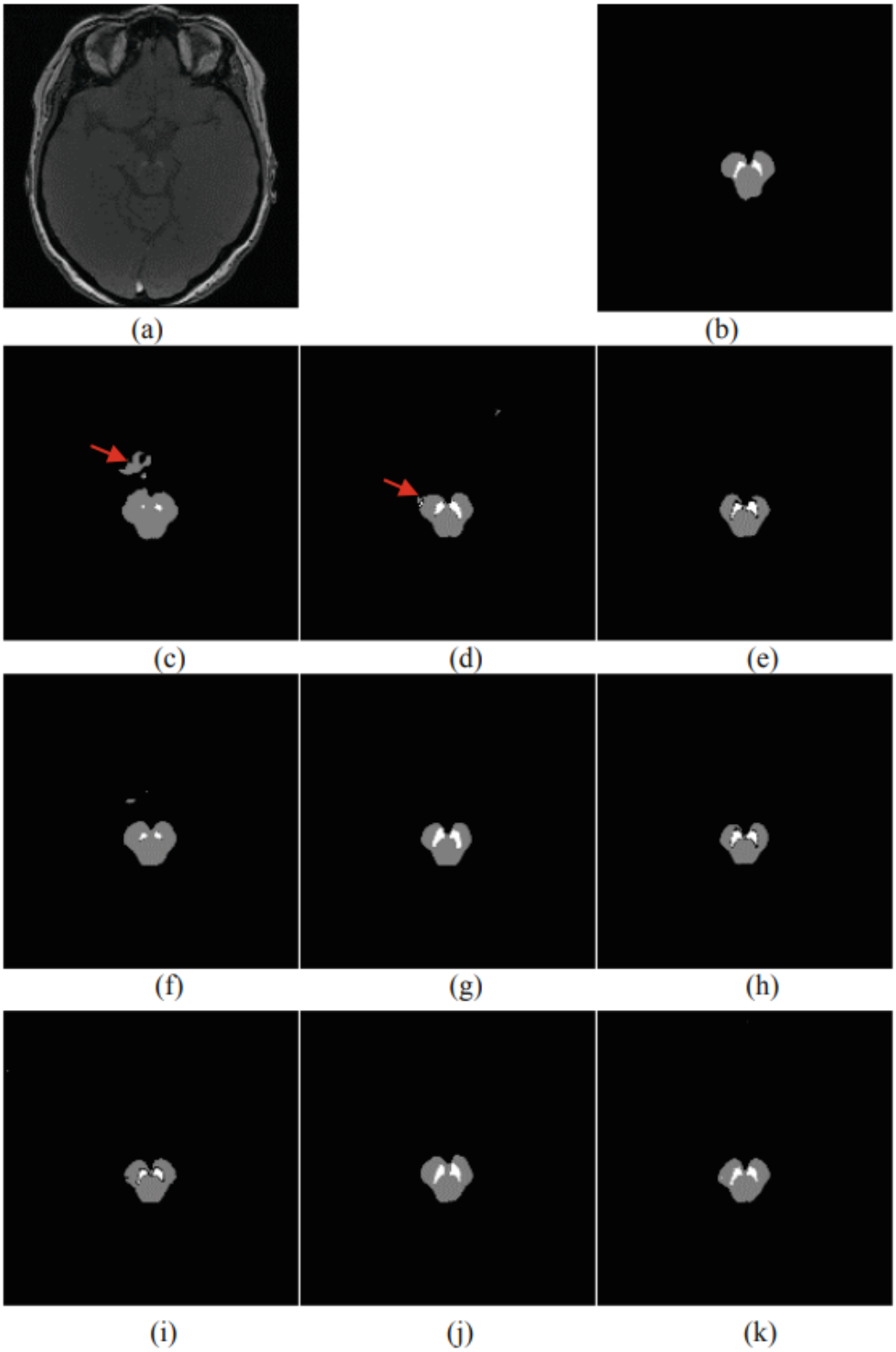


Figure 3.4: *(a) The original image. (b) The ground truth. (c) (d) (e) Segmentation results by U-net/BRU-net/FU-net respectively using 50 training samples. (f) (g) (h) Segmentation results by U-net/BRU-net/FU-net respectively using 100 training samples. (i) (j) (k) Segmentation results by U-net/BRU-net/FU-net respectively using 200 training samples.*

nificance. When training using only 50 images, the performance of FU-net remained high (Dice = 0.7087) which is much higher than the BRU-net (Dice = 0.6387) and U-net (Dice = 0.4831).

Some visual examples are also provided to demonstrate the advantages of the proposed method. In Figure 3.4, the segmentation results of an example image are presented based on 50, 100 and 200 training images. Figure 3.4(a) and (b) are the original image and ground truth annotation respectively. In Figure 3.4(b), the darker region is the MB and lighter region is the SN. Figure 3.4(c), (d) and (e) are the segmentation results for U-net, BRU-net and FU-net respectively using 50 training examples. Some false positives and false negatives can be easily identified for the U-net and BRU-net methods, as indicated by red arrows. Figure 3.4(f), (g) and (h) and Figure 3.4(i), (j) and (k) are the results for the three methods using 100 and 200 training examples respectively. Similarly, the FU-net results visually provide more similar outputs to the ground truth image than the other two methods. This is consistent with the numerical results reported in Table 3.3.

As it is show in the Fig. 3.4, the number of class samples are not balanced in different classes. In this dataset, the distribution of class for background is approximately 98.2%, class for MB is 1.62%, and class for SN is only 0.17%.

3.4 Conclusions and Discussion

In this chapter, the basic structure of U-net is adopted, and the cross-entropy cost function is improved by proposing a method to generate dynamic weight map. This method enables the prediction accuracy at each training iteration to be used for regionally focused training. The proposed method has been evaluated on Skin lesion, Retinal data, and a challenging multi-class brain tissue segmentation task. Based on the results, FU-net significantly outperforms the original U-net and an improved version of U-net (BRU-net). It is shown that FU-net is a generic and useful technique for model training with unbal-

anced class labels and with smaller number of training examples. It can be easily applied to any DCNN based segmentation framework as long as cross-entropy is used as the loss function. In addition to proposing a loss function to increase robustness and efficiency, there is still a need of a light-weight and more accurate network to be considered as the backbone model. Thus, in the next chapter, a novel network architecture is proposed based on two recent networks, i.e. Residual network (ResNet) and densely connected network (DenseNet) to improve the training efficiency and performance.

Chapter 4

DRU-net: An Efficient Deep Convolutional Neural Network for Medical Image Segmentation

4.1 Introduction

U-net [33] that is trained in an encoder-decoder architecture, has outperformed previous works in terms of the number of required training samples, memory and computational time. Several variants of U-net have been proposed, which mainly focus on improving segmentation accuracy and efficiency of feature information passing within and across layers. For instance, H-DenseU-net [90] combines the idea of densely connected network [34] and U-net for 3D liver and tumour segmentation. Alternatively, U-net++ [123] connects the encoder to decoder by using dense skip connections between different layers. MDU-net [136] is proposed to add three multiscale dense connections in U-net simultaneously, i.e. dense encoder, decoder and the connection between them. FU-net [130] modifies U-net by proposing a dynamically weighted cross-entropy loss function. Generally, based on previous studies, the encoder decoder DCNN architecture has shown a superior performance against other architectures due to its capability of capturing features in a multi-scale manner. Most methods

achieve better performance by integrating sophisticated network blocks (e.g. dense network [34], attention network [132], etc.), which require more parameters to be learned. However, it would be more desirable to reduce the number of model parameters and still achieve a similar or better performance. Therefore, in this chapter, the performance of residual network (ResNet) [40] and dense network (DenseNet) [34] is compared and further propose some modifications to improve the efficiency of the network.

As the main contribution of this chapter, a simple but efficient network block is proposed which is able to achieve better image segmentation performance than DenseNet, ResNet and an attention net based method. More importantly, the proposed method requires fewer model parameters than these methods (except for ResNet). This method is published in IEEE 17th International Symposium on Biomedical Imaging (ISBI), and is titled as DRU-Net: an efficient deep convolutional neural network for medical image segmentation [12].

4.2 Methodology

U-net [33] is chosen as the basic structure for performance comparison of different network blocks for image segmentation. Additional to the original U-net, a batch normalisation (BN) operation [131] is added to each layer, as it is a well-known strategy to achieve faster convergence and enable stabilized network training. Different from conventional Conv-ReLU operation, ResNet [40] adds a shortcut connection from the input to the output of a Conv-ReLU-Conv-ReLU process as illustrated in Figure 4.1-(a). Furthermore, DenseNet architecture [34], which has outperformed most previous networks (including ResNet) in ILSVRC challenge, consists of several interconnected dense-blocks. Each dense-block has a classical Conv-BN-ReLU process and a "bottle neck" block, as illustrated in Figure 4.1-(b). The input of each dense-block is a concatenated feature map of the outputs of all previous dense-blocks together with

the original input of the first dense-block, as illustrated by the top diagram in Figure 4.1-(b).

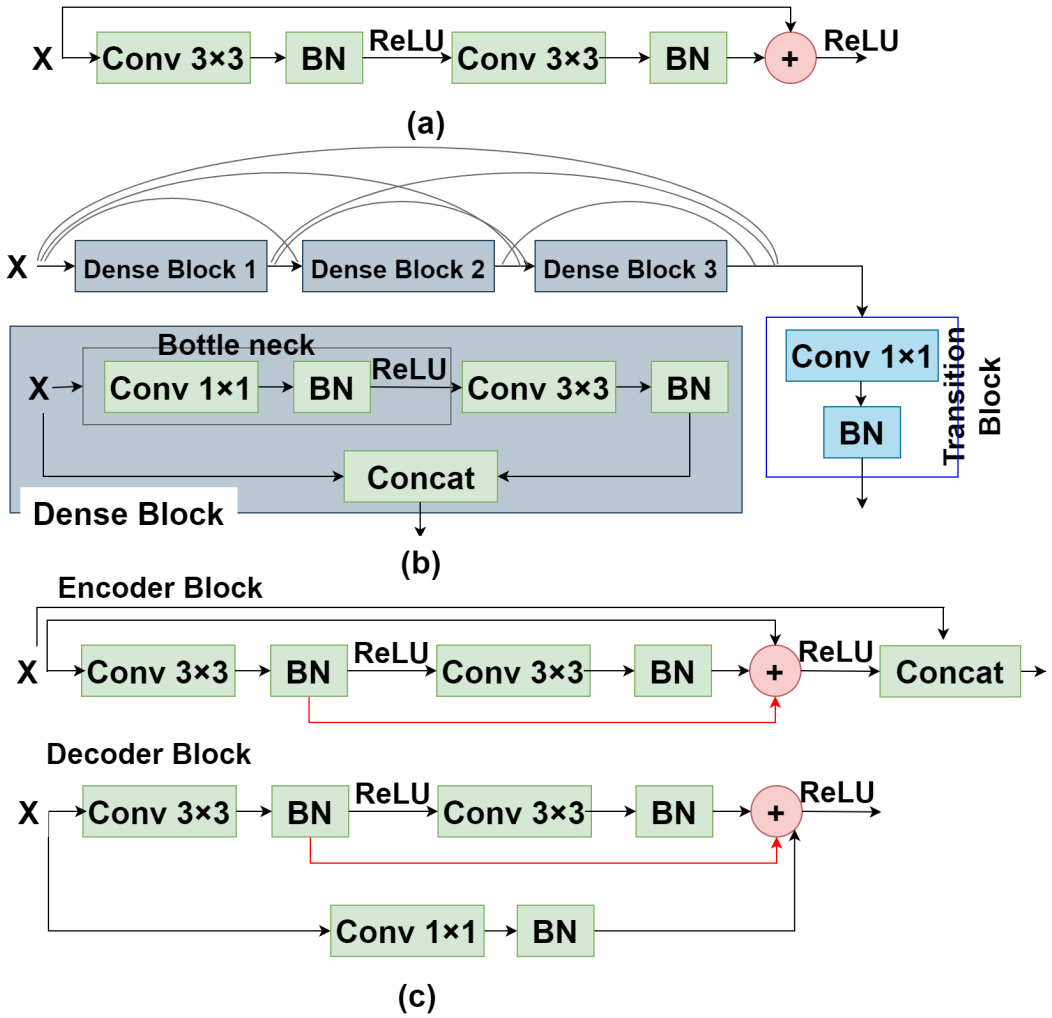


Figure 4.1: *a) Residual block. b) Dense net blocks. c) Encoder and decoder blocks in DRU-net.*

One drawback of ResNet, compared with DenseNet, is the lack of dense concantenation of all previous convolutional outputs to the successive feature maps. The original ResNet contains two Conv operations inside the skip connection (Figure 4.1-(a)). If gradients vanish at the second Conv during back-propagation, the parameters in the first Conv can not be updated neither. In contrast, DenseNet utilises information from all previous convolutional operations, which enables richer information to be combined for feature extraction in concecutive layers and flow through several paths. However, as shown in Figure 4.1-(b), it requires extra "bottle neck" and "transition" blocks to reduce

the feature map channels. This effectively performs feature map aggregation in a learnable way. However, these additional feature aggregation steps lead to extra parameters and heavier computational load, especially when several layers of DenseNet are used. By considering the advantages and disadvantages of these two networks, a novel method is proposed by adding an additional connection between the output of first Conv-BN operations to the last Conv-BN output (red shortcut connection in Figure 4.1-(c)) with a summation operation for feature map aggregation, as illustrated in Figure 4.1-(c). The justification behind the idea is that the additional shortcut connection allows the parameters to be updated in the first Conv even the gradients in the second Conv approach zeros. This is sufficient to allow the gradients to be backpropagated efficiently. Unlike the DenseNet that uses multiple learnable 1×1 Conv to aggregate feature maps, here a simple summation is used to combine feature maps (same as ResNet). It can be observed from the evaluation results that using these learnable aggregation operations in DenseNet cannot lead to a better segmentation performance but with more parameters and longer training time. Additionally, for each layer in the encoder path of the proposed method, a concatenation is applied to combine the input and output (see Encoder Block in Figure 4.1(c)). The combined feature map is then fed into the next layer. This concatenation is not only to include more information from the input of each layer but also to make the dimension of feature maps directly compatible with the feature maps in the next layer without extra parameters (unlike the transition block in DenseNet). Moreover, in the decoder block as shown in Figure 1-(c), a Conv 1×1 is used to reduce the number of channels of input instead of cropping it to make the dimension compatible to the layer output for a summation operation. In a general case, if more than two Conv-BN blocks are used, the outputs from all Conv-BN operations need to be added to the final Conv-BN output. This would not increase the number of parameters and computational time. By integrating the proposed encoder and decoder blocks into the U-net architecture, the proposed method is named as DRU-net

(Dense Residual U-net). Based on the above modifications, the overall scheme of DRU-net is illustrated in Figure 4.2.

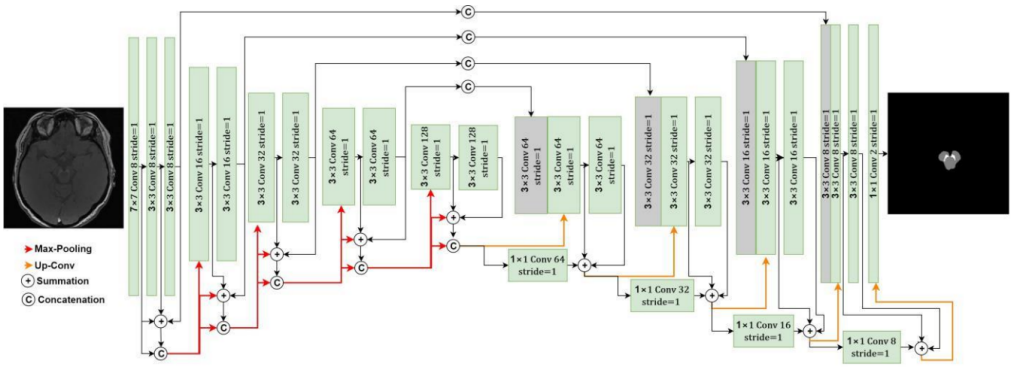


Figure 4.2: *The overall scheme of DRU-net.*

4.3 Method Evaluation

In this section, material, experimental design and network parameters are described. Based on ISIC 2018 challenge dataset for lesion segmentation [1], the proposed DRU-net is compared with the original U-net [33], ResNet-based U-net (RU-net), DenseNet-based U-net (DU-net) and an attention network based method [132] (AttnU-net) in terms of segmentation accuracy measured by Dice coefficient, Jaccard, precision and recall. Besides, a challenging brain MRI dataset is utilised to demonstrate the performance of the proposed method by varying the size of training data in a multi-class segmentation task. At the end of this chapter, the proposed method is also assessed on two different modalities of a local kidney 3D MRI dataset.

4.3.1 Datasets and Preprocessing

ISIC 2018

Skin Lesion Analysis Toward Melanoma Detection grand challenge dataset [1] consists of 2594 RGB images of skin lesions with binary annotations. For the experiments, the images were firstly converted to gray level and resized

to 192×256 pixels with a split of 70%-10%-20% for training, validation and testing (same as in [132]).

Brain Data

T1-weighted brain MRIs for segmentation of Midbrain (MB) and Substantia Nigra (SN) [4] were acquired in Nottingham Hospital and were approved by the local ethics committee for this research. The dataset is composed of 102 subjects each has 30 axial image slices. 3 or 4 slices were manually selected by a radiologist that contain both the MB and SN. A total number of 310 2D slices are available for the experiments. An example pair of original brain image (input) and the segmented image (output) is shown in Figure 4.2. In the segmented image, three classes are presented: background (black), MB (gray), and SN (white). This dataset is challenging due to the extremely unbalanced number of pixels in each class. Three experiments are performed for each method using different training/ testing sizes. 10 images were randomly extracted first for model validation. The remaining 300 images were randomly split into 200 training/ 100 testing, 100 training/ 200 testing and 50 training/ 250 testing. This allows the comparison of different methods when the size of training data varies.

Kidney Dataset

Balanced turbo field echo (bTFE) coronal localizer scans were used to quantify kidney volume. Kidney volume masks are generated by manually tracing the kidney on the bTFE localizer images (Analyze9VR, AnalyzeDirect, Overland Park, KS, USA). In total, 342 3D volumes are available which include 175 3T, 83 1.5T, 20 AKI (acute kidney injury is a sudden episode of kidney failure), and 64 CKD (chronic kidney disease is a long-term condition where the kidneys are not able to work well), all in 3D version were available. Furthermore, the model is also evaluated on a publicly available CT data from the KiTS19 challenge in MICCAI 2019 [137]. The CT dataset includes annotations for

arterial phase abdominal CT scans of 210 unique kidney cancer patients.

Network Settings

The number of layers was 5 for each of the encoder and decoder paths for all compared methods. For U-net, RU-net, DU-net and the proposed method, cross entropy loss function was used and Adam optimiser [135] was applied with an initial learning rate of 0.001. As described in the original paper [132], AttnU-net used a new loss function based on Tversky index with three hyper parameters. It also added attention blocks and optimised by stochastic gradient descent with momentum. For the ISIC 2018 dataset, all methods were trained for 50 epochs with a batch size of 4. All methods for brain data were trained with batch size of 2 for 200 epochs. For Kidney dataset, batch size is 4 and the model is trained for 200 epochs.

