BLURRY BOUNDARY DELINEATION AND ADVERSARIAL CONFIDENCE LEARNING FOR MEDICAL IMAGE ANALYSIS

Dong Nie

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Computer Science.

Chapel Hill
2019

Approved by:

Dinggang Shen

Marc Niethammer

Martin Styner

Jun Lian

Jan-Michael Frahm

ABSTRACT

Dong Nie: Blurry Boundary Delineation and Adversarial Confidence Learning for Medical Image Analysis
(Under the direction of Dinggang Shen)

Low tissue contrast and fuzzy boundaries are major challenges in medical image segmentation which is a key step for various medical image analysis tasks. In particular, blurry boundary delineation is one of the most challenging problems due to low-contrast and even vanishing boundaries. Currently, encoder-decoder networks are widely adopted for medical image segmentation. With the lateral skip connection, the models can obtain and fuse both semantic and resolution information in deep layers to achieve more accurate segmentation performance. However, in many applications (*e.g.*, images with blurry boundaries), these models often cannot precisely locate complex boundaries and segment tiny isolated parts. To solve this challenging problem, we empirically analyze why simple lateral connections in encoder-decoder architectures are not able to accurately locate indistinct boundaries. Based on the analysis, we argue learning high-resolution semantic information in the lateral connection can better delineate the blurry boundaries. Two methods have been proposed to achieve such a goal. a) A high-resolution pathway composed of dilated residual blocks has been adopted to replace the simple lateral connection for learning the high-resolution semantic features. b) A semantic-guided encoder feature learning strategy is further proposed to learn high-resolution semantic encoder features so that we can more accurately and efficiently locate the blurry boundaries. Besides, we also explore a contour constraint mechanism to model blurry boundary detection. Experimental results on real clinical datasets (infant brain MRI and pelvic organ datasets) show that our proposed methods can achieve state-of-the-art segmentation accuracy, especially for the blurry regions. Further analysis also indicates that our proposed network components indeed contribute to the performance gain. Experiments on an extra dataset also validate the generalization ability of our proposed methods.

Generative adversarial networks (GANs) are widely used in medical image analysis tasks, such as medical image segmentation and synthesis. In these works, adversarial learning is usually directly applied to the original supervised segmentation (synthesis) networks. The use of adversarial learning is effective in improving visual perception performance since adversarial learning works as realistic regularization for supervised generators. However, the quantitative performance often cannot be improved as much as the qualitative performance, and it can even become worse in some cases. In this dissertation, I explore how adversarial learning could be more useful in supervised segmentation (synthesis) models, *i.e.*, how to synchronously improve visual and quantitative performance. I first analyze the roles of discriminator in the classic GANs and compare them with those in supervised adversarial systems. Based on this analysis, an adversarial confidence learning framework is proposed for taking better advantage of adversarial learning; that is, besides the adversarial learning for emphasizing visual perception, the confidence information provided by the adversarial network is utilized to enhance the design of the supervised segmentation (synthesis) network. In particular, I propose using a fully convolutional adversarial network for confidence learning to provide voxel-wise and region-wise confidence information for the segmentation (synthesis) network. Furthermore, various loss functions of GANs are investigated and the binary cross entropy loss is finally chosen to train the proposed adversarial confidence learning system so that the modeling capacity of the discriminator is retained for confidence learning. With these settings, two machine learning algorithms are proposed to solve some specific medical image analysis problems. a) A difficulty-aware attention mechanism is proposed to properly handle hard samples or regions by taking structural information into consideration so that the irregular distribution of medical data could be appropriately dealt with. Experimental results on clinical and challenge datasets show that the proposed algorithm can achieve state-of-the-art segmentation (synthesis) accuracy. Further analysis also indicates that adversarial confidence learning can synchronously improve the visual perception and quantitative performance. b) A semi-supervised segmentation model is proposed to alleviate the everlasting challenge for medical image segmentation - lack of annotated data. The proposed method can automatically recognize well-segmented regions (instead of the entire sample) and dynamically include them to increase the label set during training. Specifically, based on the confidence map, a region-attention based semi-supervised learning strategy is designed to further train the segmentation network. Experimental results on real clinical datasets show that the proposed approach can achieve better segmentation performance with extra unannotated data.

To someone

ACKNOWLEDGEMENTS

First, I would like to express my sincere gratitude to my advisor Dr. Dinggang Shen for his continuous support of my PhD study during the past four years. I am grateful to Dinggang for introducing me to the field of medical image analysis and for largely improving and refining my writing skills. I am also grateful to him for giving me freedom to explore my own research interest and pursue various career opportunities. I would like to express my special thanks to Dinggang for his implicit influence on me which has turned me to a dedicated and patient person. His guidance, encouragement and support are always priceless to me.

I would also like to thank my committee chair, Dr. Marc Niethammar, for his support to my academic and graduation affairs, his patient revision for my dissertation, and especially for his excellent courses. Actually, my knowledge about image processing is mainly learned from Marc's courses. Also, I would like to thank Dr. Martin Styner and Dr. Jan-Michael Frahm for their suggestions and helps all the way during my graduation. Specifically, Dr. Martin Styner has been very helpful in making me to work towards understanding-oriented research and in letting me know the bases of imaging technologies. Dr. Jan-Michael Frahm is a very nice professor and he has taught me the basic knowledge about computer vision in his course. Dr. Jun Lian is not only a committee member and a tutor but also a friend to me. I indeed thanks for his kindness in teaching me the knowledge about image-guided radiotherapy, for his suggestions about career development, and for his encouragement when I fall in troubles.

Besides, I would like to thank the co-authors and co-workers of my publications, including Dr. Li Wang, Dr. Han Zhang, Dr. Yaozong Gao, Dr. Qian Wang, Dr. Ehsan Adeli, Dr. Pew-Thian Yap, Dr. Gang Li, Dr. Mingxia Liu and all the people who ever worked in the IDEA lab of BRIC. In particular, Li taught me a lot in the first year when I participated in IDEA lab; Han helped extend my work on 3D neural networks to brain tumor applications; Yaozong, Qian and Ehsan helped me a lot for revising the research paper and providing invaluable discussions on my research work; Yap, Gang and Mingxia taught me a lot for preparing good research talks and provided me many useful suggestions for the dissertation.

I am also lucky and grateful to have many friends in our IDEA lab, including, Roger Trullo, Ehsan Adeli, Jialin Peng, Lin Wang, Xiaofen Ma, Lichi Zhang, Yongqin Zhang, Kim Han Thung, Shu Liao, Sihang Zhou, Rico Zhang, Xiaohuan Cao, Lei Xiang, Luyan Liu, Renping Yu, Dingna Duan, Zhaoyu Li, Longwei Fang, Xuyun Wen, Yujie Liu, Yanting Zheng, Xuhua Ren, KeLei He, Shuai Wang, Xiaoxia Zhang, Zhengwang Wu, Deqiang Xiao, Geng Chen, Yoonmi Hong, Yu Zhang, Chunfeng Lian, Jie Xue, Shujun Liang, Guannan Li, Xiaodan Sui and many others. Without you, my PhD life in the lab will be eclipsed.

I would like to thank my fellow classmates and friends in UNC, who helped me get through these years. To name a few, Qiuyu Xiao, Xu Han, Lei Huang, Licheng Yu, Yipin Zhou, Ziqiao Zhou, Sheng Liu, Enliang Zheng, Hongkun Ge, Yue Guo, Zhipeng Ding, Qian Zhang, Zhengyang Shen, Ruibing Ma, Yang Li, Zhengyang Fang etc. I would like to give my special thanks to Yaozong Gao, Yu Meng, Peiyao Wang, Zhenghan Fang and Xiaoyang Chen who worked with me in the same lab. Without you, I may feel lonely working as a research assistant outside the computer science department.

I would also like to express my special thanks to those particular persons. I would like to thank Dr. Vladimir Jojic for offering me the opportunity to study at University of North Carolina and supporting my research in the first year. I would like to thank Dr. Tian Cao for discussing with me to choose medical image analysis as my research topic after the first year of study in UNC, and for providing me numerous guidance for research and life. I would like to thank Dr. Xiaofen Ma for driving me to emergency room and taking care of me for several times.

Last, but no the least, I would like to thank my parents, Mr. Shengxue Nie and Mrs. Yuhong Yu, for their everlasting support and love throughout the years. I would like to give my speical thanks to my wife, Dr. Lingzi Hong, for always supporting me, encouraging me and loving me in all these years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

2D	Two Dimensional
3D	Three Dimensional
ACM	Auto-Context Model
ASD	Average Surface Distance
CSF	Cerebrospinal Fluid
CT	Computed Tomography
DSC	Dice Similarity Coefficient
FCN	Fully Convolutional Network
FOV	Field of View
GAN	Generative Adversarial Network
GDL	Gradient Difference Loss
GM	Gray Matter
MAE	Mean Absolute Error
MR	Magnetic Resonance
MRI	Magnetic Resonance Imaging
MSE	Mean Squared Error
PSNR	Peak Signal to Noise Ratio
RF	Random Forest
SSIM	Structure Similarity Index
SOTA	State-of-the-Art
SR	Sparse Representation
WM	White Matter

CHAPTER 1: INTRODUCTION

1.1 Medical Image Segmentation

1.1.1 Motivation and Challenges

The goal of image segmentation is to divide an image into a set of semantically meaningful non-overlapping regions of similar attributes, such as intensity and texture [45]. The segmentation result can be represented as an image of labels identifying homogeneous regions or a set of contours describing region boundaries. Segmentation is one of the fundamental problems in medical image analysis and is critical for tasks [191, 185, 134, 50, 194], such as quantitative analysis of tissue volume, diagnosis, localization of diseased tissue, study of anatomical structures, treatment planning, partial volume effect correction of fMRI data, calculation of functional imaging data, computer guided surgery and so on. However, manual segmentation is tedious because it usually involves voxel-wise annotation. Moreover, medical image annotation requires expert knowledge, unlike annotation of natural images. Therefore, efficient and automatic segmentation methods are often desired.

However, it is very difficult to automatically segment the tissues/organs from medical images due to various challenges, which can be categorized as follows:

1. Low contrast: Low tissue contrast is a major hindrance to effective medical image segmentation. For instance, the tissue or organ boundaries usually exhibit extremely low contrast, making it very difficult to reasonably delineate these boundaries (*e.g.*, infant brain MRI in Fig. 1.1 and the prostate boundaries in Fig. 1.2);

2. Noise: Noise increases uncertainty and hence the difficulty of image segmentation (*e.g.*, infant brain MRI in Fig. 1.1);

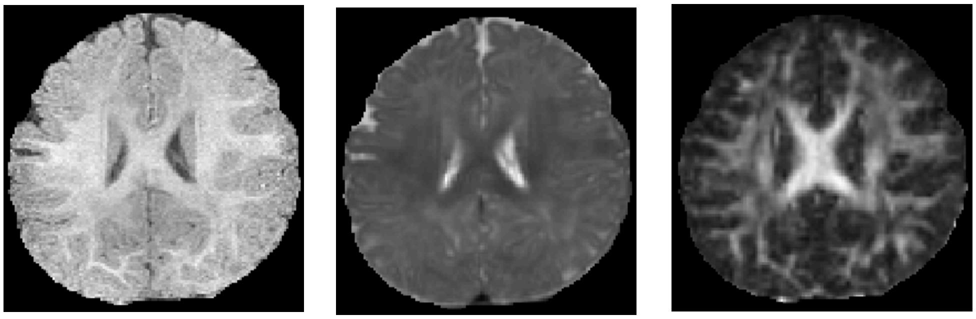


Figure 1.1: Multi-modality MRI data of an infant subject scanned at 6 months old (isointense phase). From left to right: T1-weighted, T2-weighted, and FA image.

3. Inhomogeneous image contrast: Different regions of tissues/organs can exhibit inhomogeneous image contrast that may potentially confuse the segmentation model (*e.g.*, the rectum in Fig. 1.2);

4. Large shape variability: The shapes of some organs (*e.g.*, the bladder and prostate in Fig. 1.2) can vary significantly across different subjects or even for the same subject across time, creating problems for methods based on shape priors;

5. Lack of annotated data: Supervised models for segmentation, especially deep networks, usually require large annotated datasets, which can be difficult to obtain for medical images;

6. Sample imbalance: There are two kinds of sample imbalance issues when it comes to segmentation: a) the number of voxels (pixels) of one class dominates over other classes; b) the number of samples with regular distribution (*e.g.*, normal voxels) dominates over that with irregular distribution (*e.g.*, rarely appeared voxels).

Deep segmentation networks have been shown to be able to partially solve these mentioned challenges [164, 3, 171]. In this dissertation, I will mainly focus on the following three specific challenges that have not been sufficiently addressed:

1. **Blurry Boundaries:** Medical images, such as MRI and CT, sometimes have blurry and vanishing boundaries, *i.e.*, the pelvic area (see Fig. 1.2) and the white matter

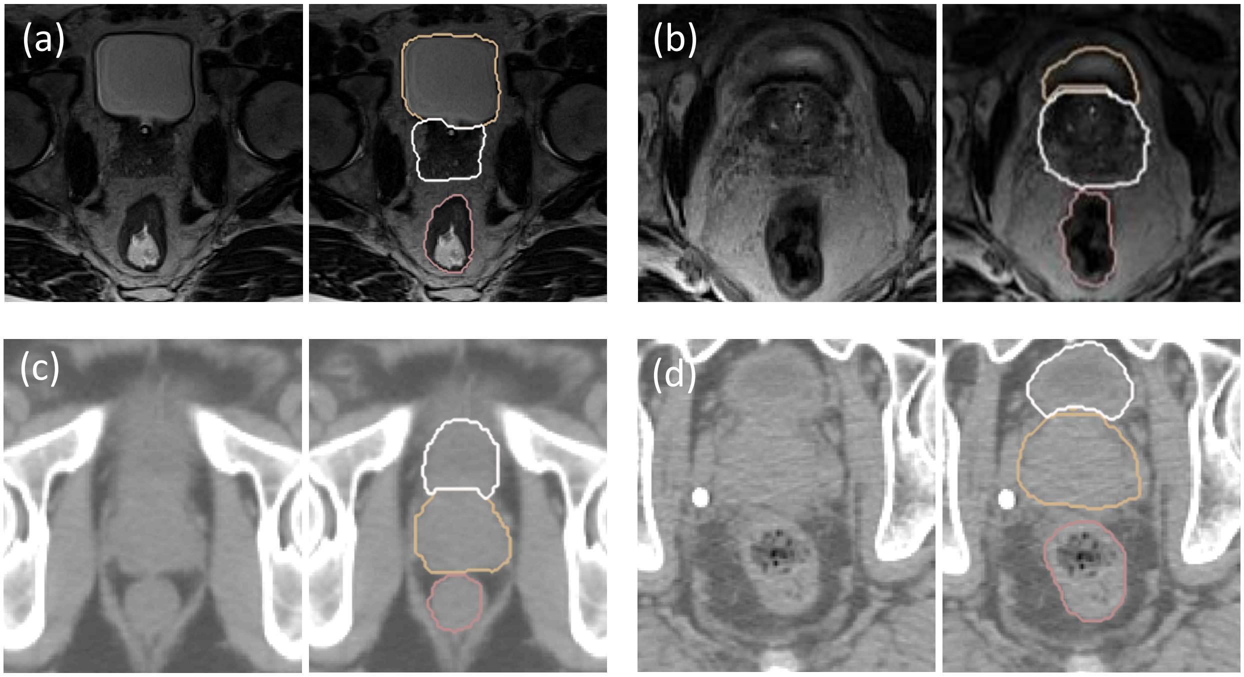


Figure 1.2: (a) and (b) are typical pelvic MRI and their corresponding manual segmentations of bladder (orange), prostate (silver), and rectum (pink), where the two columns in each panel show an MRI slice and the same slice overlaid with manual segmentations. (c) and (d) are two typical pelvic CT images and their corresponding manual segmentations.

and gray matter (see Fig. 1.1). This poses severe challenges for image segmentation algorithms. The regions around organ boundaries of medical images sometimes lack rich and stable texture information, especially for soft tissues. As a consequence, different organs can be labeled as one (*i.e.*, shown by (a) and (c) in Fig. 1.2), while a single organ can be split into multiple parts (*i.e.*, shown by (b) and (d) in Fig. 1.2). The clues for correct localization of boundaries can be unreliable (see Fig. 1.2).

2. **Lack of Labeled Data:** Training supervised segmentation (synthesis) models usually requires a large amount of labeled data, which can be difficult to obtain due to the following factors: a) Unlike annotation of natural images, annotation of medical images requires expert knowledge; b) It is time-consuming and tedious to annotate pixel-wise (voxel-wise) since medical images are usually three-dimensional; c) Anno-

tations can vary significantly between observers and even for the same observer at different time points.

3. **Easy-Sample Dominance:** Learning-based models can be easily dominated by easy samples and cannot properly handle hard samples.

1.1.2 Previous Work

Medical image segmentation is a longstanding problem in the medical imaging community. Segmentation algorithms can be categorized as 1) unsupervised model-based methods; 2) multi-atlas methods; 3) conventional learning-based methods and 4) deep learning based methods. These methods are briefly described below.

1.1.2.1 Unsupervised Model-Based Segmentation Methods

Otsu *et al.* [156] developed a component-specific thresholding algorithm for image segmentation. Chan *et al.* [26] further enforced spatial regularization for segmentation based on the Otsu method. Kass *et al.* [96] proposed an energy based active contour model, called snakes, for contour delineation. Pizer *et al.* [159] proposed a shape model named M-reps for segmenting the prostate, bladder and rectum from pelvic images. Li *et al.* [113] proposed a level set evolution method for boundary detection. Unger *et al.* [184] proposed total-variation based method to interactively segment the region of interest from the image. Lucchi *et al.* [126] proposed a superpixel method to segment irregular shape of cells. Markov random fields (MRFs) are also utilized to segment brain images [75]. Ian *et al.* [178] proposed a graph cut algorithm to extract the prostate surface. Rother *et al.* [165] further proposed a graph cut to interactively segment objects from images. Conditional random fields (CRFs) are widely used to segment medical images [17]. The main limitation of these model-based algorithms is that they cannot handle well sophisticated images (*e.g.*, low-contrast images). Also, many of these approaches are time consuming, limiting their use in many scenarios.

1.1.2.2 Multi-Atlas Segmentation Methods

In multi-atlas methods, segmentation labels are transferred and fused from multiple expert-labeled atlases to a target image [163, 101, 72, 168]. Initially, single-atlas based methods were developed [36, 44]. They are however quite sensitive to the choice of the atlas. Multi-atlas based methods were then proposed to solve this issue [163, 101, 72, 168]. Most of the atlas based segmentation algorithms concentrate on the design of sophisticated atlas selection or the mechanism of label fusion. For instance, Yan *et al.* [206] proposed an atlas selection with a label constraining and label fusion method to segment prostate MRs. During the atlas selection, label images are used to constrain the manifold projection (*i.e.*, to project a image to a point in the manifold space) of intensity images, which can alleviate the misleading projection due to other anatomical structures. Ou *et al.* [157] proposed to gradually (*i.e.*, in a cascade manner) improve the registration based on the prostate vicinity between the target and atlas images for iteratively carrying out the multi-atlas label fusion. Shan *et al.* [168] proposed a multi-atlas based segmentation method with non-local patch-based label fusion to segment MR knee images. The main issue of these algorithms is the high computation cost and sensitivity to registration accuracy.

1.1.2.3 Conventional Learning-based Segmentation Methods

Segmentation can be formulated as an optimization problem to find the best shape model for fitting the target image. This requires the definition of image-to-image similarity measures, which often requires careful feature engineering. Ayachi *et al.* [7] adopted a support vector machine (SVM) [28] for brain tumor segmentation with intensity and texture features. Toth *et al.* [180] proposed to incorporate different features in the context of active appearance models (AAMs) to improve the prostate segmentation performance.

Wang *et al*. [192] proposed a general framework that adopts a sparse representation to fuse multi-modality image information with the anatomical constraints for brain tissue segmentation. Gao *et al*. [57] proposed sparse representation based classification method to segment prostate from CT images. Wang *et al*. [190] proposed to integrate information from multi-source images together for an accurate tissue segmentation by combining random forest and auto-context model [181] with Haar features. Dictionary learning is also a widely adopted method to segment the medical images [179] with intensity features or other well-designed features. However, all these methods require well-designed features. In addition, the feature learning procedure is not directly optimized towards the classification process, which could largely suppress the power of the whole system.

1.1.2.4 Deep Learning-Based Segmentation Methods

Fully convolutional networks (FCN) [124], a variant of convolutional neural networks (CNN), is a recently common choice for semantic image segmentation in computer vision. FCN trains a neural network in an end-to-end fashion without using fully connected layers as in CNN by directly optimizing intermediate feature layers for segmentation, making it outperform traditional methods that often regard the feature learning and segmentation as two separate tasks. Apart from computer vision, FCN-based methods have also shown great success in medical image segmentation [139, 153, 29, 214]. However, FCNs (Note, FCNs in this dissertation means the original FCN-based networks, not including the UNet or dilated FCNs) cannot perform well for localization precision due to the designed pooling layers in this architecture. To extend FCNs and address the drawbacks of FCNs, lots of works have been proposed. Generally, these works can be categorized into two mainstreams: 1) encoder-decoder architectures and 2) dilated FCNs.

1. **Encoder-decoder architectures:** The typical FCN based encoder-decoder architecture is UNet [164], which is an evolutionary variant of FCN and has also achieved excellent performance in many tasks by effectively combining high-level and low-level

features in the network architecture. Compared to FCN, UNet can improve the localization accuracy near organ boundaries. Nie *et al.* [151] designed a transformation module and fusion module to alleviate the bias effect (*i.e.*, the information from the encoder (shallow) layers is quite different from the decoder (deep) layers) during information fusion. Similarly, Milletari *et al.* [139] proposed VNet using residual module and a Dice loss to improve the segmentation performance. Lin *et al.* [121] introduced a well-designed encoder-decoder architecture to fuse the high-resolution feature maps from the encoder pathway and the highly semantic feature maps from the decoder pathway in the 'RefineNet'. Generally, while effective, all these methods depend on the information from the lower layers to provide localization precision.

2. **Dilated FCNs:** The typical work for this category is the 'Deeplab' series [32, 33, 34], in which, atrous convolution is proposed to replace the pooling layers to increase the theoretical receptive field fast (in this way, the practical receptive field can become large enough fast), so that the localization precision (because we do not need to use pooling operations in this system) can be improved without losing classification accuracy. Chen *et al.* [33] further implemented an atrous spatial pyramid pooling module to increase the context information in a multi-scale manner and applied Dense Conditional Random Fields (Dense-CRF) [104] to refine the segmentation results. In the recent 'PSPNet', Zhao *et al.* [227] proposed a pyramid pooling module to aggregate the background (context) information and auxiliary losses to intermediately supervise the segmentation task.

1.1.2.5 Semi-Supervised Learning for Deep Segmentation Networks

Semi-supervised learning is a promising solution to address the aforementioned lack of labeled data issue [13, 14, 221, 200]. To relieve the demand for large-scale labeled data, Bai *et al.* [13] proposed a semi-supervised deep learning framework for cardiac MR image segmentation, in which the automatically segmented label maps from unlabeled data are incrementally included into the training set to refine the network. Baur *et al.* [14] introduced auxiliary manifold embedding (to minimize the discrepancy between similar inputs for both labeled and unlabeled data) in the latent space to FCN for semi-supervised learning for MS lesion segmentation. Zhang *et al.* [221] proposed a new deep adversarial network model to attain consistently good segmentation results on both annotated and unannotated images for biomedical image segmentation. Xiao *et al.* [200] proposed a semi-supervised segmentation method combined with transfer learning which transfers the learned knowledge from a few strong categories with pixel-level annotations to unseen weak categories with only image-level annotations. Ganaye [55] proposed to take advantage of the invariant nature (*i.e.*, structural invariance of the segmentation map) of anatomical structures to form a semantic constraint for semi-supervised segmentation. In all these cases, the unlabeled data information is entirely involved in the model learning. Meanwhile, certain parts of the segmented maps are not segmented well enough to be used to incrementally refine the segmentation network. Thus, the current semi-supervised neural network models need more investigation.

1.1.2.6 Focal Loss for Deep Segmentation Networks

The above-mentioned deep segmentation networks cannot properly handle hard-to-segment samples (or regions). One reason is that the training of the network is dominated by easy-to-segment samples [1]. This easy-to-segment sample dominance phenomenon often occurs in medical image segmentation tasks due to the irregular distribution of some medical images which may be caused by the different lesion abnormalities or imaging factors, such as devices from different vendors or different imaging protocols. Several works have been proposed in the literature to address the aforementioned challenges [172, 122, 1]. To achieve better performance on hard-to-segment (or detect) samples, Shrivastava *et al.* [172] proposed a simple strategy by automatically selecting hard samples for further training to tune the networks. To prevent the vast number of easy samples from overwhelming the network during training, Lin *et al.* [122] proposed focal loss for dense object detection and achieved promising results. In another work, Zhou *et al.* [1] introduced focal loss for biomedical image segmentation. However, focal loss has some shortcomings when applied to medical image segmentation due to its **use** of predicted probability on the samples as the hard-or-easy evaluator which could neglect the structural information and also suffers from multi-category competition issues in some cases.

1.1.3 Low-Contrast Isointense Infant Brain MRI Segmentation

The increasing availability of non-invasive infant brain MR images affords unprecedented opportunities for precise charting of dynamic early brain developmental trajectories in understanding normative and aberrant brain growth [116]. For example, the recently-awarded Baby Connectome Project (BCP)[1], will acquire and release cross-sectional and longitudinal multimodal MRI data from 500 typically-developing children from birth to 5 years of age. This will greatly increase our limited knowledge on normal early brain development, and will also provide important insights into the origins and aberrant growth trajectories of neuro-developmental disorders, such as autism and schizophrenia. For instance, autistic children are reported to experience brain overgrowth associated with an increase in cortical surface area before 2 years of age [67]. As current treatments for many neuro-developmental disorders are ameliorative rather than curative, identifying early neuromarkers of risk for these disorders will allow designing targeted preemptive intervention strategies to improve prognosis or even prevent the disorders. To measure early brain development and identify biomarkers, accurate segmentation of MRI into different regions of interest (ROIs), *e.g.*, white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF), is the most critical step. It will allow for volumetric quantification and also more sophisticated quantification of the structures of gray and white matters, such as cortical

[1] http://babyconnectomeproject.org/

thickness, surface area, and gyrification, which may provide important indications of very early neuro-anatomical developmental events [129, 56].

The first year of life is the most dynamic phase of postnatal human brain development. This is mainly because brain tissues grow rapidly, while cognitive and motor functions undergo a wide range of development [102]. Accurate tissue segmentation of infant brain MR images in this phase is of great importance in studying normal and abnormal early brain development [58, 114, 115]. It is recognized that the segmentation of infant brain MRI is considerably more difficult than the segmentation of adult brain MRI, due to reduced tissue contrast [195], increased noise, severe partial volume effect [204], and ongoing WM myelination [195, 62]. Fig. 1.1 shows examples of T1-weighted MRI, T2-weighted MRI, and fractional anisotropy (FA) images acquired at around 6 months of age. It can be observed that WM and GM exhibit almost the same intensity levels (especially in cortical regions), resulting in the lowest tissue contrast and hence significant difficulty for tissue segmentation.

Recently, deep learning based methods have achieved great success in image segmentation, including infant brain tissue segmentation. Zhang *et al.* [219] first proposed using a deep CNN to segment isointense-phase brain images, in which a hierarchy of increasingly complex features from MR images were learned. They used patch-level learning by sliding windows in the 2D space of the images. Their methods took the center voxel tissue label as the label for the whole patch during learning. Consequently, their method was somehow sensitive to the patch size, especially for the voxels on the boundaries of WM or GM. In fact, such methods (based on the sliding windows) have to tradeoff between localization and classification accuracy, as utilizing large patches will lead to a loss in localization accuracy due to more pooling layers (we have to use more pooling layers to cover the context of the input patch), while using small patches will yield perception of much less context information and thus will become sensitive to noise. Moreover, these methods contain a large number of parameters due to the existence of fully connected layers, which overburdens the convergence of the network. Moeskops *et al.* [141] further proposed a multi-scale CNN for infantile brain tissue segmentation. Nie *et al.* [153] proposed a multi-pathway FCN to segment 2D slices of infant brain tissue. This model is very memory-costly due to the use of a multi-pathway architecture, and thus not suitable for 3D MRI brain tissue segmentation. Chen *et al.* [30] introduced a residual learning technique to help the FCN training for adult brain tissue segmentation. However, all the above-mentioned methods have overlooked the fact that CNN or FCN will lose information due to adoption of pooling operations, which will affect localization accuracy.

To overcome the above-mentioned challenges, I propose to employ and further extend the UNet [164] for the segmentation of infant brain images. Although UNet is able to remedy the loss of spatial details of an FCN, it cannot solve the problem of low-contrast infant brain MR images since the features provided by the shallower layers are fuzzy and thus cannot help precise localization. As a result, I propose a multi-modal UNet to overcome such a challenge by exploiting the complementary information from multiple modalities in Chapter 3.1. Also, I have designed a transformation block composed of convolutional layers to learn semantic high-resolution features and a fusion block to better combine encoder and decoder features to evolve the multi-modal UNet.

1.1.4 Blurry Boundary Delineation for Pelvic Images

One of the major challenges for medical image segmentation is the blurry nature of medical images (*e.g.*, CT, MR, PET and microscopic images) in some cases, which can often result in low-contrast and even vanishing boundaries, for example, pelvic organ boundaries as shown in Fig. 1.2.

Many encoder-decoder networks have been proposed for semantic segmentation [124, 164, 214] and achieved very promising performance on various tasks. UNet [164], a typical encoder-decoder architecture which combines shallow and deep features with a skip connection, is widely used in many image segmentation tasks. Some works have been proposed to enhance the UNet [166, 155]. However, Heller *et al.* [76] found that deep segmentation models are robust on the non-boundary regions, but not very robust to boundaries. Actually, these models usually fail to properly segment the **low-contrast** boundaries, especially for the case with extremely low tissue contrast. For example, prostate boundaries in MR or CT pelvic images are often low-contrast and even blurry. To solve this challenge, I argue high resolution with rich semantic based feature learning is desired.

Besides the variants of UNet, to better delineate the boundaries, Ravishankar *et al.* [162] proposed a multi-task network to segment the organs by jointly regressing the boundaries and foreground. Zhu *et al.* [229] proposed a boundary-weighted domain adaptive neural network to accurately extract the boundaries of the prostate MRI. However, all these methods do not consider the fact that voxels around low-contrast boundaries are highly similar. Thus, it is better not to classify the voxels to be on the boundary or not.

In Chapter 3 of this dissertation, I propose a concept that learning high-resolution semantic features can potentially solve this **blurry boundary delineation** problem. Accordingly, I propose two methods to learn such semantic meaningful and detail-reserving features. In particular, I propose to use a series of dilated residual blocks to form a high resolution pathway to enhance the raw skip connection (*i.e.*, the skip connection without any additional operations) in the first method. I further propose a novel semantic-guided encoder feature learning mechanism to improve the skip connection in previous encoder-decoder architectures, so that it can work better for low-contrast medical image segmentation at a low cost. The design of the proposed network is mainly based on the idea of explicitly utilizing high-resolution semantic information to compensate for the deficiency on inaccurate boundary delineation of the existing encoder-decoder networks. Specifically, I propose to concatenate the low-layer (encoder) feature maps and the high-layer (decoder) feature maps, and then design a channel-wise attention and spatial-wise attention to help learn (which can also be viewed as a kind of feature selection) the high-resolution semantic encoder feature maps. With these better learned encoder feature maps, I further concatenate (or element-wisely add) it to the corresponding decoder layers in the encoder-decoder framework. Moreover, I propose using soft label (*i.e.*, probabilities) to indicate the probability of a voxel being on the boundary. Accordingly, a soft cross-entropy loss is proposed as a metric for the low-contrast boundary delineation problem.

1.2 Medical Image Synthesis

1.2.1 Motivation and Challenges

Medical imaging is crucial for the diagnosis and treatment of different diseases. Usually more than one imaging modality is required for imaging based clinical decision making because different modalities often provide different and complementary insights. Computer tomography (CT), for example, has the advantage of providing electron density and physical density of the tissues, which is indispensable for dosage planning in radiotherapy treatment of cancer patients. However, CT suffers from the disadvantage of lacking good contrast in soft tissues. The radiation exposure during acquisition may also increase the risk of secondary cancer especially for young patients [174]. Magnetic resonance imaging (MRI), on the other hand, gives very good contrast of soft tissues. Compared to CT, MRI is also much safer and does not involve any radiation; but it is much more costly than CT and does not have the electron density information that is needed for radiation therapy planning or PET image reconstruction [99].

In a second example, the acquired images cannot well depict rich details of anatomical structures and abnormality. For instance, it is difficult to delineate small brain structures such as the hippocampus in 3T MR images because of the limited signal-to-noise ratio [15, 11, 10]. 7T MRI, on the contrary, provides much better image quality than 3T MRI by revealing certain texture information within the hippocampus. This allows better imaging of the anatomy and thus contributes to better utilization of the imaging data. Yet 7T MRI is much more expensive and not widely accessible.

In a third example, a combination of sequences can provide complimentary information to support the radiologists to better understand the soft tissue and perform certain diagnoses. For example, the uniform signal on T2-weighted (T2) MRI for tumor patients can be a reliable indicator for a benign lesion, while it is not a good indicator for malignant tumors due to the inhomogeneous signals [27]. On the other hand, T1-with-contrast-enhanced (T1c) MRI can be used as an indicator for a malignant tumor and to assess growth/shrinkage because T1c provides clear demarcation of enhancing region around the tumor. However, in clinical, sequences which are routinely acquired may be unusable or missing due to various factors, such as limited available scan time, scan corruption, artifacts, wrong machine settings, allergies to certain contrast agents and so on [169].

The above observations reflect a general dilemma, where a certain modality is desired but infeasible to acquire in practice. To this end, a system being able to synthesize images-of-interest from different sources, *e.g.*, image modalities and acquisition protocols, can be of great benefit. It may provide the highly demanded imaging data for certain clinical usage, without incurring in additional cost/risk of performing a real acquisition.

However, medical image synthesis is very challenging to solve directly since the mapping from the source image to the target image (or its inverse) is usually of high dimensionality and ill-posed [52, 61, 73]. As shown in Fig. 1.3(a) and 1.3(b), CT and MRI data of the same subject have quite different appearances. And thus the mapping from MRI to CT has to be highly nonlinear in order to bridge the significant appearance gap between the two modalities, which requires a lot of effort to model. Shown in Fig. 1.3(c), the 7T MRI has much higher resolution and much clearer contrast compared to the 3T MRI, which makes the mapping from 3T MRI to 7T MRI very challenging. In addition, certain regions in the images (such as tumor in Fig. 1.3(d)) may have completely different image contrast and appearance.

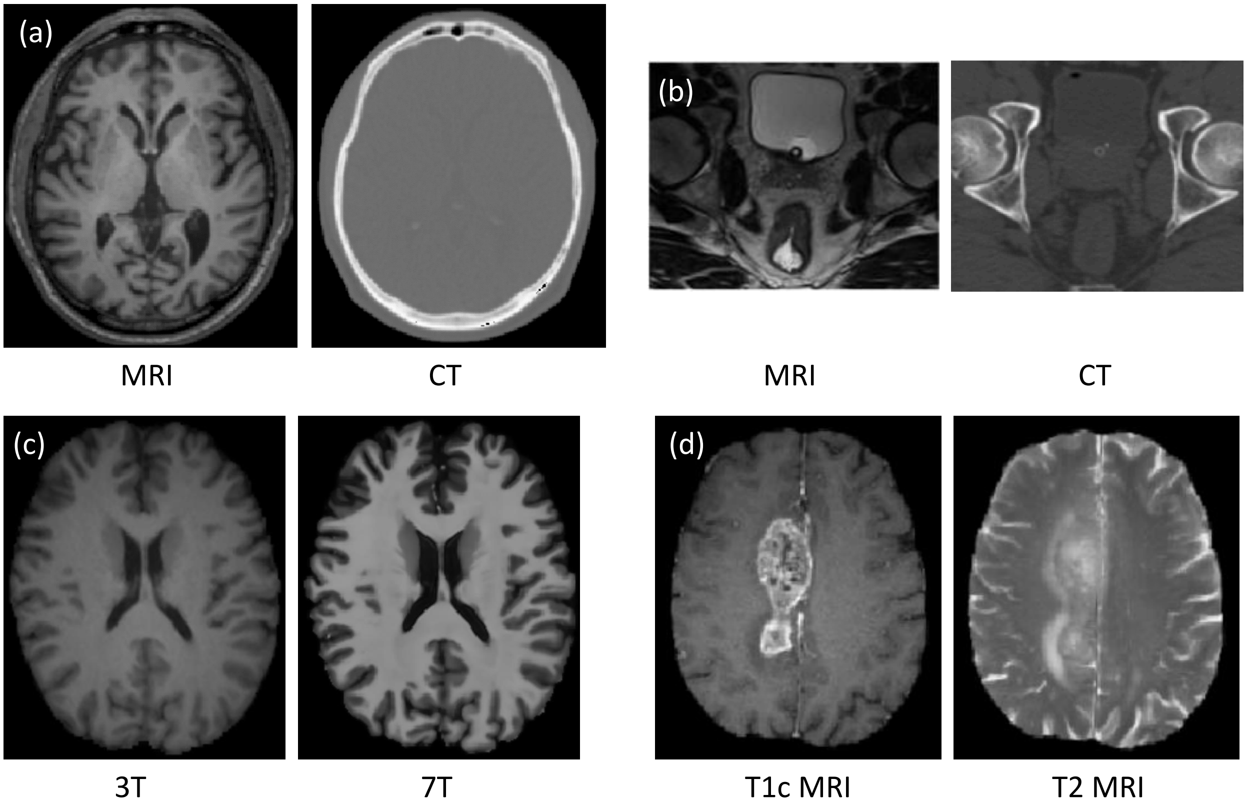


Figure 1.3: Four pairs of corresponding source (left) and target (right) images from the same subjects. (a) shows a pair of MRI/CT brain images; (b) shows a pair of MRI/CT pelvic images; (c) shows a pair of 3T/7T brain MRI; (d) shows a pair of T1c/T2 brain tumor MRI. For all of them, the source and target modalities have quite different appearances.

1.2.2 Previous Work

Recently, many researches have focused on estimating one modality image from another and proposed many methods to address this challenge [16, 216, 25, 78, 223]. Berker *et al.* [16], for example, proposed to treat the MRI-to-CT problem as a segmentation task by segmenting MRI images into different tissue classes and then assigning each class with a known attenuation property. This method highly depends on the segmentation accuracy and always needs manual work for the accuracy of the results. On the other hand, atlas-based methods have also been used in the literature. In [25], the authors proposed to register an atlas of MRI to the new subject's source image and then warp the corresponding target image of the atlas as the estimated target image.

In [210], the authors proposed an extension of the well known Label Propagation (LP) segmentation algorithm. They called it Modality Propagation which propagates intensity value from one modality to another modality and provided a generalization of LP allowing to work with continuous data instead of only categorical segmentation labels. Similarly, in [21], the authors proposed an information propagation scheme, here for a given source patch, the system looks for similar patches in the source dataset, and constructs the target image based on their corresponding target (which is known in the training set). However, these methods are quite time-consuming.

Besides, learning-based methods have also been explored to model a nonlinear mapping from source image to target image, to alleviate some of the previous drawbacks [131, 93, 41, 86, 2, 224, 198, 217, 53, 170]. For instance, Jog *et al.* [93] learned a nonlinear regression with random forest to carry out cross-modality synthesis of high resolution images from low resolution scans. Huynh *et al.* [86] presented an approach to synthesize CT from MRI using random forest as well. Unsupervised methods have also been used. In [188] for example, the authors proposed a framework where for each voxel in the source image, a set of target candidate values was generated by a nearest neighbor search in the training set of target images. Note that since there is no paired data, they need a similarity measure that is somewhat robust to changes in modality. In this case they use mutual information. Then, they select the best candidates by maximizing a global energy function that takes into account the mutual information between the source and target, and also the spatial consistency in the generated target. These methods often have to first represent the source image by features and then map them to generate the target image. Thus, the performances of these methods are bounded to the manually engineered features as well as the quality of the representation of the source image based on the extracted features.

Nowadays, deep learning has become very popular in computer vision and medical image analysis, achieving state-of-the-art results in both fields without the need of handcrafted features [105, 109, 71, 120]. In the particular case of image synthesis, Dong *et al.* [47]

proposed to use Convolutional Neural Networks (CNNs) for single image super-resolution. Kim *et al.* [98] further improved the super-resolution algorithm by proposing a recursive CNN which can boost performance without increasing parametric complexity. Li *et al.* [117] applied a similar deep learning model to estimate the missing PET image from the MRI data of the same subject. Huang *et al.* [83] proposed to simultaneously conduct super-resolution and cross-modality medical image synthesis by the weakly-supervised joint convolutional sparse coding.

One potential problem of CNN is that it tends to neglect neighborhood information in the predicted target image, especially when the input size is small. To overcome this, FCNs, which can preserve structural information, have been utilized for image synthesis [48, 83]. Typically, the L_2 distance between the predicted target image and the ground truth is used as the loss function to train the CNNs and FCNs, which tends to yield blurry target images especially in multi-modal distributions [133]. Minimizing the L_2 loss is equivalent to maximizing the peak signal-to-noise rate (PSNR); however, as it has been pointed out in [111], a higher PSNR does not necessarily provide a perceptually better result.

1.3 Adversarial Learning for Medical Image analysis

Generative Adversarial Network (GAN) [60] is currently a very popular and successful unsupervised model that can generate samples following an implicit distribution. The GAN framework consists of two competing networks: a generator and a discriminator, both of which are involved in an adversarial two-player game, in which the generator aims to learn the data distribution while the discriminator estimates the probability of a sample coming from the training data or the generator. Adversarial learning, derived from GAN [60], has been widely applied to the supervised models (such as segmentation and generation models) with purpose of enhancing models' capacity and achieved great success in image generation and segmentation [60, 133, 103, 148, 205, 221, 230]. Many works have demonstrated that adversarial learning can contribute to generate much more perceptually realistic images or videos [60, 133, 148, 88], in which the generation can even fool human. It is also shown that adversarial learning can help improve the segmentation performance, for instance, fixing the obvious segmentation errors [142, 103, 152]. However, the performance gain brought by adversarial learning is usually inconsistent (or limited) across different metrics, for instance, the generated images are becoming much more realistic, while the performance in terms of quantitative metric cannot have an obvious improvement and may even become worse [88, 94, 152]. Moreover, it is quite challenging to train such a GAN framework due to the difficulty of balancing the generator and discriminator (*i.e.*, since discriminator has an easier job compared to the generator, it may face problem of vanishing gradient for the generator) [60, 5, 63, 132]. Though various methods have been proposed to solve this problem [5, 63, 132], this issue has been alleviated but still not solved [137, 127]. Besides, mode collapse phenomenon occurs quite often in practice when training GAN systems [138].

To address such issues, in Chapter 5, I conduct an analysis for the roles of discriminators in the classic GANs and make a comparison with those in supervised adversarial systems. Based on the analysis, I propose adversarial confidence learning to upgrade the adversarial learning in the supervised adversarial systems, *i.e.*, besides the adversarial gradient as in the classic GANs, I also rely on the confidence information which depicts how well the images are segmented or synthesized provided by the discriminator to improve the supervised generator. In particular, I propose a difficulty-aware attention mechanism based on confidence learning for medical image segmentation (synthesis). Specifically, apart from the segmentation network, I propose a fully convolutional adversarial network to work as confidence network to learn how well the local regions are segmented or synthesized (i.e., the confidence map generated by the confidence network can provide us the trustworthy and untrustworthy regions in the segmented (synthesized) label map from the segmentation (synthesis) network). Based on the confidence map, two machine learning algorithms are proposed for medical image analysis aiming at solving two challenges:

a) I propose a difficulty-aware attention mechanism to adaptively assign region-level and voxel-level importance to improve the design of the supervised segmentation (synthesis) network. Since a difficulty-aware mechanism is adopted to further enhance the segmentation network, the **easy-sample dominance issue** can be alleviated accordingly. In the proposed framework, the visual perception performance gain is mainly coming from the adversarial learning, and the quantitative performance improvement is mainly from the improved design of the supervised generator. As a consequence, the proposed system is less sensitive to training imbalance between generator and discriminator. b) I also propose a confidence-aware semi-supervised segmentation algorithm which could adaptively select the well-segmented regions to expand the label set. Since my method includes the well segmented regions instead of entire segmented map in a dynamic way, my proposed method is a better solution to alleviate the **lack of labeled data** problem. Besides, I also investigate the loss functions for the confidence network, *i.e.*, to guarantee a powerful discriminator, I do a survey on various objective functions for the adversarial learning and further propose using binary cross entropy loss as in the original classic GAN, instead of using the widely adopted Wasserstein distance [5]. To this end, the proposed adversarial confidence learning framework can take better advantage of adversarial learning but avoid or alleviate the drawbacks of the adversarial learning. My proposed algorithm has been applied to several medical image segmentation and synthesis tasks, such as pelvic organ segmentation, which is critical for guiding both biopsy and cancer radiation therapy, and brain tumor image cross-modality synthesis, which can help diagnose the brain lesions. Experimental results indicate that my proposed algorithm can improve not only the visual perception performance but also the quantitative segmentation (synthesis) accuracy, compared to other state-of-the-art methods. In addition, the semi-supervised approach is validated to be effective in taking advantage of unlabeled data for better medical image modeling.

1.4 Thesis

Thesis: *1. Current encoder-decoder networks cannot accurately delineate boundaries with low contrast due to the fuzzy information in the encoder layers. Learning high-resolution semantic features is a potential solution. Semantic-guided encoder feature learning can endow these architectures with high-resolution semantic features and thus can well handle the blurry boundary delineation problems. 2. Adversarial learning for supervised models work as realistic regularization. In this learning schema, the quantitative performance gain does not well align with the visual perception improvement. Adversarial confidence learning could achieve a synchronous performance gain by utilizing the confidence information to enhance the design of the supervised model while retaining the realistic effect. More importantly, adversarial confidence learning can provide a platform for building difficulty-aware attention mechanism and confidence-aware semi-supervised algorithm to address easy sample dominance issue and lack of labeled data, respectively, for deep learning based medical image segmentation and synthesis.*

The first research topic of this dissertation is about low-contrast boundary delineation. It analyzes the limitation of the widely used encoder-decoder architecture based networks for blurry boundary delineation problems, and then proposes the idea that high-resolution semantic features could address this problem. Accordingly, this dissertation proposes three algorithms to efficiently learn detail-aware semantic features so that blurry boundary delineation could be well addressed.

The second research topic of this dissertation is about how to align the quantitative performance gain with visual perception improvement with adversarial learning for medical image analysis, how to solve easy-sample dominance issue and how to alleviate lack of annotated label problem. By investigating the roles of discriminators in classic GANs and comparing them with those in supervised adversarial learning systems, this dissertation figures out that adversarial learning works as realistic regularization for supervised models. It further proposes adversarial confidence learning framework to simultaneously achieve performance improvement for quantitative and qualitative metrics. This framework utilizes the dense confidence information to improve the supervised generator with balancing the adversarial learning for the generator. It can be suited to difficulty-aware mechanism for alleviating the easy sample dominance issue for medical image analysis. It can also be adjusted to solve the lack of labeled data problem by adapting a semi-supervised learning algorithm.

1. **Specific Aim (Blurry boundary delineation).** The currently widely used encoder-decoder architectures (UNet is a typical example) can improve the segmentation accuracy by a large margin compared to conventional methods. While the encoder-decoder architecture cannot well handle the blurry boundary delineation problem because the information provided by the encoder layers is fuzzy itself and thus cannot accurately localize the boundaries. For multi-modal data, designing a multi-modal network by taking the complementary information from multiple modalities into consideration is a good choice to address the low-contrast issue. For single-modal data, learning high resolution and rich semantic encoder features is a reasonable solution to address this problem. Accordingly, high-resolution pathway and semantic-guided encoder feature learning could both lead to learn high-resolution semantic features in the lateral connection and thus well fight against the blurry boundary delineation.

2. **Specific Aim (Adversarial confidence learning).** Adversarial learning can be utilized to enhance the training of the neural networks for medical image analysis. Direct application of adversarial learning could lead to visual perception improvement, while the quantitative performance cannot have a synchronous growth and even become worse with adversarial learning. The adversarial learning works a realistic regularization for networks by enforcing the generated image to follow the distribution of real data in a entire image manner. However, it breaks the sample-to-sample correspondence which is the basis of the conventional loss for objective functions, which could thus limit the quantitative performance. This problem could

be alleviated by exploiting the confidence information from the discriminator to design a more powerful supervised generator.

(a) **Specific Aim (Difficulty-aware attention mechanism for medical image analysis).** In medical image analysis, easy-to-segment (easy-to-synthesize) sample dominance phenomenon often occurs due to the irregular distribution of some medical images which may be caused by the different abnormal degree of the lesion or the imaging factors, such as different vendor devices or imaging protocols. To address this problem, conventional methods usually apply more weights to the difficult regions. However, it is hard to dynamically recognize the difficult-regions and determine the weights. By adopting the discriminator to encode the concatenation of probability map of segmentation and original input, the designed fully convolutional confidence network can obtain structured difficult regions and learn reasonable confidence information in a dynamic manner.

(b) **Specific Aim (Confidence-aware semi-supervised learning for medical image segmentation).** Deep learning based medical image segmentation requires large scale of annotated data. However, labeled data in medical image segmentation is usually challenging to obtain. Semi-supervised learning is a common choice to take advantage of unlabeled medical data. However, it is hard to determine which is a good segmented sample for unlabeled data and conventional semi-supervised segmentation models usually include the entire samples which have been certified to be well segmented. It largely limits the power of semi-supervised models due to the poorly segmented regions in the entire samples could provide wrong training signals. The adversarial confidence learning could provide a real-time well-or-poorly-segmented determination mechanism for unlabeled data. More importantly, since this mechanism can select the high-confidence segmented regions, the problem by using the entire images as training signals could be elegantly solved.

In order to support the thesis and the above two main specific aims, the detailed contributions of this dissertation include:

1. A multi-modal evolutionary UNet is proposed to address the problems of low-contrast multi-modal medical image segmentation. Multi-modal information is utilized to address the low-constrast issue for medical image segmentation and fusion strategy is further explored in the environment of multi-modal neural networks. Besides, a transformation module composed of convolutional operations is proposed to alleviate the channel information bias between encoder and decoder features; a fusion module is proposed to better fuse information from encoder and decoder layers; **(Aim 1)**

2. High-resolution encoder-decoder networks are proposed to better delineate the blurry boundaries for medical images, in which, dilated residual module is specifically designed to replace the skip connection in encoder-decoder networks; **(Aim 1)**

3. Semantic-guided encoder feature learning is proposed to efficiently learn high-resolution semantic features to endow encoder-decoder networks the capacity of blurry boundary delineation; **(Aim 1)**

4. Contour-sensitive loss functions, including regression and soft classification, are explored for better modeling the boundaries to more clearly recognize the boundaries; **(Aim 1)**

5. Extensive experiments on several large medical segmentation datasets (infant brain segmentation in MRI, public challenge and self-owned pelvic datasets in both MRI and CT) indicate that the proposed method can accurately delineate the blurry boundaries with outperforming many existing methods in these tasks; **(Aim 1)**

6. Deep learning based method is proposed for cross-modality medical image synthesis. Adversarial learning is utilized in the synthesis model, targeting at generating more

realistic images. Auto-context model is also explored for alleviating the long-range information dependency problem. **(Aim 2)**

7. Analysis is conducted to explore the roles of the discriminator in classic GANs and comparison is done with those roles in supervised adversarial learning systems. Together with further experiments, adversarial learning is certified to work as realistic regularization in supervised adversarial learning systems; **(Aim 2)**

8. Adversarial learning can be utilized as realistic regularization for supervised models. The problem of adversarial learning is proposed: visual perception improvement does not well align with quantitative performance gain with standard adversarial learning; **(Aim 2)**

9. Adversarial confidence learning framework is proposed to address the inconsistency of performance gain in different metrics. By adopting a fully convolutional adversarial network, dense confidence information can be utilized to better design the supervised generator networks to improve the quantitative performance, in the meantime, the realistic regularization with adversarial learning is retained; **(Aim 2)**

10. Following adversarial confidence learning, difficulty-aware attention mechanism is proposed for medical image segmentation and synthesis, especially hard-to-segment samples and lesion medical image synthesis; **(Aim 2)**

11. Confidence-aware semi-supervised segmentation networks are proposed to adaptively recognize the well segmented samples and regions. It targets at including the well-segmented regions instead of the entire sample to dynamically increase the labeled set; **(Aim 2)**

12. Extensive experiments that are conducted on several datasets indicate the effectiveness of my proposed approaches, especially the adversarial confidence learning framework. **(Aim 2)**

1.5 Overview of Chapters

The remaining chapters of this dissertation are organized as follows.

