FROM DATA-DRIVEN TO DATA-CENTRIC MEDICAL IMAGE SEGMENTATION

By

Hao Li

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Electrical and Computer Engineering

August 9, 2024

Nashville, Tennessee

Approved:

Ipek Oguz, Ph.D.

Benoit Dawant, Ph.D.

Yuankai Huo, Ph.D.

Daniel Moyer, Ph.D.

Jeffrey Long, Ph.D.

Bin Lou, Ph.D.

This dissertation is dedicated to my wise and knowledgeable advisor, Ipek Oguz; to my family, all who have deeply loved me, and my loved ones—all of whom illuminate this beautiful world; and to the person I will be when revisiting these pages in the future.

ACKNOWLEDGMENTS

It has been a long journey for me as a student, and looking back on it, I realize that the emotions and moments I have experienced over the years have shaped who I am today.

This dissertation marks the end of my time as a student, and it would not have been possible to complete without help. I sincerely appreciate everyone who has supported me. I thank them for their guidance and support, which motivated me to complete it.

First and foremost, I would like to express my gratitude to my advisor, Ipek Oguz, for being both a knowledgeable guide in my research and a wonderful friend for discussions and advice. You have consistently allowed me to explore various research possibilities, always supporting me and seldom saying no. Your expert advice has frequently illuminated my path through confusion. You always respond to my questions with patience and are readily available, which I deeply appreciate. Thank you for the scientific impact you have had on me. I would not have achieved so much these past years without your support and mentorship. It is not easy to find an advisor who is not only an expert in their field but also genuinely cares about their students. Your inclusive attitude has profoundly influenced me in many ways, helping me to become a kinder person. I am truly grateful for everything you have done for me during this journey, especially for guiding me toward success.

To my committee members, Benoit Dawant, Bin Lou, Daniel Moyer, Jeffrey Long, and Yuankai Huo, I extend my heartfelt thanks for providing me with invaluable scientific suggestions as mentors and collaborators. I have gained scientific knowledge from you and a deeper understanding of objectively evaluating the contribution and significance of my research project. The new perspectives and ways of thinking you introduced are truly indispensable.

Additionally, I am deeply grateful to my mentors, Jerry Prince and Aaron Carass from Johns Hopkins University and Ali Kamen from Siemens Healthineers, for your comprehensive mentorship in the field of medical image analysis.

I am thankful to the people in my lab, the Medical Imaging Computing Lab (MedICL), for the straightforward discussions that have improved my work and the support you have provided in my research and personal life. I appreciate the days and nights we have gone through together and have truly enjoyed working with all of you as a team. Additionally, I would like to thank all my friends in the hall for the time we have spent together and the memorable moments we have shared.

I would like to give my deepest gratitude to my family. Thank you for raising me, supporting me, always having my back, and consistently loving me. The love across distances made me never feel alone throughout this journey. I do not know how I would have done this without you.

Lastly, special thanks for the time and experiences over these years, which have made me tougher and stronger. I continue to write my own story and am on the way to being proud of my lab, my family, and myself.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Part I

Introduction

CHAPTER 1

Introduction to Medical Imaging and Segmentation

1.1 Overview

Medical imaging is the process of creating visual representations of the interior of a body for clinical analysis and medical intervention. The results of this process, medical images, are crucial for diagnosing diseases, planning treatments, and monitoring health, allowing non-invasive insight into the structure and function of the human body. In addition, medical image segmentation is an essential process in the analysis of these images, acting as a bridge between raw imaging data and actionable insights in clinical and research settings. It involves partitioning images into multiple segments to accurately identify and delineate the boundaries of various anatomical structures. This method improves the visualization of complex tissues, facilitating accurate diagnosis, treatment planning, and tracking of disease progression. Although manual segmentation is possible, it is subjective, time-consuming, expensive, requires expert knowledge, and is not reproducible. Therefore, automatic segmentation methods are more desirable. However, such methods may lack robustness due to poor contrast, high variability, and noise in medical images.

The evolution from manual segmentation by experts to automatic methods has been marked by significant milestones, including the advent of machine learning and, more recently, deep learning approaches [137]. Deep learning, especially through convolutional neural networks (CNNs) like U-Net [180, 35], has revolutionized segmentation by providing highly accurate, efficient, and automatic solutions adaptable to the detailed variability in medical images. This method is currently state-of-the-art in medical image segmentation, with various approaches quantitatively evaluating different aspects of segmentation algorithms. As the essential task of medical image analysis, segmentation enables advancements that facilitate detailed anatomical and functional assessments critical for patient care.

In this dissertation, I develop and innovate deep learning methods for medical image segmentation. My research transitions from ***data-driven*** to ***data-centric*** aspects, a shift that serves as the foundation of my work (see Fig. 1.1). This transition highlights the critical importance of data quality in enhancing the efficacy of segmentation techniques. While data-driven strategies (Fig. 1.1(a)) aim to improve the performance of CNNs with existing datasets, data-centric methods (Fig. 1.1(b)) highlight the importance of data quality and diversity as essential for improving segmentation algorithm performance.

For the data-driven aspect, I proposed CNNs with state-of-the-art performance to produce robust segmentations for specific tasks [124, 126, 125, 127, 116, 123]. These automatic techniques are designed for

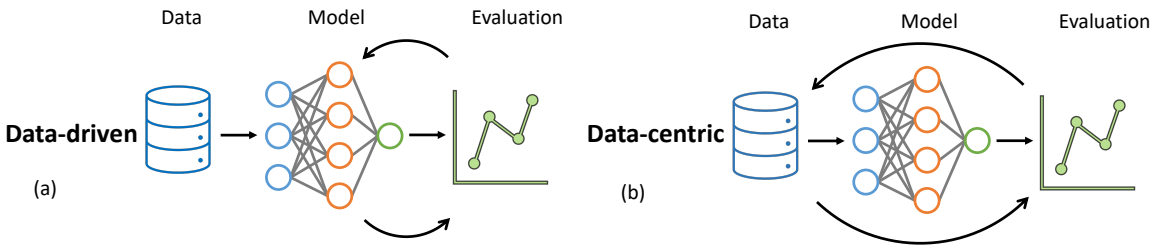


Figure 1.1: Illustration of (a) data-driven and (b) data-centric approaches. Data-driven focuses on modifying networks with the given datasets, while data-centric approaches focus more on improving the quality and diversity of input data.

various practical applications, such as neurodegenerative diseases including Huntington's disease (HD) and Aicardi Goutières Syndrome (AGS) [124, 126, 125, 127], to better monitor disease progression by quantifying the biomarkers of these diseases. They also achieve accurate segmentations of vestibular schwannoma (VS), cochlea, multiple abdominal organs, and the prostate, where the large variations between individuals are significant [116, 123].

For the data-centric aspect, rather than focusing on developing the architecture of CNNs, I proposed novel deep learning methods for medical image segmentation to tackle the common issue of domain shift in practical settings by improving the quality, consistency, and diversity of input data [117, 119, 121, 120, 122]. Domain shift in medical image segmentation arises when models trained on data from one distribution (the source domain) are applied to data from a different, previously unseen distribution (the target domain). To address this, I developed an unsupervised domain adaptation technique to align the distribution of CNN input target data with images from the source domain. Furthermore, I introduced a test-time adaptation method for situations where source domain data is not available, a common scenario in practice due to data privacy concerns. These approaches improve the quality and consistency of input data. Additionally, as one of the most recent innovations, I incorporated domain knowledge from human experts into CNNs as additional input data. By using visual prompts for interactive segmentation, this method increases the diversity of input data. Furthermore, it can improve the robustness of the model to domain shifts between different medical imaging modalities and quickly adapt to the unique characteristics of each modality, ensuring accurate and reliable segmentation. The method is also effective for challenging segmentation tasks within the same modality.

In conclusion, this dissertation represents a significant stride towards robust medical image segmentation with deep learning, transitioning from data-driven to data-centric approaches. This shift targets more practical applications, and the methods proposed will benefit future research. As the field continues to evolve, the potential for these methods, especially for data-centric approaches, to tackle unexplored challenges and improve patient outcomes remains vast, which indicates an exciting direction for subsequent research.

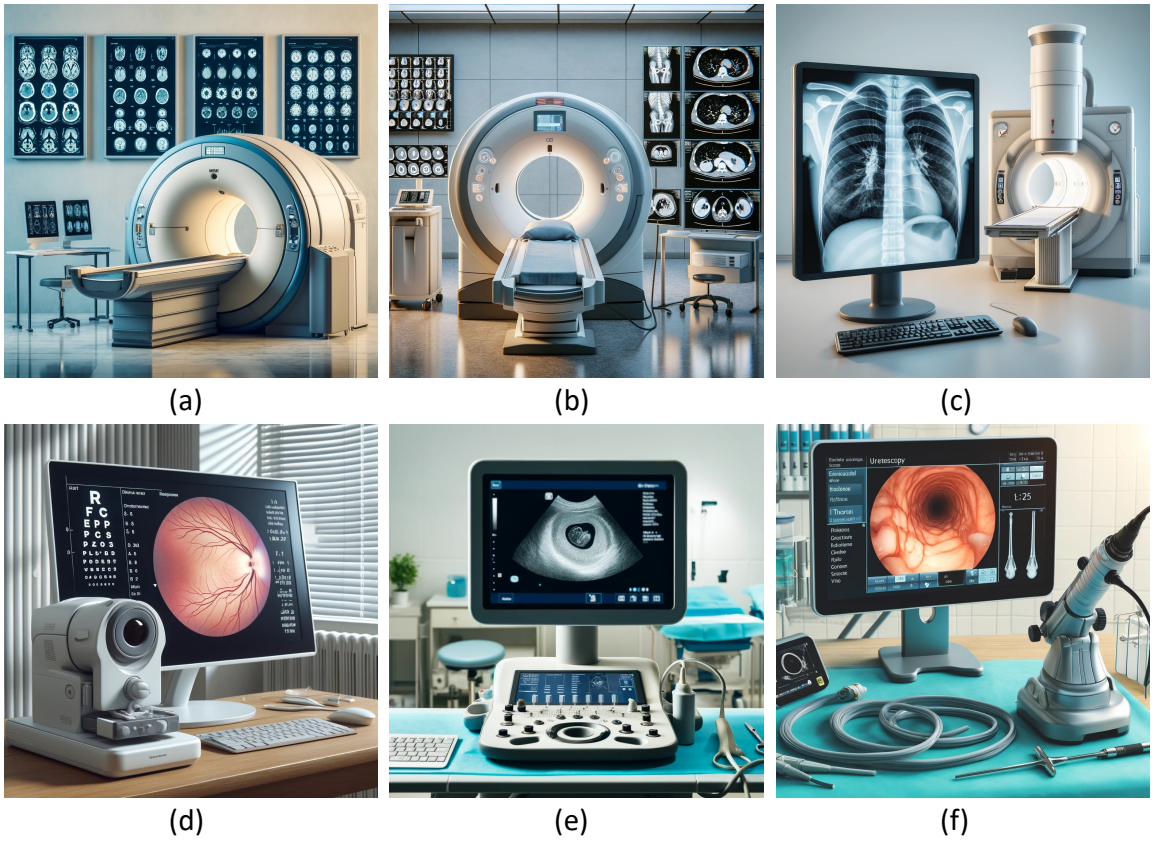


Figure 1.2: Examples of medical imaging systems include: (a) MRI, (b) CT, (c) X-ray, (d) OCT, (e) ultrasound, and (f) endoscopy. Images are generated by DALL·E [178].

1.2 Medical Imaging

Medical imaging is a fundamental aspect of healthcare, which provides clinicians or radiologists with visual representations of the internal structures and functions of the body. The examples of these representations can be viewed in Fig. 1.3. This field contains a wide range of modalities (see Fig. 1.2), such as magnetic resonance imaging (MRI) and computed tomography (CT) [177], each offering unique strengths and applications. These are widely used in radiology imaging and are the main focus of this dissertation.

MRI [177] (Fig. 1.2(a)) offers superior soft tissue contrast, which makes it highly effective for examining the brain, spinal cord, joints, and soft tissues. Additionally, MRI is crucial for the detection of brain tumors, spinal cord injuries, ligament tears, and cartilage damage. These are achieved via various imaging sequences, including T1-weighted (T1w) and T2-weighted (T2w) images, each highlighting different aspects of tissue properties. T1w images are particularly useful for visualizing normal anatomy and assessing post-contrast changes, making them essential for identifying pathological conditions in the central nervous system and evaluating musculoskeletal disorders. On the other hand, T2w images provide detailed contrast between different types of soft tissues, and they are suitable for detecting edema, inflammation, and certain types

of tumors. Diffusion-weighted imaging (DWI) provides insights into the molecular diffusion of water in tissue. It can detect acute ischemic stroke through areas of restricted diffusion and characterizes tumors and other lesions using their diffusion properties. These MRI techniques improve the understanding of various pathological conditions, leading to superior diagnostic accuracy.

CT [177] (Fig. 1.2(b)) plays an important role in medical imaging by providing clinicians with detailed cross-sectional body images. This modality quickly delivers an extensive overview of skeletal structures, organs, and tissues, demonstrating its essential role for both emergency situations and routine diagnostic assessments. CT is highly effective in identifying complex fractures, internal injuries, and bleeds, as well as in measuring tumor size and location, thus playing a key role in devising effective treatment plans. A core strength of CT imaging is its capacity to differentiate subtle tissue density variations. An example includes distinguishing between the dense bone of the spine and the softer tissue of the spinal cord. This differentiation results from analyzing X-ray (Fig. 1.2(c)) measurements from various angles, which are then integrated into a single, detailed image by computational processing.

Other imaging modalities offer unique benefits (Fig. 1.2(c)-(e)). Optical coherence tomography (OCT) [151] provides high-resolution imaging of microscopic structures and is vital for retinal scans in ophthalmology and blood vessel assessments in cardiology. Ultrasound [177] employs sound waves to visualize organs and fetal development in real-time and assists in biopsies. It serves a key role in obstetrics and cardiology, offering a safe, non-invasive diagnostic option across many medical areas. Endoscopy [151] uses a camera-tipped tube to observe internal cavities directly.

1.3 Medical Image Segmentation

Medical image segmentation plays an important role in medical image analysis to better assist clinicians and radiologists in disease detection and diagnosis. Its primary objective is to transform and simplify the visual representation of medical images into formats that are easier to analyze by dividing the image into multiple segments, regions, or structures. In Fig. 1.3, each segment corresponds to a different object or part of the image, facilitating the identification of various structures across different medical images. More specifically, medical image segmentation assigns a label to every pixel/voxel in a 2D/3D image and groups pixels/voxels under the same label based on shared characteristics.

For challenging tasks such as tumor segmentation (Fig. 1.3), medical image segmentation can improve clinician accuracy in disease detection and assist in quantifying the target region. Subcortical or whole-brain segmentation (Fig. 1.3) are important for understanding disease progression through the quantification of different biomarkers and aiding in the study of neuroanatomy and pathology detection, like Huntington's disease (HD) and Aicardi-Goutières syndrome (AGS). Accurate segmentation acts as a key factor in devel-

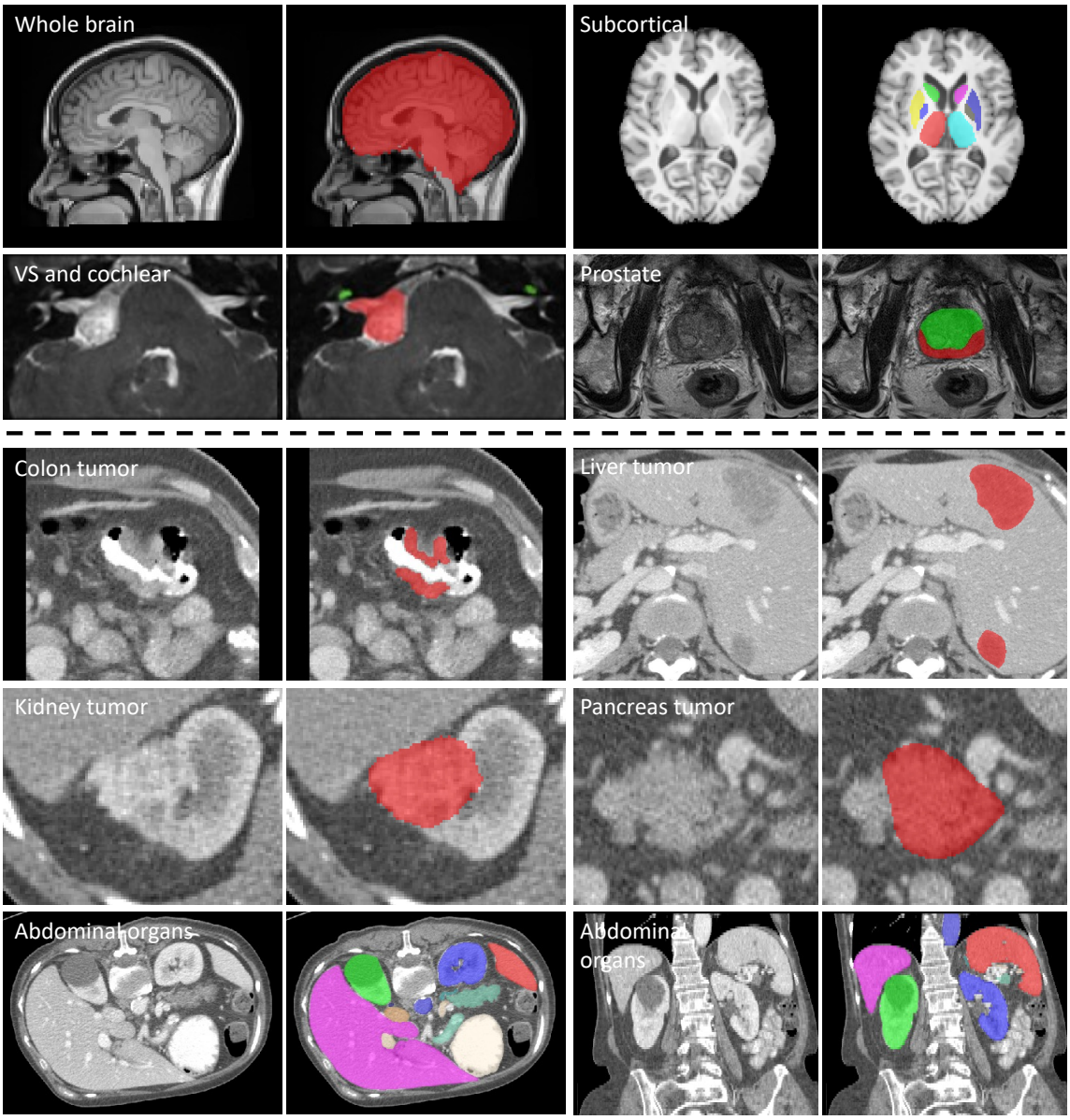


Figure 1.3: Medical images appear on the left side of each pair, while their segmentations are shown on the right. Top and bottom parts show the MRI and CT images, where the target regions are labeled. The details of whole brain segmentation can be found in Chapter 9 and 14. Thalamus, caudate, pallidum and putamen are segmented in Chapter 5, 6, 7 and 8. Vestibular schwannoma (VS) and cochlear [42], prostate (peripheral and transition zones) [4], and abdominal multple organs [112] are used for Chapter 10, 11, 12 and 13. Colon, liver, kidney and pancreas tumor [4, 10, 74] segmentations are extracted in Chapter 15, 16, and 17.

oping effective treatment plans and monitoring disease progression, highlighting the importance of advanced segmentation techniques in facilitating detailed and accurate analysis of complex medical images. Generally, it is applicable to analyzing a variety of imaging data types, such as MRI and CT across different tasks.

Labeling large-scale medical images can be categorized into three aspects: manual, semi-automatic, or

fully automatic methods, the last of which is particularly valuable for effectively assisting clinicians. Traditionally, segmentation by human experts has been the benchmark for accuracy. However, it requires a significant amount of time and expense, especially for large datasets. In response, semi-automatic or fully automatic algorithms offer an effective alternative, delivering outcomes that closely match manual accuracy, as measured by metrics such as the Dice coefficient [137], while significantly reducing both time and expense. Initial segmentation methods relied on techniques such as edge detection [23], multi-atlas, i.e., template matching [91], statistical shape models [73], and active contours [175]. Recently, deep learning-based methods [137] have dominated the medical image segmentation field with their broad generalizability. These methods have achieved superior accuracy in producing segmentations. Additionally, compared to some traditional segmentation methods, such as the multi-atlas segmentation approach, once training is complete, the inference time is significantly lower, thus increasing the efficiency of medical image analysis and making real-time deployment in clinical settings a possibility.

1.4 Deep Learning in Medical Image Segmentation

Deep learning [113] is a subfield of machine learning, which itself is a subset of artificial intelligence (AI). It is inspired by the structure and function of the human brain, specifically the interconnections among neurons, and it is used to model and understand complex patterns in the given datasets.

In recent years, deep learning has demonstrated state-of-the-art performance in different medical image analysis tasks due to its capability to learn complex patterns and structures from medical datasets. It leverages neural network architectures to automatically identify anatomical structures and pathological regions within medical imaging data [92, 137], including the various imaging modalities discussed in Sec. 1.2 [241, 240, 134, 218, 259, 21, 138, 79, 78, 81, 80, 160, 198]. Among these works, the CNN is a prevalent type of deep learning network, and it is widely used for various medical imaging tasks, particularly medical image segmentation. As a fundamental unit of deep learning, CNNs contain multiple layers with a large number of learnable parameters to enable the recognition of complex patterns and features to specific segmentation tasks, such as distinguishing between tumor tissue and healthy tissue or identifying anatomical boundaries. In addition, they are structured hierarchically to facilitate the transformation of images into high-dimensional feature maps. Specifically, each CNN layer transforms its input data into a slightly more abstract representation than the previous one. Generally, a CNN (Fig. 1.4) includes:

1. Convolutional layers that identify different features in the medical image by using convolution operations to automatically detect and learn spatial patterns, such as edges and textures, within data.

2. Normalization layers are applied after the convolutional layers to standardize the input features and

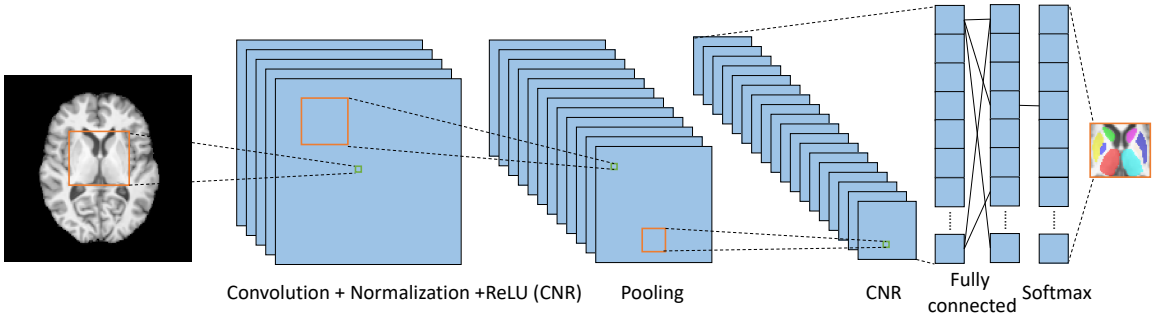


Figure 1.4: An example of CNN sequence to segment subcortical structure. Orange and green boxes denote the size of the kernel and pixel, respectively.

improve the stability and efficiency of the network.

3. Nonlinear activation functions to enhance the capability of a network to recognize complex patterns. The normalization and nonlinear activation layers are often employed after the convolutional layer.

4. Pooling layers to reduce data dimensionality and cost reduction.

5. Fully connected layers at the end of a network aim to classify images based on extracted high-level features.

6. An activation layer to assign final classification probabilities of pixels/voxels to each target class.

The high-dimensional feature maps are better represented by tuning the weights of these cascaded hierarchical layers during training, and they can be converted back to their original resolution representation based on the specific practical applications and their architectural design. Recently, visual transformers [43, 141] are widely used with the CNN for image segmentation tasks to better capture the long-term dependencies.

Training a CNN for segmentation tasks typically via supervised learning, which requires a labeled dataset, such as a collection of images where each pixel/voxel is labeled with its corresponding true class label. Initially, the CNN processes each input image, where the initial weights are assigned randomly or follow a certain distribution [66]. The output for each pixel/voxel is then compared to the corresponding true class labels. The method of backpropagation plays a crucial role in this context. It calculates the gradient of the loss function in relation to the parameters within the model. This calculation aids the optimization algorithm in making adjustments to these parameters with the goal of minimizing the loss [148]. This refinement process is achieved iteratively by allowing it to gradually refine its internal parameters, such as weights, and align the output segmentation map more closely with the label map. Through this iterative training process, the CNN learns to adjust its parameters effectively, which improves its capability to segment new, unseen data accurately. When the network processes this data, it produces precise segmentation outputs. The

effectiveness of a deep learning model is significantly influenced by the quality and size of the given dataset, the architecture of the network, and the optimization strategies implemented during training [92]. The current deep learning approaches can be divided into three perspectives based on the settings of the training dataset [137]:

1. Supervised learning: Train the models with labeled samples.
2. Semi-Supervised learning: Train the models with both labeled and unlabeled samples.
3. Unsupervised learning: Train the models with unlabeled samples.

Additionally, self-supervised learning [110, 119], which aims for CNNs to learn complex features from unlabeled input samples and further refine features for specific downstream segmentation tasks, can be classified as unsupervised learning.

In medical image segmentation, loss functions need to effectively handle specific challenges such as class imbalance between different types of tissues or structures [148]. Commonly used loss functions include but are not limited to cross-entropy loss, Dice loss, and a combination of different loss functions, such as Dice loss combined with cross-entropy loss. These loss functions play an important role in ensuring that the segmentation model not only learns to approximate the general shape and location of the structures in the images but also refines its predictions to align closely with the precise contours and features of those structures. This careful tuning and choice of loss functions are pivotal in achieving high precision and reliability in medical image segmentation outputs.

Deep learning is continually pushing the limits of medical image segmentation, offering automatic segmentation tools that are not only efficient but also capable of handling complex segmentation tasks with impressive accuracy. Recently, the Segment Anything Model (SAM [108]) has revolutionized the field of image segmentation. It has demonstrated wide generalizability and impressive performance by training on massive amounts of data to learn general representations. Moreover, SAM takes various visual prompts, such as points and bounding boxes, as additional inputs, forming an interactive model that segments the target of interest based on the provided prompts. As these deep learning-based interactive models advance (Chapter 15, 16, and 17), they are set to become even more integrated into clinical workflows to improve diagnostic capabilities and ultimately for patient care outcomes, particularly for challenging tasks. This shift from conventional data-driven to more advanced data-centric strategies represents a significant evolution in the processing and interpretation of medical imaging data. The introduction of data-driven and data-centric medical image segmentation will be discussed in Chapter 2 and 3, respectively.

CHAPTER 2

Introduction to Data-driven Medical Images Segmentation

In the field of medical imaging, non-learning-based methods, such as thresholding [162, 187], edge detection [23, 175], and graph-based [73, 235] methods, rely on predefined rules, manual feature engineering, and domain-specific knowledge to segment images. However, they lack generalizability and robustness. In contrast, the data-driven approach using deep learning leverages algorithms and models that analyze and interpret data from extensive datasets ((Fig. 1.1(a)). This approach usually employs supervised learning algorithms, with CNNs being a popular choice due to their effectiveness in processing segmentation tasks. In medical image segmentation, these models are trained on labeled datasets where accurate segmentations from human expert serve as labels. The use of a data-driven methodology offers substantial benefits in terms of efficiency, scalability, and the potential to achieve high accuracy, making it a powerful tool in the field of medical image computing.

2.1 Data Representation

Data representation (Fig. 2.1) in medical image segmentation using CNNs refers to the way the input images and output labels are formatted and preprocessed. For visual transformers, the input data is divided into non-overlapping sub-regions, denoted as token [43]. Proper data representation is crucial for training the CNN effectively. Typically, the input to a CNN is a 2D/3D grayscale image, such as a slice/volume of an MRI or CT scan, represented as a 2D/3D tensor where each pixel/voxel represents its intensity value. For example, in a CT image of the abdomen, these intensity values differentiate between various organs and structures due to their varying densities. The output is a mask that labels each pixel (for 2D) or voxel (for 3D) of the input image with the class it belongs to [116, 123]. In addition, the output is typically represented in the same shape as the input with a single channel. Each value in this channel represents the class of the corresponding pixel or voxel. For instance, in a binary segmentation task, the mask might contain the value 0 for the background and 1 for the foreground. Additionally, values such as 0, 1, 2, and 3 could represent the background, cerebrospinal fluid, gray matter, and white matter, respectively, in tissue classification tasks [7]. There are a few key steps in input data preparation before being fed into the segmentation network.

The patch-based approach is a common method used in the application of CNNs for medical image segmentation [100, 38, 124, 7, 64, 63, 92]. This method is particularly helpful when dealing with 3D medical images as input data, such as high-resolution MRI scans, that are too large to be processed as a whole by the network due to memory constraints. The image patches could be 2D slices, 3D patches, and any format in

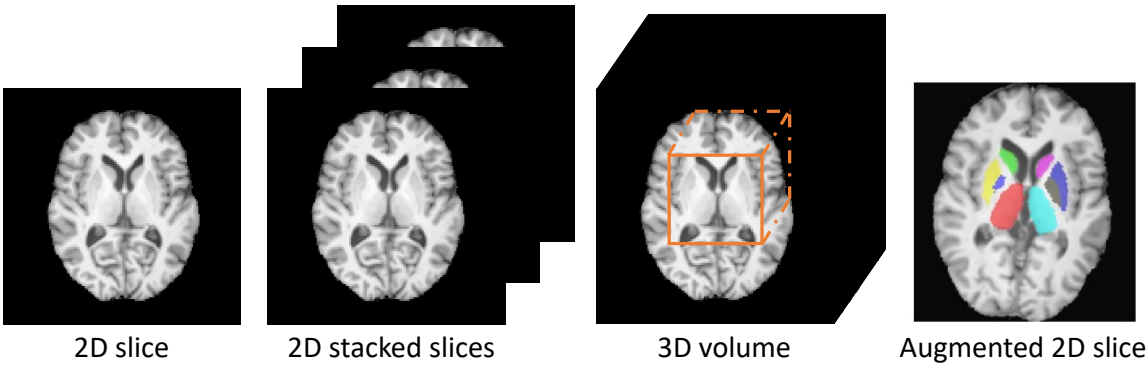


Figure 2.1: 2D axial view illustration of data representation and augmentation. Orange box denotes the patch-based sampling.

between (Fig. 2.1). The choice of patches would affect the performance of networks for a given dataset and task [92]. Specifically, for most of the 3D medical images, such as MRI and CT, compared to 3D patches, 2D slices have the advantage of lighter computational load during training. 2D slices are preferred for the 2D image data where the region of interest is effectively captured in individual slices, and it can be used in 3D data as well [241, 28]. However, for most 3D medical image data, contextual depth information along the third axis is missing, which is useful in modalities like MRI and CT scans where depth information is vital for understanding the structure and progression of diseases or pathology such as tumors. In contrast, 3D patches involve extracting small, cube-shaped segments from three-dimensional medical images, which leverage data from all three axes, but they require more computational resources. As a compromise between 2D and 3D patches, "2.5D" approaches have been proposed, by taking 2D slices from a 3D image and along all three orthogonal views [241, 240]. In addition, multiple adjacent slices are stacked along with the selected slices. This method still captures some contextual information from stacked adjacent slices without the computational intensity required for 3D convolution.

Selecting patches for training CNNs in medical image segmentation is a critical process that can significantly impact the performance of the model. The method of selection should ensure that the CNN learns to recognize relevant features accurately across varied conditions and anomalies in medical images. Random selection of patches from the images is a straightforward method but might lead to an imbalance in the types of features and classes the model encounters. Imbalanced datasets in the medical image domain are characterized by unequal numbers of positive and negative (foreground and background) samples at the pixel level, where the background significantly outweighs the foreground. This may lead to underperformance caused by overfitting [48]. Overfitting occurs when a model learns the training data too well, capturing noise and specific patterns that do not generalize to new data, leading to poor performance on unseen test data. Such cases have led to the development of various patch extraction strategies to achieve robust segmentation.

An alternative strategy is to focus on specific areas or anomalies indicated by experts or preliminary analysis as regions of interest, which are specifically used for patch extraction [211, 126, 117]. In addition to such holistic selection strategies, choosing a voxel from either the foreground or background with equal probability during each training iteration and then selecting a patch centered on that voxel can increase variability [101, 38, 124, 64, 116], and the sliding window approach is employed during inference for these methods.

Data augmentation (Fig. 2.1) is widely used in medical image segmentation by applying a series of transformations to the input images to generate modified versions of the training images. This expands the size of the training dataset, serving as an effective strategy to avoid the overfitting problem and increase the generalizability of model for producing feasible segmentation on unseen samples. This approach is important for handling real-world variability in patient scans [33, 244, 163]. It is particularly beneficial in medical imaging, where annotated data can be limited and imbalanced. The common data augmentation strategies can be classified into three categories:

1. Spatial augmentation [126, 39, 12, 9, 184], such as random image flip, rotation, scale, and deformation (Fig. 2.1) [126, 39, 12, 9, 184].