4.3.2 Experimental Results

Skin Data

Table 4.1 lists the results for the skin lesion dataset. The results show that U-net produced the lowest dice score and Jaccard values. RU-net achieved better performance than U-net for dice, Jaccard and precision measures. AttnU-net and DU-net produced similar results in terms of dice and Jaccard measures, which were better than U-net and RU-net. DRUnet significantly (statistically significant using Wilcoxon signed rank test with $p < 0.001$) outperformed all other methods for dice, Precision and Jaccard measures. AttnU-net achieved the highest recall. This may happen due to the settings of the three hyper-parameters for the Tversky loss function favour the recall measurement (see [132] for details). Some visual results of using U-net, RU-net, DU-net and DRU-net are presented in Figure 4.3. It is seen that the segmentation results of DRU-net (Figure 4.3-(f)) is more similar to the ground truth mask (Figure 4.3-(b)) than the other methods, especially for the region around the lesion

boundary. The segmentation result of U-net (Figure 4.3-(c)) and RU-net (Figure 4.3- (d)) fail to recover precise lesion boundaries. DU-net (Figure 4.3-(e)) had a more precise segmentation result than U-net and RU-net, but still produced more false positives than DRUnet around the lesion boundary.

Table 4.1: Comparison of U-net, RU-net, DU-net, AttnU-net and DRU-net on ISIC 2018. The mean dice coefficient values, precision, recall, and Jaccard index are reported.

Method	**Dice**	**Precision**	**Recall**	**Jaccard**
U-net	0.840	0.865	0.869	0.724
RU-net	0.848	0.892	0.857	0.743
DU-net	0.855	0.894	0.865	0.748
AttnU-net	0.856	0.858	**0.897**	0.748
DRU-net	**0.861**	**0.919**	0.882	**0.755**

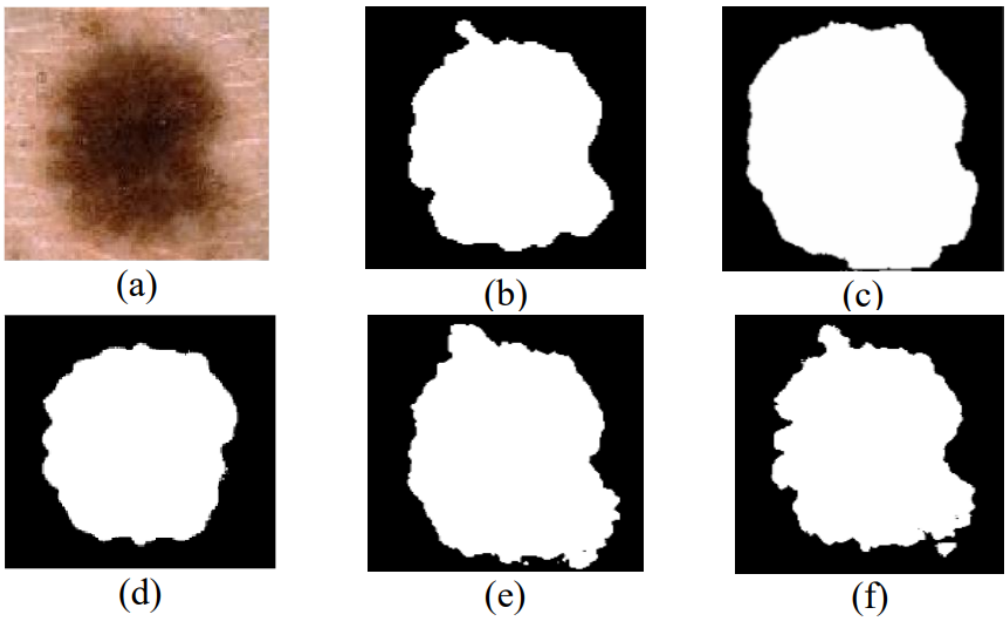


Figure 4.3: *a) Original image. b) Ground truth. c) Segmentation result by U-net. d) Segmentation result by RU-net. e) Segmentation result by DU-net. f) Segmentation result by DRU-net.*

Brain Data

The methods are compared in a multi-class segmentation task using the brain MRI data for MB and SN segmentation. Segmentation accuracy (i.e. dice) and training time are reported in Table 4.2 for different experiments by varying the size of training set as described in previous section.

Table 4.2: Comparison of U-net, RU-net, DU-net and DRU-net on different sizes of training data using brain MRI dataset. The mean dice coefficient values with Wilcoxon signed rank test results and training time are reported. Based on Wilcoxon signed rank test: $*$ indicates result of DRU-net is different from U-net with statistical significance ($p < 0.01$); $\neg$ indicates result of DRU-net is different from RU-net with statistical significance ($p < 0.01$); $\circ$ indicates result of DRU-net is different from DU-net with statistical significance ($p < 0.01$).

Number of training / testing	**Method**	**Dice**		**Training Time (Min.)**
		MB	**SN**	
200/100	U-net	0.8850	0.7383	117.69
	RU-net	0.9068	0.7782	148.87
	DU-net	**0.9167**	0.7908	238.36
	AttnU-net	0.9088	0.7690	223.40
	DRU-net	0.9132$*^{\neg}$	**0.7956**$*^{\neg}$	164.63
100/200	U-net	0.8473	0.7018	68.53
	RU-net	0.8578	0.7126	70.24
	DU-net	0.8764	0.7572	136.62
	AttnU-net	**0.8788**	0.7623	112.59
	DRU-net	0.8643$*^{\neg\circ}$	**0.7713**$*^{\neg\circ}$	96.62
50/250	U-net	0.7984	0.5877	68.53
	RU-net	0.8101	0.6439	58.90
	DU-net	0.8220	0.6578	90.78
	AttnU-net	0.8187	0.6746	91.24
	DRU-net	**0.8428**$*^{\neg\circ}$	**0.7240**$*^{\neg\circ}$	50.64

From Table 4.2, it is concluded that DRU-net is significantly better than RU-net and U-net for both MB and SN classes despite of training size. DRU-net was significantly better than DU-net when the number of training images were reduced to 50 for MB class. The segmentation of SN class is challenging as there are very few pixels in this class. The proposed DRU-net outperformed all other methods with statistical significance in using 100 and 50 training samples for the SN segmentation. The performance differences became more significant when the training size was decreased. In terms of training time, U-net, RU-net and DRU-net were similar and significantly faster than DU-net. This is due to the larger number of parameters required by the DU-net for the "bottle neck" and "transition" blocks. The number of parameters for RU-net, DU-net and DRU-net were 2 million, 6 million and 3 million respectively. Therefore, it is obvious that the proposed DRU-net is an efficient network in

terms of memory, training time and segmentation accuracy.

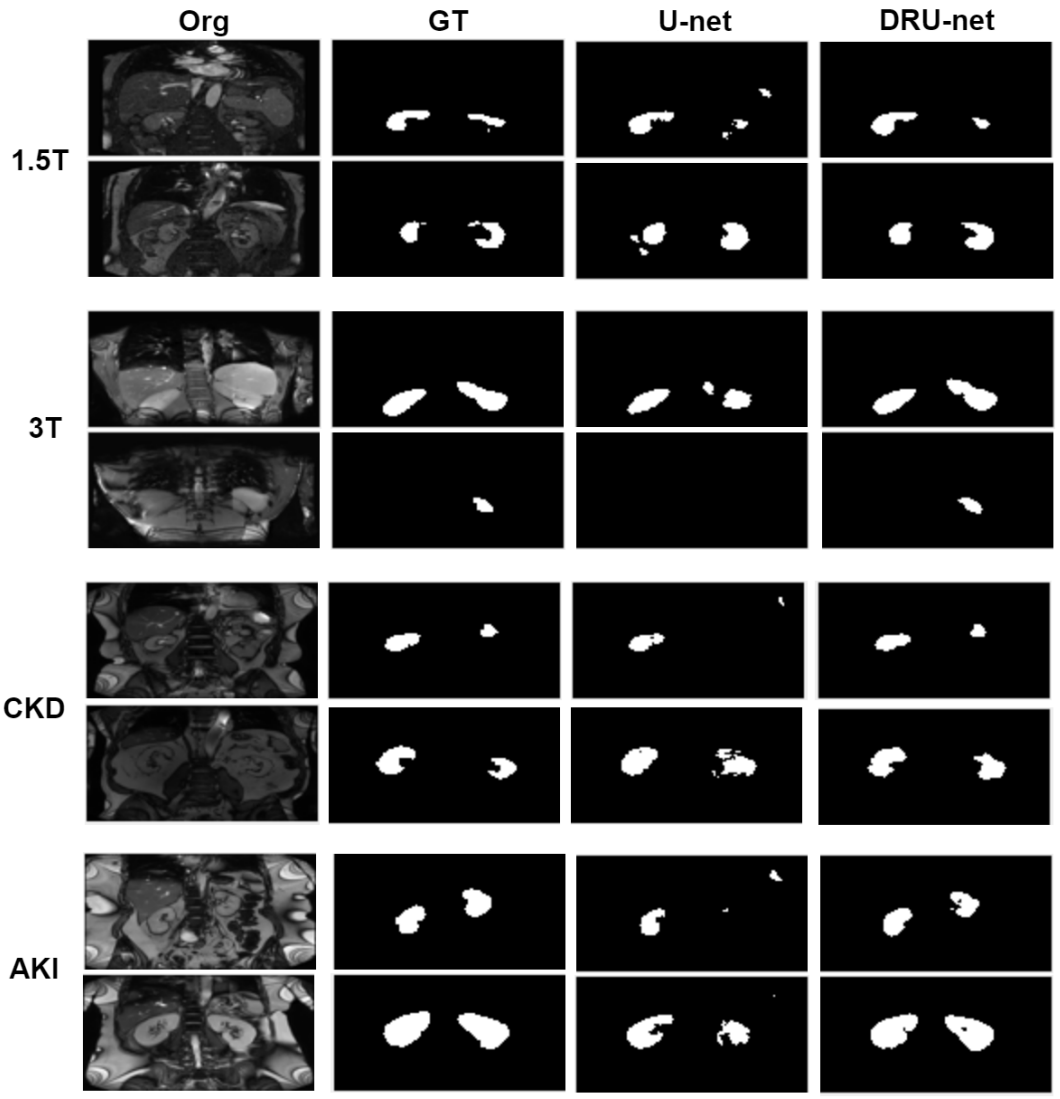


Figure 4.4: *Visual results for each comparative methods 1.5T, 3T, CKD, and AKI modalities respectively. First column is original image (Org), second column ground truth (GT), third column shows the results of U-net and last column represents the results of DRU-net*

Kidney Data

In Table 4.3, U-net and DRU-net are compared on 3D kidney dataset with four different MRI sequences, i.e. 3T, 1.5T, AKI, and CKD to show the performance of the proposed method on 3D volumes. For the training, all the sequences are mixed and included to the training set. At the test time, the performance for each of the image sequences was reported separately. According to the table, DRU-net outperformed U-net with a significant difference in terms of

Table 4.3: Segmentation results of kidney dataset trained on 1.5T, 3T, CKD, and AKI modalities based on different methods. The results are reported based on Dice, Precision, Recall, ans ASSD. Mean ± standard deviation values for all testing images are reported.

Method	Dice	Precision	Recall	ASSD
1.5T Healthy MRI				
Unet	0.884 ± 0.027	0.852 ± 0.028	$\mathbf{0.921 \pm 0.054}$	0.814 ± 0.334
DRUnet	$\mathbf{0.910 \pm 0.025}$	$\mathbf{0.905 \pm 0.029}$	0.918 ± 0.051	$\mathbf{0.674 \pm 0.352}$
3T Healthy MRI				
Unet	0.918 ± 0.024	0.885 ± 0.033	$\mathbf{0.954 \pm 0.028}$	0.535 ± 0.190
DRUnet	$\mathbf{0.936 \pm 0.015}$	$\mathbf{0.940 \pm 0.017}$	0.933 ± 0.025	$\mathbf{0.361 \pm 0.076}$
3T CKD MRI				
Unet	0.814 ± 0.236	0.870 ± 0.084	$\mathbf{0.806 \pm 0.245}$	1.352 ± 1.697
DRUnet	$\mathbf{0.861 \pm 0.188}$	$\mathbf{0.912 \pm 0.029}$	0.824 ± 0.217	$\mathbf{0.926 \pm 1.003}$
3T AKI MRI				
Unet	0.829 ± 0.087	$\mathbf{0.939 \pm 0.007}$	0.750 ± 0.131	1.152 ± 0.311
DRUnet	$\mathbf{0.868 \pm 0.035}$	0.899 ± 0.004	$\mathbf{0.858 \pm 0.058}$	$\mathbf{0.812 \pm 0.070}$

Table 4.4: Segmentation results of kidney dataset based on different methods using 5-fold cross validation. The results are reported based on Dice, Precision, Recall, and average symmetric surface distance (ASSD). Mean ± standard deviation values for all testing images are reported.

Train Data	**Test Data**	**Dice**	**Precision**	**Recall**	**ASSD**
DRU-net 2D					
CT	CT	0.915 ± 0.065	0.935 ± 0.031	0.904 ± 0.109	0.485 ± 0.467
MRI	MRI	0.929 ± 0.039	0.948 ± 0.027	0.913 ± 0.063	0.401 ± 0.113
MRI	CT	0.381 ± 0.179	0.878 ± 0.095	0.262 ± 0.150	2.440 ± 1.270
CT	MRI	0.533 ± 0.135	0.832 ± 0.093	0.402 ± 0.123	2.588 ± 1.178
DRU-net 3D					
CT	CT	0.911 ± 0.163	0.933 ± 0.092	0.906 ± 0.194	0.421 ± 0.247
MRI	MRI	0.935 ± 0.028	0.936 ± 0.026	0.936 ± 0.042	0.410 ± 0.223
MRI	CT	0.630 ± 0.144	0.830 ± 0.089	0.530 ± 0.175	2.527 ± 1.184
CT	MRI	0.436 ± 0.156	0.705 ± 0.136	0.334 ± 0.154	3.677 ± 1.268

all evaluation measures except Recall on 1.5T and 3T. Besides, in Figure 4.4, the visual results based on different sequences are shown. As it can be seen from this figure, the output of DRU-net is more similar to the ground truth.

Table 4.4 reports DRU-net results on 2D/3D and CT/MRI kidney data. In this table, in one experiment, the train set includes images from CT or MRI domains, and at the test time, the method is evaluated on unseen images from the same domain (CT or MRI) as training. In another experiment, the test images are coming from a different domain (the opposite domain of training). As it is obvious, DRU-net achieved a high performance on both modalities and both 2D and 3D volumes when the source and target are from the same modality (i.e. dice score 0.915 and 0.911 on CT images 2D and 3D respectively, and 0.929 and 0.935 on MRI 2D and 3D respectively). However, when the source and target are coming from two different modalities, DRU-net cannot segment the unseen images properly as the images are coming from different distributions. This is one of the limitations of DRU-net and other supervised networks.

4.4 Conclusions and Discussion

A simple but efficient network is proposed that is designed to take the advantages of DenseNet and ResNet. Additional minor modifications like aggregation of feature maps using summation and concatenation of the input to the output at each layer make the feature learning process more efficient. To be specific, the additional shortcut connection allows the gradients to be back-propagated efficiently even if the gradient approaches zeros in the first Conv. The concatenation that is used in encoder path is to include more information from the input of each layer and make the dimension of feature maps compatible for the next layer without extra parameters. In the decoder block a Conv 1×1 is used to reduce the number of channels of input instead of cropping it to make the dimension compatible to the layer output. By evaluating the

methods on a public skin lesion dataset, a local brain MRI dataset, and a public and a local kidney dataset, the proposed DRUnet has shown to outperform the other similar networks significantly, especially for the label class that has small number of pixels and with small number of training examples (brain data). For the skin lesion dataset, and brain dataset, the method also outperformed one of the state-of-the-art methods using attention network and Tversky loss function. However, as it has been discussed in last experiment, the proposed method is not able to segment another domain at the test time. Therefore, in the next chapter, a multi-modality image segmentation method is proposed to segment images from another domain at the test time. Besides, the efficiency of the proposed network is demonstrated in the context of image segmentation within an encoder-decoder DCNN structure. The applicability and efficiency of applying the proposed network to other scenarios (e.g. object classification) with larger datasets are worth exploring in future work.

Chapter 5

CIDA: Cross-modality Image Segmentation via Domain Adaptation of Unpaired Images

5.1 Introduction

The recent success of supervised deep learning models is attributed to (i) sufficient amount of annotated data, and (ii) sampling of the training and testing images are from the same data distribution [33]. However, in medical applications, image annotations require careful and time-consuming analysis from trained experts. Further, in medical imaging contexts, there are various imaging modalities, scanners and protocols [138] that capture various physical properties and provide different information for disease diagnosis and treatment. Therefore, the training and testing instances can come from different underlying data distributions (known as domain shifts). In clinical studies, patients often have different scanning modalities, thus constructing a general segmentation model that is suitable for various modalities with the ability to handle such significant domain shift is essential and extremely useful in practice. Motivated by this, the aim is to train a generalised model to segment images from different domains (source and target) where image annotations

are only available in the source domain for model training.

Early work in the literature addressed the multi-domain image segmentation problem by incorporating complementary information, such as adding the target domain images to an additional input channel [139] and information fusion of both domains [140]. These methods are supervised and require co-registered images as a guarantee of visual appearance alignment between the source and target domains. Fine-tuning the pre-trained models on the source data is another idea to cope with domain discrepancy [141]. However, a large annotated dataset of the target domain is still required for these methods.

Recently, research attention has been drawn to unsupervised domain adaptation, which aims to not only address the problem of scarcity of annotated data, but also overcome the domain shift issue. Most of these unsupervised domain adaptation methods utilised generative models (e.g. generative adversarial network (GAN) [142], auto-encoder [143], etc) to learn feature representation in the target domain in an unsupervised manner while adapting the information from a supervised learning in the source domain. There are a number of techniques for addressing domain adaptation, including those based on discrepancy-based methods [42], self-ensembling methods [43], methods built upon adversarial learning and cycle consistency [44] [45], and disentanglement strategies [46] [47]. Although some progress has been achieved in tackling this problem, these frameworks normally consist of multiple network modules and various aspects of adversarial learning. The implementation of such a complicated framework even in 2D is memory intensive and training prohibitive, and it is almost impossible to work with 3D volumes using a single GPU workstation.

Different from the aforementioned methods, a simple and efficient framework is proposed to tackle the cross-modality image segmentation problem, where image annotations for only one domain is available for model training. The main contributions of this chapter are summarised as follows. (1) Instead of using complex disentangle network (e.g. DADR [46] that disentangle do-

main into the domain-invariant and domain-specific features, normally using two different networks) to remove appearance discrepancies of images in different domains, Canny edge detection is applied to remove intensity information, and guide image segmentation only using edges of objects. (2) A generative adversarial network (GAN) based image reconstruction method is utilised to learn a joint latent representation of the edge information of all domains in an unsupervised manner. More importantly, a novel loss function is proposed that forces the feature representation of images in different domains to be close to each other in the latent space. This greatly improves the generalisation of the trained image segmentation model. (3) The proposed method was evaluated using a public CT dataset (KiTS19, MICCAI 2019 challenge) and a local MRI dataset for kidney segmentation. The model was trained using CT images and tested on MRIs. Through various ablation studies, the necessity of the above mentioned improvements is demonstrated. The proposed method has also outperformed other state-of-the-art methods with statistical significance, while remaining a simpler network architecture with fewer number of network parameters.