1. Chapter 2 presents the background of related techniques and evaluation metrics used in the dissertation. The techniques include convolutional neural networks (CNN), generative adversarial networks (GAN) and attention models. Under CNN, the basic knowledge and mathematical notations are presented, followed by the application of CNN to classification, regression, image segmentatino and image synthesis problems. Under GAN, the basic theory about classic GANs and adversarial learning in supervised models is introduced. The application of adversarial learning to image segmentation and synthesis is also presented. Under attention models, I introduce the basic theory about these models, and then elaborate the details about spatial attention mechanism, channel-wise attention mechanism and mixed attention mechanism. Finally, several evaluation metrics are introduced. They are used to evaluate the proposed segmentation and synthesis methods and compare them with other existing methods.

2. Chapter 3 presents three proposed deep networks based methods to address the challenging low-contrast medical image segmentation and blurry boundary delineation problems. For multi-modal data, I propose to take advantage of multi-modal information with a multi-modal enhanced UNet for low-contrast medical image segmentation due to the complementary information provided by the multi-modal images. For single-modal data, I argue conventional encoder-decoder networks (*e.g.*, UNet) cannot well delineate the blurry boundaries from medical images. After analysis of the encoder-decoder architectures, I propose learning high-resolution rich-semantic encoder features could well solve this problem. The first proposed method is to mitigate the information gap between the encoder and decoder layers by learning high-resolution semantic features in the lateral connection. The main idea is to use several

dilated residual modules to form a high-resolution pathway so that I can quickly enlarge the receptive field without pooling and make the training easier. The second proposed algorithm is to develop a semantic-guided encoder feature learning for blurry boundary delineation which elegantly learns the precise boundary information with attention mechanism. Extensive experimental results are presented to evaluate each design of the proposed methods and to show the superior performance of my proposed methods over several existing methods.

3. Chapter 4 introduces deep residual adversarial networks for medical image synthesis. This chapter describes the deep learning methods for medical image synthesis, and elaborates the usage of adversarial learning in deep synthesis networks. In addition, auto-context model is adopted to alleviate the long-range information dependency issue. A large number of experiments are conducted to investigate the advantages and disadvantages of adversarial learning for medical image synthesis and to validate the effectiveness of auto-context refinement.

4. Chapter 5 presents the adversarial confidence learning framework and two following applications to solve easy-sample-dominance issue and lack of labeled medical data problem. This chapter first analyzes the roles discriminator in classic GANs and compares with those in supervised adversarial learning systems. Then I introduce the adversarial confidence learning framework with an example of segmentation task and also have a discussion of selection of loss functions for adversarial learning. Moreover, I elaborate the difficulty-aware attention mechanism based on adversarial confidence learning for medical image segmentation and synthesis, especially for those hard-to-segment or hard-to-synthesis regions. Confidence-aware semi-supervised deep segmentation models are followed up to solve the lack of labeled medical data challenge. Extensive experiments are conducted to validate the effectiveness of my proposed framework and the superiority of my proposed methods.

5. Chapter 6 concludes the dissertation and discusses the limitations of the proposed methods and frameworks as well as future work. In the conclusion, the methods proposed in this dissertation for blurry boundary delineation and adversarial confidence learning for medical image analysis are briefly summarized. Their limitations are also discussed. Furthermore, interesting future directions and potential strategies to improve the proposed methods are discussed.

CHAPTER 2: BACKGROUND

2.1 Convolutional Neural Networks

Deep learning models can learn a hierarchy of features, *i.e.*, high-level features building on low-level ones. The CNNs [110, 105, 18] are a type of deep models, in which trainable filters and local neighborhood pooling operations are applied in an alternating sequence starting with the raw input images. This results in a hierarchy of increasingly complex features. One property of CNNs is that they can capture highly nonlinear mappings between inputs and outputs [110]. When trained with appropriate regularization [175], CNNs can achieve superior performance on visual object recognition and image classification tasks [110, 105, 71, 82]. CNNs have also been used in other applications. In [89, 90, 182, 77], CNNs were applied to restore and segment the volumetric electron microscopy images; Ciresan *et al.* [38] applied deep CNNs to detect mitosis in breast histology images by using pixel classifiers based on patches. CNNs [47, 117, 48] were also employed to reconstruct images of desired modality. Yang *et al.* [208] utilized CNN to predict the parameters in a registration model to achieve more efficient medical image registration.

2.1.1 The Basics of Convolutional Neural Networks

CNN is a class of deep, feed-forward (not recurrent) artificial neural networks. A modern CNN is composed of several essential building blocks. We will elaborate them one by one.

To ease the description, we first give the notations used throughout the chapter. Let x be an image and let k be the kernel function. (i, j) indexes the location in the image. f is an activation function which will specified in the following subsection.

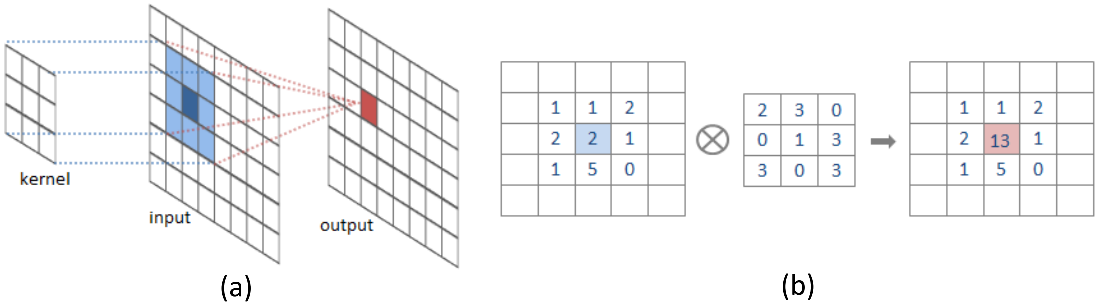


Figure 2.1: Illustration of convolution. (a) shows convolution operation in neural networks. (b) presents an example of how convolution works.

2.1.1.1 Convolution

Convolution is the core building block in CNN which could well capture the spatial relations across pixels in images. Then the discrete 2D convolution is:

$$(x * k)\,(i, j) = \sum_{-\infty}^{\infty} \sum_{-\infty}^{\infty} x_{i-u, j-v} k_{u,v}, \tag{2.1}$$

At a convolution layer, the previous layer's feature maps are convolved with learnable kernels and put through the activation function to form the output feature maps, as shown in Fig. 2.1 (figures are from this website[1]). Each output map may combine convolutions with multiple input maps, which can be formally expressed as Eq. 2.2.

$$x_j^l = f\left(\sum_{i \in M_j} x_i^{l-1} * k_{ij}^l + b_j^l\right), \tag{2.2}$$

where M_j represents a selection of input maps. Each output feature map is given an additive bias b, however for a particular output map, the input maps will be convolved with distinct kernels. In other words, if output feature map j and h both sum over input map i, then the kernels applied to feature map i are different for output feature maps j and h.

Many works have been done to improve the convolution operation. Locally connected convolution, which is different from plain convolution in the fact that every location in the

[1] http://intellabs.github.io/RiverTrail/tutorial/

feature map learns a different set of filters, is proposed for considering the location information and used to process images where the key points are relatively fixed across images, such as face recognition and prostate segmentation [196, 152]. Deformable convolution [42] is a generic operation which can model geometric transformations without additional supervision and can thus contribute to forming networks of more capacity. Depth-wise convolution, a kind of separable convolution, takes full advantage of group convolution [105] to save computation consumption. Combined with point-wise convolution which is actually convolution operation with 1×1 kernel, efficient lightweight networks [79] can be built to work on mobile devices.

2.1.1.2 Pooling

Pooling operation is designed to downsample the feature maps, that is, if there are N input feature maps, then there will be exactly N corresponding output feature maps with downsampled size. Pooling layers are periodically inserted in-between successive convolution layers in CNNs, with the purpose of progressively reducing the spatial size of the representation so that we can reduce the amount of parameters and computation cost in the network, and hence to also control overfitting to some extent. Formally,

$$x_j^l = f\left(\beta_j^l down\left(x_i^{l-1}\right) + b_j^l\right), \tag{2.3}$$

where $down\left(\cdot\right)$ represents a sub-sampling function. The widely used ones are max pooling and average pooling.

2.1.1.3 Activation function

Activation function is one of the most important building blocks in neural networks, which is put at the end of or in between neural networks to help decide if the neuron would fire or not as shown in Fig. 2.2(a). The activation function is non-linear transforma-

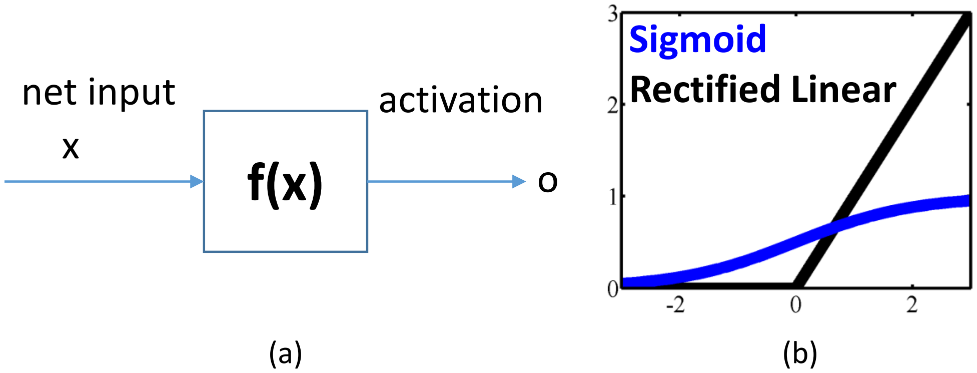


Figure 2.2: (a) shows activation function in neural networks and (b) displays typical activation functions.

tion that we do over the input signal. There are several widely used activation functions, such as sigmoid (Eq. 2.4), Rectified Linear Unit (denoted as ReLU) [144] (Eq. 2.5), tanh (Eq. 2.6) and so on.

$$sigmoid(x) = 1/\left(1 + e^{-x}\right), \tag{2.4}$$

$$\mathrm{ReLU}(x) = \max\left(0, x\right), \tag{2.5}$$

$$\tanh\left(x\right) = \frac{2}{1 + e^{-2x}} - 1. \tag{2.6}$$

Compared to sigmoid which is the previously frequently used activation function, ReLU is recently more frequently adopted as activation function in CNNs (shown in Fig. 2.2(b)), because ReLU can alleviate the gradient vanishing or gradient explosion problems which often occur using sigmoid [202] (Vanishing (exploding) gradients is a well known problem in deep neural networks. As the gradient information is back-propagated, repeated multiplication or convolution with small (big) weights lead to ineffectively small (big) gradients in shallow layers). Some activation functions are also proposed to further improve the ReLU, such as LeakyReLU [130] (Eq. 2.7), Parametric Rectified Linear Unit (denoted as PReLU) [70] (Eq. 2.8, note θ is a learned vector which has the same size with

x), exponential linear unit(denoted as elu) [39] (Eq. 2.9) and so on.

$$\mathrm{LeakyReLU}\left(x\right)=\begin{cases}\alpha x & ,x<0\\ x & ,x\geq 0\end{cases}, \tag{2.7}$$

$$\mathrm{PReLU}\left(x\right)=\begin{cases}\theta x & ,x<0\\ x & ,x\geq 0\end{cases}, \tag{2.8}$$

$$\mathrm{elu}\left(x\right)=\begin{cases}\alpha\left(e^{x}-1\right) & ,x<0\\ x & ,x\geq 0\end{cases}. \tag{2.9}$$

2.1.1.4 Normalization

To further address the gradient vanishing or exploding issues, many normalization techniques are proposed to help better optimize the training of networks. Ioffe *et al.* [87] proposed batch normalization to reduce internal covariate shift which refers to change in the input distribution to internal layers of a deep network and thus accelerate the training of neural networks. Followed batch normalization, instance normalization [183] and layer normalization [19] are proposed for better training networks in certain conditions. Later, group normalization [199] is proposed for cases when batch size is small. Currently, it is a standard way to include one normalization technique to make the networks easier to train.

2.1.1.5 Architectures

With years of development, numerous network architectures have been proposed for different applications, in which, AlexNet [105], VGGNet [173], Inception [177], ResNet [71] and DenseNet [82] are the milestone architectures which inspire the computer vision field to move forward. I briefly cover two of them: Alexnet and ResNet.

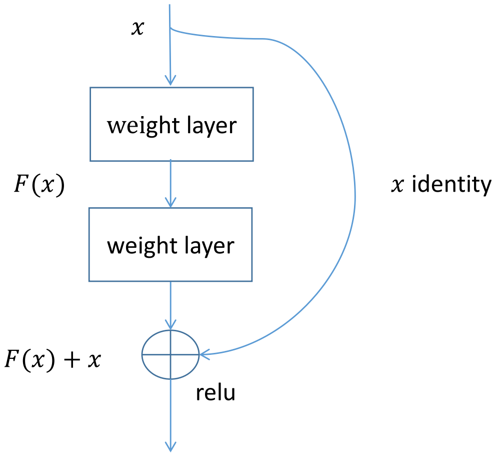


Figure 2.3: Illustration of a typical residual block.

AlexNet is the most representative CNN architecture in the early stage, which consists of 5 convolution layers, 3 fully connected layers, 3 pooling layers. Dropout [175] is adopted to enhance model's generalization ability. To save GPU memory, the convolutions are partitioned into two groups. AlexNet achieve the state-of-the-art performance on image recognition challenges at that time.

ResNet [71] is another milestone for deep learning. With residual learning, the performances of many computer vision and medical image analysis tasks have been largely improved [71, 68, 152]. In ResNet, the authors design a residual block and use series of residual blocks to form the residual networks. Fig. 2.3 shows a typical residual block with two intermedia layers. The essence of residual block is identity mapping, which can be mathematically formulated as follows:

$$\mathbf{Z} = F\left(\mathbf{A}, \{\theta_i\}\right) + \mathbf{A}, \tag{2.10}$$

where $\{\theta_i\}$ is the set of convolutional filters in the bottleneck residual unit, F is the convolutional layers in a residual block, and $\mathbf{A}$ and $\mathbf{Z}$ are the input and output feature maps, respectively.

2.1.1.6 Regularization

There are several techniques developed to regularize the networks to prevent overfitting. L_2 regularization, also called weight decay is widely adopted regularization term for neural networks. Besides, Dropout [175], a very sophisticated design by randomly dropping part of neurons at each training iteration, is a strong regularization for training neural networks which can efficiently solve the overfitting problem if well used. Early stopping is another choice for improving generalization ability of network models.

2.1.1.7 Loss function

Cross entropy loss is widely adopted to optimize the proposed CNN model for classification problems. The mathematical formulation is given by Eq. 2.11.

$$H\left(y,p\right) = -\sum_{i=1}^{k} y_i log\left(p_i\right), \tag{2.11}$$

where y_i and p_i are the ground truth and the CNN score for category i in $\{1, 2, ..., k\}$, with k denoting the number of distinct classes.

2.1.2 Deep Networks for Image Segmentation

Recent developments in deep learning have largely boosted the state of the art of segmentation methods [124, 164]. Fully convolutional network (FCN) [124], a variant of CNN by replacing fully connected layers with convolutional or deconvolutional layers, is a recent popular choice for semantic image segmentation in both computer vision and medical image fields [124, 164, 214, 158, 209, 200]. FCN trains neural networks in an end-to-end fashion by directly optimizing intermediate feature layers, allowing it to outperform traditional methods that often regard feature learning and segmentation as two separate tasks. UNet [164], an evolutionary variant of FCN, has achieved excellent performance for medical image segmentation, by effectively combining high-level and low-level features in the

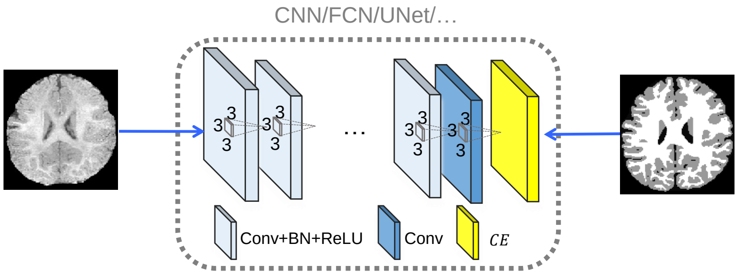


Figure 2.4: A typical example for medical image segmentation.

network architecture. Compared to FCN, UNet can improve the localization accuracy, especially near organ boundaries. Lin *et al.* [121] introduced a generic multi-path refinement network with carefully designed encoder/decoder modules to increase the capacity of U-shape network. Chen *et al.* [35] proposed using atrous separable convolution to enhance the encoder-decoder networks for semantic segmentation. Chen *et al.* [31] proposed to use contour-aware knowledge to help accurately segment gland images. Zhu *et al.* [229] introduced a boundary-weighted domain adaptive network to accurately delineate the boundaries of MRI prostate. Apart from the architecture exploration, some works are also proposed to enhance the UNet [166, 155] with the idea of applying attention mechanism for better feature learning. A typical medical image segmentation neural network is shown in Fig. 2.4.

2.1.3 Deep Networks for Image Synthesis

It is often quite challenging to directly synthesize high-quality demanded medical modality images. Convolutional neural network (CNN) provides a new way for learning highly non-linear relationships because of employing multiple-layer mapping [65, 83, 149, 226, 197, 40]. Dong *et al.* [47] proposed to use Convolutional Neural Networks (CNNs) for single image super-resolution. Kim *et al.* [98] further improved the super-resolution algorithm by proposing a recursive CNN which can boost performance without increasing parametric complexity. Li *et al.* [117] applied a similar deep learning model to estimate the missing PET image from the MRI data of the same subject. Nie *et al.* [145] proposed

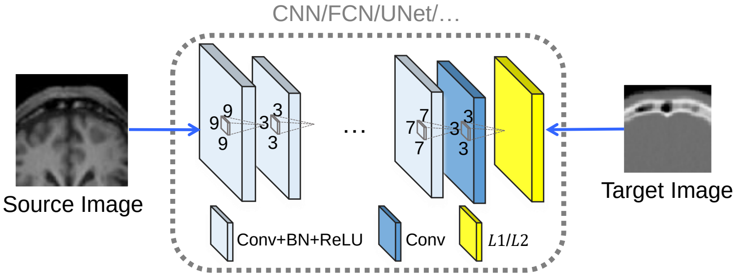


Figure 2.5: A typical example of deep networks for medical image synthesis.

using CNN for cross-modality medical image synthesis. Huang *et al.* [83] proposed to simultaneously conduct super-resolution and cross-modality medical image synthesis by the weakly-supervised joint convolutional sparse coding. Han *et al.* [65] proposed to employ UNet to synthesize CT from MRI for the brain tumor. A typical deep learning based medical image synthesis framework is shown in Fig. 2.5.

2.2 Generative Adversarial Networks

2.2.1 Classic GAN

GANs [60] are efficient unsupervised models that can generate samples following an implicit distribution given a set of data. GANs work by training two different networks: a generator network G, and a discriminator network D. G is typically a FCN which generates images $G(z)$ from a random noise vector z, and D is a CNN which estimates the probability that an input image x is drawn from the distribution of real images; that is, it can classify an input image as real or synthetic. Both networks are trained simultaneously with D trying to correctly discriminate between real and synthetic data, while G is trying to produce realistic images that will confuse D. More formally, for D, we would like to find its parameters:

$$\max_{D} \ \log(D(x)) + \log(1 - D(G(z))), \tag{2.12}$$

For G, we would like to optimize

$$\max_{G} \ \log(D(G(z))). \tag{2.13}$$

To improve the training of GANs, Radford *et al*. [161] explored unsupervised learning with CNN and introduced a class of CNNs called as deep convolutional GANs (DC-GANs) with certain architectural constraints to be strong candidates for unsupervised learning. To solve the gradient vanishing issues in GANs, Arjovsky *et al*. [5] proposed Wasserstein distance as a metric to measure how close the generated distribution and the real distribution are. Many other works are also proposed to solve or alleviate this problem [127, 132, 160]. Metz [138] proposed the unrolled GAN to stabilize the training of GAN and mitigate the mode collapse phenomenon.

2.2.2 Adversarial Learning in Supervised Systems

Adversarial learning has been also widely extended to supervised models, such as image segmentation and synthesis. Isola *et al*. [88] used conditional GAN on image-to-image translation problems, in which, a L_1 or L_2 loss is also applied to train the generator and this supervised loss can guarantee the correspondence between input modality and output modality, while the adversarial loss contributes to generate realistic style images. Nie *et al*. [148] proposed to use adversarial learning together with a L_1 or L_2 loss and gradient different loss to generate realistic-like CT images from MRI. However, the performance in terms of quantitative metrics become even worse when using adversarial learning. Luc *et al*. [125] proposed to use adversarial learning to help segmentation tasks, and the authors argue that adversarial learning works as a regularization to enforce higher-order spatial consistency among different classes because adversarial learning can assess the joint configuration of many label variables, which is beyond the capacity of a typical segmentation objective function - cross-entropy loss. Adversarial learning is also used to improve

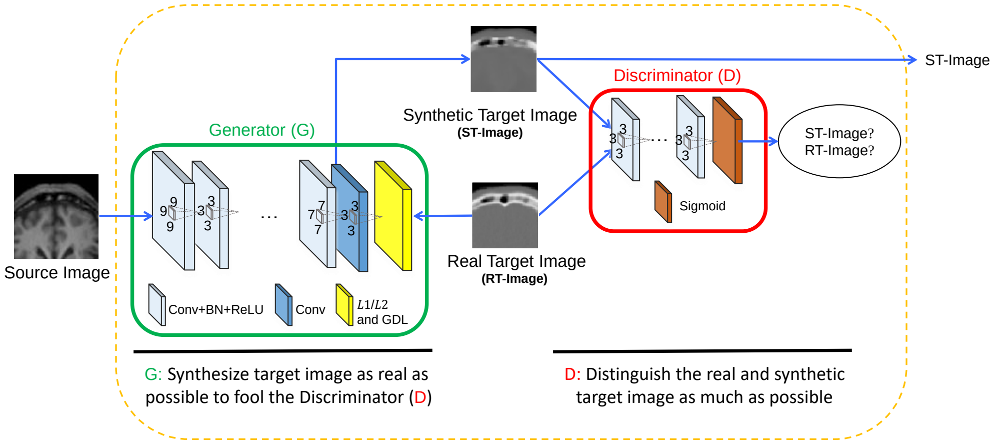


Figure 2.6: A typical example of GAN for medical image synthesis.

the medical image segmentation [205, 152]; however, the performance gain by adversarial learning is actually limited to fix the obvious segmentation errors. A typical medical image synthesis task with adversarial learning is shown in Fig. 2.6.

2.3 Attention Models

“Attention” is defined as the “active direction of the mind to an object”[2]. The essence of the attention mechanism is to draw inspiration from the human visual attention mechanism. Generally, when we visually perceive objects, we do not usually see a scene from the beginning to the end, but often focus on a specific part of the demand. When we find that a scene often appears in a certain part of what we want to observe, we will learn to put attention on that part when similar scenes appear in the future. This is thought to be the essence of the attention mechanism.

In machine learning field, attention mechanism is first introduced in [140] for image classification with visual attention on recurrent models. Current attention mechanisms are actually originated from natural language processing models which develops machine translation models with 1D data [9, 186].

[2] https://skymind.ai/wiki/attention-mechanism-memory-network

Given a set of queries, *i.e.*, n query vectors, which are usually features in different positions (spatial domain) or different channels (channel domain) (Taking spatial domain query vectors as an example: suppose there are d feature maps in a certain convolutional layer, then the features to represent a certain query position can be a d-dim vector. If there are n query positions, then there are n query vectors), $q_1, q_2, \ldots, q_n \in R^d$, a set of keys, *i.e.*, m key vectors, which are usually features in the current positions (spatial domain) or channels (channel domain) (The definition is similar with query vectors and the difference is we only consider certain key positions or channels), $k_1, k_2, \ldots, k_m \in R^p$ ($p = d$ if it is attention within the same layer which is usually the case for computer vision but not for natural language processing), and m value vectors $v_1, v_2, \ldots, v_m \in R^p$, the attention mechanism is to compute a bank of output vectors $o_1, o_2, \ldots, o_n \in R^q$ by combining the transformed value vectors $g\left(v_i\right) \in R^q$ in a linear manner. It is worth noting that the coefficients of the linear combination are usually approximated by the relations between the corresponding query vector and each key vector. This process can be formally expressed as Eq. 2.14,

$$o_j = \frac{1}{C} \sum_{i=1}^{m} f\left(q_j, k_i\right) g\left(v_i\right), \tag{2.14}$$

where $f\left(q_j, k_i\right)$ characterizes the relation (*e.g.*, similarity) between q_j and k_i, $g\left(\cdot\right)$ is usually a linear transformation $g\left(v_i\right) = W_v v_i \in R^q$, where $W_v \in R^{q \times p}$, and $C = \sum_{i=1}^{m} f\left(q_j, k_i\right)$ is a normalization factor. A commonly suggested similarity function is the embedded Gaussian [193], usually defined as Eq. 2.15.

$$f\left(q_j, k_i\right) = \exp\left(\varphi(q_j)^T \phi\left(k_i\right)\right), \tag{2.15}$$

where $\varphi\left(\cdot\right)$ and $\phi\left(\cdot\right)$ are linear transformations, which are defined as $\varphi\left(q_j\right) = W_q q_j$ and $\phi\left(k_i\right) = W_k k_i$, respectively.

Attention mechanism has been extended to deal with 2D images and 3D video process models [231, 193]. When dealing with 2D data, the inputs to the attention operator can be represented as 3D tensors $Q \in R^{h \times w \times c}$, $K \in R^{h \times w \times c}$ and $V \in R^{h \times w \times c}$, where h, w, and c represent the height, width, and number of channels, respectively (Note, we have assumed the three tensors having the same shape for simplicity). The output vectors are also a 3D tensor $O \in R^{h \times w \times q}$. We further elaborate them as below in the perspective of different domains.

2.3.1 Spatial Domain Attention Mechanism

Wang *et al.* [193] proposed non-local block to capture long-range dependencies; in other words, it is designed to aggregate information from other locations to enhance the features at the current location. The non-local operation computes the response at a position as a weighted sum of the features at all positions, as shown in Fig. 2.7(a). If the input feature maps are $x = \{x_i\}_{i=1}^{N_p}$, N_p is the number of pixels in this feature map, and z is the output feature map. Then we can formally define the non-local operation as Eq. 2.16.

$$z_i = x_i + W_z \sum_{j=1}^{N_p} \frac{f\left(x_i, x_j\right)}{C\left(x\right)} (W_v \cdot x_j), \tag{2.16}$$

where $f\left(x_i, x_j\right)$ computes the relation between pixels at position i and j, which is usually instantiated by embedded Gaussian as defined in Eq. 2.15 and $C\left(x\right)$ is the normalization factor. W_z and W_v denote linear transform matrices (*e.g.*, 1×1 convolution).

The similar idea has been applied to image segmentation tasks. For example, Yuan *et al.* [215] proposed to address the scene parsing problem by first aggregating context and then conducting the segmentation step which could be more robust compared to the original pixel-wise segmentation. Inspired by the self-attention mechanism, this study proposed computing a similarity map for each pixel p, where each similarity score indicates the degree that each corresponding pixel and the pixel p belongs to the same category. Such similarity map is called as object context map, which serves as a surrogate of the true object context.

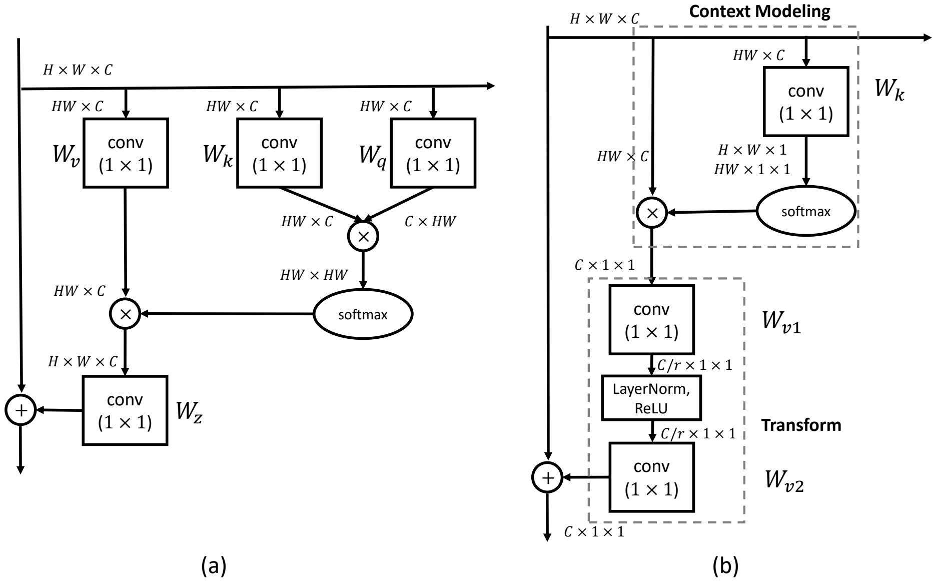


Figure 2.7: Illustration of two attention blocks. (a) shows a non-local block. (b) shows a global context block.

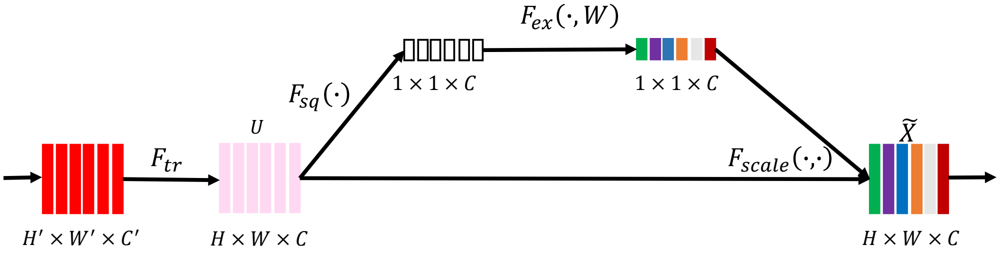


Figure 2.8: A Squeeze-and-Excitation block.

2.3.2 Channel Domain Attention Mechanism

In each convolutional layer of CNN, we have a number of channels. The characteristics of each channel actually represent the components of the image on different convolution kernels. If we compute a weight to each channel's signal to represent the channel's relevance to the key information, the greater the weight, the higher relevance to the key information will be. As a result, this weight can represent the attention we should pay to the corresponding channel.

Squeeze-and-Excitation Network (SENet) [81] is typical well designed channel domain attention model, including squeeze and excitation parts. A flowchart to depicts the SE block is shown in Fig. 2.8. The squeeze part is actually formulated as a global average pooling operation shown in Eq. 2.17, which squeezes each channel to a single numeric value.

$$z_c = F_{sq}\left(u_c\right) = \frac{1}{W \times H} \sum\nolimits_{i=1}^{W} \sum\nolimits_{j=1}^{H} u_c\left(i, j\right), \tag{2.17}$$

where u_c is feature map for channel c.

The excitation is implemented by two fully connected (FC) layers. The first FC layer compresses C channels into C/r channels to reduce the amount of computation (followed by ReLU), the second FC layer is restoring back to C channels (followed by Sigmoid), and r is compression proportion. Formally,

$$s = F_{ex}\left(z, W\right) = \sigma\left(g\left(z, W\right)\right) = \sigma\left(W_2 \delta\left(W_1 z\right)\right), \tag{2.18}$$

where z is the vector of squeezed single numeric values for all the channels, W_1 and W_2 are the weights for the first and second FC layers, respectively.

To this end, the weighting coefficient for each feature map of the convolutional block is learned and we can finally weight these feature maps so that we can emphasize or suppress certain channels.

Another typical work about channel attention mechanism is originated from DANet [54]. In DANet, the channel attention mechanism mainly targets at modeling interdependencies between channels since the mechanism wants to improve the feature representation of specific semantics (Each feature map of a high layer can be regarded as a class-specific response, and different semantic responses are associated with each other). Fig. 2.9 shows the structure of the channel attention module. The attention map $w \in R^{C \times C}$ is calculated from the original feature maps $x \in R^{C \times H \times W}$. In particular, x is shaped to $R^{C \times N}$, and then a matrix multiplication is performed between x and x^T. Finally, we apply a softmax

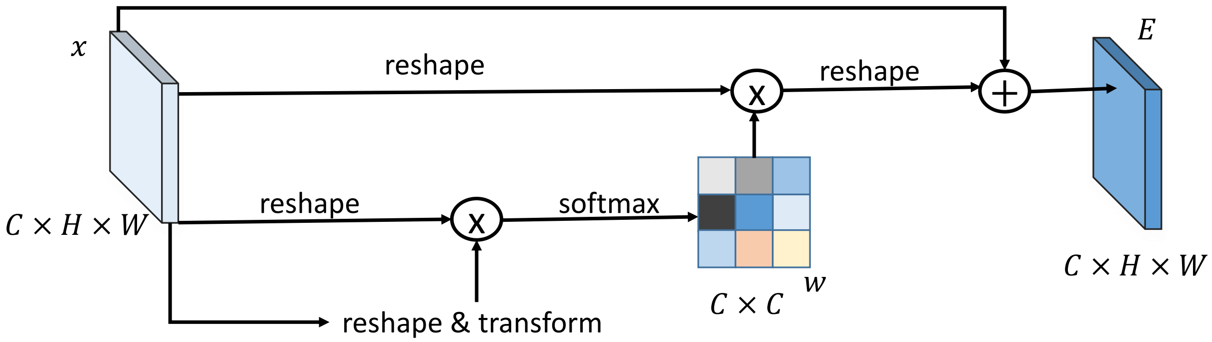


Figure 2.9: A channel attention block from DANet.

operation to obtain the channel attention map on w. Formally,

$$w_{ji} = \frac{\mathrm{e}^{(x_i \cdot x_j)}}{\sum_{k=1}^{C} \mathrm{e}^{(x_i \cdot x_k)}}, \tag{2.19}$$

where x_j and x_i represent the j^{th} and i^{th} feature maps, respectively, and w_{ji} measures the j^{th} channel's impact on the i^{th} channel.

In addition, a matrix multiplication is performed between the transpose of w and x, and the result is then reshaped to $R^{C \times H \times W}$. Then a scale parameter (β) is further applied on the result. Together with an element-wise sum operation with x, the final output is obtained by Eq. 2.20.

$$E_i = \beta \sum_{i=1}^{C} (x_j w_{ji}) + x_j. \tag{2.20}$$

The channel-wise attention mechanisms are widely used in computer vision and medical image analysis algorithms [232, 220, 225].

2.3.3 Mixed Domain Attention Mechanism

The spatial domain attention mechanism does well in global context modeling but has a huge computation cost, while the channel domain attention mechanism just needs small computation cost but it cannot take full advantage of global context information. To take advantage of both mechanisms and alleviate the shortcomings, Cao *et al*. [24] analyzed both non-local networks and squeeze-excitation networks, and proposed global context networks (GCNet) to go beyond the drawbacks of both networks. In GCNet, global context block is proposed for effective modeling on long-range dependency with lightweight computation. As shown in Fig. 2.7(b), global attention pooling is used for global context modeling, bottleneck transform is used to capture channel-wise dependency and broadcast element-wise addition is used for feature fusion.

Similar ideas have been used for natural image segmentation and achieved great success [54].

Different from the idea of context modeling, Roy *et al.* [166] successfully extended the SE module to spatial domain. Their proposed concurrent spatial and channel SENet has achieved better performance for medical image segmentation, in which the squeeze excitation idea is adjusted to spatial domain to capture the spatial details so that we can use the learned important local regions in the feature maps to enhance the training of networks. At the same time, the channel domain attention mechanism is retained and thus could model the channel-wise dependency information.

2.4 Easy-Sample Dominance Issue

The easy-to-segment (or easy-to-synthesize) sample dominance phenomenon often occurs in deep learning based medical image analysis tasks due to the irregular distribution of medical images, which is usually attributed to the different abnormal degree of the lesion or the imaging factors, such as different imaging protocols or vendor devices. Several works have been proposed in the literature to address this problem [172, 122, 1]. To achieve better performance on hard-to-segment (or hard-to-detect) samples, Shrivastava *et al.* [172] proposed a simple strategy to automatically select hard samples for further tuning the networks. Kumar *et al.* [106] made use of hard example mining technique to develop an incremental learning framework that can adapt to new medical data while retaining existing knowledge.

To prevent the vast number of easy samples from overwhelming the networks during training, Lin *et al.* [122] designed focal loss for detection and achieved promising results. In particular, the authors proposed to add a modulating factor to the cross entropy loss, with tunable focusing parameter. More formally,

$$FL\left(p_t\right) = -(1 - p_t)^{\gamma} \log\left(p_t\right), \tag{2.21}$$

where t indexes the category, and γ is the scaling factor.

The scaling factor of $(1 - p_t)^{\gamma}$ largely suppresses the contribution of easy-to-classify samples to the training loss (*e.g.*, when $p_t = 0.9$, the scaling factor is 0.01 supposing γ is 2). It can also lightly suppress the contribution of hard-to-segment samples (*e.g.*, when $p_t = 0.1$, the scaling factor is 0.81). Therefore, the sample attention mechanism can adaptively shift the training focus to hard-to-classify samples and address the issue of dominance by easy samples.

In another work [1], the authors introduced to directly apply focal loss for the biomedical image segmentation. However, the focal loss has some shortcomings when applied to medical image segmentation due to its usage of predicted probability on the samples as the hard-or-easy evaluator which could neglect the structural information and may also suffer from multi-category competition issues.

2.5 Evaluation Metrics

2.5.1 Evaluation for Medical Image Segmentation

In medical image segmentation experiments, Dice similarity score (DSC), modified Hausdorff distance (MHD) [49] and Average Surface Distance (ASD) [211] are three widely used performance metrics to quantitative evaluate the performance.

Formally, DSC is defined by,

$$DSC = 2\frac{|A \cap B|}{|A| + |B|}, \tag{2.22}$$

where A and B denote the binary segmentation labels generated manually and computationally in terms of image segmentation, respectively, $|A|$ denotes the number of positive elements in the binary segmentation A, and $|A \cap B|$ is the number of shared positive elements by A and B.

Supposing that C and D are the two sets of positive pixels identified manually and computationally, respectively, for one tissue class of a subject, the MHD can then be defined as Eq. 2.23.

$$MHD\left(C, D\right) = \max\left(d\left(C, D\right), d\left(D, C\right)\right), \tag{2.23}$$

where $d\left(C, D\right) = \frac{1}{N_c}\sum_{c \in C} d\left(c, D\right)$, and the distance between a point c and a set of points D is defined as $d\left(c, D\right) = \min \sum_{d \in D} \|c - d\|$.

Mathematically, ASD is defined by Eq. 2.24.

$$ASD = \frac{1}{2}\left(\frac{\sum_{V_i \in S_A} \min_{V_j \in S_B} d\left(V_i, V_j\right)}{|S_A|} + \frac{\sum_{V_j \in S_B} \min_{V_i \in S_A} d\left(V_j, V_i\right)}{|S_B|}\right), \tag{2.24}$$

where $\mathbf{S_A}$ is the surface of the manual segmentation map, $\mathbf{S_B}$ is the surface of the automatic segmentation map, and $d\left(\mathbf{V}_j, \mathbf{V}_i\right)$ indicates the Euclidean distance from vertex $\mathbf{V}_j$ to the vertex $\mathbf{V}_i$.

2.5.2 Evaluation for Medical Image Synthesis

For medical image synthesis tasks, mean absolute error (MAE), peak signal to noise ratio (PSNR) and structural similarity (SSIM) index are the widely used evaluation metrics to quantitatively measure the performance. Suppose I is the ground-truth image with size of $m \times n$, and U is the synthesized image.

$$MAE = \frac{1}{mn} \sum\nolimits_{i=0}^{m-1} \sum\nolimits_{j=0}^{n-1} |I\left(i,j\right) - U\left(i,j\right)|, \tag{2.25}$$

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right), \tag{2.26}$$

where MAX_I is the maximum possible pixel value of the image, and MSE is mean squared error:

$$MSE = \frac{1}{mn} \sum\nolimits_{i=0}^{m-1} \sum\nolimits_{j=0}^{n-1} \left[I\left(i,j\right) - U\left(i,j\right)\right]^2. \tag{2.27}$$

$$SSIM\left(x,y\right) = \frac{\left(2\mu_x u_y + C_1\right)\left(2\sigma_{xy} + C_2\right)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)}, \tag{2.28}$$

where μ_x and μ_y are the mean of image x and y, respectively. σ_x and σ_y are the variance of x and y, respectively. σ_{xy} is covariance between x and y. C_1 and C_2 are two variables to stabilize the division with weak denominator with $C_1 = \left(k_1 L\right)^2$ and $C_2 = \left(k_2 L\right)^2$ where L the dynamic range of the pixel-values typically this is $2^n - 1$ where n is the number of bits per pixel), $k_1 = 0.01$ and $k_2 = 0.03$. Note, it will return a local SSIM map by Eq. 2.28 and a further average step should be applied to obtain a global SSIM (scalar).

CHAPTER 3: DEEP NEURAL NETWORKS FOR BLURRY BOUNDARY DELINEATION

As introduced in Sec. 1.1.1, low contrast is a major challenge for medical image segmentation. To overcome this challenge, this chapter first describes a deep learning based method for multi-modal low-contrast medical image segmentation:

1. 3D-TFmUNet[1]: I propose a multi-modal UNet to overcome the low tissue contrast challenge in isointense infant brain MRI segmentation by fully exploiting complementary information in the multiple modalities, which is introduced in Sec. 1.1.3. Since, the UNet exist a huge information gap between the shallow and deep layers which may have a negative impact for the information fusion, I propose using a transformation block and a fusion layer to remedy the information gap. Also, I have explored how patch size could impact the segmentation performance. Moreover, initialization, pooling and upsampling strategies (*i.e.*, the basic network components) are also empirically discussed in this study.

As described in Sec. 1.1.4, blurry boundary delineation is also very challenging. In this chapter, I further propose two novel algorithms to improve the encoder-decoder networks so that I can better deal with single-modal low-contrast medical images, especially those with blurry boundaries.

1. High-resolution encoder-decoder networks[2]: To overcome the challenges of low contrast medical image segmentation (for instance, pelvic organ segmentation as introduced in Sec. 1.1.1), a high-resolution multi-scale encoder-decoder network was

[1] This work was published in IEEE Transactions on Cybernetics [151]. This chapter uses parts of text descriptions and figures from the published paper.

[2] This work was published in IEEE Transactions on Imaging Processing [228]. This chapter uses parts of text descriptions and figures from the published paper.

proposed. In particular, a high-resolution branch, a specially-designed pathway with several dilate residual blocks, was put in the lateral connection for high-resolution semantic information exploitation. In addition, contour regression was formulated as an additional task to help accurately localize the boundaries.

2. Semantic-guided encoder feature learning for encoder-decoder networks[3]: To efficiently delineate blurry boundaries from low-contrast medical images (for example, prostate boundary delineation introduced in Sec. 1.1.4), I proposed semantic-guided encoder feature learning strategy to endow the encoder-decoder networks with capacity for blurry boundary delineation. Specifically, empirical study was carried out to analyze why the encoder-decoder networks cannot well model blurry boundary delineation problems. Then I argue learning high-resolution semantic information could be a potential solution. Hence, I proposed using the semantic information from decoder layers as guidance to boost the feature learning of the encoder features. In particular, I designed a channel-wise and spatial-wise attention mechanism for the concatenated features (from both encoder and decoder layers) to efficiently learn high-resolution semantic features. Finally, these meaningful features were provided to the decoder layers to help more precisely delineate the fuzzy boundaries.

Sec. 3.1 introduces the proposed 3D-TFmUNet together with ablation study and experimental results for multi-modal isointense infant brain MRI segmentation which mainly fight against the low tissue contrast of gray matter and white matter in the images. Sec. 3.2 presents the high-resolution encoder-decoder networks for pelvic organ segmentation. I have also vastly discussed when and how I should use the proposed approach. Sec. 3.3 introduces a carefully designed semantic-guided encoder feature learning strategy, which is actually a lightweight pathway in the lateral connection, for blurry boundary delineation. Experimental results and ablation study are also reported in this section.

[3] This work was under review in a conference [147]. This chapter uses parts of text descriptions and figures from the manuscript.

3.1 Multi-Modal Neural Networks for Low-Contrast Medical Image Segmentation

As mentioned in Sec. 1.1.3, multi-modal images could provide complementary information to overcome the low tissue contrast in single modal images. Since UNet cannot well address the segmentation for low-contrast medical images, I propose to exploit multi-modal information to address the low-contrast issue in medical image segmentation by developing a multi-modal UNet. In addition, I propose a transformation module as well as a fusion module to alleviate the potential information bias between shallow layers and deep layers.

In the following, I first introduce 3D-mUNet architecture (as shown in Fig. 3.1) by extending the conventional UNet [124] architecture to 3D case with multi-modal input in Sec. 3.1.1, and then present the further extended 3D-mTFUNet model for considering information balance between lower and deeper layers in Sec. 3.1.2. Later, I give details of training in Sec. 3.1.3 and Sec. 3.1.4. Experimental results and ablation study are reported in Sec. 3.1.5.

3.1.1 The Designed Basic 3D-mUNet Architecture

One of the most challenging steps in adopting deep learning framework is the design of network architecture. The conventional UNet [164] utilizes pooling operations 4 times which aims to highly abstracting the input information. However, the localization information could be seriously lost. I try to alleviate this problem by designing a basic framework to make a tradeoff between the resolution and abstraction of input information. Inspired by Simonyan *et al.*'s work [173] and Vijay *et al.*'s work [8], I design the 3D-architecture for MR images with groups of convolutional layers and de-convolutional layers, as shown in Fig. 3.1. Note, I only employ pooling operation 3 times, and use the smallest convolution filters $3 \times 3 \times 3$ which are believed to be able to capture the details better [173]. Reducing pooling operations will definitely mitigate the loss of resolution and I can have more

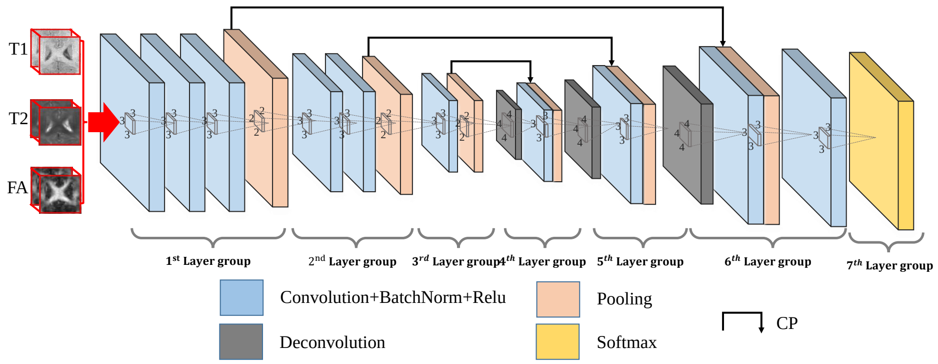


Figure 3.1: Illustration of 3D-mUNet architecture.

layers with the small convolution filters which will retain the abstraction of information. This network applies softmax loss to the top layer of the networks. As demonstrated by Wang *et al.* [192], the complementary information from multiple imaging modalities is beneficial to dealing with insufficient tissue contrast. I thus feed three modality images as inputs to the neural network to learn complementary information from each other.

In my 3D-mUNet, the 1st group of layers consists of three convolutional layers (each containing 96 filters), followed by a pooling layer. These feature maps are fed into the 2nd group of layers consisting of two convolution layers (each with 128 filters), followed by a pooling layer. In the 3rd group of layers, one convolutional layer with 128 filters is applied, followed by a pooling layer. Note that all convolution filters are with a size of $3 \times 3 \times 3$, and rectified linear unit (ReLU) [144] is employed as an activation function after each convolution layer. I use one voxel as stride and add one voxel as pad for all convolution layers.

Then, the output feature maps from the 3rd group of layers are up-sampled through a deconvolution layer to the 4th layer group with 32 filters. The 5th and 6th layer groups are both deconvolution layers with 32 filters, following the 4th layer group. The deconvolution filters are all of size of $4 \times 4 \times 4$. In addition, the last deconvolution layer owns 4 filters, which correspond to 4 categories such as CSF, GM, WM, and background. It is worth noting that, similar to many previous works [124, 173, 8], convolution layers are added after each deconvolution operation. Finally, a layer consisting of the softmax units is applied at the end of the network, aiming at predicting one label (out of 4 possible labels) for each voxel.

The network minimizes the cross entropy loss between the predicted labels and ground-truth labels. Suppose k is the number of categories, m is the samples in a batch, x and y corresponds to predicted probability and ground truth label respectively. Formally, the objective function can be described by Eq. 3.1.

$$J\left(\theta\right) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} I\left\{y^{(i)}, j\right\} \log \frac{e^{\theta_j^T x^{(i)}}}{\sum\limits_{l=1}^{k} e^{\theta_l^T x^{(i)}}}\right], \tag{3.1}$$

where

$$I\left\{y^{(i)}, j\right\} = \begin{cases} 1, y^{(i)} = j \\ 0, y^{(i)} \neq j \end{cases}. \tag{3.2}$$

3.1.2 3D-TFmUNet

The combination (concatenation) operation in UNet cannot deal with the feature maps with different numbers well (*i.e.*, the original deeper layers usually have much less feature maps than the shallower layers), and thus the signals from the deeper layers can be ignored during the fusion operation. To address it, I propose a transformation module, for example, using additional convolutional layers with 1 × 1 × 1 kernels, to adjust the number of feature maps from lower layers to be comparable to the number of the corresponding higher layers. Moreover, these additional convolutional layers can also work as *transformation modules* to boost the low-level features to be complementary for the high-level features. Note that this transformation block will not change the size of the feature maps, except adjusting just their numbers. In addition, I provide *fusion modules* (extra 1 × 1 × 1 or 3 × 3 × 3 convolutional layers are utilized to work as the fusion modules in this study and we can also apply attention blocks to work as the fusion modules in future work.) after

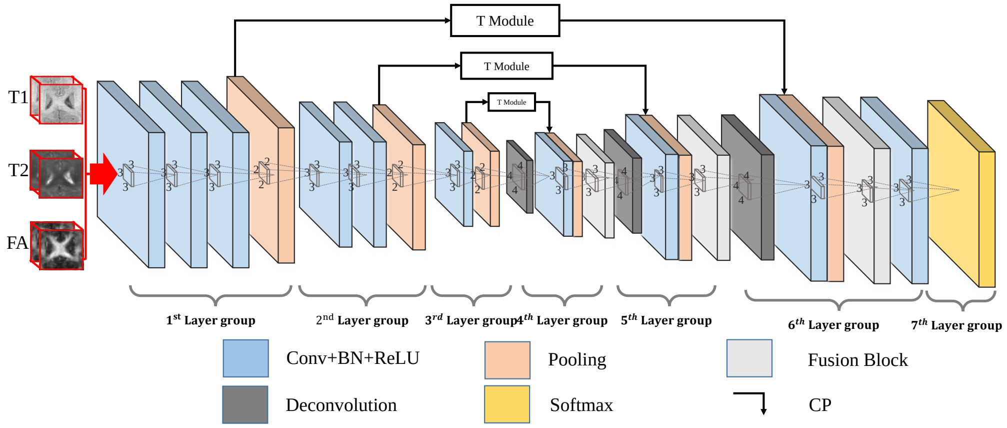


Figure 3.2: Illustration of proposed architecture for multi-modal infant brain MRI segmentation. Here, CP denotes a copy and concatenate subprocedure.

the concatenation layers to enhance the fusion of the low-level and high-level features. In this way, the resolution information from shallow-layer feature maps is smoothly passed through the network and used in the up-sampling phases to help achieve more accurate localization. As a consequence, the convolution layers during the up-sampling phase can generate more precise outputs based on the assembled feature maps. The respective architecture is shown in Fig. 3.2. I denote this model as 3D-TFmUNet for short. Furthermore, a batch normalization [87] operation is adopted after each convolution operation to make the network easier to converge.