2. Image appearance augmentation includes, but not limited to random gamma correction, intensity scale, and intensity shift [39, 117, 240, 155].

3. Image quality augmentation includes random Gaussian blur, random noise addition, and image sharpening [117, 240].

These augmentation techniques can be applied individually or in combination. The main idea is to simulate the variations that the model might encounter in the real world, without altering the underlying truth that the images represent.

In practice many studies contain multiple types of data, such as multiple modalities and/or multiple time points from per subject. Compared to single modality or cross-sectional studies, this additional information is clearly valuable and can often be leveraged to improve segmentation performance. This approach could lead the transition of deep learning from data-driven to data-centric, focusing not only on optimizing the performance of CNNs for a given dataset but also on improving the quality of the input data. For instance, different imaging modalities provide distinct visualizations of various tissue types, and multi-modality datasets can thus be leveraged to enhance segmentation accuracy [238, 241]. Data from multiple timepoints are crucial for tracking the longitudinal changes in a subject, where the additional timepoints can also serve as temporal context to improve segmentation for each time point ((Fig. 2.3)(b)) [12, 51, 125].

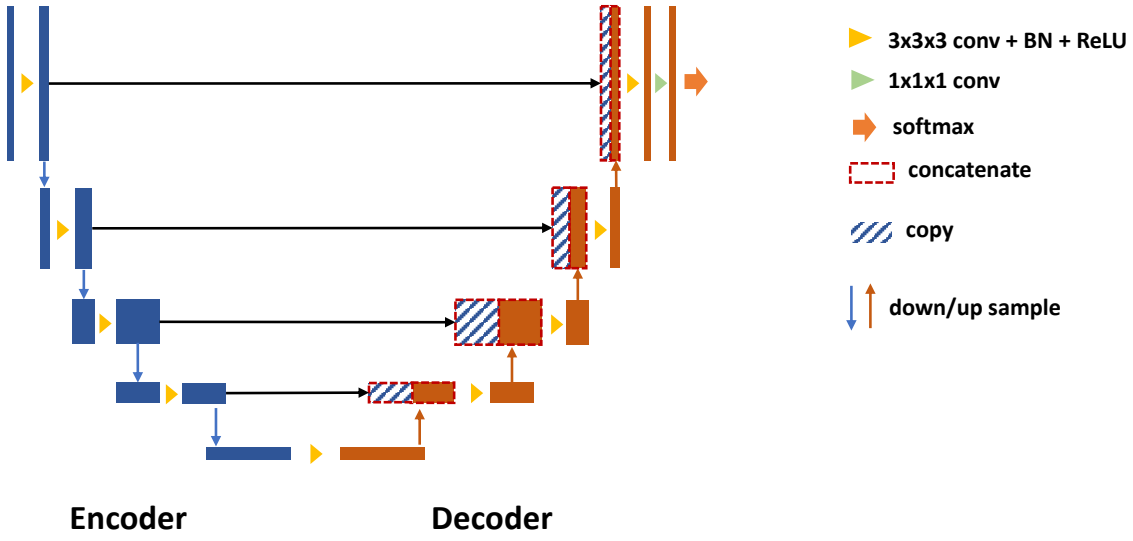


Figure 2.2: Example of an U-shaped network.

2.2 Network Architecture

Different CNN architectures for a single medical image segmentation task can lead to varying results, and selecting the appropriate CNN can be a complex process. This selection depends on various factors such as the specific problem at hand, the type and amount of available data and the computational resources [92]. The development of networks for medical image segmentation can be categorized into three main types: single-path networks [100], encoder-decoder networks [180, 35, 92], and transformers [63], all of which are popular backbones for medical image segmentation challenges in the past decade.

Single-path networks, as shown in Fig. 1.4, have been developed at the early-stage, which regards segmentation as pixel-level classification task. These types of networks typically consist of a sequence of convolutional, pooling, and fully connected layers that form a single path to output segmentations [100]. Additionally, patches are often used as inputs instead of entire images due to the trade-off between segmentation accuracy and the local receptive field. Consequently, these networks are lightweight and computationally efficient. However, a major drawback is the lack of global context from the original image because of the input local patches. This limitation could lead to noisy segmentations, such as undesired islands of false-positive voxels that need to be removed in post-processing [38]. Various efforts have been proposed to compensate for the missing global context, including using spatial coordinates as additional input channels for patches [124] and employing a multi-path network that includes both global and local paths [100, 101, 65].

To address the limitations of the single-path networks, particularly their lack of global information, encoder-decoder networks were developed. These networks designed as U-shaped, with the U-Net [180] being the most popular model for medical image segmentation.

The U-Net (Fig. 2.2) consists of two main parts: the encoder and the decoder, connected by skip connections. Both the encoder and the decoder can be regarded as the single-path networks but incorporate downsampling and upsampling operations across different scales of feature maps. Instead of classifying a local region, the decoder converts the features extracted by the encoder to produce a segmentation that matches the original resolution of input image. The skip connections improve the segmentation performance by transferring feature maps directly from the encoder to the decoder, which helps in retaining the details necessary for accurate segmentation. Furthermore, the 3D U-Net [35] was later introduced to facilitate volumetric segmentation, adapting the model to effectively learn from three-dimensional images.

To date, numerous models developed are designed as variations of the U-Net, including encoder-decoder paths [142, 180]. These designs support end-to-end training that directly transforms images into segmentation maps. A common modification in the U-Net is the incorporation of convolutional modules, such as residual blocks [67], dense blocks [85], and attention modules [184, 161]. These modules can replace standard convolution operations or be integrated into the skip connections. Residual blocks could help the gradient vanishing problem during training by skipping connection, e.g., adding the input of the module to its output, which also contributes to the speed of convergence [67]. In this configuration, the network can be built deeper by using residual connections or blocks instead of regular convolutions in the architecture, enabling robust segmentation of various brain structures [93, 24, 22, 100, 126]. Dense blocks could increase the effectiveness of feature propagation and leverage feature reuse to improve segmentation accuracy [241, 240, 98]. However, they require more computational resources than residual blocks during training. The attention module can be categorized into spatial attention and channel attention modules, which are commonly used tools in segmentation to focus on salient features [126, 117, 213, 242, 77, 95, 256, 195].

The Transformer model [208], originally developed for natural language processing tasks, has been adapted for computer vision tasks, including medical image segmentation [226, 189, 63, 172]. Transformer-based models offer several potential advantages over traditional CNNs for image segmentation tasks. Vision Transformers (ViT) process an image as a sequence of patches rather than as a whole image. The primary advantage of ViTs is their ability to model long-range dependencies and interactions between all pixels in the image, unlike CNNs, which have a localized field of view. This feature can be especially beneficial in medical image segmentation where contextual information from distant parts of the image is crucial. For example, the TransUNet [28] combines the U-Net architecture with a transformer. Initially, a CNN is used to extract local features from the image, followed by a transformer that captures global dependencies among these features. Subsequently, the decoder upsamples these features back to the original resolution to produce the final segmentation map. Additionally, the Swin-transformer [18, 63, 203, 123], which utilizes shifted windows for efficient and flexible multi-scale processing, has been widely adopted in medical image segmentation.

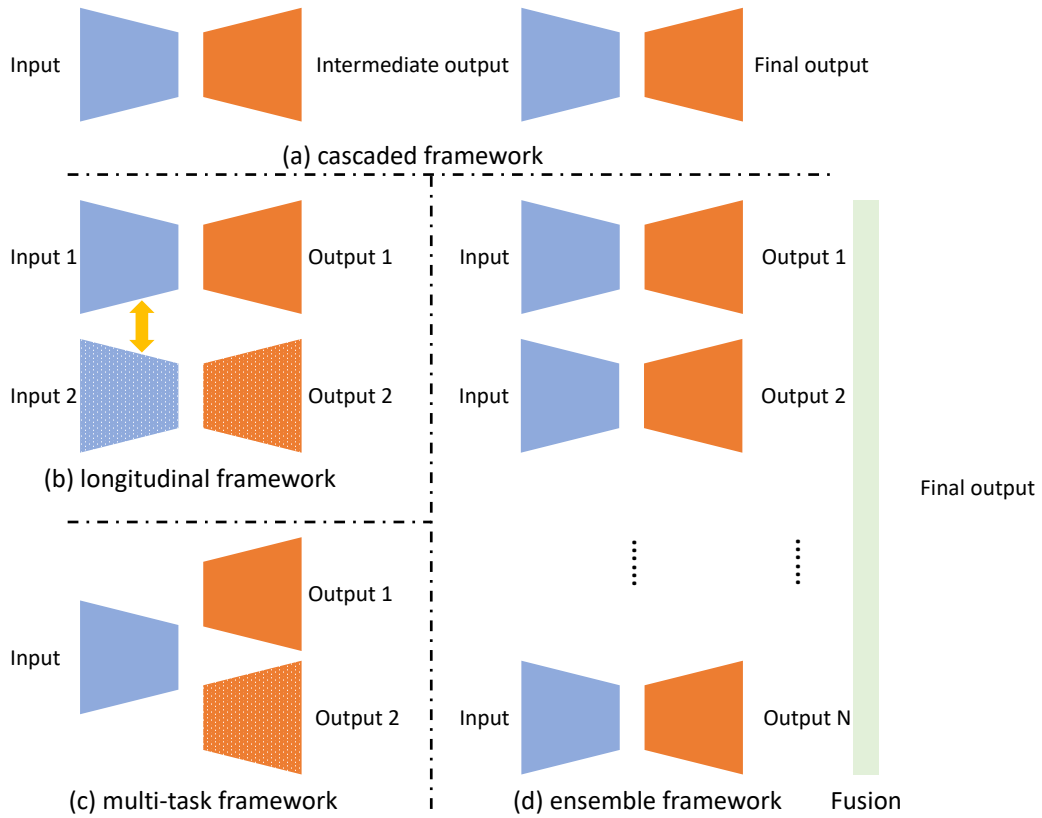


Figure 2.3: Examples of different frameworks. Yellow indicates the interaction between two networks for two different time points. Green denotes the fusion module for ensemble framework, such as majority vote.

The Swin-transformer has demonstrated state-of-the-art results on various benchmark datasets for both image classification and segmentation tasks. In practice, transformer-based models can be viewed under three categories: they can be integrated into existing architectures [53, 206], combined with them [247, 116, 123], or used to replace the convolutional operations to create a purely transformer-based architecture [18, 253].

2.3 Framework Configuration

For certain complex medical image segmentation tasks, using just one network might lead to missed opportunities for capturing relevant information because a single network usually focuses on a specific task during training. To address this issue and improve segmentation accuracy, frameworks that include multiple encoders and decoders have been developed [126, 153, 89].

Multi-task framework (Fig. 2.3(c)) aim to address a main task and auxiliary tasks simultaneously, instead of focusing solely on a single segmentation task. These networks typically feature a shared encoder and several decoders for different tasks, which can help manage class imbalance. Compared to single-task networks, the learning capabilities of this shared encoder are improved by the additional auxiliary tasks, potentially

boosting segmentation performance. Additionally, learning multiple tasks concurrently can also improve the generalizability of model [153, 155, 119].

The cascaded framework (Fig. 2.3(a)) is a series of connected networks such that the input of each downstream network is the output from an upstream network. For example, a coarse-to-fine segmentation strategy can be used to reduce the high computational cost of training for 3D images [211, 126]. In this scenario, an upstream network could take downsampled images as input to roughly locate the target structures or region, allowing the images to be cropped to the region of interest for the downstream network. The downstream network could then produce high-quality segmentation in full resolution. Another advantage of this approach is to reduce the impact of volume imbalance between foreground and background classes. However, the upstream network would determine the performance of the whole framework, and some global information is missing in the downstream networks.

Ensembling strategies (Fig. 2.3(d)) are commonly employed in practical applications to achieve robust segmentation. A favored approach involves aggregating outputs from multiple independent networks, where no weights are shared between the networks [99, 102, 251, 89, 117]. Participants in many medical image segmentation challenges often use model ensembling to achieve superior performance [55, 117, 135, 136].

2.4 Challenge 1: Suboptimal Performance in Specific Medical Image Segmentation Tasks

While the data-driven approach using deep learning represents a significant advancement in medical image segmentation, it also poses a series of challenges that can affect its effectiveness and practical application. One critical issue is that the performance of widely used models may not be optimal for specific tasks given the dataset, necessitating custom-designed solutions. Additionally, the prevalent "one-size-fits-all" approach in model selection often fails to deliver optimal results for all segmentation tasks. Specifically, complex models, while capable of capturing detailed patterns, require large volumes of labeled data to train effectively. However, such datasets are not always available. Without sufficient data, these models are prone to overfitting, resulting in poor generalization on new, unseen datasets, such as those involving different patient demographics or imaging techniques. On the other hand, simpler models might not capture all the necessary details for more complex tasks, though they are quicker to train and less data-intensive.

Moreover, there is significant variation between tasks, ranging from skull stripping to tumor segmentation (Fig. 1.3). This variation suggests that complex models, designed to capture intricate details, may be more suitable for certain tasks, while simpler models could be adequate for others.

Additionally, the framework needs to be specifically tailored for particular tasks. For example, leveraging longitudinal data can improve connections between different timepoints for better outcomes, rather than relying solely on cross-sectional data.

CHAPTER 3

Introduction to Data-centric Medical Images Segmentation

As data volume grows in the field of medical imaging, several key factors significantly contribute to this increase. First, advancements in imaging technologies have led to faster scanning capabilities, which allow more scans within a shorter amount of time, thus producing more data. Second, these advancements have also led to an increased frequency of medical, with more image data available. Lastly, modern medical research often involves multi-modal imaging, e.g., using various imaging techniques for the same subject, which increases the amount of data generated. Together, these factors drive the continuous growth of data volume.

Data-driven methods often lack generalization, such that models trained on specific datasets can perform poorly on data collected under different conditions or from different patient demographics. By improving the quality and comprehensiveness of the data, data-centric strategies enhance the ability of models to generalize across various conditions and lead to more robust and versatile systems. Medical image segmentation requires high levels of precision due to its critical role in diagnosis, treatment planning, and other clinical applications. Data-centric approaches focus on improving data quality through better annotation, noise reduction, and increased dataset diversity (Fig. 1.1(b)).

Regardless of the complexity of a deep learning model, its performance is inherently limited by the quality of its training data. Data-centric strategies emphasize sourcing, cleaning, and annotating high-quality data to mitigate this limitation. These improvements in data quality directly contribute to the increased accuracy and reliability of the segmentation outcomes.

3.1 Domain Shift in Medical Image Segmentation

In deep learning, domain shift refers to the discrepancies between the data distribution in the training set (source domain) and the test set (target domain), which can significantly impact model performance and highlight the need for domain adaptation to ensure robustness. Domain shifts are caused by variations between different scanners, imaging sites, hospitals, and patient demographics within the same imaging modality, as well as by the inherent differences between various imaging modalities [60]:

1. Scanner variations (Fig. 3.1(a)): Different imaging centers may use various scanner types or brands, leading to noticeable differences in image appearance. For example, differences might include variations in contrast, noise levels, or resolution.

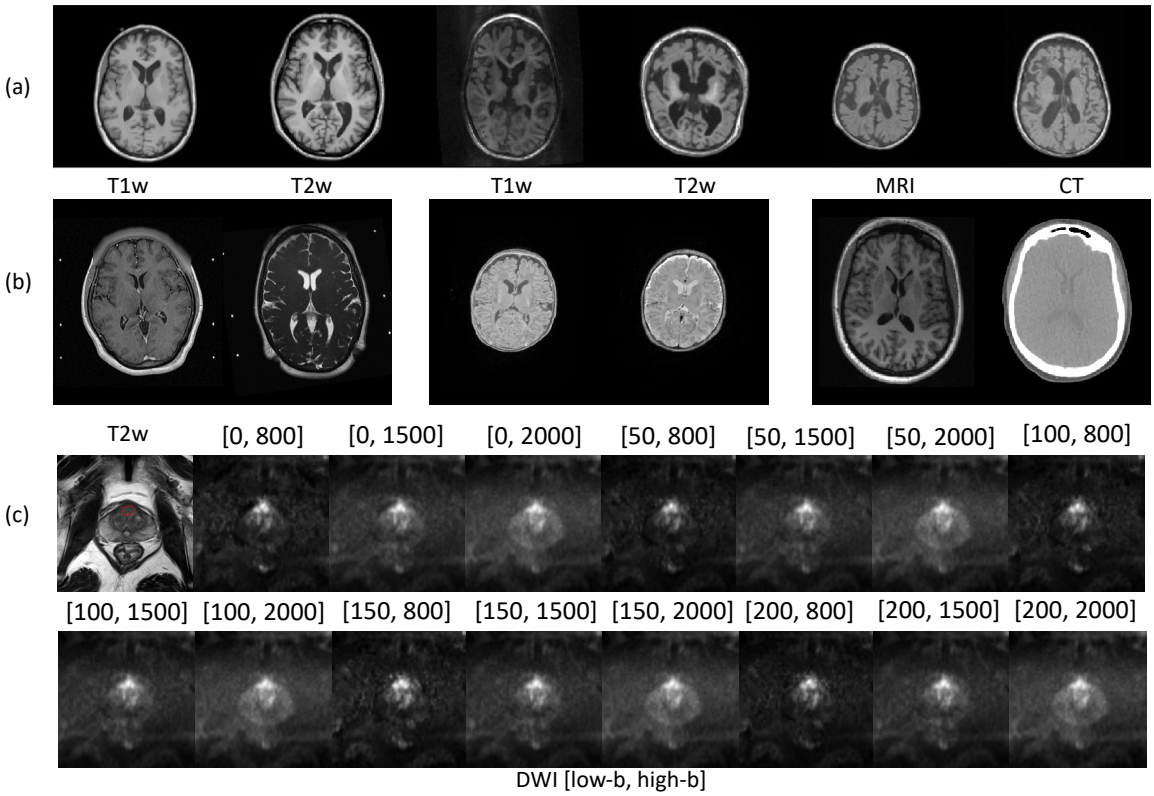


Figure 3.1: Illustration of domain shift in medical imaging. (a) represents the domain shift in a single modality, e.g., scanner variations among T1-weighted MRIs. In addition, the population domain shift can be viewed in (a) and (b) between adult (first two images in (a), left and right paris in (b)) and pediatric subjects (right four images in (a) and middle pair in (b)). Moreover, (b) indicates cross-modality domain shifts, where the middle pair represents paired data (from same subject), while the others are unpaired. (c) shows the domain shift caused by protocol variation for diffusion-weight MRI (DWI), where the low-b and high-b values are labeled.

2. Protocol variations (Fig. 3.1(c)): Even with the same scanner type, differing imaging protocols (e.g., different contrast or scanning parameters) can alter image characteristics.

3. Population variations (Fig. 3.1(a) and (b)): The patient population can also influence domain shift. For instance, pediatric brain images differ markedly from adult brain images, and images from different disease stages may vary significantly.

4. Multiple modality imaging variation (Fig. 3.1(b)): The use of different imaging modalities, e.g., MRI, CT, to study the same subject can introduce variations as each modality might depict the same anatomical structures in very different contrast or appearance, which can affect how images need to be interpreted and segmented.

Such domain shifts present substantial challenges in medical image segmentation, as models trained in

one domain often perform poorly on data from another domain. This leads to poor generalization and unreliable outcomes when these models are used in real-world settings, which is a limitation for data-driven segmentation methods.

Unseen domain adaptation is the process of adapting a deep learning model trained in one domain (source) to perform effectively in another domain (target) for which it was not originally trained. This adaptation is vital in the field of medical imaging, where training and test data frequently scan from different distributions or domains, especially for large-scale studies, where the datasets are often collected from multiple imaging sites.

3.2 Domain Adaptation Settings

In practical, domain adaptation approaches can be categorized into different groups according to different scenarios. Depending on the availability of labeled data in the source and target domains, domain adaptation methods can be broadly categorized into three settings: supervised, semi-supervised, and unsupervised:

1. Supervised domain adaptation: Labeled data available for both the source and target domains. The model can therefore learn from both domains and transfer relevant knowledge from the source domain to the target domain. However, in practice, this setting is not very common, especially in fields like medical imaging where getting labeled data is expensive and time-consuming.

2. Semi-supervised domain adaptation: This is a more common scenario where there is large amount of labeled data in the source domain, and a smaller amount of labeled data in the target domain. The goal here is to leverage the labeled data from both domains and the unlabeled data from the target domain to achieve good performance on the target domain.

3. Unsupervised domain adaptation: In this setting, which is the most challenging and commonly encountered scenario that only source domain has labeled data while none in the target domain. The objective is to learn a model on the source domain that can generalize well to the target domain, despite the lack of labels in the target domain. In this dissertation, I focus on unsupervised domain adaptation (UDA) to tackle the common practical issues in medical imaging, e.g., label availability and domain shift. The UDA framework is shown in Fig. 3.2.

In addition, depending on whether the source and target domains use the same or different modalities, domain adaptation methods in medical imaging can be categorized into single-modality and cross-modality methods:

1. Single-modality ((Fig. 3.1(a))): In this case, the source and target domains both come from the same

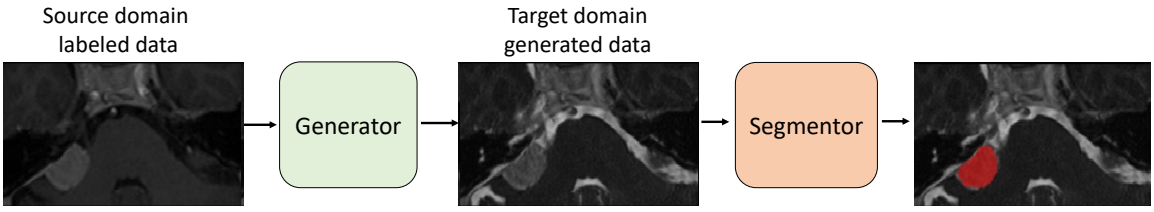


Figure 3.2: Unsupervised domain adaptation framework. Generator translates source domain labeled data into the style of target domain, which are used to train the segmentor. On segmentor is train, and it can be applied to target domain test data to obtain the segmentation.

imaging modality, e.g., both are T1w MRI scans. The domain shift could be due to other factors, like differences in patient populations, scanner hardware, or acquisition protocols.

2. Cross-modality (Fig. 3.1(b)): The source and target domains come from different imaging modalities, e.g., the source domain might be CT scans and the target domain might be MRI scans, or T1w vs. T2w MRI scans. The domain shift in this case can be more significant, as different modalities can have very different image characteristics.

The source domain data itself can come from single imaging site or multiple imaging sites. There is typically greater variability in the data that is collected from multiple sites. Furthermore, most of the current works focus on single-site domain adaptation in medical image analysis.

The above described problems are under the condition that source data is available. However, in practice, source domain data could be unavailable to researchers/clinicians between imaging sites due to privacy issues. In contrast, a pretrained model from the source domain is often easier to obtain.

3.3 Current Domain Adaptation Methods

To tackle the domain shift problem, data augmentation and transfer learning are often used, where transfer learning by finetuning the model with labeled target domain data could achieve high performance for supervised and semi-supervised domain adaptation problems despite the modality differences. These methods aim to enable the trained model to be more robust to variations in the data and perform well across different domains.

In medical imaging, obtaining labeled data can be expensive, time-consuming, and subject to privacy regulations. In addition, the diversity of medical imaging scanners and techniques indicates domain shifts. These factors limit the performance of transfer learning. Unsupervised domain adaptation (UDA) (Fig. 3.2) enables the use of abundant unlabeled data in the target domain and help to bridge this gap without the need for extensive manual labeling.

UDA techniques generally involve learning to align the feature distributions of the source and target domains in a shared feature space [45, 14, 49]. By reducing the domain discrepancy at the feature or image level, the model trained on the source domain can better generalize to the target domain. Techniques such as adversarial training [258], where a model is trained to fool a domain discriminator that tries to distinguish between the source and target domain features, where predictions under different perturbations are made consistent, are commonly used.

In addition, image level adaptation is often used to minimize domain shift by translating the style of images from the source domain into the style of target domain. Generative Adversarial Networks (GANs) [57, 96] are a popular choice for image-level domain adaptation. They consist of two networks, a generator and a discriminator, which are trained simultaneously. As the name suggests, the generator transforms source images to resemble target images, while the discriminator tries to distinguish between real target images and transformed source images. CycleGAN [258] is one common type of GAN used for unsupervised image-to-image translation in medical imaging, where the aim is to learn to translate between source and target domains without paired examples. The transformed source images can then be used along with their labels to train a segmentation model, which can be expected to perform better on the target domain. It is worth to note that the whole process of UDA does not require human label from target domain.

3.4 Challenge 2: Unreliable Performance for Unseen Data

For data-driven methods, the capability of CNNs is highly dependent on the amount of training data, and it is desirable for the training data to be tightly matched to the test data. If the available training data is limited, these models may not learn the diverse range of features necessary to generalize well to unseen data. Also, the performance of CNN drops and produces inaccurate results when encountering domain shift or distribution shift, which significantly contributes to the problem of unreliable performance for unseen data.

In addition, a major limitation of image-level UDA methods is their applicability in multi-domain scenarios, such as when images are collected from multiple imaging sites. Such single-domain mapping methods require training multiple models for each pair of domains, which is time-consuming and often not feasible in practice. Moreover, these methods are designed for use when source images are available. However, this may not be practical due to data privacy issues.

3.5 Interactive Segmentation and Prompt Engineering

As discussed previously, recent years have seen deep learning-based methods achieve state-of-the-art performance in various segmentation tasks [137, 92]. However, robust outcomes are still challenging to obtain due to significant anatomical variations among individuals and ambiguous boundaries in medical images. An

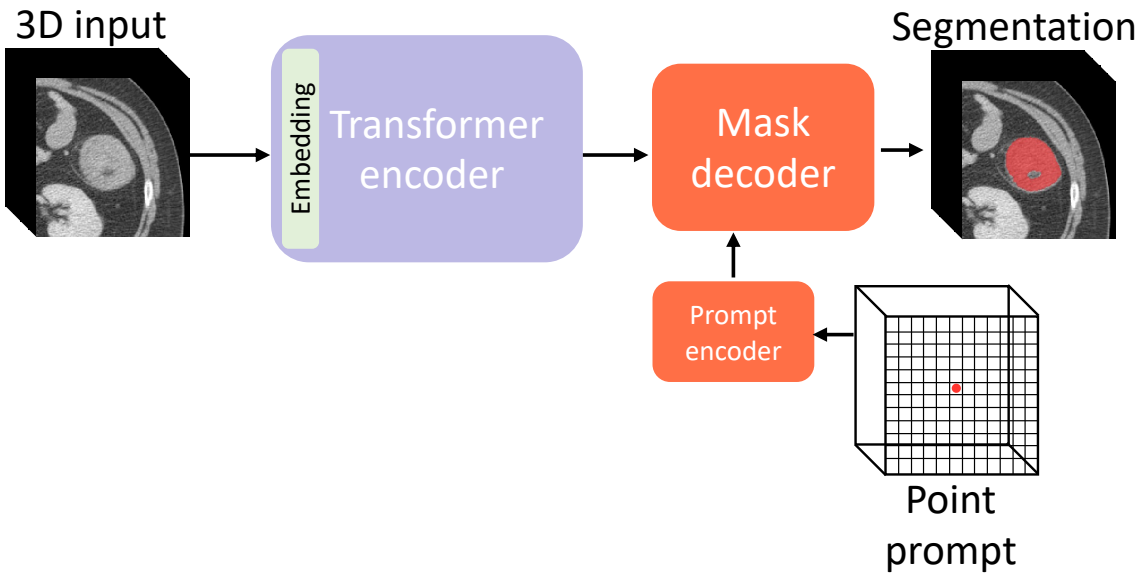


Figure 3.3: Deep learning interactive segmentation model, where transformer encoder and prompt encoder extract the information from input image and given point prompt, respectively.

alternative approach, interactive segmentation, could offer a solution. This method involves users providing real-time user inputs to guide and refine the segmentation process, helping to address some limitations of fully automated algorithms that may struggle with images featuring poor contrast, noise, or ambiguous boundaries between different targets of interest. Moreover, this interaction e.g., providing prompts, can help handle the challenging segmentation cases and domain shifts (Sec. 3.1) by providing additional context and correcting errors in real-time, making the segmentation process more accurate and reliable.

Traditional approaches for medical image segmentation, such as snakes [229] and graph cuts [15], deliver reliable outcomes but often lack consistency and can be time-consuming, especially in complex cases. However, deep learning-based interactive models (Fig. 3.3), such as the Segment Anything Model (SAM) [108], provide a promising solution to these issues. Furthermore, SAM is trained on a rich dataset of natural images, demonstrating wide generalizability and impressive performance by learning general representations from massive amounts of data. Prompt engineering further improves the segmentation capability of these models. Given proper visual prompts, such as points and a bounding box, as additional inputs, these models can handle various zero-shot tasks across domains and produce reliable segmentations during inference. By improving data quality and relevance through strategic input configuration, prompt engineering plays a crucial role in improving model performance in segmentation tasks, without necessarily altering the network architecture. This approach ensures that models not only perform well but are also more context-aware, adaptable to various segmentation scenarios.

Prompt engineering is a key technique in deep learning, originally used in natural language processing (NLP) applications like GPT (Generative Pre-trained Transformer) and other language models. It involves designing user inputs (or prompts) that effectively guide the deep learning model to produce the desired output. While it is traditionally linked with NLP, the principles of prompt engineering can significantly improve the performance of models working with image data as well.

In medical image segmentation, prompt engineering refines how data (input image and prompt) is represented to improve the ability of model to precisely identify and delineate relevant medical structures. This approach includes carefully configuring input features to significantly improve both the accuracy and efficiency of the model by integrating human knowledge into the deep learning framework. Additionally, efficiency in prompt engineering is important for optimizing computational and labor resources and reducing costs. As an example in Fig. 3.3, a point prompt is provided from user to indicate the target region. Such interaction not only helps correct errors on the spot but also improves the robustness of the model to domain shifts between different medical imaging modalities, such as MRI and CT, T1-w and T2-w scans or generated target domain images. By continuously learning from expert feedback, the model adapts to the unique characteristics of each modality, ensuring reliable and accurate segmentation across diverse medical imaging scenarios. The method also handles challenging segmentation tasks within the same modality.

Interactive segmentation and prompt engineering in a data-centric approach are important in the field of medical imaging. This integration can improve the capability of deep learning models to quantify imaging biomarkers with higher accuracy and efficiency. The focus on improving data quality through prompt engineering could redefine how deep learning frameworks are trained and deployed in clinical environments and produce robust and reliable segmentation results by integrating human knowledge.

3.6 Challenge 3: Ineffective Segmentation Performance and Inefficient Prompt Configuration

Unlike the broad successes of SAM [108] in the field of nature image segmentation, medical imaging often faces challenges such as costly data acquisition and time-consuming annotation processes, resulting in a scarcity of massive public datasets available for training. Therefore, it is advantageous to utilize transfer learning from the natural image domain for robust medical image segmentation [152]. However, directly applying pretrained 2D natural image foundation models to 3D medical image segmentation frequently produces suboptimal results [69]. This issue primarily arises due to several key factors: (1) medical images have distinct contrast and texture characteristics that differ significantly from those seen in natural images; (2) anatomical variations among individuals increase complexity of segmentation tasks; and (3) relying on slice-wise (2D) segmentation techniques with transfer learning tends to ignore essential depth-related spatial context, which is important for accurately interpreting and segmenting 3D medical data. Adapting pretrained

models effectively to achieve robust 3D medical segmentations remains a significant challenge.

Recent studies [121] have leveraged insights from pretrained natural image foundation models, SAM [108], to facilitate robust medical image segmentation through parameter-efficient transfer learning techniques. This method uses point prompts to achieve robust performance due to the strong prior provided by the interactive prompts during test-time. Points are the most efficient prompts for medical image segmentation, especially for 3D medical images. However, in the context of interactive segmentation models, determining precise key points during inference can be elusive, especially in medical images characterized by low quality, poor contrast, and ambiguous boundaries. Furthermore, subjectivity leads to variability in prompt choice, and this can be exacerbated by different user expertise levels, resulting in different segmentation outcomes at test-time [185, 233]. Thus, selecting the optimal prompt points for pretrained interactive segmentation models during test-time to achieve better outcomes is also challenging.

Lastly, a robust interactive segmentation model should effectively respond to visual prompts from users with minimal interaction. Inspired by SAM [108], numerous interactive segmentation methods have been proposed in the field of medical imaging. However, these methods generally rely on a single type of prompt [220, 56, 233, 52, 121, 36, 149, 26, 246], and their performance has not met the high standards required for interactive models. Moreover, most of these models do not include iterative human involvement in the loop [56, 52, 246, 121, 36, 220, 149, 26, 233], a critical feature for practical applications which often necessitate iterative adjustments based on new user inputs to achieve satisfactory results [197]. The combined challenges of limited prompt flexibility, insufficient performance standards, and the absence of iterative human interaction highlight why improving these models remains a significant challenge.