5.2 Methodology

The overall structure of the proposed method (named as Cross-modality Image Segmentation via Domain Adaptation of Unpaired Images(CIDA)) is illustrated in Figure 5.1. The cross-modality image segmentation problem can be formally described as: X_s and X_t represent a set of images from the source domain (e.g. CT) and the target domain (e.g. MRI) respectively, which have significantly different intensity distributions but sharing common semantic content (e.g. same organ of interest). The source domain consists of N_s labeled source samples $\{(x_s^i, y_s^i)\}_{i=1}^{N_s}$ with images, x_s^i, and their corresponding annotations, y_s^i. The target domain includes N_t unlabeled samples $\{x_t^i\}_{i=1}^{N_t}$. Z_s and Z_t represent the learned latent representations of the source and target domains

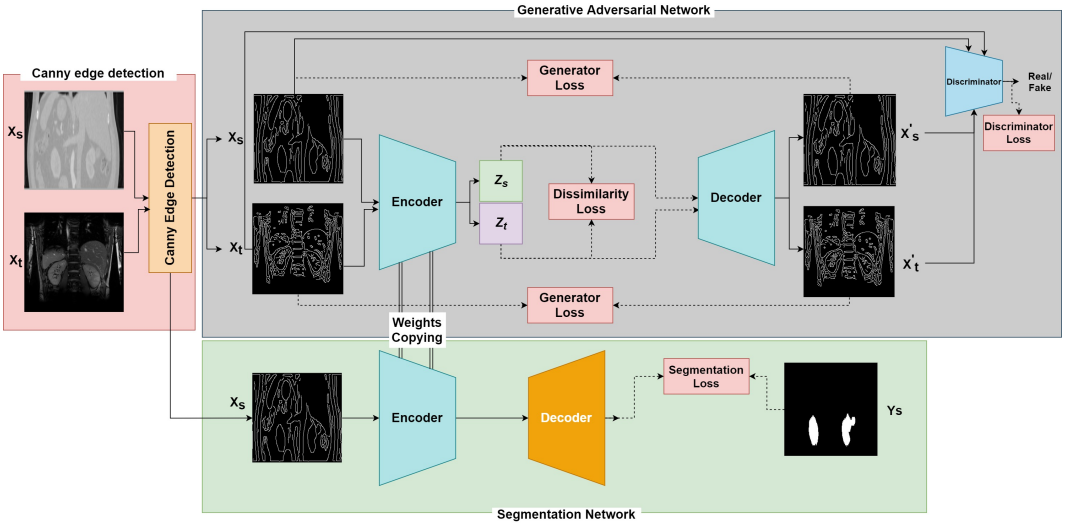


Figure 5.1: *The overall scheme of the proposed method. Generator and discriminator loss applied to encourage the generator to generate synthesized images that has high probabilities of being classified as real by the discriminator. Dissimilarity loss is added to minimise the absolute distances of the latent representations (Z_s and Z_t) of images in the source and target domains. Segmentation loss is require as the task is segmentation.*

respectively.

As illustrated in Figure 5.1, images from both domains are firstly pre-processed by applying Canny edge detection [144] to remove the cross-domain intensity discrepancy (pink box on the left). Other gradient-based edge detection methods (e.g. Sobel filter) are tried, however, Canny edge detection is chosen as it is more robust in connecting local discontinuities for weak edges. The proposed method mainly consists of two steps: (1) GAN based image reconstruction method to learn a generalised encoder for mapping the edge images of different domains into a common latent representation (gray box); (2) an image segmentation network with an encoder copied from step (1) and a trainable decoder (green box). The network is then trained by using images and labels from the source domain. The details of each step is described in the following subsections.

5.2.1 GAN based Unsupervised Feature Learning

In this step, a GAN-based auto-encoder is applied to learn the latent representations of images in both source and target domains via image reconstruction

in an unsupervised manner. The overall loss function is defined as the weighted sum of three terms:

$$L_{total} = L_{adv} + \alpha L_{rec} + \beta L_{latent} \tag{5.1}$$

where α and β are the weights that balance the loss terms. L_{adv} expressed in the following, is the adversarial loss that encourages the generator to generate synthesized images that has high probabilities of being classified as real by the discriminator [145].

$$L_{adv} = E_{x_i \sim X_s, X_t}[\, log(1 - D(G(x_i))] \tag{5.2}$$

where G is the generator, and D is the discriminator. x_i is sampling from source and target domains.

L_{rec} is the pixel-wise reconstruction loss that has been used in conventional auto-encoders [145]. As expressed in the following equation, the absolute error between the reconstructed image $G(x_i)$ and the input image x_i is minimised.

$$L_{rec} = E_{x_i \sim X_s, X_t}[\, |G(x_i) - x_i|] \tag{5.3}$$

By only utilising the L_{adv} and L_{rec} losses, the network learns to reconstruct images in their own domains. The learned latent representations of different domains remain separating from each other(Figure 5.1- pink box). Hence, a segmentation model trained using images in one domain would not achieve a good performance when is applied to images in another domain. As one of the contributions of the method, another loss term L_{latent} is proposed to minimise the absolute distances of the latent representations (Z_s and Z_t) of images in the source and target domains, as expressed in the next equation. During model training, it requires image batches from both domains to be input to the network for this loss calculation at each iteration.

$$L_{latent} = |Z_s - Z_t| \tag{5.4}$$

5.2.2 Cross-modality Image Segmentation

An encoder-decoder network (e.g. U-net [33]) is adopted to achieve the image segmentation. Once the image reconstruction network is trained (the previous section), the encoder in the network can be used to map the input images of both domains into a common latent space (an example is shown in Figure 5.3 (b)). Hence the learned weights of the ready trained encoder is copied to the encoder of the segmentation network. Only the decoder needs to be trained using images and labels from one domain. Two strategies are implemented to train the segmentation network: (1) fix the encoder and only train the decoder from scratch. (2) inspired from [146], only fix the batch normalisation parameters and fine tuning the other weights in the encoder, while the decoder is trained from scratch. The experiments showed that the second strategy achieved better performance. Hence consistently it is used for the experiments in next section. The loss for the supervised image segmentation network is a weighted sum of both cross entropy and Dice coefficient as suggested in [14], as expressed as follows.

$$L_{seg} = \gamma E_{x_i \sim X_s}[-y_i log(\hat{y}_i)] + \delta E_{x_i \sim X_s}[-\frac{2\hat{y}_i y_i}{y_i + \hat{y}_i}] \tag{5.5}$$

where γ and δ control the contribution of each loss term. $\hat{y}_i$ is the predicted output and y_i is the corresponding ground truth label.

5.2.3 Network Architecture

Both the generator and image segmentation network adopt the same Residual U-net (RU-net) architecture [130]. The encoder of RU-net is used for the discriminator in the image reconstruction network. In the encoding and decoding path, each layer consists of two convolutions (Conv) in a residual block [40].

Each Conv is followed by a batch normalisation and a rectified linear unit (ReLU). In the encoding path, the feature map for each successive layer is down-sampled by using a max pooling of stride 2 with doubled feature channels. In the decoding path, the output feature map in each layer is up-sampled using 2×2 up-convolution with half of feature channels in the previous layer. Besides, the feature map at each encoder layer is concatenated to the corresponding decoder feature map. All convolutions use kernel size of 3×3 and the initial channel number is 16. The number of layers for both paths is 5. Finally, a 1×1 convolution is used in the last layer to convert the dimension of feature maps to the number of classes in the segmentation network.

5.3 Method Evaluation

5.3.1 Datasets, Preprocessing and Parameter Settings

CIDA is validated using a publicly available CT data from KiTS19 challenge in MICCAI 2019 and a local MRI dataset. The dataset includes annotations for arterial phase abdominal CT scans of 300 unique kidney cancer patients. MRI kidney volumes were acquired by Balanced turbo field echo (bTFE) coronal localizer scans, and kidney volume masks were generated by manually tracing the kidney on the bTFE localizer images. CT images and MRI are used as the source domain and target domain respectively. For training the GAN, 1200 2D slices of MRI and 1200 2D slices of CT images are randomly chosen for training set, and 50 2D slices of CT and 50 2D slices of MRI were used for validation set. For the segmentation task, the same 1200 CT slices and 50 CT slices with their corresponding labels were used as training and validation sets respectively, and the test set includes 100 unseen MRIs. Besides, in all sets, there are images that contain no kidneys. All images were resized to 256×256, and the intensity of all sets is normalised using zero-mean normalisation.

For both the GAN and segmentation networks, the networks are trained

for 200 epochs using Adam optimiser with an initial learning rate of 10^{-4} and decayed by multiplying 0.8 for every 10 epochs. For the GAN-based feature

learning network, the batch size was 8 (4 CT and 4 MRI to calculated the L_{latent} loss in equation (4)). The batch size for the segmentation network was 4. The hyper parameters, α, β, γ and δ, were experimentally determined using the validation set and set to 40, 20, 0.5 and 0.5 respectively.

5.3.2 Experimental Results

For method comparison, firstly a baseline method (RU-net) is established: a RU-net model using original CT images for training and MRIs for testing. As the second method, the edge images of CT were used for training a RU-net model, and it was tested on edge images of MRIs, named as RU-net-C. As the performance upper bound, also a RU-net model is trained using original MRIs and tested on MRIs, named as RU-net-MRI. Furthermore, the methods are compared with a widely used image-to-image translation method (CycleGAN [93]) and a recent domain adaptation network (DAPNet [14]). For the CycleGAN, both the CT and MRI training images were used to learn the image translations between the two modalities. Then, a segmentation network RU-net was trained using CT images only. The MRI test images were segmented by firstly synthesizing their corresponding CT images using the trained CycleGAN model, and then input them to the segmentation model. The test of the DAPNet method was the same as the proposed method: CT images with labels were used as the source domain and MRI images were the target domain to be segmented. Furthermore, two ablation studies are performed to test the contributions of Canny edge detection and the proposed L_{latent} loss. CIDA-CN represents the use of the proposed method without Canny edge detection (i.e. the original images with zero-mean normalisation were used). CIDA-LN is the proposed method without using the L_{latent} loss term.

Table 5.1: Segmentation results of different methods. The results are reported based on Dice, Jaccard, Precision, and Recall.

Method	Dice	Jaccard	Precision	Recall
RU-net-MRI	0.8521	0.8287	0.8291	0.8796
RU-net	0.2453	0.1600	0.4762	0.1879
RU-net-C	0.3308	0.2242	0.7543	0.2299
CycleGAN	0.3066	0.1957	0.6576	0.2114
DAPNet	0.5323	0.3992	0.4832	**0.6115**
CIDA-CN	0.3529	0.2347	0.3314	0.4165
CIDA-LN	0.5058	0.3797	0.7113	0.4225
CIDA	**0.6364**	**0.5290**	**0.7545**	0.5566

Dice, Jaccard, precision and recall of the above methods are reported in Table 6.1. Visual results of three example images for each method are shown in Figure 5.2. It is seen from Table 5.1 and Figure 5.2 that a simple RU-net trained on CT images performs poorly when tested on MRIs. The use of edge information instead of the original intensity improves the robustness of the model (RU-net-C). The image-to-image translation method (CycleGAN) did not achieve a good performance, even worse than the RU-net-C method. It is observed that the global intensity/style of the synthesized image using CycleGAN is satisfactory. However the intensities of specific tissue types were incorrect (e.g. bone), hence leading to a poor performance. The domain adaptation method DAPNet performed much better than the RU-net, RU-net-C and CycleGAN methods in terms of Dice, Jaccard and recall, but with a lower precision. This indicates the DAPNet produced more false positives which can also be seen in Figure 5.2. Overall, the proposed method (CIDA) achieved significantly (statistically significant measured by Wilcoxon signed rank test with $p < 0.01$) better performance than other methods in terms of Dice, Jaccard and precision, with the second best for recall. When Canny edge detection was removed from the proposed method, the performance dropped from 0.6364 to 0.3529 (CIDA-CN) for Dice value, indicating the effectiveness of removing the intensity discrepancy using edge detectors.

The L_{latent} loss also contributed significantly in the method, which boosted

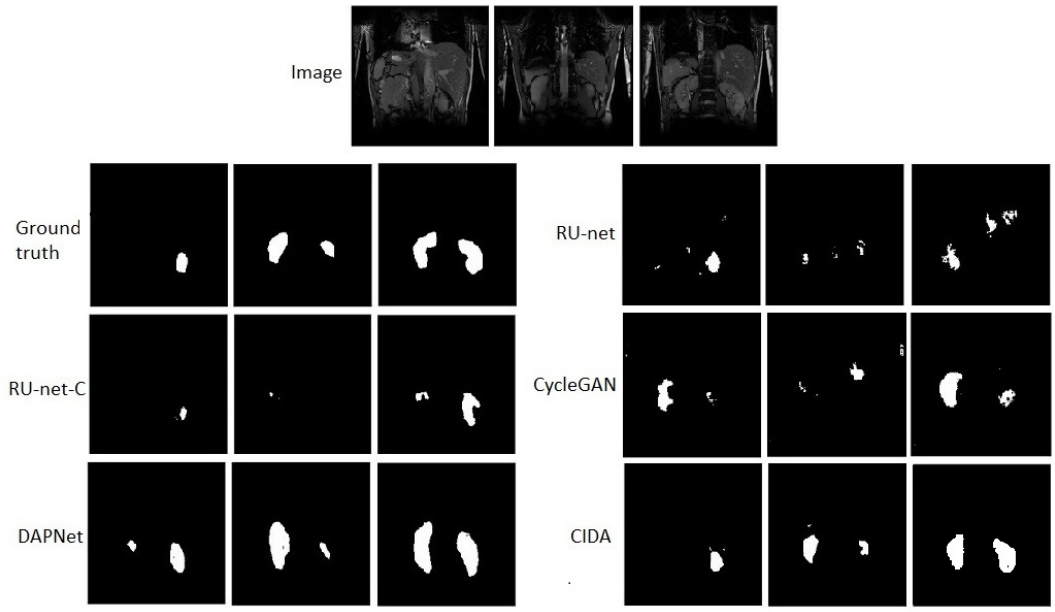


Figure 5.2: *Visual results of three example images for each comparative methods.*

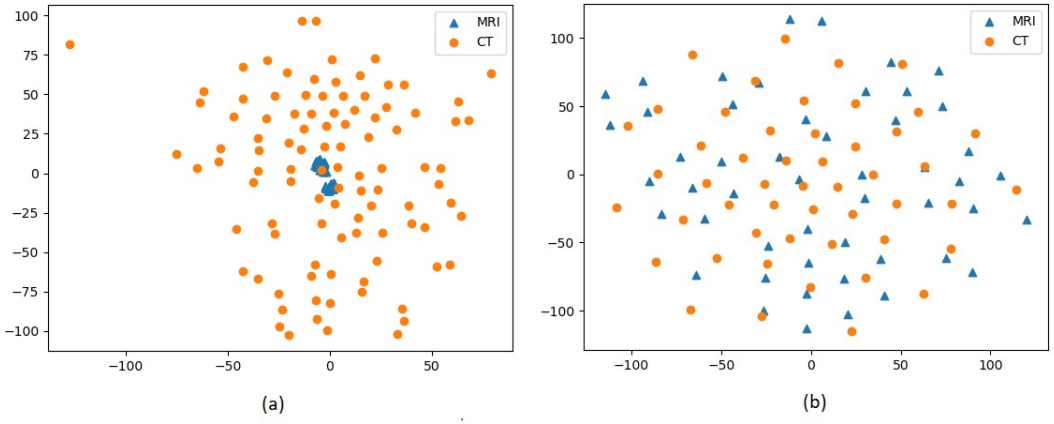


Figure 5.3: *Visualization of the latent feature space of the segmentation network using t-SNE for CT and MRI edge images, without (a) and with (b) the proposed* L_{latent} *loss.*

the performance from 0.5058 (CIDA-LN) to 0.6364 for Dice value. The learned latent features (output of the encoder in the segmentation network) of 50 CT and 50 MRI edge images from test sets using t-SNE [147] is visualized with and without the L_{latent} loss, as shown in Figure 5.3. It can be observed that the L_{latent} loss successfully mixed the feature space of images from both domains, which helps the cross-domain image segmentation. However, compared with the fully supervised learning method (RU-net-MRI), there is a room for improvement.

5.4 Conclusions and Discussion

In this chapter, a novel method is proposed to address the cross-modality image segmentation problem, where the image labels are only available from one modality for model training. The effectiveness of using Canny edge detection is demonstrated to minimise the intensity discrepancy of two domains, which significantly improved the model robustness. Furthermore, a GAN-based feature learning method is modified by adding a new latent space similarity term. Based on both improvements, the proposed method outperformed the state-of-the-art solutions for this challenging task. However, there are some limitations, the first one is the usage of canny edge detection to minimise the intensity discrepancy as this method may remove some of the essential details, therefore, there is a need of a more powerful method to minimise the intensity discrepancy between domain and keep all the relevant details. Another improvement of CIDA can be related to two different stage of training, i.e. one for reconstruction (training the encoder and decoder), one for segmentation (training of a new decoder). Thus, in the next chapter, an improved version of CIDA is proposed to address the aforementioned challenges and increase the performance and decrease the model complexity. Besides, as a future investigation, it is also noteworthy to mention that the proposed method can easily be extended to a multi-domain scenario (e.g. CT, 1.5T MRI, 3T MRI, 7T MRI etc.), as long as the edges of the organ can be extracted.

Chapter 6

LMISA: A Lightweight Multi-modality Image Segmentation Network via Domain Adaptation using Gradient Magnitude and Shape Constraint

6.1 Introduction

In the previous chapter, a novel method is proposed for cross-modality image segmentation. In this chapter, another novel method is proposed to address the multi-modality image segmentation, which is an extension on CIDA (the previous chapter), and called LMISA. LMISA is a more accurate and light-weight version of CIDA that is evaluated on different datasets and scenarios.

Briefly, LMISA is a simple and efficient framework to achieve multi-modality image segmentation for both 2D and 3D images, in which image annotations of only one domain are available for model training. In this work, instead

of using complex networks to remove appearance discrepancies across different domains, a novel multi-resolution locally-normalised gradient magnitude (MLNGM) approach is applied as a pre-processing step. Next, a hybrid network, including image reconstruction and image segmentation tasks is utilised to learn a joint latent representation of the images of both domains simultaneously. Additionally, a shape constraint is added to the model by comparing the predicted segmentation results of the target domain with the ground truth labels of the source domain using a discriminator in the context of adversarial learning. This helps in preserving the general shape of the object and removing false positives in the background. The proposed method was evaluated using a public CT dataset (KiTS19, MICCAI challenge 2019) and a local MRI dataset for kidney segmentation, and using the Multi-Modality Whole Heart Segmentation (MMWHS challenge 2017) dataset comprising MRI and CT data for cardiac organ segmentation. The model was trained using CT and MRI as the source and the target domain respectively and then vice versa.

The main contributions of this chapter are summarised as follows:

- An efficient end-to-end domain adaptation network is proposed by combining a simple image pre-processing step based on multi-resolution local gradient magnitude and a shape constraint in adversarial learning.

- Experiments have been performed on two segmentation tasks, i.e. kidney (binary class) and cardiac organ segmentation (multiple classes). Moreover, a comprehensive ablation study has been conducted to demonstrate the effectiveness of each of the proposed elements. The proposed method has achieved superior segmentation accuracy. Furthermore, the capability of directly applying the trained model is demonstrated to images of a new target domain without model retraining, which achieved remarkable performance and significantly better than other state-of-the-art methods.