3.1.3 Weighting the Loss

In many real applications, like the infant brain segmentation, the numbers of data samples from different categories are often quite different, and thus the distributions of the data among different classes are not balanced. This may cause over-fitting to a specific category (usually the dominant category), which is the so-called imbalanced data problem. To avoid this phenomenon, a class balance strategy is proposed. Specifically, a weighting scheme is adopted for the loss to address the class imbalance problem during training. The weighted cross entropy loss can also be formed as Eq. (3.1) but with Eq. (3.3) as the weight parameter (C_j is the loss weight for a specific category j, and it can be given by inversely proportional to the fraction of samples in each corresponding category.).

$$I\left\{y^{(i)}, j\right\} = \begin{cases} C_j, y^{(i)} = j \\ 0, y^{(i)} \neq j \end{cases} . \tag{3.3}$$

3.1.4 Training the Proposed Networks

As it is the case with almost every medical imaging application, the dataset used to train a model is often limited, while almost all deep models require a huge number of samples for training [59, 105]. On the other hand, it is better for 3D-TFmUNet to operate on full images in order to have large receptive field and cover broad context information. To remedy these challenges and also to train a reliable deep model, a tradeoff between receptive field and the dataset size is made. Specifically, the overlapping patches of size 32 × 32 × 32 are extracted for both original images and manually-segmented images. To augment the number of training patches, the patches are slid throughout the entire image with a certain step size. In this way, a sufficient number of training patches could be generated.

With the deep architecture shown in Fig. 3.2, the total number of parameters is $2,534,276$. To train the designed network, the convolution kernels are initialized with Xavier algorithm [59], which can automatically determine the scale of initialization based on the numbers of input and output neurons. The network biases are initialized to 0. Then, a coarse linear search is conducted to determine the initial learning rate and also the weight decay parameters. The learning rate is initially set to 5×10^{-3}, and then decreased by 10 times for a fixed step size during training. The proposed model is trained by back propagation [108]. Specially, Caffe [92], a commonly-used deep learning framework, is adopted to implement the proposed method with minor modification.

3.1.5 Experimental Results

In this study, I use data containing multi-modality infant brain MR images to run the experiments and compare my proposed method with the baseline and state-of-the-art methods. I first introduce the dataset and the respective preprocessing steps, followed by description of evaluation metrics.

3.1.5.1 Multi-modality Infant Brain Dataset

For this dataset, T1, T2, and diffusion-weighted MR images of 11 healthy infants were acquired using a Siemens 3T head-only MR scanner. Both T2 image and the fractional anisotropy (FA) image (derived from distortion-corrected DWI) were first rigidly aligned with T1 image of the same infant and further up-sampled into an isotropic resolution of $1 \times 1 \times 1$ mm^3. T1 images were acquired with 144 sagittal slices using parameters: TR/TE $= 1900/4.38$ ms, flip angle $= 7°$, resolution $= 1 \times 1 \times 1$ mm^3. T2 images were acquired with 64 axial slices using parameters: TR/TE $= 7380/119$ ms, flip angle $= 150°$ and resolution $= 1.25 \times 1.25 \times 1.95$ mm^3. Diffusion-weighted MR images were acquired with 60 axial slices using parameters: TR/TE $= 7680/82$ ms, resolution $= 2 \times 2 \times 2$ mm^3, 42 noncollinear diffusion gradients, and b $= 1000$ s/mm^2; and seven non-diffusion-weighted reference scans were also acquired. The skull, cerebellum and brain stem finally removed from the aligned T2 and FA images with in-house tools. One example is shown in Fig. 1.1.

To generate manual segmentation for training, initial segmentation was first obtained with a publicly available infant brain segmentation software, iBEAT[4] [43]. Then, manual editing was carefully performed by an experienced rater according to T1, T2 and FA images for correcting possible segmentation errors.

[4] http://www.nitrc.org/projects/ibeat

3.1.5.2 Evaluation Metrics and Comparison Methods

In the experiments, I adopt DSC (Eq. 2.22) and MHD (Eq. 2.23) as the quantitative performance metrics.

To show the superiority of my proposed method, the following methods are adopted as comparison methods in the experiments:

1. FMRIB's Automated Segmentation Tool (FAST) [222].

2. Majority Voting (MV): For the 11-subject dataset, I employed a leave-one-strategy (for any testing image, the remaining 10 subjects are used as atlases) and I used ANTs [6] for registration based on T1 MRI.

3. Random Forest (RF): In my implementation, for each tissue type, I randomly selected 10,000 training voxels for each class label from each training subject. Then, from the $7 \times 7 \times 7$ patch of each training voxel, 10,000 random Haar-like features were extracted from all source images: T1, T2, FA images, and then I trained 20 classification trees. I stopped the tree growth at a certain depth (*i.e.*, D = 50), with a minimum number of 8 samples for each leaf node (smin = 8). Note that these settings have been optimized in LINKS [190].

4. Random Forest with auto-context model (LINKS) [190]: Based on the above-mentioned RF, I then apply the auto-context model to iteratively refine the results. Specifically, I not only extracted Haar-like features from all source images: T1, T2, FA images, but also three probability maps of WM, GM and CSF. All the features were used to train the RF (Note, it is now auto-context refined, and called LINKS). In each iteration, the training of model followed the RF settings introduced above.

5. Training TFmUNet for 2D patches along each dimension, after which I perform majority voting for the results from different TFmUNet models (2D-TFmUNet).

6. 3D-CNN: The network shares the same three layer groups in Fig. 3.1, and followed by three fully connected layers, and the output corresponds to the center voxel of the input patch.

In the following sections, I first discuss and evaluate the parameter selections as well as training strategies, and then present comparison experiments with state-of-the-art methods.

3.1.5.3 Impact of Patch Size

Patch size plays an important role in CNN-based methods, since it regulates the tradeoff between localization accuracy and the use of context [38]. Fortunately, UNet-based architectures can slightly mitigate the impact of patch size on the segmentation tasks [153, 164]. To investigate the impact of patch size in this project, I conduct several experiments using 4 different input patch sizes: $16 \times 16 \times 16$, $32 \times 32 \times 32$, $48 \times 48 \times 48$ and $64 \times 64 \times 64$, for training the same UNet architecture shown in Fig. 3.2. Fig. 3.3 shows the respective results (*i.e.*, Dice ratio of segmentation as a function of patch size).

As shown in Fig. 3.3, the segmentation performance is the worst with the patch size of $16 \times 16 \times 16$ due to the smallest context information. With the patch size of $32 \times 32 \times 32$, the segmentation task obtains the best performance. It is interesting to note that the Dice ratio becomes slightly worse when the patch size grows. This may be due to the fact that, when I use larger patch size, I will have smaller available number of patches to train the model, thus resulting in a little lower performance. The result also shows that the UNet-based architecture is somewhat robust to patch size when the patch size is larger than a certain value. This advantage roots from the fact that the localization accuracy is stable when using UNet-based architectures. On the other hand, when the patch size is larger, the improvement may be limited by the number of extracted patches. With the insight gained through this experiment, I set the patch size to be $32 \times 32 \times 32$ throughout all experiments below.

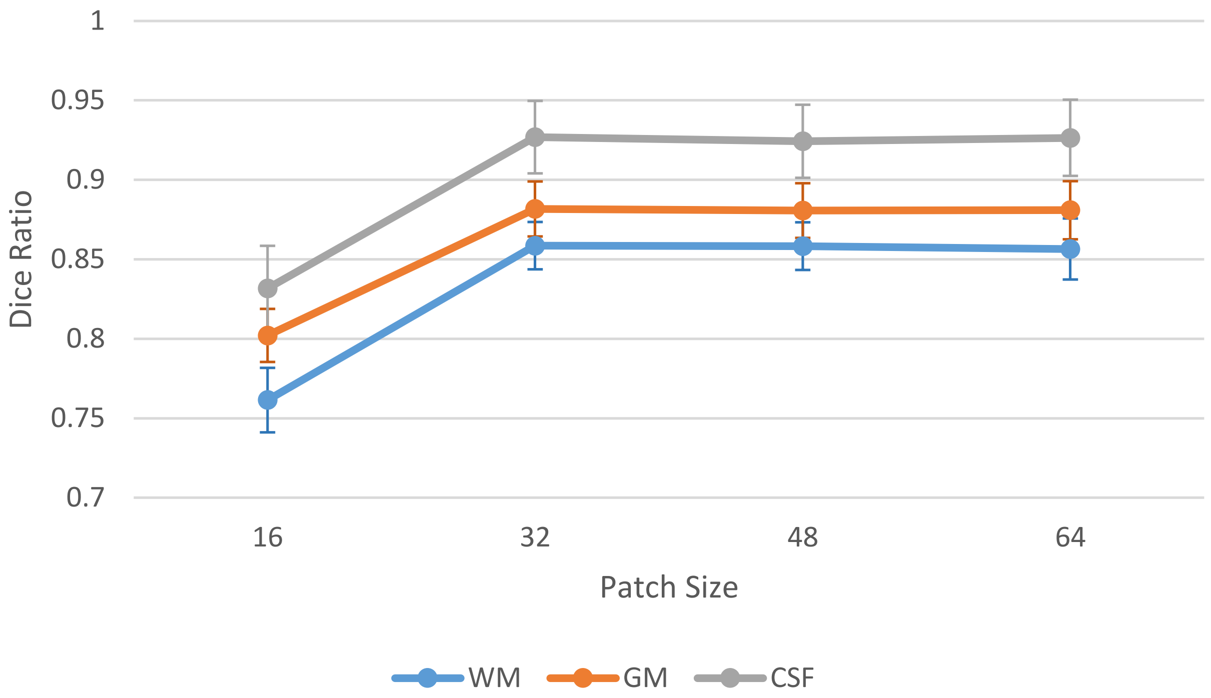


Figure 3.3: Changes of Dice ratios of WM, GM and CSF on 11 isointense subjects, with respect to different patch sizes. Here, leave-one-subject-out cross validation is used.

3.1.5.4 Importance of Multi-modality Information

To demonstrate the effectiveness of using multi-modality data (*i.e.*, T1, T2 and FA in the experiments), I run the same model for each imaging modality separately, or together. Fig. 3.4 illustrates the Dice ratios of my proposed method with respect to different combinations of three imaging modalities. It can be seen that using more imaging modalities generally results in more accurate segmentations than using any single imaging modality. Moreover, using all three imaging modalities provides the best performance, compared to the cases of using any two imaging modalities. This indicates that the multi-modality information is useful for guiding tissue segmentation. The same conclusion can also be drawn by looking into the experimental results on the second dataset as described below.

It is worth noting that I adopt *early fusion* strategy for multi-modal data fusion in this study. Actually, *late fusion* is another common choice for multi-modal data fusion, which employs multi-pathway encoder to model multi-modal data and adopts a concatenation

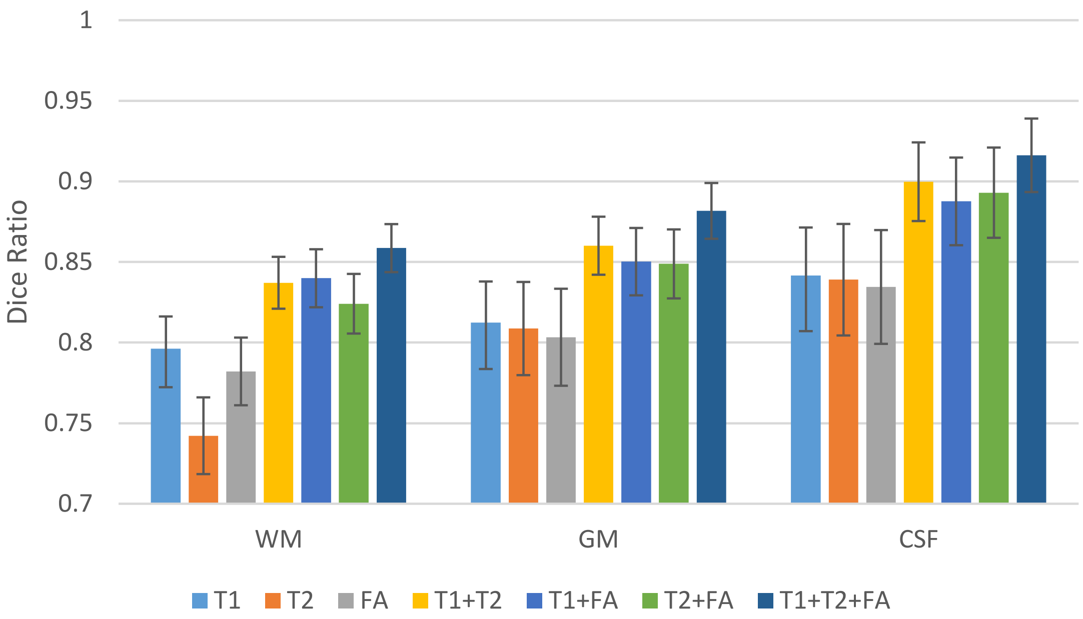


Figure 3.4: Average Dice ratios of the proposed method with respect to different combinations of 3 imaging modalities.

layer at a certain decoder layer to fuse the information from different branches. In our experimental settings, the two architectures actually present very similar performance which is consistent with [46]. Since the multi-pathway encoder based network occupies much more memory and include more parameters, we select the early fusion strategy in this study, as shown in Fig. 3.2.

3.1.5.5 Importance of Using the Transformation and Fusion Modules

As described in Sec. 3.1.2, I design a transformation block composed of convolution layers to operate on the shallow-layer feature maps before concatenating them with deep-layer feature maps. This block is used to reduce the potential bias effect when directly copying the signals to the higher layer. The fusion module is also utilized to better fuse the combined features. Experiments are carried out to compare the networks with (3D-TFmUNet) and without (3D-mUNet) the transformation and fusion modules on the 11-subject dataset. All experimental settings are the same except using the transformation and fusion modules. The experimental results in terms of Dice ratio are shown in Fig. 3.5.

As can be seen, 3D-TFmUNet outperforms 3D-mUNet on all three brain tissues, suggesting that the application of the designed transformation and fusion modules contributes to enhancing the discrimination capability. It is worth noting that the model complexity (*i.e.*, the number of parameters) only increases by 1.39% with using the designed transformation and fusion modules.

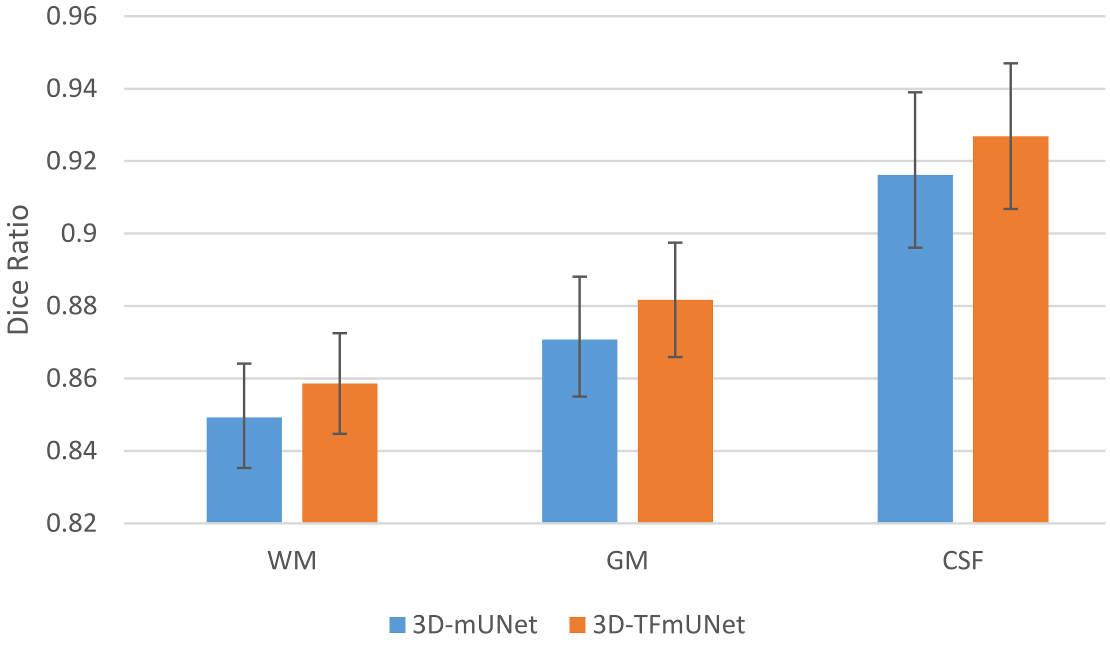


Figure 3.5: Comparison of 3D-TFmUNet and 3D-mUNet.

3.1.5.6 Impact of Network Initialization Strategies

The impact of initialization strategies is explored towards the network training and the segmentation performance. Specifically, the following strategies are used to initialize the network, respectively, i.e., Constant (0), Gaussian (0,0.01), and Xavier [59]. All other settings are kept the same, and the convergence with respective to different initialization strategies is presented in Fig. 3.6(a). Obviously, the "Constant" initialization cannot guarantee a converged training, while "Gaussian" and "Xavier" initializations can both result in good convergence. As it is obvious in Fig. 3.6(b), using "Xavier" to initialize the network leads to the best performance. Hence, Xavier initialization algorithm is adopted to initialize the networks in this study.

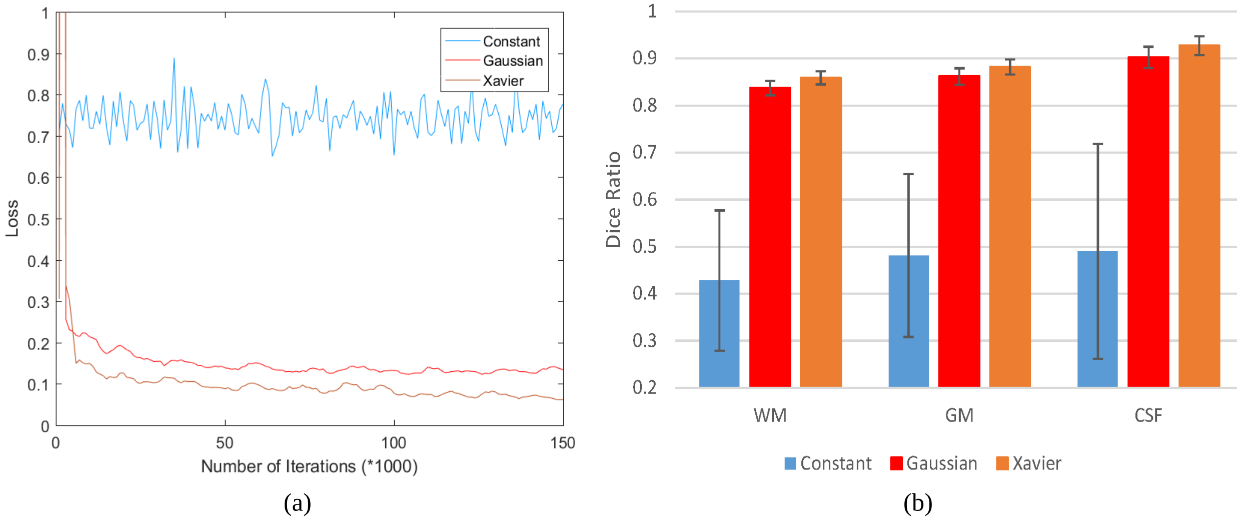


Figure 3.6: Comparison of different initialization strategies in the proposed model. The diagram (a) depicts the convergence situation, and (b) shows the segmentation performance.

3.1.5.7 Impact of Pooling Layers: Pooling-Included vs. Pooling-Excluded Networks

In the FCN-based network architectures (Pooling-Included network, *e.g.*, the proposed designed network), pooling is an important component to increase the receptive field dramatically and produce invariant feature representation. However, it will also result in the loss of spatial information. In the proposed designed network (Fig. 3.2), skip connections are utilized to aggregate the shallower (high-resolution) layers and the deeper (highly-semantic) layers, aiming at making up the lost information. Another possible way to avoid the loss of spatial information is to exclude the pooling layers and only include the convolutional layers (Pooling-Excluded network). To see the importance of pooling layers, I conducted comparison experiments between the proposed designed network with pooling layers and two Pooling-Excluded networks: 1) Self-Designed: the network is with 11 convolutional layers, with each layer having $3 \times 3 \times 3$ convolution filter (to satisfy the requirement of receptive field, as the input patch size is $32 \times 32 \times 32$), but no pooling layer; 2) DeepMedic: the popular 3D multi-scale CNN (DeepMedic) [95], which was well-designed and showed excellent performance in lesion labeling tasks. The experimental results are given in Fig. 3.7. The Pooling-Included network (*i.e.*, the proposed model) works better than the other two Pooling-Excluded networks, indicating that the use of Pooling-Included networks with skip-connection is a better choice for the semantic segmentation tasks. DeepMedic works better than the Self-Designed network, because multi-scale designed network partially alleviates the insufficient receptive field problem and learns invariant features.

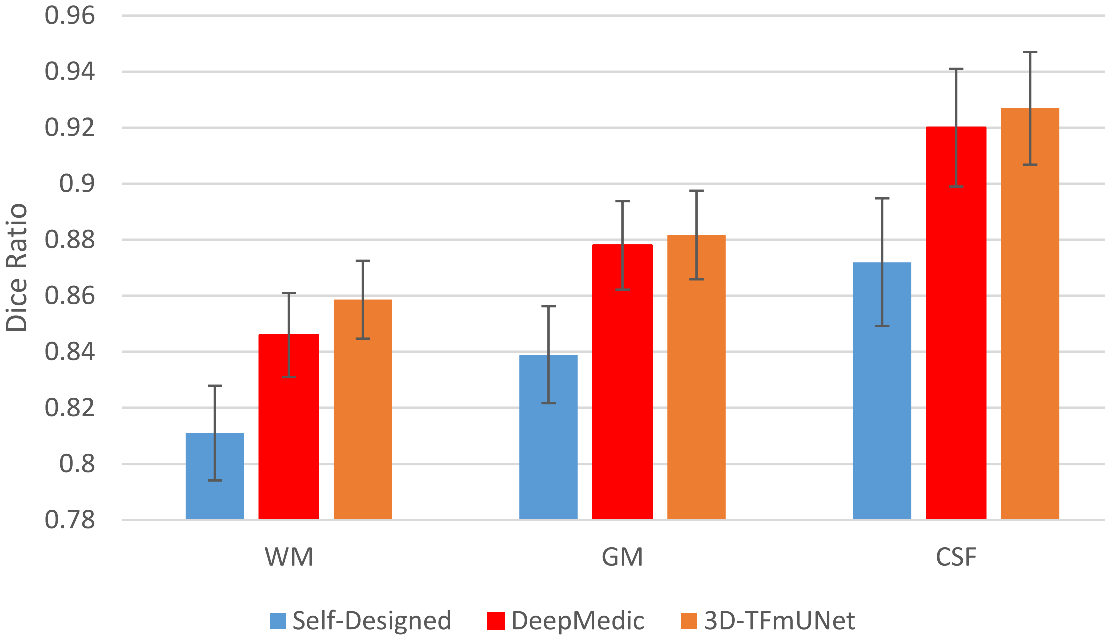


Figure 3.7: Comparison between Pooling-Included network and Pooling-Excluded networks (Self-Designed and DeepMedic).

3.1.5.8 Impact of Upsampling Strategies

Upsampling layers (*e.g.*, deconvolution layer in the proposed model) play an important role in the deep learning based segmentation models. There are several upsampling strategies: 1) "Multi-Linear" (using multi-linear interpolation to upsample the input feature maps), 2) "Deconvolution" (upsampling by learning to deconvolve the input feature map) [124], and 3) “Index-Upsampling” (using the max pooling indices to upsample (without learning) the feature map(s) and convolve with a bank of trainable filters) [8]. Experiments are conducted to investigate which up-sampling strategy works best for the segmentation task. Note, all the other settings are the same, except that I use different upsampling strategies for the comparison experiments. The comparison experimental results are shown in Fig. 3.8, indicating that Deconvolution works best, and Index-Upsampling provides close performance to Deconvolution, while Multi-linear interpolation leads to the worst results. Thus, I select Deconvolution to work as the upsampling strategy in this study.

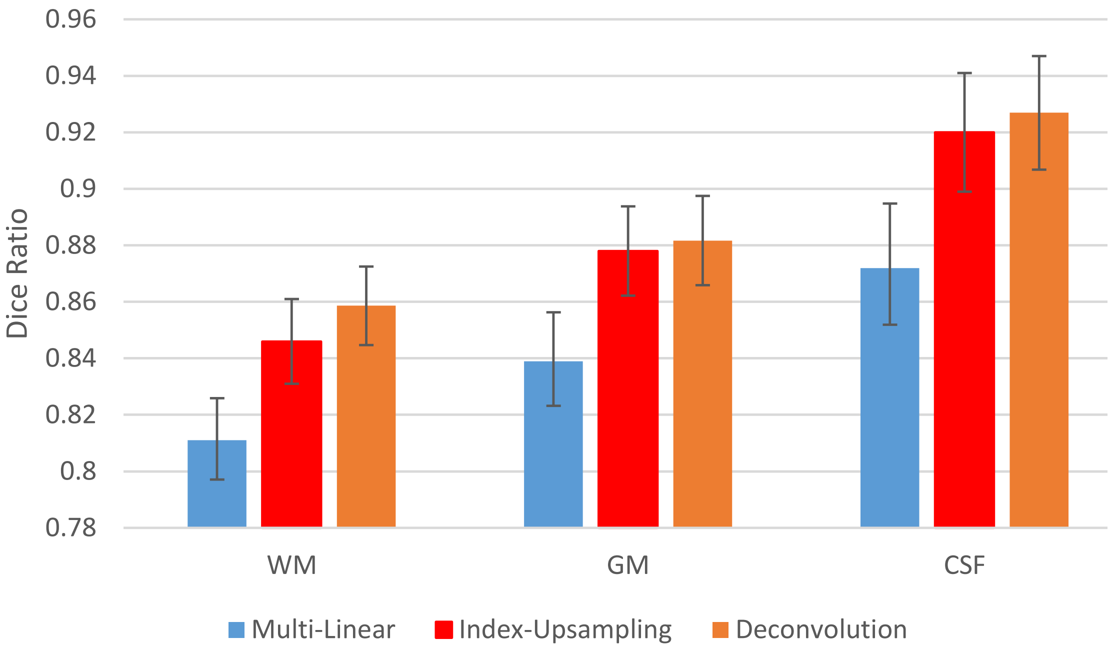


Figure 3.8: Segmentation performance with respect to the use of different upsampling strategies for the proposed model.

3.1.5.9 Patch Merging Strategy

The 3D-TFmUNet is trained in a patch level. In the testing stage, I have to first partition a whole image (T1,T2 and FA) into overlapping or non-overlapping patches, and then feed these patches into trained networks. The corresponding output patches are further merged to form a fully-predicted label map. It is worth noting that the extent of patch overlapping can largely affect the final performance, in terms of both segmentation accuracy and computational complexity. Thus, the patch overlapping extent is explored in the testing stage. Specifically, the patches are extracted from MR images with different overlapping extents, *i.e.*, a step size of 32 (non-overlapping), 16, 8 and 4, respectively. After these patches are fed into the trained model, all the predicted label patches from the same subject are combined into a single label image by averaging (or majority voting) the label values of the overlapping image regions. The performance, in terms of Dice ratio and time cost, are given in Fig. 3.9 and Fig. 3.10, respectively.

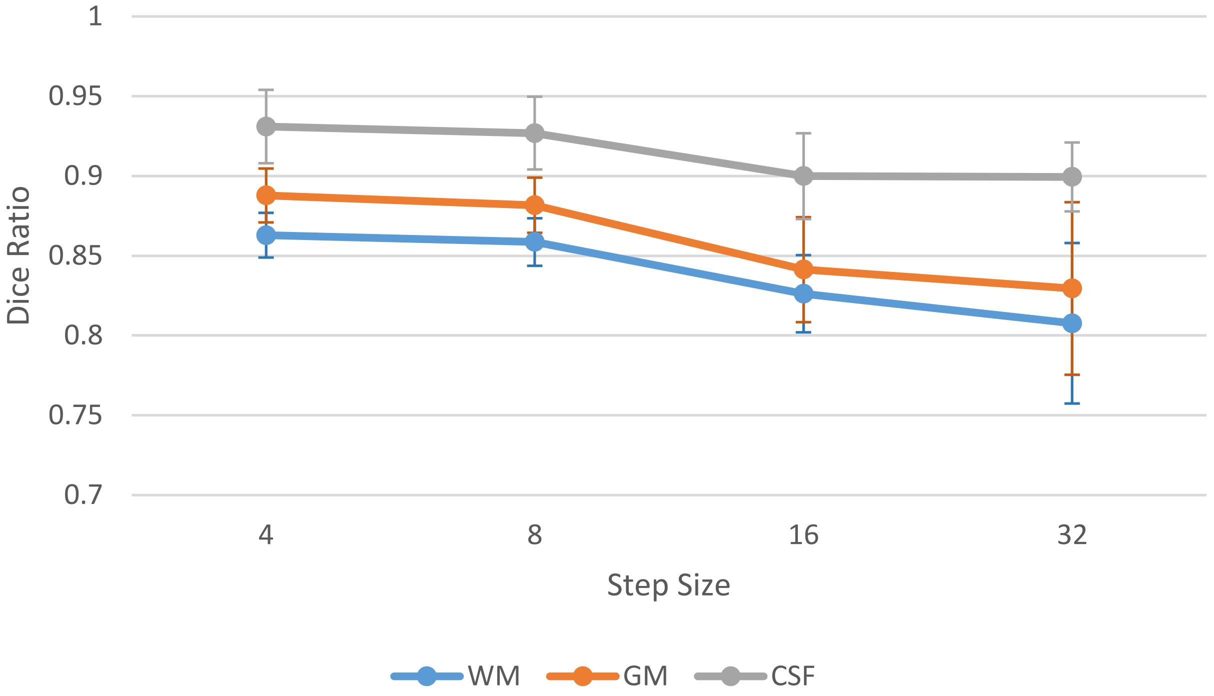


Figure 3.9: Changes of Dice ratios of WM, GM and CSF on 11 isointense subjects, with respect to different step size at the testing stage. Leave-one-subject-out cross validation is used.

As shown by both figures (Fig. 3.9 and Fig. 3.10), the smaller the step size, the better the segmentation accuracy. This is benefiting from the ensemble effect since more predicted labels can be averaged when the step size is smaller. However, small step size brings heavy work load and makes the computational cost increase dramatically. To this end, I

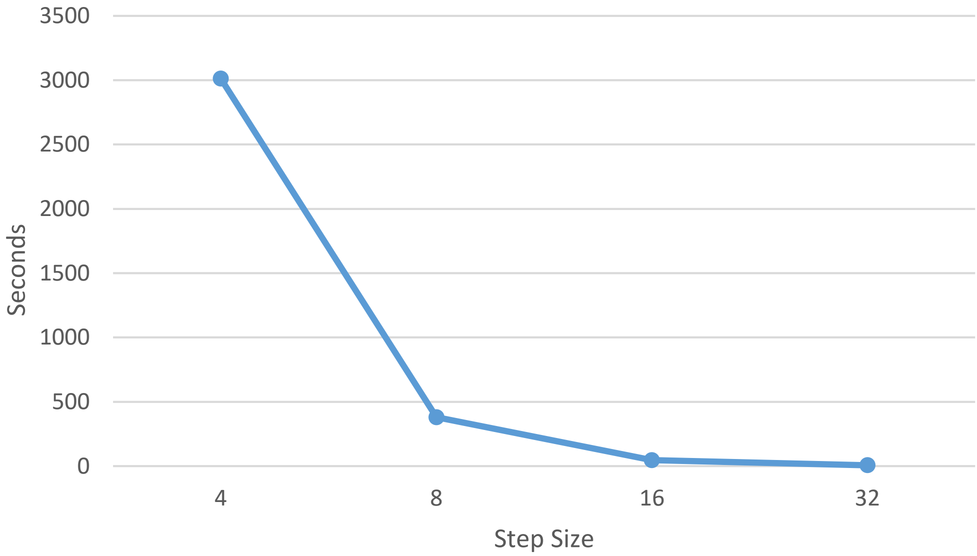


Figure 3.10: Changes of average time cost of tissue segmentation for one subject, with respect to different step size at the testing stage.

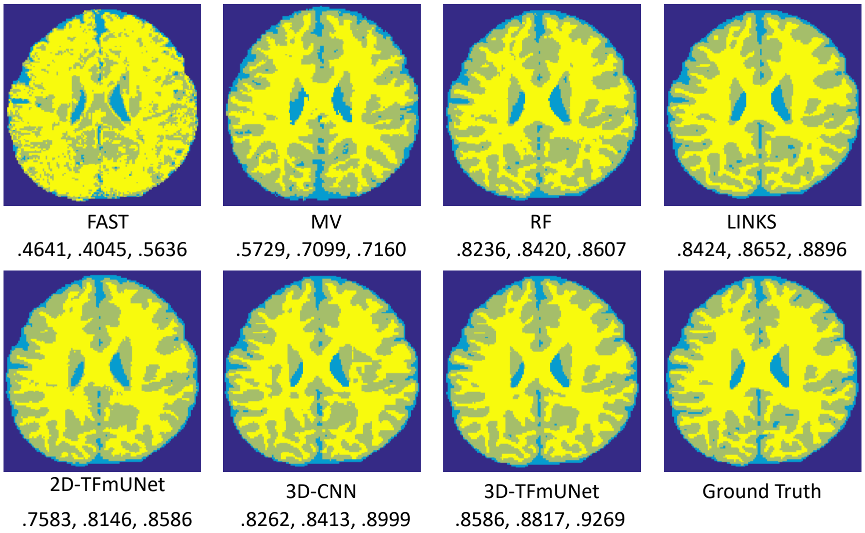
conduct a tradeoff between computational cost and prediction accuracy, and finally select the step size as 8 for the testing stage.

3.1.5.10 Experimental Results on the 1st Dataset

Experiments are first carried out on the 1st dataset containing 3 imaging modalities (T1, T2 and FA). The proposed 3D-TFmUNet is used to perform voxel-wise tissue segmentation with a leave-one-subject-out strategy. It takes approximately 130 hours to train the designed neural networks on a Titan X GPU.

To qualitatively demonstrate the advantage of the proposed 3D-TFmUNet on this dataset, I first show the segmentation results of different tissues for a typical subject in Fig. 3.11. The proposed method could achieve better visual results, especially for the tiny tissues.

To quantitatively evaluate segmentation performance, I use Dice ratio to measure the overlap ratio between automated and manual segmentation results. I report the segmentation performance in Table 3.1. I can observe that 3D-TFmUNet outperforms other methods (p=0.0371, performed by a paired t-test). Specifically, 3D-TFmUNet could achieve the


FAST
.4641, .4045, .5636

MV
.5729, .7099, .7160

RF
.8236, .8420, .8607

LINKS
.8424, .8652, .8896

2D-TFmUNet
.7583, .8146, .8586

3D-CNN
.8262, .8413, .8999

3D-TFmUNet
.8586, .8817, .9269

Ground Truth

Figure 3.11: Comparison of segmentation results by different methods, along with manual ground truth on a typical infant subject as shown in Fig. 1.1. The Dice ratios for WM, GM and CSF are listed below each method, respectively.

average Dice ratios of 0.9269 for CSF, 0.8817 for GM, and 0.8586 for WM, from these 11 subjects. In contrast, one of the state-of-the-art methods, *i.e.*, Random Forest with auto-context model (LINKS) [190], achieved the overall Dice ratios of 0.8896, 0.8652, and 0.8424 for CSF, GM and WM, respectively.

Table 3.1: Segmentation performance in terms of Dice ratio and standard deviation, achieved by the baseline comparison methods and my 3D-TFmUNet on 11 subjects. The highest performance in each tissue class is highlighted in bold.

	WM	GM	CSF
FAST	.4641(.0791)	.4045(.0979)	.5636(.2334)
MV	.5729(.0327)	.7099(.0412)	.7160(.0386)
RF	.8236(.0164)	.8420(.0126)	.8607(.0170)
LINKS	.8424(.0183)	.8652(.0154)	.8896(.0251)
DeepMedic [95]	.8458(.0150)	.8608(.0158)	.9196(.0210)
3D-UNet [37]	.8418(.0161)	.8680(.0172)	.9012(.0225)
2D-TFmUNet	.7583(.0215)	.8146(.0115)	.8586(.0101)
3D-CNN	.8262(.0242)	.8413(0.0329)	.8999(.0269)
3D-TFmUNet	**.8586**(.0139)	**.8817**(.0158)	**.9269**(.0201)

I also provide WM surfaces obtained by different methods in Fig. 3.12. These WM surfaces qualitatively demonstrate the advantage of the proposed method, as it achieves the best visual results.

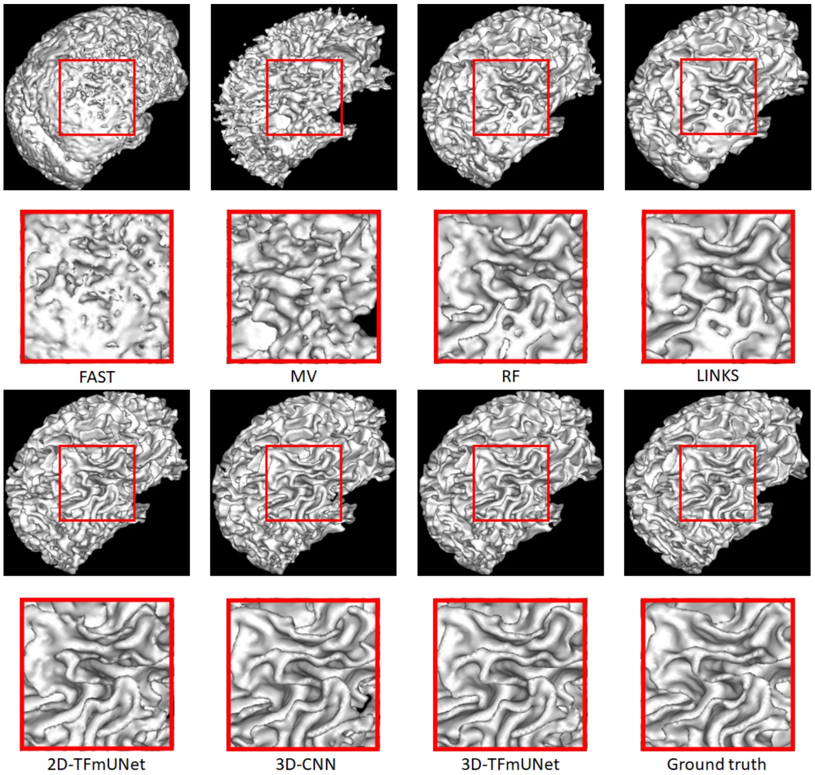


Figure 3.12: Comparison of WM surfaces obtained by different methods on a typical subject shown in Fig. 1.1.

Table 3.2: Segmentation performance in terms of MHD and standard deviation, achieved by the baseline comparison methods and my 3D-TFmUNet on 11 subjects. The highest performance in each tissue class is highlighted in bold.

	WM	GM	CSF
FAST	1.7052(.0092)	1.0083(.0269)	1.8276(.0882)
MV	1.7204(.4780)	1.1840(.1639)	1.9323(.2075)
RF	.6837(.0893)	.5625(.0502)	.4624(.0669)
LINKS	.5659(.0794)	.4827(.0460)	.3321(.0426)
DeepMedic [95]	.5154(.0540)	.4878(.0414)	.3621(.0495)
3D-UNet [37]	.5951(.0488)	.4420(.0415)	.3630(.0488)
2D-TFmUNet	.7773(.1907)	.6011(.1082)	.5379(.1210)
3D-CNN	.5802(.1386)	.4967(.0839)	.3561(.0817)
3D-TFmUNet	**.3423**(.0358)	**.3108**(.0256)	**.3230**(.0788)

The MHD comparison is also shown in Table 3.2. It can be seen again that the proposed method produces more competitive performance compared to all the state-of-the-art methods.

3.1.5.11 Comparison of 3D-based Model and 2D-based Model

As medical image data is often acquired in 3D, 3D operations are assumed to offer better performance than 2D operations. So, it is important to highlight how the 3D architecture improves the performance, compared to the conventional 2D architectures. Specifically, I run the proposed network with exactly same architecture, but with all 2D filters, and then provide results in Table 3.1. As can be seen, the 3D model performs approximately 11% better than the 2D approach. This advantage comes from the fact that 3D convolution filters can consider the 3-dimensional structures, as input images are in 3D. Thus, 3D operations can model internal structures much better. Furthermore, adopting 3D operations can avoid inconsistency in the 3rd dimension of the images, which is usually a problem when simply applying 2D filters to the 3D images.

3.1.5.12 Time and Computational Complexity

As the model training can be completed offline, the testing time cost is often more important for a segmentation task. Thus, in Table 3.3, the average time cost of segmenting one test subject by each segmentation approach is provided. Note that this experiment is performed on the same PC with the following settings – Memory: 16GB Quad Channel DDR4; Video Card: Nvidia Titan X; Processor: Intel i7-5820K; Operation System: Ubuntu 14.04. The running time in minutes are listed in Table 3.3 for different segmentation methods. As can be seen, the proposed 3D-TFmUNet method is significantly faster than any of the other methods.

Table 3.3: Average Time cost (in minutes) of each test subject and standard deviation, by the baseline comparison methods and my proposed 3D-TFmUNet on 11 subjects in the 1st dataset.

FAST	MV	RF	LINKS	DeepMedic [95]	3D-UNet [37]	2D-TFmUNet	3D-CNN	3D-TFmUNet
10.02(0.14)	367.82(10.20)	8.31(0.07)	17.20(0.10)	15.70(0.16)	7.30(0.18)	240.76(5.14)	1660.85(1.13)	**6.71**(0.12)

3.1.5.13 Results on the 2nd Dataset

To show the generalization ability of the proposed 3D-TFmUNet architecture, I further conduct experiments on the 2nd large dataset with 50 other isointense-phase subjects, where each subject has both T1 and T2 images, but no FA images. 5-fold cross-validation is performed. Note that, in each fold, it takes approximately 25 hours to train the proposed designed neural networks on a Titan X GPU.

In Fig. 3.13, I visualize the segmentation results of a typical subject by different methods in the 2nd dataset. Obviously, the proposed method has provided much better qualitative performance, especially in the tiny regions.

To quantitatively evaluate segmentation performance, Dice ratios are reported in Table 3.4. The proposed 3D-TFmUNet again achieves the best performance in segmenting WM (0.9190±0.0085), GM (0.9401±0.0052) and CSF (0.9610±0.0090), compared to the state-of-the-art method [190] which uses multi-scale RF and auto-context model to take advantage of multi-source information and refine the results. The proposed 3D-TFmUNet also outperforms the conventional CNN, indicating that the proposed 3D-TFmUNet is more capable in this segmentation task. Furthermore, the proposed 3D-TFmUNet works better than 2D-TFmUNet in all three tissues, indicating that 3D deep learning architectures are better for 3D segmentation tasks than the 2D deep learning architectures. I also provide WM surfaces obtained by different segmentation methods in Fig. 3.14, which further demonstrates the advantage of the proposed 3D-TFmUNet. Moreover, the MHD comparison is further provided in Table 3.5, where the proposed 3D-TFmUNet produces less errors compared to all the state-of-the-art methods.

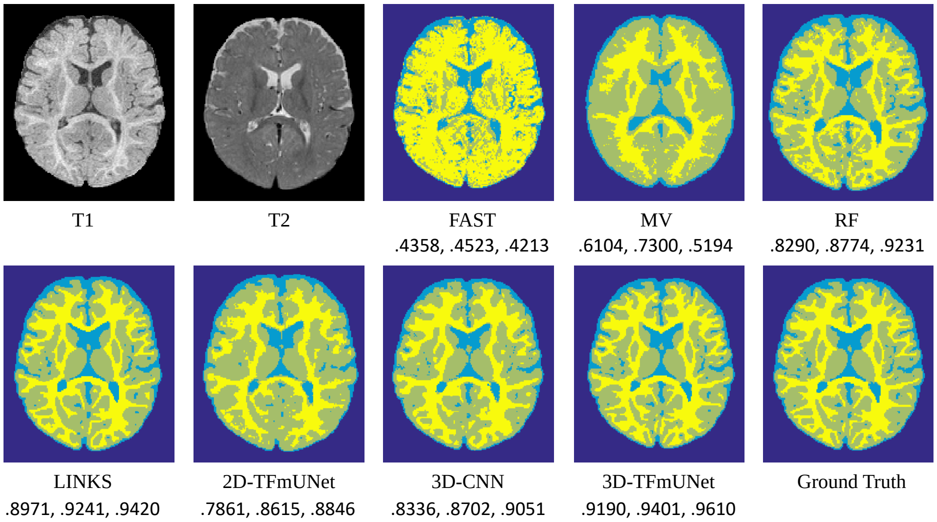


Figure 3.13: Comparison of segmentation results by baseline comparison methods and the proposed 3D-TFmUNet, along with manual ground truth, on a typical subject in the 2$^{\text{nd}}$ dataset. The Dice ratios for WM, GM and CSF are listed below each method, respectively.

Table 3.4: Segmentation performance in terms of Dice ratio and standard deviation, obtained by the baseline comparison methods and the proposed 3D-TFmUNet on 50 subjects. The highest performance in each tissue class is highlighted in bold.

	WM	GM	CSF
FAST	.4358(.0903)	.4523(.1397)	.4213(.3282)
MV	.6104(.0163)	.7300(.0196)	.5194(.0316)
RF	.8290(.0150)	.8774(.0059)	.9231(.0127)
LINKS	.8971(.0074)	.9241(.0043)	.9420(.0074)
DeepMedic [95]	.8943(.0088)	.9265(.0051)	.9484(.0086)
3D-UNet [37]	.8907(.0087)	.9228(.0049)	.9465(.0095)
2D-TFmUNet	.7861(.0240)	.8615(.0080)	.8846(.0177)
3D-CNN	.8336(.0177)	.8702(.0070)	.9051(.0168)
3D-TFmUNet	**.9190**(.0085)	**.9401**(.0052)	**.9610**(.0090)

Table 3.5: Segmentation performance in terms of MHD and standard deviation, obtained by the baseline comparison methods and the proposed 3D-TFmUNet on 50 subjects. The highest performance in each tissue class is highlighted in bold.

	WM	GM	CSF
FAST	1.7161(.2884)	1.0184(.3712)	1.1053(.5241)
MV	1.4101(.0841)	1.0962(.2187)	1.3438(.6356)
RF	.7757(.0571)	.6513(.0321)	.3090(.0250)
LINKS	.4515(.0217)	.3961(.0108)	.2565(.0282)
DeepMedic [95]	.4601(.0303)	.3990(.0210)	.2736(.02922)
3D-UNet [37]	.4948(.0230)	.4003(.0119)	.2285(.0187)
2D-TFmUNet	.8424(.0373)	.8766(.0424)	.9300(.0138)
3D-CNN	.6867(.0443)	.6912(.0576)	.7401(.0756)
3D-TFmUNet	**.3676**(.0223)	**.3530**(.0110)	**.1890**(.0122)

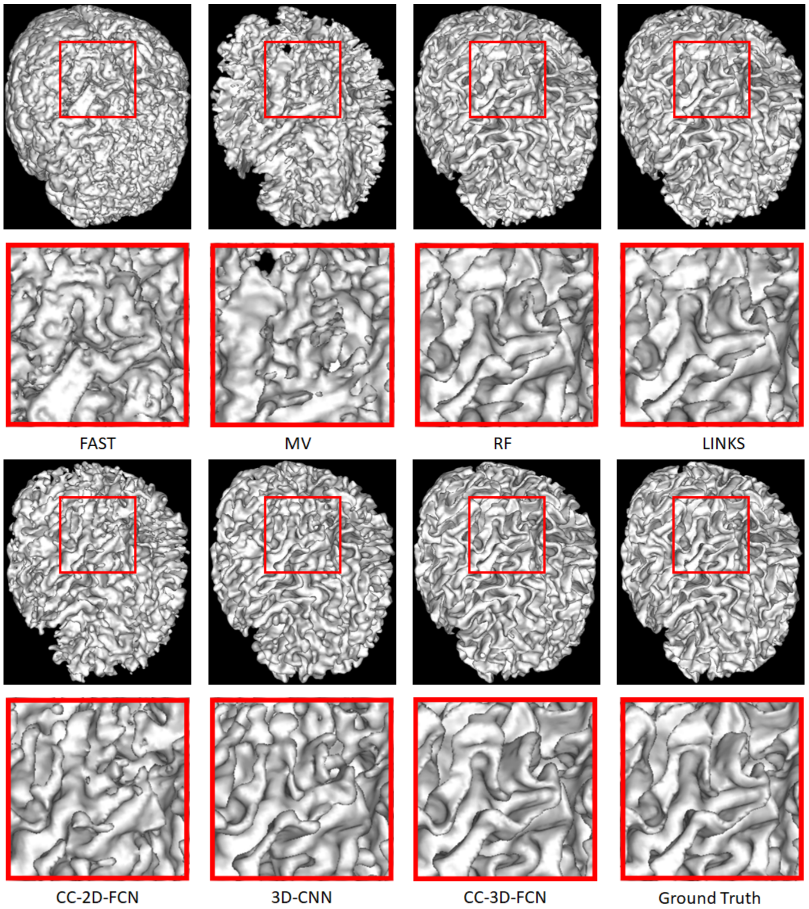


Figure 3.14: Comparison of WM surfaces obtained with different methods on a typical subject (of the 2$^{\text{nd}}$ dataset) shown in Fig. 3.13.

3.1.5.14 Limitation of the Proposed Method

The main problem of the proposed method is that the segmented infant brain images are a little bit smooth, especially for the tissues near the boundaries, as shown in Fig. 3.11 and Fig. 3.13. This is mainly due to the use of convolution and pooling in the convolutional networks. On the other hand, although high localization accuracy can be obtained because of using local image patch, global context information is missed to guide the spatial consistency of tissue segmentation. This is because patch size is much smaller compared to the whole image size, and thus I can not use the whole image information to train the proposed model; on the other hand, if using the whole image for training, the number of training images is too small to train the neural network. I will investigate these two issues in the future work.

3.2 High-Resolution Encoder-Decoder Networks for Low-Contrast Medical Image Segmentation

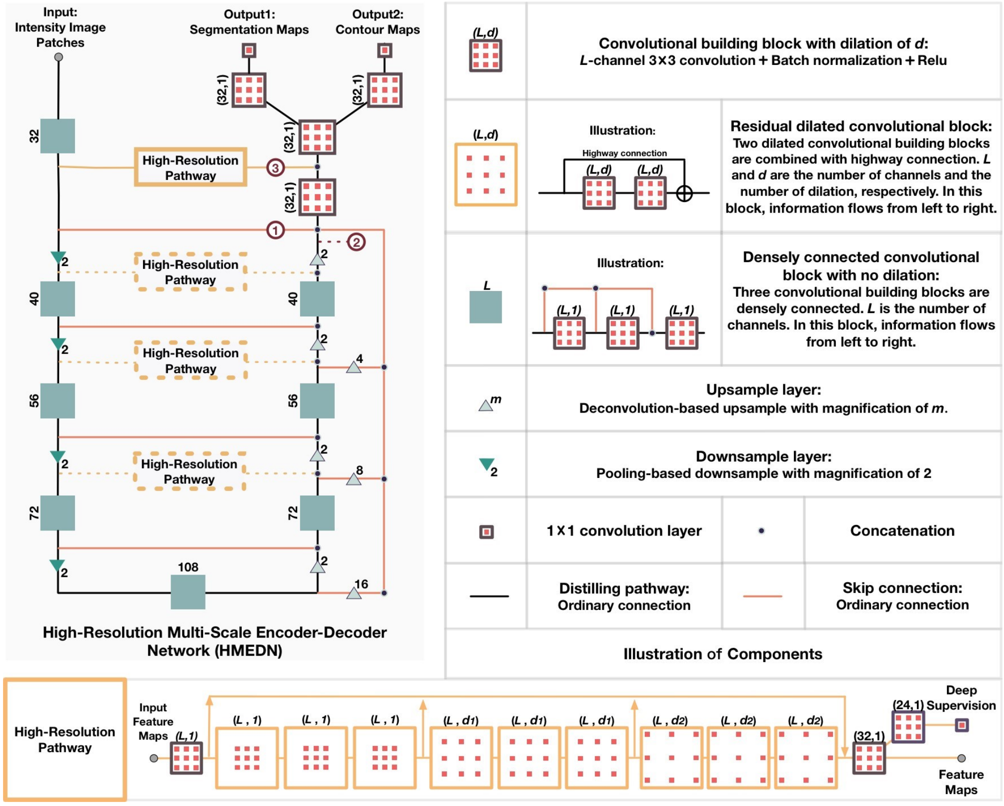
In this section, the proposed High-Resolution Multi-Scale Encoder-Decoder Network (HMEDN) is introduced for segmentation of single-modal low-contrast medical images which have been mentioned in Sec. 1.1.4. Specifically, four strategies are proposed to solve the low-contrast issue for segmentation, especially the low-contrast regions around the boundaries. First, the distilling network is introduced, in which semantic information is carefully distilled and preserved. Then, the high-resolution pathway, constructed by densely connected dilated convolution operations for high-resolution semantic information exploitation, is elaborated. Next, the task of contour regression with the task of organ segmentation is integrated for accurate boundary localization. Finally, the network is forced to concentrate more on the ambiguous boundary area by designing a difficulty-guided cross-entropy loss function. Fig. 3.15 illustrates the proposed network. Comparison experiments and ablation study are carried out in Sec. 3.2.5.