CHAPTER 4

Contributed Work

I have developed robust 3D medical image segmentation approaches, which are categorized into two main types of contributions: data-driven Contribution 1 discussed in Sec. 4.1, and data-centric Contributions 2 and 3 discussed in Sec. 4.2 and Sec. 4.3, respectively.

For data-driven contributions, I addressed the Challenge 1 (Sec. 2.4) by developing single-path, U-shaped, and transformer-based networks for various challenging medical image segmentation tasks, achieving state-of-the-art performance. Notably, these contributions tackle different diseases, including neurodegenerative disorders such as Huntington's disease (HD) and Aicardi-Goutières Syndrome (AGS), as well as vestibular schwannoma (VS) and other tumors and organs in practical clinical applications.

For data-centric contributions, I addressed the Challenges 2 (Sec. 3.4) and 3 (Sec. 3.6) by improving the quality and diversity of the input data, thereby providing robust and feasible solutions across various real-world medical imaging conditions.

For Challenge 2 (Sec. 3.4), I developed: (1) an unsupervised cross-modality domain adaptation method from T1-weighted MRI to T2-weighted MRI to segment VS; (2) an unsupervised domain adaptation method using a unified model suitable for multi-domain scenarios, validated with a large-scale dataset; and (3) a test-time domain adaptation method with pretrained models from the source domain for common practical applications, which is useful when source domain image data is unavailable due to the privacy.

For Challenge 3 (Sec. 3.6): (1) I developed a parameter-efficient fine-tuning for medical image segmentation by adapting the pretrained weight from image foundation model (SAM) with different strategies of prompt engineering; (2) I assessed the variability of interactive segmentation model at test-time to improve the reliability of 3D medical image segmentation under different conditions of prompt engineering; and (3) I develop a promptable and robust interactive segmentation model, which takes various visual prompts and involves the human expert into the interaction loop to achieve optimal results.

4.1 Contribution 1: Generalizable Model for Robust 3D Medical Segmentation

To produce robust segmentation for specific medical image segmentation tasks, I addressed the Challenge 1 as introduced in Sec. 2.4 with generalizable models in three aspects:

1. Single-path network: I explored various modifications of the LiviaNET model [38] as the earliest work, which is commonly employed for subcortical segmentation, to improve its generalizability to HD patients under significant neurodegeneration. Specifically, I integrated residual blocks (Res-blocks)

into the convolutional neural network and experimented with manipulating the input of the network and applying random elastic deformations for data augmentation. The training was conducted on images from control subjects, and I tested the model on both control and HD subjects. Our variants showed improved accuracy in the segmentation of most structures for both control and HD groups. Notably, the caudate, which is severely atrophied in HD patients, exhibited the most significant improvement. This improvement suggests that our modifications could improve the ability of LiviaNET to generalize to populations with marked neurodegenerative changes, without the need for retraining.

This work is published at Machine Learning in Clinical Neuroimaging (MLCN) workshop 2020 [124], and the details can be viewed in Chapter 5.

2. U-shaped network: Building on the single-path network, I further developed a U-shaped network to improve subcortical structure segmentation in HD patients. This network mirrors the single-path design, featuring both encoder and decoder elements, but it also incorporates skip connections between them. This advanced cross-sectional method has been validated using a large-scale dataset and achieves state-of-the-art performance. Additionally, I developed a framework for HD that uses longitudinal data as inputs, providing more robust segmentation results. These results align more closely with clinical outcomes compared to the cross-sectional methods.

   For subjects with Aicardi-Goutières Syndrome (AGS), I developed an automatic deep learning framework using a U-shaped network for brain extraction from T1-weighted MRIs. This framework addresses the shortcomings of current methods, which often fail to generalize well to more challenging datasets such as those involving pediatric cases, severe pathology, or heterogeneous data.

   These works are published at SPIE Medical Imaging 2021, 2022 [126, 125, 127], and the details can be viewed in Chapter 6 ,7, 8, and 9.

3. Transformer: I developed a framework, CATS (Complementary **C**NN **a**nd **T**ransformer encoders for **S**egmentation) to offer better performance. This U-shaped CNN is augmented with an independent transformer encoder. The transformer path improves the CNN by modeling long-range dependencies and capturing low-level details. I evaluated this framework across three public datasets on different segmentation tasks and achieved results superior to those of existing state-of-the-art methods.

   Additionally, I extended the vision transformer in CATS using a Swin-transformer [141] to build CATS *v2*. This version incorporates a shifted window scheme, which computes hierarchical representations within shifted windows rather than applying self-attention across the entire image. CATS *v2* was tested on three different segmentation tasks involving abdominal organs, VS, and the prostate, where large

inter-subject variations are present. Our comparisons with state-of-the-art models on the same datasets showed that CATS *v2* consistently outperforms them in terms of Dice scores, indicating that the Swin-transformer improves the segmentation capabilities of networks using hybrid encoders.

These works are published at the International Symposium on Biomedical Imaging (ISBI) 2022 and SPIE Medical Imaging 2023 [116, 123], and the details can be viewed in Chapter 10, and 11.

4.2 Contribution 2: Robust Solutions for Unsupervised Domain Adaptation

To achieve reliable performance on unseen data, I addressed the Challenge 2 as introduced in Sec. 3.4 by using data-centric approaches, e.g., cross-modality UDA, multi-site UDA and test-time adaptation, to tackle common practical domain shift issues. Notably, these contributions are achieved where labels from the target domain are unavailable.

1. Cross-modality UDA: I developed a UDA framework to segment the VS and cochlea for the common cross-modality domain shift scenario. The framework leverages information from contrast-enhanced T1-weighted (ceT1-w) MRIs and its labels and produces segmentations for T2-w MRIs without any labels in the target domain. Initially, a generator was used for image-to-image translation from ceT1-w to T2-w. Then, outputs from an ensemble of models were combined to produce the final segmentation. Various "online" data augmentations were applied during training to accommodate MRIs from different sites/scanners, capturing geometric variability and differences in image appearance and quality.

   This work is published at Brainlesion workshop 2021 [117], and the details can be viewed in Chapter 12.

2. Multi-site UDA: I developed another UDA framework with a unified model to segment and detect prostate lesions from diffusion-weighted MRIs (DWI) across multiple imaging sites. A large-scale dataset containing 5,150 cases (14,191 samples) under various imaging protocols with 34 different b-values was utilized to demonstrate the feasibility of this approach. It is worth highlighting that the proposed solution is computationally efficient, which is particularly beneficial when dealing with a large number of domains, as it avoids the need for training multiple models for each domain pair. Additionally, this method does not require annotations from the test data and can be applied to any pretrained network with generated DWIs for practical use without retraining.

   This work is submitted to Radiology: Artificial Intelligence, and the details can be viewed in Chapter 13.

3. Test-time adaptation: I developed a test-time adaptation method for volumetric medical image segmentation that does not require any source domain data for adaptation nor target domain label for offline

training. This is particularly relevant in scenarios where source domain data may not be available due to privacy issues. The method utilizes pretrained models from the source domain, overcoming the limitations of current UDA-based methods which require access to source domain data. At test-time, the target images are aligned with the source domain on both image and latent feature levels. A translation network is also trained at test-time to align the target image to the source domain at the image level using autoencoders, which is trained in the source domain to optimize the similarity between input and reconstructed output. This study included healthy controls, adult HD patients, and pediatric AGS patients using different scanners and MRI protocols. A subset of healthy controls and adult HD patients, and the entire pediatric AGS dataset served as the target domain data, highlighting significant intensity and geometry shifts between the source and target domains. The results indicate that our method significantly improves CNN performance in the presence of domain shifts during test-time.

This work is published at the Machine Learning in Clinical Neuroimaging (MLCN) workshop 2022 [119], and the details can be viewed in Chapter 14.

4.3 Contribution 3: Effective and Efficient Interactive Segmentation with Prompt Engineering

To develop effective and efficient interactive segmentation models, I addressed Challenge 3, as introduced in Sec. 3.6, by incorporating human knowledge into the model to enhance its segmentation capability and generalizability. This approach can be viewed as using an additional data modality to improve the quality and comprehensiveness of data, aligning with the data-centric strategy. My contributions can be summarized in three main aspects:

1. Parameter-efficient model adaptation: I developed ProMISe, a prompt-driven 3D medical image segmentation model using only a single point prompt to leverage knowledge from a pretrained 2D image foundation model. In particular, we use the pretrained vision transformer from the SAM. Additional lightweight adapters are integrated to extract depth-related (3D) spatial context without updating the pretrained weights. These plug-and-play adapters optimize knowledge transfer across domains and capture fine-grained features effectively. This strategy is compatible with various pretrained models and is easy to implement and train. ProMISe consistently outperforms state-of-the-art methods, as validated on two public datasets for tumor and pancreas segmentation.

2. Optimal test-time prompt selection: I assessed test-time variability for interactive 3D medical image segmentation with diverse point prompts using the ProMISe model to segment colon tumors, where ambiguous boundaries are present. Variability in user expertise and inherently ambiguous boundaries in medical images can lead to inconsistent prompt selections, potentially affecting segmentation ac-

curacy. My goal was to develop a straightforward and efficient strategy for optimal prompt selection during test-time, considering three aspects: (1) determining the necessary number of prompts, (2) identifying effective prompt placement locations, and (3) formulating a strategy for prompt selection. Experimental findings suggest a simple strategy for test-time prompt selection that significantly improves segmentation results over random prompt placement.

These works (aspects 1 and 2) are accepted at the International Symposium on Biomedical Imaging (ISBI) 2024 , and the details can be viewed in Chapter 15, and 16.

3. Effective interactive segmentation model: I developed PRISM, a Promptable Robust Interactive Segmentation Model for 3D medical images. PRISM is designed to respond effectively to visual prompts from users with minimal interaction, ideal for "human-in-the-loop" scenarios. It uses various types of visual prompts (points, boxes, scribbles, masks) to refine outcomes progressively to match inter-rater variability. Trained through iterative and confidence learning, PRISM allows for continuous improvements and robust performance. Each user input improves the performance progressively reach expert-level precision. Following each segmentation iteration, PRISM employs a shallow corrective refinement network to reassign mislabeled voxels. The effectiveness of the proposed PRISM has been validated on four public datasets for segmentation of the colon, pancreas, liver, and kidney, significantly outperforming all compared methods.

   This work is early accepted by the international conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2024, and the details can be viewed in Chapter 17.

4.4 Organization

The rest of this dissertation is organized as follows:

- Part II: Data-driven Medical Images Segmentation:
  - Chapter 5 presents the approach using a single-path network for subcortical segmentation among HD patients.
  - Chapter 6 introduces the proposed U-shaped network for subcortical segmentation.
  - Chapter 7 validates the proposed U-shaped network with a large-scale dataset for HD patients.
  - Chapter 8 presents the approach to longitudinal subcortical segmentation.
  - Chapter 9 presents the variant of the proposed U-shaped network to extract brain.
  - Chapter 10 presents proposed hybrid network, CATS.

  - Chapter 11 introduces CATS *v2*, an extension of the original proposed segmentation model, CATS.

- Part III: Data-centric Medical Images Segmentation:
  - Chapter 12 presents the approach to cross-modality unsupervised domain adaptation.
  - Chapter 13 details the novel strategy for multi-domain unsupervised domain adaptation.
  - Chapter 14 presents the approach for test-time domain adaptation.
  - Chapter 15 introduces the proposed ProMISe model for 3D medical segmentation.
  - Chapter 16 accesses the test-time variability of interactive segmentation model.
  - Chapter 17 presents the proposed promptable and robust interactive segmentation model, PRISM.
- Part IV: Conclusion and Future Work.
  - Chapter 18 summarizes this dissertation and lists the future work for each part.

This work of this dissertation was supported, in part, by NIH U01-NS106845, R01-NS094456 and NSF grant 2220401. The PREDICT-HD study was funded by the NCATS, the NIH (NS040068, NS105509, NS103475) and CHDI.org. The code is publicly available at https://github.com/MedICL-VU

Part II

Data-driven Medical Image Segmentation

CHAPTER 5

Generalizing MRI Subcortical Segmentation to Neurodegeneration

Many neurodegenerative diseases like Huntington's disease (HD) affect the subcortical structures of the brain, especially the caudate and the putamen. Automated segmentation of subcortical structures from MRI scans is thus important in HD studies. LiviaNET is the state-of-the-art deep learning approach for subcortical segmentation. As all learning-based models, this approach requires appropriate training data. While annotated healthy control images are relatively easy to obtain, generating such annotations for each new disease population can be prohibitively expensive. In this work, we explore LiviaNET variants using well-known strategies for improving performance, to make it more generalizable to patients with substantial neurodegeneration. Specifically, we explored Res-blocks in our convolutional neural network, and we also explored manipulating the input to the network as well as random elastic deformations for data augmentation. We tested our method on images from the PREDICT-HD dataset, which includes control and HD subjects. We trained on control subjects and tested on both controls and HD patients. Compared to the original LiviaNET, we improved the accuracy of most structures, both for controls and for HD patients. The caudate has the most pronounced improvement in HD subjects with the proposed modifications to LiviaNET, which is noteworthy since caudate is known to be severely atrophied in HD. This suggests our extensions may improve the generalization ability of LiviaNET to cohorts where significant neurodegeneration is present, without needing to be retrained.

5.1 Introduction

Quantifying the atrophy to subcortical structures from MRI scans is key for many neurodegenerative diseases, such as Alzheimer's disease, Parkinson's disease and Huntington's disease (HD) [8]. Automated segmentation of these structures has thus been an active field of research for many decades [156, 219]. In recent years, learning-based methods have dominated this field due to their superior performance, not only in terms of the accuracy of the output, but also the computational efficiency of prediction on test datasets once a model has been trained. However, while the deep learning methods can be highly accurate, they tend to be dependent on the availability of appropriate training data that is tightly matched to the target test data. When such training data is unavailable, possible options include using poorly matched training data, or investing in creating a well-matched training dataset. Since neurodegenerative diseases typically affect the subcortical structures,

This work is published at MLCN 2020.
Li, Hao, et al. "Generalizing MRI subcortical segmentation to neurodegeneration." Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-oncology: Third International Workshop, MLCN 2020, and Second International Workshop, RNO-AI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 3. Springer International Publishing, 2020.

a model trained on healthy controls can be inappropriate for use on a disease population. Generating new training data is typically time-consuming and requires expert manual annotations. Creating training data for each new disease population and acquisition protocol can thus be prohibitively expensive. It is thus desirable to improve the generalizability of deep learning models trained on healthy control datasets to disease populations.

In this paper, we explore various approaches to extend the popular LiviaNET model [38], which is the current state-of-the-art for this task, in an attempt to boost the generalizability of subcortical segmentation models trained on healthy controls for use on disease populations. Specifically, we explored using 9 convolution Res-blocks [67] with kernel size 3 in our convolutional neural network. We introduced the spatial coordinates of voxels as additional feature channels to improve the results by providing global context to the patch-based LiviaNET model. Finally, we explored random elastic deformations as a data augmentation strategy to mimic atrophied subcortical structures.

We apply our method on data from the PREDICT-HD dataset, where we trained on only healthy control subjects and tested on both control and HD subjects. Compared to the original LiviaNet, we find that our extensions improve the Dice score for all 8 considered subcortical structures (left-right pairs of thalamus, caudate, pallidum and putamen). It is well known that for the HD pathology, the caudate and putamen are the most severely atrophied subcortical structures [145, 146, 170]. Consistent with our hypothesis, our proposed extensions improve LiviaNET segmentation accuracy in HD patients especially for the caudate and the putamen. These findings suggest our approach may be more generalizable to cohorts where significant neurodegeneration is present, without needing to be retrained on disease-specific data.

5.2 Methods

5.2.1 Original LiviaNET

LiviaNET [38] is the current state-of-the-art method for subcortical segmentation. Building on Kamnitsas at al.[101], the LiviaNET uses multi-feature concatenation instead of multi-channel, which preserves the features at different scales. By using a small size kernel ($3 \times 3 \times 3$), LiviaNET goes deeper to perform better.

5.2.2 Manipulating the Network Input

In the first LiviaNET variant we explore, we have used a 4-channel 3D patch as input. The 4 channels are the image patch itself and three spatial coordinate patches encoding the x, y and z coordinate of each voxel within the whole volume. Each channel has the fixed size, $37 \times 37 \times 37$, which was determined empirically to balance the trade-off between the amount of local context and the computational cost of a bigger input patch size. Patches are randomly sampled such that the foreground (8 subcortical structures) and the background

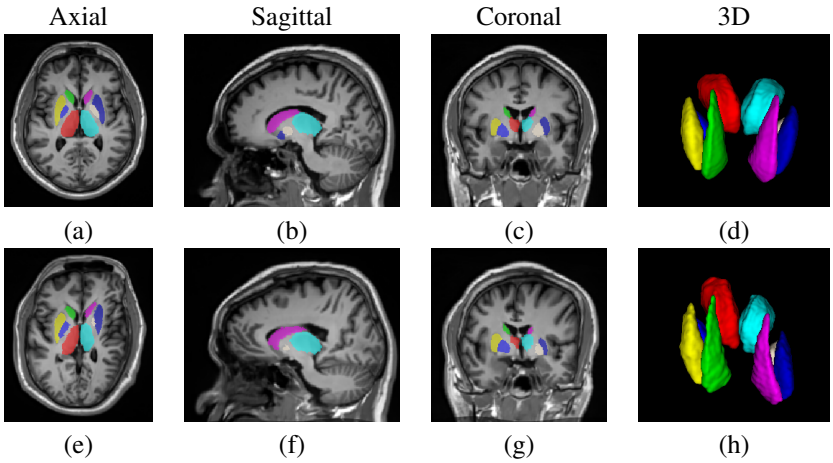


Figure 5.1: Data augmentation with random elastic deformations. **(a-d)**, original image. **(e-h)** deformed image. We note that while the deformation is relatively exaggerated near the periphery of the brain, it is plausible near the subcortical structures.

(within skullstrip mask but outside the subcortical structures) are evenly represented.

5.2.3 Data Augmentation

In another LiviaNET variant, we use random elastic deformations to augment the data. Each training image and its corresponding label map are deformed along the same randomly generated deformation field. Our hypothesis is that the elastic deformation may be able to mimic the atrophy of subcortical structures in neurodegeneration, effectively augmenting the training dataset. We note that because subcortical structures are situated within close proximity of each other near the center of the brain, large amounts of global deformation are needed to create suitable atrophy in this region. The purpose of data augmentation is to increase the variance of model to better represent atrophied brains, even if the deformed images may be non-realistic. Figure 5.1 shows an example deformed image from our augmented dataset. While this image is a representative example, we note that other deformed images in our dataset have somewhat more pronounced deformations.

5.2.4 Network Architecture

We also explored a variation to the LiviaNET architecture. The proposed network architecture is shown in Figure 5.2. We kept the backbone from the LiviaNet [38] and modified each convolution layer into a Res-block. The Res-block contains 3 convolution layers, with no zero-padding for first layer and zero-padding applied to following 2 layers during convolution. While the general intuition with deep learning models is, 'the deeper, the better', in practice, the increased number of layers may cause a degradation problem. With the skip connection of Res-block, the degradation problem is alleviated. Kernel size $3 \times 3 \times 3$ is employed on all convolution kernels in convolution blocks, and $1 \times 1 \times 1$ convolution kernels are used to replace the

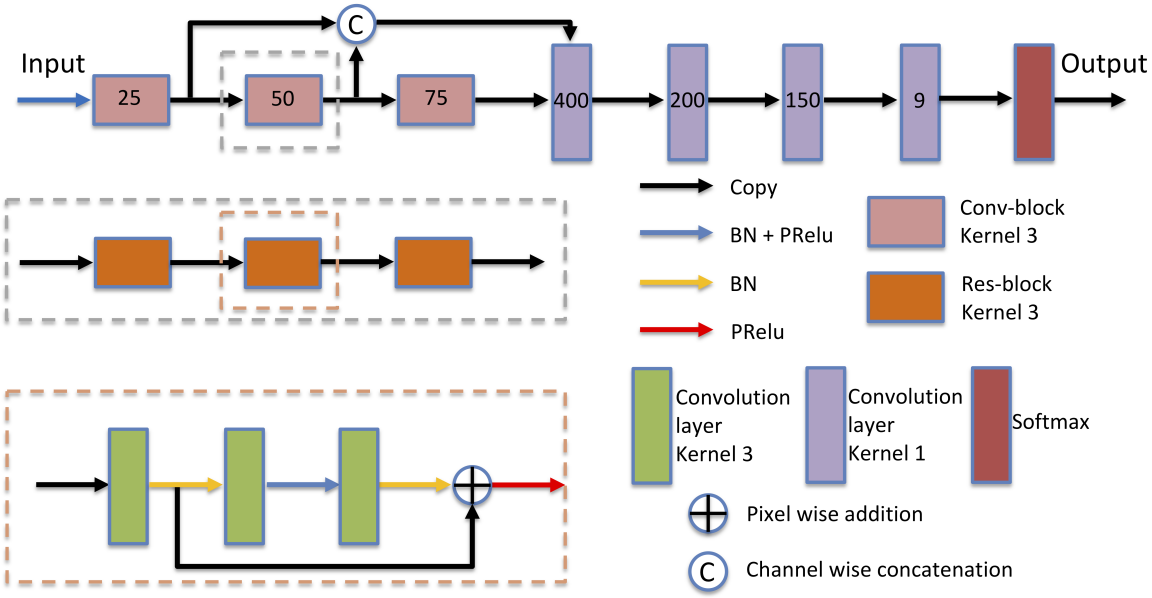


Figure 5.2: The architecture of proposed LiviaNET variant. The network contains LiviaNet and Res-blocks. There are 3 convolution blocks in each convolution layer, and each block consists of 3 Res-blocks which are composed by 3 convolution layers with kernel size 3. The arrows denote different operations. The numbers on each block denote the number of channels. The dashed boxes are expanded for more detail in subsequent rows.

fully connected layer to form the model as fully convolutional. We kept the multi-feature concatenation, which keeps the features from previous block. Before the convolution operation, the batch normalization and non-linear activation function (PReLU) are applied to inputs to minimize effects of intensity and contrast variation. The output with size $9 \times 19 \times 19 \times 19$ is following the softmax layer, where the 9 channels represent each label (8 foreground + 1 background) in the resulting segmentation.

5.2.5 Post-processing

The segmentation result from the deep learning model often contains some disconnected islands that are considered noise. We note that the original LiviaNET [38] suffers from the same problem at a more pronounced level. We use a simple post-processing step to address this issue, namely, extracting the largest connected component for each label. 6-connectivity was used. This post-processing step was used for all the models presented in Sec. 5.3, including LiviaNET.

5.2.6 Experimental Setup

Data. As a preliminary study, we use a subset of the PREDICT-HD database. The full size of PREDICT-HD database in this dissertation can be view in the later comprehensive study Sec. 7.2.2 from Chapter 7. This subset consists of 37 control subjects, as well as 13 subjects with pre-manifest HD (prior to HD

motor diagnosis based on the UHDRS Diagnostic Confidence Level), in each of the low-CAP, medium-CAP, high-CAP categories, as well as 13 HD subjects with clinical diagnosis. CAP is an HD progression marker based on age and CAG expansion at study entry [248]. For each subject, T1-w MPRAGE images from two different timepoints were used. Since PREDICT-HD is a multi-site study, the images were collected with a variety of 3.0T MRI scanners from different vendors (e.g., GE, Phillips, and Siemens), which means the intensity appearance is highly heterogeneous.

Preprocessing. The preprocessing has been conducted with the BRAINSAuto-Workup pipeline [106, 176]. This pipeline consists of: (1) denoising with non-local means filter, (2) anterior/posterior commissure and intra-subject alignments with rigid transformation, (3) bias-field correction, and (4) subcortical segmentation with a multi-atlas method [107]. The multi-atlas segmentation results were visually quality controlled and used as the silver standard. All images were then intensity-normalized to the [0, 4096] range. Finally, skull-stripping was applied.

Implementation details. We trained our models on NVIDIA RTX 2080 8GB GPU for a total of 600 epochs. Inside each epoch, we randomly sampled 500 patches whose centers are inside the skull, and we fed those patches into the model. The learning rate started from 0.0001, and was decayed by factor 0.5 every 30 epochs. We used an Adam optimizer with weight decay of 0.0001. The training loss function is an equally weighted combination of 3 different loss functions: cross entropy (CE) loss, L1 loss and L2 loss. With batch size 16, each training epoch takes approximately 1 minute. Throughout the training process, the loss decreased in an exponential manner. After approximately 200 epochs, the loss had almost reached its minimum already, and the model was very close to the final model. However, for the following 400 epochs, the results continued to slightly improve. Therefore, all the models were trained for 600 epochs. All the training implementation was implemented in PyTorch.

Training, validation and testing. We used 20 of the control subjects as the training dataset, and 2 additional control subjects for validation. The testing dataset consists of the remaining (disjoint) set of 15 control subjects, as well as all 52 HD subjects: 13 subjects in each of low, medium and high CAP pre-manifest categories and 13 patients with clinical diagnosis. The two timepoints for each subject were either both in the training set or both in the test dataset.

5.3 Results and Discussion

The quantitative results of our experiments are presented in Table 5.1. We used the Dice similarity coefficient as our performance evaluation metric.

LiviaNET + Spatial coordinate ($L+S$). In the results from Table 5.1, it can be seen that when using the spatial coordinates as additional input channels, the model is able to learn a 'prior' on the relative locations

Table 5.1: Dice scores of segmentation results, presented as $\frac{mean}{std.\ dev.}$. We compare the original LiviaNET to variants with spatial coordinates, Res-blocks, elastic data augmentation and different loss function combination; statistically significant improvements over original LiviaNET are presented in **bold**, determined by two-tailed, paired t-test ($p < 0.05$). The loss function used in each model is L1+L2+CE, unless noted otherwise.

	Control Subjects Dice Score (right / left)			
	Thalamus	Caudate	Putamen	Pallidum
Original LiviaNET	0.965 / 0.964 0.008 / 0.006	0.951 / 0.951 0.031 / 0.019	0.962 / 0.964 0.009 / 0.009	0.938 / 0.937 0.012 / 0.011
L+S	**0.969 / 0.971** **0.007 / 0.005**	**0.961 / 0.958** **0.017 / 0.017**	**0.972 / 0.974** **0.011 / 0.008**	**0.950 / 0.954** **0.021 / 0.014**
L+R	**0.971 / 0.971** **0.006 / 0.005**	**0.956** / 0.953 **0.029** / 0.028	**0.972 / 0.973** **0.008 / 0.008**	**0.953 / 0.955** **0.014 / 0.012**
L+S+R	**0.970 / 0.970** **0.008 / 0.006**	**0.962 / 0.959** **0.018 / 0.016**	**0.972 / 0.972** **0.007 / 0.007**	**0.951 / 0.954** **0.015 / 0.012**
L+S+R+E	**0.972 / 0.972** **0.006 / 0.005**	**0.961 / 0.956** **0.011 / 0.020**	**0.970 / 0.971** **0.007 / 0.007**	**0.952 / 0.955** **0.014 / 0.010**
L+S+R+E (CE only)	0.961 / 0.962 0.025 / 0.024	0.955 / 0.948 0.009 / 0.020	0.961 / 0.960 0.019 / 0.014	0.940 / 0.930 0.016 / 0.023

	HD Diagnosed Subjects Dice Score (right / left)			
	Thalamus	Caudate	Putamen	Pallidum
Original LiviaNET	0.955 / 0.957 0.021 / 0.013	0.820 / 0.868 0.240 / 0.149	0.924 / 0.921 0.060 / 0.067	0.855 / 0.887 0.110 / 0.056
L+S	0.957 / 0.958 0.031 / 0.023	**0.859 / 0.888** **0.215 / 0.119**	**0.911** / 0.915 **0.110** / 0.082	**0.818** / 0.857 **0.186** / 0.133
L+R	0.958 / **0.961** 0.041 / **0.014**	0.856 / 0.864 0.199 / 0.194	0.930 / 0.917 0.060 / 0.095	**0.882 / 0.906** **0.100 / 0.048**
L+S+R	**0.963 / 0.963** **0.020 / 0.015**	**0.875 / 0.894** **0.181 / 0.128**	0.931 / 0.933 0.064 / 0.055	0.882 / 0.901 0.123 / 0.058
L+S+R+E	**0.966 / 0.965** **0.016 / 0.011**	**0.893** / 0.887 **0.135** / 0.167	**0.945 / 0.934** **0.039 / 0.058**	**0.920 / 0.911** **0.043 / 0.055**
L+S+R+E (CE only)	**0.959** / 0.959 **0.016** / 0.010	**0.883** / 0.872 **0.143** / 0.153	0.936 / 0.921 0.030 / 0.076	**0.897** / 0.895 **0.032** / 0.040

between subcortical structures, which compensates for the global context missing in the patch-based method. The impact seems most pronounced for the caudate in diagnosed subjects, which is worth noting since the caudate is known to be severely atrophied in HD populations.

LiviaNET + Res-block ($L+R$). With the increased number of layers, the model can learn deeper features than the original LiviaNET. The skip connection from Res-block can help handle the degradation problem. After post-processing, Table 5.1 shows this model also has better performance than the original LiviaNET, again most notable in the caudate of diagnosed subjects. However, this model seems more prone to overfitting, which means the raw segmentation result (without post-processing, data not shown) contains more spurious islands than the original LiviaNET.

LiviaNET + Spatial coordinate + Res-block ($L+S+R$). As expected from (L+S) results, the spatial coordinate gave global information to the (L+S+R) model. This model has much less noise at the raw segmentation stage compared to the model without spatial coordinate as input (L+R). The improvement of this model over original LiviaNET is over 5 Dice points for the caudate in diagnosed patients. Qualitative results from this variant can be seen in Figure 5.3, which shows that our caudate segmentation is visibly

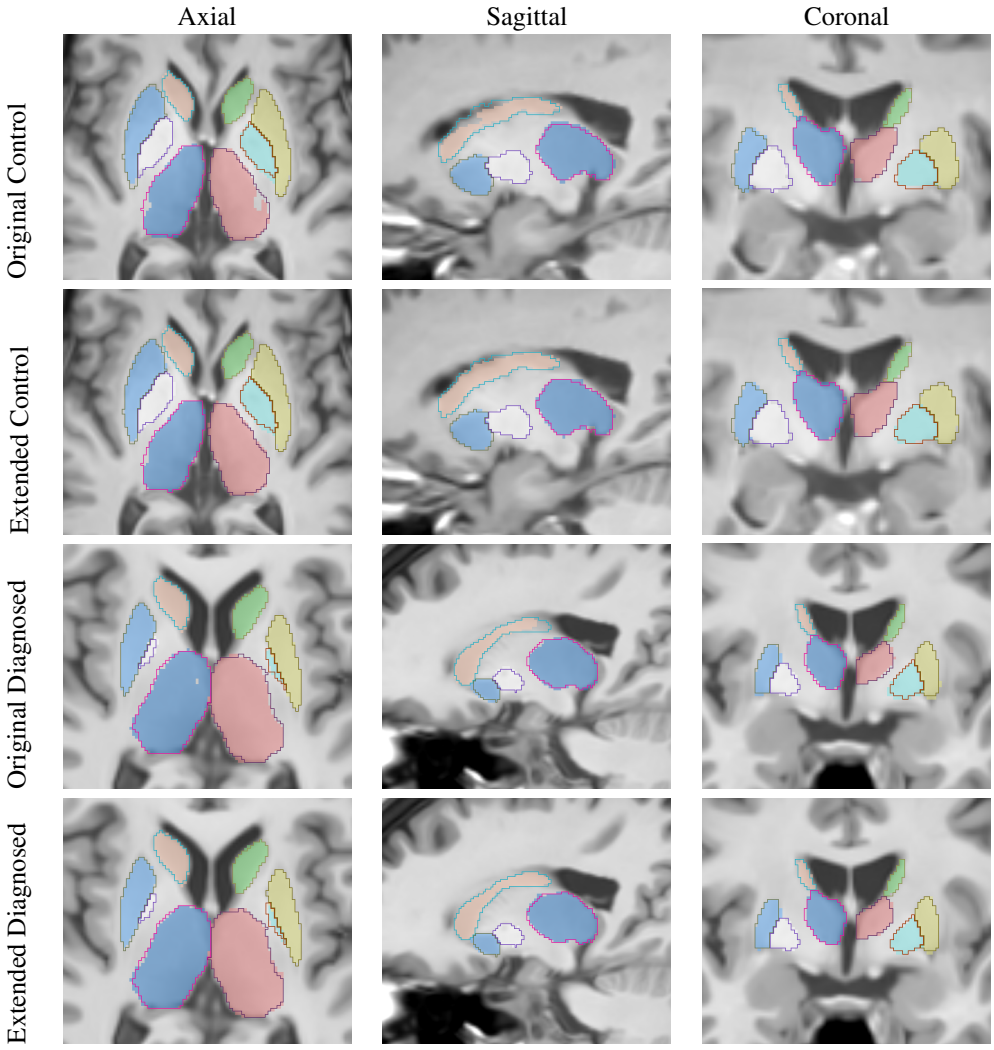


Figure 5.3: Qualitative results from the original LiviaNET and the L+S+R variant. The model results are shown in solid colors, whereas the ground truth is shown as contrasting outline. We note the noisy thalamus segmentations in the original are improved in the L+S+R variant. The caudate of the control subject is undersegmented by the original method, improved by the L+S+R extension. Perhaps most importantly, we also note the tail of the caudate in the diagnosed subject is missed by original but is picked up by the L+S+R extension.

improved for both control and HD subjects. After post-processing, the pre-manifest HD detailed comparison with LiviaNET is shown in Figure 5.4, where it can be clearly seen that the L+S+R variant improves the segmentation of atrophied structures, especially caudate and putamen.