- By achieving better segmentation performance than other state-of-the-art methods, the proposed end-to-end model requires fewer network pa-

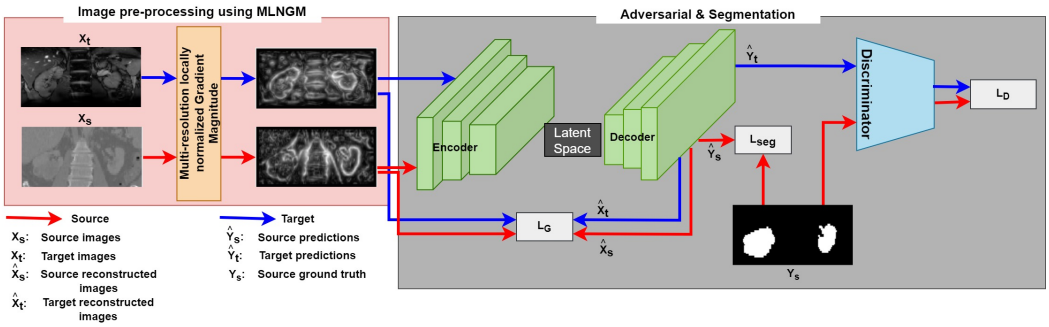


Figure 6.1: *The overall scheme of LMISA. 1) The images from both domains are input to the image pre-processing method (Multi-resolution Locally normalised Gradient Magnitude (MLNGM)). 2) The output of MLNGM is then input to an encoder-decoder network as a batch of images from both domains. L_G, L_{seg} and L_D are the loss terms to be optimised during model training.*

rameters, less memory usage and shorter training time, and is implemented for both 2D and 3D images.

This method is published in medical image analysis journal, and is titled as LMISA: A Lightweight Multi-modality Image Segmentation Network via Domain Adaptation using Gradient Magnitude and Shape Constraint [148].

6.2 Methodology

The cross-modality image segmentation problem is formally described as follows. X_s and X_t represent a set of images from the source domain and the target domain respectively, which have significantly different intensity distributions but sharing common semantic content. The source domain consists of N_s labeled source samples $\{(x_s^i, y_s^i)\}_{i=1}^{N_s}$ with images, x_s^i, and their corresponding annotations, y_s^i. The target domain includes N_t unlabeled samples $\{x_t^i\}_{i=1}^{N_t}$. The goal is to train a model that can correctly predict the label of any given images in either the source or the target domain (more difficult). The overall structure of the proposed proposed method, named as Lightweight Multi-modality Image Segmentation network via domain Adaptation (LMISA), is presented in Figure 6.1.

In Figure 6.1, images from two different domains (source and target) are firstly pre-processed using the proposed proposed multi-resolution locally nor-

malised gradient magnitude (MLNGM), which is explained in the next section, method to minimise the cross-domain intensity discrepancy (pink box). Next, the pre-processed images are input to an encoder-decoder based network, which mainly performs two tasks with shared model weights: (1) it perform an image segmentation task: the multi-class cross entropy loss function is used based on the images and their associated labels in the source domain; (2) it serves as the generator in a GAN-based model: it learns the reconstruction of the input images from both domains. Therefore, the encoder-decoder network outputs both the segmentation prediction ($\hat{Y}_s$ and $\hat{Y}_t$ in Figure 6.1) and the reconstructed images ($\hat{X}_s$ and $\hat{X}_t$ in Figure 6.1) of both domains. $\hat{X}_s$ and $\hat{X}_t$ are used to calculate the loss function L_G in equation (6.5). $\hat{Y}_s$ is used to calculate the segmentation loss function L_{seg} in equation (6.1). The segmentation output of the target domain ($\hat{Y}_t$) and the ground truth label (Y_s) in the source domain are input to a discriminator to calculate the loss L_D in equation (6.7). The network is trained by using images with labels from the source domain and images without labels from the target domain. The detail of each component is described in the following subsections.

6.2.1 Multi-resolution Locally normalised Gradient Magnitude (MLNGM)

One of the key challenges in cross-domain image segmentation is the intensity discrepancies between images in different domains. In computer vision, it is well known that the image gradient is more robust than the original intensity to illumination changes for feature representation. Such as the famous histogram of oriented gradient (HOG) [149] that utilises local distribution of gradient information to represent features for object detection.

Moreover, difference of Gaussian (DoG or Laplacian) pyramid is widely used to represent the key information of the image in a multi-scale manner. DoG utilise two different Gaussian blurs on the image, with a different blur-

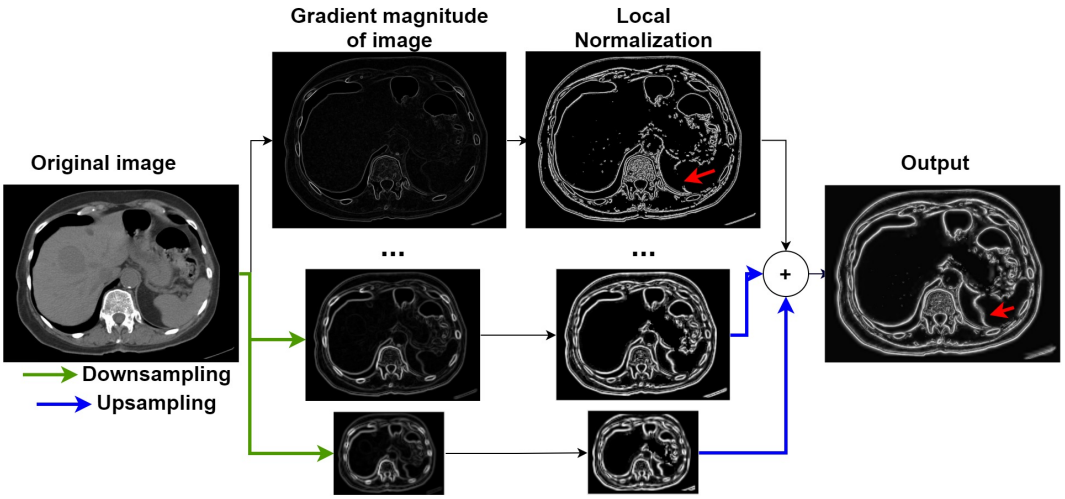


Figure 6.2: *Multi-resolution locally normalised gradient magnitude (MLNGM) scheme.*

ring radius for each, and performs subtraction between consecutively Gaussian blurred versions of images. This has been used in one of the most popular feature descriptors (SIFT [80]) as a scale estimation method.

By adopting from the ideas of gradient representation and image pyramid, a method is proposed that is based on multi-resolution locally normalised gradient magnitude to minimise the intensity discrepancy of images in the source and target domains, as illustrated in Figure 6.2. In this method, a multi-resolution pyramid is constructed with K scales by down-sampling the original image using bicubic image interpolation and calculating a locally normalised gradient magnitude image at each scale. K is determined by the size of the input image. Similar to the DoG method, if only the global gradient magnitude is used, some local edges with smaller gradient values can not be enhanced. Hence, a locally normalised gradient magnitude is necessary to provide much clearer representation of local object contours. Additionally, the images at different scales allow the combination of different levels of details, as highlighted by the red arrows in Figure 6.2.

The pseudo-code of this proposed method is presented in Algorithm 1. First, the input image I is down-sampled to a multi-level image pyramid; then, the gradient magnitude is independently calculated for each scale i, denoted as I^i_{GM}. Next, each I^i_{GM} is locally normalised if it satisfies a threshold condition,

Algorithm 1: Multi-resolution Locally Normalized Gradient Magnitude

Input: I

Output: Multi-resolution locally normalized gradient magnitude image(I_{MLNGM})

Initialize $Norm_{GM}$ to zeros with the same size of I

for $i = 1$ to K **do**

  $I^i \leftarrow$ Scale image to level $\frac{1}{2^{i-1}}$

  $I^i_{GM} \leftarrow$ Compute gradient magnitude of I^i

  $Std^i_{GM} \leftarrow$ Compute standard deviation of I^i_{GM}

  for each patch, j in I^i_{GM} **do**

    $Std^{i,j}_{GM} \leftarrow$ Compute standard deviation of $I^{i,j}_{GM}$

    if $Std^{i,j}_{GM} < \frac{Std^i_{GM}}{thr}$

      $Norm^{i,j}_{GM} = 0$

    else

      $Norm^{i,j}_{GM} = \frac{I^{i,j}_{GM} - min(I^{i,j}_{GM})}{max(I^{i,j}_{GM}) - min(I^{i,j}_{GM})}$

    end

  end

  $Norm_{GM} \leftarrow Norm_{GM} +$ Scale $Norm^{i,j}_{GM}$ to original size

end

$$I_{MLNGM} = \frac{Norm_{GM} - \min(Norm_{GM})}{\max(Norm_{GM}) - \min(Norm_{GM})}$$

return I_{MLNGM}

denoted as $Norm_{GM}^{i,j}$. The threshold is based on the standard deviation of I_{GM}^{i} divided by thr. thr is a positive real value that is experimentally determined depending on the object of interest. When thr is determined, it is used consistently for all images in the dataset. A regional higher standard deviation corresponds to a local region with high intensity variations. For a local $r \times r$ (or $r \times r \times r$ in 3D) region j, if the standard deviation $Std_{GM}^{i,j}$ is smaller than the threshold, the output value is set to 0. Otherwise, it is normalised by the min-max intensity normalisation method. With this mechanism, a higher thr allows to retain more detailed local changes, while a lower thr helps to remove these local responses by assuming they are noises.

Finally, $Norm_{GM}^{i,j}$ is upsampled in multiple hierarchies to the original image resolution using bicubic image interpolation and add them to obtain a single image. The proposed MLNGM algorithm effectively detects object contours of all regions, which is more robust in connecting local discontinuities of weak edges (indicated by the red arrows in Figure 6.2) than the global gradient magnitude method. The MLNGM method is compared to single scale gradient magnitude and Canny edge detection methods as the pre-processing step for model training. The results are reported in the results section.

6.2.2 Image Segmentation Network

As illustrated in Figure 6.1, an encoder-decoder network is utilised to achieve the image segmentation. This architecture consists of a two-part network with a feature extractor (encoder) to extract meaningful features from the images, and a decoder to perform the transformation of these features into a segmentation map corresponding to the input image. This task is trained with the source domain sample pairs $\{(x_s^i, y_s^i)\}_{i=1}^{N_s}$ by minimising a multi-class cross entropy [13] loss as expressed in the following equation:

$$L_{seg} = -\frac{1}{N_s} \sum_{i=1}^{N_s} \sum_{c=1}^{C} y_s^{i,c} log(\hat{y}_s^{i,c}) \tag{6.1}$$

where c and i repeat over all classes and pixels. $\hat{y}_s^{i,c}$ is the predicted output, and $y_s^{i,c}$ is a one-hot encoding representation of corresponding ground truth label for the input sample image x_s.

The encoder-decoder network is adopted from the Dense Residual U-net (DRU-net) architecture [12]. In the encoding and decoding path, each layer consists of two convolutions (Conv) in a residual block [40]. Each Conv is followed by a leaky rectified linear unit (leaky-ReLU). In the encoding path, the feature map for each successive layer is down-sampled by using a max pooling operation with doubled feature channels. In the decoding path, the output feature map in each layer is up-sampled using up-convolution with half of feature channels from the previous layer. Besides, the feature map at each encoder layer is concatenated to the corresponding decoder feature map. All convolutions use kernel size of 3×3 for 2D image and $3 \times 3 \times 3$ for 3D image. Finally, a 1×1 ($1 \times 1 \times 1$ for 3D image) convolution is used in the last layer to convert the dimension of the feature maps to the number of classes in the segmentation network.

6.2.3 GAN based Unsupervised Feature Learning

Although the learned model from the previous step can effectively segment the images of source domain, the model performance decreases in the target domain due to domain shift between the source and target domain. In this step, to mitigate the effects of domain shift, a GAN-based auto-encoder is applied to map the image samples from both domains into a common feature space and learn the latent representations of images in both the source and target domains in an unsupervised manner.

A typical GAN [142] consists of a generator and a discriminator. The generator aims to generate image samples from input as realistically as possible to fool the discriminator by estimating the underlying distribution of the target domain, whereas the discriminator aims to distinguish between the real

samples and the synthetic samples. The GAN network is optimised through a competition framework by a minimax game between the generator and the discriminator. Therefore, by supposing that generator G intends to minimise the cost function whereas the discriminator D aims to maximise it, the objective function is expressed as follows in the context of the proposed method:

$$\begin{aligned} &min_{G} max_{D} E_{x \sim P_r}[\, log(D(x))] \\ &+ E_{\hat{x} \sim P_g}[\, log(1 - D(\hat{x}))] \end{aligned} \tag{6.2}$$

where P_r represents the distribution of the input images, which are from both the source and target domains. P_g is the distribution of the generated images. $\hat{x} = G(x)$ indicates the samples produced by the generator. The encoder and decoder network of the segmentation task in the previous step is used as the generator G (using shared weights with no additional network modules). The same encoder structure of the segmentation network, but with only one Conv in each layer, is used as the discriminator D.

Although GANs have achieved the state-of-the-art performance in many applications, the original version suffers from some limitations: (i) since generator and discriminator update their loss functions alternatively, convergence is not guaranteed; (ii) the vanishing gradient problem can occur in different occasions; (iii) the generator can not generate sufficiently diverse samples to represent the whole distribution of the target domain (known as mode collapse). This is the result of trapping the generator to a local minimum of the cost function; therefore, similar samples are synthesized. The Wasserstein Generative Adversarial Network—Gradient Penalty (WGAN-GP) [150] is an improved version of GAN to solve some of the training difficulties of a typical GAN. In this version, the Jensen Shannon divergence is replaced by the Wasserstein distance function, which enforces a 1-Lipschitz constraint on the discriminator, and uses a gradient penalty to satisfy a soft version of the constraint of Lipschitz continuity. In WGAN-GP, the goal of the generator is

changed to minimise the following loss function, which is used in the proposed method:

$$L_G = -E_{\hat{x}\sim P_g}[\, D(\hat{x})] \tag{6.3}$$

Additionally, in order to learn a good feature representation in the target domain (without any labels) using the generator, inspired by the auto-encoders [145], a pixel-wise image reconstruction loss L_{rec} is added that minimises the mean squared error between the reconstructed image $G(x)$ and the input image x (as shown in equation (6.4)). This pixel-wise loss helps to improve the structural similarity between the input image and the synthesized image from the generator. L_{rec} is calculated based on images from both domains, hence a common feature space can be learned by the generator network [151].

$$L_{rec} = E_{x\sim P_r}[\, (G(x) - x)^2] \tag{6.4}$$

Therefore, the whole loss function for G is described as:

$$L_G = -E_{\hat{x}\sim P_g}[\, D(\hat{x})] \, + \alpha L_{rec} \tag{6.5}$$

where α is the weight to balance the loss terms.

Based on WGAN-GP, the discriminator loss with gradient penalty of the proposed method is calculated as follows:

$$\begin{aligned} L_D = E_{\hat{x}\sim P_g}[\, D(\hat{x})] \, - E_{x\sim P_r}[\, D(x)] \\ + \lambda E_{x'\sim P_{x'}}[\, (\| \bigtriangledown_{x'} D(x^{'}) \, \|_2 - 1)^2] \end{aligned} \tag{6.6}$$

where the first two terms perform Wasserstein distance estimation, and the last term is the network regularization gradient penalty term introduced by the WGAN-GP [150]. $x^{'}$ is uniformly sampled along straight lines between pairs of points sampled from the real data distribution and the generated data distribution. λ is the hyper-parameter for weighting the gradient penalty term, and $\| \, . \, \|_2$ is an L2 norm. In equation (6.6), the discriminator makes the

decision based on a learned feature space [98, 152].

Although domain adaption-based approaches have considerably improved the segmentation accuracy, there still remain some challenges. One of the key challenges is to maintain contextual information of shape and boundary of the segmented object that are not modeled explicitly in conventional GAN models. As the aim of the proposed application is image segmentation rather than image synthesis, an accurate shape prediction is more important than synthesizing a high quality image. Thus, it is proposed to replace the input of the discriminator from the original images and reconstructed images to the ground truth labels in the source domain and the predicted segmentation labels in the target domain. This helps to guide the predicted shape in the target domain to be similar to the ground truth shape in the source domain.

In previous works (e.g. [13, 15]), the discriminator in a GAN has been used to differentiate the output predictions between the source and target domains for more effective feature alignment. In the proposed framework, the discriminator is also applied in the semantic prediction space; however, the source labels are utilised as a shape guidance for the predictions of the target domain.

As shown in Figure 6.1, a discriminator is constructed to classify the outputs between the source ground truth annotations and the predicted segmentation outputs. Therefore, the discriminator back-propagates the adversarial gradients to minimise the discrepancy between the feature distributions of the ground truth shape in the source domain and the predicted shape in the target domain. Note that the discrimination is performed in the feature space, hence the ground truth label of the source domain and the predicted label of the target domain do not need to be aligned or paired. The only assumption here is that the aim is to segment the same organs of interest in both the source domain and the target domain. Consequently, the discriminator loss in equation

(6) is modified as follows:

$$\begin{aligned} L_D = & E_{\hat{y}_t \sim P_{\hat{y}_t}}[\, D(\hat{y}_t)] \, - E_{y_s \sim P_{y_s}}[\, D(y_s)] \\ & + \lambda E_{x' \sim P_{x'}}[\, (\| \bigtriangledown_{x'} D(x^{'}) \, \|_2 - 1)^2] \end{aligned} \tag{6.7}$$

where $\hat{y}_t$ is the predicted segmentation output in the target domain, and $P_{\hat{y}_t}$ and P_{y_s} are the distributions of the predicted labels in the target domain and the ground truth labels in the source domain, respectively. $x^{'}$ is sampled uniformly from the original images and the synthesized images to calculate the penalty.

Using the discriminator to embed the shape constraint in the feature space, enables the network to learn domain invariant features and also helps to maintain shape context by removing noisy false positives in the background, especially with the absence of absolute intensity information that is removed by the proposed pre-processing step. An ablation study has been performed to demonstrate the effectiveness of the shape constraint in section 6.3.2.

6.2.4 Model Training

The proposed framework is trained in an end-to-end manner. As illustrated in Figure 6.1, the network to be learned consists of an encoder-decoder and a discriminator. The network is trained by inputting a batch of source images, their corresponding ground truth labels and target images at each iteration as a mini-batch. The weights of the encoder-decoder and the discriminator are updated alternatively in each iteration. When the encoder-decoder is updated, equation (6.8) is applied based on both the generator loss (L_G) and the segmentation loss (L_{seg}).

$$L_{encoder_decoder} = L_G + \beta L_{seg} \tag{6.8}$$

where β is the weight to balance the loss terms. L_G and L_{seg} are expressed in equation (6.5) and equation (6.1) respectively.

When the discriminator is updated, the loss function in equation (6.7) is used. The detailed parameter setting (e.g. batch size) is described in section 6.3.1.

6.3 Method Evaluation

In this section, several experiments are designed to investigate different aspects of LMISA for medical image segmentation. The performance of LMISA is evaluated on two different medical imaging datasets, kidney and cardiac organ segmentation on MRI and CT images. The images, in both datasets, are unpaired; in other words, MRI data and CT data were from different subjects and different protocols. MRI and CT are alternatively considered as the source domain and the target domain. For all experiments, labels in the target domain are only available at test time for performance evaluation purpose.

6.3.1 Datasets, Preprocessing and Parameter Settings

Kidney Dataset

Firstly, LMISA is validated using a publicly available CT data from the KiTS19 challenge in MICCAI 2019 [137] and a local MRI dataset for kidney segmentation. The CT dataset includes annotations for arterial phase abdominal CT scans of 210 unique kidney cancer patients. In this dataset, the image size is either 512×512 or 512×796 and pixel size varies from $0.43mm \times 0.43mm$ to $1.04mm \times 1.04mm$. The number of slices varies from 29 to 1059, and the slice thicknesses range from 0.5 mm to 5 mm. The kidney MRI volumes were acquired using balanced turbo field echo (bTFE) coronal localizer scans from either a 1.5T scanner or a 3T scanner. The kidney volume masks were generated by manually tracing the kidney contour on the bTFE localizer images.