3.2.1 Distilling Network

The first proposed strategy to segment low-contrast medical image is to provide a more comprehensive multi-scale information collection and fusion mechanism. In general, two structures are usually adopted for multi-scale information preservation in the literature, *i.e.*, UNet [164] and Holistically-nested Edge Detection (HED) [201]. In the UNet, multi-scale information is gradually merged by fusing the upsampled large receptive field layers with those passed through small receptive field layers via skip connection (*i.e.*, merging feature maps of the same scale at a time). In contrast, by fusing the feature maps from multiple scales into the final output at a time, the HED methods acquire multi-scale information in a more direct manner.

Without the complicated convolution operations in the decoding procedure, these networks bring the multi-scale information together in the original form. To preserve multi-scale information, UNet gradually integrates and delicately processes the information, thus making the fusion of the information sufficient. Moreover, UNet implicitly utilizes the intermediate results to guide the subsequent fusion. Whereas in the case of HED methods, since all information is processed at the same time, the fusion of multi-scale information can be done more comprehensively. To take advantage of both types of networks, the U-Net structure and the side outputs of HED networks are inherited to construct the network. Moreover, to further encourage smooth information flow between different layers and make the training of the network more manageable, dense connections [82] are proposed to replace the original plain connections.

Based on the above-mentioned intuitions, a densely connected multi-scale encoder-decoder network is proposed to comprehensively reveal the multi-level structural information. This network is denoted as distilling network, due to the use of downsampling layer, which can efficiently enlarge the receptive field and effectively filter the redundant insignificant components. As shown in Fig. 3.15, the outline of the distilling network (the black pathway, together with the orange skip connections) is a U-Net with four downsampling and four upsampling layers. However, besides the regular skip connections, three extra side channels from intermediate layers with different sizes of receptive fields are also upsampled and merged with the main channel of the network to encourage more comprehensive multi-scale information fusion. Moreover, by linking all the preceding layers to the final layer, I construct dense blocks (*i.e.*, those solid green rectangles in Fig. 3.15 and use them as the building block to encourage smooth information flow within the network.



Input:
Intensity Image
Patches
Output1:
Segmentation Maps
Output2:
Contour Maps
(32,1)
(32,1)
(32,1)
(32,1)
32
High-Resolution
Pathway
3
1
2
2
40
40
2
High-Resolution
Pathway
2
4
56
56
High-Resolution
Pathway
2
2
8
72
72
2
108
2
16
High-Resolution Multi-Scale Encoder-Decoder
Network (HMEDN)
(L,d)
Convolutional building block with dilation of d:
L-channel 3×3 convolution + Batch normalization + Relu
(L,d)
Illustration:
Highway connection
(L,d)
(L,d)
Residual dilated convolutional block:
Two dilated convolutional building blocks are combined with highway connection. L and d are the number of channels and the number of dilation, respectively. In this block, information flows from left to right.
L
Illustration:
(L,1)
(L,1)
(L,1)
Densely connected convolutional block with no dilation:
Three convolutional building blocks are densely connected. L is the number of channels. In this block, information flows from left to right.
m
Upsample layer:
Deconvolution-based upsample with magnification of m.
2
Downsample layer:
Pooling-based downsample with magnification of 2
1×1 convolution layer
Concatenation
Distilling pathway:
Ordinary connection
Skip connection:
Ordinary connection
Illustration of Components
High-Resolution
Pathway
Input
Feature
Maps
(L,1)
(L , 1)
(L , 1)
(L , 1)
(L , d1)
(L , d1)
(L , d1)
(L , d2)
(L , d2)
(L , d2)
(32,1)
(24,1)
Deep
Supervision
Feature
Maps

Figure 3.15: Illustration of the structure of the proposed high-resolution multi-scale encoder-decoder network (HMEDN). The input is a set of intensity image patches and the outputs are segmentation and contour probability maps. Rectangles and triangles represent operations in the network. Three kinds of pathways, *i.e.*, skip connection (pathway 1), distilling pathway (pathway 2) and high-resolution pathway (pathway 3) connect all kinds of operations and form the network.

3.2.2 High-Resolution Pathway

Our second (and main) strategy is to endow the network with a better capacity to extract discriminative high-resolution semantic information. In the task of segmentation, the intuitive tension between what and where has long been realized in [124]. The solution to the problem in the current literature is to combine the coarse layers with fine layers in the encoder-decoder networks by skip connections and allow the networks to make local decisions concerning the global structures. This strategy works well in the high-contrast images with clear and consistent boundaries. However, when it is applied to the images with low contrast, local appearance features extracted by lower layers may fail to refrain from the surrounding hypothetical boundaries and recognize the vanishing boundaries, causing negative effects on the accuracy of these algorithms. Consequently, to achieve accurate boundary localization in blurry images, a mechanism which can provide discriminative high-resolution contextual information is needed. To meet this special demand, the dilated convolution-based pathways are introduced. Given a 2D image X with C channels, the definition of a dilated convolution (dilation ratio as d) with kernel W of size K is defined as Eq. 3.4:

$$Z_{i,j} = \sum_{p=0}^{K} \sum_{q=0}^{K} \sum_{c=1}^{C} W_{p,q,c} X_{(i+pd),(j+qd),c}, \tag{3.4}$$

where Z is the output feature map, and (i, j) indexes the location in image X. Since this convolution can arbitrarily enlarge the receptive field by tuning the dilation ratio d, it can be used to replace the encoder-decoder structure to extract contextual information [118, 212]. This semantic information extraction procedure can deliver two merits to the corresponding network: a) Because no resolution is lost in the information processing procedure, small and thin objects that can be important for correctly understanding the image are finely preserved. b) Since no downsampling operation is included, the location information of the generated feature maps can be better conserved.

The building block in these pathways is a residual dilated convolutional block [118]. As shown in Fig. 3.15 (*i.e.*, the orange squares), it is constructed by two convolution blocks and a shortcut connection. The benefit of this block is two-fold: a) It improves the training speed and encourages smooth information flow [69]; b) Combining with the dilated convolutions, skip connections implicitly exploit and fuse information from different scales.

Moreover, to further improve the longterm information flow which is weak in the classic dilated residual network [187], I combine dense connection to allow the information from the early stage of the high-resolution pathway to be directly passed to the final layer of the module. This setting also leads to an even finer grain multi-scale information collection of the whole network. After that, to reduce the training difficulties and also to make the pathway discriminative to the true organ (or tissue) boundaries, a deep supervision mechanism is introduced. In the experiments, nine residual dilated convolutional blocks compose the pathway. The first three blocks are with the dilation of 1, the second three with 3, and the last three with 5.

3.2.3 Contour Information Integration

In recent studies, neuroscientists have investigated that, in mammal visual system, contour delineation correlates with object segmentation closely [107]. To incorporate these insights to improve the segmentation accuracy, researchers integrate the task of contour detection with the task of segmentation. The advantage of this design is three-fold. a) It provides extra robust guidance to the task of segmentation. b) It improves the generalization capacity of the corresponding network. c) Introducing a task of contour regression can help guide the network to concentrate more on the boundary of organ regions, thus helping overcome the adverse effect of low tissue contrast. In this study, as shown in Fig. 3.15, a regression task is added to the end of the network as auxiliary guidance. In the existing studies [31, 203], thanks to the high image contrast, the boundaries are usually clear and stable. As a result, authors in these studies [31, 203] modeled the contour detection as a binary classification problem. However, in the proposed application, due to the blurry nature of images, the voxels near the boundaries are usually highly similar. As a result, it will be more reasonable to model the boundary delineation task as a regression problem, which estimates the probability of each voxel being on the organ boundary.

To extract the contour for training, the boundaries of different organs were first delineated by performing Canny detector [23] on the ground-truth segmentation. Then, on this boundary map, a Gaussian filter was further exerted with a bandwidth of δ, which was empirically set as 2 in the experiments. For other datasets, the setting in landmark heat map generation [218] could be followed (*i.e.*, set δ from 2 to 3 for good performance). An approximated probability map (denoted as $\hat{p}$ which corresponds to the contour map shown in Fig. 3.15) was generated to describe the certainty of each voxel being on the boundary of an organ. Hence, the regression target was to minimize an Euclidean loss function as defined below:

$$L_{rBoundary} = \sum\nolimits_{h} \sum\nolimits_{w} \left| p_{h,w} - \hat{p}_{h,w} \right|^2, \tag{3.5}$$

where $\hat{p}_{h,w}$ indicates the probability of being boundary at location (h, w).

3.2.4 Difficulty-Guided Cross-Entropy Loss

To balance the frequency of the voxels from different classes, categorical cross-entropy loss is a common choice for multi-class segmentation [153, 31]. Different from the original cross-entropy loss, the categorical version adds a loss weight w_k for the voxels in the k^{th} category as shown in Eq. 3.1. This weight is inversely related to the portion of voxels belonging to the k^{th} category as shown in Eq. 3.3.

In a recent work, Li *et al.* [119] argued that not all voxels are equal and more attention should be paid to the difficult voxels. Inspired by this argument, a difficulty-guided weight map was proposed to guide the network and focus more on the ambiguous areas. It is evident that the error of existing networks mainly lies around the borders of both foregrounds and backgrounds. It becomes even larger at the touching boundary of soft tissues. With these observations, the weight map was constructed in three steps. a) the Canny operator was used to calculate the binary boundary image B_c of the category (*i.e.*, organ) c, according to the ground-truth segmentation. b) a Gaussian filter with bandwidth

δ_2 was used to scan each B_c and get the smoothed boundary image SB_c. c) Finally, all SB_c were summed up and then normalized to construct the final weight map. As a result, the proposed difficulty-guided weight on voxel v was defined as Eq. 3.6.

$$u^v = u_0 + \sum_{c}^{C} u_c \cdot SB_c^v, \tag{3.6}$$

where u_0 is the base weight for all the voxels and u_c is the importance balancing weight of category c, similar to what is used in Eq. 3.1. In the experiments, these hyper-parameters were set as $u_0 = 1$, and $u_1 = u_2 = u_3 = 25$ which worked as the ratio of the volume of background to the volume of foreground for prostate, bladder and rectum, respectively. The same strategy is effective for other datasets. The bandwidth δ_2 of the Gaussian filter was set as 8 to achieve a good coverage of the ambiguous boundary regions in all the experiments. In the designed map, the regions of the foreground that were far away from the boundary were treated equally with those from the background. Also, since the area emphasized by different maps could overlap around the touching border, these areas were automatically endowed with the most concentration. Replacing the categorical weight map in Eq. 3.1 with the proposed difficulty-guided weight map, the loss function was accordingly proposed in Eq. 3.7 for segmentation, which is an improved version compared to the category-based importance-aware loss (Eq. 3.1).

$$L_{diffCE}\left(X, Y; \theta_S\right) = \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{c=1}^{C} I\left\{Y_{h,w}, c\right\} V_{h,w} \log \hat{P}_{h,w,c}. \tag{3.7}$$

Combining the loss for segmentation and contour regression, the final loss function for network optimization is presented in 3.8.

$$L_{hrTotalLoss} = L_{diffCE} + \alpha L_{rBoundary}, \tag{3.8}$$

where α is hyper-parameter used to balance the importance between the terms. In the experiments, setting $\alpha = 1$ and making the normalized loss functions of segmentation and regression to be in a comparable magnitude provides preferable results.

3.2.5 Experimental Results

In this section, the effectiveness of the proposed algorithm was showcased on a pelvic CT image dataset, *i.e.*, pelvic organ segmentation introduced in Sec. 1.1.4. Specially, for the pelvic CT image dataset, considering the large size of pelvic organs, large receptive field on the axial direction was beneficial for accurate segmentation. For computational efficiency, the problem was modeled as a 2D semantic segmentation problem. In the first part of the experiment, careful ablation studies were conducted to verify the effectiveness of each component of the designed network, especially the high-resolution pathway. Then, the proposed method was compared with state-of-the-art methods.

The evaluation of the proposed method on pelvic CT image dataset starts by comparing the performance of dilated convolutional networks with their encoder-decoder counterparts. Then, the high-resolution pathway is introduced to the encoder-decoder network and the effectiveness on detecting blurry and vanishing boundaries was tested. Next, the effectiveness of the difficulty-guided cross-entropy loss function and the multi-task learning mechanism are further investigated. After that, the sensitivity of the main hyper-parameter is analyzed in the proposed algorithm. Finally, the proposed algorithm is compared with several state-of-the-art medical image segmentation methods.

3.2.5.1 Data Description and Implementation Details

The dataset used in this experiment is acquired by the North Carolina Cancer Hospital, which includes 339 CT scans from prostate cancer patients. In this task, three important pelvic organs, *i.e.*, prostate, bladder, and rectum are being segmented. All the images are normalized using their z-scores. As a result of this normalization, the normalized data will follow a normal distribution (mean 0 and standard deviation of 1). Before experiments, a simple UNet [164] is first run to extract ROIs for all the compared algorithms, as a rough initial localization. In the experiment, the network patch size is set to $144 \times 208 \times 5$. In each of the extracted patches, five consecutive slices across the axial plane are included as five different channels to introduce space information across slices and to preserve across-slice consistency in the axial direction. In the sampling procedure, the axial slices are permuted upside-down to double the number of samples for data augmentation. The data is randomly divided into the training, validation and testing sets with 180, 59 and 100 samples, respectively.

The implementations of all the compared algorithms in this part are based on the Caffe platform [92]. To train the network, Xavier method [59] is used to initialize all the parameters of convolutional layers in the compared networks. To make a fair comparison, the Adam optimization method [100] is employed for all the methods with fixed hyper-parameters. The learning rate (lr) is set to 0.001, and the step size hyper-parameter β_1 is 0.9 and β_2 equal to 0.999 in all cases. The batch size of all compared methods is 10. The models were trained for at least $200,000$ iterations until a plateau or overfitting tendency is observed according to the loss on the validation set. To evaluate the effectiveness of the proposed method extensively, the DSC (Eq. 2.22) and ASD (Eq. 2.24) are reported.

3.2.5.2 Evaluation of Dilated Convolutional Networks

First, the performance of the high-resolution dilated convolutional networks on CT pelvic organ segmentation is evaluated. To conduct such an evaluation, five baseline networks is designed as comparison methods with the proposed method. Among the compared networks, the first three are dilated convolutional networks (see Fig. 3.16 for an overview of their architecture). Their differences mainly lie in the number of residual dilated convolutional blocks (refer to Fig. 3.15 for the definition) and the dilation factors (d_1 and d_2). The first three networks are named as DilNet1, DilNet2, and DilNet3 for simplicity, respectively. Specifically, DilNet1 and DilNet2 both consist of 9 residual dilated convolutional blocks. Their dilation factors d_1 and d_2 are 3 and 5 for DilNet1, and 2, 4 for DilNet2. DilNet3 has six blocks (without three blocks within the black dotted rectangular in Fig. 3.16). Its dilation ratios d_1 and d_2 are 3 and 5, respectively. The receptive fields of these three networks are 133 × 133, 97 × 97 and 85 × 85, respectively, which are nearly in the receptive filed range of UNets [164] with 3 to 4 pooling layers. The fourth and the fifth networks are the distilling networks with four and three pooling layers, respectively. They are designed as representers for encoder-decoder networks, named as Dst-Net1 (Distilling Network 1) and Dst-Net2 (Distilling Network 2), respectively.

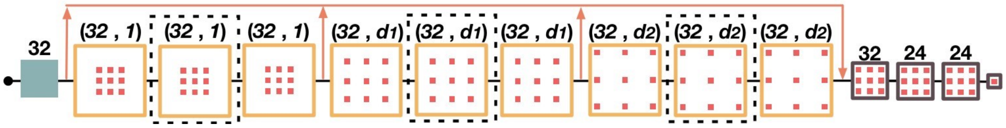


Figure 3.16: Illustration of the dilated convolutional network.

All the networks are trained in the same manner as mentioned in Sec. 3.2, with the corresponding DSC and memory cost listed in Table 3.6. Observing the experimental results, two conclusions could be made. a) Larger receptive fields and deeper network structures are essential for the performance of both dilated convolutional networks and encoder-decoder networks. b) The encoder-decoder networks in the experiments tend to provide better performance with smaller memory consumption than the compared dilated networks in CT pelvic organ segmentation. The reasons for its better result are two-fold. First, the relative plain connection and the smaller number of kernels limit the performance of the dilated convolutional network. Moreover, without the help of the downsampling operation, dilated convolutional networks are more likely to be adversely affected by the noise in CT images.

3.2.5.3 Evaluating the Effectiveness of Integrating High-Resolution Pathway

Although in the last experiment, dilated networks have shown relatively inferior performance than their encoder-decoder competitors, the capacity of providing high-resolution semantic information makes them potentially more suitable than the coarse-grained encoder-decoder networks on accurately localizing the blurry target boundaries, thus improving the segmentation performance. Here, to reveal the limitation of current encoder-decoder networks and show the effectiveness of introducing high-resolution pathways for solving the corresponding problems, two networks are constructed and compared. The baseline algorithm is the distilling network (*i.e.*, Dst-Net1) introduced in the last section. In the compared network, a high-resolution pathway is added to connect the encoder and decoder at the highest resolution in Dst-Net1, named as high-resolution distilling network (HRDN). The results are listed in Table 3.8. From the results, an approximate 1% improvement in terms of DSC can be observed on the two smaller and also more difficult organs, *i.e.*, prostate and rectum. The improvement of ASD on the high-resolution pathway enhanced network is also promising, with 0.143 on the prostate and 0.145 on the rectum. The results numerically verify the effectiveness of the high-resolution pathway.

Table 3.6: Dice ratio (%) and memory cost (Mb) comparison between dilated convolutional networks and encoder-decoder networks.

Networks	Bladder	Prostate	Rectum	Memory Cost
DilNet3	88.4	82.2	81.0	7259
DilNet2	88.5	83.4	81.9	9269
DilNet1	89.6	83.5	83.7	9269
DstNet2	92.2	85.4	85.0	5443
DstNet1	86.2	93.1	84.9	5933

Table 3.7: Result comparison between distilling network (DstNet1) and high-resolution distilling network (HRDN).

Networks	DSC(%)			ASD(mm)		
	Bladder	Prostate	Rectum	Bladder	Prostate	Rectum
DstNet1	93.1(4.5)	86.2(4.0)	84.9(5.2)	1.334(0.858)	1.585(0.437)	1.543(0.493)
HRDN	93.2(5.5)	87.5(3.8)	85.9(5.3)	1.542(2.278)	1.434(0.425)	1.395(0.617)

To further exploit the characteristics of the three kinds of basis pathways, *i.e.*, skip connection, distilling pathway and high-resolution pathway, and to intuitively reveal what limitations of the encoder-decoder network have been resolved by the high-resolution pathway, some of the salient feature maps generated by the two networks are visualized on a representative sample.

First, the information conserved by the skip connection and the distilling pathway is illustrated in Dst-Net1 and that by the high-resolution pathway in HRDN. The exact locations of where the information is collected in the corresponding networks are also marked as pathway 1 and 2, and 3 consecutively in Fig. 3.15. Three representative feature maps with high activation values on the target organs, *i.e.*, bladder, prostate, and rectum, are illustrated and compared in Fig. 3.17. In this selected sample, as pointed out by the white arrow in the intensity map, some wavy streaks appear on the three target organs and affect the boundary on the top of the prostate due to the effects of artifacts in the CT image which results in generating a small visually isolated tissue. Under such circumstance, as can be seen in the activation maps passed by the skip connection (see the first row of Fig. 3.17), although the skeletons of the organs look more evident since the surrounding small fractions of tissues are filtered, the less obvious but essential texture information is either weakened (*e.g.*, shown in the first and third sub-figures) or strengthened (*e.g.*, shown in the second sub-figure) indistinguishably. As a consequence, with the falsely included tiny texture, the isolated part looks more like a portion of bladder than prostate. Moreover, as little semantic information is contained in this pathway, no organ-specific information is incorporated, leaving the coarse-grained encoder-decoder pathway to select the correct boundary within all these closely located boundary candidates. Considering the feature maps generated by the distilling pathway (the second row of Fig. 3.17), although the maps are more semantically meaningful, the boundaries of these maps, especially those on the border between bladder and prostate, are inaccurate, since the down-sampling operations can undermine the accuracy of location information.

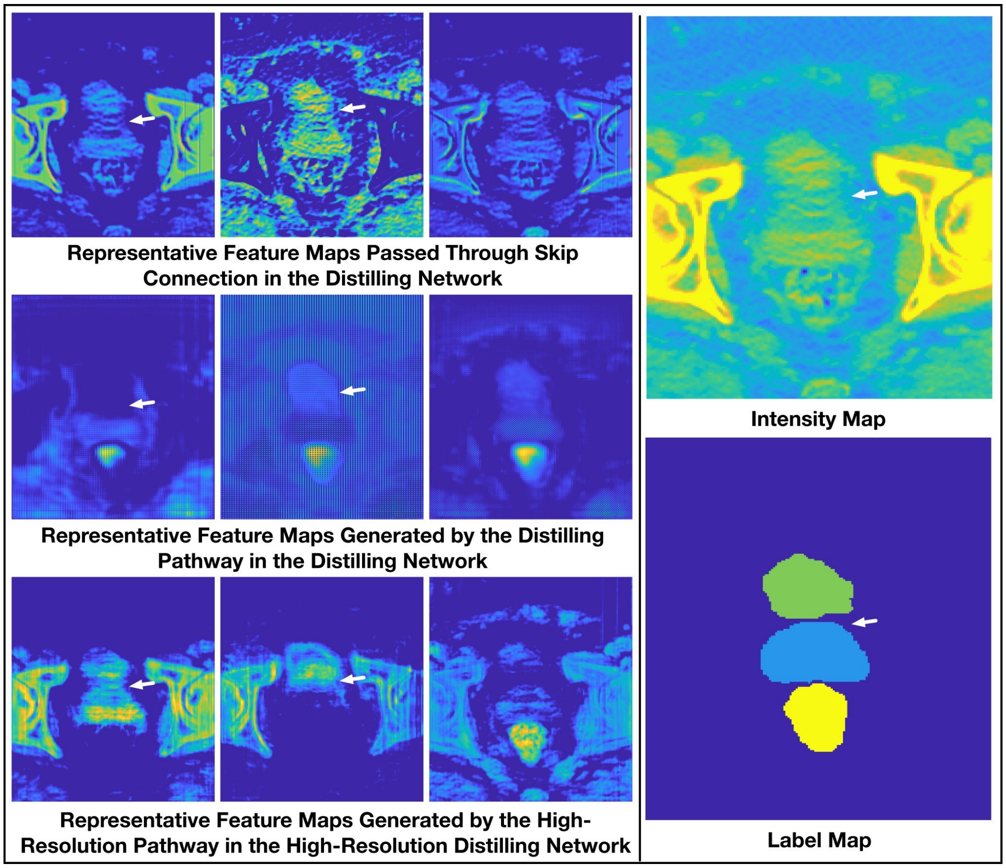

Representative Feature Maps Passed Through Skip Connection in the Distilling Network
Representative Feature Maps Generated by the Distilling Pathway in the Distilling Network
Representative Feature Maps Generated by the High-Resolution Pathway in the High-Resolution Distilling Network
Intensity Map
Label Map

Figure 3.17: Comparison of representative feature maps.

In contrast, since high-resolution semantic information is preserved, the feature maps generated by the high-resolution pathway is more like a combination of the above-mentioned two kinds of feature maps. They contain detailed textural information and yet more semantics. Besides, thanks to the integrated deep supervision mechanism, the hypothetical boundaries are finely weakened or neglected (see the first and second sub-figures of the third row in Fig. 3.17), making the boundaries in the ambiguous area clear and correct.

Similar with the intermediate activation maps, as can be seen in the final output feature maps and the corresponding prediction maps of the two networks (Fig. 3.18), due to the falsely located boundary, a large portion at the bottom of the bladder and the top of the prostate is mixed in the distilling network. Comparatively, thanks to the high-resolution pathway, the damaged boundaries are handled more appropriately in HRDN, resulting in a more feasible segmentation.

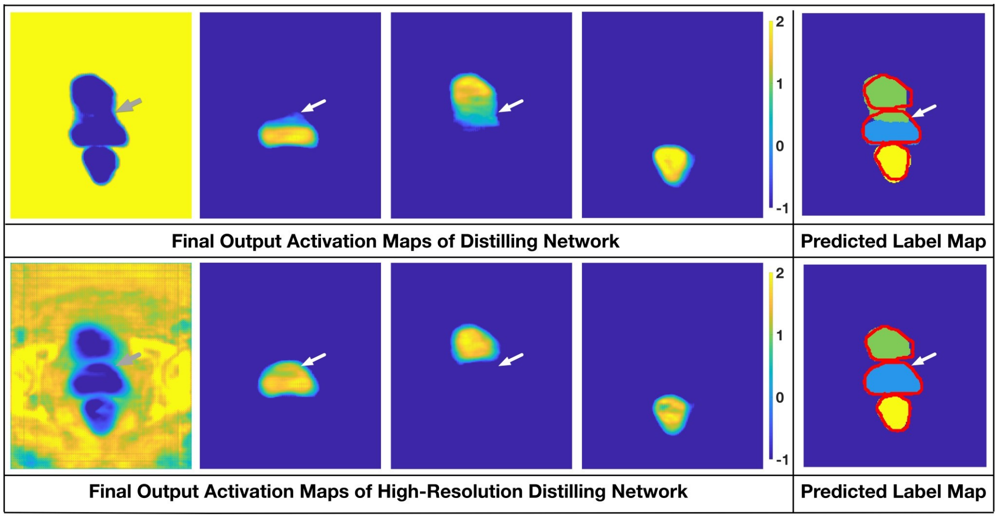

2
1
0
-1
Final Output Activation Maps of Distilling Network
Predicted Label Map
2
1
0
-1
Final Output Activation Maps of High-Resolution Distilling Network
Predicted Label Map

Figure 3.18: Comparison of representative feature maps.

The numerical and qualitative results in this section support the following arguments: a) Simple skip connections can be insufficient to detect the blurry or vanishing boundaries in pelvic CT image segmentation; b) The downsampling and upsampling operations of the encoder-decoder networks pose potential risks of inaccurate boundary localization and mis-detecting isolated portions of the target; c) By carefully combining the advantage of the dense connection, residual connection, dilated convolution and deep supervision, the high-resolution pathway can well remedy the limitation of the encoder-decoder network.

3.2.5.4 Balance Between Resolution and Network Complexity

Although the effectiveness of introducing high-resolution pathway has been validated above, the memory cost of injecting such a branch is huge due to the dilated convolution operations in the large-size feature maps. Adding such a pathway in the intermediate stages of the network is also a possible way to improve the performance of the network with smaller memory cost because it also allows us to use more complex network structure. To explore the balance between the network complexity and the resolution of the semantic feature maps, four networks were further designed. In these four networks, the high-resolution pathway is placed on the first to the fourth stage of the network, respectively. Here, the first stage indicates the feature extracting stage with no downsampling, the second stage indicates the stage with one downsampling, and so on. The feature number L of the high-resolution pathways are $32, 40, 56, 72$, respectively.

Table 3.8: Testing the balance between the network complexity and the resolution of the high-resolution pathway. The boldface results indicate significant difference from the best result (p-value $<$ 0.05 of Students t-test).

Networks		DSC(%)			ASD(mm)	
	Bladder	Prostate	Rectum	Bladder	Prostate	Rectum
HRDN-L1	**93.2(5.5)**	**87.5(3.8)**	**85.9(5.3)**	1.542(2.278)	**1.434(0.425)**	1.395(0.617)
HRDN-L2	**93.6(4.7)**	**87.5(3.9)**	**86.1(5.4)**	1.399(1.600)	**1.438(0.404)**	**1.422(0.587)**
HRDN-L3	94.0(4.3)	87.9(3.9)	86.8(5.1)	1.282(1.275)	1.427(0.483)	1.397(0.673)
HRDN-L4	**93.6(4.7)**	**87.4(4.2)**	**86.0(6.0)**	**1.362(1.810)**	1.532(0.408)	**1.488(0.745)**

In Table 3.8, HRDN-L1 to HRDN-L4 denote the HRDNs with high-resolution pathway on the first to the fourth stage, respectively. One can see that tuning the location of the high-resolution pathway does improve the performance of the network, especially on improving the overall segmentation accuracy in terms of Dice score. However, for the pelvic CT image dataset, placing the high-resolution to the third stage provides the best balance between feature resolution and network complexity.

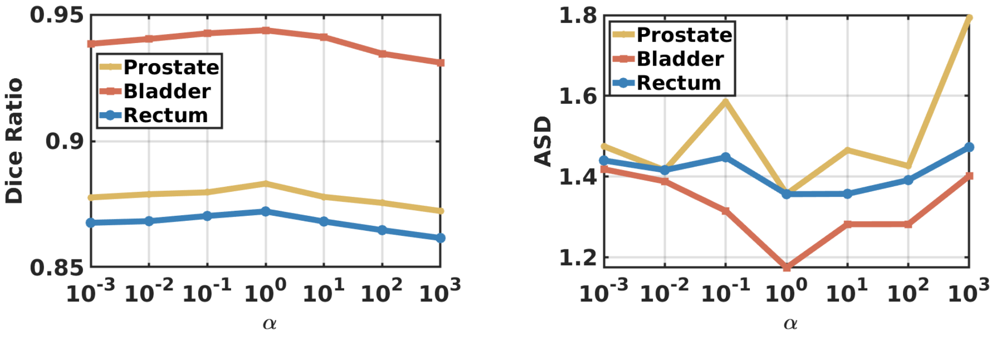
3.2.5.5 Evaluation of Difficulty-Guided Loss Function and Multi-task Learning Mechanism

To evaluate the effectiveness of the difficulty-guided loss function and the multitask learning mechanism, two networks, including a baseline High-Resolution Distilling Network (HRDN), and a multi-task HRDN with difficulty-guided cross-entropy loss (HMEDN), are designed and tested. The numerical results of these two networks are reported in Table 3.9. Since the introduced mechanism is mainly proposed to improve the performance on boundary localization, an extra metric, *i.e.*, the Hausdorff Distance [85] (Eq. 2.23), which measures the largest distance between two segmentation contours are introduced. As shown in the table, all three metrics, *i.e.*, DSC, ASD, and Hausdorff distance witnessed a stable improvement on all the three organs. Especially on ASD and the Hausdorff distance, which can be easily influenced by the inaccurately located boundaries, the average surface distance of the three organs has been improved by approximately 4%, and 15% on average, respectively.

Table 3.9: Comparison between the high-resolution distilling network (HRDN) and the multi-task HRDN with difficulty-guided cross entropy loss (HMEDN). The boldface results indicate significant difference from the best result (p-value < 0.05 in Student's t-test).

Networks	DSC(%)			ASD(mm)			HD(mm)		
	Bladder	Prostate	Rectum	Bladder	Prostate	Rectum	Bladder	Prostate	Rectum
HRDN-L1	87.9(3.9)	**94.0(4.3)**	**86.8(5.1)**	**1.427(0.483)**	**1.282(1.275)**	1.397(0.673)	**17.2(21.6)**	**21.6(21.0)**	**20.5(17.0)**
HRDN-L2	88.3(4.3)	94.4(4.2)	87.2(5.5)	1.357(0.532)	1.175(1.197)	1.357(0.796)	15.3(20.9)	17.5(16.8)	17.2(11.1)

The sensitivity of the hyper-parameter α is also investigated, which balances the importance between the tasks of segmentation and boundary regression. In this experiment, α is tuned in the range of $\{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$ to train the corresponding networks (Note, they are trained in the same manner except the α). In Fig. 3.19, it can be observed that the performance of the proposed algorithm is quite stable in a broad range of the hyper-parameter α. This reflects the tight correlation between the integrated tasks as well as the stability of the proposed algorithm. The best result is achieved when $\alpha = 1$ and the same magnitudes are adopted for the normalized segmentation and boundary regression losses.



(a) Dice variation against α.

(b) ASD variation against α

Figure 3.19: Sensitivity analysis of the proposed network on the multi-task importance balance hyper-parameter α.

3.2.5.6 Comparing with the State-of-the-art Methods

For further evaluation, the proposed network is compared with several state-of-the-art methods for medical image segmentation. These methods include:

1. UNet: UNet [164] is the pioneering work that introduces fully convolutional neural network [124] for medical image analysis. This network achieved the best performance on ISBI 2012 EM challenge dataset [4].

2. FCN: Fully convolutional neural network [124] is the first trial that allows the network directly output a segmentation mask having the same dimension of the input image. The method achieved the state-of-the-art performance on multiple popular benchmark datasets, like PASCAL VOC [51] in 2015[5].

3. DCAN: Deep contour-aware neural network [31] has won the 1st prize in 2015 MICCAI Grand Segmentation Challenge[6] and 2015 MICCAI Nuclei Segmentation Challenge[7].

4. DenseSeg: Densely convolutional segmentation neural network [20] introduces dense connections into the HED network to ensure maximum information flow. This method has won the first prize in the 2017 MICCAI grand challenge on 6-month infant brain MRI segmentation[8].

5. Proposed: the proposed high-resolution multi-scale encoder-decoder network (HMEDN) is a novel encoder-decoder network enhanced by multi-scale dense connections, high-resolution pathways, difficulty-guided cross-entropy loss function and multi-task learning mechanism.

[5] https://github.com/shelhamer/fcn.berkeleyvision.org

[6] https://www2.warwick.ac.uk/fac/sci/dcs/research/tia/glascontest

[7] http://miccai.cloudapp.net:8000/competitions/37

[8] http://iseg2017.web.unc.edu

Table 3.10 shows the segmentation results of the compared state-of-the-art methods. As can be seen, all the results of the compared algorithms are reasonably well. However, the proposed algorithm still outperforms the second best performance of the state-of-the-art methods by about 1.5% in terms of DSC and more than 10% in the average surface distance. From Fig. 3.20, it can be seen that the proposed algorithm tends to not only achieve more accurate segmentation on those easy subjects but also provide more robust results on difficult subjects. More specifically, through the visualization of the segmentation results on two representative samples in Fig. 3.20, it can be observed that the advantage of the proposed method mainly lies in two perspectives: a) It can localize the boundary better, especially on those blurry areas; b) It can better handle the CT artifacts. It is worth noting that hence no deep supervision was involved in DenseSeg [20] and FCN [124] (while DCAN [31] has the deep supervision module, as the proposed algorithm). Therefore, the performance of these two algorithms can be further improved with the deep supervision mechanism.

Table 3.10: DSC and ASD comparison with the state-of-the-art methods on pelvic CT image dataset. Note all results from comparison methods exist significant difference from the best result from the proposed method (p-value < 0.05 in Student's t-test).

Networks		DSC(%)			ASD(mm)	
	Bladder	Prostate	Rectum	Bladder	Prostate	Rectum
UNet	91.7(5.9)	86.1(5.2)	85.5(5.1)	1.773(1.851)	1.532(0.487)	1.470(0.535)
FCN	92.9(5.4)	86.2(4.6)	85.5(5.6)	1.588(2.267)	1.591(0.546)	1.479(0.587)
DCAN	92.5(7.0)	86.5(3.8)	85.2(5.5)	1.367(1.302)	1.580(0.534)	1.530(0.760)
DenseSeg	92.7(7.1)	86.8(4.3)	84.8(5.8)	1.554(0.555)	1.724(2.591)	1.851(1.132)
Proposed	94.4(4.2)	88.3(4.3)	87.2(5.5)	1.175(1.197)	1.357(0.532)	1.357(0.796)

3.3 Semantic-guided Feature Learning for Blurry Boundary Delineation

To efficiently delineate the blurry boundaries as introduced in Sec. 1.1.4, I describe a novel semantic-guided encoder feature learning strategy for encoder-decoder networks in this section. The architecture of the proposed framework is presented in Fig. 3.21, in

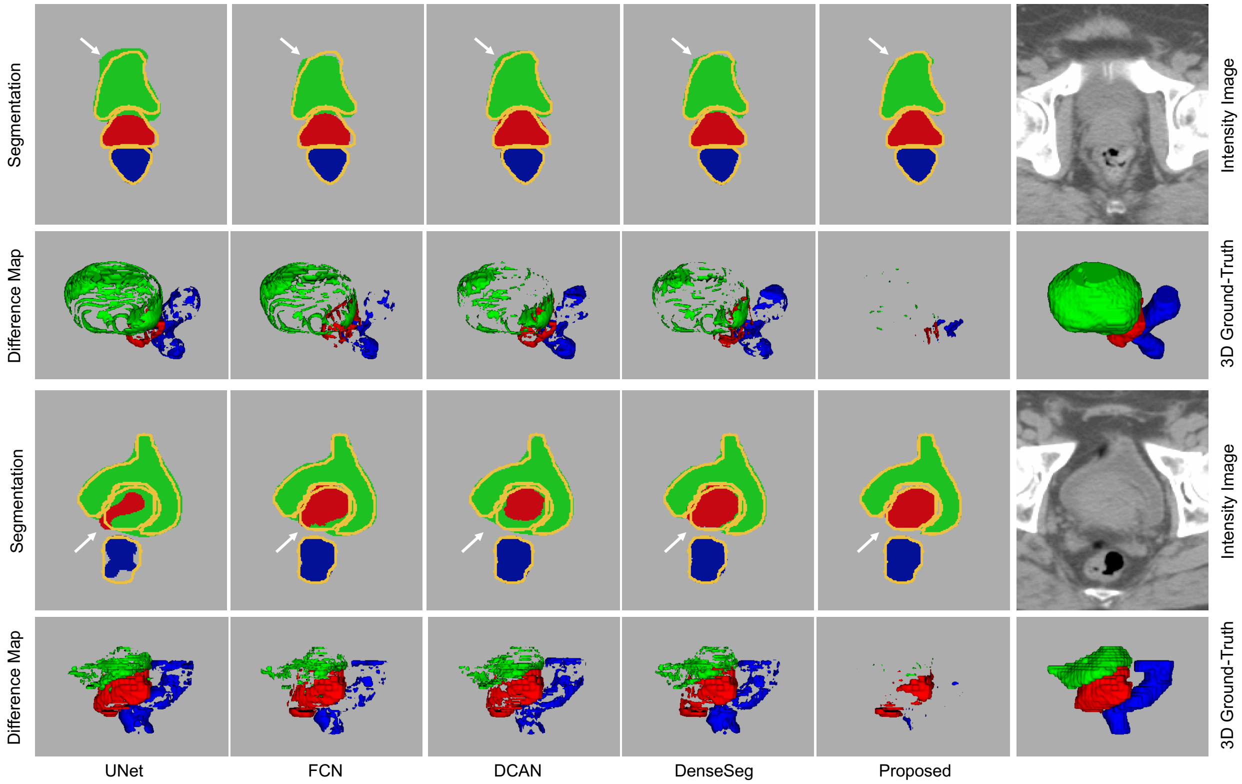


Figure 3.20: Segmentation results of the proposed algorithm and the compared state-of-the-art algorithms of two representative samples on the pelvic CT image dataset. In the first and the third rows, the segmentation masks and intensity images in the axial direction are provided. The yellow curves in the segmentation masks indicate the ground-truth contour of the target organs. The second and the fourth rows are the difference map and the segmentation ground-truth in 3D space. The green, red, and blue fragments are false predictions on prostate, bladder, and rectum, respectively.

which an encoder-decoder architecture is introduced with three tasks (segmentation, clear boundary detection, and blurry boundary detection). The proposed semantic-guided encoder feature learning module (SGM) is further highlighted in Fig. 3.23.

In the following subsections, I analyze the deficiency of the skip connection in the current encoder-decoder framework. Then, I introduce the proposed semantic-guided encoder feature learning strategy. Moreover, I describe the soft contour constraint for blurry boundary delineation. The implementation details are followed up. Finally, The comparison experiments and ablation study are introduced.

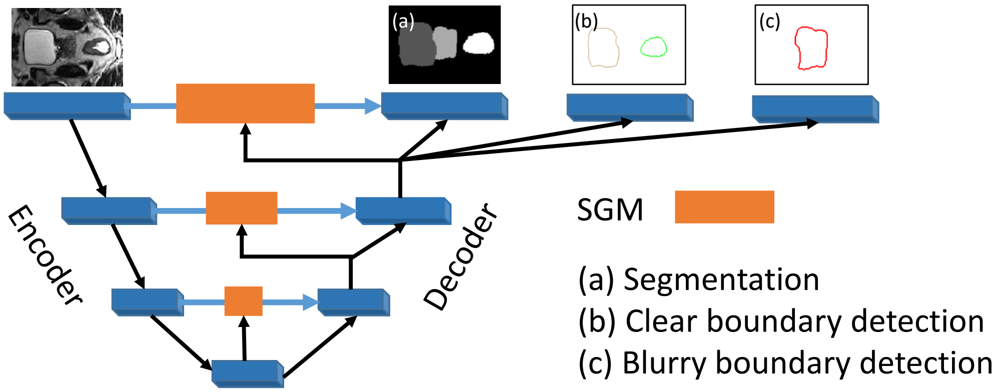


Figure 3.21: Illustration of the architecture of the proposed method, which consists of a semantic-guided module (SGM). (a) means a segmentation branch, and (b) and (c) indicate boundary detection branches.

3.3.1 Analysis of Skip Connection in Encoder-Decoder Architecture

In the classical encoder-decoder architecture [164], shallow and deep features are usually complementary to each other. For example, shallow features are rich in resolution but insufficient in semantic information, while deep features are semantically meaningful but lack of spatial details. The skip connection proposed in UNet [164] is supposed to provide high-resolution information from the shallow (encoder) layers to the deep (decoder) layers, so that the localization precision can be improved without losing classification accuracy. However, the raw (simple) skip connection has several drawbacks. a) It would bring 'noise' (unnecessary information) to the deep layers which will definitely affect the concatenation of feature maps, as shown in the visualized encoder feature maps in Fig. 3.22. b) The huge gap between shallow and deep features will decrease the power of this combination. c) Moreover, for the clear boundaries (*e.g.*, bladder and rectum), the encoder feature maps can provide sufficiently precise localization information as shown in Fig. 3.22, which can thus work well with the raw skip connection. However, the blurry boundary (*e.g.*, prostate) cannot be well described in the encoder feature maps as shown in Fig. 3.22, which thus cannot provide accurate localization information with simple skip connection. Therefore, it is highly desired to select discriminative features, not simply inhibiting indiscriminative features from shallow layers; in other words, learning discriminative high-resolution semantic features from the encoder could potentially solve this blurry boundary delineation problem. To achieve such an effect, Roy *et al.* [166] proposed concurrent spatial-and-channel squeeze and excitation module to boost meaningful features and suppress weak ones. Oktay *et al.* [155] proposed gated attention mechanism to select the salient part of the feature maps to further improve the UNet. However, in both works, the feature learning process is actually conducted in an *implicit* manner which limits the learning efficiency.

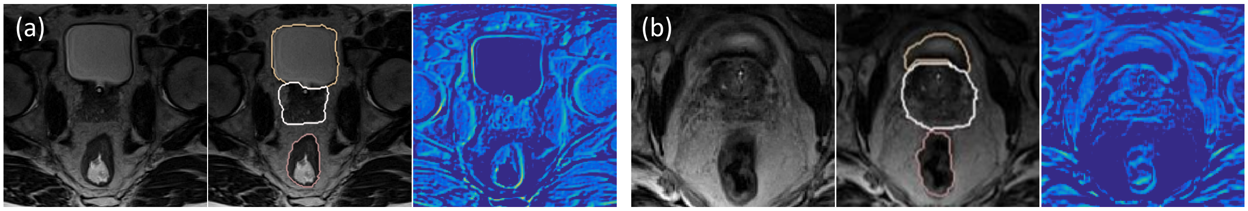


Figure 3.22: Illustration of the blurry and vanishing boundaries within pelvic MRI images, together with overlaid ground truth contour and the typical feature maps in the encoder layer of a conventional UNet. (a) and (b) are the two typical slices of two subjects, in which boundaries of bladder and rectum are relatively clear, but prostate is blurry.

3.3.2 Semantic-guided Encoder Feature Learning

To overcome the above mentioned problems, I propose to *explicitly* learn the high resolution semantic features (which are also more discriminative) from shallow (encoder) layers with semantic guidance from deep (decoder) layers. The key idea is to encode semantic concept from deep-layer features to guide the learning of shallow-layer features. As shown in Fig. 3.23, my semantic-guided feature learning module (*i.e.*, SG module or SGM) is designed to selectively enhance or suppress the features of shallow layer at each stage so that I can enhance the consistency between shallow and deep layers without losing resolution information. Besides the widely-used channel-wise encoder, I have also designed the spatial-wise encoder as described below.

I consider the feature maps of a certain encoder layer (*i.e.*, shallow features) to be $S = \{s_1, s_2, ..., s_K\}$, where $s_i \in R^{H \times W \times T}$. I also assume the up-sampled feature maps in the

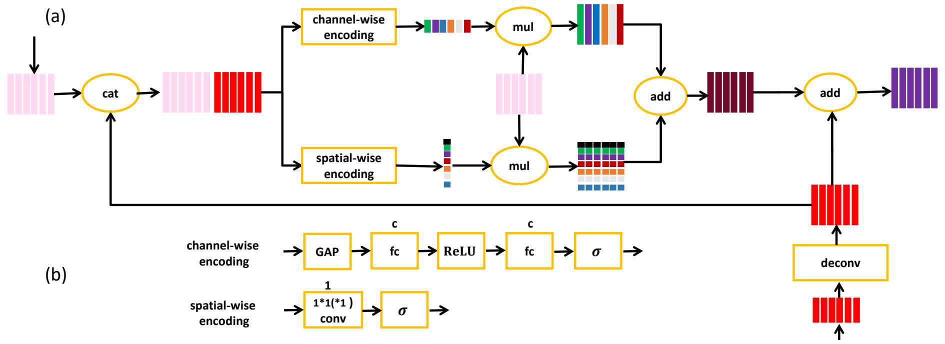


Figure 3.23: Illustration of the proposed semantic-guided module (SGM), as shown in (a). The pink blocks represent the features of shallow layers, while the red ones represent the features of deep layers. Different from direct skip connection in UNet, I propose using semantic concept from deep layers to guide feature learning in the corresponding shallow layers, for which a channel-wise encoder and a spatial-wise encoder are both proposed, as shown in (b). 'GAP' means Global Average Pooling.

corresponding decoder layer (deep features) to be $D = \{d_1, d_2, ..., d_K\}$, where $d_i \in R^{H \times W \times T}$. I concatenate the two group of feature maps together and thus result in a bank of *high-resolution and rich-semantic mixed* feature maps as shown in Eq. 3.9.

$$F = \{s_1, s_2, ..., s_K, d_1, d_2, ..., d_K\}. \tag{3.9}$$

3.3.2.1 Channel-wise Encoding

With a global average pooling layer, I obtain a vector $Q = \{q_1, q_2, ..., q_K, ...q_{2K}\}$, where q_k is a scalar and corresponds to the averaging value of the k-th feature maps in F. Then, two successive fully connected layer are adopted to fuse the resolution and semantic information: $Z = W_1 \left(\mathrm{ReLU}\left(W_2 Q\right)\right)$, with $W_1 \in R^{K \times K}$ and $W_2 \in R^{2K \times K}$. This encodes the channel-wise dependencies by considering both shallow and deep features. I apply a sigmoid activation function to map the neurons to probabilities so that I can formulate as a channel-wise importance descriptor, which can be described as $\sigma\left(Z\right)$. Thus, the semantic-guided channel-wise encoded feature maps are formulated as Eq. 3.10.

$$SGCF = \{\sigma\left(z_1\right) s_1, \sigma\left(z_2\right) s_2, ..., \sigma\left(z_K\right) s_K\}. \tag{3.10}$$

Note that the weight $\sigma\left(z_k\right)$ before the shallow feature map s_k can be viewed as an indicator of how important this specific feature map is. Thus, I argue this channel-wise encoding is actually a semantic-guided feature selection process in a channel-wise manner, which is able to ignore less meaningful feature maps and emphasize the more meaningful ones. In other words, it can help remove the 'noise' and retain the useful information. More importantly, since $\sigma\left(Z\right)$ has taken both high resolution and rich semantic information into account, it has more discriminative capacity than the case of only considering shallow layer information in [166].

3.3.2.2 Spatial-wise Encoding

Now I come to consider the spatial-wise importance to achieve better fine-grained image segmentation.

Based on the concatenated feature maps F, I apply a $2K \times 1 \times 1 \times 1$ convolution to squeeze the channels. Therefore, I can obtain a one-channel output feature map U, where $U \in R^{H \times W \times T}$. I directly apply sigmoid function to acquire a probability map for U. Similarly, the semantic-guided spatial-wise encoded shallow feature maps can be described in Eq. 3.11.

$$SGSF = \left\{\sigma\left(U\right) \otimes s_1, \sigma\left(U\right) \otimes s_2, ..., \sigma\left(U\right) \otimes s_K\right\}. \tag{3.11}$$

Since $\sigma\left(U_{h,w,t}\right)$ corresponds to the relative importance of a spatial information at (h, w, t) of a given shallow layer feature map, it can help select more important features to relevant spatial locations and also ignore the irrelevant ones. Moreover, $\sigma\left(U\right)$ is a fusion of both resolution and rich semantic information, thus it can provide a better localization capacity even for the blurry boundary regions which cannot done by [166]. As a result, I view this spatial-wise encoding as a semantic-guided recalibration process.

3.3.2.3 Combination of Encoded Feature Maps

Now I can formulate both channel-wise and spatial-wise encoding by a simple element-wise addition operation, as shown in Eq. 3.12.

$$SGF = \mathrm{SGCF} + \mathrm{SGSF}. \tag{3.12}$$

This SGF considers both channel-wise encoded and spatial-wise encoded information, thus, it contains *not only* the discriminative (semantic) features, *but also* more accurate localization information.

3.3.2.4 Final Combination with Deep-Layer Feature Maps

To this end, I can simply complete the concatenation operation or element-wise addition operation. Instead of using the shallow feature maps S, I use the channel-wise and spatial-wise encoded shallow feature maps SGF to combine with the deep-layer feature maps D (through concatenation or element-wise addition). Compared with the raw skip connection in UNet, the encoded shallow feature maps SGF has same resolution but much more semantic and precise localization information (especially for the blurry regions), and thus can make the combination more reasonable. At the same time, since the operations in the encoder are mostly $1 \times 1 \times 1$ convolution, the number of parameters just increases a little bit.

To further increase the model's discriminative capacity, I also adopt the multi-scale deep supervision strategy as in [214] after feature fusion at each stage.

3.3.3 Boundary Delineation with Soft Contour Constraint

In mammal visual system [107], contour delineation closely correlates with object segmentation. To incorporate the knowledge to improve the segmentation accuracy, I integrate the task of contour detection with the task of segmentation, assuming that introducing a task of contour detection can help guide the network to concentrate more on the boundaries of organ regions, thus helping overcome the adverse effect of low tissue contrast. In this study, as shown in Fig. 3.23, two boundary detection tasks are added to the end of the network as auxiliary guidance.

To extract the contour for training, I first delineate the boundaries of different organs by performing Canny detector on the ground-truth segmentation. For the organs with clear boundaries (*i.e.*, bladder and rectum in the proposed case), I model the problem as a classification problem. However, due to the potential sample imbalance problem, I propose using focal loss to alleviate such an issue, as shown in Eq. 3.13.

$$L_{cboundary} = -\sum_h \sum_w \sum_t \sum_{c \in csets} I_{\{Y_{h,w,t},c\}} \left(1 - \hat{p}\left(X_{h,w,t};\theta\right)\right)^{\gamma} \left(1 - \hat{p}\left(X_{h,w,t};\theta\right)\right). \quad (3.13)$$

Note that, for the regions with blurry boundaries (*i.e.*, prostate in my case), the voxels near the boundaries look almost same. As a result, it will be more reasonable to assign soft labels (instead of hard labels) around the ground-truth boundaries. Thus, I can formulate the blurry-boundary delineation task as a soft classification problem, which estimates the probability of each voxel being on the organ boundaries. Then, for these blurry boundaries, I further exert a Gaussian filter (with a bandwidth of δ, *i.e.*, empirically set to 3 in this study) on the obtained boundary map. In other words, for each voxel, I generate an approximate probability belonging the blur boundary of an organ. Hence, I can formulate soft classification as a soft cross-entropy loss function as defined in Eq. 3.14.

$$L_{bboundary} = -\sum_h \sum_w \sum_t p_{h,w,t} \left(1 - \hat{p}\left(X_{h,w,t};\theta\right)\right). \quad (3.14)$$

3.3.4 Implementation Details

Pytorch[9] is adopted to implement the proposed method shown in Fig. 5.3. The code can be obtained by this link[10]. We adopt Adam algorithm to optimize the network. The input size of the segmentation network is $144 \times 144 \times 16$. The network weights are initialized by the Xavier algorithm, and weight decay is set to be 1e-4. For the network biases, I initialize them to 0. The learning rate for the network is initialized to 2e-3, followed by decreasing the learning rate 10 times every 2 epochs during the training until 1e-7. Four Titan X GPUs are utilized to train the networks.