LiviaNET + Spatial coordinate + Res-block + Elastic deformation ($L+S+R+E$). This model examines 2 different loss functions on training process (shown in the last two rows of Table 5.1): (1) only cross-entropy (CE) loss, and (2) the combination of CE, L1 and L2 loss functions. The combination loss function works better than CE alone, since the L1 and L2 loss functions can smoothly enhance the loss of

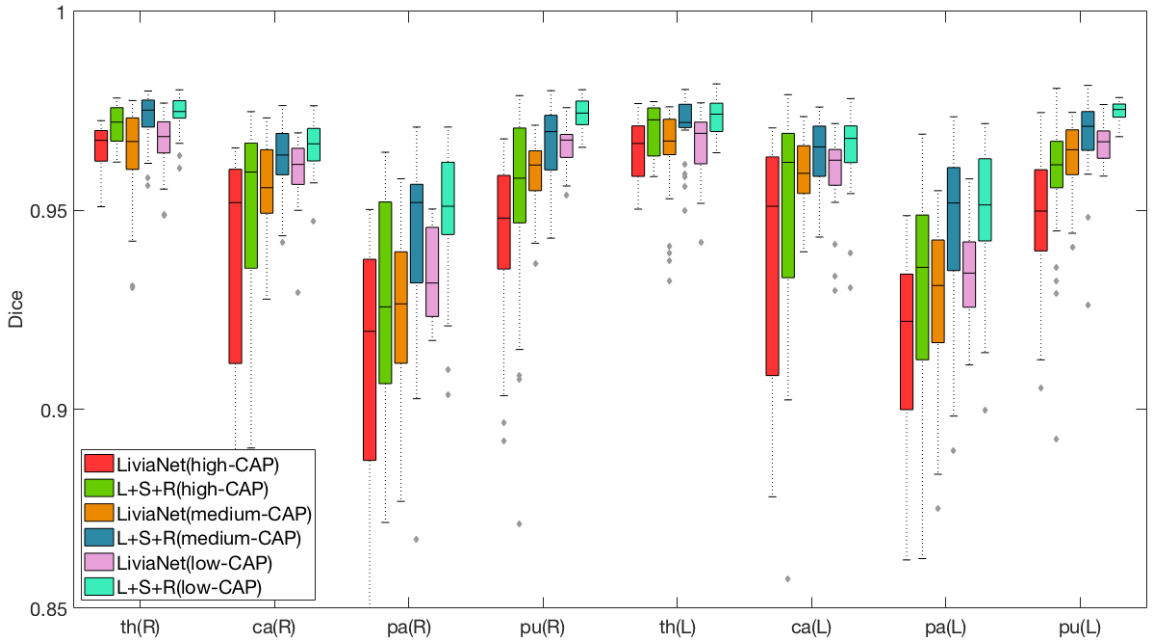


Figure 5.4: Segmentation performance in relation to disease progression. For pre-manifest HD subjects, the higher CAP score indicates more severe atrophy can be expected, and thus harder to segment using healthy training data. Compared to LiviaNet, the L+S+R variant is more robust to atrophied subcortical structures, specifically caudate and putamen which are known to be impacted by HD neurodegeneration.

false positive voxels, which produces the desired feedback to the model in back propagation. Moreover, with the combination loss function, the noise from raw segmentation is the least. While this model has the best Dice performance among the variants we explored, we have reservations about the stability of its use in practice, since the realism of the elastic data augmentation approach is somewhat questionable. A potential consequence of this instability can be observed in the putamens of HD subjects (Table 5.1, where the left putamen Dice is much lower than right; similar pattern can be observed even for the caudate of healthy controls). For these reasons, we focus on the more stable L+S+R model for Figs. 5.3 and 5.4. We nevertheless also report L+S+R+E as it may be preferable for certain applications.

5.4 Conclusion

In this work we explored different variants to the popular LiviaNET model for subcortical segmentation, with the goal of improving the generalization ability of the model to patients with neurodegeneration. Several of the modifications have indeed significantly improved the segmentation accuracy of subcortical structures in diagnosed HD patients, most prominently in the caudate, one of the most affected structures in HD. Exploring additional data augmentation strategies, such as GANs, remains as future work.

CHAPTER 6

MRI Subcortical Segmentation in Neurodegeneration with Cascaded 3D CNNs

The subcortical structures of the brain are relevant for many neurodegenerative diseases like Huntington's disease (HD). Quantitative segmentation of these structures from magnetic resonance images (MRIs) has been studied in clinical and neuroimaging research. Recently, convolutional neural networks (CNNs) have been successfully used for many medical image analysis tasks, including subcortical segmentation. In this work, we propose a 2-stage cascaded 3D subcortical segmentation framework, with the same 3D CNN architecture for both stages. Attention gates, residual blocks and output adding are used in our proposed 3D CNN. In the first stage, we apply our model to downsampled images to output a coarse segmentation. Next, we crop the extended subcortical region from the original image based on this coarse segmentation, and we input the cropped region to the second CNN to obtain the final segmentation. Left and right pairs of thalamus, caudate, pallidum and putamen are considered in our segmentation. We use the Dice coefficient as our metric and evaluate our method on two datasets: the publicly available IBSR dataset and a subset of the PREDICT-HD database, which includes healthy controls and HD subjects. We train our models on only healthy control subjects and test on both healthy controls and HD subjects to examine model generalizability. Compared with the state-of-the-art methods, our method has the highest mean Dice score on all considered subcortical structures (except the thalamus on IBSR), with more pronounced improvement for HD subjects. This suggests that our method may have better ability to segment MRIs of subjects with neurodegenerative disease.

6.1 Introduction

Subcortical structures are related to many neurodegenerative diseases, such as Alzheimer's, Parkinson's and Huntington's (HD) diseases [8]. To better understand HD, quantitative measurements of subcortical structures are essential. In the past, many automatic subcortical segmentation methods from magnetic resonance images (MRIs) were proposed [156, 219, 168, 47, 159, 158]. Many of these methods are based on multi-atlas label fusion, which is time-consuming due to the deformable registration process. In more recent years, convolutional neural network (CNN) based methods have dominated many segmentation fields with superior performance [241, 62, 71], and subcortical segmentation is no exception [38, 124, 225]. Dolz et al. [38] developed a fully convolutional neural network with small kernels and multi-concatenation to segment the subcortical area. Li et al.[124] explored variants based on the work of Dolz et al. [38] to improve the perfor-

This work is published at SPIE 2021.
Li, Hao, et al. "MRI subcortical segmentation in neurodegeneration with cascaded 3D CNNs." Medical Imaging 2021: Image Processing. Vol. 11596. SPIE, 2021.

mance of CNNs for subcortical segmentation. Wu et al. [225] proposed a joint 3D+2D framework to achieve accurate subcortical segmentation.

Hierarchically learning parameters with linear and non-linear layers, CNNs leverage local and global information from images for predicting segmentations. Once the training process is completed, CNN-based methods have high accuracy and are computationally efficient to predict the output. However, the ability of CNNs is highly dependent on the amount of training data, and it is desirable for the training data to be tightly matched to the test data. For neurodegeneration studies, small in-house datasets are widely used and it is hard to find public datasets with manual annotations. In such situations, generating new training data for each considered disease population is expensive and time-consuming. Alternatively, datasets of healthy control subjects may be used in training, and improving the generalizability of models from healthy controls to neurodegenerative disease populations is therefore important.

In this paper, we propose a 2-stage cascaded framework for subcortical segmentation, with a 3D CNN architecture at each stage. In the first stage, we downsample the original image and use it as input to create a coarse segmentation. Then we find an extended subcortical region of interest (ROI) from this coarse segmentation. In the second stage, we obtain the final segmentation from the cropped subcortical ROI at full resolution. Note that both stages use the same model, which is a 3D U-Net [35] based CNN. In our model, residual blocks [68] and attention gates [184] are used. Additionally, we connect feature maps from every level to preserve global information.

We evaluate our model on 2 datasets, the publicly available IBSR[1] and a subset of the PREDICT-HD database [143], which consists of both healthy control and HD subjects. For PREDICT-HD, we train our model only on healthy control subjects, and test on both healthy control and HD subjects to evaluate the generalization ability of our model. We use the Dice similarity coefficient as our metric to validate the results.

6.2 Methods

6.2.1 Segmentation Framework

We propose a 2-stage 3D subcortical segmentation framework (Fig. 6.1) to handle the memory limitation problem in 3D segmentation. After preprocessing, we feed downsampled images into the proposed CNN to create a coarse segmentation in the first stage. Then we upsample the coarse segmentation back to its original size, and find the bounding box of the subcortical areas. We extend the bounding box generously (see Fig. 6.1) to make sure any initially under-segmented areas are preserved. In the second stage, we crop this extended subcortical area at full resolution, and use this cropped area as the input of the proposed CNN

[1] https://www.nitrc.org/projects/ibsr

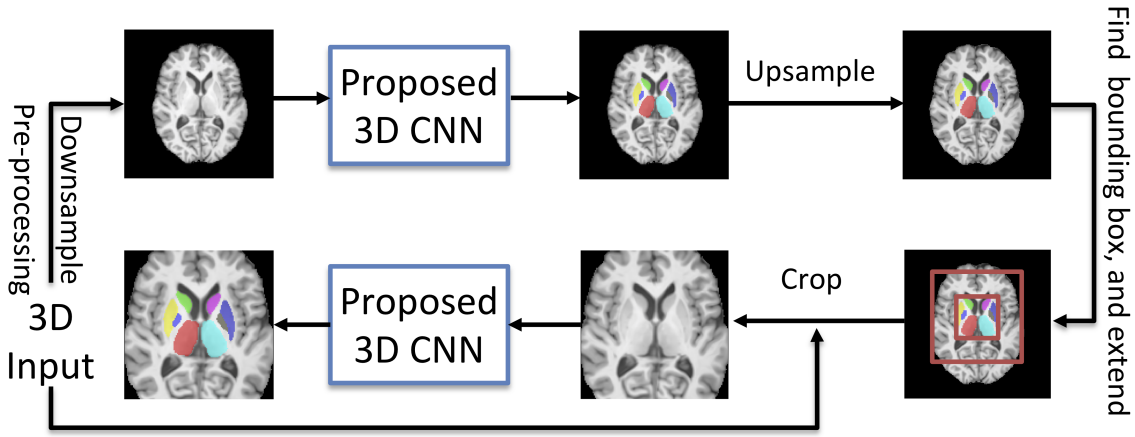


Figure 6.1: Overall workflow of the proposed framework.

to have the final segmentation.

6.2.2 Network Architecture

Our network (Fig. 6.2) is a 3D fully convolutional neural network adopted from the 3D U-Net [35]. Residual blocks [68] and attention gates [184] are used in our network to minimize the degradation problem and emphasize the ROI. In the residual blocks, batch normalization and ReLU-activation are followed by the convolution operation, and ReLU-activation is applied before outputting. We used 3D max-pooling and 3D nearest neighbor upsampling in the encoder and the decoder. Furthermore, we use skip connections between the encoder and the decoder, which preserves the coding information for the decoder. Before outputting the final segmentation, we further reinforce information by adding the outputs from the different scales.

6.2.3 Implementation Details

The Adam optimizer was used with L2 penalty 0.00001, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and initial learning rate 0.01. The learning rate was decayed by a factor of 0.5 every 50 epochs. Inspired by Milletari et al. [154], we used one minus mean of Dice coefficients from all labels as loss function during training, with equal weight ($w_{FG} = 1$) for all foreground labels and decayed weight ($w_{BG} = 0.1$) for the background. With a batch size of 2, we trained our model for 1000 total epochs and early stop was used in the training process. With the early stop, the training process is finished around 400 epochs and each epoch took 97 seconds. Total number of parameters of model is 36,290,202. The whole training process was conducted on an NVIDIA Titan RTX with 24 GB memory and was implemented using PyTorch.

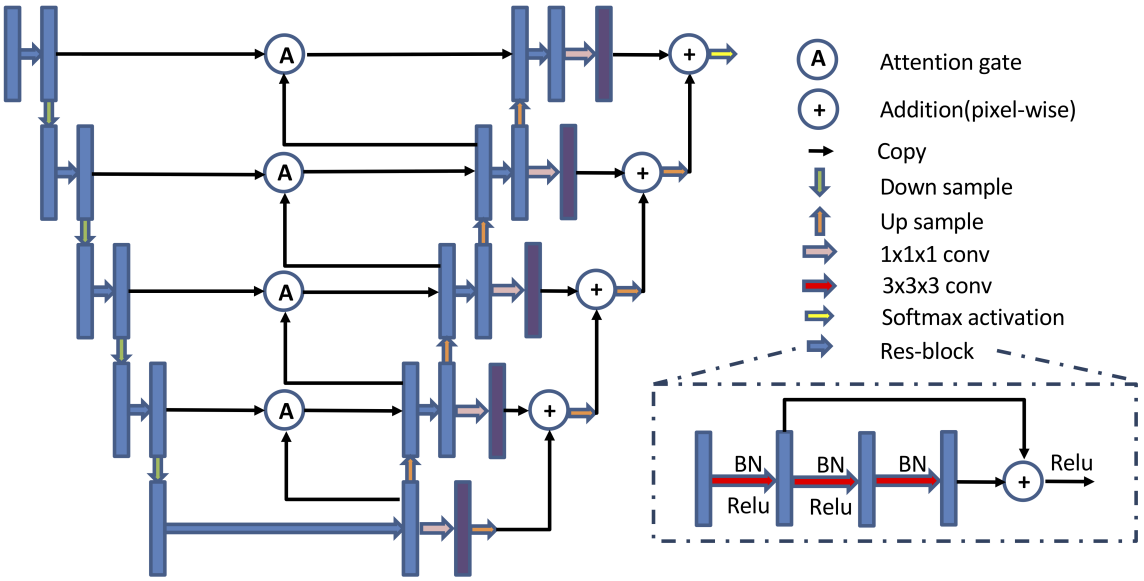


Figure 6.2: Proposed network architecture. Blue and purple boxes are feature maps. The number of channels for blue boxes at each level is 32, 64, 128, 256, 512 and the number of channels for purple boxes is 9 for all levels. The input size is $N \times 1 \times 128 \times 128 \times 96$, and the output size is $N \times 9 \times 128 \times 128 \times 96$, where N is the batch size and 9 is the number of output channels (8 subcortical structures+background).

6.2.4 Datasets and Preprocessing

Two datasets were used for our experiments. One of these is a publicly available dataset, which allows us to compare our results to state-of-the-art methods reported in the literature [225, 38]. The second dataset includes HD subjects, which allows us to assess the generalizability of our model to neurodegeneration studies.

The first dataset, IBSR, includes 18 T1w MRIs (resolution: $0.8 \times 0.8 \times 1.5\text{mm}^3$ to $1.0 \times 1.0 \times 1.5\text{mm}^3$). The publicly available manual segmentations are used as ground truth for this dataset. We randomly select 12, 3, and 3 as training, validation and testing sets respectively. Leave-three-out cross-validation is used on this dataset. For preprocessing, we use the publicly provided skull-stripped images and normalize the intensities using histogram matching.

The second dataset is a subset of the multi-site PREDICT-HD study [143] (1mm isotropic resolution T1w), which is the same as described in Sec. 5.2.6 of Chapter 5. This dataset is used to validate the innovative methodology compared to the previously proposed network for subcortical segmentation(Chapter 5), e.g, the comparison between U-shaped and single-path networks. The experimental settings and preprocessing steps are kept the same, except that histogram matching is applied. Briefly, from the healthy controls, we randomly select 20 subjects for training and 2 for validation. The remaining 15 healthy controls and all HD subjects are used for testing, to assess the generalizability of our model to atrophied brains.

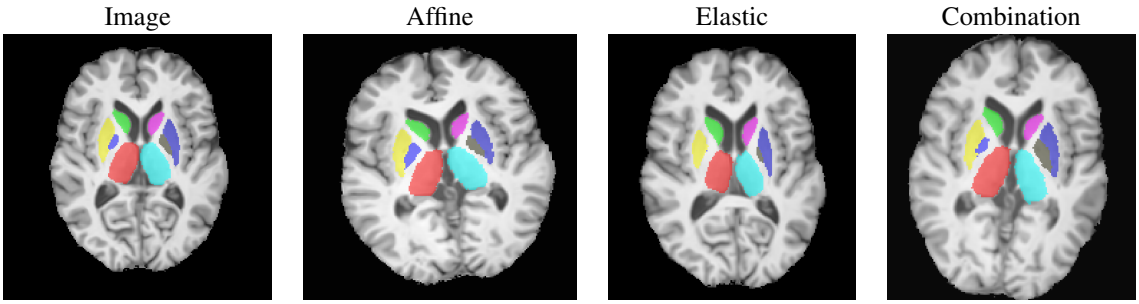


Figure 6.3: Examples of the used data augmentations. Left to right: original image, random affine, random elastic deformation, combination of random affine and random elastic deformation.

6.2.5 Data Augmentation

We used 3 types of random data augmentation: affine transformation, elastic deformation, and combination of affine and elastic deformation. In addition to reducing overfitting, these augmentations serve the purpose of mimicking atrophy present in HD, such that the model can be trained on only healthy subjects and still generalize well to HD subjects. Fig. 6.3 shows examples of augmentations.

6.3 Results

Tab. 6.1 shows the results of all 8 considered subcortical structures for the IBSR dataset. Compared to the results reported by Wu et al. [225], our method produced higher mean Dice scores except for the left thalamus. Compared to Dolz et al. [38], our method obtained superior performance for all structures except the thalamus.

For the PREDICT-HD dataset, we trained our model only on healthy control subjects, and tested on both diagnosed HD patients and healthy controls. The results are shown in Tab. 6.2. Bold numbers indicate significant improvements ($p < 0.05$, 2-tail, paired t-test) compared to the work from Li at al. [124]. Additionally, Dice scores from all subcortical structures were significantly better than the method from Dolz et al. [38] (these are not explicitly denoted on Table 6.2, for brevity). We observe that the improvement is especially pronounced for the caudate, putamen and pallidum of the HD subjects, which is noteworthy since these struc-

Table 6.1: Dice scores on IBSR, presented as $mean \pm std.\ dev.$. Highest mean Dice scores are presented in **bold**. The left and right structures are reported jointly by Dolz et al. [38]. The train/test splits may differ between the compared methods, as these are not reported [225, 38].

	R thalamus	L thalamus	R caudate	L caudate	R pallidum	L pallidum	R putamen	L putamen
Wu et al.	0.917±0.013	0.913±0.013	0.899±0.022	0.898±0.019	0.839±0.042	0.846±0.027	0.910±0.015	0.911±0.015
Dolz et al.	**0.92**		0.91		0.83		0.90	
Proposed	0.918±0.016	0.913±0.015	**0.915±0.000**	**0.909±0.008**	**0.858±0.016**	**0.876±0.026**	**0.912±0.020**	**0.917±0.007**

Table 6.2: Dice scores on PREDICT-HD, presented as $mean \pm std.\ dev.$ Compared to the method from Li et al. [124], significant improvements ($p < 0.05$ with 2-tailed paired t-test) are presented in **bold**. For all 8 structures, the proposed method significantly outperformed the Dolz et al. [38] method.

Control Subject Dice score

	R thalamus	L thalamus	R caudate	L caudate	R pallidum	L pallidum	R putamen	L putamen
Dolz et al.	0.965±0.008	0.964±0.006	0.951±0.031	0.951±0.019	0.938±0.012	0.937±0.011	0.962±0.009	0.964±0.009
Li et al.	0.970±0.008	0.970±0.006	0.962±0.018	0.959±0.016	0.951±0.015	0.954±0.012	0.972±0.007	0.972±0.007
Proposed	**0.972±0.006**	**0.973±0.005**	0.965±0.008	0.961±0.013	**0.963±0.009**	0.956±0.029	**0.976±0.004**	0.974±0.009

Diagnosed Subject Dice score

	R thalamus	L thalamus	R caudate	L caudate	R pallidum	L pallidum	R putamen	L putamen
Dolz et al.	0.955±0.021	0.957±0.013	0.820±0.240	0.868±0.149	0.855±0.110	0.887±0.056	0.924±0.060	0.921±0.067
Li et al.	0.963±0.020	0.963±0.015	0.875±0.181	0.894±0.128	0.882±0.123	0.901±0.058	0.931±0.064	0.933±0.055
Proposed	**0.970±0.007**	**0.970±0.008**	**0.925±0.080**	**0.932±0.059**	**0.933±0.028**	**0.938±0.019**	**0.959±0.018**	**0.955±0.023**

tures are known to be impacted in HD. This suggests our method may have better generalizability to brains that present neurodegeneration.

Finally, our method produced the highest mean Dice scores over all subcortical structures for the 3 pre-manifest HD populations as well, as can be seen in Figure 6.4. This further suggests that our model is more generalizable to atrophied structures. Importantly, the Dice score of our method is more consistent across the disease progression (low-, medium-, high-CAP) than the compared methods, which suggests that it is more robust to increasing amounts of atrophy.

The qualitative results are shown in Fig. 6.5. We observe that the proposed method is visually the most similar to the 'ground truth'. Our method is also more smooth than the 'ground truth', as can be seen in the thalamus of the control subject (red arrows). We further observe that the thin tail of the caudate in the HD subject (diagnosed) is captured well by the proposed approach, but it is substantially under-segmented by the other two methods (green arrows). Furthermore, our proposed approach produced superior putamen segmentations for the HD subject (diagnosed), which can be observed in the axial and coronal views (blue arrows).

It is noteworthy that Li et al. [124] used spatial coordinates to encapsulate global context, but this is a relatively inefficient approach. Our currently proposed model better preserves global context thanks to its more robust architecture, and it produces superior results as shown in Fig. 6.5 and Tab. 6.2. Finally, we note that these two methods [38, 124] include a postprocessing step to eliminate spurious islands from the CNN result, while our approach does not require such a postprocessing step.

6.4 Discussion and Conclusion

In this work, we proposed a 2-stage cascaded framework for subcortical segmentation, with a 3D CNN architecture at each stage. Our results indicate that our method is the best among the compared state-of-the-art methods. Our model can be trained efficiently on the relatively small IBSR dataset, and it generalizes well

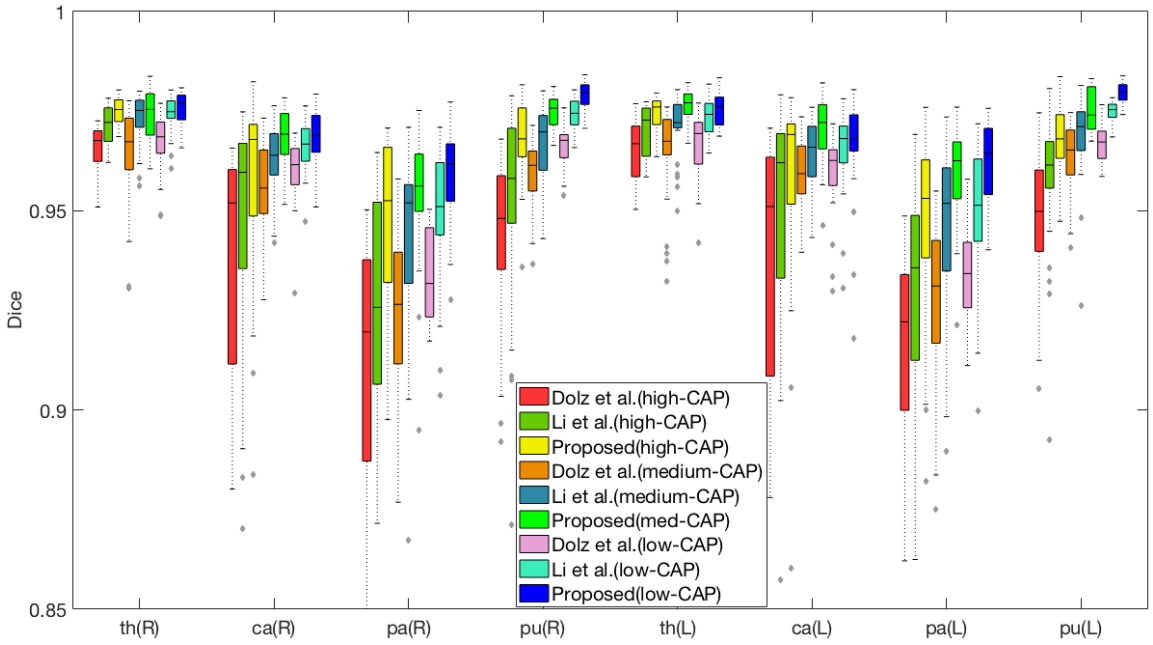


Figure 6.4: Segmentation performance of 39 HD pre-manifest subjects (13 subjects in each category). Higher CAP scores indicate patients further along the HD disease progression, and are associated with more atrophy. The horizontal axis lists the (R)ight and (L)eft pairs of thalamus, caudate, pallidum and putamen.

from healthy training subjects to HD test subjects in the PREDICT-HD dataset. Compared to the state-of-the-art methods, our model produces the highest mean Dice scores on all considered subcortical structures, except for the thalamus on the IBSR dataset. Furthermore, for HD subjects, our model has a dramatic improvement of accuracy for the caudate and the putamen, which are the most atrophied subcortical structures in HD [144, 143, 171]. These findings indicate that our method has better generalizability not only to unseen healthy subjects, but also from healthy controls to an HD population. Our findings suggest that the skip connections, residual blocks, attention gates, and output adding in our architecture all play important roles for preserving global information and improving segmentation accuracy.

This study also indicates several directions for further research. Tables 6.1 and 6.2 suggest that our method may have more consistent performance between left and right hemispheres compared to the alternative methods (e.g., diagnosed pallidum in Table 6.2). This behavior could potentially be further enhanced by developing a symmetric attention gate. Additionally, the loss function could be further investigated to improve segmentation. Finally, validating our method in a larger dataset as well as exploring multi-modal segmentation remain as future work.

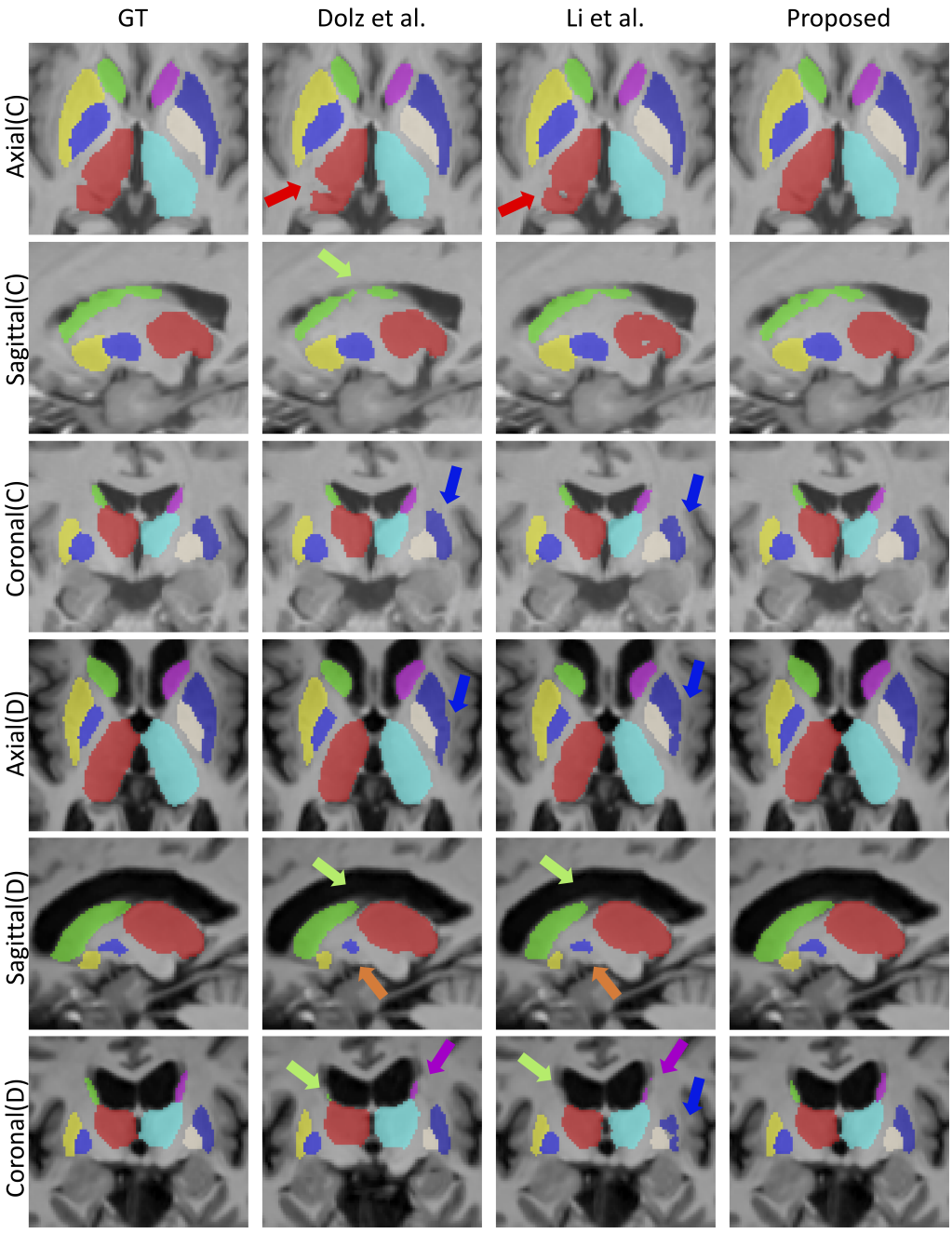


Figure 6.5: The comparison of segmentation results. Top 3 rows are from a (C)ontrol subject and bottom 3 rows are from a (D)iagnosed HD subject. Methods are marked on the top. Our method has better performance in the thalamus **(red arrows)**, the putamen **(blue arrows)**, the pallidum **(orange arrows)**, as well as on the caudate tail **(green and purple arrows)**.

CHAPTER 7

Towards Robust MRI Subcortical Segmentation in Huntington's Disease Using Deep Networks: A Large-scale Study

Subcortical brain structures are relevant for Huntington's disease (HD), and especially the caudate and putamen volumes are known to be critical imaging-based markers of disease progression. Quantitative measurements of subcortical structures from magnetic resonance (MR) images provide a better way to monitor the progression of HD, which may correlate with the motor function of HD subjects. Thus, a reliable automatic subcortical segmentation method is essential for HD studies. In this work, we validate the proposed deep learning framework with a coarse-to-fine strategy for robust subcortical segmentation in Chapter 6 with a large-scale study. This dataset includes control subjects and an HD population, covering a wide range of disease statuses and ages. The thalamus, caudate, pallidum, and putamen are segmented in both the left and right hemispheres. Compared to the current state-of-the-art methods, the comprehensive experiments show our method produces superior subcortical segmentations with higher Dice score, lower average surface distance, and lower 95th-percentile Hausdorff distance. In addition, the results of our ablation studies could benefit model selection and network configuration for future studies. Furthermore, our findings show our method to be suitable for clinical HD studies, ensuring reliable segmentation of subcortical structures across disease stages and ages. It is important to highlight that the proposed method bridges the gap in automatic segmentation algorithms designed for a large HD cohort.

7.1 Introduction

The subcortical structures of the brain are relevant to many neurodegenerative diseases, such as Huntington's disease (HD) [169, 16], which is an inherited, autosomal-dominant, progressive, neurodegenerative disorder [209]. Based on the number of repetitions of a cytosine-adenosine-guanine (CAG) trinucleotide, an expansion of 36 or more repetitions can be classified as an HD subject. Compared to healthy control (HC) subjects (CAG repeat length < 26), the motor and cognitive function are progressively impaired in HD subjects [209, 54]. Moreover, the age at the onset of HD has a strong association with CAG length [3]. The length of the CAG repeat expansion is inversely related to the age at onset of the disease. Typically, the longer the CAG repeat length, the earlier the onset of symptoms. Although age and CAG length are usually used as primary factors when predicting the clinical onset of HD, predicting the time of clinical manifestation of HD lacks precision and needs to be improved.

This work is currently being finalized for submission.