The MRI dataset consists of 170 3T and 83 1.5T volumes. In this dataset, the image size is 256×256 with pixel size of $1.5mm \times 1.5mm$, and the number of slices is from 13 to 80 with slice thickness from 5 mm to 7 mm.

Cardiac Dataset

As the second dataset, the Multi-Modality Whole Heart Segmentation (MMWHS) challenge 2017 dataset for cardiac organ segmentation [153] is chosen. This dataset includes 20 MRI and 20 CT volumes, and their corresponding segmentation ground truth. The image size is either 256×256 or 512×512, and the pixel size varies from $0.28mm \times 0.28mm$ to $1.2mm \times 1.2mm$. The number of slices varies from 112 to 363 with slice thickness varying from 0.45 mm to 1.6 mm. In this dataset, the annotations of four organs, including ascending aorta (AA), left atrium blood cavity (LAC), left ventricle blood cavity (LVC) and myocardium of the left ventricle (MYO) were provided by the dataset organizer.

Pre-processing and Parameter Settings

For parameter setting, 70% is used for training and 10% for validation, and the validation set was used for determining the parameters if not mentioned explicitly. λ (equation (6.7)) was set to 10 as suggested in the original paper [150]. Window size for MLNGM was set to 7. The batch size for 2D images was 4 (2 source data and 2 target data randomly sampled), while for 3D images was 2 (1 source data and 1 target data randomly sampled) because of GPU memory restriction. The original volumes were cropped by using a 3D bounding box to capture the same field of view, as the original images have different field of view. For instance, the kidney CT scans may include the area of hip bone or lungs. The intensity of all sets was normalised using zero-mean intensity normalisation. For LMISA, MLNGM was firstly applied then followed by zero-mean normalisation.

For the kidney dataset, each volume was firstly re-sampled to physical unit

(mm) and then cropped by a $300mm \times 200mm \times 150mm$ bounding box around the region of interest. If the original image was smaller than the bounding box, the original size was used. Finally, the images were resized to the size of $128 \times 64 \times 25$. For all 2D and 3D experiments performed in this paper, all the training, validation and testing sets were the same when testing different methods. For CT and MRI domain adaptation, only the 3T MRIs were used. The 1.5T MRIs were reserved for an independent test without model retraining. In 2D experiments, the coronal view image slices were utilised, and the slices that do not contain any kidney labels were removed for training as the same process suggested by Chen et al. [15]. For both 2D and 3D experiments, the networks were trained for 200 epochs using the Adam optimiser [135] with an initial learning rate of 10^{-3} and decayed by multiplying by 0.8 every 10 epochs. The initial channel number for the encoder-decoder architecture was 32. The number of layers for both the encoder and decoder was 3. The stride for downsampling and upsampling between different layers was [2, 2, 1] along the three dimensions. The hyper parameters, α (equation (6.5)) and β (equation (6.8)) were experimentally determined using the validation set, and set to 40 and 20, respectively. No data augmentation was performed for this dataset. For the kidney data, in addition to the performance in the target domain, also the model performance is evaluated on the source domain using the test set. This helps to investigate the impact of the performance when introducing another image modality to train the source domain model.

Each of the cardiac volumes was firstly re-sampled to physical unit (mm) and cropped by a $256mm \times 256mm \times 128mm$ bounding box around the region of interest. They were then resized to an image size of $128 \times 64 \times 64$. For this dataset, both the 2D and 3D version of LMISA were tested and the results were compared to other methods reported in the literature based on the same dataset using the same training, validation and testing sets. Due to the limited training data, data augmentation with rotation, translation, and scaling was applied during the training process. The rotation range was between $[-15°,$

15°] around one of the three axes randomly. The translation range was [-10, 10] pixels for the first image dimension (128 pixels) and [-5, 5] pixels for the second and third image dimension (64 pixels). For scaling, the range was [0.75, 1.2]. For this dataset, the network was trained for 200 epochs for 2D and 300 epochs for 3D using Adam optimiser with an initial learning rate of 10^{-3} for 2D and 10^{-4} for 3D. The learning rate was then decayed by multiplying by 0.8 every 10 epochs. The initial channel number for the encoder-decoder architecture was 16 for both 2D and 3D LMISA. The model that produced the best Dice value on the validation set was used for evaluation on the test set. The number of layers for the encoder and decoder was 4 for both 2D and 3D LMISA. The stride for downsampling and upsampling was [2,2] for 2D and [2, 2, 2] for 3D. The batch size for 2D was 64 (32 from each domain), and 2 (1 from each domain) for 3D. For both 2D and 3D LMISA, α and β were set to 40 and 40, respectively, in equation(6.5) and (6.8), based on the result of the validation set.

All experiments were performed on a workstation with Intel Xeon 4110 2.1 GHz CPU, 16 GHz memory and a single GeForce GTX 1080Ti graphic card. The code was implemented in Python based on Tensorflow 2.0 and available on Github.

Evaluation Metrics

The similarity between the predicted segmentation and the ground truth for both 2D and 3D experiments was consistently calculated in the 3D volume using the Dice coefficient (Dice), Precision, Recall, and average symmetric surface distance (ASSD). The mean ± standard deviation were reported for all testing images. All these measures were performed in the original image resolution for a fair comparison with other methods. If a downsampled image was used in the experiments, the output segmentation was resized to the original image resolution. Besides the number of model parameters, the memory usage for model training, training time and inference time were reported on kidney

data to compare the model complexity of LMISA with other state-of-the-art methods.

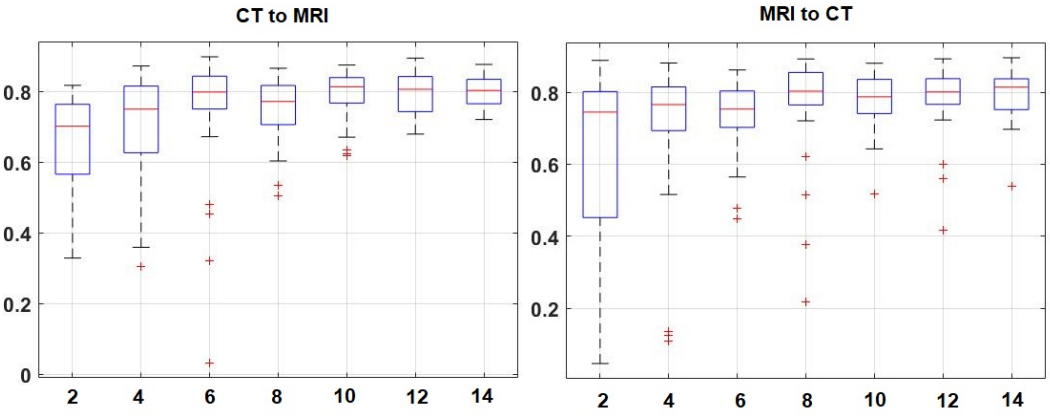


Figure 6.3: *By varying the thr value in MLNGM for both CT to MRI and MRI to CT domain adaptation, Dice scores (vertical axis) were compared using LMISA on kidney dataset.*

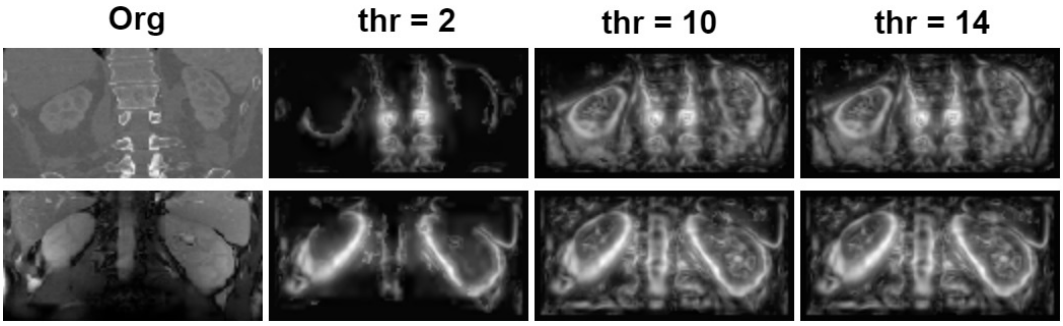


Figure 6.4: *Visualization of the original images (Org), output of MLNGM by thr = 2, MLNGM by thr = 10, MLNGM by thr = 14. The top row is CT and the bottom row is MRI.*

6.3.2 Experimental Results

Parameter optimisation for MLNGM

Since the absolute intensity information were removed from the input images using the proposed MLNGM method and only focus on local gradient information, the level of details retained in the MLNGM output may affect the segmentation quality. Hence, before performing any subsequent experiments, firstly the effect of the key parameter (*thr*) is investigated in the MLNGM method. In this experiment, the kidney dataset was divided into 70% for train-

ing, 10% for validation and 20% for testing. LMISA-3D is evaluated based on Dice score on Kidney data for both directions, CT to MRI and MRI to CT. The value of *thr* was varied from 2 to 14 at an interval of 2, as shown in Figure 6.3. Some example images of using different *thr* are shown in Figure 6.4. According to the Dice values in Figure 6.3, by considering both directions, the performance of LMISA became similar and maintained in a high level when the value of *thr* is greater than 10. The small variations were due to different model initializations. It can be seen from Figure 6.4 that by setting *thr*$>$ 10 (i.e. 10 and 14), the maximum amount of contour information was detected in both CT and MRI and no significant differences for higher *thr* values. In addition, statistical tests measured by Wilcoxon signed rank test revealed that there were no statistically significant differences of Dice scores when *thr* values were greater than 10 (with all p values greater than 0.1). This indicates that the performance of LMISA is not sensitive to the setting of *thr*, as long as the maximum amount of contour information is extracted. Thus, the value of *thr* is set to a large value (14), and all the subsequent experiments were evaluated based on this value.

Comparison of Pre-processing Methods

The proposed pre-processing method (MLNGM) is also compared with a global gradient magnitude method and the widely used Canny edge detection method, denoted as LMISA-gm and LMISA-edg respectively. In this experiment, the kidney dataset was divided into 70% for training, 10% for validation and 20% for testing. Figure 6.5 presents the original images (Org), edge images (Edges), gradient magnitude (GM) images and MLNGM images of one CT and one MRI example. Table 6.1 presents the segmentation results of using LMISA-gm, LMISA-edg and LMISA-MLNGM on the kidney dataset using the 3D version of the proposed method.

From Table 6.1, it can be observed that LMISA-edg achieved average Dice values of 0.508 and 0.553 when MRI and CT were the target domain respec-

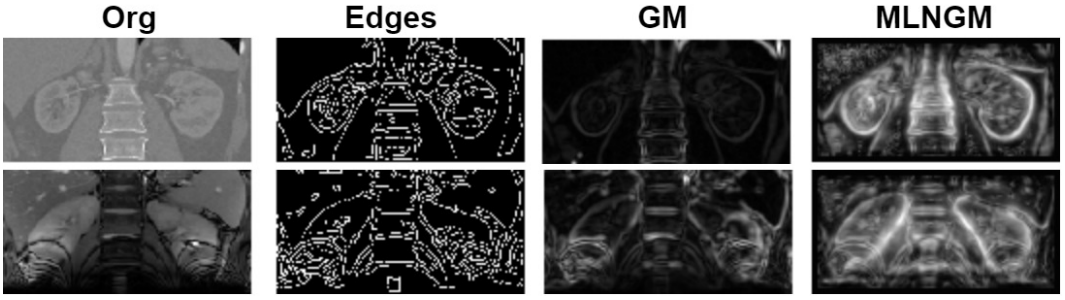


Figure 6.5: *Visualization of the original images (Org), canny edge detection (Edges), gradient magnitude (GM), and the proposed method (MLNGM). The top row is CT and the bottom row is MRI.*

tively. LMISA-gm produced an average Dice of 0.603 and 0.794 when being tested on MRI and CT images respectively. The proposed method LMISA-MLNGM outperformed both LMISA-edg and LMISA-gm significantly (statistically significant measured by Wilcoxon signed rank test with $p < 0.01$) for Dice, Recall and ASSD measurements in both CT to MRI and MRI to CT.

Table 6.1: Segmentation results of 3D kidney dataset based on different image preprocessing methods. The results are reported based on Dice, Precision, Recall, and average symmetric surface distance (ASSD). Mean ± standard deviation for all testing images are reported.

Method	Data	Dice	Precision	Recall	ASSD
CT to MRI					
LMISA-edg	CT	0.873 ± 0.041	0.856 ± 0.059	0.899 ± 0.077	0.930 ± 0.295
	MRI	0.508 ± 0.145	$\mathbf{0.907 \pm 0.061}$	0.368 ± 0.138	2.498 ± 0.750
LMISA-gm	CT	0.941 ± 0.028	0.933 ± 0.030	0.950 ± 0.042	0.477 ± 0.339
	MRI	0.603 ± 0.171	0.870 ± 0.097	0.482 ± 0.179	2.837 ± 1.263
LMISA-MLNGM	CT	0.934 ± 0.019	0.921 ± 0.032	0.950 ± 0.036	0.427 ± 0.295
	MRI	$\mathbf{0.776 \pm 0.070}$	0.894 ± 0.058	$\mathbf{0.693 \pm 0.097}$	$\mathbf{1.461 \pm 0.496}$
MRI to CT					
LMISA-edg	MRI	0.874 ± 0.037	0.861 ± 0.038	0.892 ± 0.070	0.809 ± 0.237
	CT	0.553 ± 0.168	0.660 ± 0.181	0.519 ± 0.207	2.282 ± 0.746
LMISA-gm	MRI	0.925 ± 0.031	0.930 ± 0.024	0.920 ± 0.049	0.470 ± 0.232
	CT	0.794 ± 0.151	$\mathbf{0.855 \pm 0.053}$	0.779 ± 0.204	1.473 ± 0.416
LMISA-MLNGM	MRI	0.929 ± 0.027	0.929 ± 0.018	0.931 ± 0.049	0.487 ± 0.249
	CT	$\mathbf{0.806 \pm 0.076}$	0.824 ± 0.049	$\mathbf{0.809 \pm 0.148}$	$\mathbf{1.185 \pm 0.567}$

Ablation Study

To better understand the contribution of each component in LMISA, also a number of ablation experiments are conducted based on the Dice score from CT to MRI and MRI to CT using the 3D version of the proposed method that is shown in Figure 6.6 on kidney data and Figure 6.7 on cardiac data.

In these experiments, the datasets were divided into 70% for training, 10% for validation and 20% for testing. Firstly, a baseline model was considered, which was trained using the original images without the shape guidance in the discriminator, named as *no MLNGM/no shape*. In this case, the same proposed model structure was used. For the discriminator, the inputs were the reconstructed images of both domains and the original images of both domains. Second, the effect of adding the shape aware component to the baseline (i.e. replacing the input of the discriminator by the predicted segmentation output of the target domain and the ground truth label from the source domain) is investigated. The original images were input to the encoder-decoder, which was named as *no MLNGM/shape aware*. In another experiment, the benefits of adding MLNGM to the baseline without shape constraint is investigated, and called *MLNGM/no shape*.

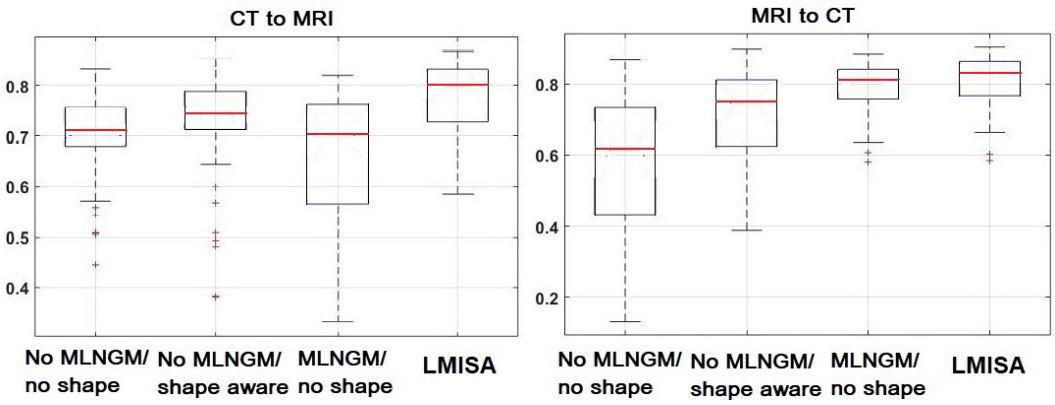


Figure 6.6: *Ablation study of the key components in the proposed method based on Dice score for both CT to MRI and MRI to CT domain adaptation on kidney dataset.*

According to Figure 6.6 for both CT to MRI and MRI to CT on kidney data, the dice values for *no MLNGM/no shape* model were lowest (median Dice scores of 0.71 and 0.61 respectively). Based on the plots of *no MLNGM/shape aware* using MRI and CT as the target domain respectively, it is obvious that the shape guidance improved the performance significantly. It can be seen that *MLNGM/no shape* model in Figure 6.6 CT to MRI did not improve the baseline method, but achieved better performance in MRI to CT. However,

together with the shape constraint, the model achieved the best performance (refer to LMISA in Figure 6.6, plot CT to MRI).

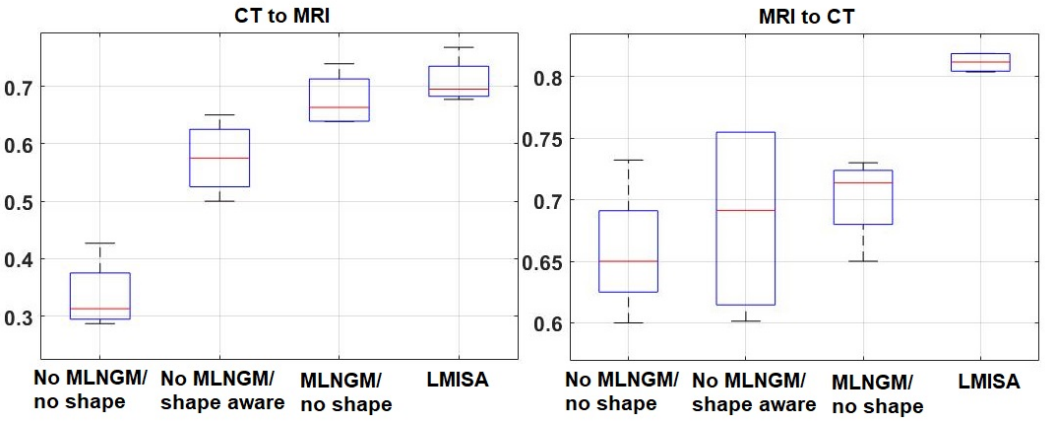


Figure 6.7: *Ablation study of the key components in the proposed method based on Dice score for both CT to MRI and MRI to CT domain adaptation on Cardiac dataset. The average Dice score of four organs are used here.*

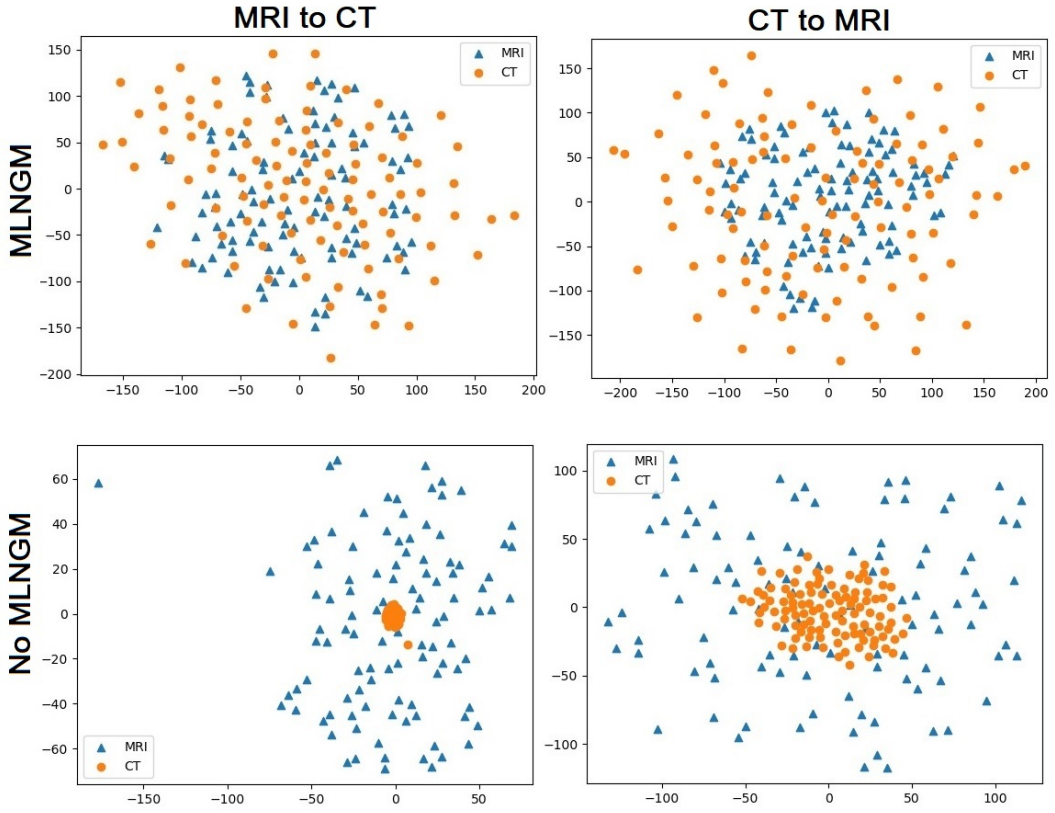


Figure 6.8: *Visualization of the latent feature space of LMISA with and without MLNGM using t-SNE for both CT to MRI and MRI to CT domain adaptation on 3D volumes.*

In Figure 6.7, based on the cardiac dataset, for CT to MRI, the dice values for *no MLNGM/no shape* model were varying from 0.29 to 0.42. Based

on the plots of *no MLNGM/shape aware* and *MLNGM/no shape* when MRI is the target, it is observed that both the shape guidance and MLNGM improved the performance significantly. When CT is the target, it can be seen that *No MLNGM/shape aware* model in comparing with the baseline method achieves a better higher median value but with a larger variation of dice scores. *MLNGM/no shape* was shown to be helpful and improved the performance in comparing with the baseline with a significant difference.