3.3.5 Experimental Results

The used pelvic dataset consists of 50 prostate cancer patients from a cancer hospital, each with one T2-weighted MR image and corresponding manually-annotated label map by medical experts. In particular, the prostate, bladder and rectum in all these MRI scans have been manually segmented, which serve as the ground truth for evaluating the proposed segmentation method. All these images were acquired with 3T MRI scanners. The image size is mostly $256 \times 256 \times (120 \sim 176)$, and the voxel size is $1 \times 1 \times 1$ mm^3. A typical example of the MR image and its corresponding label map are given in Fig. 3.22.

Five-fold cross validation is used to evaluate the proposed method. Specifically, in each fold of cross validation, I randomly chose 35 subjects as training set, 5 subjects as validation set, and the remaining 10 subjects as testing set. Unless explicitly mentioned, all the reported performance by default is evaluated on the testing set. As for evaluation metrics, I utilize DSC (Eq. 2.22) and ASD (Eq. 2.24) to measure the agreement between the manually and automatically segmented label maps.

[9] https://github.com/pytorch/pytorch

[10] https://github.com/ginobilinie/SemGuidedSeg.git

3.3.5.1 Comparison with State-of-the-art Methods

To demonstrate the advantage of the proposed method, I also compare the proposed method with other three widely-used methods on the same dataset as shown in Table 3.11: 1) SSAE [64], 2) UNet [164], 3) SResSegNet [214].

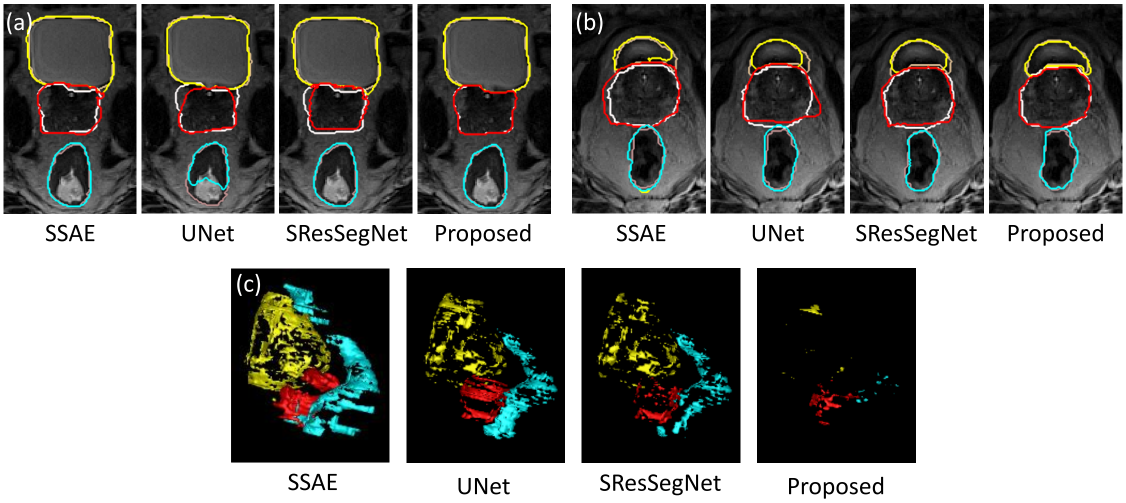


Figure 3.24: Visualization of pelvic organ segmentation results by four methods. In (a) and (b), orange, silver and pink contours indicate the manual ground-truth segmentations, while yellow, red and cyan ones indicate automatic segmentations. (a) Clear boundary case, (b) blurry boundary case, and (c) 3D renderings of difference maps between ground-truth segmentation and automatic segmentations.

Table 3.11: DSC and ASD on the pelvic dataset by four different methods.

Method	DSC			ASD		
	Bladder	Prostate	Rectum	Bladder	Prostate	Rectum
SSAE	.918(.031)	.871(.042)	.863(.044)	1.089(.231)	1.660(.490)	1.701(.412)
UNet	.896(.028)	.822(.059)	.810(.053)	1.214(.216)	1.917(.645)	2.186(0.850)
SResSegNet	.944(.009)	.882(.020)	.869(.032)	.914(.168)	1.586(.358)	1.586(.405)
Proposed	**.975(.006)**	**.932(.017)**	**.918(.025)**	**.850(.148)**	**1.282(.273)**	**1.351(.347)**

Table 3.11 quantitatively compares the proposed method with three state-of-the-art segmentation methods. We can see that my method achieves better accuracy than the other state-of-the-art methods in terms of both DSC and ASD. It is worth noting that the proposed method can achieve much better performance for the blurry-boundary organ (*i.e.*, prostate), which indicates the effectiveness of the proposed network components for blurry boundary delineation.

I also visualize some typical segmentation results in Fig. 3.24, which further show the superiority of the proposed method, especially for the blurry regions around the prostate.

3.3.5.2 Impact of Each Proposed Component

As the proposed method consists of several novel proposed components, I conduct empirical studies below to analyze them.

Impact of Proposed SG Module: As mentioned in Sec. 3.3.2, I propose a semantic-guided encoder feature learning module to learn more discriminative features in shallow layers. The effectiveness of the SG module is further confirmed by the improved performance, *e.g.*, 2.40%, 4.41% and 2.8% performance improvements in terms of DSC for bladder, prostate, and rectum, respectively, compared with the UNet with multi-scale deep supervision.

Relationship with Similar Work: Several previous work are proposed to use attention mechanism [166, 155] to enhance the encoder-decoder networks. However, the proposed work is different from them mainly in that I propose to use highly semantic information from the decoder to explicitly guide the building of attention mechanism, so that I can efficiently learn the encoder features. To further compare them, I visually present the three *typical* learned feature maps (selected by clustering) of a certain layer (*i.e.*, combined layer) in different networks at a certain training iteration (*i.e.*, 4 epochs). The methods include FCN [124], UNet [164], UNet with concurrence SE module [166] (ConSEUNet), attention-UNet [155] (AttUNet) and the proposed one(SGUNet). The visualized maps are in Fig. 3.25.

Fig. 3.25(a-e) indicates that the raw encoder-decoder network (*i.e.*, FCN and UNet) cannot handle well for the blurry boundary cases. The attention based networks can generate higher semantic maps with better localization information. Among them, the proposed method can learn more precise boundaries due to explicit semantic guidance. Also, the

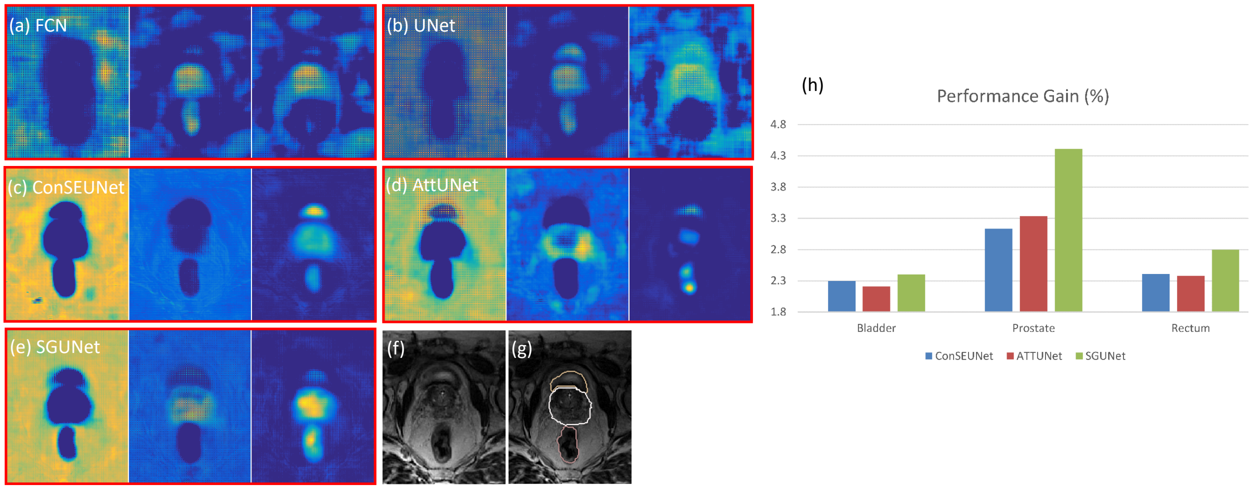


Figure 3.25: (a-e): Visualization of three typical learned feature maps of a certain layer by five different networks. (f) and (g) are the corresponding input MRI and the MRI overlaid by the ground-truth contours. (h) is the performance gain in terms of DSC with different strategies towards the UNet with *multi-scale deep supervision*.

proposed method have a faster convergence compared to other methods. Besides, the quantitative analysis in Fig. 3.25(h) is consistent with the the conclusion of qualitative analysis.

Impact of Soft Contour Constraint: As introduced in Sec. 3.3.3, I apply a hard contour constraint for clear-boundary organs while a soft contour constraint for the blurry-boundary organs. Since hard contour constraint is a widely adopted strategy, I directly compare the proposed soft contour constraint with the case of using hard constraint. With soft constraint on the prostate, I can achieve a slight performance gain such as 0.2% in terms of DSC; but I can achieve more performance gain in terms of ASD (0.8%), which is mainly because the soft contour constraint can help more accurately locate the blurry boundaries.

3.3.5.3 Validation on Extra Dataset

To show the generalization ability of the proposed algorithm, I conduct additional experiments on the PROMISE12-challenge dataset [123]. This dataset contains 50 labeled subjects where only prostate was annotated. I can achieve a high DSC (0.92), small ASD (1.57) in average based on five-fold cross validation. As for the extra 30 subjects' testing dataset whose ground-truth label maps are hidden from us, the performance of the proposed algorithm is still very competitive (we are ranking the top 5 among 290 teams with an average overall score of 89.46.) compared to the state-of-the-art methods on the 30 subjects' testing dataset [214, 229]. These experimental results indicate a very good generalization capability of the proposed algorithm.

CHAPTER 4: DEEP NEURAL NETWORKS FOR MEDICAL IMAGE SYNTHESIS

As mentioned in Sec. 1.2, traditional methods for medical image synthesis will face various problems, such as feature extraction and non-linear mapping optimization problems. The previous deep learning based methods could improve the quantitative performance, while it cannot generate better-perceived images. In this chapter, I proposed a deep residual adversarial network for medical image synthesis coupled with a novel gradient difference loss function. In addition, I designed auto-context refinement to address the long-range information dependency issue[1].

4.1 Context-aware Deep Residual Adversarial Networks for Medical Image Synthesis

To address the aforementioned problems and challenges in Sec. 1.2, I propose a deep convolutional adversarial network framework by adversarially training FCN as the generator and CNN as the discriminator. First, I propose a basic 3D FCN to estimate the target image from the corresponding source image. Note that I adopt 3D operations to better model the 3D spatial mapping and thus could solve the discontinuity problem across 2D slices, which often occurs when using the 2D CNN. Second, I utilize the adversarial learning strategy [60] for the designed network, where an additional discriminator network is modeled. The discriminator urges the generator output to be similar with the ground-truth target image perceptually. The generator is featured with incorporating the image gradient difference into the loss function, with the goal of retaining the sharpness of the

[1] This work was published in MICCAI 2017 [148] and IEEE Transactions on Biomedical Engineering [150]. This chapter uses parts of text descriptions and figures from the published papers.

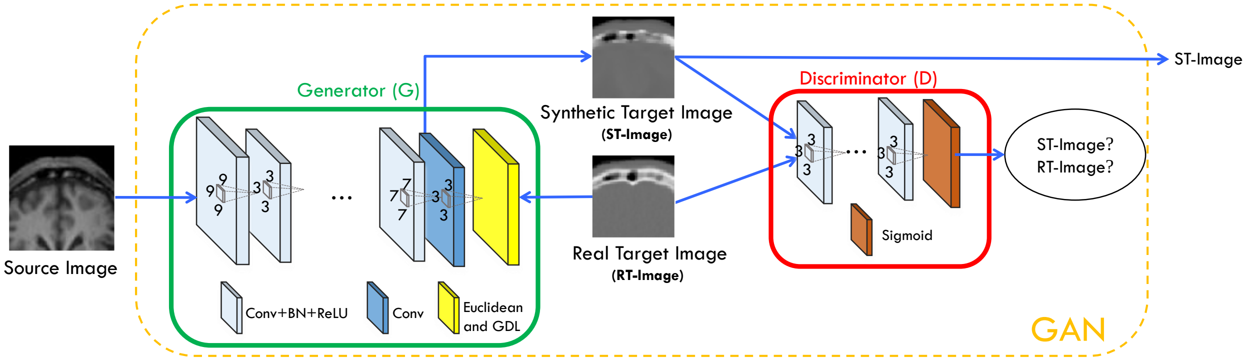


Figure 4.1: Architecture used in the deep convolutional adversarial setting for estimation of the synthetic target image.

generated target image. I further explore the long-term residual unit to train the network. Moreover, I employ auto-context model (ACM) to iteratively refine the output of the generator. At the testing stage, an input source image is first partitioned into overlapping patches, and, for each patch, the corresponding target is estimated by the generator. Then, all generated target patches are merged into a single image to complete the source-to-target synthesis by averaging the intensities of the overlapping CT regions. In the following section, I describe in detail the GAN framework used in the source-to-target image synthesis.

4.1.1 Supervised Deep Convolutional Adversarial Network

As mentioned above, I propose a supervised deep convolutional adversarial framework, which is inspired by the recent popular generative adversarial networks (GAN) [60], to complete the source-to-target synthesis as shown in Fig. 4.1. The components in Fig. 4.1 will be introduced in the following paragraphs.

Fully Convolutional Network (FCN) for Medical Image Synthesis: FCN is widely used for segmentation and reconstruction in both computer vision and medical image analysis fields [124, 153, 151, 48, 10, 65, 145], because it can preserve spatial information in local neighborhood of the image space and is also much faster compared to CNN at the testing stage. In this study, I adopt FCN to implement the image generator. A typical 3D FCN (as shown in Fig. 4.2) is proposed to perform the medical image synthesis task. I use only the convolution operations without pooling, which would potentially lead to loss of resolution.

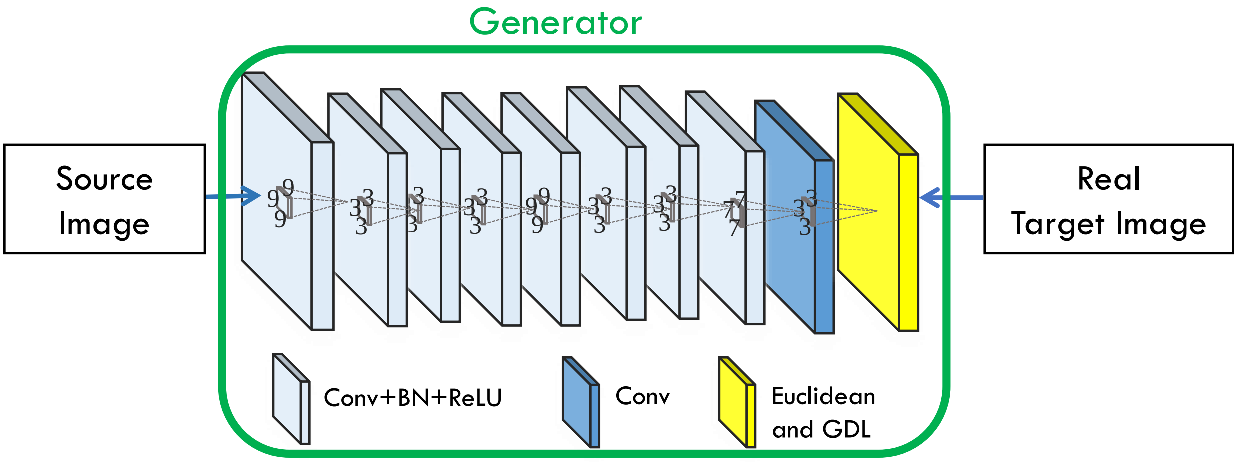


Figure 4.2: The 3D FCN architecture for estimating a target image from a source image.

As described in the Introduction section, typically a Euclidean loss is used to train the model as shown in Eq. 4.1.

$$L_G(X, Y) = \|Y - G(X)\|_2^2, \tag{4.1}$$

where Y is the ground-truth target image, and $G(X)$ is the generated target image from the source image X by the Generator network G.

Adversarial Learning: To make the generated target images better perceptually, I propose to use adversarial learning to improve the performance of FCN.

GANs have achieved the state-of-the-art results in the field of image generation by producing very realistic images in an unsupervised setting [60, 161]. Inspired by the works in [133, 60], I propose the supervised GAN to synthesize medical images. My networks include 1) the generator for estimating the target image and 2) the discriminator for distinguishing the real target image from the generated one, as shown in Fig. 4.1. The generator network G is an FCN as described above. The discriminator network D is a CNN, which estimates the probability of the input image being drawn from the distribution of real images. That is, D can classify an input image as “real” or “synthetic”.

Both networks are trained simultaneously, with D trying to correctly discriminate real and synthetic data, and G trying to produce realistic images that confuse D. Concretely, the loss function for D and G can be defined as:

$$L_D(X,Y) = L_{BCE}(D(Y),1) + L_{BCE}(D(G(X)),0), \tag{4.2}$$

where X is the source input image, Y is the corresponding target image, $G(X)$ is the estimated image by the generator, and $D(\cdot)$ computes the probability of the input to be “real”. And, L_{BCE} is the binary cross entropy defined by Eq. 4.3.

$$L_{BCE}(\widehat{Y},Y) = -\sum_i Y_i \log\left(\widehat{Y}_i\right) + (1-Y_i)\log\left(1-\widehat{Y}_i\right), \tag{4.3}$$

where Y represents the label of the input data and takes its values in $\{0,1\}$ (i.e., 0 for the generated image and 1 for the real one), and $\widehat{Y}$ is the predicted probability in $[0,1]$ that the discriminator assigns to the input of being drawn from the distribution of real images.

On the other hand, the loss term used to train G is defined as:

$$\begin{aligned} L_{G\_ADV}(X,Y) =& \lambda_1 L_{ADV}(X) \\ & + \lambda_2 L_G(X,Y) + \lambda_3 L_{GDL}(Y,G(X)). \end{aligned} \tag{4.4}$$

Specifically, I minimize the binary cross entropy (“BCE”) between the decisions of D and the correct labels (“real” or “synthetic”), while the network G minimizes the binary cross entropy between the decisions by D and the wrong labels for the generated images. The loss of G incorporates the traditional term used for image synthesis in Eq. 4.1, as well as several other terms that will be detailed later. In general, D can distinguish between the real target data and the synthetic target data generated by G. At the same time, G aims to produce more realistic target images and to confuse D.

In the case of G, I use a loss function that includes an adversarial term ("ADV") to fool D:

$$L_{ADV}(X) = L_{BCE}(D(G(X)), 1). \tag{4.5}$$

The training of the two networks is performed in an alternating fashion. First, D is updated by taking a mini-batch of real target data and a mini-batch of generated target data (corresponding to the output of G). Then, G is updated by using another mini-batch of samples including sources and their corresponding ground-truth target images.

Image Gradient Difference Loss: If I only take into account Eq. 4.5 for the generator, the system would be able to generate images that are drawn from the distribution of the target data. I further incorporate the L2 loss term of Eq. 4.1 as a data fitting term in the loss of the generator, aiming at producing realistic images. Training the system with the above mentioned losses would be able to generate a target image from its corresponding source image. Furthermore, as the L_2 loss may produce blurry images, I propose to use an image gradient difference loss ("GDL") as an additional term. It is defined as:

$$\begin{aligned} L_{GDL}(Y, \hat{Y}) = & \left| |\nabla Y_x| - \left|\nabla \widehat{Y}_x\right| \right|^2 \\ & + \left| |\nabla Y_y| - \left|\nabla \widehat{Y}_y\right| \right|^2 \\ & + \left| |\nabla Y_z| - \left|\nabla \widehat{Y}_z\right| \right|^2, \end{aligned} \tag{4.6}$$

where Y is the ground-truth target image, and $\hat{Y}$ is the estimated target by the generator network. This loss tries to minimize the difference of the magnitudes of the gradients between the ground-truth target image and the synthetic target image. In this way, the synthetic target image will try to keep the regions with strong gradients (*e.g.*, edges) for an effective compensation of the L_2 reconstruction term. By combining all losses above, the generator can thus be modeled to minimize the loss function shown in Eq. 4.4.

Architecture Details: The architecture of the proposed generator network G is showed in Fig. 4.2, where the numbers indicate the filter sizes. This network takes a source image as the input, and tries to generate the corresponding target image. The architecture is designed with empirical knowledge from the widely-used FCN architectures. As the input size of the proposed network is $32 \times 32 \times 32$ and the output size is $16 \times 16 \times 16$, I have to reduce the feature map sizes during the network inference. If keeping $3 \times 3 \times 3$ as the kernel size, I will have too many layers, which is challenging to both physical memory and optimization in training. Thus, I choose several big kernels to decrease the depth of the network in the generator. The proposed kernel size setting is empirical, and I believe that other possible configurations can also be used. Specifically, it has 9 layers containing convolution, batch normalization (BN) and ReLU operations. The kernel sizes are 9^3, 3^3, 3^3, 3^3, 9^3, 3^3, 3^3, 7^3, and 3^3 respectively. The numbers of filters are 32, 32, 32, 64, 64, 64, 32, 32, and 1, respectively, for the individual layers. The last layer only includes 1 convolutional filter, and its output is considered as the estimated target image. Regarding the architecture, I avoid the use of pooling since it will reduce the spatial resolution of the feature maps. Considering the fact that the traditional convolution operations of the generator in Fig. 4.1 cannot guarantee a sufficiently effective receptive field [128], I adopt the dilated convolution as an alternative [212] so that I can achieve enough receptive field. The dilation for the first and last convolution layers of the generator in Fig. 4.1 are 1, and 2 for all the rest convolution layers.

The discriminator D is a typical CNN architecture including three stages of convolution, BN, ReLU and max pooling, followed by one convolutional layer and three fully connected layers where the first two use ReLU as activation functions and the last one uses sigmoid (whose output represents the likelihood that the input data is drawn from the distribution of real target image). The filter size is $3 \times 3 \times 3$, the numbers of the filters are 32, 64, 128 and 256 for the convolutional layers, and the numbers of the output nodes in the fully connected layers are 512, 128 and 1.

4.1.2 Residual Learning for the Generator

CNN with residual connections has achieved promising results in many challenging generic image processing tasks [71, 97]. Residual connections, in principle, help bypass the nonlinear transformations with an identity mapping in the network and explicitly reformulates the layers as learning residuals with reference to the precedent layers [71]. Formally, the residual connection can be expressed as Eq. 4.7:

$$y = F\left(x, \{W_i\}\right) + x, \tag{4.7}$$

where W_i are the convolutional filters in the bottleneck residual unit, and x and y are the input and output feature maps, respectively.

'ResNet' demonstrates that the residual connection benefits convergence when training a very deep CNN. The residual learning unit works on a local convolutional layer by transforming it to a bottleneck architecture. Since in some tasks, the source and target images (e.g., 3T-to-7T task) are largely similar, in this study, I extend such a connection (bottleneck architecture) to skip the whole CNN (or FCN), instead of a single convolutional layer. With this long-term residual unit, the residual image is likely to be (or close to be) zero, making the network much easier to train. The long-term residual connection is illustrated as the solid-purple-line arrow in Fig. 4.3.

For the 3T-to-7T synthesis task, it might be hard for very deep networks to produce accurate results since the model will require a very long-term memory [98]. This is due to the fact that the structure of the output is very similar to the structure of the input, and the required memory might be difficult to model during the training because of vanishing gradient issues and the large number of layers. As introduced in the above paragraphs, residual learning can help to alleviate this issue by learning a residual map in the last layer [97, 71]. This is accomplished by adding a skip connection from the input to the final layer and then performing an element-wise addition. In Fig. 4.3, I show this architecture.

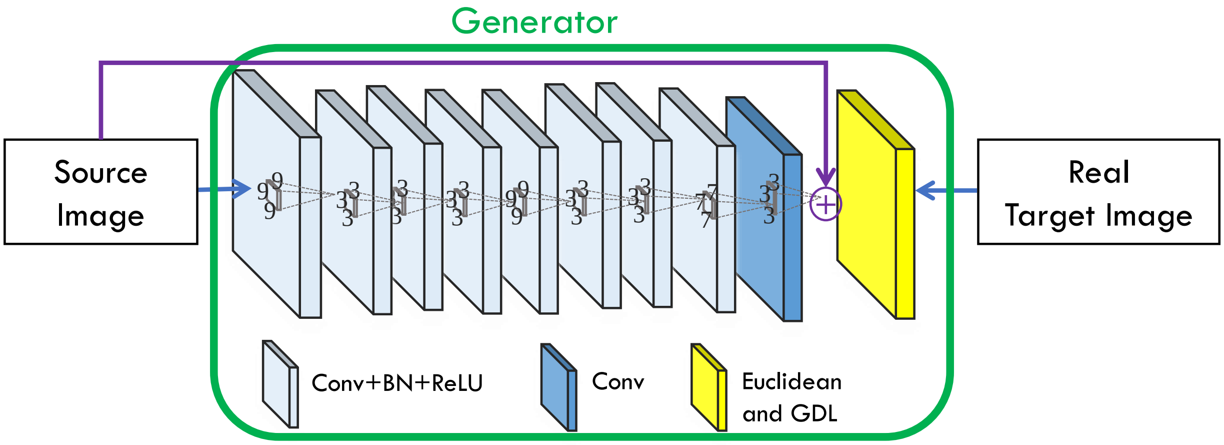


Figure 4.3: Generator architecture used in the GAN setting for estimation of synthetic target image. Note the solid-purple-line arrow from source image to the 'plus' sign, which expresses the long-term residual connection.

It is worth mentioning that this long-term residual unit only makes sense for tasks where the input is highly correlated to the output, such as 3T-to-7T synthesis, super resolution, denoising and so on. For that reason, I only use this method for the 3T-to-7T synthesis task in this study.

4.1.3 Auto-Context Model (ACM) for Refinement

Since the work in this study is patch-based, the context information available for each training sample is limited inside the patch. This obviously affects the modeling capacity of the proposed network. One remedy to enlarge the context during training is by using ACM, which is commonly used in the task of semantic segmentation [181]. The idea is to train several classifiers iteratively, where each classifier is trained not only with the feature data of original image(s) but also with the probability map outputted by the previous classifier. Note that the output of the previous classifier gives additional context information for the subsequent classifier to use. At testing time, the input will be processed for each classifier one after the other, concatenating the probability map to the initial input.

In this work, I show that the ACM can also be applied successfully to the deep learning based regression tasks. In particular, I adopt the ACM to iteratively refine the generated results, making the proposed GAN context-aware. To this end, I iteratively train several GANs that take inputs from both early synthetic target patches and source patches. These patches are then concatenated as a second channel with the source patches, which are both input for training of the next GAN. An illustration of this scheme is shown in Fig. 4.4. It is worth noting that the architectures of the GANs I use for ACM are exactly the same as shown in Fig. 4.1. The only difference is about the input to the generator, which concatenates the source MRI patch and the synthetic target patch since the 1^{st} iteration of ACM. I keep the same sizes of the input patches throughout the ACM based refinement. Since the context information is extracted from the whole previously-estimated target image, they can encode information that is not available within the initial input image patch.

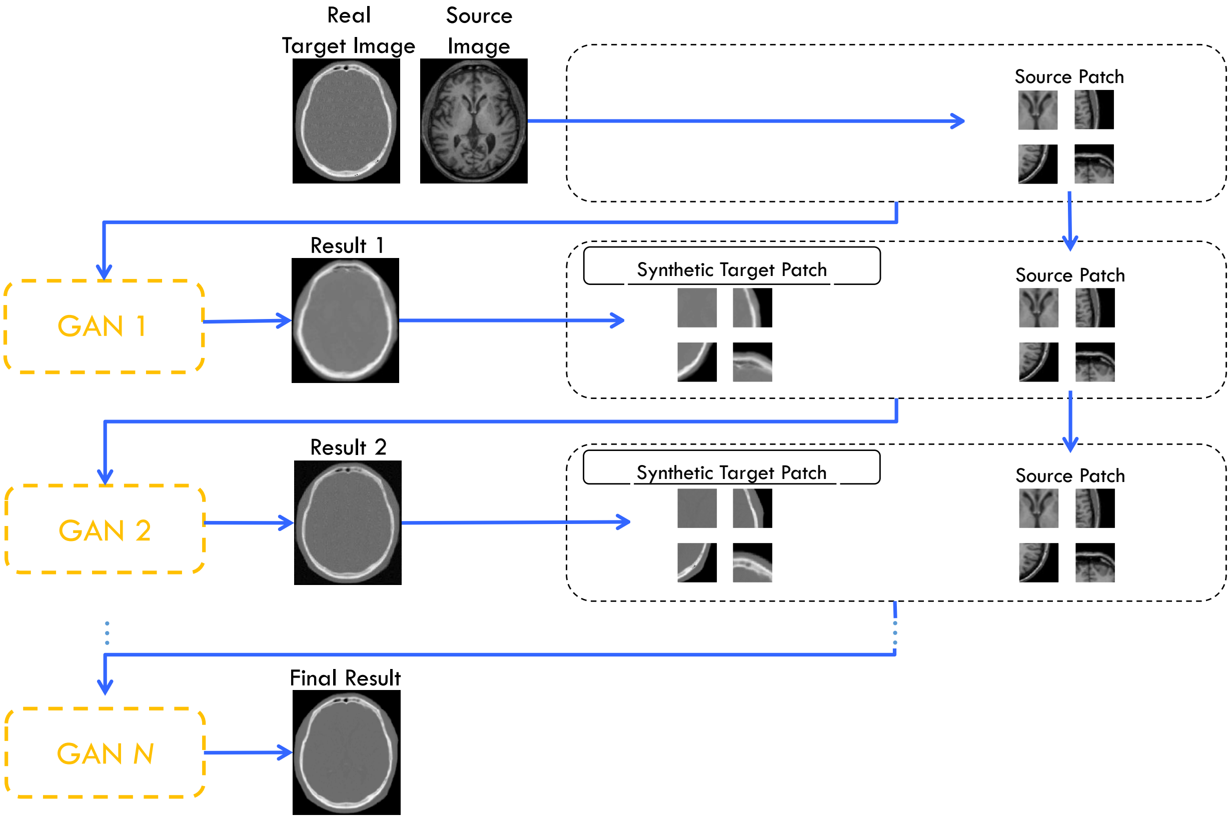


Figure 4.4: Proposed architecture for ACM with GAN.

4.2 Experimental Results for Medical Image Synthesis

I use three datasets to test the proposed method in two different tasks. First, I estimate CT images from its corresponding MRI data for both pelvic and brain datasets. Second, I estimate 7T MRI data from its corresponding 3T MRI data. I describe the experiments and the results of these two tasks separately.

4.2.1 Experiments on MR-to-CT Synthesis

- The brain dataset was acquired from 16 subjects with both MRI and CT scans in the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (see `www.adni-info.org` for details). The MR images were acquired using a Siemens Triotim scanner, with the voxel size $1.2 \times 1.2 \times 1$ mm^3 , TE 2.95 ms, TR 2300 ms, and flip angle 9°. The CT images, with the voxel size $0.59 \times 0.59 \times 3$ mm^3, were acquired on a Siemens Somatom scanner. A typical example of preprocessed CT and MR images is given in Fig. 1.3.

- The pelvic dataset consists of 22 subjects, each with MR and CT images. The spacings of CT and MR images are $1.172 \times 1.172 \times 1$ mm^3 and $1 \times 1 \times 1$ mm^3, respectively. In the training stage, I rigidly align the CT to MRI with FLIRT for each subject [74]. As there may be a large deformation on soft tissues for the prostate images, I adopt non-rigid registration (i.e., ANTs-SyN [6]) for each individual subject with careful parameter tuning. For both of these rigid and non-rigid registration steps, I use mutual information as the similarity metric to perform the registration with intensity images. To refine the registration results especially on the crucial pelvic organs between the CT and MR images, I further use Diffeomorphic Demons to register the respective manual labels of the prostate, bladder and rectum. In this way, the boundaries of these pelvic organs can be strictly aligned in the CT and MR images after final registration. After alignment, CT and MR images of the same patient have the same image size and spacing. Since only pelvic regions are concerned, I further crop the aligned CT and MR images to reduce the computational burden. Finally, each preprocessed image has a size of $153 \times 193 \times 50$ and a spacing of $1 \times 1 \times 1$ mm^3.

I first normalized the data using $\overline{X} = (X - mean)/std$, where $mean$ and std is the mean value and stand deviation across all the training data. And then I randomly extracted source patches of size $32 \times 32 \times 32$, along with their corresponding target patches of size $16 \times 16 \times 16$, by using the same center point as each pair of training samples. The networks were trained using the Adam optimizer with a learning rate of 10^{-6}, $\beta_1 = 0.5$ as suggested in [161], and mini-batch size of 10. The generator was trained using $\lambda_1 = 0.5, \lambda_2 = \lambda_3 = 1$.

The code is implemented using the TensorFlow library, and is publicly available from this github[2]. The training is done with a Titan X GPU. For the brain dataset, it costs about 10 hours to train the GAN in the 0^{th} iteration of ACM, and 3 hours to train the 1^{st} and 2^{nd} iteration of ACM based refinement, respectively. For the pelvic dataset, it costs about 12 hours to train the GAN in the 0^{th} iteration of ACM, and 3.5 hours for the 1^{st} and 2^{nd} iterations of ACM based refinement, respectively. As mentioned in Sec. 4.1, the intensities of the overlapping target image regions is averaged at the testing stage. To tradeoff between the time cost and the accuracy, I set the stride to be 8 along each direction of the image for the overlapping target image regions at the testing stage. The time cost for one testing brain MRI is about 1.2 minutes with the trained GAN model. Note, only generator is needed at the testing stage. In particular, the testing time costs increase to 2.4 and 3.6 minutes, respectively, if the 1^{st} and the 2^{nd} iterations of ACM are adopted. Similarly, the time cost for one testing pelvic MRI with the trained 0^{th}, 1^{st} and 2^{nd} iterations of ACM are 0.5, 1.0 and 1.5 minutes, respectively.

To demonstrate the advantage of the proposed method in terms of prediction accuracy, I compare it with three widely-used approaches: 1) atlas-based method [189]: specif-

[2] https://github.com/ginobilinie/medSynthesis

ically, it uses multi-atlas registration(by Demons) and intensity averaging for fusion, and the number of atlases I use is 5, 2) sparse representation based method (SR), and 3) structured random forest with ACM (SRF+) [86]. I used my own implementation of the first two methods, while for the third method (SRF+) I just show the results reported in [86]. All experiments are done in a leave-one-out fashion. The evaluation metrics are the MAE 2.25 and the PSNR 2.26.

Impact of Dilated Convolution: As mentioned in Sec. 4.1, I adopt a dilated convolution to replace the part of standard convolution operations in the generator, which could lead to a huge increase of the effective receptive field [128] (actually, with dilated convolution, the theoretical receptive field is 69) and thus make up for the insufficient receptive field of using standard convolution operation. The effect of using the dilated convolution operations is quantitatively evaluated in this study. In particular, the dilated FCN could provide PSNR of 24.7(1.4), while the FCN (with standard convolution operation) is 24.1(1.4). The GAN with dilated generator is able to achieve 25.2(1.4), in contrast, the GAN with standard generator's performance is 24.6(1.4). Thus, these experimental results further demonstrate the effectiveness of using the dilated convolution operation.

Impact of Adversarial Learning: To show the contribution of the adversarial learning, I conduct comparisons between the traditional FCN (i.e., just the generator shown in Fig. 4.1 but with dilated convolution operations) and the proposed GAN model. The PSNR values are 24.9 and 25.2 for the traditional FCN and the proposed approach, respectively. Note that these results do not include the adoption of ACM. I visualize results in Fig. 5.5, where the leftmost image is the input MRI, and the rightmost image is the ground-truth CT. I can clearly see that the generated data using the GAN approach has less artifacts than the traditional FCN, by estimating an image that is closer to the desired output quantitatively and qualitatively.

Impact of Gradient Difference Loss: To show how the proposed gradient difference loss (GDL) works in the framework, I conduct comparisons between the case of removing

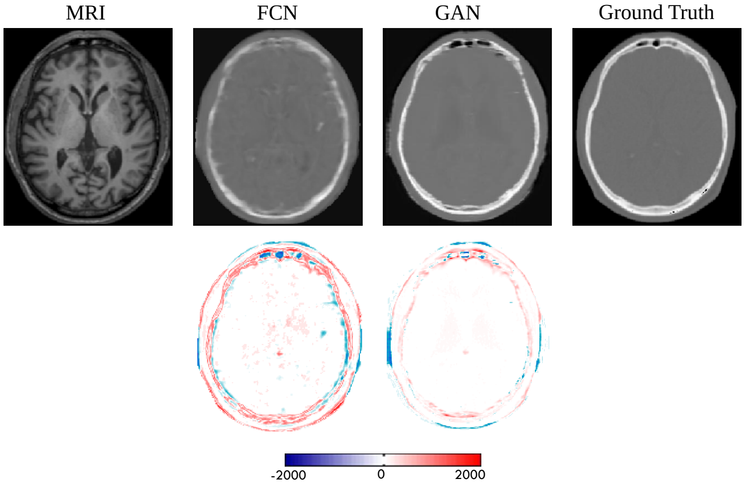


Figure 4.5: Visual comparison for impact of adversarial learning. The 1st row shows the synthetic CT by FCN and GAN, and the 2nd row shows the difference map between the synthetic CT and ground truth CT. Note that FCN means the case without adversarial learning, and GAN means the proposed method with adversarial learning.

GDL (No GDL, $\lambda_3 = 0$) and including GDL (With GDL, $\lambda_3 = 1$). The PSNR values are 25.2 and 25.9 for 'No GDL' and 'With GDL', respectively. Note again that these results do not include the adoption of ACM. I can visualize results in Fig. 4.6. It is very clear that the method 'With GDL' results in much sharper image. In contrast, the method 'No GDL' generates more blurred image. That is because GDL can enforce the gradient distribution of the generated image to be close to the gradient distribution of the real target image.

Auto-Context Model Refinement: To show the contribution of ACM, I present the performance (in terms of PSNR and MAE) of the proposed method with respect to the number of iterations of ACM in Fig. 4.7 and Fig. 4.8. I can observe that both MAE and PSNR are improved gradually and consistently with iterations, especially in the first two iterations. This is because ACM could solve the short-range dependency by providing long-range context information. Considering the trade-off between the performance and the training time, I choose 2 iterations for ACM in the proposed experiments on both datasets.

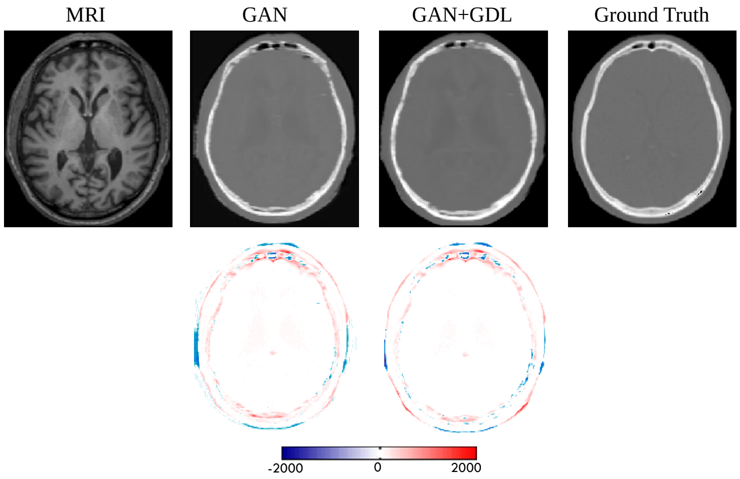


Figure 4.6: Visual comparison for impact of using the gradient difference loss (GDL). The 1st row shows the input MRI, two synthetic CT by two different methods, and the ground-truth CT. The 2nd row shows difference maps between each synthetic CT and the ground-truth CT.

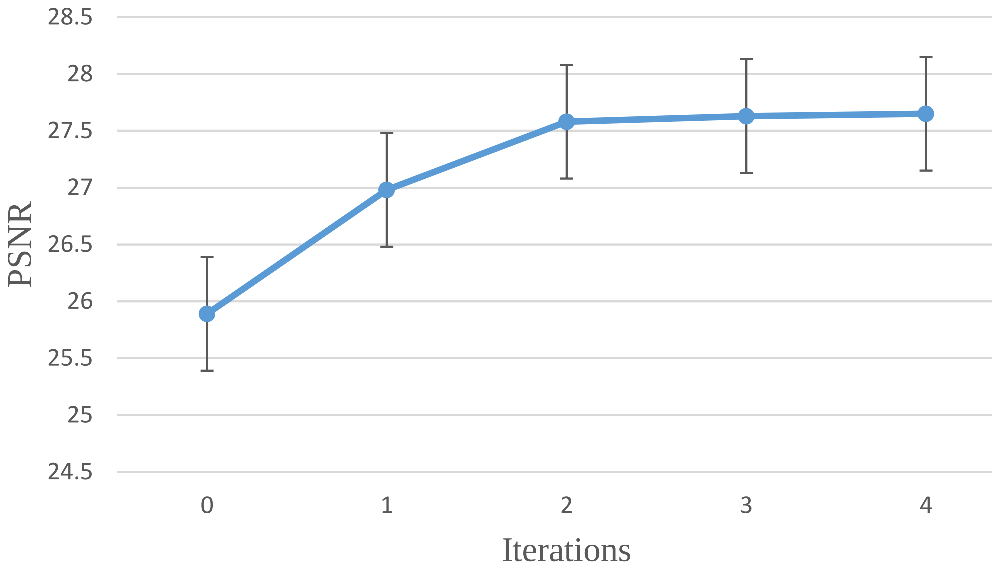


Figure 4.7: Performance (PSNR) of using ACM on the brain dataset with iterations.

In order to asses the effect of the ACM on the quality of the results, I also visualize one slice from the generated brain CT of a typical dataset in the first two stages of the framework in Fig. 4.9. The previous effects have been summarized in Table 4.1.

Comparison with Other Methods for the two MR-to-CT Synthesis datasets: To qualitatively compare the estimated CT by different methods, I visualize the generated

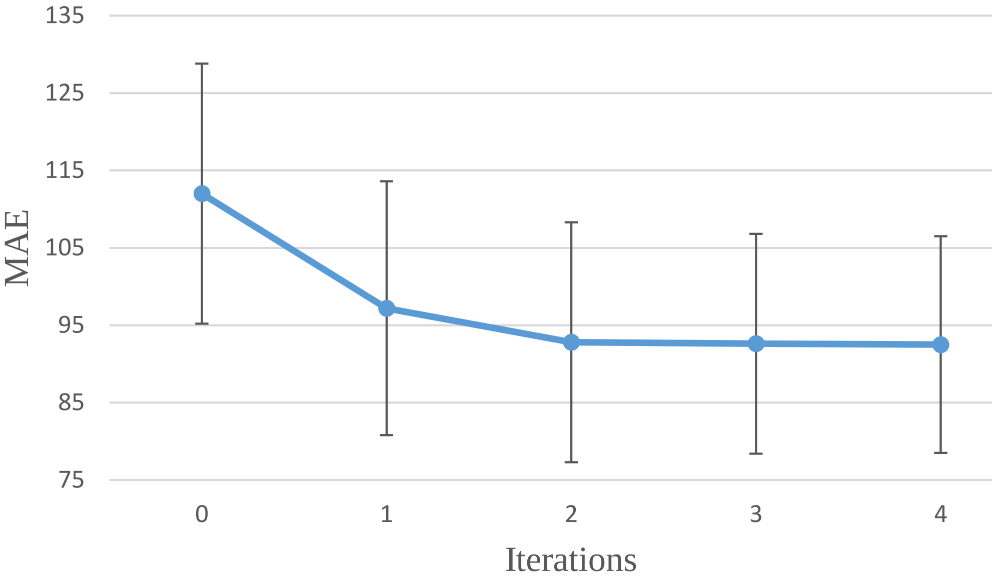


Figure 4.8: Performance (MAE) of using ACM on the brain dataset with iterations.

Table 4.1: Results summarizing different effects of the proposed method on the brain dataset in terms of PSNR.

Method	No Adv.	Adv.	Adv.+GDL	Proposed
Mean(std)	24.9(1.4)	25.2(1.4)	25.9(1.4)	27.6(1.3)

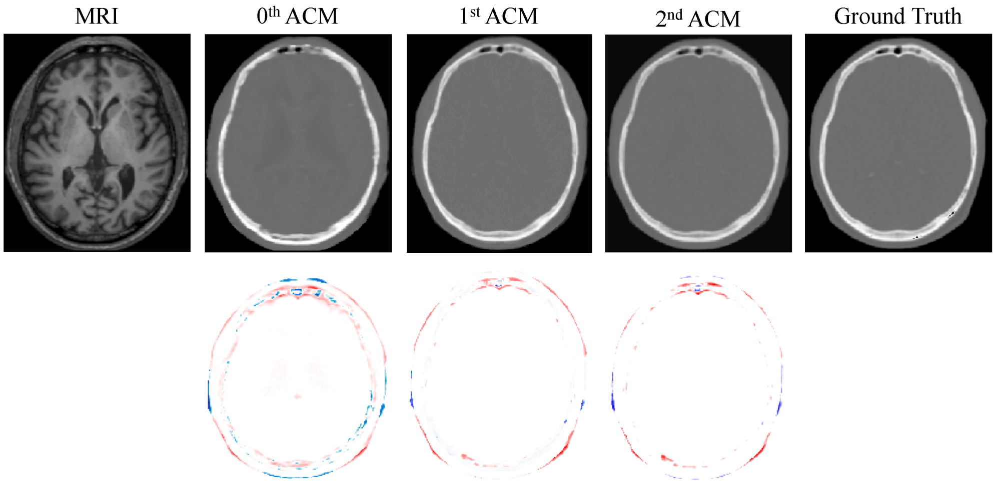


Figure 4.9: The 1st row shows visual comparison of MR image, three synthetic CT images by applying 0th, 1st and 2nd iterations of ACM, and the ground-truth CT image for a typical brain case. The 2nd row shows difference maps between each iteratively-estimated CT and the ground-truth CT.

CT with the ground-truth CT in Fig. 4.10. I can see that the proposed algorithm can better preserve the continuity and smoothness in the results since it uses image gradient difference constraints in the image patch as discussed in Section 4.1.1. Furthermore, I can conclude from the difference maps in Fig. 4.10 that the generated CT looks closer to the ground-truth CT compared to all other methods. I argue that this is due to the use of adversarial learning strategy that urges the generated images to be very similar to the real ones, so that even a complex discriminator cannot perform better than chance.

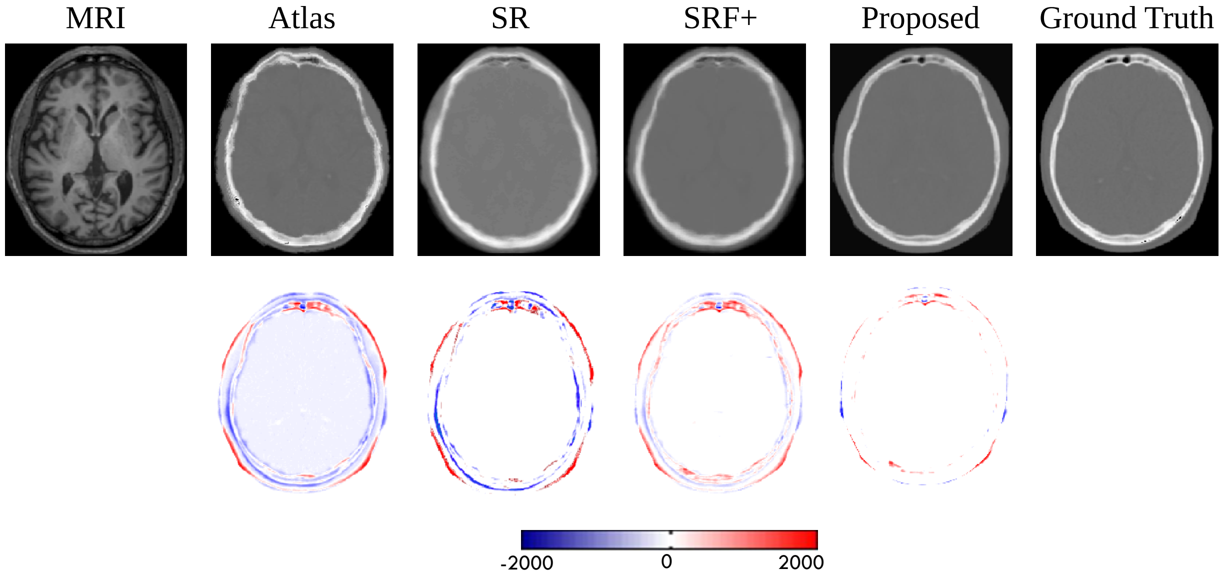


Figure 4.10: The 1st row shows visual comparison of the MR image, the four estimated CT images by other three competing methods and the proposed method, and the ground-truth CT for a typical brain case. The 2nd row shows difference maps between each estimated target CT and the ground-truth CT.

I also quantitatively compare the synthesis results in Table 4.2 using evaluation metrics, *i.e.*, PSNR and MAE. The proposed method outperforms all other competing methods in both metrics, which further demonstrates the advantage of my proposed framework.

Table 4.2: Average MAE and PSNR on 16 subjects from the brain dataset.

Method	MAE		PSNR	
	Mean (std)	Med.	Mean (std)	Med.
Atlas	171.5(35.7)	170.2	20.8(1.6)	20.6
SR	159.8(37.4)	161.1	21.3(1.7)	21.2
SRF+ [86]	99.9(14.2)	97.6	26.3(1.4)	26.3
Proposed	**92.5(13.9)**	**92.1**	**27.6(1.3)**	**27.6**

Table 4.3: Average MAE and PSNR on 22 subjects from the pelvic dataset.

Method	MAE		PSNR	
	Mean (std)	Med.	Mean (std)	Med.
Atlas	66.1(6.9)	66.7	29.0(2.1)	29.6
SR	52.1(9.8)	52.3	30.3(2.6)	31.1
SRF+ [86]	48.1(4.6)	48.3	32.1(0.9)	31.8
Proposed	**39.0(4.6)**	**39.1**	**34.1(1.0)**	**34.1**

The prediction results on the pelvic dataset by the same above methods are also shown in Fig. 4.11. It can be seen that my experimental result is consistent with the ground-truth CT. The quantitative results based on the same two evaluation metrics are shown in Table 4.3, indicating that the proposed method outperforms other competing methods in terms of both MAE and PSNR. Specifically, my method gives an average PSNR of 34.1, which is higher than the average PSNR of 32.1 obtained by the state-of-the-art SRF+ method. The MAE values (i.e., 39.0 by the proposed method, and 48.1 by the SRF+) further shows the improved effectiveness of my method.