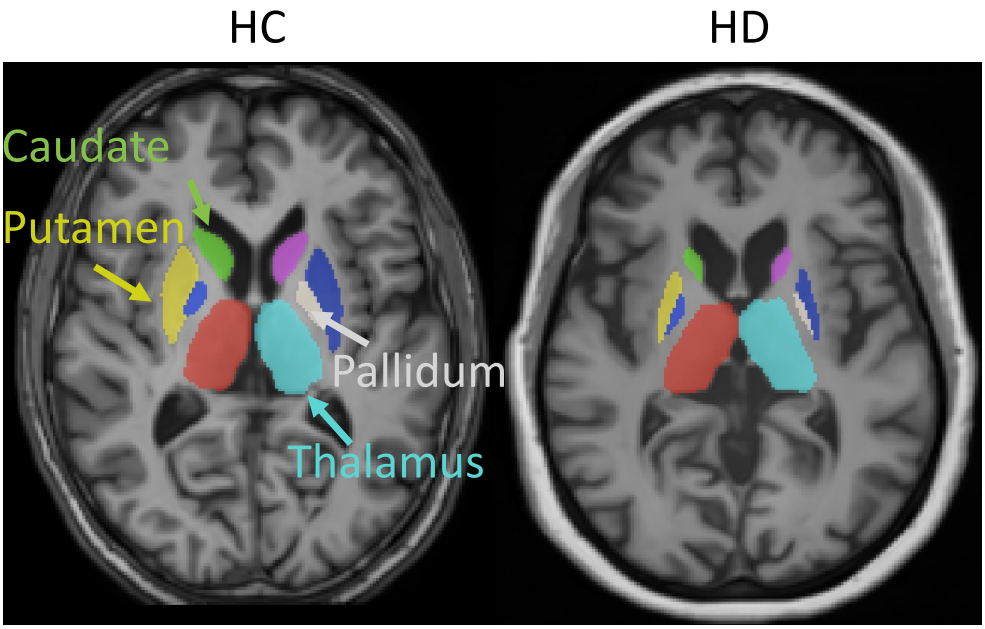


Figure 7.1: Example T1-weighted MR images of healthy control (HC) and Huntington's disease (HD) subjects are presented. The subcortical structures are marked in color. The images are produced by different scanners, resulting in different image appearances known as domain shift. In addition, structural variations based on healthy conditions are also presented. Compared to the HC subject, we observe an atrophied caudate and putamen in the HD patient; the volumes of these structures are important bio-markers of HD progression.

In practice, magnetic resonance (MR) images are widely used in HD research and even clinical trials [54]. Literature shows that subcortical structure volumes can help improve the prediction of manifest HD onset [171]. In particular, the caudate and putamen are the primary imaging-based markers of HD, showing atrophy in HD subjects throughout the disease progression (see Fig. 7.1). However, it is not easy to obtain volumes of subcortical structures from MR images to support the clinical findings of HD studies. This is because generating human delineations of subcortical structures from 3D MRIs as the gold standard is time-consuming and expensive, especially for large-scale datasets, which are often required for clinical studies. Additionally, annotators are required to have experience and anatomical knowledge. The exact size and shape of subcortical structures can vary among individuals, making labeling especially challenging for the HD cohort. Image quality and resolution influence the consistency of delineations. Furthermore, the reproducibility of human raters may not be guaranteed due to intra-rater and inter-rater variability.

To support large-scale studies, it is vital to develop a reliable automatic algorithm for subcortical segmentation. This reduces human effort and ensures consistent segmentations that are standardized and comparable to expert manual annotations, especially for extensive HD datasets.

Over the past few decades, researchers have developed automatic subcortical segmentation methods to produce high-quality results, and these methods are still employed in practice [47, 168, 156, 216, 219, 159,

158]. Many of these methods were based on multi-atlas label fusion, one of the widely used methods for medical image segmentation with superior performance at the time. By utilizing labeled training images, i.e., atlases, these methods could infer reliable segmentations for a target image via label fusion after registering multiple atlases to the target image. Multi-atlas segmentation methods rely on an accurate registration approach and advanced label fusion technique to produce a robust segmentation with spatial consistency. However, despite their good segmentation performance, these methods are time-consuming for inference due to the computational complexity of the deformable registration process.

In recent years, as convolutional neural networks (CNNs) have become increasingly popular in solving medical imaging challenges [133, 137], deep learning-based methods [113] have demonstrated their ability to learn complex patterns and structures within medical datasets. These methods have delivered superior performance in many medical image analysis tasks, particularly in segmentation [35, 101, 38, 257, 241, 92, 78, 240, 117, 134, 239, 138]. As the foundational unit of deep learning, a CNN comprises multiple convolutional, pooling, and fully connected layers that use both linear and nonlinear operations with a significant number of learnable parameters. Each layer progressively transforms its input data into a more abstract representation than the previous layer. By hierarchically learning parameters through backpropagation, a CNN predicts segmentations by leveraging both local and global information from feature maps. Once trained, the deep learning model can make predictions on unseen data by processing the input through its layers to produce an output. Deep learning-based methods have several advantages over traditional image segmentation methods: (1) Automated feature extraction: CNNs can automatically learn and extract complex features from raw image data, potentially capturing more relevant information for the segmentation task while traditional segmentation techniques often require manual extraction and definition of features; (2) High performance: CNNs have demonstrated superior performance in a wide range of medical image segmentation tasks; (3) Adaptability: CNNs can be easily adapted to different segmentation structures or tasks with appropriate training data; (4) Computation efficiency: Deep learning-based methods have a significant advantage in computing efficiency during inference.

To date, the transformer is one of the latest advanced deep learning techniques widely used in medical image segmentation [254, 28, 64, 116, 79, 204, 20, 190, 128, 123]. The transformer converts an input image into a sequence of patches rather than analyzing the entire input. Its primary advantage over CNNs is the ability to model long-range dependencies using the self-attention mechanism, allowing it to interact with all elements of the image as opposed to just a localized field of view. Such a global perspective proves especially valuable in medical image segmentation where distant parts of the image can offer crucial contextual information. In practice, they can replace CNNs [20, 254], be combined with CNNs [247, 116, 64, 123], or integrated into CNNs [206, 28].

Although medical image segmentation with deep learning achieves high performance, it still faces some remaining challenges in producing robust segmentations among a large HD population. These challenges include:

1. Structural variations: As HD progresses, it leads to significant atrophy and morphological changes in subcortical structures, especially for caudate and putamen. The ability of CNNs is highly dependent on the amount of training data, and it is desirable for the training data to be tightly matched to the test data. A model trained on data from a specific stage of the disease, or on a particular cohort including controls, may struggle to generate accurate segmentations for patients at different stages of the disease [124, 127].

2. Image appearance variations: The intensity and contrast differences are often present among different imaging sites due to different scanners and MRI protocols which is known as domain shift [60, 119, 234]. The performance of deep learning models are not guaranteed when encounter domain shift, particularly given the additional variability introduced by HD-related changes (see Fig. 7.1).

3. Data collection and label availability: Given that HD is a rare disease, significant efforts, such as data privacy and accessibility, are being made to collect large amounts of HD data from multiple sites. In addition, obtaining human delineations requires experience and anatomical knowledge, and it is both time-consuming and expensive, especially for large-scale datasets.

4. Computational requirements: deep learning models can be computationally intensive when processing large 3D volumetric data and require a large amount of computational resources. It may cause challenges in existing complex clinical workflow by disrupting and delaying various stages.

Therefore, it is important to design a viable deep learning algorithm to tackle these practical problems in support of clinical HD studies.

In this work, we propose a deep learning framework for subcortical segmentation in a large HD population, which is designed for practical clinical use. The proposed framework is formulated in a coarse-to-fine manner, with cascaded CNNs from two separate stages, to effectively and accurately segment both the left and right sides of the thalamus, caudate, pallidum, and putamen (see Fig. 7.1). To enhance efficiency, the coarse stage aims to locate a region of interest (ROI) for the target subcortical structures, and based on this identified ROI, the fine stage is designed to produce robust segmentations. As shown in Fig. 6.1, the coarse segmentation is generated using a downsampled image as input during the initial stage. Subsequently, the ROI is identified based on this coarse segmentation. Finally, this cropped ROI is utilized as input during

the fine stage, resulting in the production of the final segmentations. To ensure accuracy, histogram matching [157, 188, 179] is used in the preprocessing pipeline to tackle the domain shift problem in our multi-site dataset. We used 3D U-Net [35] as the baseline model at each stage, and the residual blocks [68], attention gates [184] and deep supervision [126] techniques are incorporated as modifications to improve segmentation results as well as the data augmentation to mimic structural variations. This is an extension of our previous conference paper [126], and the main contributions are summarized as follows:

- We proposed a deep learning framework for 3D subcortical segmentation that is designed for practical clinical use with a large-scale dataset (Chapter 6). The proposed framework employs a coarse-to-fine strategy to enhance performance and reduce computational costs. Comprehensive results in Sec. 7.3 indicate that the proposed framework also improves other segmentation methods.
- This approach aligns well with HD studies since subcortical structures are expected to atrophy during disease progression. The extensive experiments (Sec. 7.3) may assist clinicians in model configuration and selection for future HD or other neurodegenerative studies.
- We evaluated the proposed methods on a large-scale multi-site dataset, PREDICT-HD [143]. Through comparisons with the latest SOTA methods listed in Sec. 7.2.3, our extensive experimental results show that our methods offer superior performance in subcortical segmentation in the presence of appearance and structural variations. Furthermore, the effectiveness of our methods, despite using a limited amount of training data, indicates that it can handle new datasets, enhancing its value for clinical research.

7.2 Experimental Settings

7.2.1 Study Design

We comprehensively evaluated our subcortical segmentation methods proposed in Chapter. 6 using a large-scale dataset. The Dice Similarity Coefficient (DSC), Average Surface Distance (ASD), and 95th percentile Hausdorff Distance (HD95) are used as evaluation metrics to assess the accuracy of segmentations.

7.2.2 Dataset

We utilized a large-scale dataset from the PREDICT-HD study [143] that was collected from multiple imaging sites. This dataset includes 1068 labeled 3D T1-weighted MR images from 390 subjects, all of which have a $1mm$ isotropic resolution. Each subject has scans at multiple time-points. The participants fall into two main categories: healthy control subjects and HD patients. Furthermore, HD patients are categorized into either manifest or pre-manifest groups. Using the CAP (CAG-Age Product) score [248], a commonly used index for tracking HD progression, pre-manifest patients can be classified into low-CAP (CAP $<$ 290), med-CAP (290

Table 7.1: Overview of PREDICT-HD datasets [143] in the proposed work. It has 1068 labeled 3D T1-weighted MR images from 390 subjects. The CAP score is the well-known measurement for HD progression [143], e.g. CAP=$(CAG - 33.66) \times Age$, where CAG is the length of cytosine-adenosine-guanine expansion and Age is the age of the subject at the time of first MR scan [248]. Patients in the pre-manifest stage are categorized based on their CAP scores: low-CAP (CAP $<$ 290), med-CAP (290 $\leq$ CAP $<$ 368), and high-CAP (CAP $\geq$ 368).

	Total	Training	Validation	Test
Gender (Female/Male)	255/135	51/27	26/13	178/95
Age (mean/std.ev. in years)	46.80/12.59	46.76/13.37	46.54/13.37	46.85/12.41
Scanner type	25	21	21	24
Subjects (S)/Images (I)	390/1068	78/212	39/103	273/753
Control (S/I)	110/328	20/62	12/36	78/228
Low-CAP (S/I)	97/259	23/60	11/26	63/173
Med-CAP (S/I)	102/272	19/49	7/21	76/202
High-CAP (S/I)	81/211	16/41	9/20	56/150
HD-diagnosed (S/I)	55/119	13/37	8/5	34/77

Table 7.2: Compared methods. Hybrid indicates that the method employs a "CNN+transformer" architecture. $(|)$ denotes the method where the transformer replaces the CNN-based encoder, and $(||)$ indicates the method that employs the transformer as an additional encoder alongside the CNN-based encoder.

Methods	Toolbox	Pretrained	Self-configurated	Pure CNN	Pure transformer	Hybrid	Official code
FasterSurfer [75]	✓	✓	-	✓	-	-	✓
SLANT [89]	✓	✓	-	✓	-	-	✓
3D U-Net [35]	-	-	-	✓	-	-	✓
V-Net [154]	-	-	-	✓	-	-	✓
nnUnet [92]	✓	-	✓	✓	-	-	✓
LiviaNET [38]	-	-	-	✓	-	-	✓
UNETR [64]	-	-	-	-	-	✓(|)	✓
CATS [116]	-	-	-	-	-	✓(||)	✓
Swin-UNETR [204]	-	-	-	-	-	✓(|)	✓
Swin-UNET [20]	-	-	-	-	✓	-	✓

$\leq$ CAP $<$ 368), and high-CAP (CAP $\geq$ 368) groups. Details regarding population distribution are provided in Table 7.1. We randomly assigned 78/39/273 subjects, corresponding to 212/103/753 images, for training, validation, and testing, respectively. It's important to note that all time-points from a single subject were included in the same subset (whether training, validation, or testing) to prevent cross-contamination.

7.2.3 Compared Methods

We compared the proposed method to existing SOTA methods designed for medical image segmentation, which are shown in Table 7.2. The comparative methods include two widely used toolboxes: Fastersurfer [75] and SLANT [89], which are the pretrained models with the large number of training samples from T1-weighted MRI. Additionally, we also included another toolbox, nnUnet from [92], a self-configurated framework with SOTA performance in different medical image segmentation challenges. Moreover, two popular CNNs are included, e.g., the LiviaNET and V-Net from the work of [38] and [154], respectively. To date, the

widely used transformer are also included in our experiments, and they can be classified into two categories: hybrid CNN-transformer models Li et al. [116], Hatamizadeh et al. [64] and pure transformers [20]. Within the hybrid category, some utilize transformers as encoders, replacing the traditional CNN-based encoder (as seen in UNETR and Swin-UNETR), while CATS introduces an additional transformer encoder parallel to the CNN-based encoder. We ensured a fair comparison by using the official, publicly available codes of all the methods in our study. Notably, our primary distinguishing factor in this comparison was the model architecture employed in the fine-tuning stage. We excluded the Fastersurfer, SLANT, and nnUnet from this criterion, since they are designed as ready-to-use toolboxes. Specifically, Fastersurfer and SLANT are pretrained with large amount of training data and nnUnet has its own training pipeline. The original Swin-UNET was designed for the 2D task, and we implemented it in a 3D manner based on the original official implementation. Compared methods (except Fastersurfer and SLANT) maintained a consistent train-test split, and use same cropped images produced from the coarse stage of the proposed framework as input, except the nnUnet. Our experiments were focused on segmenting eight subcortical structures, including the left and right sides of the thalamus, caudate, pallidum, and putamen.

7.3 Results

7.3.1 Overall Performance

Quantitative results. Table 7.3 presents a comprehensive comparison of results for segmenting subcortical structures for the PREDICT-HD dataset. In addition, the best performance and significant improvements are denoted as bold and *, respectively. A clear observation from the table indicates that our method consistently delivers superior segmentation for all targeted subcortical structures across every evaluation metric. Specifically, we not only achieved results in terms of DSC with a higher mean but also maintained a low standard deviation. Compared to the second best method in each column, the proposed method consistently has significant improvements, which are tested by paired 2-tail t-test with p-value < 0.001. Compared to other methods, we observed expected pronounced improvements in DSC for the caudate, pallidum, and putamen. These structures are known to be impacted during HD progression, especially the caudate and putamen, and thus more precise measurements can help with downstream analysis. Furthermore, results from Table 7.3 indicate that our method effectively segments smaller structures like the pallidum and provides robust segmentation for neurodegeneration cohorts.

From Table 7.3, FasterSurfer and SLANT [75, 89] produce undesired results. These methods, designed as ready-to-use toolboxes with pretrained models trained on extensive data, suggest that such pretrained approaches might not always be robust when tested on new datasets. In contrast, other methods that trained with PREDICT-HD dataset could produce more accurate segmentations, especially for the 3D methods, which can

Table 7.3: Comprehensive quantitative results for Dice similarity coefficient (DSC), averaged surface distance (ASD) and 95-th percentage Hausdorff distance (HD95) are reported as $mean \pm std.\,dev.$ ↑ and ↓ denote higher is better and low is better, respectively. Bold text indicates the best performance. * denotes significant differences between proposed and the second best method with paired 2-tail t-test ($p < 0.001$). The subcrotical structures from left to right are: thalamus (th), caudate (ca), pallidum (pa) and putamen (pu) on the right (r) and left (l) side. The last column (overall) is derived from all considered subcortical structures.

Dice similarity coefficient ↑

	th(r)	th(l)	ca(r)	ca(l)	pa(r)	pa(l)	pu(r)	pu(l)	overall
FasterSurfer	.908±.011	.897±.012	.828±.092	.860±.040	.801±.038	.789±.041	.841±.034	.863±.029	.848±.059
SLANT	.925±.012	.918±.012	.887±.037	.905±.022	.873±.035	.882±.036	.912±.029	.916±.030	.902±.034
3D U-Net	.972±.005	.971±.006	.965±.011	.965±.011	.943±.019	.943±.021	.964±.011	.966±.010	.961±.017
V-Net	.946±.011	.941±.012	.926±.021	.932±.021	.876±.044	.874±.044	.910±.027	.912±.027	.914±.038
nnUnet (2D)	.970±.042	.970±.037	.960±.044	.961±.041	.938±.043	.937±.044	.962±.043	.962±.041	.957±.043
nnUnet (3D)	.976±.040	.977±.029	.972±.030	.974±.026	.962±.053	.961±.041	.972±.065	.974±.045	.971±.043
LiviaNET	.969±.006	.968±.006	.953±.071	.955±.037	.937±.019	.937±.054	.962±.011	.964±.012	.956±.037
UNETR	.969±.007	.969±.007	.965±.012	.965±.013	.937±.023	.936±.023	.963±.014	.961±.013	.958±.020
CATS	.975±.006	.974±.006	.971±.011	.971±.011	.955±.016	.953±.019	.973±.009	.973±.009	.968±.014
Swin-UNETR	.979±.005	.978±.005	.972±.010	.972±.010	.960±.014	.959±.016	.977±.005	.977±.008	.972±.013
Swin-UNET	.972±.005	.972±.005	.965±.011	.964±.011	.949±.017	.948±.018	.967±.010	.968±.011	.963±.015
Proposed	**.982±.004***	**.982±.004***	**.975±.009***	**.975±.009***	**.969±.011***	**.968±.012***	**.980±.006***	**.981±.006***	**.977±.010***

Averaged surface distance (*mm*) ↓

	th(r)	th(l)	ca(r)	ca(l)	pa(r)	pa(l)	pu(r)	pu(l)	overall
FasterSurfer	.420±.069	.507±.066	.750±1.06	.556±.173	.659±.152	.763±.167	.577±.118	.483±.094	.589±.411
SLANT	.308±.068	.346±.070	.190±.047	.179±.052	.258±.092	.243±.090	.228±.088	.225±.092	.247±.093
3D U-Net	.084±.022	.089±.026	.047±.021	.046±.023	.080±.042	.082±.049	.069±.029	.063±.032	.070±.036
V-Net	.217±.065	.255±.079	.134±.052	.122±.053	.261±.123	.266±.117	.239±.095	.226±.086	.215±.102
nnUnet (2D)	.145±.779	.143±.640	.203±1.47	.170±1.12	.204±1.58	.194±1.52	.214±1.77	.226±1.75	.187±1.39
nnUnet (3D)	.066±.326	.115±.612	.068±.575	.065±.615	.116±.1.45	.217±1.73	.170±1.88	.328±2.29	.122±1.04
LiviaNET	.099±.032	.103±.030	.562±6.82	.196±3.42	.099±.046	.372±5.39	.075±.027	.067±.028	.197±3.31
UNETR	.094±.033	.101±.036	.046±.024	.045±.023	.090±.049	.095±.053	.069±.039	.075±.038	.077±.043
CATS	.071±.027	.077±.029	.036±.019	.036±.021	.062±.040	.064±.053	.047±.025	.046±.027	.055±.035
Swin-UNETR	.054±.017	.056±.019	.035±.018	.034±.021	.046±.027	.048±.033	.036±.020	.036±.019	.043±.024
Swin-UNET	.079±.021	.079±.023	.046±.020	.045±.021	.064±.032	.068±.038	.057±.022	.055±.026	.062±.029
Proposed	**.046±.026***	**.044±.018***	**.027±.014***	**.028±.018***	**.032±.018***	**.035±.023***	**.028±.014***	**.028±.017***	**.033±.020***

95-th percentage Hausdorff distance (*mm*) ↓

	th(r)	th(l)	ca(r)	ca(l)	pa(r)	pa(l)	pu(r)	pu(l)	overall
FasterSurfer	1.71±0.32	1.81±0.29	2.85±1.91	2.38±0.97	2.30±0.62	3.01±0.65	1.81±0.29	2.10±0.40	2.31±0.96
SLANT	1.22±0.25	1.36±0.30	1.37±0.86	1.15±0.46	1.03±0.10	1.03±0.12	1.03±0.13	1.06±0.207	1.16±0.41
3D U-Net	1.00±0.08	1.00±.07	0.41±0.50	0.38±0.49	0.92±0.30	0.92±0.33	0.81±0.40	0.71±0.47	0.77±0.43
V-Net	1.03±0.12	1.06±0.16	1.03±0.20	1.00±0.09	1.08±0.21	1.09±0.22	1.04±0.15	1.03±0.13	1.05±0.17
nnUnet (2D)	1.24±2.30	1.25±2.12	1.00±3.78	0.94±3.68	1.28±3.49	1.20±3.00	1.25±4.33	1.29±4.71	1.18±3.54
nnUnet (3D)	0.79±0.63	1.11±2.73	0.26±1.63	0.29±2.41	0.39±2.11	0.90±4.91	0.40±2.94	1.22±6.82	0.67±3.56
LiviaNET	1.00±0.05	1.00±0.02	1.35±7.66	1.00±3.90	1.00±0.18	1.27±5.64	0.91±0.30	0.80±0.45	1.04±3.64
UNETR	1.00±0.10	1.00±0.07	0.45±0.50	0.39±0.49	0.93±0.29	0.96±0.24	0.77±0.45	0.82±0.40	0.79±0.42
CATS	0.90±0.30	0.91±0.29	0.20±0.40	0.19±0.40	0.62±0.51	0.65±0.51	0.39±0.50	0.37±0.50	0.53±0.51
Swin-UNETR	0.69±0.46	0.71±0.46	0.18±0.38	0.17±0.39	0.38±0.49	0.44±0.51	0.18±0.39	0.15±0.36	0.36±0.49
Swin-UNET	0.99±0.10	0.98±0.13	0.36±0.48	0.39±0.49	0.72±0.45	0.82±0.40	0.59±0.49	0.54±0.51	0.67±0.47
Proposed	**0.36±0.48***	**0.40±0.49***	**0.11±0.31***	**0.07±0.27***	**0.11±.31***	**0.17±0.38***	**0.06±0.31***	**0.05±0.22***	**0.17±0.37***

be viewed from the results of nnUnet [92]. In addition, the single-path network, LiviaNET [38], produces lower quality results with the comparison of 3D U-Net [35], which could be caused by lacking global information. The comparison of UNETR and CATS [64, 116] suggests that a traditional CNN block used as an encoder might be more effective than the Vision-transformer [43] for subcortical segmentation. Similarly, a comparison of between UNETR and Swin-UNETR [64, 63] indicates that the Swin-transformer [141] is

more effective. However, the pure transformer (Swin-UNET [20]) network design is not recommended for the subcortical segmentation task due to the undesired performance which can be viewed in Table 7.3.

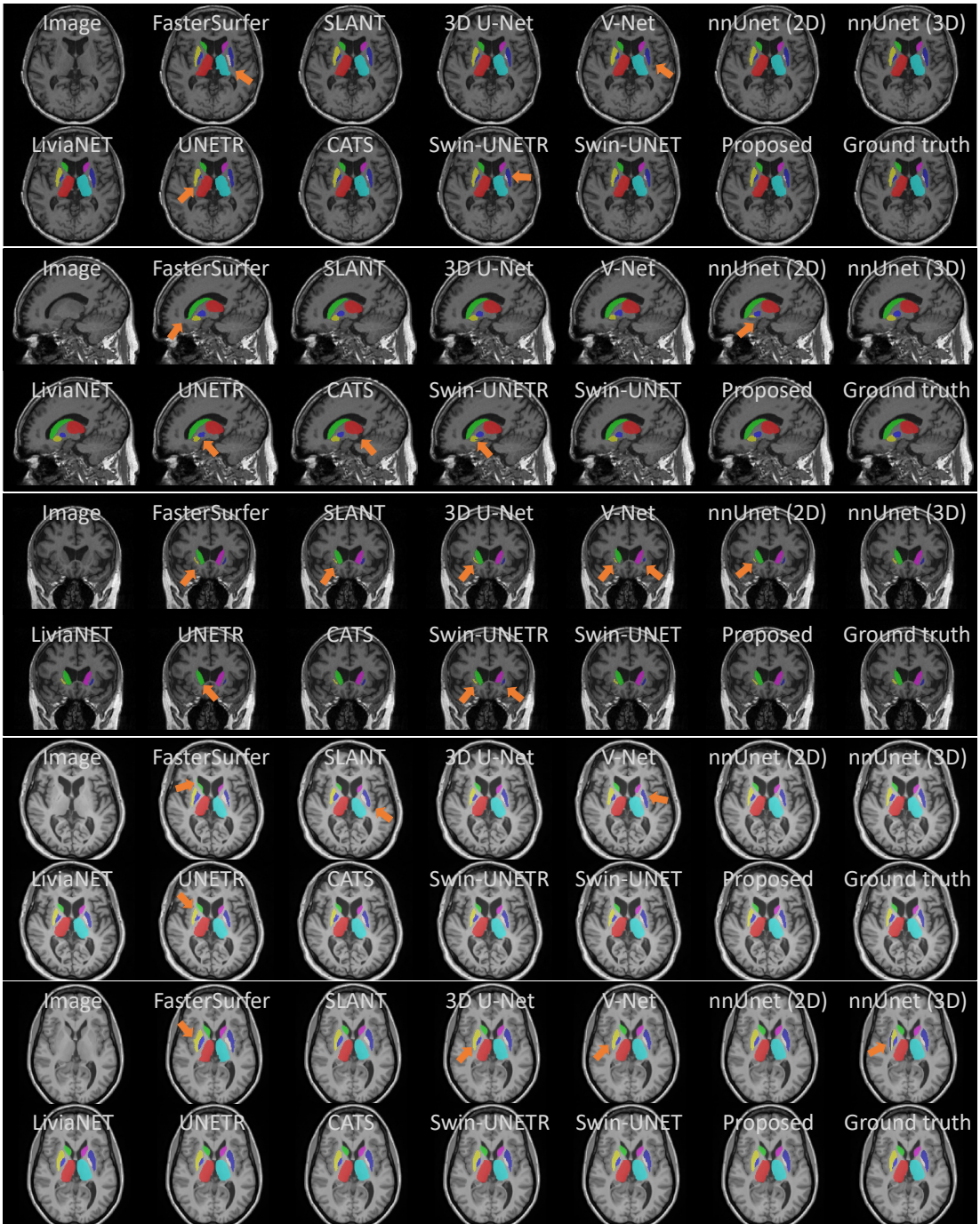


Figure 7.2: Qualitative results of all compared methods (Table 7.2). Arrows denote defects.

Qualitative results. Figure 7.2 shows the qualitative results of segmentations produced from all compared methods. Major segmentation defects are highlighted by orange arrows. From Figure 7.2, it is not hard to observe that all methods could produce plausible segmentations for the thalamus (red and light blue) with-

out major errors. However, local defects, such as under-segmentation and over-segmentation, are apparent in the results from FasterSurfer, Swin-UNETR, and CATS, especially in the axial and sagittal views (as shown in the first and second parts of Figure 7.2). Furthermore, UNETR, Swin-UNETR, and nnUnet (2D) under-segment the putamen (yellow) as seen in the sagittal view. In the coronal view, inaccurate segmentations of the putamen are noticeable in the outputs from SLANT, 3D U-Net, V-Net, nnUnet (2D), UNETR, and Swin-UNETR. Swin-UNETR also shows a subpar segmentation of the right caudate (pink) towards the bottom region. In both the first and fourth sections, UNETR and V-Net under-segment the left and right pallidum (blue and white). Incorrect segmentations between the left putamen and pallidum are evident from nnUnet (3D) in the last section of Figure 7.2, consistent with the large standard deviation displayed in Table 7.3. Although the Swin-UNET produces the plausible segmentations for all subcortical structures in Figure 7.2, it nevertheless achieves worse quantitative results than the proposed method. Mostly, our method provides more reasonable segmentations than all compared methods, which are the closest to the ground truths (see Figure 7.2). Notably, our proposed method did not deliver segmentation with local defects for caudate and putamen in any of these representative subjects. These observations are crucial for HD studies, given that these subcortical structures serve as biomarkers for HD progression.

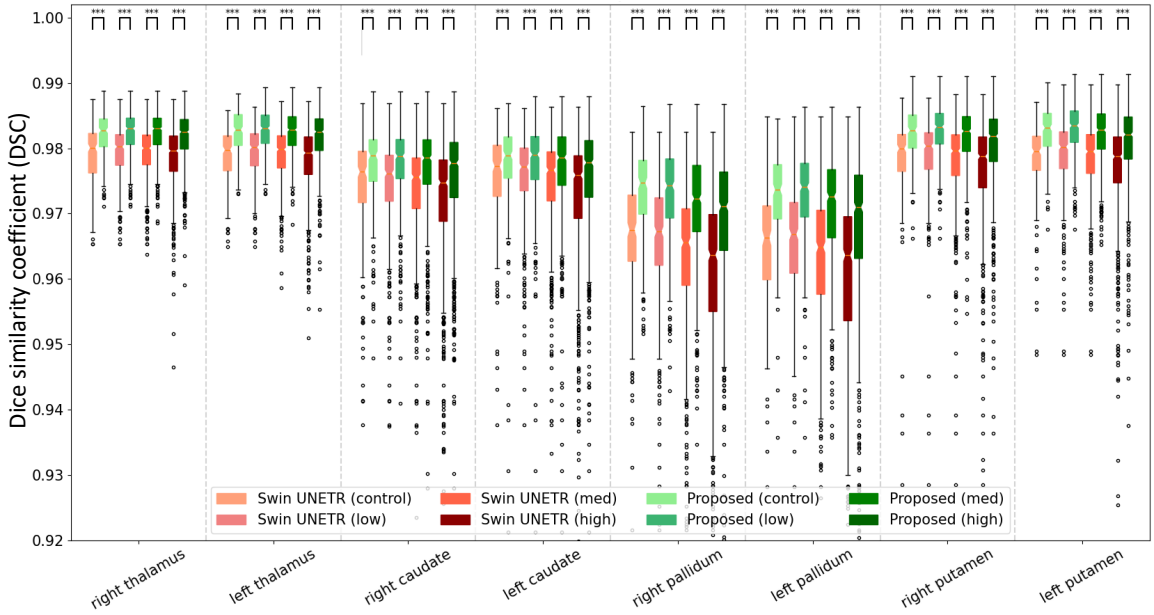


Figure 7.3: Segmentation performance for different disease groups. The significant differences are marked by *** ($p < 0.00001$). HD groups, i.e., low-, med- and high-CAP (Cytosine-Adenosine-Guanine Age Product) groups, are classified into colors from healthy (light color) to severe (dark color), where the higher CAP (CAG-Age Product) score indicates more severe disease status and likely more severe atrophy.

7.3.2 Detailed Performance Analysis

As introduced in Section 7.1, age and CAG are the main factors during the progression of HD, and the patient can be divided into three different groups based on their CAP scores (see Section 7.2.2). To better understand the capability of subcortical segmentation methods, we selected the methods with the best and second best overall performance in Table 7.3, i.e., the proposed method and Swin-UNETR [64], to further investigate the performance at the group level between healthy controls, low-, medium-, and high-CAP subjects.

Group-wise quantitative results. Figure 7.3 displays the group-wise subcortical segmentation performance in terms of DSC for selected methods, ranging four different groups from healthy controls (control) to severe HD subjects (high-CAP). For all subcortical structures and groups, the proposed method significantly outperformed Swin-UNETR (denoted as ***, $p < 0.00001$). Additionally, the fewer outliers in our results may suggest that our model is more generalizable to atrophied structures. Importantly, we further observe the DSC of the proposed method is more consistent along the disease progression (Low-, Medium-, High-CAP) than Swin-UNETR, which indicates that it is more robust to increasing amounts of atrophy. However, both methods have decreased performance for more severe atrophy cases, especially for caudate and putamen which are expected to atrophy most during disease progression.