Overall, the results in Figure 6.6 and Figure 6.7 have shown that applying MLNGM and shape constraint individually may not always improve the performance, compared to the baseline. This may be dependent on the image quality, model complexity of the shape and number of classes. However, the combination of MLNGM and shape constraint is able to produce significantly improved performance in both MRI to CT and CT to MRI domain adaptation.

Furthermore, also the feature space was visualized (using t-SNE [147]) and obtained from the encoder in LMISA for the methods with and without *MLNGM/shape aware* respectively, for both MRI to CT and CT to MRI of 3D volume data. As shown in Figure 6.8, it is observed that the MLNGM method helps in effectively "mixing" the feature representations of both CT and MRI images in the latent feature space.

Comparison to the State-of-the-art Methods

Kidney Dataset: The proposed method is compared with a baseline model using fully supervised training (DRU-net) [12], a widely used image-to-image translation method (CycleGAN [93]), and two recent domain adaptation networks called DAPNet [14], SIFA [15] and CIDA (the proposed method in the previous chapter) on the kidney dataset. The code of the methods in the literature were provided by the authors, but only 2D versions were available (the 3D versions were not implemented due to model complexity). Both the 2D and 3D versions of the proposed method were implemented for comparison.

The results reported in this section were based on five-fold cross validation.

Table 6.2: Segmentation results of kidney dataset based on different methods using 5-fold cross validation. The results are reported based on Dice, Precision, Recall, and average symmetric surface distance (ASSD). Mean ± standard deviation values for all testing images are reported. Asterisk indicates the best 2D method and the overall best domain adaptation method is highlighted in bold.

Method	**Test Data**	**Dice**	**Precision**	**Recall**	**ASSD**
Fully-supervised					
DRU-net-2D [12]	CT	0.915 ± 0.065	0.935 ± 0.031	0.904 ± 0.109	0.485 ± 0.467
DRU-net-2D	MRI	0.929 ± 0.039	0.948 ± 0.027	0.913 ± 0.063	0.401 ± 0.113
DRU-net-3D	CT	0.911 ± 0.163	0.933 ± 0.092	0.906 ± 0.194	0.421 ± 0.247
DRU-net-3D	MRI	0.935 ± 0.028	0.936 ± 0.026	0.936 ± 0.042	0.410 ± 0.223
Without Domain Adaptation					
DRU-net-2D-MRI	CT	0.381 ± 0.179	0.878 ± 0.095	0.262 ± 0.150	2.440 ± 1.270
DRU-net-2D-CT	MRI	0.533 ± 0.135	0.832 ± 0.093	0.402 ± 0.123	2.588 ± 1.178
DRU-net-3D-MRI	CT	0.630 ± 0.144	0.830 ± 0.089	0.530 ± 0.175	2.527 ± 1.184
DRU-net-3D-CT	MRI	0.436 ± 0.156	0.705 ± 0.136	0.334 ± 0.154	3.677 ± 1.268
Without Domain Adaptation + MLNGM					
DRU-net-2D-MRI	CT	0.652 ± 0.067	0.577 ± 0.096	0.797 ± 0.155	2.567 ± 0.643
DRU-net-2D-CT	MRI	0.7193 ± 0.054	0.731 ± 0.072	0.716 ± 0.079	1.607 ± 0.391
DRU-net-3D-MRI	CT	0.768 ± 0.086	0.819 ± 0.058	0.745 ± 0.149	1.383 ± 0.506
DRU-net-3D-CT	MRI	0.678 ± 0.155	0.942 ± 0.062	0.553 ± 0.173	1.988 ± 1.255
Domain Adaption **CT to MRI**					
CycleGAN [93]	MRI	0.635 ± 0.094	0.865 ± 0.065	0.514 ± 0.108	2.067 ± 0.769
DAPNet [14]	CT	0.886 ± 0.031	0.906 ± 0.037	0.923 ± 0.042	0.476 ± 0.343
	MRI	0.742 ± 0.060	0.849 ± 0.048	0.691 ± 0.080 *	1.674 ± 0.263
SIFA [15]	CT	0.886 ± 0.085	0.891 ± 0.031	0.892 ± 0.091	0.510 ± 0.340
	MRI	0.748 ± 0.057	0.877 ± 0.030	0.660 ± 0.104	1.652 ± 1.325
CIDA	CT	0.904 ± 0.073	0.932 ± 0.026	0.887 ± 0.114	0.432 ± 0.251
	MRI	0.730 ± 0.115	0.914 ± 0.045 *	0.616 ± 0.133	1.732 ± 0.447
LMISA-2D	CT	0.891 ± 0.039	0.881 ± 0.043	0.907 ± 0.066	0.472 ± 0.278
	MRI	0.756 ± 0.066 *	0.848 ± 0.056	0.690 ± 0.090	1.587 ± 0.298 *
LMISA-3D	CT	0.934 ± 0.025	0.944 ± 0.039	0.939 ± 0.043	0.391 ± 0.283
	MRI	$\mathbf{0.768 \pm 0.081}$	$\mathbf{0.922 \pm 0.033}$	$\mathbf{0.732 \pm 0.125}$	$\mathbf{1.527 \pm 0.617}$
Domain Adaption **MRI to CT**					
CycleGAN [93]	CT	0.544 ± 0.160	$\mathbf{0.828 \pm 0.066}$ *	0.428 ± 0.171	1.853 ± 0.652
DAPNet [14]	MRI	0.909 ± 0.027	0.753 ± 0.030	0.921 ± 0.034	0.368 ± 0.173
	CT	0.660 ± 0.115	0.762 ± 0.088	0.607 ± 0.144	1.817 ± 1.785
SIFA [15]	MRI	0.800 ± 0.054	0.753 ± 0.057	0.875 ± 0.054	1.251 ± 0.774
	CT	0.624 ± 0.096	0.606 ± 0.091	0.673 ± 0.123	1.794 ± 0.686
CIDA	MRI	0.903 ± 0.038	0.912 ± 0.028	0.897 ± 0.059	0.511 ± 0.182
	CT	0.741 ± 0.082	0.794 ± 0.062	0.764 ± 0.135	1.682 ± 0.424
LMISA-2D	MRI	0.880 ± 0.033	0.868 ± 0.042	0.895 ± 0.044	0.685 ± 0.246
	CT	0.756 ± 0.073*	0.715 ± 0.075	$\mathbf{0.809 \pm 0.134}$ *	1.451 ± 0.559 *
LMISA-3D	MRI	0.923 ± 0.032	0.932 ± 0.029	0.916 ± 0.047	0.479 ± 0.240
	CT	$\mathbf{0.791 \pm 0.088}$	0.813 ± 0.055	0.791 ± 0.147	$\mathbf{1.395 \pm 0.601}$

The performance of the upper bound was established by directly applying a DRU-net model using CT images for training and unseen CT images for testing, and then the same for MRI images (named as DRU-net-2D or 3D in Table 6.2). For lower bound, DRU-net without any domain adaption approach was trained on source domain and evaluated on target domain. In another experiment, DRU-net was trained on source domain which was pre-processed

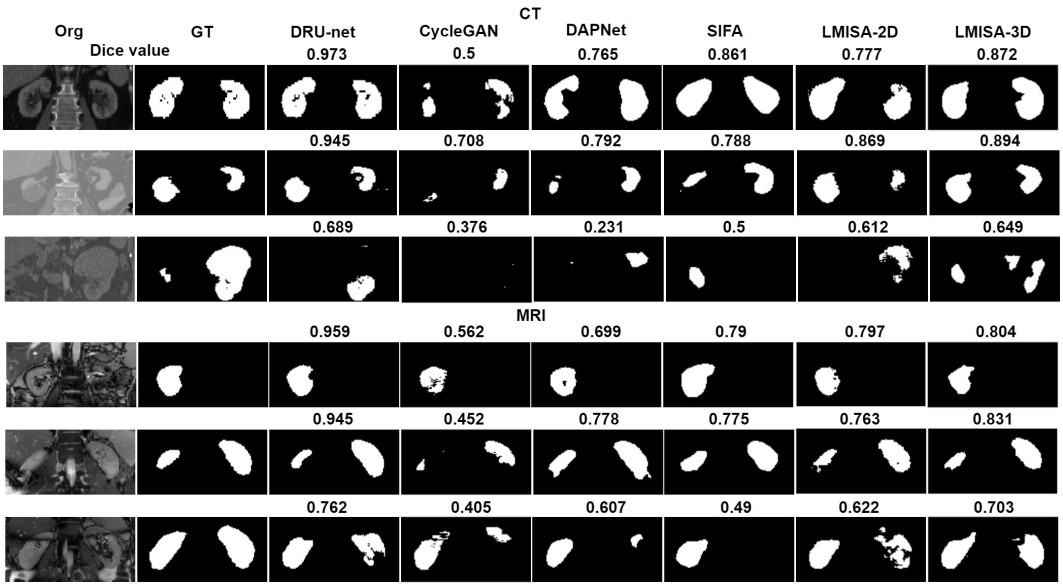


Figure 6.9: *Visual results for each comparative method using CT and MRI as the target domain respectively. The top three rows are CT, and the bottom three rows are MRI images. These images are chosen based on the best, median, and the worst dice score of DRU-net method.*

by MLNGM and assessed on target domain with MLNGM. For the CycleGAN method, both the CT and MRI training images were used to learn the image translations between the two modalities. Then, a segmentation network DRU-net was trained using the source image. During testing, the test images in the target domain were firstly converted to the source image domain using the trained CycleGAN model, and then input to the segmentation model. For testing the methods DAPNet, SIFA, CIDA and LMISA, no separate image conversion step is required.

Dice, Precision, Recall, and ASSD measures for the aforementioned methods on kidney images are reported in Table 6.2. In Table 6.2, asterisk indicates the best 2D method and the overall best domain adaptation method is highlighted in bold. It can be seen from Table 6.2 that LMISA-2D was the best domain adaptation method using 2D images. For CT to MRI, SIFA, DAPNet, CIDA and CycleGAN were ranked from the second to the fifth in terms of Dice and ASSD values for 2D methods. For MRI to CT, based on the ASSD measurement, CIDA, SIFA, DAPNet and CycleGAN were ranked from the second to the fifth in sequence. Considering both the ranking sequences based on the

Dice and ASSD measures, SIFA was the second best method after LMISA-2D in 2D methods.

LMISA-3D achieved the best overall performance in terms of Dice and ASSD measurements for both CT to MRI and MRI to CT. They were statistically significantly better than any other methods, measured by Wilcoxon signed rank test with $p < 0.01$. Moreover, the LMISA-3D produced similar performance in comparison with the upper bound baseline using DRU-net method when testing images came from the source domain. This indicates that the proposed LMISA method performed well in segmenting the target domain images without downgrading the performance in the source domain, while other methods produced poorer performance in the source domain segmentation task.

Three visual examples for each method using MRI and CT as the target domain respectively are shown in Figure 6.9. The images were chosen based on the best, median, and the worst dice scores of DRU-net. As shown in Figure 6.9, it can be seen that both LMISA-2D and LMISA-3D visually provide more similar outputs to the ground truth images than CycleGAN, DAPNet and SIFA with significantly lower false negatives (higher recall in Table 6.2). In general for the proposed LMISA method, the performance of MRI to CT segmentation is higher than CT to MRI as MRI normally contains more tissue information than CT.

In another experiment, all the methods were evaluated on another unseen target domain (1.5T kidney MRIs of 83 subjects). All the models were trained on CT as source and 3T MRI as the target domain, and then tested on 1.5T MRI without model retraining. Figure 6.10 shows the 3T and 1.5T intensity differences, and the output of MLNGM. The segmentation results are reported in Table 6.3. According to this table, LMISA-3D outperformed all the other methods with a significant difference and LMISA-2D achieved the second best place in terms of all the measurements except for the Recall.

The Dice (0.780) and ASSD (1.082) values of the proposed method (LMISA-

Table 6.3: Segmentation results of kidney dataset from an unseen target domain (1.5T MRIs of 83 subjects) based on different methods. The results are reported based on Dice, Precision, Recall, and average symmetric surface distance (ASSD). Mean ± standard deviation values for all testing images are reported. Asterisk indicates the best 2D method and the overall best domain adaptation method is highlighted in bold.

Method	Dice	Precision	Recall	ASSD
1.5T MRI				
CycleGAN [93]	0.408 ± 0.175	0.766 ± 0.156	0.293 ± 0.153	2.791 ± 0.879
DAPNet [14]	0.665 ± 0.101	0.671 ± 0.103	0.663 ± 0.111	1.576 ± 0.495
SIFA [15]	0.673 ± 0.032	0.566 ± 0.038	**0.833 ± 0.043 ***	1.785 ± 0.193
LMISA-2D	0.756 ± 0.097 *	0.856 ± 0.077 *	0.686 ± 0.124	1.178 ± 0.157 *
LMISA-3D	**0.780 ± 0.072**	**0.886 ± 0.067**	0.706 ± 0.103	**1.082 ± 0.421**

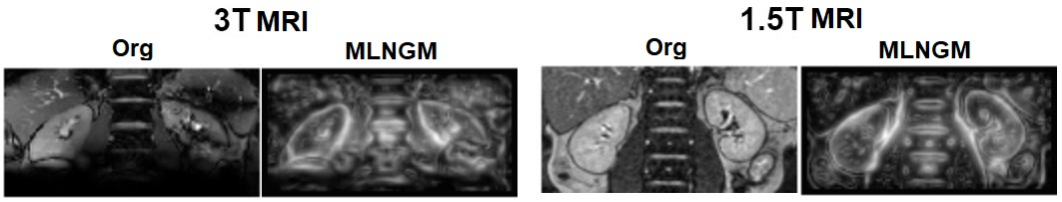


Figure 6.10: *Example slices of 3T MRI and 1.5T MRI, and their corresponding MLNGM output.*

Table 6.4: Based on the segmentation of kidney dataset, the number of parameters, memory usage, training time and inference time of different methods are reported.

Method	Number of parameters (m)	Memory usage (MB)	Training time (h)	Inference time (s)
DRU-net-2D [12]	0.758	1935.93	3.21	4.32
DRU-net-3D [12]	2.071	1963.1	2.83	4.62
CycleGAN [93]	28.396	2204.21	21.02	9.72
DAPNet [14]	2.651	1998.33	13.33	5.53
SIFA [15]	43.510	3679.28	14.56	6.68
LMISA-2D	0.858	1957.84	5.29	6.06
LMISA-3D	2.555	1975.09	6.70	6.89

3D) for 1.5T MRIs (Table 6.3) are similar to the 3T MRI results (0.768 and 1.527 in Table 6.2). However, other methods (CycleGAN, DAPNet and SIFA) produced much worse results for 1.5T MRI than 3T MRIs. It indicates that the proposed method is more robust to segment images from new domains without model retraining. This is mainly due to the MLNGM method effectively reducing the appearance discrepancy of images in different domains.

In Table 6.4, the number of model parameters, memory usage (occupied GPU memory during training), training time and inference time for all methods are compared. The fully-supervised methods DRU-net-2D and DRU-net-3D were used as baseline here, which required smaller number of parameters

and less memory than these domain-adaptation models. LMISA-2D and 3D in comparison with other domain adaptation methods had shown to be more lightweight with about 2 million parameters for 3D and 858,000 for 2D, and only required $\sim$ 6 hours of training time. Moreover, LMISA had similar memory usage and number of model parameters to the baseline DRU-net method, as LMISA uses exactly the same architecture of DRU-net by only adding an additional discriminator. According to Table 6.4, DAPNet had similar memory usage and network parameters as the DRU-net-3D and LMISA-3D methods, and the training time was doubled in comparison with LMISA-2D and 3D. This is due to that DAPNet uses two discriminators, one shared encoder, one pyramid pooling module and two decoders in the model. The results presented in this table show that SIFA was the most memory consuming method with the largest number of parameters, as it contains one GAN, one shared encoder with two decoders and three discriminators. CycleGAN was the second worst in terms of number of parameters and memory usage with the longest training time due to the requirement of more training epochs for model convergence. The inference time for all the adaptation methods is almost the same ($\sim$ 6s) except for cycleGAN (9.72s). For LMISA, the time for pre-processing (ML-NGM) was included.

Cardiac Dataset: The proposed method is also compared to other nine state-of-the-art methods based on the public multi-class cardiac dataset. The results of these methods were reported in one of the most recent domain adaptation publications (i.e. the SIFA paper by Chen *et al.*[15]). Re-implementing and evaluating these methods is not attempted, as it is fair to assume that the original authors have tried their best effort to tune the model parameters and achieved the best results using their methods. Except for the VarDA method [104], which was recently published. The original code is provided by the VarDA authors to produce the result for comparison. As the VarDA method was evaluated on the same public cardiac dataset in their paper, their default parameter settings are used in the experiment for testing the VarDA method.

Table 6.5: Segmentation results of cardiac dataset based on different methods. The results are reported based on Dice and the average symmetric surface distance (ASSD). Asterisk indicates the best 2D method and the overall best domain adaptation method is highlighted in bold. Note: N/A means that the ASSD cannot be calculated due to no prediction.