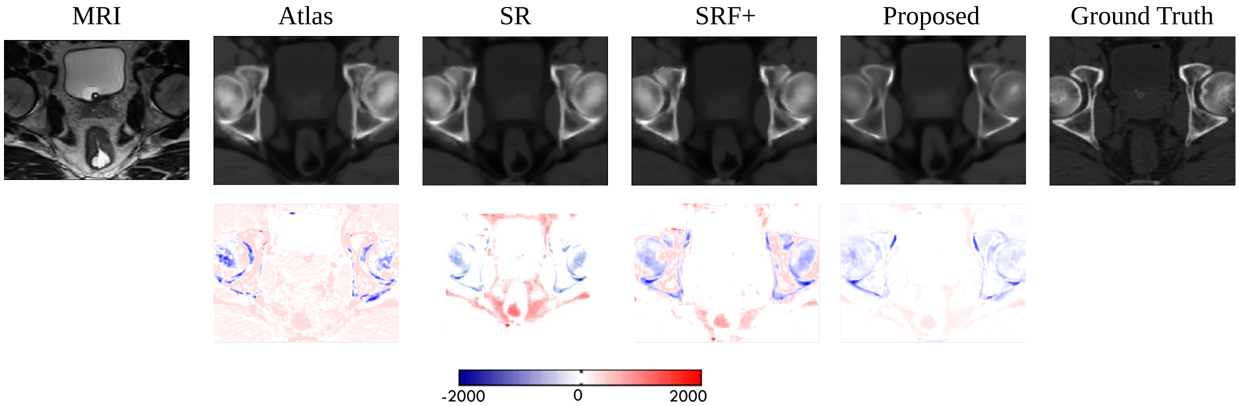


Figure 4.11: The 1st row shows visual comparison of the MR image, the estimated CT images by the proposed method and other competing methods, and the ground-truth CT image for the typical pelvic case; The 2nd row shows the difference maps between estimated CT and ground truth CT.

I further performed Wilcoxon signed-rank test to validate whether the improvement of the proposed method compared to the previous methods is significant or not. The experimental results in Table 4.4 show the statistical significant improvement ($p < 0.05$ by Wilcoxon signed-rank test).

Table 4.4: P-Values by performing Wilcoxon signed-rank Test between the proposed method and all the previous method for both PSNR and MAE values on brain and pelvic datasets.

Method	Brain		Pelvic	
	PSNR	MAE	PSNR	MAE
Atlas	<0.01	<0.01	<0.01	<0.01
SR	<0.01	<0.01	<0.01	<0.01
SRF+ [86]	<0.05	<0.05	<0.01	<0.01

4.2.2 Experiments on 3T-to-7T Synthesis

The 3T-to-7T dataset consists of 15 subjects, each with 3T MRI ($1 \times 1 \times 1$ mm^3) and 7T MRI ($0.65 \times 0.65 \times 0.65$ mm^3), scanned using 3T and 7T MRI scanners, respectively. The 7T MRI provides higher resolution and contrast than the 3T MRI, thus benefiting early diagnosis of brain diseases. These images are all linearly aligned and skull-stripped to remove non-brain regions.

Impact of Adversarial Learning: To show the contribution of the adversarial learning, I conduct comparison experiments between the traditional FCN (i.e., just the generator shown in Fig. 4.1) and the proposed GAN model. The PSNR values are 26.15(1.27) and 26.88(1.25) by the traditional FCN and the one with adversarial learning, respectively. Note that these results do not include the GDL, residual learning and ACM. I visualize results in Fig. 4.12, where the leftmost image is the 3T MRI, and the rightmost image is the ground-truth 7T MRI. I can clearly see that the generated data using the GAN approach has less artifacts than the traditional FCN, by estimating an image that is closer to the desired output quantitatively and qualitatively.

Impact of Gradient Difference Loss: To show how the proposed gradient difference loss (GDL) work in the framework, I conduct the same comparison experiments as the pre-

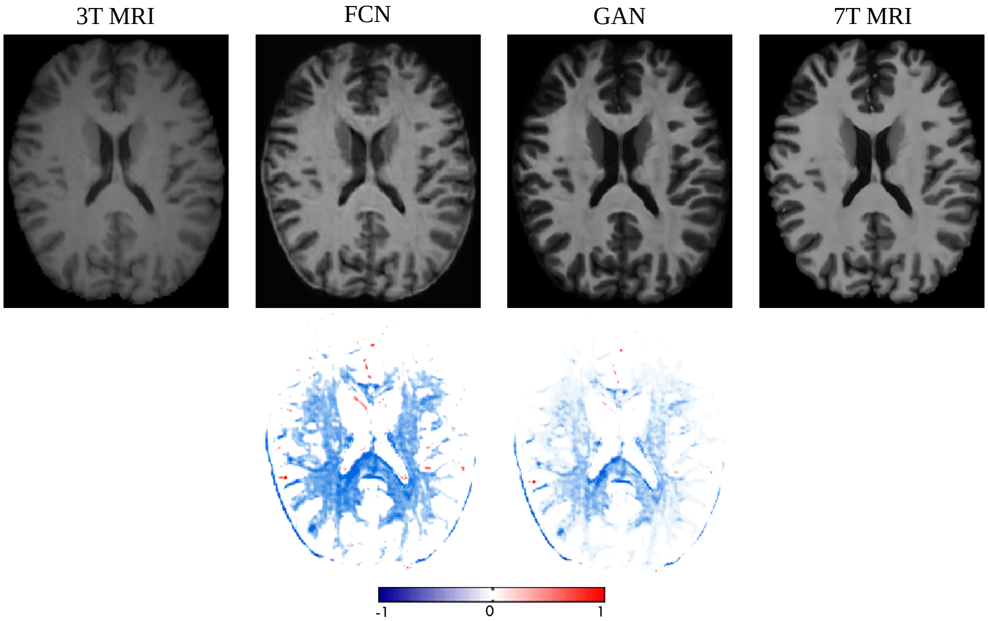


Figure 4.12: Visual comparison to demonstrate the impact of using the adversarial learning for the 3T-to-7T dataset. The 1st row shows 3T MRI, two synthetic 7T MRI by two methods, and ground-truth 7T MRI. The 2nd row shows difference maps between each synthetic 7T MRI and ground-truth 7T MRI. Note that FCN means the case without adversarial learning, and GAN means the case with adversarial learning.

vious datasets. The PSNR values are 26.83(1.25) and 27.18(1.24) for the method 'No GDL' and the method 'With GDL', respectively. These results do not include the residual learning and ACM. I visualize results in Fig. 4.13. Similar conclusions to those discussed above for the previous datasets can be made, i.e., obtaining much sharper and more realistic images.

Impact of Residual Learning: To show how the proposed long-term residual learning unit work in the framework, I conduct comparison experiments (i.e., using a GAN without this residual unit and a GAN with this unit, denoted as 'GAN' and 'ResGAN', respectively) to validate it in this dataset. The PSNR values are 27.18(1.24) and 27.69(1.22) by the method 'GAN' and the method 'ResGAN', respectively. Note that these results do not include the ACM. I have visualized the generated 7T MRI in Fig. 4.14. The 'ResGAN' generates a clearer 7T MRI compared to 'GAN', especially for the details. This is mainly

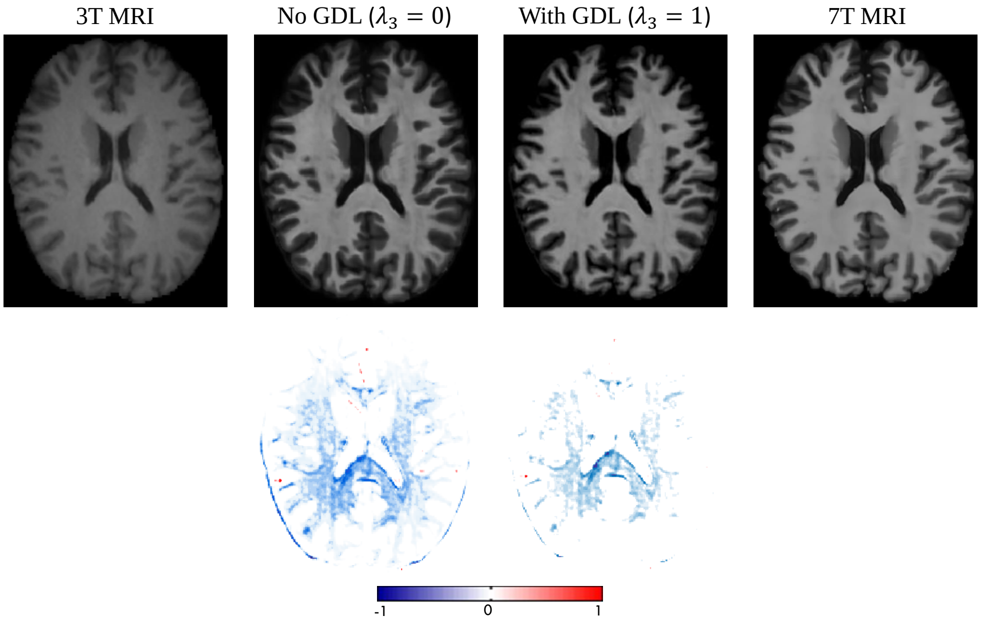


Figure 4.13: Visual comparison to demonstrate the impact of using the gradient difference loss. The image obtained via GDL is more realistic and sharper. The 1st row shows 3T MRI, the synthetic 7T MRI by two methods, and ground-truth 7T MRI; the 2nd row shows difference maps between each synthetic 7T MRI and the ground-truth 7T MRI.

Table 4.5: Results summarizing different effects of the proposed method on the 3T-to-7T dataset in terms of of PSNR.

Method	No Adv.	Adv.	Adv.+GDL	ResGAN	ResGAN+ACM
Mean(std)	26.15(1.27)	26.83(1.25)	27.18(1.24)	27.69(1.22)	27.93(1.18)

due to better convergence after using residual learning concept, which has been validated in Fig. 4.15.

More Evaluation for the Image Reconstruction Quality: Since the contrast is quite high with subcortical regions (such as thalamus and putamen) in 7T images compared to 3T, I also show a slice of the generated image in Fig. 4.16 in order to verify if this contrast is produced. The previous investigated effects have been summarized in Table 4.5.

On the other hand, it is important to notice that I have been evaluating the quality of the generated images with a global metric such as the PSNR. In a medical setting however, it is important to asses if the image is anatomically correct. Trying to evaluate the medical applicability, I decided to try a segmentation algorithm on both the generated image

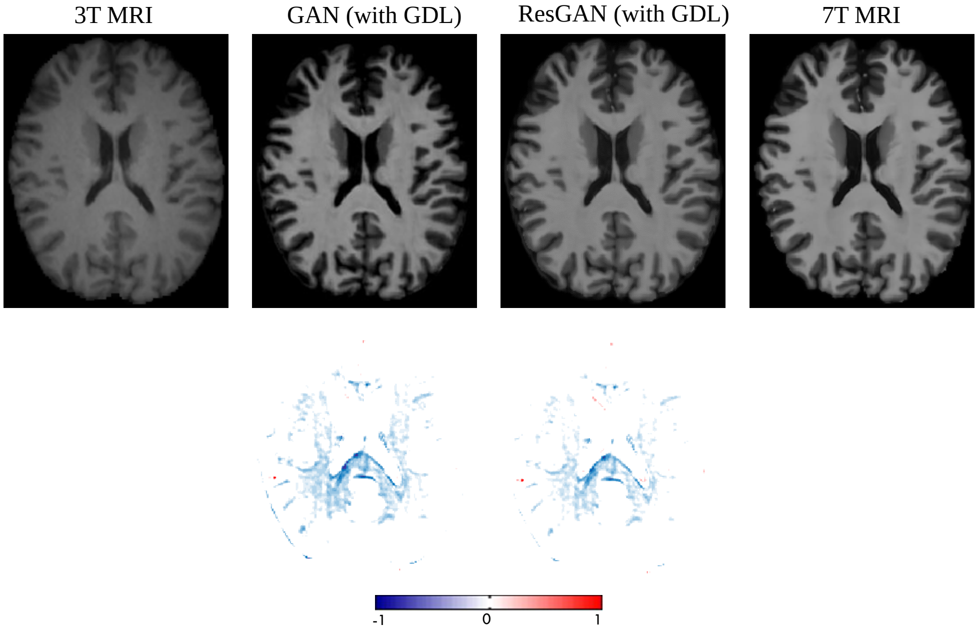


Figure 4.14: Visual comparison to demonstrate the impact of using the residual learning. The 1st row shows synthetic 7T MRI, and the 2nd row shows the difference maps between the synthetic 7T MRI and ground truth 7T MRI.

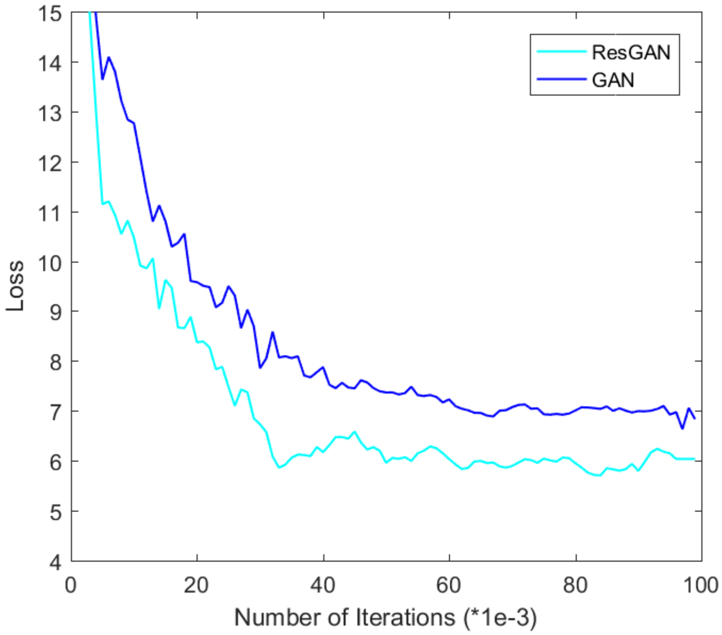


Figure 4.15: The mean squared error (MSE) of the generator in GAN and ResGAN on the testing dataset with respect to different training iterations.

and the real 7T. In particular, I train an FCN (U-NET [164]) in order to segment 7T images into White Matter (WM), Gray Matter (GM), and Cerebrospinal Fluid (CSF). I then

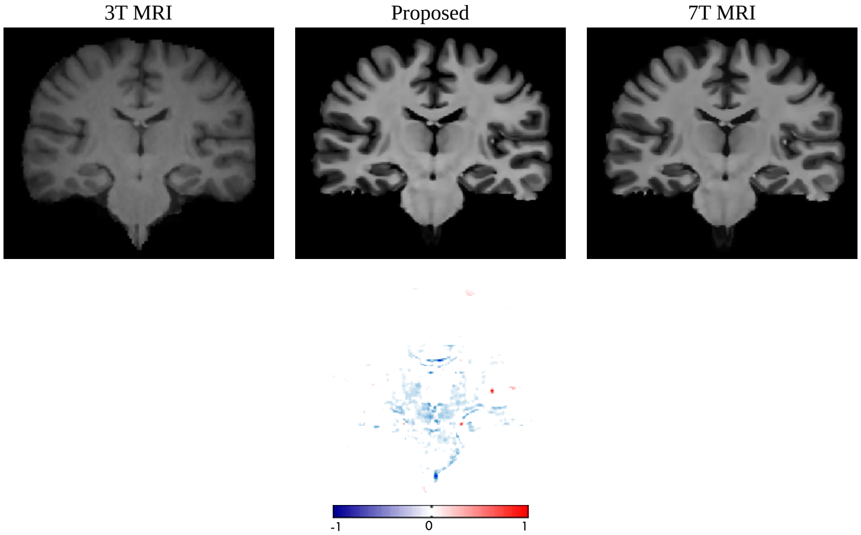


Figure 4.16: The visualization for subcortical regions with 3T MRI, the synthetic 7T MRI, and the ground-truth 7T MRI. The 2nd row shows a difference map between the ground-truth 7T MRI and the synthetic 7T MRI.

Table 4.6: Performance of segmentation on the MRI dataset in terms of Dice Index and its corresponding standard deviation.

Input	WM	GM	CSF
3T MRI	80.35(2.02)	85.49(1.08)	88.75(0.93)
Synthetic 7T MRI	86.84(1.84)	91.68(0.92)	95.96(0.88)
Ground-Truth 7T MRI	87.70(1.76)	92.33(0.86)	96.58(0.90)

evaluate the Dice Index of the segmentation maps obtained using the original 7T and the generated 7T as inputs to the network. I show a slice of the segmentation maps obtained in Fig. 4.17, and show the Dice Index in Table 4.6. The results show that the synthetic 7T MRI produces a segmentation map that is very close to that one produced by the real 7T in terms of Dice Index, and both of them largely outperform the results obtained by directly segmenting the 3T MRI. These results imply that the synthetic images have high quality and could be applicable to image segmentation.

Comparison with Other Methods: I compare the proposed method with several state-of-the-art methods: 1) HM: Histogram Matching, which matches the intensity distribution of an image with the intensity distribution of a target image; 2) LIS: Local Image Similarity [22], which synthesizes the target image using multiple atlases propagated according

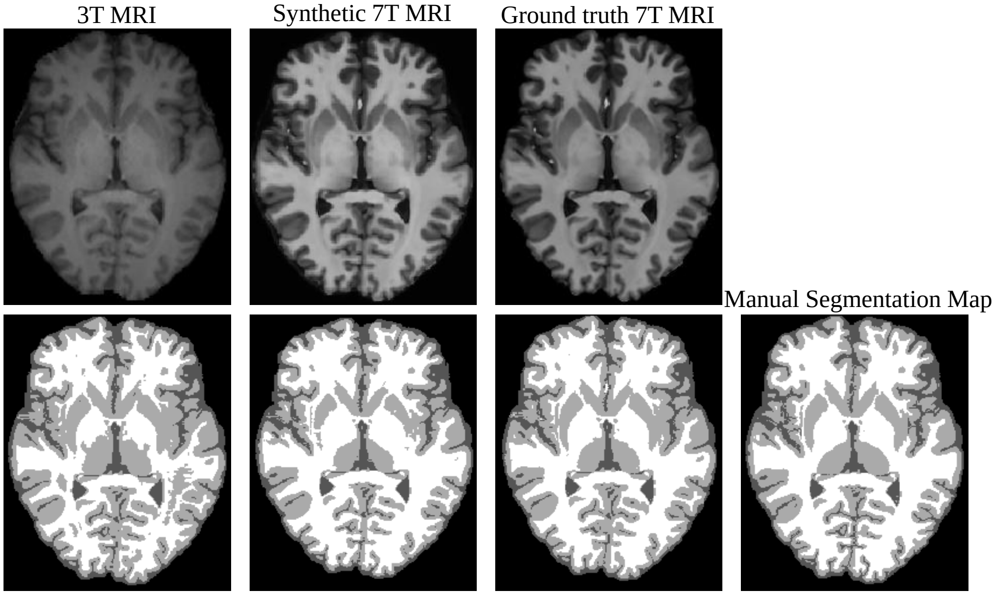


Figure 4.17: Visual comparison of segmentation results for a typical subject by using different input data (3T MRI, synthetic 7T MRI, and ground-truth 7T MRI). The 1st row is the MRI, and the 2nd shows the segmented slices as well as the manual segmentation map.

to local image similarity measures; 3) M-CCA: Multi-level CCA [12], which conducts a hierarchical reconstruction based on group sparsity in a novel multi-level CCA framework; 4) CNN: a 3D Convolutional Neural Networks [10], which learns non-linear mapping between the source image and target image. I list the experimental results in Table 4.7. The proposed framework outperforms the baselines methods by a big margin, which further validates the effectiveness of the proposed generative adversarial networks.

Table 4.7: Comparison of the performances of different methods on the 3T-to-7T dataset in terms of PSNR. The p-values by performing Wilcoxon signed-rank test between the proposed method and all other methods are also reported, and I use "*" to denote p<0.01.

Method	HM	LIS	M-CCA	CNN	Proposed
Mean (std)	21.10(1.44) *	24.33(1.26) *	25.41(1.20) *	26.50(1.22) *	**27.93(1.18)**

CHAPTER 5: ADVERSARIAL CONFIDENCE LEARNING FOR MEDICAL IMAGE ANALYSIS

As mentioned in Sec. 1.3 and Sec. 4.1, although adversarial learning could improve the visual perception to a large extent, it cannot provide quantitative performance gain at a similar extent. Thus, in this chapter, I analyzed the roles of discriminator in supervised adversarial learning systems and proposed an adversarial confidence learning framework to address such as issue. The Introduction also pointed out that training supervised models for medical image analysis could be easily dominated by the easy samples because of the distribution of irregular medical images. Accordingly, I proposed a novel difficulty-aware attention mechanism to better model the hard-to-segment (hard-to-synthesis) regions of the medical images following the adversarial confidence learning framework[1]. As figured out in Sec. 1.1.1, lack of annotated data was a ever-lasting challenge in medical image analysis. To move forward in this direction, I proposed a confidence-aware semi-supervised segmentation model[2] to provide a potential solution for this problem.

5.1 Adversarial Confidence Learning

I first present analysis for the roles of discriminator which is the basis of the proposed adversarial confidence learning. Then, I introduce the components of the proposed framework one by one with an example of medical image segmentation. Finally, I also extend the adversarial confidence learning framework to lesion image synthesis.

[1] One work was published in AAAI 2019 [154]. Another work was still under review in International Journal of Computer Vision [151]. This chapter uses parts of text descriptions and figures from these papers.

[2] This work was published in MICCAI 2018 [146]. This chapter uses parts of text descriptions and figures from the published paper.

5.1.1 Analyzing Role of Discriminator

To take better advantage of adversarial learning, I analyze the roles of discriminators in GAN systems and compare them between classic GAN and supervised adversarial learning system. Fig. 5.1(a) and (b) illustrate these two typical architectures.

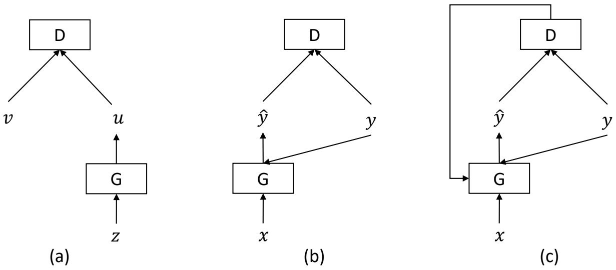


Figure 5.1: Illustration of classic GAN and supervised adversarial learning system. (a) shows a typical classic GAN, where z is an input signal following a certain distribution, u is the generated image, and v is the real image. (b) depicts a typical supervised adversarial learning system, where x is the input modality, $\hat{y}$ is the generated image, and y is the corresponding ground truth image. (c) introduces the proposed adversarial confidence learning framework which retains the adversarial learning and imposes confidence learning to enhance the supervised generator.

In classic GAN, there are two roles of the D: 1) distinguishing the real image v from the generated image u; 2) providing adversarial loss to train the G. In this unsupervised system, the training signal for G only comes from the D network, as a consequence, the generated u does not necessarily correspond to v but follow an implicit distribution of $\{v\}$. Similarly, the supervised adversarial learning system shown in Fig. 5.1(b), D also has the same roles. However, since G also benefits from the supervised loss from y besides the adversarial loss from D, the generated image $\hat{y}$ has a spatial match with the ground-truth image y. In other words, the G in supervised adversarial learning system in Fig. 5.1(b) does not rely on D as much as that in classic GANs.

Some research papers [125, 88, 152] figure out that adversarial learning in supervised models (segmentation and synthesis) work as high-order spatial consistency regularization to improve the supervised model since the traditional supervised losses (*i.e.*, cross entropy loss for segmentation and L_p loss for synthesis) for G only consider pixel-level correspondence but ignore image-level (or pairwise) match. With such adversarial learning, the qualitative performance (mainly visual perception) usually becomes better, while it cannot produce the same level of contribution to quantitative performance gain (the quantitative performance even degenerated in many cases) [88, 152].

In this study, I hope to take better advantage of adversarial learning so that the algorithm can synchronously improve both the visual perception and the quantitative performance. I propose adversarial confidence learning to achieve this goal, in which, the adversarial learning is retained by adopting a fully convolutional (dense) discriminator, and I develop confidence learning to enhance the design of the supervised generator. Specifically, I propose a fully convolutional adversarial framework as shown in Fig. 5.2. a) I adopt a full convolutional discriminator for local adversarial learning and also learn dense confidence information. b) With the well-learned confidence map, I propose difficulty-aware mechanism to improve the design of the supervised loss of the generator for medical image segmentation and synthesis. The architecture of the proposed framework is presented in Fig. 5.2, which consists of two sub-networks, i.e., a) base generator network (denoted as S for segmentation or synthesis) and b) confidence network (denoted as D).

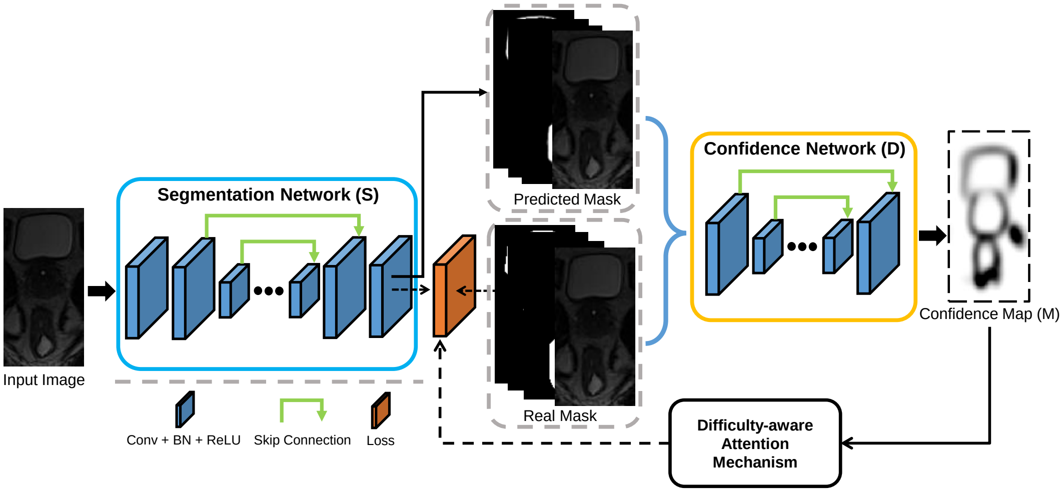


Figure 5.2: Illustration of the architecture of the proposed framework by taking segmentation as an example (although this framework can also be adapted to synthesis). This framework consists of a segmentation network (S), a confidence network (D), and the difficulty-aware attention mechanism. Note, *a perfect* D is desired in this framework.

To ease the description of the proposed algorithm, I give the formal notation used throughout this section. Given a labeled input image $\mathbf{X} \in R^{H\times W\times T}$ with corresponding ground-truth output map (segmentation or output modality) $\mathbf{Y} \in Z^{H\times W\times T}$. For segmentation map, I encode it to one-hot format $\mathbf{P} \in R^{H\times W\times T\times C}$ (by converting the label map Y into C binary label maps with one-hot encoding), where C is the number of semantic categories in the dataset. The base generator network outputs the class probability maps $\widehat{\mathbf{P}} \in R^{H\times W\times T\times C}$. The segmented label map can be obtained by $\widehat{\mathbf{Y}} = \arg\max \widehat{\mathbf{P}}$.

5.1.2 Base Generator Network for Segmentation

Since segmentation and synthesis share the same characteristic as dense prediction, the base generator network for segmentation (synthesis) can be any end-to-end dense prediction network (as shown in Fig. 5.2), such as FCN [124, 145], UNet [164, 48], VNet [139], or DSResUNet [214] (a UNet-like structure with residual learning, element-wise addition of skip connection, and deep supervision). In this study, I adopt an enhanced UNet as the segmentation network. Specifically, I replace all the convolutional layers but the last one with the residual modules [71], apply dilated residual module in the intermediate layers between encoder and decoder (the feature maps with the smallest size) [213], utilize the transformation modules in the long-skip connections [151], inject deep supervision at three scales in the decoder path [136], and propose channel attention module to better fuse the concatenated information from lower layers and higher layers [80].

5.1.2.1 Training Segmentation Network

The class imbalance problem is usually serious in medical image segmentation tasks. To overcome it, I propose using a generalized multi-class Dice loss [176] as the training loss for the proposed segmentation network, as defined below in Eq. 5.1.

$$L_{Dice}\left(\mathbf{X},\mathbf{P};\theta_{\mathbf{S}}\right)=1-2\frac{\sum\limits_{c=1}^{C}\pi_c\sum\limits_{h=1}^{H}\sum\limits_{w=1}^{W}\sum\limits_{t=1}^{T}P_{h,w,t,c}\widehat{P}_{h,w,t,c}}{\sum\limits_{c=1}^{C}\pi_c\sum\limits_{h=1}^{H}\sum\limits_{w=1}^{W}\sum\limits_{t=1}^{T}P_{h,w,t,c}+\widehat{P}_{h,w,t,c}}, \tag{5.1}$$

where π_c is the class balancing weight of category c, and $\theta_{\mathbf{S}}$ contains the parameters of segmentation network. I set $\pi_c = 1/\left(\sum\limits_{h=1}^{H}\sum\limits_{w=1}^{W}\sum\limits_{t=1}^{T}P_{h,w,t,c}\right)^2$. $\widehat{\mathbf{P}}$ is the predicted probability map from the segmentation network: $\widehat{\mathbf{P}} = S\left(\mathbf{X},\theta_{\mathbf{s}}\right)$.

Besides, I also use the multi-category cross entropy loss to form the voxel-wise measurement, as shown in Eq. 5.2.

$$L_{CE}\left(X,Y;\theta_S\right)=-\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{t=1}^{T}\sum_{c=1}^{C}I\left\{Y_{h,w,t,c}\right\}\log\widehat{P}_{h,w,t,c}. \tag{5.2}$$

To this end, the hybrid loss which leverages both losses for training the segmentation network can be concluded as in Eq. 5.3.

$$L_{Hyb}=L_{Dice}+L_{CE}. \tag{5.3}$$

5.1.3 Fully Convolutional Adversarial Confidence Learning

Adversarial learning has been shown to be effective in improving visual perception performance for segmentation and synthesis tasks [125, 142, 84, 161, 149]. In the classic adversarial networks, the discriminator is mostly a CNN-based network with the output probability of an input image belonging to be the real [167]. Obviously, the conventional discriminator only provides a global confidence over the entire image domain, without providing local confidence in the dense map, *i.e.*, voxel-wise confidence. To address this issue, I propose using a UNet-based network to model the discriminator and name it as confidence network for convenience. The output of confidence network (denoted as confidence map (M) with size $H \times W \times T \times 1$) indicates locally whether automatic segmentation (generated image) is similar to the ground-truth segmentation (real image) [112]. I argue that the confidence network can learn the structural information that can be used to regularize the output of base dense prediction network [84, 112]. More importantly, this local discriminator can mitigate the gradient vanishing issue to some degree [66, 143]. In this study, a simplified version of typical UNet [164] is adopted as the architecture of confidence network. Specifically, to save memory, I only keep one convolutional layer at each stage and also half the number of feature maps in the convolution layers across the network except the last one.

5.1.3.1 Training the Confidence Network

The training objective of the confidence network is the summation of binary cross-entropy loss over the image domain, as shown in Eq. 5.4. Here, I use S and D to denote the segmentation and confidence networks, respectively.

$$L_D(\mathbf{X}, \mathbf{P}; \theta_{\mathbf{D}}) = L_{BCE}(D(\mathbf{P}, \theta_{\mathbf{D}}), \mathbf{1}) + L_{BCE}(D(S(\mathbf{X}), \theta_{\mathbf{D}}), \mathbf{0}), \tag{5.4}$$

where

$$\begin{aligned} L_{BCE}\left(\widehat{\mathbf{Q}}, \mathbf{Q}\right) = &-\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{t=1}^{T} Q_{h,w,t} \log\left(\widehat{Q}_{h,w,t}\right) \\ &+ (1 - Q_{h,w,t}) \log\left(1 - \widehat{Q}_{h,w,t}\right), \end{aligned} \tag{5.5}$$

where $\mathbf{X}$ and $\mathbf{P}$ represent the input data and its corresponding manual label map (one-hot encoding format), respectively. $\theta_{\mathbf{D}}$ is network parameters for the confidence network.

5.1.3.2 Adversarial Loss as Realistic Regularization

For segmentation network, the above-mentioned hybrid loss as defined in Eq. 5.3 mainly targets at bringing voxel-level or organ-level match between ground-truth segmentation and automatic segmentation. However, it cannot evaluate the match of the two segmentation's in an overall sense. As a result, I propose using an adversarial loss term from D to work as a *realistic regularization*, which aims at enforcing higher-order spatial consistency between ground-truth segmentation and automatic segmentation in an implicit manner. In particular, the adversarial loss ("ADV") to improve S and fool D can be defined by Eq. 5.6:

$$L_{ADV}\left(\mathbf{X}, \theta_{\mathbf{S}}\right) = L_{BCE}\left(D\left(S\left(\mathbf{X}; \theta_{\mathbf{S}}\right)\right), \mathbf{1}\right), \tag{5.6}$$

5.1.4 Discussion for Selection of Adversarial Loss Functions

There are many well designed loss functions proposed for training the GANs [127]. Among them, the widely used loss functions are classic GAN [60], NSGAN [60], WGAN [5], WGANGP [63] and LSGAN [132], respectively.

Table 5.1: Loss functions in the adversarial learning system: GAN, NSGAN, WGAN, WGANGP and LSGAN.

GAN	Loss for Discriminator	Loss for Generator
GAN	$-E_y \log\left(D\left(y\right)\right) - E_{\hat{y}} \log\left(1 - D\left(\hat{y}\right)\right)$	$E_{\hat{y}} \log\left(1 - D\left(\hat{y}\right)\right)$
NSGAN	$-E_y \log\left(D\left(y\right)\right) - E_{\hat{y}} \log\left(1 - D\left(\hat{y}\right)\right)$	$-E_{\hat{y}} \log\left(D\left(\hat{y}\right)\right)$
WGAN	$E_{\hat{y}} D\left(\hat{y}\right) - E_y D\left(y\right)$	$-E_{\hat{y}} D\left(\hat{y}\right)$
WGANGP	$L_D^{WGAN} + \lambda E_{\hat{y}}(\left\|\nabla D\left(\alpha y + \left(1 - \alpha \hat{y}\right)\right)\right\|_2 - 1)^2$	$-E_{\hat{y}} D\left(\hat{y}\right)$
LSGAN		$-E_{\hat{y}}(D\left(\hat{y} - 1\right))^2$

Table 5.1 presents the basic discriminator and generator loss functions. Since the classic GAN does not impose any prior on the data distribution, its implicit assumption is that GAN could generate samples from any data distributions. To achieve such an effect, the classic GAN implicitly assumes that their discriminator has infinite modeling capacity which can distinguish the distributional consistency between generated and real samples [60, 160]. The non-saturate GAN (NSGAN) also has such an assumption but it instead utilizes a non-saturating loss to generate better gradient signal for the generator. To alleviate the gradient vanishing issues in classic GAN and NSGAN, the authors of

WGAN [5] proposed using Earth-Move distance to build their GAN models with Lipschitz regularity. Similar works are proposed in [63, 132]. The ideas of these works are actually to limit the infinite modeling capacity of the D so that the gradient vanishing issue can be mitigated.

In this study, apart from improving the visual perception with the adversarial learning, I also hope to improve quantitative performance with the supervised generator by utilizing the confidence information from the D. To achieve such goals, I have to retain the infinite modeling capacity of D. Therefore, I select NSGAN as the loss function to train the proposed supervised adversarial learning system.

5.2 Difficulty-Aware Attention Mechanism

5.2.1 Difficulty-Aware Attention Mechanism for Medical Image Segmentation

Focal loss has been shown effective to alleviate the overwhelming effect of easy samples in many computer vision tasks, such as image detection and segmentation [122, 1]. The success of focal loss can be attributed to its strategy that pays more attention on the recognized hard samples (pixels) and less attention to the easy samples. The key point is how to recognize difficult samples (pixels). Focal loss utilizes the predicted probability of a sample as the indicator of the difficulty degree, which may lead to some potential problems in medical image segmentation tasks. Firstly, training may be unstable due to the dominance of a certain class. Secondly, easy and hard samples may also have similar focal weights due to the potential multi-class competition. Thirdly, focal loss only provides pixel-level attention and ignores neighborhood-level attention which is usually important for dense prediction tasks. Lastly, focal loss ignores structural information because it does not consider the original input image of the segmentation network. These potential problems are mostly caused by the fact that the focal loss uses only predicted probability from the segmentation network as the standard to determine whether it is a hard or easy sample. To overcome the above-mentioned problems, I argue that *a more suitable easy-or-hard representer* is needed.

The previously described confidence learning provides a potential solution to better recognize the easy-or-hard samples. The confidence map produced by the confidence network contains the easy-or-hard information. Also, since confidence network is actually a binary classification model, it will avoid the multi-category competition issue. More importantly, the confidence map contains information from both the original input image and the predicted probability mask, and thus it can provide neighborhood and structural information about the easy-or-hard samples (regions).

To this end, I propose a difficulty-aware attention mechanism to better represent the easy-or-hard information. Specifically, I design a difficulty-aware hybrid loss for segmentation using region-level and voxel-level attentions from both predicted probability mask and confidence map. I also propose difficulty-aware mechanism for lesion image synthesis.

5.2.1.1 Difficulty-aware Attention based Segmentation Loss

First, an organ-level attention based generalized Dice loss is proposed to depict the region-level difficulty, as shown in Eq. 5.7 (Note that the loss function (Eq. 7) in the paper [154] should actually be the same with Eq. 5.7).

$$L_{FDice(X,P;\theta_S)} = \frac{1}{C}\sum\nolimits_{c=1}^{C}(1-dsc_c)^{\alpha}\left(1-\frac{\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{t=1}^{T}P_{h,w,t,c}\hat{P}_{h,w,t,c}}{\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{t=1}^{T}\left(P_{h,w,t,c}+\hat{P}_{h,w,t,c}\right)}\right), \quad (5.7)$$

where dsc_c is the average Dice similarity coefficient of a specific category c, e.g., a certain organ or tissue. α is the organ-level attention parameter with a range of $[0,5]$. Following [122], I set α to 2 in this study.

The voxel-level difficulty-aware attention from the confidence map (M) is formulated (based on Eq. 5.2) in Eq. 5.8.

$$L_{FCE}\left(X, Y; \theta_S\right) = -\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{t=1}^{T}\sum_{c=1}^{C} I\left\{Y_{h,w,t}, c\right\} F_{h,w,t} \log \widehat{P}_{h,w,t,c}, \tag{5.8}$$

where

$$F = \left(1 - M\right)^{\beta}, \tag{5.9}$$

where β is the voxel-level attention parameter, and it follows the settings of γ as described above.

Now the difficulty-aware attention mechanism with the hybrid loss can be defined as Eq. 5.10.

$$L_{DamHyb} = L_{FDice} + L_{FCE}. \tag{5.10}$$

With the difficulty-aware hybrid loss in Eq. 5.10, I can pay more attention in the lower confidently (hard) segmented regions. Note, it is different from focal loss which is defined based on the probability map (P) from the segmentation network.

5.2.1.2 Total Loss for Segmentation Network

By summing the above losses, the total loss to train the segmentation network can be defined by Eq. 5.11.

$$L_Seg = L_{DamHyb} + \lambda_1 L_{ADV}, \tag{5.11}$$

where λ_1 is the scaling factor for the regularization term of adversarial learning. It is selected as a very small value (i.e., 0.005 in this case) since it works as soft constraint. In this perspective, the adversarial loss term can be viewed as “variational” regularization term to guarantee the overall realism of the automatic segmentation.

5.2.2 Adversarial Confidence Learning for Lesion Image Synthesis

Adversarial learning has been widely used for medical image synthesis due to its capacity to generate realistic images [148, 197, 207]. However, the quantitative performance cannot be improved as much as qualitative improvement (it can even become worse with adversarial learning in many cases). Especially, for the irregular regions, such as lesion or tumor regions, both the visual perception and the quantitative performance need further improvement even with the conventional adversarial learning. To achieve this goal, I propose to use the similar framework shown in Fig. 5.2 for lesion medical image synthesis.

5.2.2.1 Basic L_p Loss for Reconstruction

As mentioned in the Introduction section, typically an L_1/L_2 loss is conventionally used to train the typical synthesis network as shown in Eq. 5.12.

$$L_G(X,Y) = \|Y - G(X)\|^p , \tag{5.12}$$

where Y is the ground-truth target image, and $G(X)$ is the generated target image from the source image X by the Generator network G and p is 1 or 2.

5.2.2.2 Realistic Regularization with Adversarial Learning

To produce realistic target modality images, Eq. 5.6 is adopted to work as an regularization term. This realistic regularization term drives the objective function of image synthesis to consider the realistic effect in an entire view instead of only optimizing towards the minimal reconstruction error in voxel (pixel) level.

5.2.2.3 Difficulty-aware Attention based L_p Loss

Due to the inhomogeneous characteristics and irregular distribution of the medical images, certain region of the images are usually more difficult to well synthesize. As a consequence, it is quite desired to build a model that can better model the hard-to-prediction regions. Since the local discriminator could provide the dense confidence information about how well each region is synthesized, I can thus pay more attention on the hard-to-predict regions (*e.g.*, lesion regions) so that these regions can be better modeled. To this end, I propose using the above-mentioned adversarial difficulty-aware attention mechanism to better represent the easy-or-hard information. Specifically, I design a difficulty-aware L_1/L_2 loss using region-level attentions from the adversarial local confidence map.

The voxel-level difficulty-aware attention from the confidence map (M) is formulated (based on Eq. 5.12) in Eq. 5.13.

$$L_{AttG}(X, Y) = F \odot \|Y - G(X)\|^p \,, \tag{5.13}$$

where $\odot$ is the element-wise multiplication and

$$F = (1 - M)^{\beta}, \tag{5.14}$$

where β is the voxel-level attention parameter. Note, F here works as a scaling factor, which largely suppresses the contribution of easy-to-synthesize regions to the training loss and emphasizes the hard-to-synthesize regions.

With the difficulty-region-aware L_1/L_2 loss in Eq. 5.13, I can pay more attention in the less confidently (*i.e.*, hard-to-predict) regions and thus better model them (*e.g.*, tumor or lesion regions). As a consequence, this adversarial difficulty-region-aware attention mechanism provides an opportunity to use voxel-wise focal loss in regression context.

5.2.2.4 Total Loss for Training Generator

To this end, the total loss for training generator includes the attention based L_1/L_2 loss, and the local adversarial loss, which can be summarized below Eq. 5.15.

$$L_G = L_{AttG} + \lambda_1 L_{ADV}, \tag{5.15}$$

In this study, the balance coefficient (λ_1) is selected at 0.005. The above training loss could encourage G to generate target images with voxel-wise correspondence to real target image. At the same time, the generated image will be constrained to be as realistic as possible so that it can fool the discriminator.

5.2.3 Implementation Details

Pytorch[3] is adopted to implement the proposed framework shown in Fig. 5.2. Part of the codes are released in the github repositories [4] and [5]. Since we desire a perfect discriminator (D), I do not adopt the traditionally used strategies to limit the D [161]. I adopt Adam algorithm to optimize the networks. For segmentation tasks, the input size of the segmentation network is 64 × 64 × 16. The network weights are initialized by the Xavier algorithm [59] and weight decay is set to be 1e-4. For the network biases, I initialize them to 0. The learning rates for the generator network and the confidence network are both initialized to 5e-3, followed by decreasing the learning rate 2 times for the S, and 5 times for the D every 3 epochs during the training until smaller than 1e-7. For synthesis tasks, the input size is set as 240 × 240 × 5. The network weights are also initialized by the Xavier algorithm [59] and weight decay is set to be 1e-4. The network biases are initialized to 0. The learning rates for the generator network and the confidence network are both initial-

[3] https://github.com/pytorch/pytorch

[4] https://github.com/ginobilinie/medSynthesisV1

[5] https://github.com/ginobilinie/diffAwareAttMedSeg

ized to 5e-4, followed by decreasing the learning rate 2 times for the S, and 5 times for the D every 3 epochs during the training until smaller than 5e-7. Then I use SGD as optimal solver to continue the training until the loss cannot decrease any more. A Titan X GPU server is utilized to train the networks.

5.3 Confidence-Aware Semi-Supervised Learning for Medical Image Segmentation

Semi-supervised algorithms provide potential solutions for the lack of annotated data problem of medical image segmentation. However, it is very difficult to automatically evaluate how well the unlabeled samples are segmented. In addition, previous semi-supervised segmentation models mostly involve the entire unlabeled sample into the learning process, which could actually introduce error signals to mis-supervise the model. Based on the adversarial confidence learning mentioned in Sec. 5.1, this section proposes a novel semi-supervised deep learning framework - "Attention based Semi-supervised Deep Networks" (ASDNet), to overcome these challenges. The proposed ASDNet consists of two subnetworks, *i.e.*, a) segmentation network (denoted as S) and b) confidence network (denoted as D). The architecture of the proposed framework is presented in Fig. 5.3.

Besides the notations described at the end of Sec. 5.1.1, I regard an unlabeled image as $\mathbf{U} \in R^{H \times W \times T}$. Therefore, the whole input image dataset can be defined by $\mathbf{O} = \{\mathbf{X}, \mathbf{U}\}$.

In the following subsections, I first introduce the segmentation network. Then, I describe the confidence network with fully convolutional adversarial learning, followed by the confidence-aware semi-supervised learning strategy. Finally, I describe the implementation details.

5.3.1 Segmentation Network with Sample Attention

In ASDNet as shown in Fig. 5.3, the segmentation network can be any end-to-end segmentation network, such as FCN [124], UNet [164], VNet [139], and DSResUNet [214].

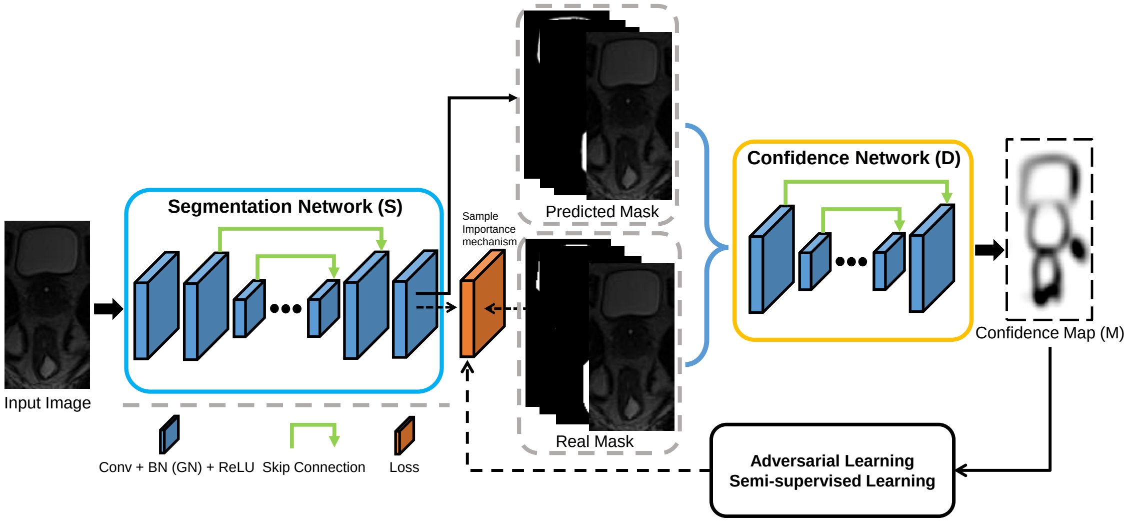


Figure 5.3: Illustration of the architecture of the proposed ASDNet, which consists of a segmentation network and a confidence network.

In this study, I adopt a simplified VNet [139] (SVNet) as the segmentation network to balance the performance and memory cost. Specifically, I remove the 4th downsampling layer, the 1st upsampling layer and the convolution layers between them from the original VNet.

5.3.1.1 Multi-class Dice loss

The class imbalance problem is usually serious in medical image segmentation tasks. To overcome it, I propose using a generalized multi-class Dice loss [176] again as the base training loss for the proposed segmentation network, as defined below in Eq. 5.1.

5.3.1.2 Multi-class Dice loss with Sample Attention

Besides the class imbalance problem, the network optimization also suffers from the issue of dominance by easy samples: the large number of easy samples will dominate network training, thus the difficult samples cannot be well considered. To address this issue, inspired by the focal loss [122] proposed to handle similar issue in detection networks, I propose a sample attention based mechanism to consider the importance of each sample during the training. The multi-class Dice loss with sample attention is thus defined below by Eq. 5.16.

$$L_{AttDice}\left(\mathbf{X},\mathbf{P};\theta_{\mathbf{s}}\right)=(1-dsc)^{\beta}\left(1-2\frac{\sum\limits_{c=1}^{C}\pi_{c}\sum\limits_{h=1}^{H}\sum\limits_{w=1}^{W}\sum\limits_{t=1}^{T}P_{h,w,t,c}\widehat{P}_{h,w,t,c}}{\sum\limits_{c=1}^{C}\pi_{c}\sum\limits_{h=1}^{H}\sum\limits_{w=1}^{W}\sum\limits_{t=1}^{T}P_{h,w,t,c}+\widehat{P}_{h,w,t,c}}\right), \tag{5.16}$$

where dsc is the average Dice similarity coefficient of the sample over different categories, e.g., different organ labels. Note that I re-compute the dsc in each iteration, but I don't back-propagate gradient through it when training the networks. β is the sample attention parameter with a range of $[0,5]$. Following [122], I set β to 2 in this study.

The scaling factor of $(1-dsc)^{\beta}$ largely suppresses the contribution of easy-to-segment samples to the training loss (*e.g.*, when $dsc=0.9$, the scaling factor is 0.01). It can also lightly suppress the contribution of hard-to-segment samples (*e.g.*, when $dsc=0.1$, the scaling factor is 0.81). Therefore, the sample attention mechanism can adaptively shift the training focus to hard-to-segment samples and address the issue of dominance by easy samples.

5.3.2 Confidence Network for Fully Convolutional Adversarial Learning

Following the adversarial confidence learning in Sec. 5.1, we employ a FCN to work as the discriminator to conduct the adversarial learning. The training objective of the confidence network is once again the summation of binary cross-entropy loss over the image domain, as shown in Eq. 5.4. The adversarial loss for training the segmentation networks is the same as Eq. 5.6.

5.3.3 Confidence-aware Semi-supervised Learning

The above adversarial loss can enforce the distribution of segmented probability maps to follow the distribution of the ground-truth label map of real data, which is similar to the usage of adversarial loss in the conventional GAN setting [60]. Different from GAN, the proposed discriminator (*i.e.*, confidence network) provides local confidence information over the image domain. As I will see below, this information can be used in the semi-supervised setting to include unlabeled data for improving segmentation accuracy, and the similar strategy has been explored in [84].

Specifically, given an unlabeled image $\mathbf{U}$, the segmentation network will first produce the probability map $\widehat{\mathbf{P}} = S(\mathbf{U})$, which will be then used by the trained confidence network to generate a confidence map $\mathbf{M} = D(\widehat{\mathbf{P}})$, indicating where are the confident regions with the prediction results close enough to the ground truth label distribution. The confident regions can be easily obtained by setting a threshold (*i.e.*, γ) to the confidence map. In this way, I can use these confident regions as masks to select parts of unlabeled data and their segmentation results to enrich the set of supervised training data. Thus, the proposed semi-supervised loss can be defined by Eq. 5.17.

$$L_{semi}(\mathbf{U}, \theta_{\mathbf{s}}) = 1 - 2\frac{\sum\limits_{c=1}^{C} \pi_c \sum\limits_{h=1}^{H} \sum\limits_{w=1}^{W} \sum\limits_{t=1}^{T} [\mathbf{M} > \gamma]_{h,w,t} \overline{P}_{h,w,t,c} \widehat{P}_{h,w,t,c}}{\sum\limits_{c=1}^{C} \pi_c \sum\limits_{h=1}^{H} \sum\limits_{w=1}^{W} \sum\limits_{t=1}^{T} [\mathbf{M} > \gamma]_{h,w,t} \left(\overline{P}_{h,w,t,c} + \widehat{P}_{h,w,t,c}\right)}, \tag{5.17}$$

where $\overline{\mathbf{P}}$ is the one-hot encoding of $\widehat{\mathbf{Y}}$, and $\widehat{\mathbf{Y}} = \arg\max(\widehat{\mathbf{P}})$. [] is the indicator function. Similar to *dsc* in Eq. 5.16, $\overline{\mathbf{P}}$ and the value of indicator function are re-computed in each iteration, and I don't back-propagate gradients through them during network training. With such a setup, the semi-supervised loss in Eq. 5.17 can actually be seen as a masked multi-class Dice loss.

5.3.3.1 Total Loss for Segmentation Network

By summing the above losses, the total loss to train the segmentation network can be defined by Eq. 5.18.

$$L_S = L_{AttDice} + \lambda_1 L_{ADV} + \lambda_2 L_{semi}, \tag{5.18}$$

where λ_1 and λ_2 are the scaling factors to balance the losses. They are selected at 0.03 and 0.3 after trails, respectively.

5.3.4 Implementation Details

Pytorch[6] is adopted to implement the proposed ASDNet shown in Fig. 5.3. The code can be obtained by this link[7]. We adopt Adam algorithm to optimize the network. The input size of the segmentation network is $64 \times 64 \times 16$. The network weights are initialized by the Xavier algorithm, and weight decay is set to be 1e-4. For the network biases, I initialize them to 0. The learning rates for the segmentation network and the confidence network are both initialized to 1e-3, followed by decreasing the learning rate 10 times every 3 epochs during the training. Four Titan X GPUs are utilized to train the networks.