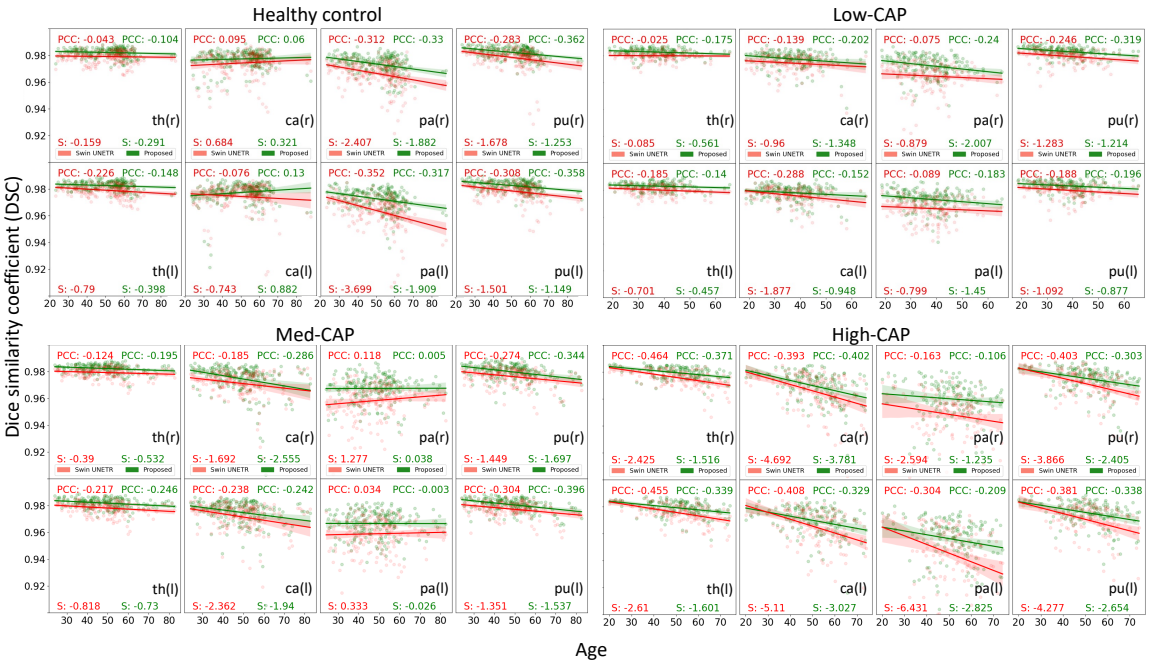


Figure 7.4: The group-wise visualizations of relationship between DSC and age (*years*). Two methods are shown: proposed method (green) and Swin-UNETR (red). The translucent bands around the regression line is 95% confidence interval that estimated using a bootstrap with 1000 as resample number. PCC and S indicate values of Pearson correlation coefficient and slope, respectively. The CAP (Cytosine-Adenosine-Guanine Age Product) groups and subcortical structures are marked as well.

Age-wise quantitative results. As previously mentioned, age is a significant factor influencing HD. Consequently, we further examine the DSC of various subcortical structures across different age groups, as illustrated in Figure 7.4. Similarly, visualizations based on the CAP scores are presented across four categories. In each plot, the regression line depicts the general trend or relationship between DSC and age, with its 95% confidence interval represented by translucent bands. Furthermore, both the Pearson correlation coefficient and the slope value are annotated to more clearly highlight the interaction. We observe that our proposed method consistently outperforms the comparison methods in most age groups with higher Dice scores, i.e., no intersections between the lines. In contrast, Swin-UNETR produced better segmentations for younger age groups ($\lessapprox 20$ *years*) in some categories, such as left caudate from healthy control and the plots under High-CAP except right pallidum. We note that even in healthy subjects, the subcortical structures are expected to be atrophied throughout aging. From Figure 7.4, the smaller slope values of the proposed method are found from 22/32 categories. These findings are similar to observations in Figure 7.3 that indicate our method is more generalizable to atrophied structures. Another important finding can be viewed between the Med- and High-CAP groups, with severe atrophied subcortical structures, our method has larger slope values, which suggests our method is robust to increasing amounts of atrophy.

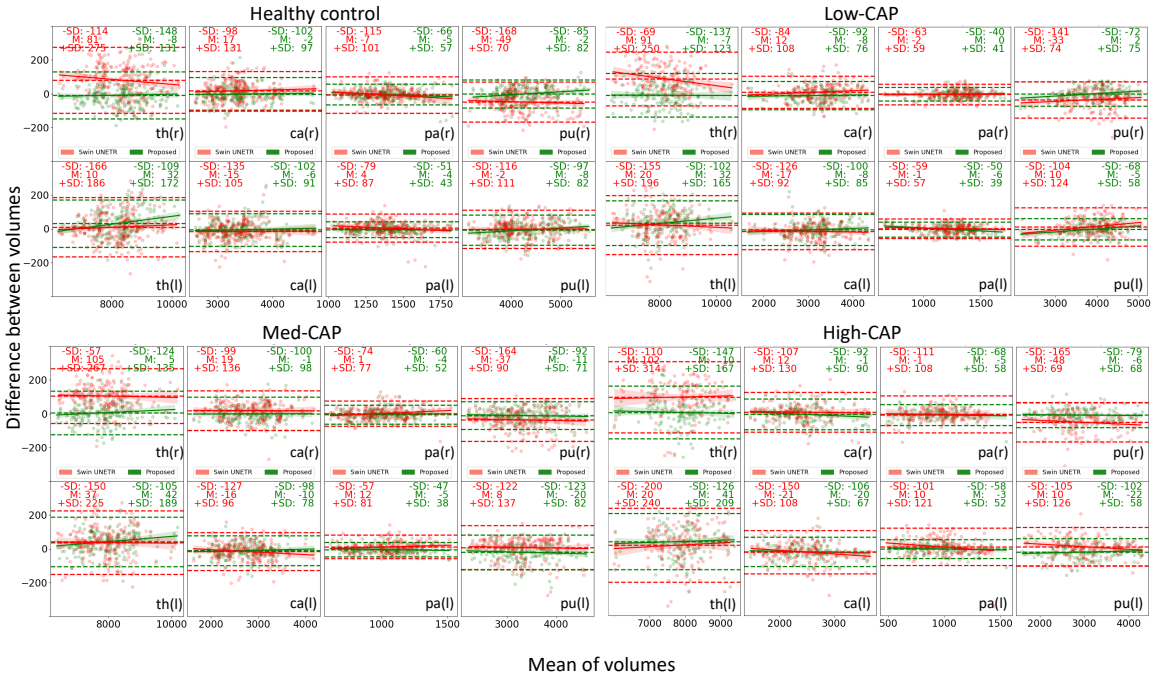


Figure 7.5: The volume Bland-Altman plots of the proposed method (green) and Swin-UNETR (red) in mm^3. M indicates mean, + and - SD represent +1.96 and -1.96 standard deviation to demonstrate the agreement between predictions and ground truths. The CAP (Cytosine-Adenosine-Guanine Age Product) groups and subcortical structures are marked. The translucent bands around the regression line is 95% confidence interval that estimated using a bootstrap with 1000 as resample number.

Volumetrics results. Figure 7.5 shows the Bland-Altman plots for different cohorts. Specifically, these plots highlight the presence of under- and over-segmentations by showing volume differences relative to the mean volume between predictions and ground truths. The plots also mark the means and standard deviation ranges. In Figure 7.5, the standard deviation range of the proposed method is narrower than that of Swin-UNETR in 20/32 different groups, and the mean values of the samples are closer to 0.

It's evident that both methods occasionally under- and over-segment larger volume structures. The right thalamus segmentations from our method are notably accurate (with minimal bias) across all groups. In contrast, the obvious bias can be viewed from the point cloud of Swin-UNETR, especially for the healthy control and low-CAP groups. Similarly, over-segmentation can be found in the left thalamus for our method.

For small volume structures, both methods achieved high performance by delivering superior segmentations. However, the unstable performance is found from Swin-UNETR as evidenced by the inconsistent volume differences corresponding to increasing mean volume, an example can be viewed as right pallidum from control cohort. It is noteworthy that the proposed method has a tighter range for pallidum segmentation. Considering that the pallidum does not significantly change during HD progression and inherently has a small volume even in control subjects, this suggests that our model adeptly segments smaller subcortical structures.

For the majority of subcortical structures in healthy control subjects, both methods offer satisfactory segmentations, with the exception of the right putamen. Here, biased results and under-segmentation emerge from our method and Swin-UNETR, respectively. Yet, with a smaller standard deviation and mean, our method tends to yield more accurate outcomes.

The severed atrophied subcortical structures are expected during the late stage of HD, and both methods can segment them in an accurate way, which can be viewed as the left-shifted distribution of mean volume from the Low- to the High-CAP group. Yet, Swin-UNETR has a broader standard deviation and a more dispersed point cloud, indicating that our method is more stable across different volume groups. For example, our model is able to segment them accurately which can be viewed as right putamen from control to CAP-high group that samples produced by our method have smaller standard deviation and fewer outliers. Although both methods can accurately segment the caudate, the segmentation of the left caudate by Swin-UNETR remains an exception.

7.3.3 Ablation Study

We further conducted ablation studies under various conditions to assess the effectiveness of the proposed method across the entire PREDICT-HD dataset.

Effectiveness of the proposed framework. The first two rows of Table 7.4 demonstrate the effectiveness of the proposed coarse-to-fine framework, where using a cropped image to produce segmentation improves

Table 7.4: Quantitative results (overall DSC) of proposed framework applied on different methods. Formatting is same as Table 7.3. Three types of networks are listed based on their overall performance in Table 7.3, and the details of methods can be found in Table 7.2. O and D represent original and downsampled image are used as inputs, respectively. P indicates the 3D patch-based training scheme on input image and cubic patch size is marked as subscript. C denotes the cropped image is used as input which is produced by coarse stage of proposed framework. Proposed (D) and (C) denote the results from coarse and fine stages. Bold indicates the best performance of each method after interacted with proposed framework.

	Dice similarity coefficient ↑								
	th(r)	th(l)	ca(r)	ca(l)	pa(r)	pa(l)	pu(r)	pu(l)	overall
Proposed (D)	.934±.006	.935±.005	.891±.018	.891±.016	.874±.019	.873±.021	.907±.013	.906±.013	.901±.027
Proposed (C)	**.982±.004**	**.982±.004**	**.975±.009**	**.975±.009**	**.969±.011**	**.968±.012**	**.980±.006**	**.981±.006**	**.977±.010**
CATS (P_{64} + O)	.963±.008	.964±.007	.949±.023	.955±.016	.936±.023	.937±.023	.945±.018	.955±.014	.950±.020
CATS (P_{64} + C)	**.970±.006**	**.971±.006**	**.963±.013**	**.966±.012**	**.950±.017**	**.949±.019**	**.965±.011**	**.967±.010**	**.962±.015**
CATS (P_{96} + O)	.974±.007	.975±.006	.969±.011	.970±.012	.955±.016	.954±.017	.972±.010	.972±.010	.968±.013
CATS (P_{96} + C)	**.975±.006**	.974±.006	**.971±.011**	**.971±.011**	.955±.016	.953±.019	**.973±.009**	**.973±.009**	.968±.014
Swin-UNETR (P_{64} + O)	.974±.005	.974±.005	.966±.013	.968±.012	.954±.017	.952±.016	.969±.009	.969±.009	.966±.014
Swin-UNETR (P_{64} + C)	**.977±.004**	**.977±.005**	**.970±.011**	**.971±.011**	**.956±.016**	**.956±.018**	**.974±.008**	**.975±.008**	**.970±.014**
Swin-UNETR (P_{96} + O)	.976±.005	.977±.005	.970±.011	.972±.010	.957±.015	.957±.018	.974±.008	.974±.009	.970±.013
Swin-UNETR (P_{96} + C)	**.979±.005**	**.978±.005**	**.972±.010**	.972±.010	**.960±.014**	**.959±.016**	**.977±.005**	**.977±.008**	**.972±.013**

Table 7.5: Quantitative results in terms of DSC of different network architectures on PREDICT-HD dataset. Formatting is same as Table 7.3. Baseline is the variant of 3D U-Net, a deeper 3D U-Net shown in Figure 6.2. R, A and D represent residual block, attention gate and deep supervision are used as modifications of baseline, respectively. Note, baseline (R + A + D) is the proposed segmentation network.

	Dice similarity coefficient ↑								
	th(r)	th(l)	ca(r)	ca(l)	pa(r)	pa(l)	pu(r)	pu(l)	overall
3D U-Net	.972±.005	.971±.006	.965±.011	.965±.011	.943±.019	.943±.021	.964±.011	.966±.010	.961±.017
baseline (deeper 3D U-Net)	.973±.005	.972±.005	.966±.013	.967±.011	.940±.025	.948±.020	.970±.009	.970±.010	.963±.018
baseline (R)	.973±.005	.967±.009	.968±.012	.965±.013	.953±.017	.959±.014	.971±.009	.972±.009	.966±.013
baseline (R + A)	.980±.005	.980±.005	.973±.010	.974±.010	.965±.014	.962±.018	.979±.007	.979±.008	.974±.012
baseline (R + A + D)	**.982±.004**	**.982±.004**	**.975±.009**	**.975±.009**	**.969±.011**	**.968±.012**	**.980±.006**	**.981±.006**	**.977±.010**

performance compared to using a downsampled image, this can be viewed in overall performance. Moreover, The largest improvements are for the left pallidum, from 0.873 to 0.968.

Additionally, Table 7.4 presents the performance of the proposed coarse-to-fine framework when applied to various methods. We replaced the CNN in the proposed framework with Swin-UNETR and CATS. Specifically, Swin-UNETR [63] uses a transformer as its encoder and CATS [116] incorporates hybrid encoders. It is evident from Table 7.4 that the proposed coarse-to-fine framework can positively contribute to all methods in terms of overall performance, except for CAT with 96 as input patch size. However, the proposed framework improves the accuracy of caudate and putamen segmentations.

Different patch sizes are investigated for the proposed framework, which is shown in Table 7.4. Noteworthy larger improvements are observed in methods with smaller patch sizes for Swin-UNETR and CATS. This indicates a simple and useful sampling strategy with limited computation resources.

Effectiveness of network modifications. The impact of the modifications on the baseline method

(deeper 3D-UNet) for subcortical segmentation is examined and presented in Table 7.5.

First, we extended 3D U-Net as the baseline (see Figure 6.2) to better capture more comprehensive input data through deeper representations. Compared to the DSC performance of the original 3D U-Net, this modification produces better segmentations except for the right pallidum.

We obtain superior pallidum segmentations with the residual blocks (R), especially for the right pallidum, which improved from 0.940 to 0.953. However, the left thalamus and caudate show decreased performance for this configuration.

The most pronounced improvements were achieved by integrating attention gates with the skip connections (R+A). As seen in Table 7.5, the Dice scores for all subcortical structures increased, alongside the overall performance. Furthermore, the consistency of the same structure for both left and right sides is also improved. For instance, the difference between the left and right sides of the pallidum is reduced from 0.008 to 0.003.

Lastly, deep supervision (R+A+D) in the decoder (see Figure6.2) is used to form the final architecture of the proposed network, which contributes positively to subcortical structure segmentations. This indicates that these hierarchical layers can retain crucial deeper-level information, facilitating robust segmentation. The most significant improvement is observed in the left pallidum, and the difference between the left and right sides narrows further.

With these modifications, our network elevates the overall performance from 0.963 to 0.977 and also achieves a reduced standard deviation.

Effectiveness of augmentation strategy. Although data augmentations are widely used, their contributions have not been discussed for subcortical segmentation in the context of neurodegeneration. Table 7.6 shows the effects of various augmentations described in Section 6.3 on the proposed method with respect to volumetric (DSC) and surface (ASD and HD95) accuracy. The formatting is consistent with Table 7.3.

As a general observation, integrating the proposed augmentation strategy, which combines affine and elastic transformations, improves both overall Dice and surface metrics.

When used alone, the affine transformation contributes more to the results than elastic deformation. Nevertheless, the performance for every metric of every subcortical structure is boosted when either of the augmentation strategies is employed, compared to no augmentation.

It is worth noting that elastic deformation has a pronounced effect on structures with smaller volumes, such as the ASD of the right pallidum.

Furthermore, for bio-markers of HD, the affine transformation proves to be more effective than elastic deformation, as seen in metrics like a higher DSC and lower ASD (except for the right putamen) and smaller HD95 for caudate and putamen. For such subcortical structures with larger volumes, the best performance

might be found by applying affine alone.

Nevertheless, the combined augmentation strategy is suggested due to its overall performance. Moreover, consistency in DSC and ASD between the left and right sides for all subcortical segmentations is desirable. For HD95, larger differences still exist, which can be observed in the caudate and pallidum.

Table 7.6: Quantitative results of different augmentations on PREDICT-HD dataset. w/o denotes no augmentation is applied. A and E indicate only affine and elastic augmentation (see Section 6.3) is applied, respectively. Comb. denotes combination of A and E.

Dice similarity coefficient ↑

	th(r)	th(l)	ca(r)	ca(l)	pa(r)	pa(l)	pu(r)	pu(l)	overall
Proposed (w/o)	.980±.004	.980±.005	.973±.010	.974±.010	.965±.013	.965±.014	.979±.007	.979±.007	.974±.011
Proposed (A)	**.982±.004**	**.982±.004**	**.976±.009**	**.976±.010**	.968±.010	.967±.015	**.981±.006**	.981±.007	.976±.010
Proposed (E)	.981±.004	.981±.004	.974±.011	.973±.011	.967±.011	.966±.013	.979±.007	.979±.007	.975±.011
Proposed (Comb.)	**.982±.004**	**.982±.004**	.975±.009	.975±.009	**.969±.011**	**.968±.012**	.980±.006	**.981±.006**	**.977±.010**

Averaged surface distance (*mm*) ↓

	th(r)	th(l)	ca(r)	ca(l)	pa(r)	pa(l)	pu(r)	pu(l)	overall
Proposed (w/o)	.059±.038	.058±.044	.039±.036	.037±.030	.043±.042	.047±.050	.040±.034	.048±.057	.046±.043
Proposed (A)	.048±.042	.047±.021	**.027±.013**	**.027±.018**	.035±.034	.037±.039	.034±.040	.028±.022	.035±.032
Proposed (E)	.047±.017	.048±.022	.032±.022	.033±.027	.036±.034	**.033±.027**	.032±.020	.039±.030	.038±.026
Proposed (Comb.)	**.046±.026**	**.044±.018**	.027±.014	.028±.018	**.032±.018**	.035±.023	**.028±.014**	**.028±.017**	**.033±.020**

95-th percentage Hausdorff distance (*mm*) ↓

	th(r)	th(l)	ca(r)	ca(l)	pa(r)	pa(l)	pu(r)	pu(l)	overall
Proposed (w/o)	.569±.500	.542±.529	.131±.338	.130±.340	.195±.412	.261±.462	.095±.299	.113±.320	.254±.447
Proposed (A)	.375±.484	**.372±.484**	**.090±.287**	.102±.309	.115±.320	.172±.400	**.057±.232**	.057±.232	.168±.377
Proposed (E)	.409±.492	.421±.494	.128±.335	.123±.343	.148±.359	.192±.415	.074±.262	.083±.284	.197±.403
Proposed (Comb.)	**.363±.481**	.397±.490	.106±.308	**.074±.267**	**.107±.310**	**.167±.375**	.061±.310	**.053±.224**	**.166±.373**

7.4 Discussion and Conclusion

In this work, we comprehensively evaluated the proposed deep learning framework with cascaded networks for subcortical segmentation. The proposed method is designed for practical clinical use among large HD population with robust segmentation. It also tackles common issues encountered in practice, such as label availability and computational resource (GPU memory) limitations for 3D image processing, leading to increased efficiency and accuracy. Unlike typical deep learning methods trained models with downsampled images to avoid GPU memory issues and later resample the results to their original resolution, which would lead to degraded performance, we instead built a 2-stage framework that consists of coarse and fine stages. The proposed framework involves a coarse stage to identify the regions of target structures and a subsequent fine stage that leverages the information from the coarse stage to generate precise segmentations. Our findings indicate that this framework not only improves training schemes based on full images but also benefits patch-based approaches by producing higher-quality segmentation. Compared to other SOTA methods, com-

prehensive evaluations reveal that our method consistently outperforms, delivering superior segmentations in terms of accuracy metrics.

This study was conducted on a large-scale PREDICT-HD dataset, which consists of various cohorts, including healthy subjects and HD patients, as detailed in Sec. 7.2.2. It demonstrated the capability of CNN for subcortical segmentation by comparing transformers, and the proposed method produced the best quantitative and qualitative results. The PREDICT-HD dataset contains a large number of HD patients, and it is important to examine the generalizability of the proposed methods. Our methods produce high-quality segmentations for neurodegeneration cohorts (different HD groups), and the results indicate our method is more robust to increasing amounts of atrophy. The superior performance is also validated in the aspects of age and volume.

The ablation study is performed to investigate the effectiveness of the proposed framework. The proposed coarse-to-fine framework benefits other segmentation methods. We further investigated the impact of different patch sizes. This addresses a frequent challenge in 3D medical image segmentation: oversized inputs that exceed GPU memory limits.

We showed improved performance by modifying network architectures and applying data augmentations. Such straightforward modifications progressively improve the overall performance, and data augmentations are useful for improving the performance in our experiments.

However, there are several limitations of the proposed work: (1) the consistency of HD95 between left and right structures required to be improved, such as caudate and pallidum, (2) The generalizability of the proposed network needs to be further evaluated on other neurodegeneration diseases, (3) the performance of proposed framework on large volume structure need to be tested. In future work, we plan to focus on the size of the training samples, aiming to achieve comparable performance with fewer training instances. To meet this objective, self-training techniques may be employed to boost generalizability. Furthermore, pseudo-label training could serve as an effective strategy for generating robust segmentations with limited training data. Moreover, we will assess the efficacy of the proposed method using clinical measurements from HD patients.

In conclusion, we proposed a straightforward yet efficient framework with cascaded networks for efficient and accurate subcortical structure segmentation in neurodegeneration. Our proposed framework is evaluated on the large-scale PREDICT-HD dataset with comprehensive experimental designs and comparisons. The superior accuracy of our framework in accuracy metrics suggests its potential for clinical application.

CHAPTER 8

Longitudinal Subcortical Segmentation with Deep Learning

Longitudinal information is important for monitoring the progression of neurodegenerative diseases, such as Huntington's disease (HD). Specifically, longitudinal magnetic resonance imaging (MRI) studies may allow the discovery of subtle intra-subject changes over time that may otherwise go undetected because of inter-subject variability. For HD patients, the primary imaging-based marker of disease progression is the atrophy of subcortical structures, mainly the caudate and putamen. To better understand the course of subcortical atrophy in HD and its correlation with clinical outcome measures, highly accurate segmentation is important. In recent years, subcortical segmentation methods have moved towards deep learning, given the state-of-the-art accuracy and computational efficiency provided by these models. However, these methods are not designed for longitudinal analysis, but rather treat each time point as an independent sample, discarding the longitudinal structure of the data. In this paper, we propose a deep learning based subcortical segmentation method that takes into account this longitudinal information. Our method takes a longitudinal pair of 3D MRIs as input, and jointly computes the corresponding segmentations. We use bi-directional convolutional long short-term memory (C-LSTM) blocks in our model to leverage the longitudinal information between scans. We test our method on the PREDICT-HD dataset and use the Dice coefficient, average surface distance and 95-percent Hausdorff distance as our evaluation metrics. Compared to cross-sectional segmentation, we improve the overall accuracy of segmentation, and our method has more consistent performance across time points. Furthermore, our method identifies a stronger correlation between subcortical volume loss and decline in the total motor score, an important clinical outcome measure for HD.

8.1 Introduction

Huntington's disease (HD) is an autosomal dominant neurodegenerative disorder known to affect subcortical structures [8]. Specifically, changes in caudate and putamen volume are the primary imaging-based markers of HD pathology even in the early stages of disease progression [144, 143, 171], whereas other structures such as the pallidum appear to be more mildly affected, and the thalamus is relatively preserved. Understanding the relationship between subcortical atrophy and clinical outcome measures such as the total motor score (TMS) is of interest for HD studies. Given the large inter-subject variability in brain anatomy and disease progression, longitudinal MRI studies are especially important in this context. While many large-scale studies

This work is published at SPIE 2021.
Li, Hao, et al. "Longitudinal subcortical segmentation with deep learning." Medical Imaging 2021: Image Processing. Vol. 11596. SPIE, 2021.

indeed collect longitudinal MRI data, many segmentation methods treat each MRI scan as an independent sample even if they belong to the same subject, and only consider the longitudinal structure of the data in post-processing, i.e., in the statistical analysis stage. Such an approach limits the benefits of longitudinal datasets.

In recent years, deep learning based segmentation methods have dominated the field, and state-of-the-art deep learning based methods produced promising results for segmenting subcortical structures [38, 124, 225, 126]. Dolz et al [38]. proposed a 3D fully convolutional neural network (FCNN) on subcortical segmentation, and they applied their method to a large-scale dataset to prove the model robustness and segmentation accuracy. Based on the work of Dolz et al. [38], Li et al [124]. further improved segmentation results by exploring variants on augmentation strategies and network architecture. Wu et al. [225] proposed a 2D + 3D framework for segmenting the subcortical structures, which demonstrated promising performance. In a concurrent submission, we propose [126] a cascaded 3D framework and a 3D FCNN to segment subcortical structures, which provides more accurate segmentations than Dolz et al. [38], Li et al. [124] and Wu et al. [225]. However, none of these methods are designed for longitudinal subcortical segmentation, and the development of such methods is an important need in the field.

In related previous studies, the convolutional long short-term memory (C-LSTM) [228] approach has been used for joint segmentation of multiple images and produced promising results [51, 70, 5] by extracting and passing useful inter-image information. He et al. [70] used C-LSTM to leverage inter-slice information and improved segmentation consistency in retinal OCT scans. Bai et al. [5] achieved aortic image sequence segmentation by applying C-LSTM. In a closely related work, Gao et al. [51] stacked C-LSTMs into an FCNN for joint 4D medical image segmentation. This allows the model to learn the overall trend and the correlations from MRIs at multiple time-points.

In this paper, we propose a longitudinal subcortical segmentation method, which uses two 3D MRIs acquired at different time points from a given subject as input and jointly computes the corresponding segmentations. Inspired by the work from Gao et al. [51], we use two different time-point MRIs for each subject and focus on the relationship between these inputs to improve the segmentation accuracy. Additionally, we use a network architecture specifically optimized for the subcortical segmentation task [126]. With the bi-directional C-LSTM blocks, our model is able to extract, pass and fuse useful longitudinal information to form segmentations. Thus, we leverage the inherent dependency between longitudinal scans instead of treating them as independent samples and discarding the useful contextual information. We test our method on the PREDICT-HD dataset and evaluate our results against a cross-sectional variant [126]. The Dice coefficient, average surface distance and 95-percent Hausdorff distance are used as evaluation metrics. We also report the correlation of the volume loss with the decline in total motor score.

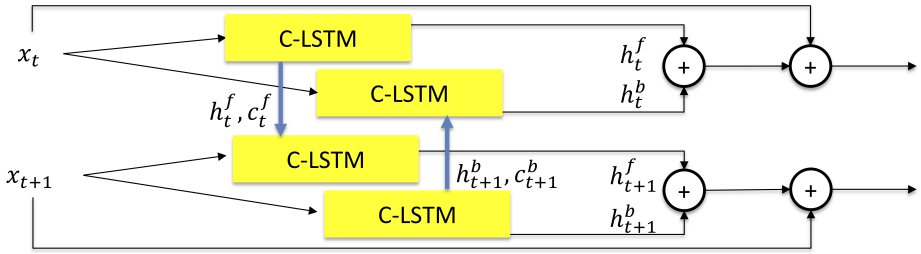


Figure 8.1: The bi-directional convolutional long short-term memory (C-LSTM) block. h^f, c^f and h^b, c^b are the output cell and hidden state in forward and backward path. t denotes the time-point.

8.2 Methods

8.2.1 Convolutional LSTM

The long short-term memory (LSTM) [76] is a specific type of recurrent neural network (RNN) for increasing learning ability based on previous information. Furthermore, the LSTM minimizes the effect of the "gradient vanishing" problem in the training process. Based on fully connected LSTM [58], Shi et al. proposed the C-LSTM [228] which can be incorporated in the fully convolutional neural network (FCNN) architecture and leverages spatiotemporal information due to the convolution operation. The C-LSTM can be described as:

$$
\begin{aligned}
i_t &= \sigma(W_{xi} * x_t + W_{hi} * h_{t-1} + b_i) \\
f_t &= \sigma(W_{xf} * x_t + W_{hf} * h_{t-1} + b_f) \\
o_t &= \sigma(W_{xo} * x_t + W_{ho} * h_{t-1} + b_o) \\
c_t &= f_t \otimes c_{t-1} + i_t \otimes \tanh(W_{xc} * x_t + W_{hc} * h_{t-1} + b_c) \\
h_t &= o_t \otimes \tanh(c_t)
\end{aligned}
\tag{8.1}
$$

where x_t is the input, c_t is the output cell state and h_t is the hidden state. c_{t-1} and h_{t-1} are the output from previous hidden layer. i_t, f_t and o_t are the input, forget and output gates respectively. $*$ denotes convolution operation, $\otimes$ denotes pixel-wise multiplication, and σ is the sigmoid function. Finally, the Ws contain the weights and the bs contain the bias.

For our work, to allow each input scan to leverage information from the other scan, we added a backward path to form a bi-directional C-LSTM, as shown in Fig. 8.1. Additionally, to reinforce current information, we used an addition before outputting. Along the training processes, C-LSTMs will extract, pass and fuse the useful longitudinal information between scans. Thus, our segmentations take into account the longitudinal context.

8.2.2 Network Architecture

Fig. 8.2 shows our framework for longitudinal subcortical segmentation. The framework contains 2 paths, such that each path receives the 3D image from one time-point as input and outputs the corresponding 3D segmentation. The output has the same size as the input volume but contains 9 channels (8 considered subcortical structures and 1 background). Similar to Gao et al. [51], in the encoder phase, the bi-directional C-LSTM blocks serve as the bridge between paths to achieve the extraction, transmission and fusion of longitudinal information. Thus, the feature maps containing both inter-scan and intra-scan information are forwarded to each decoder path separately to form segmentations. The encoder and decoder include 4 3D max-pooling and 3D nearest neighbor upsampling operations. Besides that, 8 residual blocks are stacked in encoder and decoder evenly, and 1 is used in bottleneck path. Residual blocks are modified from He et al. [68], and consist of a $3 \times 3 \times 3$ convolution operation, batch normalization and ReLU activation. Like the 3D U-Net [35], there is a skip connection with an attention gate [184] between encoder and decoder. An addition operation is applied in the decoders before the final output. The detailed network architecture is shown as Fig. 8.3, which is same as the architecture of cross-sectional subcortical segmentation method [126] (Chapter 6 and 7).

8.2.3 Experimental Setup

Dataset. Our dataset is a subset of the multi-site PREDICT-HD database [143] (Chapter 7) and contains 40 healthy control subjects and 119 HD subjects. Each subject has 2 T1-weighted MRIs with at least 2 years between the two scans. All HD subjects who meet the timeframe requirements have been selected. The HD subjects were classified into three groups based on their CAP scores, which is a marker of HD progression [248]: high-CAP (n=39), medium-CAP (n=40) and low-CAP (n=40). We randomly select 20 subjects in each category for training and 5 subjects for validation. The rest of the subjects were used for

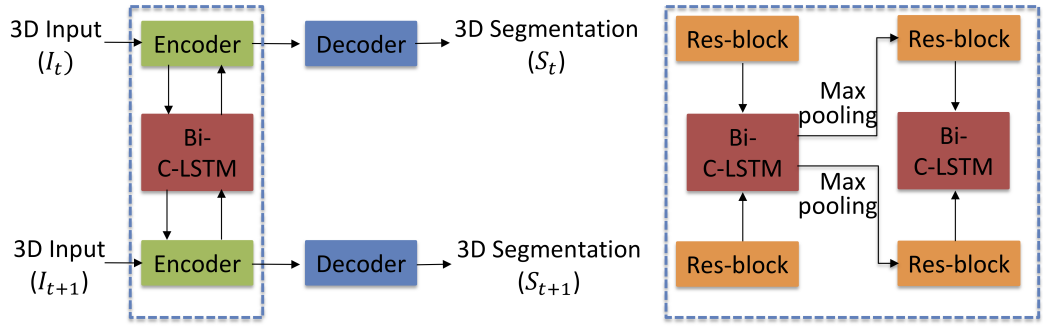


Figure 8.2: The longitudinal subcortical segmentation framework. The dashed square shows the connection between blocks. *Is* are the input 3D MRIs and *Ss* are the corresponding segmentations. t represents the time-point.