Method	Dice					ASSD				
	AA	**LAC**	**LVC**	**MYO**	**Average**	**AA**	**LAC**	**LVC**	**MYO**	**Average**
CT to MRI										
Supervised training (DRU-net)	90.9	88.5	93.7	85.2	89.5	1.4	1.8	0.7	1.1	1.2
Without Domain Adaptation	48.4	14.7	33.7	7.7	26.1	10.5	26.7	7.8	8.4	13.3
Without Domain Adaptation + MLNGM	31.9	42.4	60.9	29.8	41.2	11.3	15.1	4.2	4.5	8.7
PnP-AdaNet [13]	43.7	47.0	77.7	48.6	54.3	11.4	14.5	4.5	5.3	8.9
SynSeg-Net [95]	41.3	57.5	63.6	36.5	49.7	8.6	10.7	5.4	5.9	7.6
AdaOutput [99]	60.8	39.8	71.5	35.5	51.9	**5.7***	8.0*	4.6	4.6	5.7
CycleGAN [93]	64.3	30.7	65.0	43.0	50.7	5.8	9.8	6.0	5.0	6.6
CyCADA [96]	60.5	44.0	77.6	47.9	57.5	7.7	13.9	4.8	5.2	7.9
Prior SIFA [154]	**67.0***	60.7	75.1	45.8	62.1	6.2	9.8	4.4	4.4	6.2
SIFA [15]	65.3	62.3*	78.9	47.3	63.4	7.3	7.4	3.8	4.4	5.7
VarDA [104]	54.6	48.4	74.1	58.2	58.8	15.5	9.7	3.1	3.3	7.9
CIDA	45.1	30.5	67.1	45.2	46.9	7.1	9.6	5.2	7.4	7.3
LMISA-2D	61.6	60.0	84.1*	**65.4***	67.7*	8.8	8.2	2.2*	**2.0***	5.3*
LMISA-3D	60.7	**72.4**	**86.2**	64.1	**70.8**	6.5	**3.7**	**2.0**	2.3	**3.6**
MRI to CT										
Supervised training (DRU-net)	91.3	89.1	94.0	90.8	91.3	1.2	1.9	0.9	0.8	1.2
Without Domain Adaptation	0.0	73.2	60.3	45.0	32.6	N/A	12.1	9.5	4.8	N/A
Without Domain Adaptation + MLNGM	82.6	71.5	53.4	50.4	64.4	3.8	8.5	8.2	4.3	6.2
PnP-AdaNet [13]	74.0	68.9	61.9	50.8	63.9	12.8	6.3	17.4	14.7	12.8
SynSeg-Net [95]	71.6	69.0	51.6	40.8	58.2	11.7	7.8	7.0	9.2	8.9
AdaOutput [99]	65.2	76.6	54.4	43.6	59.9	17.9	5.5*	5.9	8.9	9.6
CycleGAN [93]	73.8	75.7	52.3	28.7	57.6	11.5	13.6	9.2	8.8	10.8
CyCADA [96]	72.9	77.0	62.4	45.3	64.4	9.6	8.0	9.6	10.5	9.4
Prior SIFA [154]	81.1	76.4	75.7	58.7	73.0	10.6	7.4	6.7	7.8	8.1
SIFA [15]	81.3	79.5*	73.8	61.6*	74.1*	7.9	6.2	5.5	8.5	7.0
VarDA [104]	81.9	66.1	56.8	42.1	61.7	8.1	12.6	4.6	3.2*	7.1
CIDA	72.4	65.3	58.2	45.3	60.3	6.0	11.7	5.8	4.6	7.0
LMISA-2D	83.5*	62.5	77.8*	60.5	71.0	3.1*	6.0	3.2*	3.5	3.9*
LMISA-3D	**84.5**	**82.8**	**88.6**	**70.1**	**81.5**	**2.0**	**3.9**	**1.5**	**1.8**	**2.3**

For the proposed method, the best result is reported for both LMIDA-2D and LMISA-3D, and used exactly the same test set as stated in the paper [15]. The Dice and ASSD were reported in the paper, hence listed in Table 6.5.

In Table 6.5, from CT to MRI, LMISA-2D achieved the best performance among all 2D methods in terms of average value of all classes for Dice and ASSD measures. LMISA-3D outperformed all the other methods based on LAC and LVC classes, as well as produced the best average values of all classes for both Dice and ASSD. However, based on AA organ, LMISA-2D and LMISA-3D ranked the forth and sixth place respectively for the Dice measurements. It is believed that this is due to the ground truth labeled region of AA in MRI is significantly larger than the AA region in CT images for most cases. With the shape constraint and minimal original intensity information used in the LMISA method, it could not accurately segment the extra AA

region in MRI which is not labeled in CT. When CT is the target, LMISA-2D ranked the 3rd place after SIFA and Prior SIFA in 2D methods based on Dice values. However, LMISA-2D was the best 2D method using ASSD measure. LMISA-3D achieved the best performance in all classes according to Dice and ASSD results. The visual results of LMISA-3D are also shown in Figure 6.11. For the visual results of other methods, the readers are encouraged to refer to Chen *et al.*[15].

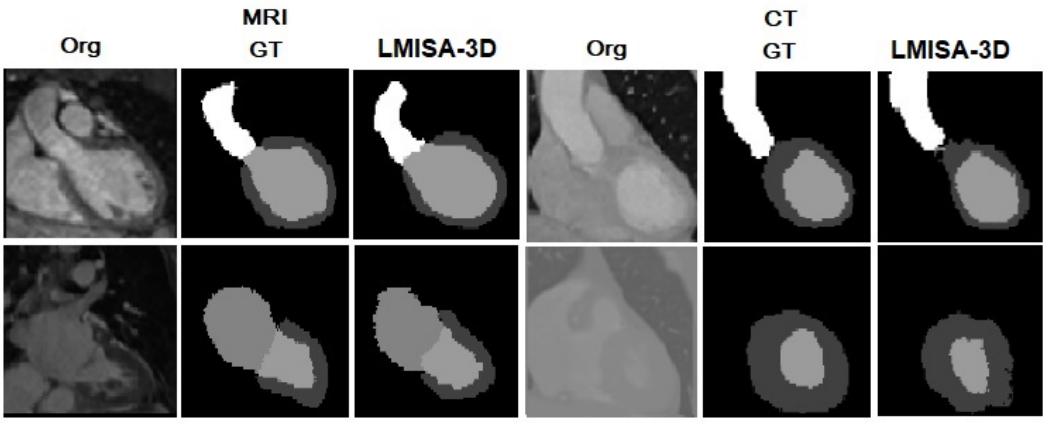


Figure 6.11: *Visual results of LMISA-3D using MRI and CT as the target domain respectively. The original image (Org), ground truth annotation (GT) and LMISA-3D prediction are shown for both MRI and CT as the target domain respectively.*

6.4 Conclusions and Discussion

A novel method is proposed to address the multi-modality image segmentation problem, in which the image labels are only available from one modality for model training. The proposed method utilises a simple and efficient pre-processing method to minimise the intensity difference between different domains. Next, a hybrid task encoder-decoder network is utilised to simultaneously learn the image segmentation in the source domain via supervised learning and learn a unified feature representation across domains using unsupervised adversarial learning. Further, a shape aware discriminator is applied to the model for guiding the shape of the predicted target segmentation to be similar to the shape of source ground truth. This is based on the assumption

that the same object of interest to be segmented and the object boundary is visible in the images of both domains. Additionally, the proposed method may be limited to the segmentation of regular shaped organs rather than objects with arbitrary shapes (e.g. vessels). In the experimental results, the robustness of the proposed method is demonstrated in both a binary class (kidney) and a multi-class (cardiac) image segmentation application. The proposed LMISA method has been implemented in both 2D and 3D versions, which have shown superior performance to other state-of-the-art methods in terms of segmentation accuracy. Remarkably, the proposed method remains the simplest model with significantly fewer model parameters, less memory consumption and faster training time. It is also noteworthy to emphasize that the proposed method can be easily extended to new unseen target domains (e.g. 3T to 1.5T MRI), even without model retraining. This superior characteristic has not been shown by other methods in the literature. The proposed method will be evaluated on more datasets with different image modalities in the future to further explore the advantages and limitations. In the next chapter, the limitations of the proposed methods and future works are elaborated.

Chapter 7

Conclusion

As mentioned in previous chapters, despite the great achievements of medical image segmentation based on deep learning, this task is still encountered difficulties in research. In summary, the segmentation accuracy is prone to failure when the model applied on imbalanced class samples data or on another image modalities. In addition, most of the methods are labour-intensive and complex. Therefore, a light-weight and robust network is required for when there is enough annotated data and when there is not enough annotated data. Aiming at the above problems, some methods are proposed in this thesis. This chapter summarises the proposed methods including the proposed FU-net, DRU-net, CIDA and LMISA. Besides, this section also discusses the limitations of the proposed approaches and provides directions for the future work.

7.1 Main Contributions and Limitations

In the following subsections, the proposed methods and limitations are explained.

7.1.1 FU-Net: Multi-class Image Segmentation using Feedback Weighted U-Net

FU-net is based on a well-established supervised end-to-end DCNN model, known as U-net. FU-net is proposed to improve the original U-net in two aspects: (1) improve the architecture of U-net by adding batch normalisation and residual block to achieve faster convergence, and (2) modify the cross-entropy loss function to a weighted version that allows the training to be performed based on a weighting scheme. The weighting scheme is calculated dynamically based on the pixel-wise prediction accuracy and updated iteratively in the training process (feedback weight) to avoid overwhelming the training process on easy examples. The proposed method was evaluated on a publicly available Skin lesion (ISIC 2018 grand-challenge), retinal image database (i.e. DRIVE), and a local MRI brain dataset. The proposed FU-net is compared with other methods. The evaluation results show that FU-net achieved significantly better performance than the other methods. FU-net is fast and easy to implement, and can be applied to real-time case scenarios. FU-net is similar to Focal loss [155], when Focal loss was published FU-net was also under the review. Besides, Focal loss is proposed for object detection, FU-net is for segmentation.

Limitations: Although the FU-net shows good results, there are several limitations of this approach as follows:

- This is working well when there are enough annotated data.
- It is accurate enough when the testing and training images are coming from the same distributions.
- The method cannot work well on different datasets, and need more modifications.
- In comparison with the original cross-entropy, it is more time-consuming.
- There is a need of setting a hyper-parameter.

7.1.2 DRU-net: An Efficient Deep Convolutional Neural Network for Medical Image Segmentation

Residual network (ResNet) and densely connected network (Dense- Net) have significantly increased the efficiency for training of deep DCNNs mainly for object classification tasks. Several studies have applied ResNet and DenseNet for different image segmentation problems. DRU-net is proposed as an efficient alternative network that is integrated into an encoder-decoder DCNN architecture for medical image segmentation. Based on ResNet, the proposed network adds more skip connections between all convolutional outputs to the final network output. This simple modification allows more feature maps to be combined and enables more efficient gradient backpropagation. DRU-net is evaluated on a challenging brain dataset, skin lesion dataset, and kidney including different modalities (i.e. CT, MRI). The results are compared with other methods such as ResNet and DenseNet, and attention net within the same encoder-decoder network structure. With similar number of network parameters, the proposed network achieved the best segmentation accuracy with less training time.

Limitations: DRU-net has some limitations as follows:

- This is an accurate and efficient method when there are enough annotated data.
- DRU-net perfoms well when the testing and training are from the same distributions.
- DRU-net is more time-consuming and requires more number of parameters in compare with RU-net and U-net.

7.1.3 CIDA: Cross-modality Image Segmentation via Domain Adaptation of Unpaired Images

In medical image segmentation, supervised machine learning model trained using one image modality (e.g. computed tomography image (CT)) is often prone to failure when it is applied to another image modality (e.g. magnetic resonance image (MRI)) even for the same type of organ. This is due to large geometric differences across different subjects and significant intensity variations of different image modalities. CIDA is a novel method for cross-modality image segmentation of unpaired images, where image labels of only one modality is available for model training. In this proposed method, firstly, Canny edge detection is used as a preprocessing step to be applied to images of both domains to remove the intensity discrepancy. Next, a GAN is utilised to learn the latent representation of the images of both domains in an unsupervised manner. A new loss term is added to the GAN loss for minimising the feature distances of both domains in the latent space, so that images from different domains are close to each other in the latent space. The learned encoder in GAN is reused as the encoder of an encoder-decoder segmentation network (i.e. U-net). Next, the segmentation network is trained using images and their corresponding labels of one domain only (source domain). Eventually, images from the other domain can be segmented, as they share similar latent feature representations with images of the learned domain. CIDA is assessed using a public CT dataset (KiTS19, MICCAI 2019) and a local MRI dataset for kidney segmentation. The segmentation model was trained only using CT images and able to segment kidneys in MRI images. The proposed CIDA achieved significantly higher performance in comparison with other state-of-the-art methods. This method can be applied to both 2D and 3D images, it is easy to implement, and fast to be trained in comparison with other state-of-the-art domain adaption based methods.

Limitations: The limitations of CIDA are listed as follows:

- Canny edge detection to reduce the intensity difference of the domains results in removing all the details and lead to performance drop.

- The framework of CIDA includes two different stages, one for reconstruction (training encoder and decoder) and another one for segmentation (training the decoder only).

7.1.4 LMISA: A Lightweight Multi-modality Image Segmentation Network via Domain Adaptation using Gradient Magnitude and Shape Constraint

LMISA is an extension of CIDA in terms of improving the performance and model complexity. LMISA is a novel end-to-end method to achieve multi-modality image segmentation of unpaired images. First, a multi-resolution locally normalised gradient magnitude approach is used on images of both domains for removing the intensity discrepancy. Then, a hybrid task encoder-decoder network including image segmentation and reconstruction is applied to effectively adapt a segmentation network to the unlabeled target domain. Additionally, a shape constraint is imposed by leveraging adversarial learning. Finally, images from the target domain can be segmented, as the network learns a consistent latent feature representation with shape awareness from both domains. LMISA is implemented in both 2D and 3D versions, which had been evaluated based on CT and MRI images for kidney and cardiac organ segmentation tasks. For kidney, a public CT dataset (KiTS19, MICCAI 2019) and a local MRI dataset were utilised, and for cardiac the Multi-Modality Whole Heart Segmentation (MMWHS) challenge 2017 dataset was used. Experimental results revealed that our proposed method achieved significantly higher performance with a much lower model complexity in comparison with other state-of-the-art methods.

Limitations: Although LMISA is so powerful, there are some limitations as follows:

- The domain adaptation assumes segmentation of the same organ, which can not be adapted to images without shared shapes. e.g. pathology image.

- Requires object edges to be identifiable using MLNGM.

7.2 Future Works

According to the limitations summarised above, to further improve the proposed methods and provide more investigations on other dataests, the future works are listed as follows:

- One of the future studies is related to FU-net to improve this method to achieve a better performance on the other datasets.

- FU-net can be combined to the multi-modality image segmentation methods, i.e. CIDA and LMISA.

- Another future work will be devoted to improving the performance of LMISA by not cropping the images and adding another stage LMISA as alignment. This can be performed by adding another module before LMISA to do the alignment of the images.

- Another investigation can be related to evaluation of LMISA on other multi-domain modalities such as Philips, GE, and Siemens.

- Finally, this thesis only applied the proposed CIDA and LMISA approaches to medical images. However, it may be useful for other field of studies too. Thus, from the perspective of application, as future work, CIDA and LMISA can be applied and evaluated to other case scenarios.

Bibliography

[1] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data*, vol. 5, no. 1, pp. 1–9, 2018.

[2] J. Staal, M. D. Abràmoff, M. Niemeijer, M. A. Viergever, and B. Van Ginneken, "Ridge-based vessel segmentation in color images of the retina," *IEEE transactions on medical imaging*, vol. 23, no. 4, pp. 501–509, 2004.

[3] S. Will, P. Martirosian, C. Würslin, and F. Schick, "Automated segmentation and volumetric analysis of renal cortex, medulla, and pelvis based on non-contrast-enhanced t1-and t2-weighted mr images," *Magnetic Resonance Materials in Physics, Biology and Medicine*, vol. 27, no. 5, pp. 445–454, 2014.

[4] S. T. Schwarz, Y. Xing, P. Tomar, N. Bajaj, and D. P. Auer, "In vivo assessment of brainstem depigmentation in parkinson disease: potential as a severity marker for multicenter studies," *Radiology*, vol. 283, no. 3, pp. 789–798, 2017.

[5] R. Kohler, "A segmentation system based on thresholding," *Computer Graphics and Image Processing*, vol. 15, no. 4, pp. 319–338, 1981.

[6] R. Adams and L. Bischof, "Seeded region growing," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 16, no. 6, pp. 641–647, 1994.

[7] U. Laube and M. E. Nebel, "Maximum likelihood analysis of the ford–fulkerson method on special graphs," *Algorithmica*, vol. 74, no. 4, pp. 1224–1266, 2016.

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

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

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

[11] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[12] M. Jafari, D. Auer, S. Francis, J. Garibaldi, and X. Chen, "Dru-net: An efficient deep convolutional neural network for medical image segmentation," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 2020, pp. 1144–1148.

[13] Q. Dou, C. Ouyang, C. Chen, H. Chen, B. Glocker, X. Zhuang, and P.-A. Heng, "Pnp-adanet: Plug-and-play adversarial domain adaptation network with a benchmark at cross-modality cardiac segmentation," *arXiv preprint arXiv:1812.07907*, 2018.

[14] X. Hou, J. Liu, B. Xu, B. Liu, X. Chen, M. Ilyas, I. Ellis, J. Garibaldi, and G. Qiu, "Dual adaptive pyramid network for cross-stain histopathology image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019, pp. 101–109.

[15] C. Chen, Q. Dou, H. Chen, J. Qin, and P. A. Heng, "Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation," *IEEE Transactions on Medical Imaging*, 2020.

[16] D. Marr and E. Hildreth, "Theory of edge detection," *Proceedings of the Royal Society of London. Series B. Biological Sciences*, vol. 207, no. 1167, pp. 187–217, 1980.

[17] J. Canny, "A computational approach to edge detection," *IEEE Transactions on pattern analysis and machine intelligence*, no. 6, pp. 679–698, 1986.

[18] D. Ziou, "Line detection using an optimal iir filter," *Pattern Recognition*, vol. 24, no. 6, pp. 465–478, 1991.

[19] S. Hojjatoleslami and J. Kittler, "Region growing: a new approach," *IEEE Transactions on Image processing*, vol. 7, no. 7, pp. 1079–1084, 1998.

[20] B. Cernuschi-Frias and D. B. Cooper, "3-d space location and orientation parameter estimation of lambertian spheres and cylinders from a single 2-d image by fitting lines and ellipses to thresholded data," *IEEE transactions on pattern analysis and machine intelligence*, no. 4, pp. 430–441, 1984.

[21] T. Matsuyama, "Expert systems for image processing: knowledge-based composition of image analysis processes," *Computer Vision, Graphics, and Image Processing*, vol. 48, no. 1, pp. 22–49, 1989.

[22] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[23] F. Visin, M. Ciccone, A. Romero, K. Kastner, K. Cho, Y. Bengio, M. Matteucci, and A. Courville, "Reseg: A recurrent neural network-based model for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 41–48.

[24] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3640–3649.

[25] G. B. Coleman and H. C. Andrews, "Image segmentation by clustering," *Proceedings of the IEEE*, vol. 67, no. 5, pp. 773–785, 1979.

[26] R. Katyal, S. Paneri, and M. Kuse, "Gaussian intensity model with neighborhood cues for fluid-tissue categorization of multisequence mr brain images," *Proceedings of the MICCAI Grand Challenge on MR Brain Image Segmentation (MRBrainS'13)*, 2013.