5.4 Experiments for Segmentation Tasks

To evaluate the proposed method, I apply the proposed algorithm on two different datasets. The first dataset is our own pelvic dataset and the other one is a publicly available challenge dataset which will be introduced in later subsection.

The pelvic dataset consists of 50 prostate cancer patients from a Cancer Hospital, each with one T2-weighted MR image and its corresponding manually-labeled map by a medical expert. The images were acquired with 3T magnetic field strength, while different patients were scanned with different MR image scanners (i.e., Siemens Medical Systems and Philips Medical Systems). Under such a situation, the challenge for the segmentation task increases since both shape and appearance differences are large. The prostate, bladder, and rectum in all MRI scans have been manually segmented, which serve as the ground truth for evaluating the proposed segmentation method. The image size is mostly $256 \times 256 \times (120 \sim 192)$, and the voxel size is mainly $1 \times 1 \times 1$ mm^3.

[6] https://github.com/pytorch/pytorch

[7] https://github.com/ginobilinie/asdnet

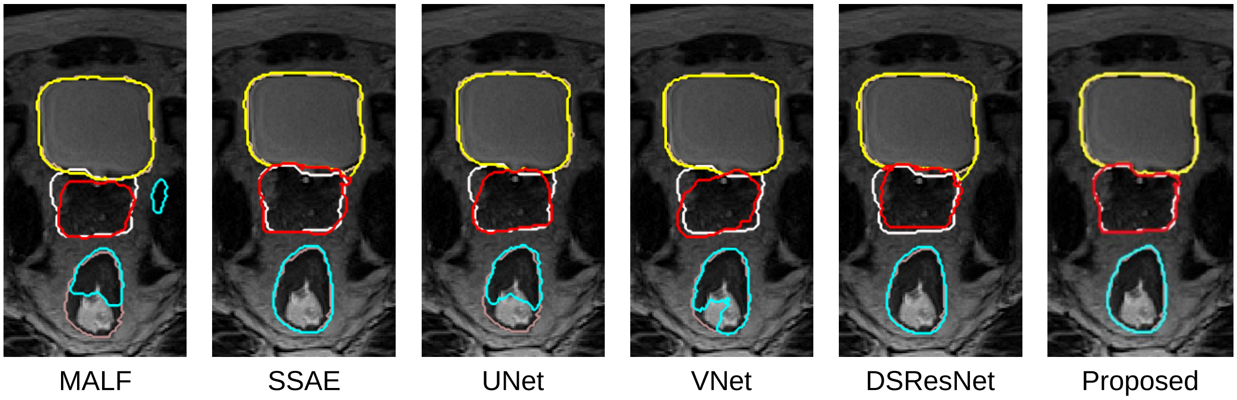


Figure 5.4: Pelvic organ segmentation results of a typical subject by different methods. Orange, silver and pink contours indicate the manual ground-truth segmentations, and yellow, red and cyan contours indicate automatic segmentations.

Five-fold cross-validation is used to evaluate the proposed method. Specifically, in each fold of cross-validation, I randomly chose 35 subjects as the training set, 5 subjects as the validation set, and the remaining 10 subjects as the testing set. I use sliding windows to go through the whole MRI for prediction for a testing subject. Unless explicitly mentioned, all the reported performance by default is evaluated on the testing set. As for evaluation metrics, I utilize DSC (Eq. 2.22) and ASD (Eq. 2.24) to measure the agreement between the manually and automatically segmented label maps.

5.4.1 Comparison with State-of-the-art Methods

To demonstrate the advantage of the proposed method, I compare the proposed method with other five widely-used methods on the same dataset as shown in Table 5.2: 1) multi-atlas label fusion (MALF), 2) SSAE [64], 3) UNet [164], 4) VNet [139], and 5) DSResUNet [214]. Also, I present the performance of the proposed method.

We visualize some typical segmentation results in Fig. 5.4, which can show the superiority of the proposed method, especially for the hard-to-segment regions, *i.e.*, prostate and rectum. I also present the quantitative comparison of the proposed method with the five state-of-the-art segmentation methods in Table 5.2. I can see that the proposed method achieves better accuracy than the five state-of-the-art methods in terms of both DSC and

Table 5.2: DSC and ASD on the pelvic dataset by different methods.

Method	DSC (%)			ASD (in mm)		
	Bladder	Prostate	Rectum	Bladder	Prostate	Rectum
MALF	86.69(6.81)	79.28(8.72)	76.43(11.88)	1.641(.360)	2.791(.930)	3.210(2.112)
SSAE	91.75(3.10)	87.07(4.24)	86.38(4.41)	1.089(.231)	1.660(.490)	1.701(.412)
UNet	89.57(2.83)	82.22(5.88)	81.04(5.31)	1.214(.216)	1.917(.645)	2.186(0.850)
VNet	92.61(1.84)	86.40(3.61)	83.16(4.12)	1.023(.186)	1.725(.457)	1.969(.449)
DSResUNet	94.43(.90)	88.24(2.01)	86.91(3.24)	.914(.168)	1.586(.358)	1.586(.405)
Proposed	**97.48(.65)**	**92.11(1.70)**	**91.05(2.47)**	**.850(.146)**	**1.297(.276)**	**1.387(.346)**

Table 5.3: Quantitative comparison between the proposed method and other methods on the prostate challenge testing dataset.

Method	DSC (%)			ASD (in mm)			95HD			aRVD			Score(std)
	whole	base	apex	whole	base	apex	whole	base	apex	whole	base	apex	
pxl_mcg	91.23	89.07	88.54	1.60	1.76	1.57	4.47	4.48	3.64	2.08	-0.07	2.23	88.98(3.41)
Isensee	91.61	90.29	88.05	1.52	1.65	1.64	4.21	4.20	3.85	3.42	1.86	3.48	88.84(2.94)
whu_mlgroup(2)	91.42	89.41	88.51	1.54	1.79	1.57	4.21	4.88	3.82	5.27	4.00	6.43	88.72(4.36)
Proposed	90.12	88.95	87.71	1.84	1.73	1.68	5.36	4.43	3.99	4.99	2.19	6.65	88.28(3.02)
tbrosch	90.46	88.51	85.29	1.70	1.91	1.90	4.91	5.04	4.57	2.14	7.22	-4.93	87.24(4.46)
whu_mlgroup(1)	90.26	89.15	88.36	1.86	1.79	1.62	5.57	4.83	3.90	9.74	10.73	9.64	87.04(5.79)
AutoDenseSeg	90.14	88.09	86.79	1.83	1.94	1.79	5.36	5.13	4.32	4.53	5.19	2.05	87.19(4.25)
CUMED	89.43	86.42	86.81	1.95	2.13	1.74	5.54	5.41	4.29	6.95	11.04	15.18	86.65(4.42)
SCIRESU	90.24	88.98	83.30	1.74	1.81	2.11	4.93	4.51	5.34	6.01	8.18	-7.33	86.41 (3.49)
QUILL(M2)	88.81	87.39	85.46	1.97	2.01	1.91	5.29	5.07	4.35	6.97	4.76	5.85	85.93(4.97)

ASD, especially for the prostate and rectum which are believed more difficult to segment. In contrast, the VNet works well in segmenting bladder and prostate, but it cannot work very well for rectum (which is often more challenging to segment due to the long and narrow shape). DSResUNet presents good performance in the bladder and rectum regions, but it cannot well model the prostate region which is the most difficult but important region. More importantly, thanks to the adversarial confidence learning framework, the quantitative performance gain can now align with the visual perception improvement.

5.4.2 Impact of the Realistic Regularization

To investigate how adversarial learning helps the segmentation model, I visually check two typical subjects in Fig. 5.5. In Fig. 5.5(a), enUNet gives similar (or a little better) segmentation results as enUNet+LocalD, which means adversarial learning can still provide subtle improvement even if enUNet has already produced very similar organs to the manual ground truth. In Fig. 5.5(b), I can clearly see that adversarial learning has corrected obvious errors in the segmented organs by enUNet. To summarize, adversarial learning can supply reasonable visual perception improvement, especially when the results from the segmentation model have obvious structural errors.

On the other hand, with the adversarial learning, the DSC values are only improved by 0.2%, 0.2% and 0.1% in average for bladder, prostate and rectum, respectively. The experimental results indicate that adversarial learning based realistic regularization can contribute to performance gain but in a subtle manner which does not correspond to the visual perception improvement. I argue adversarial learning provides a way of minimizing the "variational" loss by enforcing higher-order consistency between ground-truth segmentations and automatic segmentations. As a result, the visual perception performance can improve in a larger degree due to such an emphasis on the image-level similarity. While the quantitative performance, for instance, DSC, is actually included by the original objective function of the segmentation network, it thus cannot benefit as much as visual perception performance.

To further explore the effectiveness of the realistic regularization, I ask a physician to select the segmentations from UNet and UNet with realistic regularization (Note, the physician does not know which method produced the segmentations beforehand). About 65% of segmentations chosen by the physician are those segmented by UNet with realistic regularization, which validates that realistic regularization can improve visual perception for medical image segmentation.

5.4.3 Impact of the Difficulty-aware Attention Mechanism

As mentioned in the Method Section, I propose an enhanced UNet with several widely used techniques injected, and I further propose a difficulty-aware attention mechanism to assign different importance for different samples (regions) so that the network can concentrate on hard-to-segment examples and thus avoid dominance by easy-to-segment samples.

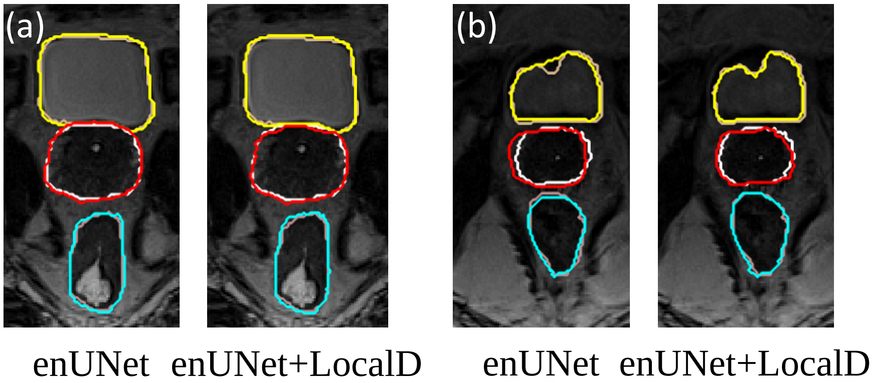


Figure 5.5: Visual inspection of segmentation improvements by adversarial learning on two different cases. Here, enUNet means the proposed networks without adversarial learning, and enUNet+LocalD means the proposed networks with adversarial learning. In (a), adversarial learning does not help much, as enUNet already gives good results. In (b), adversarial learning can help to correct the segmented organs obviously, due to large segmentation errors by enUNet.

I visualize the performance comparison among the basic UNet, enUNet and the one with difficulty-aware attention mechanism (enUNet+dam) in Fig. 5.6. (Note, I use the hybrid loss to train UNet and enUNet). Actually, in this case, the enhancement for the UNet with certain modules as introduced before contribute most to the performance gain. The effectiveness of difficulty-aware attention mechanism is also confirmed by the improved performance as shown in Fig. 5.6. It is worth noting that the proposed difficulty-aware attention mechanism contributes more performance gain for prostate and rectum compared with the bladder. It is consistent with the proposed assumption that difficulty-aware attention mechanism could pay more attention to difficult samples (regions) and thus can handle difficult samples (regions) much better.

5.4.3.1 Comparison with the Focal Loss

Since the proposed difficulty-aware attention mechanism is designed based on the focal loss, it is interesting to explore the difference of the proposed module against focal loss for medical image segmentation.

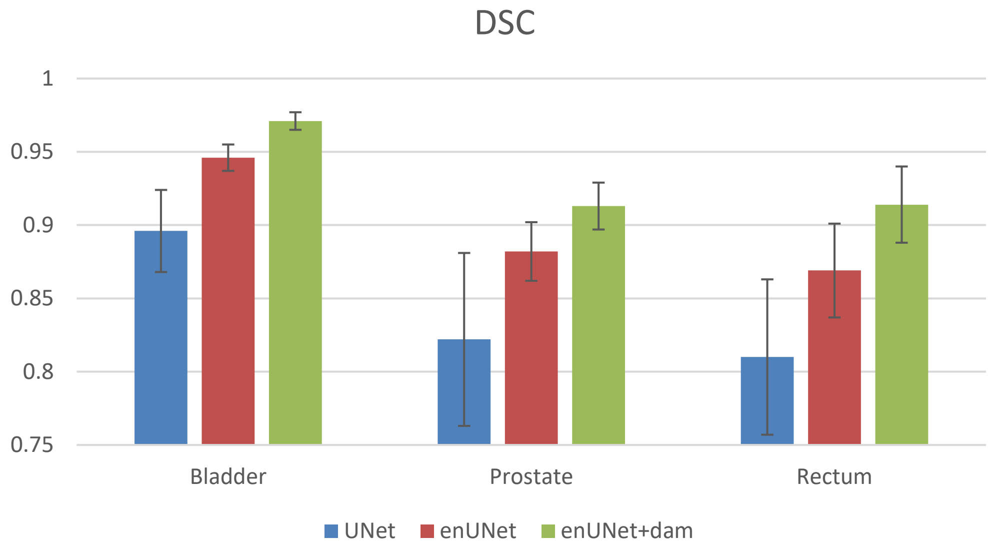


Figure 5.6: Average Dice ratios of different methods.

To better understand the two strategies, I first visualize the difficulty-aware mask (i.e., $(1-M)$) and the focal mask (i.e., $\left(1-\widehat{P}\right)$) in Fig. 5.7. The focal mask mainly focuses on the regions with low predicted probability from segmentation network which needs more attention. Since it is directly related with predicted probability map, it can reflect the difficult regions more precisely in *voxel-level*. On the contrary, difficulty-aware mask reflects the difficulty regions in a more *structured* manner, in which it focuses more on the regions with lower confidence ratios from confidence network. The reason behind it is that I have a professional hard-or-easy recognizer: The D can represent the input containing both the predicted probability mask from segmentation network and the original input image by confidence learning so that I can have a more expert hard-or-easy representation, as expressed in Eq. (5.19):

$$M = D(\widehat{P} \cup X), \tag{5.19}$$

where $\cup$ denotes the concatenation operation.

I further conducted experiments with these different strategies to segment the prostate only, since the prostate is traditionally thought to be hard to segment. To make a fair comparison, I use the same architecture (enUNet) as the basis to conduct the experiments.

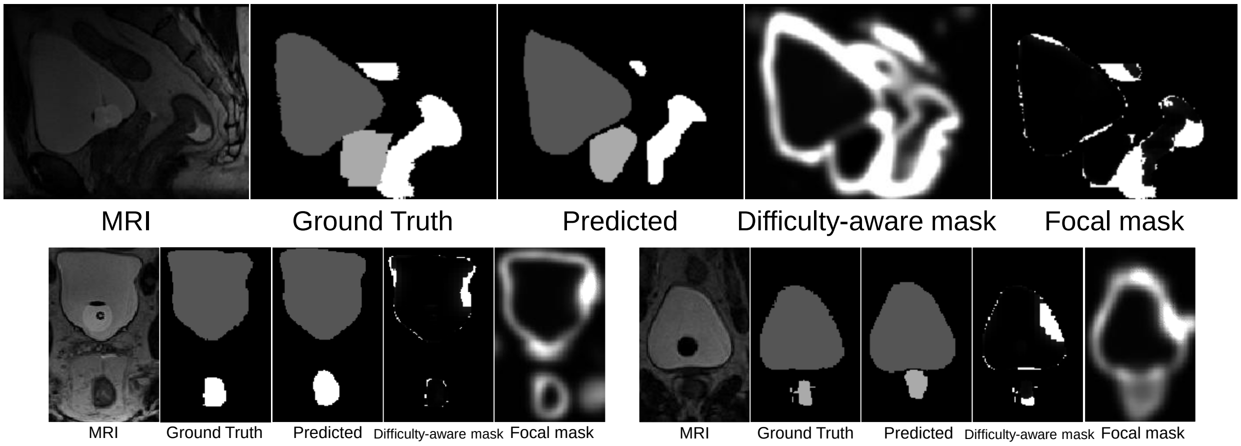


Figure 5.7: Visualization of the difficulty-aware mask and the focal mask, obtained after training the network for 5 epochs. The first row is the sagittal view. The second row contains the axial and coronal views.

Table 5.4: Comparison of different strategies in segmenting prostate on the pelvic dataset in terms of DSC (%).

Method	Base	Middle	Apex
enUNet	86.70(4.91)	87.91(4.83)	83.92(5.87)
enUNet+Focal	88.24(4.53)	89.21(3.20)	86.83(4.90)
enUNet+Hybrid	88.12(4.19)	90.08(2.70)	86.71(5.47)
Proposed	89.41(3.68)	90.90(2.37)	88.21(4.14)

Due to computational times, I only do a two-fold cross-validation for these comparison experiments. To better depict the difficult parts of the prostate, I partition the prostate into three parts: apex (first 1/3 of the prostate volume), base (last 1/3 of the prostate volume) and middle (the rest). The performance of the enUNet with different strategies is listed in Table 5.4.

As described in Table 5.4, the focal loss can help improve the performance, especially for the base and apex parts of the prostate, since it pays more attention to the hard voxels. The hybrid loss described in Eq. 5.3 can achieve similar performances with the focal loss since the hybrid loss can capture the organ structure as well as the voxel-level information. The proposed method (difficulty-aware attention mechanism) achieves the largest performance gain, since it can not only capture the difficult regions in a structured way but also absorb the advantage of the hybrid loss. This demonstrates that the proposed difficulty-aware attention mechanism can work better than the focal loss in medical image segmentation tasks.

5.4.4 Validation on Prostate Challenge Dataset

I have also evaluated the proposed method on the prostate segmentation challenge dataset[8]. The ground-truth label maps for the testing set are hidden from the participants. The official evaluation metrics used in this challenge include the DSC, the average over the shortest distance between the boundary (surface) points of the volumes (ABD or ASD), the percentage of the absolute difference between the volumes (aRVD), and the 95% Hausdorff distance (95HD). It is worth noting that the organizers not only calculate the evaluation metrics on the whole prostate, but also on the apex and base parts of the prostate that are believed to be the most difficult regions for segmentation. Besides, an overall score (shown as the last column in Table 5.3) combining the above-mentioned evaluation metrics is also provided to rank the submitted methods (please refer to [123] for the details about the evaluation metrics).

The quantitative results of the proposed method and our competitors are shown in Table 5.3. (Note, the results were directly obtained from the organizers based on my submission in Sep. 2018). There were more than 150 teams successfully submitting their results and being listed in the leaderboard at that time. Note I only list top 10 teams in the Table for convenience, and please refer the entire leaderboard through this link[9]. The proposed method ranks 4^{th} in terms of the overall score among all the 150 participants. It is worth noting that the top 3 methods all ensemble their results from different deep networks. In contrast, my submission is a single model as presented in this study. More importantly, my proposed method presents a much lower standard deviation value compared to the other top 8 methods. (Note, the minimum standard deviation comes from

[8] https://promise12.grand-challenge.org/

[9] https://promise12.grand-challenge.org/evaluation/results/

the 2nd ranked team who has assembled results from 20 segmentation networks), which further indicates the effectiveness and robustness of my proposed method.

It is interesting to note that my proposed method achieves a very competitive performance on the base and apex parts which are usually thought to be the most difficult segmented regions, and it further validates that the proposed difficulty-aware attention mechanism indeed contributes to the performance gain.

5.5 Experiments on Synthesis Tasks

I choose the BRATS dataset to evaluate the proposed method, which is a publicly available dataset of MRI from brain tumor patients [135]. A total of 354 pairs of T1 MRI and T2 MRI were assembled, where 200 subjects were used for training and 60 for validation, and the rest 94 for testing.

To demonstrate the advantage of the proposed method in terms of synthesis accuracy, I compare it with four widely-used approaches: atlas-based, FCN, UNet [65], UNet with CNN-based global adversarial learning (UNet+GlobalD or AdUNet) [150], and UNet with FCN-based local adversarial learning (UNet+LocalD) [112]. For fair comparison, I use the UNet to work as the image synthesis network (generator). I adopt the widely used MAE (Eq. 2.25), PSNR (Eq. 2.26) and SSIM (Eq. 2.28) as the quantitative evaluation metrics.

5.5.1 Impact of Realistic Regularization

To explore the contribution of realistic regularization, comparison experiments are conducted among the UNet, UNet+GlobalD and UNet+LocalD on the BRATS dataset. The visual comparison is shown in Fig. 5.9. Obviously, both the global and local adversarial learning can largely improve the visual perception performance.

Besides the above-mentioned qualitative comparison, a quantitative measurement is designed to further investigate the realistic effect of adversarial learning. Specifically, 100 slices are randomly sampled from the ground-truth images. Then the corresponding 100

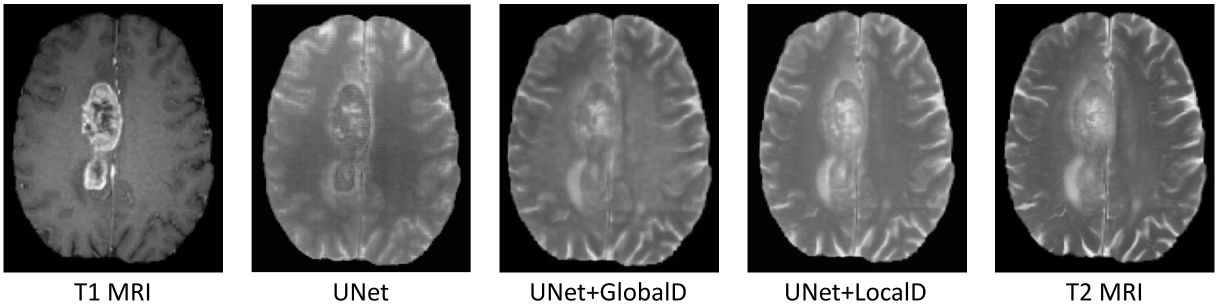


Figure 5.8: Visual evaluation of the effect of the realistic regularization. Using the proposed realistic regularization, the respective results (third column) looks more similar to the real target T2 MRI (fourth column), compared to the case without using the mechanism.

slices are sampled from the corresponding synthetic images by UNet and UNet+GlobalD, respectively. Next, two selection games are designed: a radiologist is asked to select 'real' slice between ground-truth slice and synthetic slice by UNet; the radiologist is also asked to select 'real' slice from the ground-truth slice and synthetic slice by UNet+GlobalD (Note, the radiologist does not know the ground-truth images beforehand). As a result, 12% of the UNet based synthetic slices are chosen by the radiologist, in other words, 12% of the synthetic slices could confuse the expert (*i.e.*, the confusion rate is 12%). In contrast, 32% of the UNet+GlobalD based synthetic slices are chosen the by the radiologist (*i.e.*, the confusion rate is 32%). Two facts can be observed from the experiment. a) The synthetic slices cannot still work as well as the ground-truth slices (both of the confusion rates do not overcome 50%). b) The adversarial learning can largely improve the visual effect of synthetic MRI, which means adversarial learning works as realistic regularization for medical image synthesis by improving visual perception.

The PSNR values are 26.2dB, 25.9dB and 26.0dB in average for these three methods, respectively. The average SSIM values are 0.862, 0.871 and 0.873 for these three methods, respectively. Note that these results are achieved with the ordinary L_1 loss for the generator. Compared to UNet, the adversarial learning seems not able to improve the quantitative performance in terms of PNSR. It can only improve a little bit the SSIM values with adversarial learning though the SSIM evaluates the structural similarity. This is

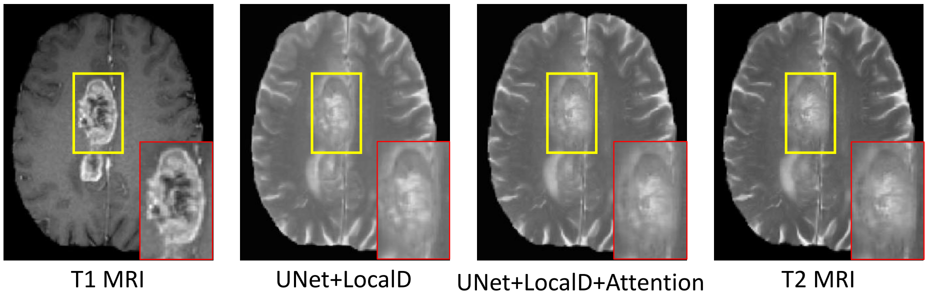


Figure 5.9: Visual evaluation of the proposed difficult-region-aware attention mechanism. Using the proposed mechanism, the respective results (third column) is more similar to the real target T2 MRI (fourth column), compared to the case without using the proposed mechanism (second column).

consistent with the objective functions of these three methods, since UNet only optimizes towards minimizing the L_1 loss which actually directly maximizes the PSNR, while UNet with adversarial learning is also constrained by the realistic regularization.

5.5.2 Impact of Difficulty-Region-Aware Attention Mechanism

To show the impact of the proposed difficult-region-aware attention mechanism, I first conduct experiments to compare the performance for cases with/without this mechanism on the BRATS dataset. The experimental results indicate that the performance could be improved by 0.8dB in terms of PSNR using the proposed attention mechanism. To further investigate the impact of the proposed mechanism, I focus on evaluating the synthesis performances only on tumor regions. By using the manually segmented tumor regions in this database, I compute PSNR on tumor regions of testing set, obtaining 1.2dB performance gain in average.

We also visualize results in Fig. 5.9. I can clearly see that the generated image by using the proposed difficult-region-aware attention mechanism (*i.e.*, 'UNet+LocalD+Attention') could recover much more details, compared to the results without using the proposed mechanism (*i.e.*, 'UNet+LocalD'), especially for the tumor regions.

To better understand why the difficult-region-aware mechanism works, I also analyze the confidence map generated by the local discriminator (*i.e.*, LocalD). I find that, initially, the tumor regions are evaluated to be poorly synthesized as indicated by local confidence, and thus more attention is paid to tumor regions in later training of the generator network. In the end of training, tumor regions can also be better synthesized.

5.5.3 Comparison with Other Methods

To qualitatively compare the image synthesis results by different methods, I show synthetic target image, along with real target image, in Fig. 5.10. I can see that the proposed algorithm can better preserve the continuity, coalition and smoothness in the synthetic results, since it uses both global and local adversarial learning constraints in the image patch. More importantly, the tumor region of the generated T1 MRI can recover much more details than other methods, and thus looks much closer to the real T2 MRI compared to all other methods. I argue that this is due to the difficult-region-aware attention mechanism which reweight more on the recognized hard-to-synthesis regions, *i.e.*, tumor regions.

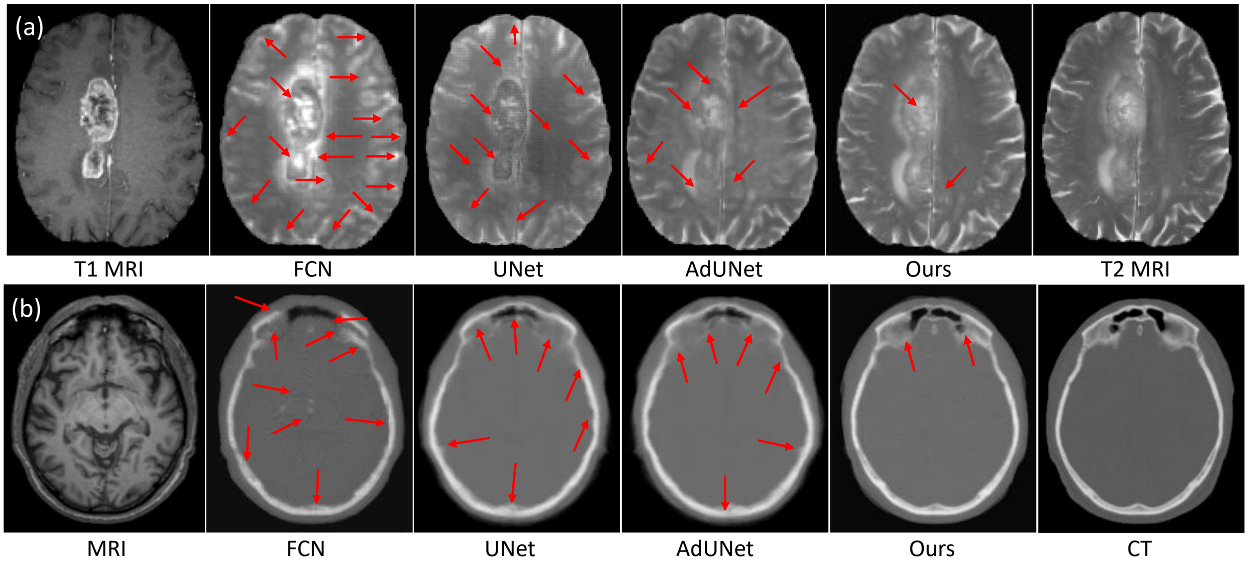


Figure 5.10: Visual comparison of results by different methods for two cases of application: (a) T1 MRI to T2 MRI synthesis, and (b) MRI to CT synthesis. Red arrows indicate poorly-synthesized regions.

Table 5.5: Average MAE, PSNR and SSIM on 94 testing subjects from the BRATS dataset.

Method	MAE	PSNR	SSIM
FCN	34.5(8.6)	25.0(2.3)	.785(.014)
UNet	28.8(6.9)	26.2(1.8)	.862 (.009)
AdUNet	29.1(5.7)	26.0(**1.5**)	.878 (.009)
Ours	**27.3**(5.2)	**26.9**(1.6)	**.913(.008)**

Table 5.6: Average MAE, PSNR and SSIM on 16 subjects from the brain dataset.

Method	MAE	PSNR	SSIM
FCN	24.4(15.1)	22.7(3.2)	.834(.018)
UNet	21.8(12.8)	26.7(2.1)	.908(.010)
AdUNet	21.9(11.3)	26.8(**1.7**)	.914(.011)
Ours	**20.8(10.8)**	**27.3**(1.8)	.932(.009)

We also quantitatively compare the predicted results in Table 5.5, in terms of both PSNR and MAE. The proposed method outperforms all other competing methods in both metrics. It is worth noting that the quantitative performance cannot improve much with only adversarial learning (it may even become worse), while my method can improve both the quantitative performance and the qualitative performance due to characteristic of the proposed adversarial confidence learning.

Fig. 5.10(a) shows synthesis results on BRATS dataset (with brain tumors) by different methods. It can be seen that the result by the proposed method is more consistent with the real T2 MRI (right).

To show the generalization ability of my proposed method, I also evaluate it on another brain dataset for synthesizing CT from MRI. Fig. 5.10(b) shows CT synthesis results by different methods, and Table 5.6 gives quantitative comparison results. It is clear that the proposed method can work better than the state-of-the-art methods, demonstrating the good generalization of the proposed method to other datasets for other image synthesis tasks.

5.6 Experiments for Confidence-aware Semi-Supervised Learning

The pelvic dataset for experiments consists of 50 prostate cancer patients from a cancer hospital, each with one T2-weighted MR image and corresponding manually-annotated label map by medical experts. In particular, the prostate, bladder and rectum in all these MRI scans have been manually segmented, which serve as the ground truth for evaluating the proposed segmentation method. Besides, I have also acquired 20 MR images from additional 20 patients, without manually-annotated label maps. All these images were acquired with 3T MRI scanners. The image sizes are mostly $256 \times 256 \times (120 \sim 176)$, and the voxel size is $1 \times 1 \times 1$ mm^3. We first conduct a rigid registration for aligning all images to the same space [91]. A typical example of the MR image and its corresponding label map are given in Fig. 5.12(a).

Five-fold cross validation is used to evaluate the proposed method. Specifically, in each fold of cross validation, I randomly chose 35 subjects as training set, 5 subjects as validation set, and the remaining 10 subjects as testing set. I use sliding windows to go through the whole MRI for prediction for a testing subject. Unless explicitly mentioned, all the reported performance by default is evaluated on the testing set. As for evaluation metrics, I utilize DSC (Eq. 2.22) and ASD (Eq. 2.24) to measure the agreement between the manually and automatically segmented label maps.

5.6.1 Comparison with State-of-the-art Methods

To demonstrate the advantage of the proposed method, I also compare the proposed method with other five widely-used methods on the same dataset as shown in Table 5.7: 1) multi-atlas label fusion (MALF), 2) SSAE [64], 3) UNet [164], 4) VNet [139], and 5) DSResUNet [214]. Also, I present the performance of the proposed ASDNet.

Table 5.7 quantitatively compares the proposed method with the five state-of-the-art segmentation methods. We can see that the proposed method achieves better accuracy

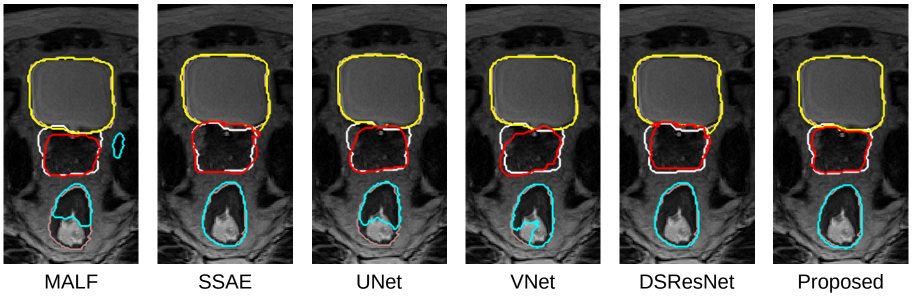


Figure 5.11: Pelvic organ segmentation results of a typical subject by different methods. Orange, silver and pink contours indicate the manual ground-truth segmentation, and yellow, red and cyan contours indicate automatic segmentation.

Table 5.7: DSC and ASD on the pelvic dataset by different methods.

Method	DSC			ASD		
	Bladder	Prostate	Rectum	Bladder	Prostate	Rectum
MALF	.867(.068)	.793(.087)	.764(.119)	1.641(.360)	2.791(.930)	3.210(2.112)
SSAE	.918(.031)	.871(.042)	.863(.044)	1.089(.231)	1.660(.490)	1.701(.412)
UNet	.896(.028)	.822(.059)	.810(.053)	1.214(.216)	1.917(.645)	2.186(0.850)
VNet	.926(.018)	.864(.036)	.832(.041)	1.023(.186)	1.725(.457)	1.969(.449)
DSResUNet	.944(.009)	.882(.020)	.869(.032)	.914(.168)	1.586(.358)	1.586(.405)
Proposed	**.970(.006)**	**.911(.016)**	**.906(.026)**	**.858(.144)**	**1.316(.288)**	**1.401(.356)**

than the five state-of-the-art methods in terms of both DSC and ASD. The VNet works well in segmenting bladder and prostate, but it cannot work very well for rectum (which is often more challenging to segment due to the long and narrow shape). Compared to UNet, DSResUNet improves the accuracy by a large margin, indicating that residual learning and deep supervision bring performance gain, and thus it might be a good future direction for us to further improve my proposed method. We also visualize some typical segmentation results in Fig. 5.11, which further show the superiority of the proposed method.

5.6.2 Ablation Study

As the proposed method consists of several designed components, I conduct ablation studies below to analyze them.

5.6.2.1 Impact of Sample Attention

As mentioned in Sec. 5.3.1, we propose a sample attention mechanism to assign different importance for different samples so that the network can concentrate on hard-to-segment examples and thus avoid dominance by easy-to-segment samples. The effectiveness of sample attention mechanism (i.e., used in the SVNet network, thus namely AttSVNet) is further confirmed by the improved performance, *e.g.*, 0.82%, 1.60% and 1.81% DSC performance improvements (as shown in Table 5.8) for bladder, prostate and rectum, respectively.

5.6.2.2 Impact of Fully Convolutional Adversarial Learning

We conduct more experiments for comparing with the following three networks: 1) only segmentation network; 2) segmentation network with a CNN-based discriminator [60]; 3) segmentation network with a FCN-based discriminator (*i.e.*, confidence network). Performance in the middle of Table 5.8 indicates that adversarial learning contributes a little bit to improving the results as it provides a regularization to prevent overfitting. Compared with CNN-based adversarial learning, the proposed FCN-based adversarial learning further improves the performances by 0.90% in average. This demonstrates that fully convolutional adversarial learning works better than the typical adversarial learning with a CNN-based discriminator, which means the FCN-based adversarial learning can better learn structural information from the distribution of ground-truth label map.

5.6.2.3 Impact of Confidence-aware Semi-supervised Learning

I apply the semi-supervised learning strategy with the proposed ASDNet on 50 labeled MRI and 20 extra unlabeled MRI. The comparison methods are semiFCN [13] and semiEmbedFCN [14]. I use the AttSVNet as the basic architecture of these two methods for fair comparison. The evaluation of the comparison experiments are all based on the

Table 5.8: Comparison of the performance of methods with different strategies on the pelvic dataset in terms of DSC.

Method	Bladder	Prostate	Rectum
VNet	.926(.018)	.864(.036)	.832(.041)
SVNet	.920(.015)	.862(.037)	.844(.037)
AttSVNet	.931(.010)	.878(.028)	.862(.034)
AttSVNet+CNN	.938(.010)	.884(.026)	.874(.031)
AttSVNet+FCN	.944(.008)	.893(.022)	.887(.025)
semiFCN	.959(.006)	.895(.024)	.885(.030)
semiEmbedFCN	.964(.007)	.902(.022)	.891(.028)
AttSVNet+Semi	.937(.012)	.878(.036)	.865(.041)
Proposed	**.970(.006)**	**.911(.016)**	**.906(.026)**

labeled dataset, and the unlabeled data involves only in the learning phase. The experimental results in Table 5.8 show that the proposed semi-supervised strategy works better than the semiFCN and the semiEmbedFCN. Moreover, it is worth noting that the adversarial learning on the labeled data is important to the proposed semi-supervised schema (*i.e.*, Proposed or AttSVNet+Ad+Semi). If the segmentation network does not seek to fool the discriminator (*i.e.*, AttSVNet+Semi), the confidence maps generated by the confidence network would not be meaningful.

To further investigate how the semi-supervised loss works, I visualize the confidence map generated by the confidence network in Fig. 5.12(b). Obviously, the unsure regions (black) are mostly around the organ boundaries, and the confirmed regions (bright) are mainly the background and inside the organs. It demonstrates that the confidence network can approximately indicate the trustworthy regions and can thus guide the training of segmentation network, since I can then use the highly-confidently-segmented regions as the supervision signals to train the segmentation network with the unlabeled data.

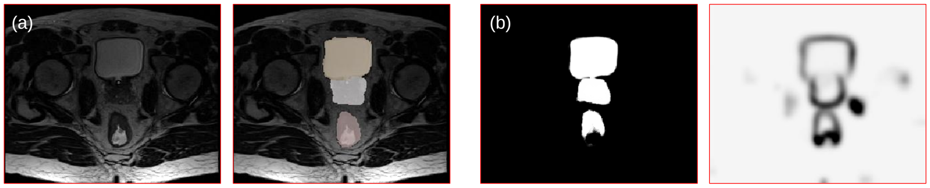


Figure 5.12: (a) A typical pelvic MRI and its manual segmentations of bladder (orange), prostate (silver), and rectum (pink). b) The probability maps (to save space, I combine the probability maps for different organs into one probability map) generated by the segmentation network on the same slice with a), and its corresponding confidence map from the confidence network. All these are obtained after 3 epochs' training. In the confidence map, the brighter regions indicate that they are closer to the ground truth distribution.

5.6.2.4 Exploration of the Confidence-aware Semi-supervising Scheme

The above subsection has validated the effectiveness of the proposed semi-supervised learning strategy towards other methods. This subsection wants to explore when the proposed semi-supervised scheme can contribute to performance gain.

To make an investigation, I randomly sample 1/16, 1/8, 1/4, and 1/2 images as labeled data from the labeled training dataset, and use the rest of training images as unlabeled data. I compare the performance of the segmentation algorithms with (*i.e.*, 'AttSVNet+Ad+Semi') and without (*i.e.*, 'AttSVNet') the proposed semi-supervised learning strategy on the prostate. For fair evaluation, all other settings are the same (Note, the unlabeled data is not working for performance evaluation).

The experimental results are showed in Fig. 5.13. It can be observed that a) the proposed confidence-aware semi-supervised scheme can always contribute to better segmentation performance; b) the proposed strategy works best when the ratio of labeled data over the all training data is ranged from 1/8 to 3/4; c) the proposed semi-supervised scheme cannot work well the the labeled data set is too small because the confidence information cannot well be learned in this setting.

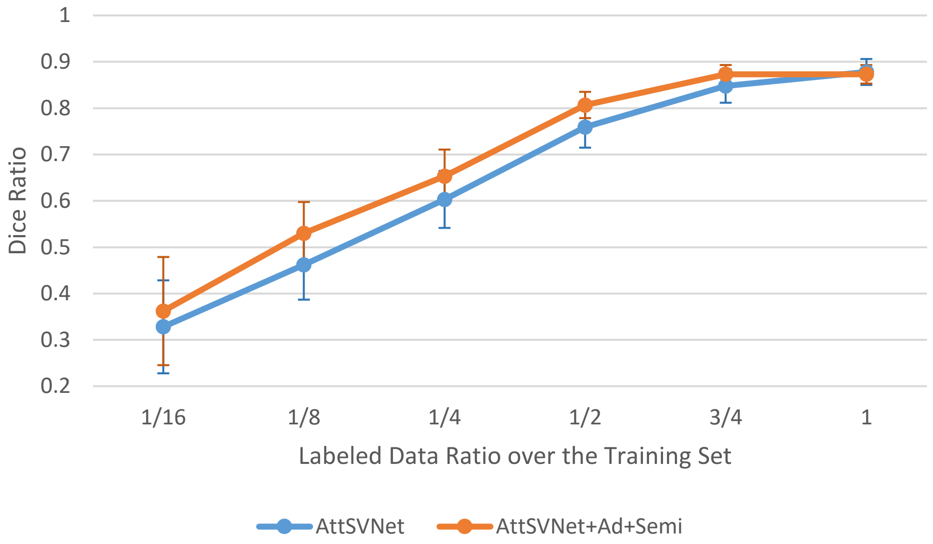


Figure 5.13: Comparison of performance on prostate by the method with the proposed semi-supervised learning strategy and the one without the strategy on different labeled-to-unlabeled ratios.

CHAPTER 6: SUMMARY, DISCUSSION AND FUTURE WORK

6.1 Summary

6.1.1 Summary

This dissertation mainly focused on two research topics: how to accurately delineate blurry organ boundaries and adversarial confidence learnings.

Blurry Boundary Delineation: In this research topic, I have described three solutions to delineate blurry boundaries for different medical images.

1. Accurate segmentation of infant brain images into different regions of interest (ROIs) is one of the most important fundamental steps in studying early brain development. In the isointense phase (approximately 6-8 months of age), white matter (WM) and grey matter (GM) exhibit similar levels of intensities in Magnetic Resonance (MR) images, due to the ongoing myelination and maturation. This results in extremely low tissue contrast and thus makes tissue segmentation very challenging. Existing methods for tissue segmentation in this isointense phase usually employ patch-based sparse labeling on single modality. To address the challenge, we propose a novel 3D multi-modal evolutionary UNet (3D-TFmUNet) for segmentation of isointense phase brain MR images. Specifically, we extend the conventional UNet from 2D to 3D, in addition, I further propose a transformation module to better connect the aggregating layers; I also propose a fusion module to better serve the fusion of feature maps. We compare the performance of my approach with several baseline and state-of-the-art methods on two sets of isointense phase brain images. The comparison results show that my proposed 3D-TFmUNet outperforms all previous methods by a large

margin in terms of segmentation accuracy. In addition, the proposed framework also achieves faster segmentation results compared to all other methods. The experiments further demonstrate that a) transformation and fusion module can reasonably enhance the UNet by endowing the UNet the capacity of dealing with tiny objects; and b) integrating multimodal information can further boost the segmentation performance.

2. One major challenge for medical image segmentation is the blurry nature of organ boundaries in medical images (e.g., CT, MR and, microscopic images), which can often result in low-contrast and vanishing boundaries. With recent advances in CNN, vast improvements have been made for image segmentation, mainly based on the skip-connection-linked encoder-decoder deep architectures. However, in many applications (with adjacent targets in blurry images), these models often fail to accurately locate complex boundaries and properly segment tiny isolated parts. Hence, we aim to provide a method for blurry medical image segmentation and argue that skip connections are not enough to help accurately locate indistinct boundaries. Accordingly, I propose a novel high-resolution multi-scale encoder-decoder network (HMEDN), in which multiscale dense connections are introduced for the encoder-decoder structure to finely exploit comprehensive semantic information. Besides skip connections, extra deeply-supervised high-resolution pathways (comprised of densely connected dilated convolutions) are integrated to collect high-resolution semantic information for accurate boundary localization. These pathways are paired with a difficulty-guided cross-entropy loss function and a contour regression task to enhance the quality of boundary detection. Extensive experiments on a pelvic CT image dataset show the effectiveness of my method for blurry boundary delineation, respectively. My experimental results also show that besides increasing the network complexity, raising the resolution of semantic feature maps can largely affect the overall model performance.

And for different tasks, finding a balance between these two factors can further improve the performance of the corresponding network.

3. Encoder-decoder architectures are widely adopted for medical image segmentation tasks. With the lateral skip connection, the models can obtain and fuse both semantic and resolution information in deep layers to achieve more accurate segmentation performance. However, in many applications (*e.g.*, images with blurry boundary), these models often can not precisely locate complex boundaries and segment tiny isolated parts. To solve this challenging problem, I first analyze why simple skip connections are not sufficient to help accurately locate indistinct boundaries (the information in the skip connection provided from the encoder layers is fuzzy). Then I propose a semantic-guided encoder feature learning strategy to learn high resolution semantic encoder features so that we can more accurately locate the blurry boundaries, which can also enhance the network by selectively learning discriminative features. Besides, I further propose a soft contour constraint mechanism to model the blurry boundary detection. Experimental results on real clinical datasets show that my proposed method can achieve state-of-the-art segmentation accuracy, especially for blurry regions. Further analysis also indicates that my proposed network components indeed contribute to the performance gain. Experiments on additional datasets validate the generalizability of my proposed method.

Adversarial Confidence Learning: Generative adversarial networks (GAN) are widely used in medical image analysis tasks, such as medical image segmentation and synthesis. In these works, adversarial learning is directly applied to the original supervised segmentation (synthesis) networks. The use of adversarial learning is effective in improving visual perception performance since adversarial learning works as realistic regularization for supervised generators. However, the quantitative performance often cannot improve as much as the qualitative performance, and it can even become worse in some cases. In this study, I explore how we can take better advantage of adversarial learning in supervised segmentation (synthesis) models. I analyze the roles of discriminator in the classic GANs and compare them with those in supervised adversarial systems. Based on this analysis, I propose an adversarial confidence learning framework, *i.e.*, besides the adversarial learning for emphasizing visual perception, I use the confidence information provided by the adversarial network to enhance the design of supervised segmentation (synthesis) network. In particular, I propose using a fully convolutional adversarial network for confidence learning to provide voxel-wise and region-wise confidence information for the segmentation (synthesis) network. With these settings, I propose two machine learning algorithms to solve specific medical image analysis problems.

1. I propose a difficulty-aware attention mechanism to properly handle hard samples or regions by taking structural information into consideration so that we can well handle the irregular distribution of medical data. Furthermore, I investigate the loss functions of various GANs and propose using the binary cross entropy loss to train the proposed adversarial system so that I can retain the unlimited modeling capacity of the discriminator. Experimental results on clinical and challenge datasets show that my proposed network can achieve state-of-the-art segmentation (synthesis) accuracy. Further analysis also indicates that adversarial confidence learning can both improve the visual perception performance and the quantitative performance.

2. Recently, deep neural networks yield promising solutions for automatic image segmentation. The network training usually involves a large scale of training data with corresponding ground truth label maps. However, it is very challenging to obtain the ground-truth label maps due to the requirement of expertise knowledge and also intensive labor work, although medical images (*i.e.*, unlabeled data) are relatively easier to obtain. To address such challenges and utilize as much data as possible, I propose a novel semi-supervised deep learning framework, called "Attention based Semi-supervised Deep Networks" (ASDNet), to fulfill the segmentation tasks in an end-to-end fashion. Specifically, compared to the previous deep learning based seg-

mentation networks, I propose a fully convolutional confidence network to adversarially train the segmentation network. Based on the confidence map resulted from the confidence network, I then propose a region-attention based semi-supervised learning strategy to include the unlabeled data for further training the segmentation network. Experimental results on real clinical datasets show that the ASDNet can achieve state-of-the-art segmentation accuracy. Further analysis also indicates that the proposed network components contribute most to the improvement of performance.

6.1.2 Contributions

The contributions of this dissertation are as follows:

1. *A multi-modal evolutionary FCN is proposed to address the problems of low-contrast multi-modal medical image segmentation. Multi-modal information is utilized to address the low-contrast issue for medical image segmentation and fusion strategy is further explored in the environment of multi-modal neural networks. Besides, a transformation module composed of convolutional operations is proposed to alleviate the information bias between encoder and decoder features; a fusion module is proposed to better fuse information from encoder and decoder layers.*

   The multi-modal evolutionary FCN was presented in Sec. 3.1. With empirical study, we argue early-fusion based multi-modal networks is the reasonable choice for multi-modal medical image segmentation. Besides, a transformation module is proposed and applied on the lateral connection to remedy the information bias between encoder and decoder feature maps. Instead of directly concatenating the transformed encoder feature maps and decoder feature maps together, I propose a fusion block which is $1 \times 1 \times 1$ convolution operation to channel-wise fuse the information. In this study, I explored the impact from different patch sizes and argued $32 \times 32 \times 32$ was a reasonable choice for infant brain segmentation. I also investigated the influence of adopting different initialization, pooling and upsampling strategies for the networks. Experiments in Sec. 3.1.5 validate the effectiveness of the multi-modal fusion strategy in segmentation to overcome the challenges brought by the low-contrast medical images. The ablation study also shows the segmentation performance can be improved by the proposed transformation and fusion module. Also, my proposed method can achieve state-of-the-art performance in terms of accuracy and time cost for the isointense infant brain segmentation at that time.

2. *High-resolution encoder-decoder networks are proposed to better delineate the blurry boundaries for medical images, in which, dilated residual module is specifically designed to enhance the skip connection in encoder-decoder networks.*

   In Sec. 3.2, I began to work on blurry boundary delineation. I analyzed the limitation of UNet for blurry boundary delineation problems and figured out that simply fusing encoder and decoder layers for segmentation could work for images with clear boundaries but fail for blurry boundaries. I proposed high-resolution encoder-decoder network to better delineate the blurry boundaries. Namely, high-resolution pathway is designed to endow the lateral connection in encoder-decoder networks with capacity to learn high-resolution semantic features. In particular, two dilated residual blocks are used to form one high-resolution module and several high-resolution modules with different dilation ratios are then sequentially stacked to work as the high-resolution pathway which is put on the lateral connection. Since dilated convolution can effectively increase the theoretical receptive field, sufficient context information could be easily obtained. Moreover, the optimization will be much easier with residual learning. Experimental results on pelvic CT images in Sec. 3.2.5 validates that my proposed method can well delineate the organ boundaries, even for those blurry boundaries. I also carried out experiments to verify that the proposed high resolution pathway can indeed learn high resolution semantic features. Moreover, experimental analysis suggested that using the high-resolution pathway on only one lateral connection could also achieve reasonable performance. As for the disadvantages, although I have controlled the number of additional network parameters, it is still a little more and the memory cost using the proposed method is still huge due to the dilation operation.

3. *Semantic-guided encoder feature learning is proposed to efficiently learn high-resolution semantic features to endow encoder-decoder networks the capacity of blurry boundary delineation.*

   To propose better solution for the blurry boundary delineation problem, I further analyzed the encoder-decoder networks and pointed out that the rich semantic information in decoder layers can be exploited to learn the high-resolution semantic encoder features. In Sec. 3.3, I have presented a novel semantic-guided encoder feature learning strategy to learn both highly semantic and rich resolution information features, so that we can better deal with the blurry-boundary delineation problem. In particular, I introduced semantic features from decoder layers to the spatial-detail reserved encoder features, and designed channel-wise encoding and spatial-wise encoding blocks to form the semantic-guided module. My semantic-guided module can improve the raw skip connection of the classic encoder-decoder networks by enhancing the discriminative features while compressing the less informative features. Experimental results in Sec. 3.3.5 demonstrated that my proposed framework has achieved promising improvement compared to other methods, in terms of both accuracy and robustness, also on the extra public dataset. Additional experiments and analysis further validated my proposed semantic-guided module can learn the high-resolution semantic feature very well, more importantly, my proposed method only increases a little bit of the network complexity and needs very subtle extra computational cost.