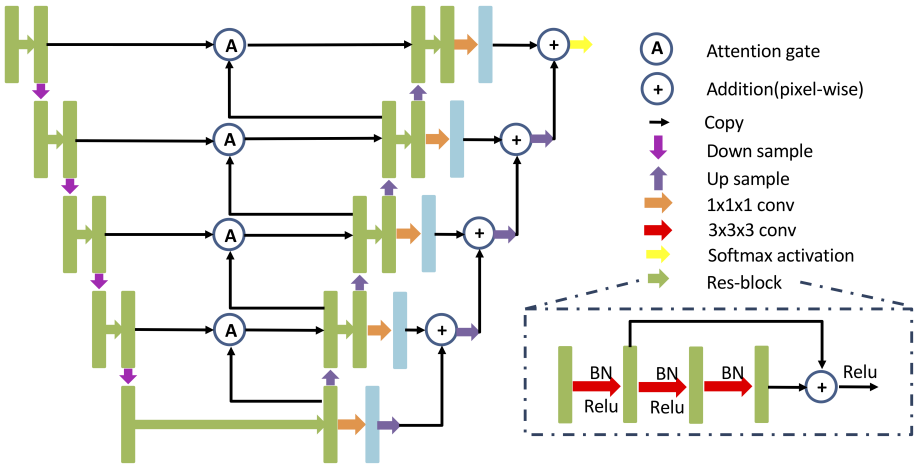


Figure 8.3: The network architecture. The feature maps are represented by rectangular boxes. The green boxes at each level contain 32, 64, 128, 256, 512 channels respectively. For all levels, light blue boxes have 9 channels (8 subcortical structures and background).

testing. The subcortical structures that are considered in our study are the left and right pairs of thalamus, caudate, pallidum and putamen. The preprocessing steps and data augmentation strategies are kept same as Chapter 7.

Implementation Details. During training, we used the Dice Loss[154] as our loss function, with weight=1 for all foreground labels and weight=0.1 for background. Other details are same as Chapter 6. The weights of bi-directional C-LSTM blocks and C-LSTM cells are not shared. With a batch size of 2, the whole training process took around 450 epochs with early stop. However, we set the maximum epoch number to 1000, in case the early stop condition is not triggered. We randomly switched the input order in each iteration to increase the model robustness and boost the ability of the bi-directional C-LSTM for extracting useful information. Due to GPU memory limitation, we applied the method described in[126] to automatically crop input images to a region of $96 \times 96 \times 48 \text{mm}^3$ which includes the subcortical area. The model is implemented using PyTorch and trained with a NVIDIA Titan RTX.

Evaluation methods. We compare the results of our longitudinal subcortical segmentation method to its cross-sectional counterpart [126]. We note that this cross-sectional method [126] outperforms previous state-of-the-art methods such as those presented in [38, 124, 225]. All preprocessing and data augmentation pipelines as well as train/validation/test splits of the data were the same between the cross-sectional and longitudinal methods, with the only difference being the introduction of a second path in the longitudinal network as well as the bi-directional C-LSTM blocks connecting the two paths. For evaluation, we use the Dice coefficient, the average surface distance and the 95-percent Hausdorff distance as our metrics. Statistical significance was determined using a 2-tail, paired t-test with significance threshold $p < 0.05$. We also report

Table 8.1: **(Top)** Segmentation Dice scores. Statistically significant improvements (2-tailed paired t-test, $p < 0.05$) over the cross-sectional method [126] are denoted in **bold**. **(Bottom)** The performance consistency, computed as the absolute difference of Dice score between two time-points of a subject. Underlined entries highlight better performance consistency. For both tables, results are presented as $mean \pm std.\ dev.$

Dice score

	R thalamus	L thalamus	R caudate	L caudate	R pallidum	L pallidum	R putamen	L putamen
Cross-sec.[126]	0.979±0.005	0.980±0.004	0.975±0.008	0.974±0.007	0.967±0.011	0.965±0.014	0.980±0.005	0.980±0.006
Longitudinal	**0.980±0.004**	**0.980±0.004**	**0.976±0.007**	**0.975±0.007**	**0.969±0.009**	**0.968±0.011**	**0.981±0.004**	**0.981±0.006**

Absolute difference of Dice score between time points ($\times$ 0.1)

	R thalamus	L thalamus	R caudate	L caudate	R pallidum	L pallidum	R putamen	L putamen
Cross-sec.[126]	0.021±0.020	0.029±0.024	0.040±0.036	0.048±0.046	0.067±0.052	0.069±0.059	0.033±0.026	0.036±0.029
Longitudinal	0.023±0.016	0.028±0.025	0.036±0.040	0.035±0.034	0.061±0.050	0.068±0.063	0.036±0.027	0.034±0.027

the Pearson's correlation coefficient between the change in total motor score (TMS) between the two visits and the volume loss between these two time-points.

8.3 Results

The Dice results for all subjects (control and HD) are shown in Tab. 8.1. The top panel shows that the Dice score of our longitudinal approach was significantly higher than the cross-sectional analysis for all 8 structures. To assess the consistency of the performance, we report the absolute value of the difference between the Dice scores of the two time-points, averaged over all subjects. These results are shown in the bottom panel of Tab. 8.1, where we observe that the longitudinal approach was more consistent for 6 out of 8 structures.

It is well known that the Dice score cannot capture small features that do not contribute substantially to overall volume, such as the thin tail of the caudate. For this reason, we also report the surface distances in Tab. 8.2, specifically, the average surface distance and the 95-percent Hausdorff distance. Both metrics are improved in the longitudinal segmentation for all 8 structures, with the difference reaching statistical significance for 10 out of the 16 comparisons.

Fig. 8.4 shows the breakdown of Dice scores across the 3 disease stages (low-, med-, high-CAP) as well as controls. We note that the longitudinal accuracy is consistently superior to cross-sectional segmentation, and it is more robust to increased amounts of atrophy known to be present in later disease stages.

The qualitative results are shown in Fig. 8.5. From axial and sagittal slices, we can see the clear improvements of pallidum (blue zoomed-in panels, orange arrows). The yellow and pink panels highlight improvements in the putamen segmentation accuracy (red arrows), where the most noticeable changes visible in sagittal slices. Furthermore, axial slices show our method delivered more plausible thalamus segmentations, since the boundary between the left and right thalamus should follow the mid-sagittal plane.

Table 8.2: **(Top)** Average surface distance. **(Bottom)** 95-percent Hausdorff distance. For both tables, results are presented as $mean \pm std.\ dev.$ Statistically significant improvements (2-tailed paired t-test, $p < 0.05$) over the cross-sectional method [126] are denoted in **bold**.

Average surface distance

	R thalamus	L thalamus	R caudate	L caudate	R pallidum	L pallidum	R putamen	L putamen
Cross-sec.[126]	0.051±0.016	0.051±0.018	0.030±0.016	0.030±0.012	0.035±0.014	0.040±0.030	0.032±0.035	0.029±0.011
Longitudinal	**0.049±0.017**	**0.048±0.017**	0.029±0.014	**0.028±0.019**	**0.031±0.012**	**0.035±0.019**	0.028±0.015	0.028±0.014

95-percent Haussdorf distance

	R thalamus	L thalamus	R caudate	L caudate	R pallidum	L pallidum	R putamen	L putamen
Cross-sec.[126]	0.585±0.495	0.568±0.497	0.093±0.292	0.119±0.325	0.161±0.369	0.253±0.448	0.051±0.221	0.076±0.267
Longitudinal	**0.508±0.502**	**0.492±0.502**	**0.051±0.221**	0.093±0.292	0.102±0.304	**0.169±0.377**	**0.008±0.092**	0.059±0.237

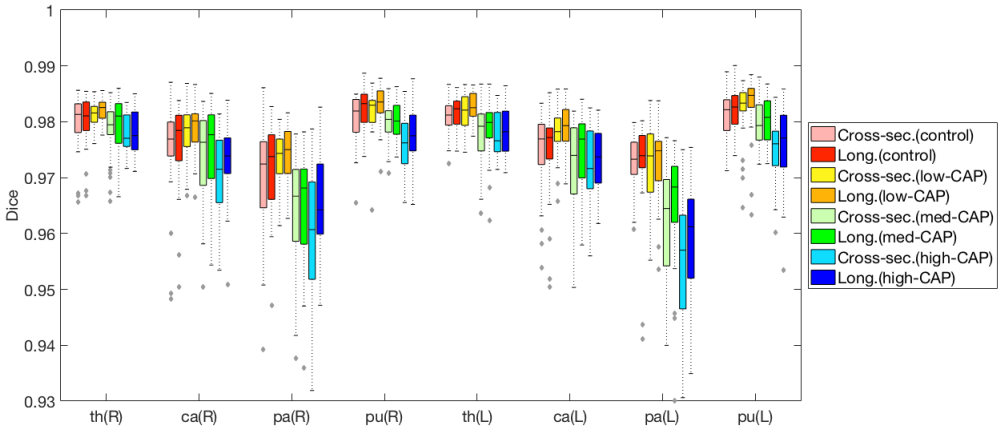


Figure 8.4: Comparison of segmentation performance. In HD subjects, many subcortical structures are increasingly atrophied in higher CAP groups. (L)eft and (R)ight pairs of (th)lamus, (ca)daute, (pa)llidum and (pu)tamen are considered. Compared to the cross-sectional method [126], our proposed longitudinal method obtained consistently superior Dice scores.

In Tab. 8.3, we report the Pearson's correlation coefficient between the change in TMS between the two visits and the volume loss between these two time points. The volumes were normalized by the total brain volume prior to this analysis. We expect to see strong correlations for the caudate and putamen which are known to be affected in HD, and an increase in correlation strength in later disease stages. We note that the findings in Tab. 8.3 are preliminary, since the number of test subjects in each group was limited, and several subjects had to be excluded due to missing TMS scores. Nevertheless, we note that the longitudinal segmentation method produced stronger correlations between volume loss and TMS decline for most of the comparisons in the caudate and putamen, the most affected structures. Interestingly, we also find robust correlations in the pallidum, which suggests that our methods might be more sensitive to HD changes in the pallidum than previously reported HD studies [144, 143, 171] that relied on an older generation of segmentation algorithms such as multi-atlas methods. In line with the literature [144, 143, 171], the thalamus does not appear to be

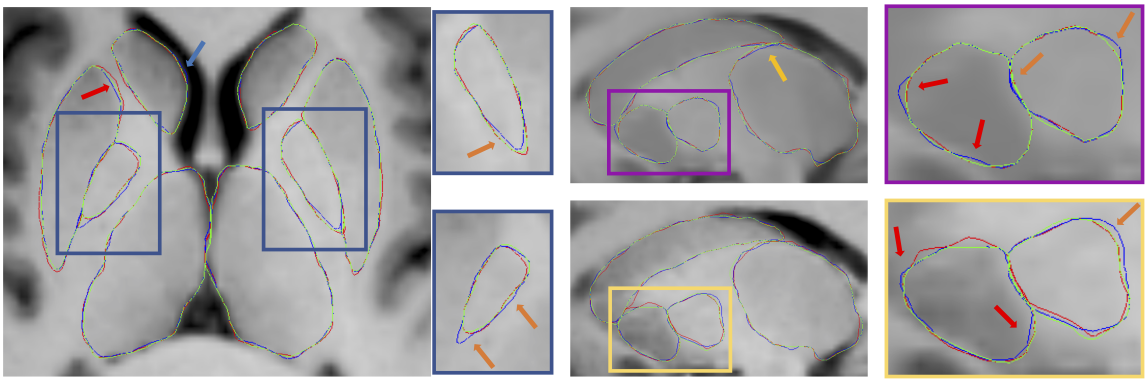


Figure 8.5: Qualitative results. Red, blue and green lines represent "ground truth", cross-sectional [126] and proposed longitudinal segmentations respectively. Yellow, blue, orange and red arrows highlight the improvements of thalamus, caudate, pallidum and putamen, respectively.

strongly associated with the HD pathology, although the correlation in the high-CAP group suggests that this structure may become affected in later disease stages. Replicating these preliminary findings in a larger dataset remains as future work.

Table 8.3: Pearson's correlation coefficient between volume loss between two time-points and the TMS decline during the same time period. **Bold** numbers denote the method identifying a stronger correlation. We observe that some structures like the thalamus show weak correlations with both methods, in line with the existing HD literature. Note that subjects with missing TMS data were excluded from this analysis (n=3, 2, 2 for high-CAP, medium-CAP and low-CAP categories respectively).

All HD Subjects (low-CAP, med-CAP, high-CAP)

	R thalamus	L thalamus	R caudate	L caudate	R pallidum	L pallidum	R putamen	L putamen
Cross-sec.[126]	**-0.1985**	**-0.0232**	-0.3800	-0.2826	-0.3233	-0.4280	**-0.3750**	-0.3431
Longitudinal	-0.1345	0.0743	**-0.3809**	**-0.3654**	**-0.3712**	**-0.4587**	-0.3152	**-0.3482**

Low-CAP Subjects

	R thalamus	L thalamus	R caudate	L caudate	R pallidum	L pallidum	R putamen	L putamen
Cross-sec.[126]	-0.0639	**-0.2707**	-0.2342	-0.1888	**-0.4166**	-0.2540	**-0.2987**	-0.4645
Longitudinal	**-0.0840**	-0.2680	**-0.2764**	**-0.2871**	-0.4079	**-0.3044**	-0.2922	**-0.4730**

Med-CAP Subjects

	R thalamus	L thalamus	R caudate	L caudate	R pallidum	L pallidum	R putamen	L putamen
Cross-sec.[126]	**-0.1258**	**0.3095**	-0.3268	-0.2676	-0.3877	-0.4911	**-0.3082**	-0.3169
Longitudinal	-0.1034	0.4050	**-0.3270**	**-0.3185**	**-0.4349**	**-0.5535**	-0.2327	**-0.4000**

High-CAP Subjects

	R thalamus	L thalamus	R caudate	L caudate	R pallidum	L pallidum	R putamen	L putamen
Cross-sec.[126]	**-0.5348**	**-0.2763**	-0.627	-0.5513	-0.4162	-0.6226	**-0.6836**	**-0.4264**
Longitudinal	-0.3254	-0.0960	**-0.6549**	**-0.6006**	**-0.6017**	**-0.6232**	-0.5589	-0.3597

8.4 Discussion and Conclusion

In this work, we proposed a 3D subcortical segmentation method leveraging longitudinal information. We used two 3D scans of a given subject as inputs and took advantage of the longitudinal context by using the bi-directional C-LSTM, such that information from both time points were learned by our model jointly. With the longitudinal information, the model learns the relationship between the scans and achieves superior segmentation performance with higher accuracy and better consistency for the considered subcortical structures. Additionally, our segmentations better correlate with TMS decline in a limited dataset. We note that our bi-directional C-LSTM blocks and C-LSTM cells are not using shared weights in our experiments. However, the shared weights blocks and cells produced nearly identical results to those reported in this paper. Extending our framework to allow more than 2 time-points per subject remains as future work; an important step towards this goal will be to optimize the network architecture to avoid GPU memory limitation issues. Another potential extension might be a multi-task network that handles the registration of the two time-points along with the longitudinal segmentation.

CHAPTER 9

Human Brain Extraction with Deep Learning

Brain extraction, also known as skull stripping, from magnetic resonance images (MRIs) is an essential preprocessing step for many medical image analysis tasks and is also useful as a stand-alone task for estimating the total brain volume. Currently, many proposed methods have excellent performance on T1-weighted images, especially for healthy adults. However, such methods do not always generalize well to more challenging datasets such as pediatric, severely pathological, or heterogeneous data. In this paper, we propose an automatic deep learning framework for brain extraction on T1-weighted MRIs of adult healthy controls, Huntington's disease patients and pediatric Aicardi Goutières Syndrome (AGS) patients. We examine our method on the PREDICT-HD and the AGS datasets, which are multi-site datasets with different protocols and scanners. Compared to current state-of-the-art methods, our method produced the best segmentations with the highest Dice score, lowest average surface distance and lowest 95-percent Hausdorff distance on both datasets. These results indicate that our method has better accuracy and generalizability for heterogeneous T1-w MRI datasets.

9.1 Introduction

In many medical image analysis tasks, such as segmentation or registration, brain extraction (skull stripping) is widely used as a preprocessing step for removing non-brain tissue from the image. [124, 125, 241] In previous decades, many algorithms for brain extraction were proposed [193, 196, 90] with promising results, and many of these classical methods are still widely used today. Among these, Shattuck et al. [193] proposed the Brain Surface Extractor (BSE), which uses diffusion filters, edge detection, and morphological operations for skull-stripping. Smith developed the Brain Extraction Tool (BET) [196], which applies a deformable model to fit the brain surface by adaptive forces. The RObust Brain EXtraction (ROBEX) algorithm, proposed by Iglesias et al. [90], combines a discriminative model, a generative model, and a refinement step. In recent years, deep learning-based methods have outperformed these traditional methods for both healthy subjects and certain diseases, in both pediatric and adult cohorts [109, 245, 50, 94]. Notably, Isensee et al. developed the HD-BET [94], which is a deep learning model for skull-stripping that was trained on a large number of subjects with multiple MRI modalities. In addition to their high accuracy, deep learning models also have the advantage of being very rapid for skull-stripping MRIs in the test set once trained.

This work is published at SPIE 2022.
Li, Hao, et al. "Human brain extraction with deep learning." Medical Imaging 2022: Image Processing. Vol. 12032. SPIE, 2022.

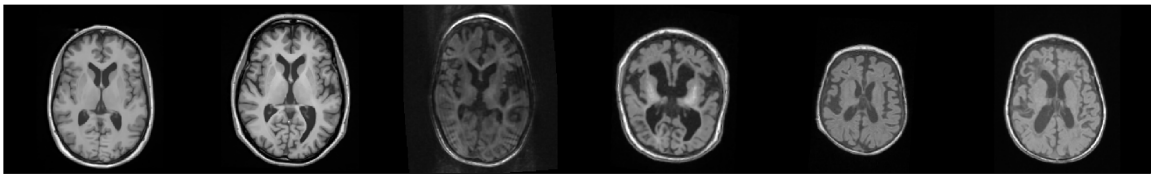


Figure 9.1: Anatomical and contrast variability across study datasets. (a) Healthy control subject, (b) HD patient, (c-f) AGS patients.

Despite their performance and computational efficiency, deep learning-based skull-stripping methods require appropriate training data that accurately represents the anatomical variability and acquisition protocols of the test data and do not always generalize well between protocols/scanners. In this work, we propose **a single skull-stripping method that can be applied to a variety of age groups, disease status, and acquisition protocols**. Specifically, we consider a multi-site adult dataset of healthy controls and Huntington's disease (HD) patients [143] as well as a pediatric dataset of Aicardi Goutières Syndrome (AGS) patients. HD is a fatal genetic disorder, with relatively gradual neurodegeneration through the disease progression, affecting the striatum and cortical thickness. Because of the cortical atrophy, HD patients typically have more cerebrospinal fluid (CSF) around the brain than healthy controls (Fig. 9.1, a-b). AGS is a pediatric heritable disorder of excessive interferon production affecting the brain, especially the white matter and the striatum [207]. AGS patients typically range from infants to adolescents, which is more challenging for brain extraction due to the different brain sizes, shapes, and tissue contrasts, as well as the severe atrophy present in some AGS patients (Fig. 9.1, c-f). Currently, **no automated methods** can reliably segment the brain in patients with severe AGS, which is a significant barrier in the study of this disease since even basic morphometry tools such as total brain volume estimation is challenging. Additionally, without robust skull stripping, downstream tasks such as registration and segmentation become even more difficult.

In this paper, we propose a fully convolution neural network for human brain extraction. Our model derives from the 3D U-Net [35] with modifications to improve the model accuracy and generalizability. Specifically, we use residual blocks, [68] convolutional blocks in skip connections, and add outputs from each level. [126] We evaluate our model on multi-site adult and pediatric datasets, consisting of different acquisition protocols and scanners.

9.2 Methods

9.2.1 Data and Preprocessing

Adult dataset. We use a subset of the multi-site PREDICT-HD database, with 3D T1-w 3T MRIs of 40 healthy control subjects and 119 HD subjects (Table 9.1), the details are same as Chapter 5. In addition, the

Table 9.1: Overview of datasets in our experiment.

	Age mean (stdev.)	Gender (F/M)	1.5T scanners	3T scanners
Control	45.87 (11.95)	26/14	0	12
HD	45.71 (12.89)	74/45	0	17
AGS	4.47 (5.08)	38/36	11	13

experimental settings and preprocessing steps are kept same as Chapter 6. Briefly, we randomly select 20 subjects for training and 2 for validation from healthy controls. The remaining 15 healthy controls and all HD subjects are used for testing.

Pediatric dataset. We use a private multi-site dataset of 3D T1-w MRIs (1.5T and 3T) from 74 AGS patients (Table 9.1). We randomly select 20/5 MRIs as training/validation data, and the rest for testing. **Preprocessing.** The preprocessing for this dataset contains the following steps: (1) rigid registration to the template, (2) N4 bias field correction, (3) repeat rigid registration to the template (for additional robustness), and (4) histogram matching. To create the ground truth brain masks, every 5^{th} slice of the MRIs were manually segmented, and the remaining slices were filled in via interpolation.

Template. We used the symmetric UNC pediatric template[1] for registration, which is created using 10 4-years-old subjects. We resampled the template to isotropic resolution $1 \times 1 \times 1\text{mm}^3$, resulting in $176 \times 192 \times 176$ voxels.

9.2.2 Network Architecture

Fig. 9.2 shows our proposed network architecture. Our model is adapted from 3D U-Net with 3D MRIs as inputs and 3D brain masks as output. Like the 3D U-Net, the network consists of an encoder and a decoder. 3D max-pooling and 3D nearest neighbor upsampling were used in the encoder and decoder, respectively. There are 4 skip connections between encoder and decoder at each level for passing the coded information to the decoder. To improve the accuracy and generalizability of the model, we modified the network as follows: (1) we applied the convolution operator with kernel size 5 in the skip connection, followed by batch normalization and ReLU activation, (2) we stacked residual blocks in the encoder and decoder path, (3) to preserve the global information, we added the feature maps at each level in the decoder path to form the final segmentation. Details of the residual blocks can be found in Fig. 9.2.

[1] https://www.nitrc.org/frs/?group_id=288

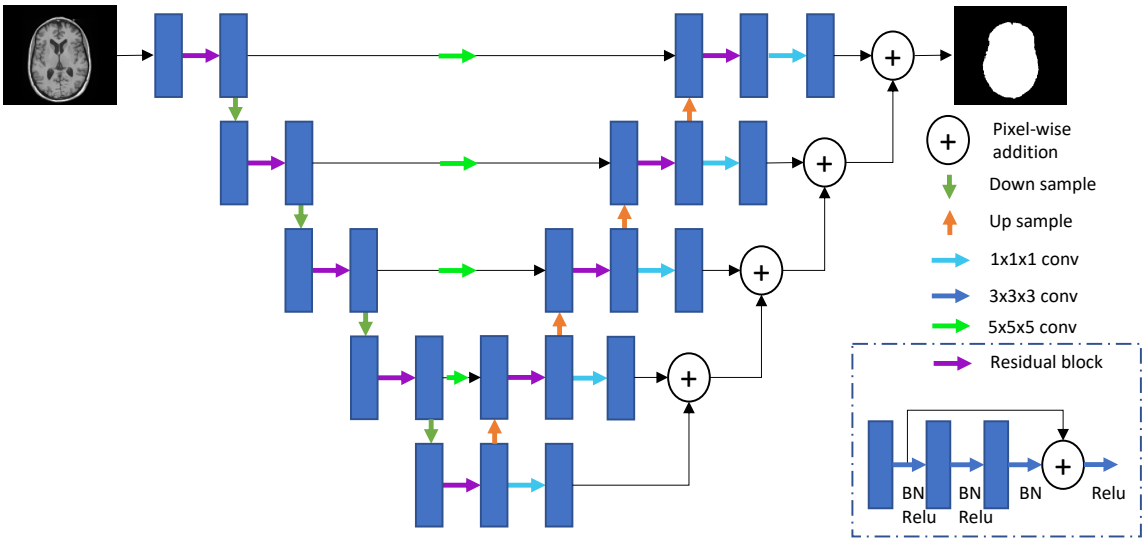


Figure 9.2: Proposed network architecture. Blue boxes are feature maps, with 16, 32, 64, 64, 128 channels at respective levels. The details of the residual block are shown in the dashed inset.

9.2.3 Implementation Details

We used the Dice loss [154] as our loss function, and set the weights to $w_b = 0.1, w_f = 1$ for background and foreground classes. We used the Adam optimizer with L2 penalty 0.00001, $\beta_1 = 0.9$, $\beta_2 = 0.999$. The initial learning rate was set to 0.0001. Additionally, we selected the epoch with the best results during the validation process as the final model. For a fair comparison, we set the same hyper parameters for all models. Besides these, the same early stop conditions are used in the training phase. The segmentations are obtained after a post-processing step, which extracts the largest connect component. We used a batch size of 2. We implemented the models using an NVIDIA TITAN RTX and Pytorch.

9.3 Results

Quantitative results in the adult dataset. The Dice coefficient, average surface distance (ASD), and 95-percent Hausdorff distance (95HD) are used as our metrics for the PREDICT-HD dataset (healthy controls and HD patients). We compare our proposed method to BET [196], BSE [193], ROBEX [90], HD-BET [94], and 3D U-Net [35]. These quantitative results are presented in Table 9.2. Compared to these state-of-the-art methods, our proposed method produced the best segmentations, evidenced by the highest Dice scores and the lowest ASD/95HD scores. We applied the 2-tailed paired t-test for statistical analysis, which showed that the proposed method has significantly improved ($p < e^{-10}$) all metrics compared to each alternative method (for brevity, we do not list each of these p-values in Table 9.2).

Qualitative results in the adult dataset. We present typical results from the PREDICT-HD dataset in

Table 9.2: Quantitative results for PREDICT-HD and AGS datasets. Bold numbers indicate best performance. Significant improvements between the proposed method and 3D U-Net (2-tailed paired t-test, $p < 0.05$) are denoted via $*$.

	HD dataset			AGS dataset		
Method	Dice	ASD	95HD	Dice	ASD	95HD
BET	89.10 (.0388)	5.332 (2.326)	25.8 (9.79)	81.81 (.1494)	7.916 (7.352)	26.9 (16.4)
BSE	93.76 (.0312)	2.566 (2.287)	10.2 (9.62)	94.30 (.0478)	2.061 (1.832)	8.14 (7.86)
ROBEX	95.23 (.0110)	1.358 (0.378)	5.08 (1.71)	94.52 (.0337)	1.510 (0.738)	5.07 (2.02)
HD-BET	96.01 (.0073)	1.113 (0.250)	3.96 (0.87)	95.69 (.0282)	1.204 (0.952)	4.85 (4.02)
3D U-Net	98.21 (.0040)	0.360 (0.131)	1.61 (0.41)	95.07 (.0363)	1.772 (1.856)	7.41 (10.0)
Proposed	**98.32 (.0040)**$*$	**0.322 (0.121)**$*$	**1.45 (0.44)**$*$	**96.12 (.0163)**	**1.146 (0.627)**	**3.97 (3.58)**

Fig. 9.3. Arrows highlight errors in segmentation; we note that such local errors may not be fully reflected in the Dice score, since Dice is relatively insensitive to small local errors in large structures. We also note that the "ground truth" we used is from the multi-atlas segmentation method and includes small defects. It is not hard to see that BSE and ROBEX delivered plausible segmentations overall. However, these segmentations are partially incorrect, notably near the dura. BET often over-segmented near the neck area. Although the 3D U-Net has better results than conventional methods, it still made some errors. Despite some local errors, HD-BET delivers reliable brain masks when directly applied to this HD dataset. Compared with all these alternative methods, the segmentations produced by the proposed method are the closest to the "ground truth", and appears to even compensate for the small mistakes in the "ground truth". We note that our segmentation is also more smooth than the "ground truth".

Quantitative results in the pediatric dataset. Table 9.2 also shows the quantitative results in the pediatric AGS dataset. Our method achieved the best performance among the state-of-the-art methods for each metric. The improvement over the baseline methods is more pronounced in this more challenging dataset, indicating the robustness of our method. We also note that our method has good ability to deal with the heterogeneous multi-site MRI dataset, as evidenced by its small standard deviation.

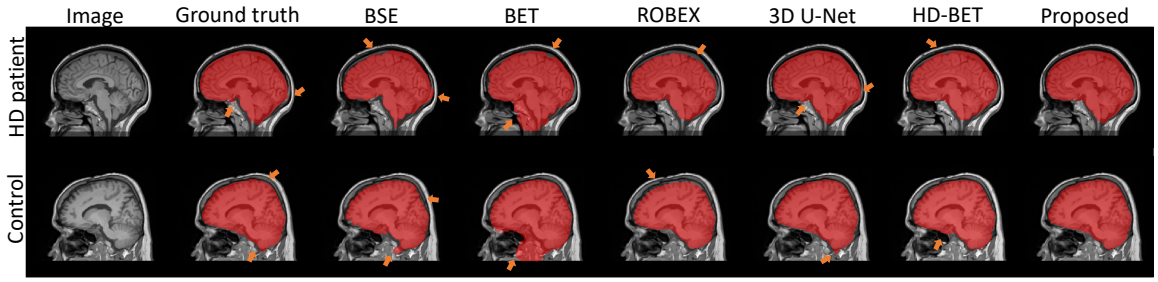


Figure 9.3: Qualitative results from the PREDICT-HD dataset. Top: HD patient. Bottom: healthy subject. Our proposed method produced the best segmentations. Segmentation errors are marked by arrows.

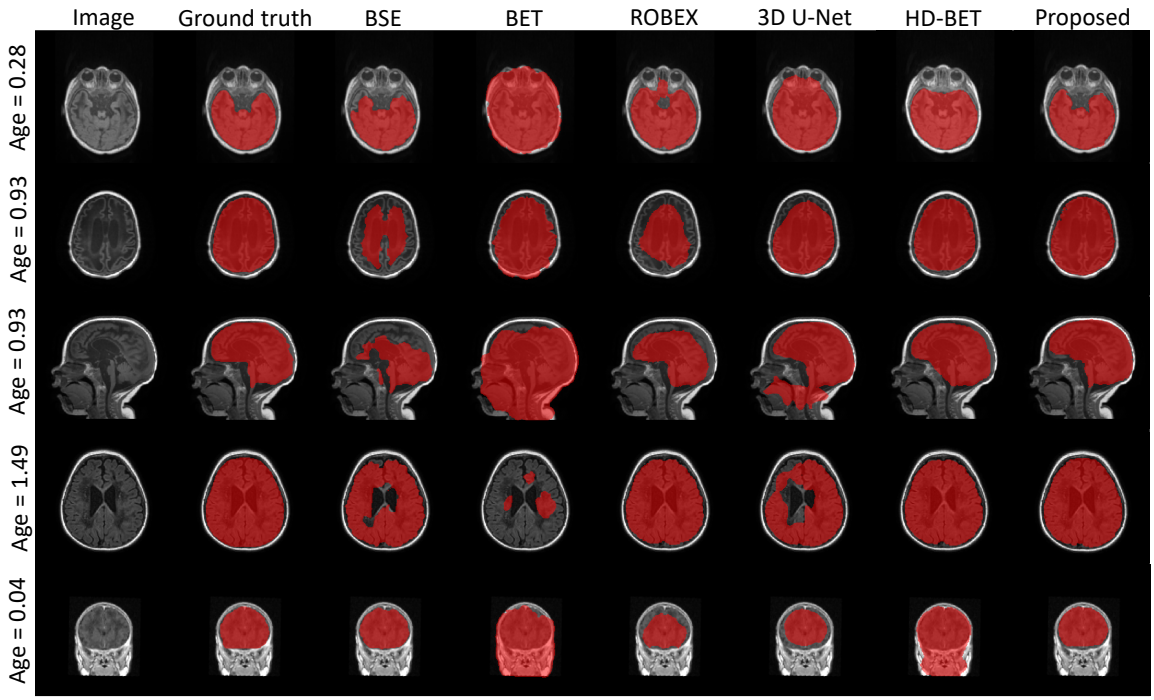


Figure 9.4: Qualitative results of AGS dataset, ages are posted on the side.

Qualitative results in the pediatric dataset. The pediatric dataset contains a wide age range from infants to adolescents and beyond, which means the images present a wide range of brain sizes and tissue contrasts. Many AGS patients also present severe brain atrophy. Fig. 9.4 shows the qualitative results in representative subjects. Small segmentation errors are present for all compared methods in this challenging dataset. However, BSE produces severely under-segmented brain masks. BET failed on some cases with either over-segmented or under-segmented segmentations. Although ROBEX has a good performance on the adult dataset, it does not appear to generalize well to these pediatric patients. Deep learning methods outperformed the conventional methods in adult dataset, but struggle in this challenging dataset. 3D U-Net produces unsatisfactory masks on cases with severe atrophy, enlarged ventricles, and/or reduced tissue contrast. HD-BET also outputs unreliable brain masks for such cases. In contrast, our proposed method delivers plausible segmentations which are the most similar to the ground truth, despite some local errors.

9.4 Discussion and Conclusion

In this paper, we proposed a deep learning framework for human brain extraction. We tested our proposed method on the PREDICT-HD dataset and the AGS dataset. For both datasets, our method produced the best results among the compared state-of-the-art methods. For the adult dataset, although 3D U-Net produces promising segmentions, our proposed model significantly improves the segmentations and qualitative results

show more accurate local boundaries. For the pediatric dataset, our model handles challenging cases well, and outputs plausible brain masks. Importantly, this is the **first automated method for reliably segmenting the brain in AGS patients**, which allows the total brain volume to be automatically estimated and is an enabling step for further MRI analysis in AGS cohorts, such as tissue classification and regional segmentation.