[27] B. Tsagaan, A. Shimizu, H. Kobatake, and K. Miyakawa, "An automated segmentation method of kidney using statistical information," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2002, pp. 556–563.

[28] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[29] P. Surlakar, S. Araujo, and K. M. Sundaram, "Comparative analysis of k-means and k-nearest neighbor image segmentation techniques," in *2016 IEEE 6th International Conference on Advanced Computing (IACC)*. IEEE, 2016, pp. 96–100.

[30] C. Jin, F. Shi, D. Xiang, X. Jiang, B. Zhang, X. Wang, W. Zhu, E. Gao, and X. Chen, "3d fast automatic segmentation of kidney based on modified aam and random forest," *IEEE transactions on medical imaging*, vol. 35, no. 6, pp. 1395–1407, 2016.

[31] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual review of biomedical engineering*, vol. 19, pp. 221–248, 2017.

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

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

[34] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[35] N. Baker, H. Lu, G. Erlikhman, and P. J. Kellman, "Deep convolutional networks do not classify based on global object shape," *PLoS computational biology*, vol. 14, no. 12, p. e1006613, 2018.

[36] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, vol. 1. Ieee, 2001, pp. I–I.

[37] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2009.

[38] S. Rota Bulo, G. Neuhold, and P. Kontschieder, "Loss max-pooling for semantic image segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2126–2135.

[39] D. Rengasamy, M. Jafari, B. Rothwell, X. Chen, and G. P. Figueredo, "Deep learning with dynamically weighted loss function for sensor-based prognostics and health management," *Sensors*, vol. 20, no. 3, p. 723, 2020.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[41] S. M. Sajjadian, M. Jafari, and X. Chen, "Automatic architectural drawing labelling using deep convolutional neural network," in *Sustainability in Energy and Buildings 2021*. Springer, 2022, pp. 69–78.

[42] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *European conference on computer vision*, 2016, pp. 443–450.

[43] C. S. Perone, P. Ballester, R. C. Barros, and J. Cohen-Adad, "Unsupervised domain adaptation for medical imaging segmentation with self-ensembling," *NeuroImage*, vol. 194, pp. 1–11, 2019.

[44] Y. Zhang, S. Miao, T. Mansi, and R. Liao, "Task driven generative modeling for unsupervised domain adaptation: Application to x-ray image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2018, pp. 599–607.

[45] H. Zhao, H. Li, S. Maurer-Stroh, Y. Guo, Q. Deng, and L. Cheng, "Supervised segmentation of un-annotated retinal fundus images by synthesis," *IEEE transactions on medical imaging*, vol. 38, pp. 46–56, 2018.

[46] J. Yang, N. C. Dvornek, F. Zhang, J. Chapiro, M. Lin, and J. S. Duncan, "Unsupervised domain adaptation via disentangled representations: Application to cross-modality liver segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019, pp. 255–263.

[47] C. Chen, Q. Dou, Y. Jin, H. Chen, J. Qin, and P.-A. Heng, "Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019, pp. 447–456.

[48] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

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

[50] Z. Hao, L. Lin, C. J. Post, E. A. Mikhailova, M. Li, Y. Chen, K. Yu, and J. Liu, "Automated tree-crown and height detection in a young forest plantation using mask region-based convolutional neural network (mask r-cnn)," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 178, pp. 112–123, 2021.

[51] S. U. Lee, S. Y. Chung, and R. H. Park, "A comparative performance study of several global thresholding techniques for segmentation," *Computer Vision, Graphics, and Image Processing*, vol. 52, no. 2, pp. 171–190, 1990.

[52] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *Journal of Electronic imaging*, vol. 13, no. 1, pp. 146–165, 2004.

[53] S. F. Gull and J. Skilling, "Maximum entropy method in image processing," in *IEE Proceedings F-Communications, Radar and Signal Processing*, vol. 131, no. 6. IET, 1984, pp. 646–659.

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

[55] W.-X. Kang, Q.-Q. Yang, and R.-P. Liang, "The comparative research on image segmentation algorithms," in *2009 First International Workshop on Education Technology and Computer Science*, vol. 2. IEEE, 2009, pp. 703–707.

[56] W. Gao, X. Zhang, L. Yang, and H. Liu, "An improved sobel edge detection," in *2010 3rd International conference on computer science and information technology*, vol. 5. IEEE, 2010, pp. 67–71.

[57] W. Rong, Z. Li, W. Zhang, and L. Sun, "An improved canny edge detection algorithm," in *2014 IEEE international conference on mechatronics and automation*. IEEE, 2014, pp. 577–582.

[58] L. J. Van Vliet, I. T. Young, and G. L. Beckers, "An edge detection model based on non-linear laplace filtering," in *Machine Intelligence and Pattern Recognition*. Elsevier, 1988, vol. 7, pp. 63–73.

[59] A. Mallick, S. Roy, S. S. Chaudhuri, and S. Roy, "Optimization of laplace of gaussian (log) filter for enhanced edge detection: A new approach," in *Proceedings of The 2014 International Conference on Control, Instrumentation, Energy and Communication (CIEC)*. IEEE, 2014, pp. 658–661.

[60] T. Lindeberg and M.-X. Li, "Segmentation and classification of edges using minimum description length approximation and complementary junction cues," *Computer Vision and Image Understanding*, vol. 67, no. 1, pp. 88–98, 1997.

[61] S. Hore, S. Chakraborty, S. Chatterjee, N. Dey, A. S. Ashour, L. Van Chung, and D.-N. Le, "An integrated interactive technique for image segmentation using stack based seeded region growing and thresholding." *International Journal of Electrical & Computer Engineering (2088-8708)*, vol. 6, no. 6, 2016.

[62] P. Asmussen, O. Conrad, A. Günther, M. Kirsch, and U. Riller, "Semi-automatic segmentation of petrographic thin section images using a "seeded-region growing algorithm" with an application to characterize wheathered subarkose sandstone," *Computers & Geosciences*, vol. 83, pp. 89–99, 2015.

[63] Z. Huang, X. Wang, J. Wang, W. Liu, and J. Wang, "Weakly-supervised semantic segmentation network with deep seeded region growing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7014–7023.

[64] A. Q. Al-Faris, U. K. Ngah, N. A. M. Isa, and I. L. Shuaib, "Breast mri tumour segmentation using modified automatic seeded region growing based on particle swarm optimization image clustering," in *Soft Computing in Industrial Applications*. Springer, 2014, pp. 49–60.

[65] L. Najman and H. Talbot, "Introduction to mathematical morphology," *Mathematical morphology: From theory to applications*, pp. 1–33, 2013.

[66] A. Serna and B. Marcotegui, "Detection, segmentation and classification of 3d urban objects using mathematical morphology and supervised learning," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 93, pp. 243–255, 2014.

[67] R. A. Husain, A. S. Zayed, W. M. Ahmed, and H. S. Elhaji, "Image segmentation with improved watershed algorithm using radial bases function neural networks," in *2015 16th International conference on sciences and techniques of automatic control and computer engineering (STA)*. IEEE, 2015, pp. 121–126.

[68] L. Vincent and P. Soille, "Watersheds in digital spaces: an efficient algorithm based on immersion simulations," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 13, no. 06, pp. 583–598, 1991.

[69] S. Beucher and F. Meyer, "The morphological approach to segmentation: the watershed transformation," in *Mathematical morphology in image processing*. CRC Press, 2018, pp. 433–481.

[70] J. Cousty, G. Bertrand, L. Najman, and M. Couprie, "Watershed cuts: Minimum spanning forests and the drop of water principle," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 8, pp. 1362–1374, 2008.

[71] R. Barnes, C. Lehman, and D. Mulla, "Priority-flood: An optimal depression-filling and watershed-labeling algorithm for digital elevation models," *Computers & Geosciences*, vol. 62, pp. 117–127, 2014.

[72] A. V. Goldberg and R. E. Tarjan, "Efficient maximum flow algorithms," *Communications of the ACM*, vol. 57, no. 8, pp. 82–89, 2014.

[73] T. Roughgarden, "Cs261: A second course in algorithms lecture# 3: The push-relabel algorithm for maximum flow," 2016.

[74] C. Wang, N. Komodakis, and N. Paragios, "Markov random field modeling, inference & learning in computer vision & image understanding: A survey," *Computer Vision and Image Understanding*, vol. 117, no. 11, pp. 1610–1627, 2013.

[75] Z. Ji, Y. Xia, Q. Sun, G. Cao, and Q. Chen, "Active contours driven by local likelihood image fitting energy for image segmentation," *Information Sciences*, vol. 301, pp. 285–304, 2015.

[76] S. Zhou, J. Wang, S. Zhang, Y. Liang, and Y. Gong, "Active contour model based on local and global intensity information for medical image segmentation," *Neurocomputing*, vol. 186, pp. 107–118, 2016.

[77] K. H. Cha, L. Hadjiiski, R. K. Samala, H.-P. Chan, E. M. Caoili, and R. H. Cohan, "Urinary bladder segmentation in ct urography using deep-learning convolutional neural network and level sets," *Medical physics*, vol. 43, no. 4, pp. 1882–1896, 2016.

[78] A. Tsai, A. Yezzi, W. Wells, C. Tempany, D. Tucker, A. Fan, W. E. Grimson, and A. Willsky, "A shape-based approach to the segmentation of medical imagery using level sets," *IEEE transactions on medical imaging*, vol. 22, no. 2, pp. 137–154, 2003.

[79] N. Paragios, "Curve propagation, level set methods and grouping," in *Handbook of mathematical models in computer vision*. Springer, 2006, pp. 145–159.

[80] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[81] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[82] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, 2017, pp. 240–248.

[83] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1915–1929, 2012.

[84] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.

[85] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," *Advances in neural information processing systems*, vol. 24, pp. 109–117, 2011.

[86] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation," in *Deep learning and data labeling for medical applications*. Springer, 2016, pp. 179–187.

[87] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2016, pp. 424–432.

[88] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 565–571.

[89] X. Xiao, S. Lian, Z. Luo, and S. Li, "Weighted res-unet for high-quality retina vessel segmentation," in *2018 9th international conference on information technology in medicine and education (ITME)*. IEEE, 2018, pp. 327–331.

[90] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes," *IEEE transactions on medical imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.

[91] N. Ibtehaz and M. S. Rahman, "Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation," *Neural Networks*, vol. 121, pp. 74–87, 2020.

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

[93] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[94] Z. Zhang, L. Yang, and Y. Zheng, "Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9242–9251.

[95] Y. Huo, Z. Xu, H. Moon, S. Bao, A. Assad, T. K. Moyo, M. R. Savona, R. G. Abramson, and B. A. Landman, "Synseg-net: Synthetic segmentation without target modality ground truth," *IEEE transactions on medical imaging*, vol. 38, pp. 1016–1025, 2018.

[96] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *International conference on machine learning*, 2018, pp. 1989–1998.

[97] J. Jiang, Y.-C. Hu, N. Tyagi, P. Zhang, A. Rimner, G. S. Mageras, J. O. Deasy, and H. Veeraraghavan, "Tumor-aware, adversarial domain adaptation from ct to mri for lung cancer segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2018, pp. 777–785.

[98] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7167–7176.

[99] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7472–7481.

[100] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li, "Deep reconstruction-classification networks for unsupervised domain adaptation," in *European Conference on Computer Vision*, 2016, pp. 597–613.

[101] A. Rozantsev, M. Salzmann, and P. Fua, "Beyond sharing weights for deep domain adaptation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, pp. 801–814, 2018.

[102] B. Bhushan Damodaran, B. Kellenberger, R. Flamary, D. Tuia, and N. Courty, "Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 447–463.

[103] F. Wu and X. Zhuang, "Cf distance: A new domain discrepancy metric and application to explicit domain adaptation for cross-modality cardiac image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 12, pp. 4274–4285, 2020.

[104] ——, "Unsupervised domain adaptation with variational approximation for cardiac segmentation," *IEEE Transactions on Medical Imaging*, 2021.

[105] C. Pei, F. Wu, L. Huang, and X. Zhuang, "Disentangle domain features for cross-modality cardiac image segmentation," *Medical Image Analysis*, vol. 71, p. 102078, 2021.

[106] A. Mansoor, J. J. Cerrolaza, R. Idrees, E. Biggs, M. A. Alsharid, R. A. Avery, and M. G. Linguraru, "Deep learning guided partitioned shape model for anterior visual pathway segmentation," *IEEE transactions on medical imaging*, vol. 35, no. 8, pp. 1856–1865, 2016.

[107] S. A. Eslami, N. Heess, C. K. Williams, and J. Winn, "The shape boltzmann machine: a strong model of object shape," *International Journal of Computer Vision*, vol. 107, no. 2, pp. 155–176, 2014.

[108] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3d shapenets: A deep representation for volumetric shapes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1912–1920.

[109] K. Li, B. Hariharan, and J. Malik, "Iterative instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3659–3667.

[110] H. Ravishankar, R. Venkataramani, S. Thiruvenkadam, P. Sudhakar, and V. Vaidya, "Learning and incorporating shape models for semantic segmentation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2017, pp. 203–211.

[111] A. BenTaieb and G. Hamarneh, "Topology aware fully convolutional networks for histology gland segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 460–468.

[112] H. Chen, X. Qi, L. Yu, Q. Dou, J. Qin, and P.-A. Heng, "Dcan: Deep contour-aware networks for object instance segmentation from histology images," *Medical image analysis*, vol. 36, pp. 135–146, 2017.

[113] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. De Marvao, T. Dawes, D. P. O'Regan *et al.*, "Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation," *IEEE transactions on medical imaging*, vol. 37, no. 2, pp. 384–395, 2017.

[114] T. F. Cootes and C. J. Taylor, "Combining point distribution models with shape models based on finite element analysis," *Image and Vision Computing*, vol. 13, no. 5, pp. 403–409, 1995.

[115] C. Davatzikos, X. Tao, and D. Shen, "Hierarchical active shape models, using the wavelet transform," *IEEE transactions on medical imaging*, vol. 22, no. 3, pp. 414–423, 2003.

[116] M. Avendi, A. Kheradvar, and H. Jafarkhani, "A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac mri," *Medical image analysis*, vol. 30, pp. 108–119, 2016.

[117] M. Tang, S. Valipour, Z. Zhang, D. Cobzas, and M. Jagersand, "A deep level set method for image segmentation," in *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, 2017, pp. 126–134.

[118] S. M. R. Al Arif, K. Knapp, and G. Slabaugh, "Spnet: Shape prediction using a fully convolutional neural network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 430–439.

[119] O. U. Aydin, A. A. Taha, A. Hilbert, A. A. Khalil, I. Galinovic, J. B. Fiebach, D. Frey, and V. I. Madai, "On the usage of average hausdorff distance for segmentation performance assessment: hidden error when used for ranking," *European Radiology Experimental*, vol. 5, no. 1, pp. 1–7, 2021.

[120] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Transactions on image processing*, vol. 10, no. 2, pp. 266–277, 2001.

[121] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 11, pp. 1222–1239, 2001.

[122] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation," 2018.

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

[124] J. Zhuang, "Laddernet: Multi-path networks based on u-net for medical image segmentation," *arXiv preprint arXiv:1810.07810*, 2018.

[125] K.-K. Sung, "Learning and example selection for object and pattern detection," 1996.

[126] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 142–158, 2015.

[127] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[128] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 761–769.

[129] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[130] M. Jafari, R. Li, Y. Xing, D. Auer, S. Francis, J. Garibaldi, and X. Chen, "Fu-net: multi-class image segmentation using feedback weighted u-net," in *International Conference on Image and Graphics*. Springer, 2019, pp. 529–537.

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

[132] N. Abraham and N. M. Khan, "A novel focal tversky loss function with improved attention u-net for lesion segmentation," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 683–687.

[133] C. Wang, T. MacGillivray, G. Macnaught, G. Yang, and D. Newby, "A two-stage 3d unet framework for multi-class segmentation on full resolution image," *arXiv preprint arXiv:1804.04341*, 2018.

[134] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

[135] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[136] J. Zhang, Y. Jin, J. Xu, X. Xu, and Y. Zhang, "Mdu-net: Multi-scale densely connected u-net for biomedical image segmentation," *arXiv preprint arXiv:1812.00352*, 2018.

[137] N. Heller, N. Sathianathen, A. Kalapara, E. Walczak, K. Moore, H. Kaluzniak, J. Rosenberg, P. Blake, Z. Rengel, M. Oestreich *et al.*, "The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes," *arXiv preprint arXiv:1904.00445*, 2019.

[138] E. Gibson, Y. Hu, N. Ghavami, H. U. Ahmed, C. Moore, M. Emberton, H. J. Huisman, and D. C. Barratt, "Inter-site variability in prostate segmentation accuracy using deep learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2018, pp. 506–514.

[139] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle, "Brain tumor segmentation with deep neural networks," *Medical image analysis*, vol. 35, pp. 18–31, 2017.

[140] T. Joyce, A. Chartsias, and S. A. Tsaftaris, "Robust multi-modal mr image synthesis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2017, pp. 347–355.

[141] M. Ghafoorian, A. Mehrtash, T. Kapur, N. Karssemeijer, E. Marchiori, M. Pesteie, C. R. Guttmann, F.-E. de Leeuw, C. M. Tempany, B. van Ginneken *et al.*, "Transfer learning for domain adaptation in mri: Application in brain lesion segmentation," in *International conference on medical image computing and computer-assisted intervention*, 2017, pp. 516–524.

[142] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[143] M. Kan, S. Shan, and X. Chen, "Bi-shifting auto-encoder for unsupervised domain adaptation," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3846–3854.

[144] G. Parthasarathy, L. Ramanathan, K. Anitha, and Y. Justindhas, "Predicting source and age of brain tumor using canny edge detection algorithm and threshold technique," *Asian Pacific Journal of Cancer Prevention: APJCP*, vol. 20, no. 5, p. 1409, 2019.

[145] H.-Y. Lee, H.-Y. Tseng, Q. Mao, J.-B. Huang, Y.-D. Lu, M. Singh, and M.-H. Yang, "Drit++: Diverse image-to-image translation via disentangled representations," *International Journal of Computer Vision*, pp. 1–16, 2020.

[146] J. Choi, T. Kim, and C. Kim, "Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6830–6840.

[147] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, pp. 2579–2605, 2008.

[148] M. Jafari, S. Francis, J. M. Garibaldi, and X. Chen, "Lmisa: A lightweight multi-modality image segmentation network via domain adaptation using gradient magnitude and shape constraint," *Medical Image Analysis*, vol. 81, p. 102536, 2022.

[149] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. Ieee, 2005, pp. 886–893.

[150] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Advances in neural information processing systems*, 2017, pp. 5767–5777.

[151] D. Bank, N. Koenigstein, and R. Giryes, "Autoencoders," *arXiv preprint arXiv:2003.05991*, 2020.

[152] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, pp. 2096–2030, 2016.

[153] X. Zhuang and J. Shen, "Multi-scale patch and multi-modality atlases for whole heart segmentation of mri," *Medical image analysis*, vol. 31, pp. 77–87, 2016.

[154] 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," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, pp. 865–872.

[155] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

Appendix A

List of Abbreviations

CNN	Convolutional Neural Network
DCNN	Deep Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long-Short Term Memory
GAN	Generative Adversarial network
SAE	Stacked Auto Encoder
FC	Fully Connection
RBM	Restricted Boltzmann Machine
CRF	Conditional Random Fields
PCA	Principal Component Analysis
HOG	Histogram of Oriented Gradients
SVM	Support Vector Machine
IoU	Intersection over Union
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
CT	Computed Tomography
MRI	Magnetic Resonance Imaging
AKI	Acute Kidney Injury
CKD	Chronic Kidney Disease