4. *Contour-sensitive loss functions, including regression and soft classification, are explored for better modeling the boundaries to more clearly recognize the boundaries.*

To further improve the boundary localization accuracy, I explored how to take better advantage of contour information to achieve better performance for boundary delineation. I first designed a boundary regression branch parallel to the decoder path together with a boundary regression loss in Sec. 3.2.3. To make the optimization of the complicated network easier, in Sec. 3.3.3, I further model the blurry-boundary detection using a soft contour constraint, while an ordinary hard contour constraint to model the clear-boundary detection. The experimental results in Sec. 3.2.5 and Sec. 3.3.5.2 validated the effectiveness of my contour detection based approaches for helping boundary localization and alleviate inter-class mis-classification.

5. *Extensive experiments on several large medical segmentation datasets (infant brain segmentation in MRI, public challenge and self-owned pelvic datasets in both MRI and CT) indicate that the proposed method can accurately delineate the blurry boundaries with outperforming many existing methods in these tasks.*

   I totally developed three methods to delineate blurry boundaries for medical images in Chapter 3. The first method, *i.e.*, 3D-TFmUNet is validated on isointense infant brain segmentation dataset. Due to the low tissue contrast and partial volume effect, part of the boundaries between different tissues are even vanished. My proposed method can achieve state-of-the-art performances at that time. The second method, namely, high-resolution encoder-decoder networks, is validated on a on a large pelvic CT dataset (> 300 patients). The boundaries of the prostate are very blurry and even vanished. The proposed high-resolution pathway learns the high-resolution and semantic features and thus achieves promising results on this task. The last method is mainly about to use lightweight computational cost and less parameters to well delineate the fuzzy boundaries. Experiments on a pelvic dataset and an extra public dataset testify that efficiency of this method.

6. *Deep learning based method is proposed for cross-modality medical image synthesis. Adversarial learning is utilized in the synthesis model, targeting at generating more realistic images. Auto-context model is also explored for alleviating the long-range information dependency problem.*

   In Chapter 4, I proposed a deep residual adversarial network for cross-modality medical image synthesis. The FCN (or UNet) is first proposed to work as the basic synthesis network. Adversarial learning is directly applied to train the supervised synthesis network. I also propose an auto-context framework for the adversarial synthesis network with the purpose of enlarging the context information. My proposed method is validated to be effective on two MRI-to-CT datasets and one 3T-to-7T datasets. More importantly, for the first time, adversarial learning is certified to be able to work well in supervised adversarial model and to contribute to generating more realistic medical images. The auto-context refinement strategy is also validated to be useful for capturing long-range information.

7. *Analysis is conducted to explore the roles of the discriminator in classic GANs and comparison is done with those roles in supervised adversarial learning systems. Together with further experiments, adversarial learning is certified to work as realistic regularization in supervised adversarial learning systems.*

   GANs are used as unsupervised model to automatically generate samples following an implicit data distribution and they are also used in supervised models as adversarial learning. However, very few work analyzed the difference of these two schema. In Sec. 5.1.1, to take better advantage of adversarial learning, I analyzed the roles of discriminators in GAN systems and compared them between classic GAN and supervised adversarial learning system. I argued that adversarial learning provides the sole training signals for unsupervised generator and works as a regularization for supervised generator. With these analysis, I further proposed a potential adversarial learning schema to take better advantage of discriminator.

8. *Adversarial learning can be utilized as realistic regularization for supervised models. The problem of adversarial learning is proposed: visual perception improvement does not well align with quantitative performance gain with raw adversarial learning.*

   In Sec. 5.1.1, Sec. 5.3.2 and Sec. 5.2.2, I figured out that adversarial learning could encourage the generator to generate realistic images since the adversarial loss for generator drives the generated image as a whole to be similar with the ground-truth target image. Hence, adversarial learning is concluded to work as realistic regularization for supervised medical image segmentation and synthesis models. Experiments in Sec. 5.4.2 and Sec. 5.5.1 again certified this argument. With further analysis about the experimental results, I found a phenomenon that adversarial learning could usually bring more visual perception improvement than the quantitative performance gain.

9. *Adversarial confidence learning framework is proposed to address the inconsistency of performance gain in different metrics. By adopting a fully convolutional adversarial network, dense confidence information can be utilized to better design the supervised generator networks to improve the quantitative performance, in the meantime, the realistic regularization with adversarial learning is retained.*

   I analyzed the output of the discriminator and argued that the output of the discriminator could work as confidence information to indicate how well the corresponding image is synthesized or segmented. Based on this analysis, I proposed adversarial confidence learning framework in Sec. 5.3.2. I adjusted the discriminator to learn local confidence information, that is, I adopt a dense network (such as FCN and UNet) as the discriminator instead of using CNN. To this end, the confidence information can be used to design better generator network while the realistic regularization from the adversarial leaning is retained. As a consequence, the quantitative performance gain could potentially align with the visual perception improvement. Also, it is worth noting this is a flexible framework and various specific machine learning techniques can be injected to solve different problems.

10. *Following adversarial confidence learning, difficulty-aware attention mechanism is proposed for medical image segmentation and synthesis, especially hard-to-segment samples and lesion medical image synthesis.*

    Medical image segmentation is a key step for various applications, such as image-guided radiation therapy and diagnosis. Recently, deep neural networks provided promising solutions for automatic image segmentation; however, they often perform well on regular samples (*i.e.*, easy-to-segment samples), since the datasets are dominated by easy and regular samples. For medical images, due to huge inter-subject variations or disease-specific effects on subjects, there exist several difficult-to-segment cases that are often overlooked by the previous works. In Sec. 5.2, I proposed difficulty-aware attention mechanism for medical image segmentation which used the confidence information from the dense output of the discriminator to enhance the design of the objective function of the supervised generator so that we could better handle the hard-to-segment regions. The experiments in Sec. 5.4.3 on the clinical datasets validated that my proposed framework can synchronously improve the quantitative and visual perception performance for medical image segmentation. I further explored the difference between my proposed difficulty-aware attention mechanism with focal loss and figured out that the confidence map could learn structure information if including original input image as input for the discriminator.

    In Sec. 5.2.2, I extended the difficulty-aware attention mechanism to lesion medical image synthesis by developing a weighted L_p loss. Experiments stated in Sec. 5.5 indicated that the proposed framework could be a potential solution for cross-modality lesion medical image synthesis.

11. *Confidence-aware semi-supervised segmentation networks are proposed to adaptively recognize the well segmented samples and regions. It targets at including the well-segmented regions instead of the entire sample to dynamically increase the training labeled set.*

    Training deep segmentation networks usually requires a large scale of training data with corresponding ground truth label maps. However, it is very challenging to obtain the ground-truth label maps due to the requirement of expertise knowledge and also intensive labor work, although medical images (*i.e.*, unlabeled data) are relatively easier to obtain. To address such challenges and utilize as much data as possible, in Sec. 5.3, I have presented a novel attention-based semi-supervised deep networks to segment medical images. Specifically, the semi-supervised learning strategy is implemented by fully convolutional adversarial learning, and also region-attention based semi-supervised loss is adopted to effectively address the insufficient data problem for training the complex networks. More importantly, my proposed semi-supervised segmentation algorithm can dynamically recognize the well segmented regions and thus include them to increase the labeled dataset. By integrating these two components into the framework, my proposed method has achieved significant improvement in terms of both accuracy and robustness, as reported in Sec. 5.6.

12. *Extensive experiments that are conducted on several datasets indicate the effectiveness of my proposed approaches, especially the adversarial confidence learning framework.*

    Generally, in Chapter 4, I have investigated the roles of discriminator in the classic GANs and compared them with those in supervised adversarial learning systems. With the analysis and experiments, I certify that adversarial learning in supervised models actually works as realistic regularization, which aims at constraining the outputs of generator to be as real as possible in an entire view. To align the quantitative performance with the visual perception performance, I propose an adversarial confidence learning framework to take better advantage of adversarial learning for medical image segmentation and synthesis. The proposed difficulty-aware attention mechanism uses the confidence information from the dense output of the discriminator to enhance the design of the objective function of the supervised generator so that we can better handle the hard-to-segment (or hard-to-synthesize) regions. The experiments on the clinical datasets validate that my proposed framework can improve the quantitative performance and visual perception for both medical image segmentation and synthesis models.

These contributions support the thesis statement in Chapter 1, which I revisit here: *1. Current encoder-decoder architecture based networks cannot well solve low-contrast boundary delineation tasks due to the fuzzy information in the encoder layers. Learning high-resolution semantic features is a potential solution. Semantic-guided encoder feature learning can endow these architectures with high-resolution semantic features and thus can well handle the blurry boundary delineation problems. 2. Adversarial learning for supervised models work as realistic regularization. In this learning schema, the quantitative performance gain does not well align with the visual perception improvement. Adversarial confidence learning could achieve a synchronous performance gain by utilizing the confidence information to enhance the design of the supervised model while retaining the realistic effect. More importantly, adversarial confidence learning could provide a chance for building difficulty-aware attention mechanism and confidence-aware semi-supervised algorithm to address easy sample dominance issue and lack of labeled data, respectively, for deep learning based medical image segmentation and synthesis.*

6.2 Discussion

6.2.1 Generalization

As far as my understanding, the proposed two algorithms for blurry boundary delineation can be suited for all kinds of boundary delineation problems, though the experiments are carried out with pelvic datasets. Besides, these algorithms can also be directly applied to many medical image segmentation tasks as long as owning sufficient annotated data, such as abdominal organ segmentation, knee tissue segmentation, gland segmentation, vessel segmentation and so on, since the encoder-decoder architectures are quite suitable for medical image segmentation and my proposed methods are developed to enhance these architectures without any additional supervision.

The adversarial confidence learning framework can be applied to many dense prediction tasks with simple adjustment, such as medical image segmentation and synthesis. In particular, the difficulty-aware attention mechanism relies on the proposed dense confidence network, thus, it is limited to work together with the adversarial confidence learning framework. Moreover, if we have to make this mechanism work well, we need to learn sufficiently good confidence map. Similarly, the confidence-aware semi-supervised learning also needs to go with the adversarial confidence learning and depends heavily on the well-learned confidence map. Since in both scenarios, the dense confidence map is the key indicator to determine which regions are well-segmented and which regions are poorly-segmented.

6.2.2 Resources

Most of my early projects, for instance, the infant brain segmentation project, are based on caffe platform [92]. Since medical images are usually in 3D format, I have developed a 3D caffe library based on the official 2D caffe, in which, many useful functions are implemented. I have released the code in github[1]. My recent projects, such as pelvic organ segmentation and medical image synthesis, are mainly developed based on the pytorch platform[2]. The source code is available via my github page[3].

6.2.3 Limitations

6.2.3.1 Limitations of Deep Learning

In this dissertation, I am using deep learning techniques to build models for most projects, as a result, they suffer from almost all the limitations of deep learning:

1. Deep learning is data hungry. It is almost a common sense that we need large scale of data for training neural networks to obtain a well-performed model. On the other side, human beings can learn abstract relationships in a few trials. Hence, I think some mechanism should be designed to endow deep learning the capacity for learning abstractions through unsupervised data (maybe together with a few labeled data).

2. Deep models are usually black-box in nature and lack interpretability.

3. Training deep networks have huge computational cost. Even with a Titan XP GPU, it will cost tens of hours to train a deep segmentation model, and it may cost hundreds of hours to train a segmentation model with only CPU. The computational cost not only have an impact on the training phase, the inference phase is also affected, for example, deploying deep models requires reasonable computational platform which is usually not supported by hospitals.

[1] https://github.com/ginobilinie/caffe3D

[2] https://pytorch.org/

[3] https://github.com/ginobilinie

6.2.3.2 Model Limitations

The segmentation models can achieve very good performance in terms of blurry boundary delineation. However, my models are still not perfect. It can fail in some cases, such as testing images from different domain, testing images with much noise, the boundaries of the testing images are almost totally invisible and so on. As for the medical image synthesis work, my models are far away from perfect. Currently, the generated images look authentic but cannot necessarily replace ground-truth images.

6.2.3.3 Data Limitations

My models, especially the blurry boundary delineation models, have very high demand for accurate ground-truth label maps, which is very difficult to obtain in practice for 3D medical images.

6.3 Future Work

6.3.1 Future Work

6.3.1.1 Domain Knowledge Integration

My segmentation models cannot produce perfect segmentation so far. Observing the failed samples of the proposed algorithm on the pelvic organ segmentation, I found that the algorithm is inclined to fail in cases where the boundaries are totally invisible due to significant amounts of noise incurred by low dose, metal, and motion artifacts, and so forth. To solve these problems, in the future I would like to combine my algorithm with shape-based (or registration-based) segmentation methods and incorporate more robust shape and structural information of target organs.

6.3.1.2 Medical Imaging Speedup

One of the fields this dissertation mainly focuses on is medical image synthesis, however, just like the statement in the limitations, currently, the proposed models cannot generate the completely real cross-modality medical images though these models have achieved state-of-the-art performances. However, the proposed methods can be used for medical imaging speedup, that is, I can try to improve the MRI process, CT process and PET process. This is not just my own desire. In fact, we are now near the dawn of a new era for faster and safer medical imaging. Researchers are combining advanced artificial intelligence techniques to improve the medical imaging process, which can not only accelerate the imaging process but also reduce the side effect of medical imaging. Hence, my future work includes how to take advantage of the proposed cross-modality synthesis algorithms to speed up the imaging process of a certain modality and how to take advantage of utilizing information from a safer modality to reduce the risk of the imaging process of a dangerous modality.

6.3.1.3 More Elegant Solution for Adversarial Learning

This dissertation has pointed out that adversarial learning could not increase the quantitative performance gain as much as the visual perception improvement. To achieve a synchronous performance increment, we propose a adversarial confidence learning framework, followed with difficulty-aware attention mechanism and confidence-aware semi-supervised learning. Though this framework can indeed improve the quantitative performance and retain the qualitative performance, I believe there should be another more elegant solution to this problem. The essence of the inconsistency performance phenomenon of the adversarial learning is the inconsistency between the objective function of generator and the adversarial loss. It will be beneficial to investigate more about the adversarial loss and further develop a compatible adversarial learning system which could not only improve the visual perception as the traditional adversarial learning but also increase the quantitative performance gain. Actually, I have worked on this topic for some time, but cannot find a reasonable solution towards this direction so far.

6.3.2 Future Directions

6.3.2.1 Network Compression

Currently, the network models are usually as large as $10 - 100MB$ containing millions of network parameters. That's the reason why the deep models are so powerful, however, that's also the burden of computation. For the proposed medical image segmentation and synthesis models, I indeed need to reduce the computational cost if we hope to deploy it in hospitals. Network compression is a reasonable solution to achieve such an effect, that is to say, how to further save computational cost and compress the network without losing much performance for the medical image analysis models.

6.3.2.2 Interpretability

Most machine learning models, especially the recent popular deep learning models, require lots of parameters and a considerable amount of data in order to obtain a high performance. This makes their predictions extremely complex, and often impossible to interpret for their decisions and actions. This lack of interpretability is currently one of the major challenges of applying artificial intelligence on many daily scenarios, including medical image analysis. Governments are becoming aware of the central role that AI-based systems will play in our societies and are starting to take actions. For instance, the European Union has established regulations to require a right of explanation for output of an machine learning algorithm which can make the automated decision-making process more transparent. As a result, I hope I can conduct the research on explainable machine learning, targeting applications in medical imaging where explanations for decisions and actions are highly demanded. For example, I can use the layer-wise relevance propagation as the basis to conduct further study for the interpretation of black-box machine learning models. Generally, I need to explore the exact reason behind every prediction made by the AI agent for biomedical image and health data analysis. For example, I need to know why the agent predicts the patient to be with malignant tumor, and I need to know why the agent determines the proposal as a lung nodule.

REFERENCES

[1] Focal fcn: towards small object segmentation with limited training data. *arXiv preprint arXiv:1711.01506*, 2017.

[2] Daniel C Alexander, Darko Zikic, Jiaying Zhang, Hui Zhang, and Antonio Criminisi. Image quality transfer via random forest regression: applications in diffusion mri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 225–232. Springer, 2014.

[3] Fouzia Altaf, Syed MS Islam, Naveed Akhtar, and Naeem K Janjua. Going deep in medical image analysis. 2019.

[4] Ignacio Arganda-Carreras, Srinivas C Turaga, Daniel R Berger, Dan Cireşan, Alessandro Giusti, Luca M Gambardella, Jürgen Schmidhuber, Dmitry Laptev, Sarvesh Dwivedi, Joachim M Buhmann, et al. Crowdsourcing the creation of image segmentation algorithms for connectomics. *Frontiers in neuroanatomy*, 9:142, 2015.

[5] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[6] Brian B Avants, Charles L Epstein, Murray Grossman, and James C Gee. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis*, 12(1):26–41, 2008.

[7] Raouia Ayachi and Nahla Ben Amor. Brain tumor segmentation using support vector machines. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 736–747. Springer, 2009.

[8] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*, 2015.

[9] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[10] Khosro Bahrami, Feng Shi, Islem Rekik, and Dinggang Shen. Convolutional neural network for reconstruction of 7t-like images from 3t mri using appearance and anatomical features. In *International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, pages 39–47. Springer, 2016.

[11] Khosro Bahrami, Feng Shi, Xiaopeng Zong, Hae Won Shin, Hongyu An, and Dinggang Shen. Hierarchical reconstruction of 7t-like images from 3t mri using multi-level cca and group sparsity. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 659–666. Springer, 2015.

[12] Khosro Bahrami, Feng Shi, Xiaopeng Zong, Hae Won Shin, Hongyu An, and Dinggang Shen. Reconstruction of 7t-like images from 3t mri. *IEEE transactions on medical imaging*, 35(9):2085–2097, 2016.

[13] Wenjia Bai and *et al.*. Semi-supervised learning for network-based cardiac mr image segmentation. In *MICCAI*, pages 253–260. Springer, 2017.

[14] Christoph Baur and *et al.*. Semi-supervised deep learning for fully convolutional networks. In *MICCAI*, pages 311–319. Springer, 2017.

[15] Roland Beisteiner, S Robinson, M Wurnig, M Hilbert, K Merksa, J Rath, I Höllinger, N Klinger, Ch Marosi, Siegfried Trattnig, et al. Clinical fmri: evidence for a 7t benefit over 3t. *Neuroimage*, 57(3):1015–1021, 2011.

[16] Yannick Berker, Jochen Franke, André Salomon, Moritz Palmowski, Henk CW Donker, Yavuz Temur, Felix M Mottaghy, Christiane Kuhl, David Izquierdo-Garcia, Zahi A Fayad, et al. Mri-based attenuation correction for hybrid pet/mri systems: a 4-class tissue segmentation technique using a combined ultrashort-echo-time/dixon mri sequence. *Journal of nuclear medicine*, 53(5):796–804, 2012.

[17] Chetan Bhole, Nicholas Morsillo, and Christopher Pal. 3d segmentation in ct imagery with conditional random fields and histograms of oriented gradients. In *International Workshop on Machine Learning in Medical Imaging*, pages 326–334. Springer, 2011.

[18] Jake Bouvrie. Notes on convolutional neural networks. Technical report, 2006.

[19] Juliette Bruce and Daniel Erman. A probabilistic approach to systems of parameters and noether normalization. *arXiv preprint arXiv:1604.01704*, 2016.

[20] Toan Duc Bui, Jitae Shin, and Taesup Moon. 3d densely convolutional networks for volumetric segmentation. *arXiv preprint arXiv:1709.03199*, 2017.

[21] N. Burgos, M. J. Cardoso, K. Thielemans, M. Modat, S. Pedemonte, J. Dickson, A. Barnes, R. Ahmed, C. J. Mahoney, J. M. Schott, J. S. Duncan, D. Atkinson, S. R. Arridge, B. F. Hutton, and S. Ourselin. Attenuation correction synthesis for hybrid pet-mr scanners: Application to brain studies. *IEEE Transactions on Medical Imaging*, 33(12):2332–2341, Dec 2014.

[22] Ninon Burgos, M Jorge Cardoso, Kris Thielemans, Marc Modat, Stefano Pedemonte, John Dickson, Anna Barnes, Rebekah Ahmed, Colin J Mahoney, Jonathan M Schott, et al. Attenuation correction synthesis for hybrid pet-mr scanners: application to brain studies. *IEEE transactions on medical imaging*, 33(12):2332–2341, 2014.

[23] John Canny. A computational approach to edge detection. In *Readings in computer vision*, pages 184–203. Elsevier, 1987.

[24] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. *arXiv preprint arXiv:1904.11492*, 2019.

[25] Ciprian Catana, Andre van der Kouwe, Thomas Benner, Christian J Michel, Michael Hamm, Matthias Fenchel, Bruce Fischl, Bruce Rosen, Matthias Schmand, and A Gregory Sorensen. Toward implementing an mri-based pet attenuation-correction method for neurologic studies on the mr-pet brain prototype. *Journal of Nuclear Medicine*, 51(9):1431–1438, 2010.

[26] Tony F Chan and Luminita A Vese. Active contours without edges. *IEEE Transactions on image processing*, 10(2):266–277, 2001.

[27] Wing P Chan. Magnetic resonance imaging of soft-tissue tumors of the extremities: A practical approach. *World journal of radiology*, 5(12):455, 2013.

[28] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.

[29] Hao Chen, Qi Dou, Xi Wang, Jing Qin, Jack CY Cheng, and Pheng-Ann Heng. 3d fully convolutional networks for intervertebral disc localization and segmentation. In *International Conference on Medical Imaging and Virtual Reality*, pages 375–382. Springer, 2016.

[30] Hao Chen, Qi Dou, Lequan Yu, and Pheng-Ann Heng. Voxresnet: Deep voxelwise residual networks for volumetric brain segmentation. *arXiv preprint arXiv:1608.05895*, 2016.

[31] Hao Chen, Xiaojuan Qi, Lequan Yu, and Pheng-Ann Heng. Dcan: Deep contour-aware networks for accurate gland segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2487–2496, 2016.

[32] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[33] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016.

[34] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[35] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–818, 2018.

[36] Gary E Christensen, Sarang C Joshi, and Michael I Miller. Volumetric transformation of brain anatomy. *IEEE transactions on medical imaging*, 16(6):864–877, 1997.

[37] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 424–432. Springer, 2016.

[38] Dan Ciresan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. In *Advances in neural information processing systems*, pages 2843–2851, 2012.

[39] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

[40] Pedro Costa, Adrian Galdran, Maria Ines Meyer, Meindert Niemeijer, Michael Abràmoff, Ana Maria Mendonça, and Aurélio Campilho. End-to-end adversarial retinal image synthesis. *IEEE transactions on medical imaging*, 37(3):781–791, 2018.

[41] Pierrick Coupé, José V Manjón, Maxime Chamberland, Maxime Descoteaux, and Bassem Hiba. Collaborative patch-based super-resolution for diffusion-weighted images. *NeuroImage*, 83:245–261, 2013.

[42] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.

[43] Yakang Dai, Feng Shi, Li Wang, Guorong Wu, and Dinggang Shen. ibeat: a toolbox for infant brain magnetic resonance image processing. *Neuroinformatics*, 11(2):211–225, 2013.

[44] Christos Davatzikos. Spatial normalization of 3d brain images using deformable models. *Journal of computer assisted tomography*, 20(4):656–665, 1996.

[45] Ivana Despotović, Bart Goossens, and Wilfried Philips. Mri segmentation of the human brain: challenges, methods, and applications. *Computational and mathematical methods in medicine*, 2015, 2015.

[46] Jose Dolz, Karthik Gopinath, Jing Yuan, Herve Lombaert, Christian Desrosiers, and Ismail Ben Ayed. Hyperdense-net: A hyper-densely connected cnn for multi-modal image segmentation. *IEEE transactions on medical imaging*, 2018.

[47] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision*, pages 184–199. Springer, 2014.

[48] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2016.

[49] M-P Dubuisson and Anil K Jain. A modified hausdorff distance for object matching. In *Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision &amp; Image Processing., Proceedings of the 12th IAPR International Conference on*, volume 1, pages 566–568. IEEE, 1994.

[50] Juergen Dukart and Alessandro Bertolino. When structure affects function–the need for partial volume effect correction in functional and resting state magnetic resonance imaging studies. *PloS one*, 9(12):e114227, 2014.

[51] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015.

[52] Sina Farsiu, Dirk Robinson, Michael Elad, and Peyman Milanfar. Advances and challenges in super-resolution. *International Journal of Imaging Systems and Technology*, 14(2):47–57, 2004.

[53] Qianjin Feng, Mark Foskey, Wufan Chen, and Dinggang Shen. Segmenting ct prostate images using population and patient-specific statistics for radiotherapy. *Medical physics*, 37(8):4121–4132, 2010.

[54] Jun Fu, Jing Liu, Haijie Tian, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. *arXiv preprint arXiv:1809.02983*, 2018.

[55] Pierre-Antoine Ganaye, Michaël Sdika, and Hugues Benoit-Cattin. Semi-supervised learning for segmentation under semantic constraint. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 595–602. Springer, 2018.

[56] Wei Gao, John H Gilmore, Dinggang Shen, Jeffery Keith Smith, Hongtu Zhu, and Weili Lin. The synchronization within and interaction between the default and dorsal attention networks in early infancy. *Cerebral cortex*, 23(3):594–603, 2012.

[57] Yaozong Gao, Shu Liao, and Dinggang Shen. Prostate segmentation by sparse representation based classification. *Medical physics*, 39(10):6372–6387, 2012.

[58] John H Gilmore, Feng Shi, Sandra L Woolson, Rebecca C Knickmeyer, Sarah J Short, Weili Lin, Hongtu Zhu, Robert M Hamer, Martin Styner, and Dinggang Shen. Longitudinal development of cortical and subcortical gray matter from birth to 2 years. *Cerebral Cortex*, 22(11):2478–2485, 2012.

[59] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256, 2010.

[60] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[61] Hayit Greenspan. Super-resolution in medical imaging. *The Computer Journal*, 52(1):43–63, 2009.

[62] Laura Gui, Radoslaw Lisowski, Tamara Faundez, Petra S Hüppi, François Lazeyras, and Michel Kocher. Morphology-driven automatic segmentation of mr images of the neonatal brain. *Medical image analysis*, 16(8):1565–1579, 2012.

[63] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *NIPS*, pages 5767–5777, 2017.

[64] Yanrong Guo et al. Deformable mr prostate segmentation via deep feature learning and sparse patch matching. *IEEE TMI*, 35:1077–1089, 2016.

[65] Xiao Han. Mr-based synthetic ct generation using a deep convolutional neural network method. *Medical Physics*, 44(4):1408–1419, 2017.

[66] Corentin Hardy, Erwan Le Merrer, and Bruno Sericola. Md-gan: Multi-discriminator generative adversarial networks for distributed datasets. *arXiv preprint arXiv:1811.03850*, 2018.

[67] Heather Cody Hazlett, Michele D Poe, Guido Gerig, Martin Styner, Chad Chappell, Rachel Gimpel Smith, Clement Vachet, and Joseph Piven. Early brain overgrowth in autism associated with an increase in cortical surface area before age 2 years. *Archives of general psychiatry*, 68(5):467–476, 2011.

[68] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[69] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[70] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[71] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[72] Rolf A Heckemann, Joseph V Hajnal, Paul Aljabar, Daniel Rueckert, and Alexander Hammers. Automatic anatomical brain mri segmentation combining label propagation and decision fusion. *NeuroImage*, 33(1):115–126, 2006.

[73] Alaa A. Hefnawy. *Super Resolution Challenges and Rewards*, pages 163–206. Atlantis Press, Paris, 2010.

[74] Mattias P Heinrich, Mark Jenkinson, Manav Bhushan, Tahreema Matin, Fergus V Gleeson, Michael Brady, and Julia A Schnabel. Mind: Modality independent neighbourhood descriptor for multi-modal deformable registration. *Medical Image Analysis*, 16(7):1423–1435, 2012.

[75] Karsten Held, E Rota Kops, Bernd J Krause, William M Wells, Ron Kikinis, and H-W Muller-Gartner. Markov random field segmentation of brain mr images. *IEEE transactions on medical imaging*, 16(6):878–886, 1997.

[76] Nicholas Heller et al. Imperfect segmentation labels: How much do they matter? In *MICCAI workshop*, pages 112–120. Springer, 2018.

[77] Moritz Helmstaedter, Kevin L Briggman, Srinivas C Turaga, Viren Jain, H Sebastian Seung, and Winfried Denk. Connectomic reconstruction of the inner plexiform layer in the mouse retina. *Nature*, 500(7461):168–174, 2013.

[78] Matthias Hofmann, Florian Steinke, Verena Scheel, Guillaume Charpiat, Jason Farquhar, Philip Aschoff, Michael Brady, Bernhard Schölkopf, and Bernd J Pichler. Mri-based attenuation correction for pet/mri: a novel approach combining pattern recognition and atlas registration. *Journal of nuclear medicine*, 49(11):1875–1883, 2008.

[79] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[80] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks.

[81] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[82] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[83] Yawen Huang, Ling Shao, and Alejandro F Frangi. Simultaneous super-resolution and cross-modality synthesis of 3d medical images using weakly-supervised joint convolutional sparse coding. *arXiv preprint arXiv:1705.02596*, 2017.

[84] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:1802.07934*, 2018.

[85] Daniel P Huttenlocher, William J Rucklidge, and Gregory A Klanderman. Comparing images using the hausdorff distance under translation. In *Proceedings 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 654–656. IEEE, 1992.

[86] Tri Huynh, Yaozong Gao, Jiayin Kang, Li Wang, Pei Zhang, Jun Lian, and Dinggang Shen. Estimating ct image from mri data using structured random forest and auto-context model. *IEEE transactions on medical imaging*, 35(1):174–183, 2016.

[87] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

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

[89] Viren Jain, Joseph F Murray, Fabian Roth, Srinivas Turaga, Valentin Zhigulin, Kevin L Briggman, Moritz N Helmstaedter, Winfried Denk, and H Sebastian Seung. Supervised learning of image restoration with convolutional networks. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[90] Viren Jain and Sebastian Seung. Natural image denoising with convolutional networks. In *Advances in Neural Information Processing Systems*, pages 769–776, 2009.

[91] Mark Jenkinson and Stephen Smith. A global optimisation method for robust affine registration of brain images. *Med. Image Anal.*, 5(2):143–156, 2001.

[92] Yangqing Jia et al. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678. ACM, 2014.

[93] Amod Jog, Aaron Carass, and Jerry L Prince. Improving magnetic resonance resolution with supervised learning. In *Biomedical Imaging (ISBI), 2014 IEEE 11th International Symposium on*, pages 987–990. IEEE, 2014.

[94] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[95] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical Image Analysis*, 36:61–78, 2017.

[96] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International journal of computer vision*, 1(4):321–331, 1988.

[97] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1646–1654, 2016.

[98] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1637–1645, 2016.

[99] Paul E Kinahan, DW Townsend, T Beyer, and D Sashin. Attenuation correction for a combined 3d pet/ct scanner. *Medical physics*, 25(10):2046–2053, 1998.

[100] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[101] Arno Klein, Brett Mensh, Satrajit Ghosh, Jason Tourville, and Joy Hirsch. Mindboggle: automated brain labeling with multiple atlases. *BMC medical imaging*, 5(1):7, 2005.

[102] Rebecca C Knickmeyer, Sylvain Gouttard, Chaeryon Kang, Dianne Evans, Kathy Wilber, J Keith Smith, Robert M Hamer, Weili Lin, Guido Gerig, and John H Gilmore. A structural mri study of human brain development from birth to 2 years. *The Journal of Neuroscience*, 28(47):12176–12182, 2008.

[103] Simon Kohl et al. Adversarial networks for the detection of aggressive prostate cancer. *arXiv preprint arXiv:1702.08014*, 2017.

[104] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pages 109–117, 2011.

[105] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[106] Pratyush Kumar and Muktabh Mayank Srivastava. Example mining for incremental learning in medical imaging. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 48–51. IEEE, 2018.

[107] Victor AF Lamme, Valia Rodriguez-Rodriguez, and Henk Spekreijse. Separate processing dynamics for texture elements, boundaries and surfaces in primary visual cortex of the macaque monkey. *Cerebral Cortex*, 9(4):406–413, 1999.

[108] Y LeCun, L Bottou, and G Orr. Efficient backprop in neural networks: Tricks of the trade (orr, g. and müller, k., eds.). *Lecture Notes in Computer Science*, 1524.

[109] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[110] Yann LeCun et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[111] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*, 2016.

[112] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision*, pages 702–716. Springer, 2016.

[113] Chunming Li, Chenyang Xu, Changfeng Gui, and Martin D Fox. Level set evolution without re-initialization: a new variational formulation. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 430–436. IEEE, 2005.

[114] Gang Li, Jingxin Nie, Li Wang, Feng Shi, Weili Lin, John H Gilmore, and Dinggang Shen. Mapping region-specific longitudinal cortical surface expansion from birth to 2 years of age. *Cerebral cortex*, 23(11):2724–2733, 2013.

[115] Gang Li, Li Wang, Feng Shi, Weili Lin, and Dinggang Shen. Constructing 4d infant cortical surface atlases based on dynamic developmental trajectories of the cortex. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 89–96. Springer, 2014.

[116] Gang Li, Li Wang, Feng Shi, Amanda E Lyall, Weili Lin, John H Gilmore, and Dinggang Shen. Mapping longitudinal development of local cortical gyrification in infants from birth to 2 years of age. *Journal of Neuroscience*, 34(12):4228–4238, 2014.

[117] Rongjian Li, Wenlu Zhang, Heung-Il Suk, Li Wang, Jiang Li, Dinggang Shen, and Shuiwang Ji. Deep learning based imaging data completion for improved brain disease diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 305–312. Springer, 2014.

[118] Wenqi Li, Guotai Wang, Lucas Fidon, Sebastien Ourselin, M Jorge Cardoso, and Tom Vercauteren. On the compactness, efficiency, and representation of 3d convolutional networks: brain parcellation as a pretext task. In *International Conference on Information Processing in Medical Imaging*, pages 348–360. Springer, 2017.

[119] Xiaoxiao Li, Ziwei Liu, Ping Luo, Chen Change Loy, and Xiaoou Tang. Not all pixels are equal: Difficulty-aware semantic segmentation via deep layer cascade. *arXiv preprint arXiv:1704.01344*, 2017.

[120] Shu Liao, Yaozong Gao, Aytekin Oto, and Dinggang Shen. Representation learning: a unified deep learning framework for automatic prostate mr segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 254–261. Springer, 2013.

[121] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of*

the IEEE Conference on Computer Vision and Pattern Recognition, pages 1925–1934, 2017.

[122] Tsung-Yi Lin et al. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*, 2017.

[123] Geert Litjens et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *MedIA*, 18(2):359–373, 2014.

[124] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[125] Pauline Luc, Camille Couprie, Soumith Chintala, and Jakob Verbeek. Semantic segmentation using adversarial networks. *arXiv preprint arXiv:1611.08408*, 2016.

[126] Aurélien Lucchi, Kevin Smith, Radhakrishna Achanta, Vincent Lepetit, and Pascal Fua. A fully automated approach to segmentation of irregularly shaped cellular structures in em images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 463–471. Springer, 2010.

[127] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are gans created equal? a large-scale study. In *Advances in neural information processing systems*, pages 700–709, 2018.

[128] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 4898–4906, 2016.

[129] Amanda E Lyall, Feng Shi, Xiujuan Geng, Sandra Woolson, Gang Li, Li Wang, Robert M Hamer, Dinggang Shen, and John H Gilmore. Dynamic development of regional cortical thickness and surface area in early childhood. *Cerebral cortex*, 25(8):2204–2212, 2014.

[130] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3, 2013.

[131] José V Manjón, Pierrick Coupé, Antonio Buades, D Louis Collins, and Montserrat Robles. Mri superresolution using self-similarity and image priors. *Journal of Biomedical Imaging*, 2010:17, 2010.

[132] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *ICCV*, pages 2813–2821. IEEE, 2017.

[133] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015.

[134] M Mazonakis, J Damilakis, H Varveris, P Prassopoulos, and N Gourtsoyiannis. Image segmentation in treatment planning for prostate cancer using the region growing technique. *The British journal of radiology*, 74(879):243–249, 2001.

[135] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE TMI*, 34(10):1993, 2015.

[136] Jameson Merkow, Alison Marsden, David Kriegman, and Zhuowen Tu. Dense volume-to-volume vascular boundary detection. In *MICCAI*, pages 371–379. Springer, 2016.

[137] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *ICML*, pages 3478–3487, 2018.

[138] Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163*, 2016.

[139] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016.

[140] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212, 2014.

[141] Pim Moeskops, Max A Viergever, Adriënne M Mendrik, Linda S de Vries, Manon JNL Benders, and Ivana Išgum. Automatic segmentation of mr brain images with a convolutional neural network. *IEEE transactions on medical imaging*, 35(5):1252–1261, 2016.

[142] Pim Moeskops et al. Adversarial training and dilated convolutions for brain mri segmentation. *arXiv preprint arXiv:1707.03195*, 2017.

[143] Gonçalo Mordido, Haojin Yang, and Christoph Meinel. Dropout-gan: Learning from a dynamic ensemble of discriminators. *arXiv preprint arXiv:1807.11346*, 2018.

[144] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[145] Dong Nie, Xiaohuan Cao, Yaozong Gao, Li Wang, and Dinggang Shen. Estimating ct image from mri data using 3d fully convolutional networks. In *Deep Learning and Data Labeling for Medical Applications*, pages 170–178. Springer, 2016.

[146] Dong Nie, Yaozong Gao, Li Wang, and Dinggang Shen. Asdnet: Attention based semi-supervised deep networks for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 370–378. Springer, 2018.

[147] Dong Nie and Dinggang Shen. Semantic-guided encoder feature learning for blurry boundary delineation. *arXiv preprint arXiv:1906.04306*, 2019.

[148] Dong Nie, Roger Trullo, Jun Lian, Caroline Petitjean, Su Ruan, Qian Wang, and Dinggang Shen. Medical image synthesis with context-aware generative adversarial networks. In *MICCAI*, pages 417–425. Springer, 2017.

[149] Dong Nie, Roger Trullo, Jun Lian, Caroline Petitjean, Su Ruan, Qian Wang, and Dinggang Shen. Medical image synthesis with context-aware generative adversarial networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2017.

[150] Dong Nie, Roger Trullo, Jun Lian, Li Wang, Caroline Petitjean, Su Ruan, Qian Wang, and Dinggang Shen. Medical image synthesis with deep convolutional adversarial networks. *IEEE Transactions on Biomedical Engineering*, 65(12):2720–2730, 2018.

[151] Dong Nie, Li Wang, Ehsan Adeli, Cuijin Lao, Weili Lin, and Dinggang Shen. 3-d fully convolutional networks for multimodal isointense infant brain image segmentation. *IEEE transactions on cybernetics*, (99):1–14, 2018.

[152] Dong Nie, Li Wang, Yaozong Gao, Jun Lian, and Dinggang Shen. Strainet: Spatially varying stochastic residual adversarial networks for mri pelvic organ segmentation. *IEEE transactions on neural networks and learning systems*, 2018.

[153] Dong Nie, Li Wang, Yaozong Gao, and Dinggang Sken. Fully convolutional networks for multi-modality isointense infant brain image segmentation. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 1342–1345. IEEE, 2016.

[154] Dong Nie, Li Wang, Lei Xiang, Sihang Zhou, Ehsan Adeli, and Dinggang Shen. Difficulty-aware attention network with confidence learning for medical image segmentation. In *AAAI*, 2019.

[155] Ozan Oktay et al. Attention u-net: learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[156] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

[157] Yangming Ou, Jimit Doshi, Guray Erus, and Christos Davatzikos. Multi-atlas segmentation of the prostate: A zooming process with robust registration and atlas selection. *Medical Image Computing and Computer Assisted Intervention (MICCAI) Grand Challenge: Prostate MR Image Segmentation*, 7:1–7, 2012.

[158] Tianxiang Pan, Bin Wang, Guiguang Ding, and Jun-Hai Yong. Fully convolutional neural networks with full-scale-features for semantic segmentation. 2017.

[159] Stephen M Pizer, P Thomas Fletcher, Sarang Joshi, Andrew Thall, James Z Chen, Yonatan Fridman, Daniel S Fritsch, A Graham Gash, John M Glotzer, Michael R Jiroutek, et al. Deformable m-reps for 3d medical image segmentation. *International journal of computer vision*, 55(2-3):85–106, 2003.

[160] Guo-Jun Qi. Loss-sensitive generative adversarial networks on lipschitz densities. *arXiv preprint arXiv:1701.06264*, 2017.

[161] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[162] Hariharan Ravishankar et al. Joint deep learning of foreground and shape for robust contextual segmentation. In *IPMI*, pages 622–632. Springer, 2017.

[163] Torsten Rohlfing, Robert Brandt, Randolf Menzel, and Calvin R Maurer Jr. Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. *NeuroImage*, 21(4):1428–1442, 2004.

[164] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[165] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM transactions on graphics (TOG)*, volume 23, pages 309–314. ACM, 2004.

[166] Abhijit Guha Roy et al. Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks. In *MICCAI*.

[167] Mohammad Sabokrou, Masoud Pourreza, Mohsen Fayyaz, Rahim Entezari, Mahmood Fathy, Jürgen Gall, and Ehsan Adeli. Avid: Adversarial visual irregularity detection. *ACCV*, 2018.

[168] Liang Shan, Christopher Zach, Cecil Charles, and Marc Niethammer. Automatic atlas-based three-label cartilage segmentation from mr knee images. *Medical image analysis*, 18(7):1233–1246, 2014.

[169] Anmol Sharma and Ghassan Hamarneh. Missing mri pulse sequence synthesis using multi-modal generative adversarial network. *arXiv preprint arXiv:1904.12200*, 2019.

[170] Dinggang Shen, Zhiqiang Lao, Jianchao Zeng, Wei Zhang, Isabel A Sesterhenn, Leon Sun, Judd W Moul, Edward H Herskovits, Gabor Fichtinger, and Christos Davatzikos. Optimized prostate biopsy via a statistical atlas of cancer spatial distribution. *Medical Image Analysis*, 8(2):139–150, 2004.

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

[172] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *CVPR*, pages 761–769, 2016.

[173] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[174] Rebecca Smith-Bindman, Jafi Lipson, Ralph Marcus, Kwang-Pyo Kim, Mahadevappa Mahesh, Robert Gould, Amy Berrington De González, and Diana L Miglioretti. Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer. *Archives of internal medicine*, 169(22):2078–2086, 2009.

[175] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[176] Carole H Sudre et al. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *DLMIA*. Springer, 2017.

[177] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[178] Zhiqiang Tian, Lizhi Liu, Zhenfeng Zhang, and Baowei Fei. Superpixel-based segmentation for 3d prostate mr images. *IEEE transactions on medical imaging*, 35(3):791–801, 2016.

[179] Tong Tong, Robin Wolz, Pierrick Coupé, Joseph V Hajnal, Daniel Rueckert, Alzheimer's Disease Neuroimaging Initiative, et al. Segmentation of mr images via discriminative dictionary learning and sparse coding: application to hippocampus labeling. *NeuroImage*, 76:11–23, 2013.

[180] Robert Toth and Anant Madabhushi. Multifeature landmark-free active appearance models: application to prostate mri segmentation. *IEEE Transactions on Medical Imaging*, 31(8):1638–1650, 2012.

[181] Zhuowen Tu and Xiang Bai. Auto-context and its application to high-level vision tasks and 3d brain image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(10):1744–1757, 2010.

[182] Srinivas C Turaga, Joseph F Murray, Viren Jain, Fabian Roth, Moritz Helmstaedter, Kevin Briggman, Winfried Denk, and H Sebastian Seung. Convolutional networks can learn to generate affinity graphs for image segmentation. *Neural Computation*, 22(2):511–538, 2010.

[183] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[184] Markus Unger, Thomas Pock, Werner Trobin, Daniel Cremers, and Horst Bischof. Tvseg-interactive total variation based image segmentation. In *BMVC*, volume 31, pages 44–46. Citeseer, 2008.

[185] Bram Van Ginneken, Mikkel B Stegmann, and Marco Loog. Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. *Medical image analysis*, 10(1):19–40, 2006.

[186] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[187] Andreas Veit, Michael J Wilber, and Serge Belongie. Residual networks behave like ensembles of relatively shallow networks. In *Advances in Neural Information Processing Systems*, pages 550–558, 2016.

[188] R. Vemulapalli, H. V. Nguyen, and S. K. Zhou. Unsupervised cross-modal synthesis of subject-specific scans. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 630–638, Dec 2015.

[189] Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. Diffeomorphic demons: Efficient non-parametric image registration. *NeuroImage*, 45(1):S61–S72, 2009.

[190] Li Wang, Yaozong Gao, Feng Shi, Gang Li, John H Gilmore, Weili Lin, and Dinggang Shen. Links: Learning-based multi-source integration framework for segmentation of infant brain images. *NeuroImage*, 108:160–172, 2015.

[191] Li Wang, Gang Li, Feng Shi, Xiaohuan Cao, Chunfeng Lian, Dong Nie, Mingxia Liu, Han Zhang, Guannan Li, Zhengwang Wu, et al. Volume-based analysis of 6-month-old infant brain mri for autism biomarker identification and early diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 411–419. Springer, 2018.

[192] Li Wang, Feng Shi, Yaozong Gao, Gang Li, John H Gilmore, Weili Lin, and Dinggang Shen. Integration of sparse multi-modality representation and anatomical constraint for isointense infant brain mr image segmentation. *NeuroImage*, 89:152–164, 2014.

[193] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.

[194] Simon K Warfield, Ferenc A Jolesz, and Ron Kikinis. Real-time image segmentation for image-guided surgery. In *SC'98: Proceedings of the 1998 ACM/IEEE Conference on Supercomputing*, pages 42–42. IEEE, 1998.

[195] Neil I Weisenfeld and Simon K Warfield. Automatic segmentation of newborn brain mri. *Neuroimage*, 47(2):564–572, 2009.

[196] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016.

[197] Jelmer M Wolterink et al. Generative adversarial networks for noise reduction in low-dose ct. *TMI*, 36(12).

[198] Yao Wu, Wei Yang, Lijun Lu, Zhentai Lu, Liming Zhong, Ru Yang, Meiyan Huang, Yanqiu Feng, Wufan Chen, and Qianjin Feng. Prediction of ct substitutes from mr images based on local sparse correspondence combination. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 93–100. Springer, 2015.

[199] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[200] Huaxin Xiao, Yunchao Wei, Yu Liu, Maojun Zhang, and Jiashi Feng. Transferable semi-supervised semantic segmentation. *arXiv preprint arXiv:1711.06828*, 2017.

[201] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015.

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

[203] Yan Xu, Yang Li, Mingyuan Liu, Yipei Wang, Maode Lai, I Eric, and Chao Chang. Gland instance segmentation by deep multichannel side supervision. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 496–504. Springer, 2016.

[204] Hui Xue, Latha Srinivasan, Shuzhou Jiang, Mary Rutherford, A David Edwards, Daniel Rueckert, and Joseph V Hajnal. Automatic segmentation and reconstruction of the cortex from neonatal mri. *Neuroimage*, 38(3):461–477, 2007.

[205] Yuan Xue, Tao Xu, Han Zhang, L Rodney Long, and Xiaolei Huang. Segan: Adversarial network with multi-scale l 1 loss for medical image segmentation. *Neuroinformatics*, pages 1–10, 2018.

[206] Pingkun Yan, Yihui Cao, Yuan Yuan, Baris Turkbey, and Peter L Choyke. Label image constrained multiatlas selection. *IEEE transactions on Cybernetics*, 45(6):1158–1168, 2015.

[207] Qingsong Yang, Pingkun Yan, Yanbo Zhang, Hengyong Yu, Yongyi Shi, Xuanqin Mou, Mannudeep K Kalra, Yi Zhang, Ling Sun, and Ge Wang. Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss. *IEEE transactions on medical imaging*, 37(6):1348–1357, 2018.

[208] Xiao Yang, Roland Kwitt, Martin Styner, and Marc Niethammer. Quicksilver: Fast predictive image registration–a deep learning approach. *NeuroImage*, 158:378–396, 2017.

[209] Xin Yang, Lequan Yu, Lingyun Wu, Yi Wang, Dong Ni, Jing Qin, and Pheng-Ann Heng. Fine-grained recurrent neural networks for automatic prostate segmentation in ultrasound images. In *AAAI*, pages 1633–1639, 2017.

[210] Dong Hye Ye, Darko Zikic, Ben Glocker, Antonio Criminisi, and Ender Konukoglu. *Modality Propagation: Coherent Synthesis of Subject-Specific Scans with Data-Driven Regularization*, pages 606–613. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[211] Varduhi Yeghiazaryan and Irina Voiculescu. An overview of current evaluation methods used in medical image segmentation. Technical report, Tech. Rep. CS-RR-15-08, Department of Computer Science, University of Oxford, Oxford, UK, 2015.

[212] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[213] Fisher Yu, Vladlen Koltun, and Thomas A Funkhouser. Dilated residual networks.

[214] Lequan Yu, Xin Yang, Hao Chen, Jing Qin, and Pheng-Ann Heng. Volumetric convnets with mixed residual connections for automated prostate segmentation from 3d mr images. In *AAAI*, pages 66–72, 2017.

[215] Yuhui Yuan and Jingdong Wang. Ocnet: Object context network for scene parsing. *arXiv preprint arXiv:1809.00916*, 2018.

[216] Habib Zaidi, Marie-Louise Montandon, and Daniel O Slosman. Magnetic resonance imaging-guided attenuation and scatter corrections in three-dimensional brain positron emission tomography. *Medical physics*, 30(5):937–948, 2003.

[217] Yiqiang Zhan and Dinggang Shen. Automated segmentation of 3d us prostate images using statistical texture-based matching method. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 688–696. Springer, 2003.

[218] Jun Zhang, Mingxia Liu, Li Wang, Si Chen, Peng Yuan, Jianfu Li, Steve Guo-Fang Shen, Zhen Tang, Ken-Chung Chen, James J Xia, et al. Joint craniomaxillofacial bone segmentation and landmark digitization by context-guided fully convolutional networks. In *International conference on medical image computing and computer-assisted intervention*, pages 720–728. Springer, 2017.

[219] Wenlu Zhang, Rongjian Li, Houtao Deng, Li Wang, Weili Lin, Shuiwang Ji, and Dinggang Shen. Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. *NeuroImage*, 108:214–224, 2015.

[220] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6848–6856, 2018.

[221] Yizhe Zhang, Lin Yang, Jianxu Chen, Maridel Fredericksen, David P Hughes, and Danny Z Chen. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 408–416. Springer, 2017.

[222] Yongyue Zhang, Michael Brady, and Stephen Smith. Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging*, 20(1):45–57, 2001.

[223] Yu Zhang, Guoxu Zhou, Jing Jin, Xingyu Wang, and Andrzej Cichocki. Optimizing spatial patterns with sparse filter bands for motor-imagery based brain–computer interface. *Journal of neuroscience methods*, 255:85–91, 2015.

[224] Yu Zhang, Guoxu Zhou, Jing Jin, Qibin Zhao, Xingyu Wang, and Andrzej Cichocki. Sparse bayesian classification of eeg for brain–computer interface. *IEEE transactions on neural networks and learning systems*, 27(11):2256–2267, 2016.

[225] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 286–301, 2018.

[226] Can Zhao et al. A deep learning based anti-aliasing self super-resolution algorithm for mri. In *MICCAI*, pages 100–108. Springer, 2018.

[227] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2881–2890, 2017.

[228] Sihang Zhou, Dong Nie, Ehsan Adeli, Jianping Yin, Jun Lian, and Dinggang Shen. High-resolution encoder-decoder networks for low-contrast medical image segmentation. *IEEE Transactions on Image Processing.*, 2019.

[229] Qikui Zhu et al. Boundary-weighted domain adaptive neural network for prostate mr image segmentation. *arXiv preprint arXiv:1902.08128*, 2019.

[230] Wentao Zhu, Xiang Xiang, Trac D Tran, Gregory D Hager, and Xiaohui Xie. Adversarial deep structured nets for mass segmentation from mammograms. In *ISBI*, pages 847–850. IEEE, 2018.

[231] Xizhou Zhu, Dazhi Cheng, Zheng Zhang, Stephen Lin, and Jifeng Dai. An empirical study of spatial attention mechanisms in deep networks. *arXiv preprint arXiv:1904.05873*, 2019.

[232] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8697–8710, 2018.