In future work, we will explore domain adaptation techniques to reduce the contrast variation further. Additionally, the convolutional blocks and the loss function need to be further developed to improve the model generalizability and accuracy. Finally, applying our method to a larger cohort remains as future work.

CHAPTER 10

CATS: Complementary CNN and Transformer Encoders for Segmentation

Recently, deep learning methods have achieved state-of-the-art performance in many medical image segmentation tasks. Many of these are based on convolutional neural networks (CNNs). For such methods, the encoder is the key part for global and local information extraction from input images; the extracted features are then passed to the decoder for predicting the segmentations. In contrast, several recent works show a superior performance with the use of transformers, which can better model long-range spatial dependencies and capture low-level details. However, transformer as sole encoder underperforms for some tasks where it cannot efficiently replace the convolution based encoder. In this paper, we propose a model with double encoders for 3D biomedical image segmentation. Our model is a U-shaped CNN augmented with an independent transformer encoder. We fuse the information from the convolutional encoder and the transformer, and pass it to the decoder to obtain the results. We evaluate our methods on three public datasets from three different challenges: BTCV, MoDA and Decathlon. Compared to the state-of-the-art models with and without transformers on each task, our proposed method obtains higher Dice scores across the board.

10.1 Introduction

In recent years, convolutional neural networks (CNNs) with U-shaped structures have dominated the medical image segmentation field [35, 184, 94]. The U-shaped networks consist of an encoder and a decoder, with skip connections in between. The encoder extracts information by consecutive convolution and downsampling operations. The encoded information is sent to the decoder via skip connections to obtain the segmentation result. The U-Net [35] and its many variants have shown great performance on many medical image segmentation tasks [192, 78, 126, 240].

However, convolutional encoders are somewhat limited for modeling long-range dependencies, due to the local receptive field of the convolution kernels. A potential solution is the transformer, which was originally proposed in the context of nature language processing, and has been used for image segmentation and classification [43, 141, 227, 206, 189] due to its ability to better capture global information and modeling long-range context. For medical images, Chen et al. proposed the transUnet [28], which is the first segmentation framework with transformers. They added a transformer layer to the CNN-based encoder to better leverage global information. The UNEt TRansformers (UNETR) [64] is an architecture that completely re-

This work is published at ISBI 2022.
Li, Hao, et al. "Cats: complementary CNN and transformer encoders for segmentation." 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). IEEE, 2022.

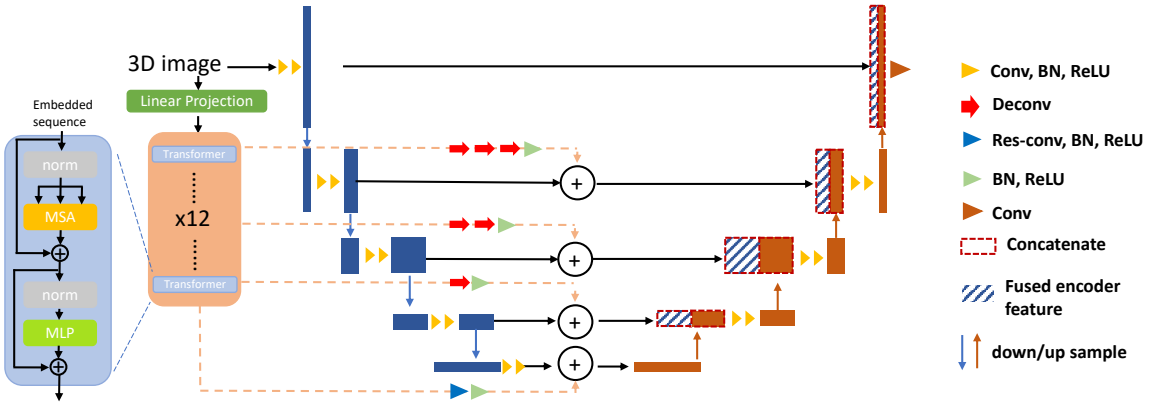


Figure 10.1: Proposed network architecture with two independent encoder paths: a CNN-based encoder and a transformer encoder.

places the CNN-based encoder with a transformer. The UNETR passes the multi-level scale information from this transformer to CNN-based decoder, and it is a state-of-the-art model for multi-organ segmentations on computed tomography (CT) images. Similar to U-Net, Cao et al. proposed Swin-Unet [18] which is a pure transformer-based U-shaped architecture for medical image segmentation. Since the transformer has limited ability to encode high-level information, these networks with transformers may not work well for some medical image segmentation tasks. To overcome this, Zhang et al. proposed a 2D architecture that combines a CNN-based encoder and a transformer-based segmentation network in parallel [247].

In this paper, we propose CATS (Complementary CNN and Transformer Encoders for Segmentation), a U-shaped architecture with double encoders. Inspired by UNETR [64] and TransFuse [247], we use a transformer as an independent encoder in addition to the CNN encoder. However, unlike the TransFuse, we use multi-scale features from the transformer rather than only the highest level features (i.e., the transformer output). The proposed CATS is a straightforward way to combine CNN and transformer without requiring a complex network architecture, such as the attention blocks and the BiFusion module in TransFuse. We use a 3D architecture and train from scratch instead of pre-training. Unlike the UNETR, we include both a transformer-based encoder and a CNN-based encoder. Multi-scale features extracted from the transformer are added with CNN features. The fused information is delivered to the CNN-based decoder for segmentation. We compare our model to state-of-the-art models with and without transformers on three public datasets.

10.2 Methods

10.2.1 Framework Overview

The proposed CATS framework is shown as Fig. 10.1. Our model contains two encoder paths, a CNN path and a transformer path. For the CNN-based encoder, the information is gradually coded by the convolution and

down-sampling operations. For the transformer path, to make sure the low-level details are well-preserved, we directly send the raw input to the transformer. Then, the information from the two paths are fused by addition operations at each level, and delivered to the CNN-based decoder to predict the final segmentation.

10.2.2 Transformer

We begin with an input 3D image $x \in \mathbb{R}^{C\times W\times H\times D}$, where W, H and D are the image dimensions and C is the number of channels. We construct $x_p = \{x_p^i | i \in [1,N]\}$, which is a set consisting of N non-overlapping patches $x_p^i \in \mathbb{R}^{(P^3\times C)}$, where $N = \frac{W\times H\times D}{P^3}$ and each x_p^i is a patch of P^3 voxels with C channels. Next, we send the patches to the linear projection layer to obtain the embedded projection $\boldsymbol{E}$ with dimension $M = P^3C$. To keep the position information, we add the position embedding $\boldsymbol{E_p}$ to form the transformer input:

$$z_0 = [x_p^1\boldsymbol{E}; x_p^2\boldsymbol{E}; ...; x_p^N\boldsymbol{E}] + \boldsymbol{E_p} \tag{10.1}$$

where $\boldsymbol{E} \in \mathbb{R}^{(P^3\times C)\times M}$ and $\boldsymbol{E_p}, z_0 \in \mathbb{R}^{N\times M}$.

The encoder path of transformer (Fig. 10.1) has L layers of Multihead Self-Attention (MSA) and Multi-Layer Perceptron (MLP) blocks:

$$z_l^{MSA} = MSA(norm(z_{l-1})) + z_{l-1} \tag{10.2}$$

$$z_l = z_l^{MLP} = MLP(norm(z_l^{MSA})) + z_l^{MSA} \tag{10.3}$$

where z_l^{MSA} and z_l^{MLP} are the outputs from MSA and MLP blocks, $norm(\cdot)$ denotes the layer normalization and l is the layer index. The MLP blocks contain two linear layers followed by the GELU activation functions.

There are n Self-Attention heads (SAs) in the MSA block to extract global information from the embedded sequence:

$$SA(z_i) = softmax(\frac{qk^T}{\sqrt{M_n}})v \tag{10.4}$$

where $z_i \in \mathbb{R}^{N\times M}$, q, k and v are the query, key and value of z_i respectively, $q = W_q z_i$, $k = W_k z_i$ and $v = W_v z_i$. Ws are the three weight matrices and $\sqrt{M_n} = \frac{M}{n}$ is the scaling factor. The output of the $softmax(\cdot)$ function is the similarity weight between q and k. Then the MSA is defined as:

$$MSA(z) = [SA_1(z); SA_2(z); ...; SA_n(z)]W_{msa} \tag{10.5}$$

where W_{msa} are the trainable weights.

Inspired by the UNETR [64], we use the same strategy for visualizing the multi-scale features z_3, z_6, z_9, z_{12} from transformer encoder path. The feature size of each z_i is $N \times M = \frac{H}{P} \times \frac{W}{P} \times \frac{D}{P} \times M$. We upsample

the z_i into the same size as the corresponding outputs from CNN-based encoder by deconvolution (deconv) operations, batch normalization (BN) and RELU activation function. The details can be viewed in Fig. 10.1. For the last level of the transformer features (z_{12}), we directly apply a residual convolution (res-conv) [67], BN and RELU for reshaping.

10.2.3 Convolutional Neural Network Architecture

Our CNN (Fig. 10.1) is adapted from the 3D U-Net [35]. There are two parts of the CNN model, the encoder and the decoder. Four max-pooling and deconvolution operations are used for down-sampling and upsampling respectively. The feature maps from the top level are directly forwarded to the decoder, and the rest of the feature maps from the lower levels are fused with the encoded information from transformer path by addition. Then, the fused information is delivered to the decoder by skip connections that follows the way of 3D U-Net.

10.3 Results

10.3.1 Datasets and Implementation Details

We used three public datasets for our experiments, and followed the evaluation metrics of each challenge.

Beyond the Cranial Vault (BTCV)[1] dataset contains 30/20 subjects with abdominal CT images for training/testing, with 13 different organs labeled by experts. To preprocess, we resampled the images and clipped HU values to range [-175, 250]. Three data augmentations were used: random flip, rotation and intensity shift. UNETR [64] is the winner for this challenge, and we compared our results to the publicly available UNETR implementation [2]. We also compare to the TransUNet model [28]. We note that both of these methods have been shown to be superior to CNN-only models such as 3D U-Net for this challenge [28, 64].

Cross-Modality Domain Adaptation for Medical Image Segmentation (CrossMoDA)[3] has 105 contrast-enhanced T1-weighted MRIs with manual labels for vestibular schwannomas (VS). We split the dataset into 55/20/30 for training/validation/ testing. The preprocessing pipeline consists of rigid registration to MNI space and cropping. Adam optimizer and Dice loss are used. We again compare our results to TransUNet[28] and UNETR [64], as well as to a 2.5D CNN [192] which was provided as the baseline for the challenge. We also report the 95% Hausdorff distance in this dataset in addition to the metrics used in the challenge.

Task 5 of Medical Segmentation Decathlon (MSD-5)[4] consists of 32 MRIs with manual prostate labels. 2 MRIs in validation were excluded due to the wrong labels being provided in the public dataset. We compare

[1] https://www.synapse.org/#!Synapse:syn3193805/wiki/217753
[2] https://github.com/Project-MONAI/tutorials/blob/master/3d_segmentation/unetr_btcv_segmentation_3d.ipynb
[3] https://crossmoda.grand-challenge.org/
[4] http://medicaldecathlon.com/

Table 10.1: Mean Dice scores in BTCV dataset. Bold numbers denote the highest Dice scores. The results of TransUNet are directly copied from [28]. The experiments of UNETR and proposed method use the public pipeline of UNETR. The organs from left to right are: spleen, right and left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, portal vein and splenic vein, pancreas, right and left adrenal gland, and overall average.

Method	Spl	RKid	LKid	Gall	Eso	Liv	Sto
TransUNet [28]	85.1	77.0	81.9	63.1	-	94.1	75.6
UNETR [64]	93.4	85.5	87.6	61.9	74.7	95.7	76.8
CATS	**95.8**	**90.2**	**93.4**	**65.9**	**77.1**	**96.8**	**83.0**
Method	Aor	IVC	Veins	Pan	RAG	LAG	Avg.
TransUNet [28]	87.2	-	-	55.9	-	-	77.5
UNETR [64]	85.2	77.2	69.8	61.5	64.4	59.4	76.9
CATS	**88.6**	**83.1**	**76.9**	**73.8**	**70.2**	**62.6**	**81.4**

to the TransFuse [247] model and the nn-Unet [94], which was the top-performing approach of the challenge at the time of the initial competition for this task. For a direct comparison with these two models, we use this dataset in a 5-fold cross-validation framework, and follow the setting in [94].

Implementation details. The training batch size was 2 for all three experiments, and constant learning rate was 0.0001. All intensities were normalized to range [0, 1]. We used Pytorch, MONAI and an Nvidia Titan RTX GPU.

10.3.2 BTCV Results

The Dice score is used to evaluate the BTCV experiment; the results can be viewed in Tab. 10.1. Our proposed method outperformed the state-of-the-art transformer-based models for each organ in this dataset. The most dramatic improvements (Dice improvement $> 5\%$) between UNETR and our proposed method are in left kidney, stomach, inferior vena cava, portal vein and splenic vein, pancreas and right adrenal gland. It is noteworthy that our method improves the segmentation accuracy not only on larger organs such as stomach and kidney, but also on small organs such as the right adrenal gland.

Fig. 10.2 shows qualitative results. Compared to the UNETR, our proposed model produces smoother segmentation for the stomach and liver (axial view, arrow). We can also see (coronal view, arrows) that UNETR undersegments the right kidney and liver, unlike our proposed model.

10.3.3 CrossMoDA Results

The quantitative results of CrossMoDA dataset are shown in Tab. 10.2. We report the Dice score, average surface distance (ASD) and 95-percent Hausdorff distance (HD95) as metrics. It is easy to observe that the CNN-only network has better performance than the transformer-only encoders for this task. However, our

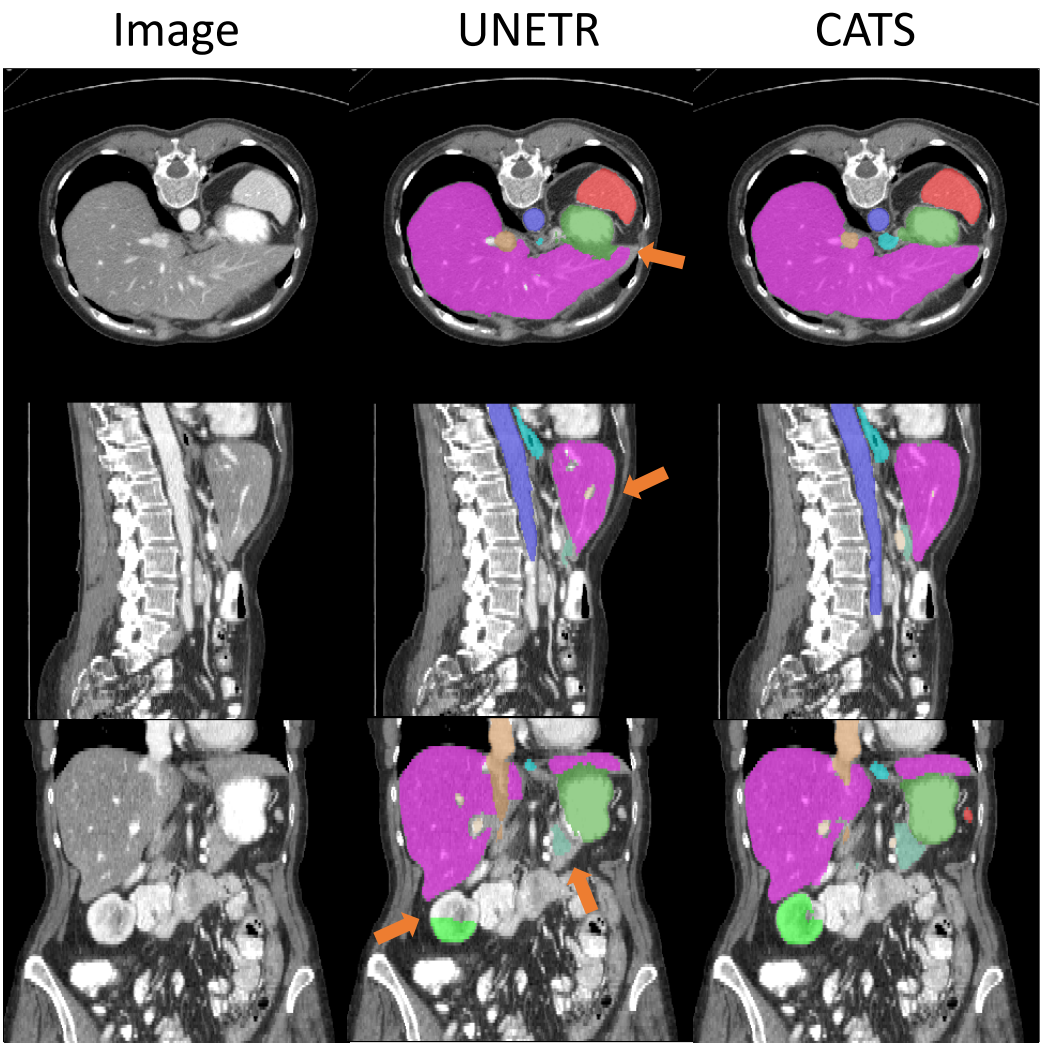


Figure 10.2: Qualitative results in BTCV test set. Some major differences are highlighted by orange arrows.

proposed CATS model outperformed the 2.5D CNN [192], which was specifically designed for segmenting VS from MRIs with large difference between in-plane resolution and slice thickness, as is the case for this dataset.

Fig. 10.3 shows the qualitative results of VS segmentation. While all methods appear to undersegment the VS, our proposed model most closely resembles the ground truth segmentations.

10.3.4 MSD-5 Results

We compare the nnUnet and TransFuse for the prostate segmentation in Tab. 10.3. Our proposed method has the highest Dice scores on all labels. Moreover, we improved the peripheral zone (PZ) nearly 4% compared to the performance of the TransFuse model.

Table 10.2: Quantitative results in CrossMoDA dataset, presented as *mean*(*std.dev.*). Bold numbers indicate the best performance.

Method	Dice	ASD	HD95
2.5D CNN [192]	0.856 (1.000)	0.69 (1.20)	3.5 (5.2)
TransUNet [28]	0.792 (0.234)	7.86 (27.6)	12 (31)
UNETR [64]	0.772 (0.139)	7.95 (14.2)	26 (43)
CATS	**0.873 (0.088)**	**0.48 (0.63)**	**2.6 (3.6)**

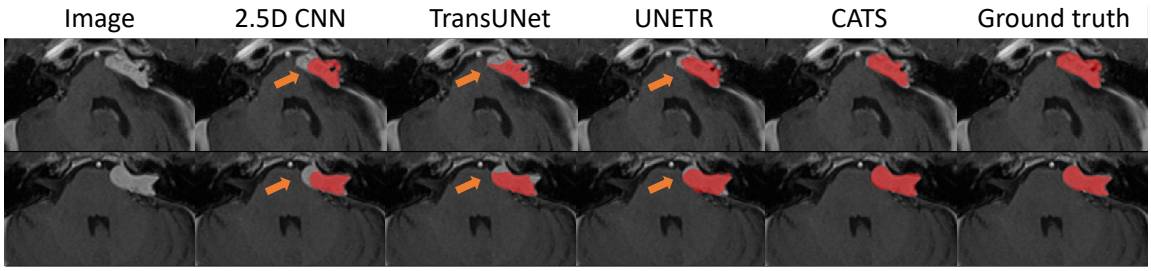


Figure 10.3: Quantitative results in CrossMoDA. Local segmentation errors are highlighted with arrows.

10.4 Discussion and Conclusion

In this paper, we propose a convolutional neural network with a transformer as an independent encoder. The transformer can complement the CNN by modeling long-range dependencies and capturing low-level details. We evaluate our proposed method on three public datasets: (1) BTCV, (2) CrossMoDA and (3) Decathlon. Compared to the state-of-the-art models which also attempt to incorporate transformers into the segmentation networks in various ways, our proposed model has superior performance on each task. We believe this is due to our efficient integration of the transformer and CNN encoders, as well as our use of a 3D architecture compared to 2D models. In future work, we will apply our method to larger public datasets. Additionally, others transformer layers may be helpful to further improve the performance.

Table 10.3: Mean Dice scores in MSD-5 dataset. PZ and TZ denote the peripheral zone and the transition zone of the prostate, respectively.

Method	PZ	TZ	Avg.
2D nnUnet [94]	0.6285	0.8380	0.7333
3D nnUnet full [94]	0.6663	0.8410	0.7537
TransFuse-S [247]	0.6738	0.8539	0.7639
CATS	**0.7136**	**0.8618**	**0.7877**

CHAPTER 11

CATS v2: Hybrid Encoders for Robust Medical Segmentation

Convolutional Neural Networks (CNNs) exhibit strong performance in medical image segmentation tasks by capturing high-level (local) information, such as edges and textures. However, due to the limited field of view of convolution kernels, it is hard for CNNs to fully represent global information. Recently, transformers have shown good performance for medical image segmentation due to their ability to better model long-range dependencies. Nevertheless, transformers struggle to capture high-level spatial features as effectively as CNNs. A good segmentation model should learn a better representation from local and global features to be both precise and semantically accurate. In our previous work, we proposed CATS, which is a U-shaped segmentation network augmented with transformer encoder. In this work, we further extend this model and propose CATS *v2* with hybrid encoders. Specifically, hybrid encoders consist of a CNN-based encoder path paralleled to a transformer path with a shifted window, which better leverage both local and global information to produce robust 3D medical image segmentation. We fuse the information from the convolutional encoder and the transformer at the skip connections of different resolutions to form the final segmentation. The proposed method is evaluated on three public challenge datasets: Beyond the Cranial Vault (BTCV), Cross-Modality Domain Adaptation (CrossMoDA) and task 5 of Medical Segmentation Decathlon (MSD-5), to segment abdominal organs, vestibular schwannoma (VS) and prostate, respectively. Compared with the state-of-the-art methods, our approach demonstrates superior performance in terms of higher Dice scores. Our code is publicly available at https://github.com/MedICL-VU/CATS.

11.1 Introduction

In recent years, deep learning (DL) has shown excellent performance in many medical image segmentation tasks [137]. Inspired by the success of the convolutional neural networks (CNNs) [180, 240, 126, 125, 78, 127, 119, 134] and Vision Transformer (ViT) [43, 189, 28, 64]. We proposed CATS, a hybrid network with CNN and Vit encoders for medical image segmentation (Chapter 10). However, ViT is computationally intensive and struggles to capture local information, especially for high-resolution medical data. As a variant of the ViT, the Swin Transformer [141] has shown good performance by computing representations hierarchically within shifted windows instead of applying self-attention to the entire image. Compared to ViT, the Swin Transformer reduces computational redundancies using the shifted window scheme, and it has been utilized

This work is published at SPIE 2024.
Li, Hao, et al. "CATS v2: Hybrid encoders for robust medical segmentation." Medical Imaging 2024: Image Processing. Vol. 12926. SPIE, 2024.

in medical applications to produce robust segmentations from high-resolution medical data [63, 173, 19, 253]. In addition to preserving global information, the shifted window approach also enhances the capture of local details. Given the importance of precise segmentation of anatomical structures and pathological regions in medical imaging, the ability to focus on fine-grained details is particularly advantageous for tasks like tumor and multi-organ segmentation [63, 173, 19, 253].

Although the shifted window approach is effective, it may still not match the local specificity of a carefully designed CNN for certain medical image segmentation tasks, because fine-level details can be of paramount importance in medical imaging. Thus, a hybrid approach that combines the strengths of both CNN and transformer might provide an optimal solution [247, 116, 121]. This raises the question: could hybrid encoders incorporating Swin Transformers enhance the current segmentation networks used for 3D medical image segmentation?

In this work, we introduce a 3D segmentation network with hybrid encoders named CATS *v2*. This is an improved version of our previous work, CATS (complementary CNN and transformer encoders for segmentation) [116], and offers better performance. In particular, we replace the ViT with the Swin Transformer, which is used as an additional independent encoder in a U-shaped CNN. The multi-scale features extracted from the Swin Transformer are fused with the features from the CNN and then delivered to the CNN-based decoder for segmentation. We evaluate the proposed methods on three different segmentation tasks, including abdominal organs, vestibular schwannoma (VS), and prostate, where large inter-subject variations are present. We compare our model to state-of-the-art models on three public datasets. The better performance of the proposed method in terms of Dice scores indicates that Swin Transformer improves the segmentation ability of existing segmentation networks with hybrid encoders. Moreover, our method has the potential to serve as a backbone for recent methods[121, 120, 233, 217] based on the Segment Anything Model (SAM [108]) in the field of medical image segmentation.

11.2 Methods

11.2.1 Framework Overview

Fig. 11.1 (a) shows the proposed segmentation network with hybrid encoders. Our model consists of two encoder paths: a CNN path and a transformer path with shifted window. The CNN-based encoder progressively encodes information using convolution and downsampling operations. On the Transformer path, the input images pass through the patch partition layer to reduce the dimension and visualize high-level features by a convolution operation and are then fed into the transformer blocks. The information from both paths is fused at each level using addition operations, and this combined information is delivered to the CNN-based decoder to predict the final segmentation.

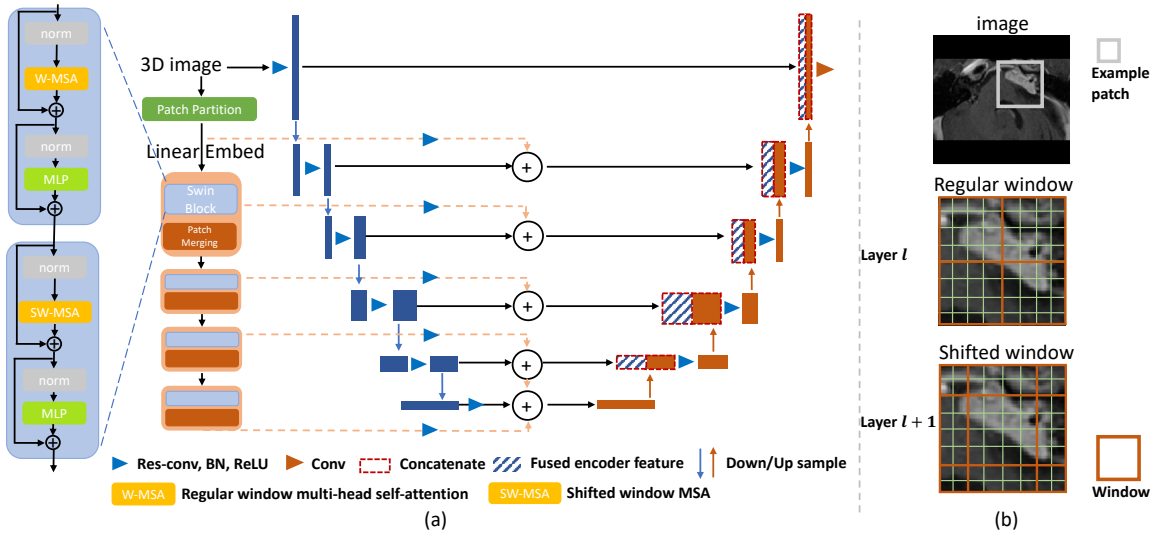


Figure 11.1: **(a)** Proposed network architecture. **(b)** 2D illustrations of shifted window where self-attention is only computed within each non-overlapping local window. Note that the patch sizes vary.

11.2.2 Swin Transformer Encoder

The proposed Swin Transformer encoder is adopted from [141, 63]. Specifically, the input of the Swin Transformer encoder is a 3D image, and a patch partition layer is applied to create a sequence of 3D patches/tokens with a given patch size. However, unlike ViT that flattens these patches and feeds them directly into the Transformer, non-overlapping local windows are created for efficient patch interaction modeling. Each local window goes through a linear projection layer to transform it into a sequence of token vectors. The transformed vectors are then processed by the self-attention mechanism of the transformer. Our encoder has four Swin blocks and each contains two successive transformer layers, i.e., regular window multi-head self-attention (W-MSA) and shifted window MSA (SW-MSA), which are shown in Fig. 11.1 (a).

Fig. 11.1 (b) demonstrates the shifted window scheme for subsequent transformer layers. In the layer l (W-MSA), we evenly partition the patch into subregions with same window size at each dimension. In the subsequent layer, $l+1$, the partitioned window regions are shifted by half of window size. The position of the windows is shifted to allow the model to gradually increase its receptive field and incorporate a more global context into its representations. To preserve the hierarchical structure of the encoder, a patch merging layer is employed at the end of each stage. This reduces the resolution of feature representations by a factor of 2, thereby decreasing the complexity and increasing the efficiency of the model. Following Hatamizadeh et al. [63], the embedding layer reduces the dimension of its input by half. Note that the linear projection layer enables the model to efficiently handle high-resolution inputs by reducing the dimensionality.

11.2.3 Convolutional Neural Network Architecture

Fig. 11.1 (a) also shows the proposed CNN, which is adapted from the 3D U-Net and its variants [35, 117]. Max-pooling and deconvolution operations are employed for down-sample and up-sample, respectively. The feature maps from the highest level are sent directly to the decoder, while feature maps from the lower levels are combined with encoded information from the Swin Transformer encoder path via addition. This fused information is then delivered to the decoder using skip connections, following the pattern of the 3D U-Net [35] to produce the final segmentation.

11.2.4 Implementation Details

We use three publicly available datasets, BTCV [112], CrossModa [42], and MSD-5 [4], in our experiments, which are same as Chapter 10. Dice score, average surface distance (ASD) and 95-percent Hausdorff distance (HD95) are used as evaluation metrics. The details of preprocessing steps for all datasets can be found in the original CATS paper [116] and Chapter 10.

We followed the implementation settings in CATS [116] for our experiments for a fair comparison. Briefly, we normalized the image intensity to range [0, 1]. The constant learning rate was set to 0.0001. Training batch size was 2 for all experiments which are conducted on Pytorch, MONAI and an NVidia Titan RTX GPU.

11.3 Results

11.3.1 BTCV Results

The quantitative and qualitative results of BTCV dataset are shown in Tab. 11.1 and Fig. 11.2, respectively. The compared methods include TransUNet [28], UNETR [64], Swin UNETR [63], CATS [116], and the proposed CATS *v2*. Briefly, UNETR [64] is composed of a ViT encoder and a CNN decoder, while Swin UNETR replaces the ViT with a Swin encoder. Similarly, CATS [116] is built upon the 3D U-Net [35] and integrates a ViT encoder. The proposed CATS *v2* employs a Swin encoder as the upgrade.

From Tab. 11.1, the proposed CATS *v2* achieves the best overall performance among the state-of-the-art compared methods (the 'Avg.' column). In the comparison between Swin UNETR and proposed CATS *v2*, we observe the improvements in 8 out of 13 organs when a CNN encoder is integrated. Furthermore, the proposed CATS *v2* outperforms original CATS in 7 out of 13 organs, with larger improvements observed in organs of smaller volume, such as the gallbladder, and the right and left adrenal glands. These improvements suggest that the Swin encoder could further refine the local details. Fig. 11.2 shows qualitative results, with major differences highlighted by orange arrows. Compared to the Swin UNETR and the original CATS, our proposed model produces smoother results.

Table 11.1: Mean Dice scores in BTCV dataset. Bold numbers denote the highest Dice scores. The results of TransUNet are directly copied from [28]. The experiments follow the public pipeline of Swin UNETR[63]. The organs from left to right are: spleen, right and left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, portal vein and splenic vein, pancreas, right and left adrenal gland, and overall average. Bold numbers indicate the best performance. The results can be found on the official leaderboard.

Method	Spl	RKid	LKid	Gall	Eso	Liv	Sto
TransUNet [28]	85.1	77.0	81.9	63.1	-	94.1	75.6
UNETR [64]	93.4	85.5	87.6	61.9	74.7	95.7	76.8
Swin UNETR [63]	**95.9**	87.8	92.9	65.7	77.2	96.5	83.3
CATS	95.8	**90.2**	**93.4**	65.9	77.1	**96.8**	83.0
CATS *v2*	94.8	87.1	93.2	**70.7**	**78.1**	96.7	**85.8**
	Aor	IVC	Veins	Pan	RAG	LAG	Avg.
TransUNet [28]	87.2	-	-	55.9	-	-	77.5
UNETR [64]	85.2	77.2	69.8	61.5	64.4	59.4	76.9
Swin UNETR [63]	85.5	82.8	75.1	72.5	**74.0**	**72.0**	81.6
CATS	**88.6**	**83.1**	76.9	73.8	70.2	62.6	81.4
CATS *v2*	88.0	82.5	**77.0**	**76.1**	72.2	66.3	**82.2**

Table 11.2: Quantitative results in CrossMoDA dataset, presented as $mean(std.dev.)$. Bold numbers indicate the best performance.

Method	Dice	ASD	HD95
2.5D CNN [192]	0.856 (1.000)	0.69 (1.20)	3.5 (5.2)
TransUNet [28]	0.792 (0.234)	7.86 (27.6)	12 (31)
UNETR [64]	0.772 (0.139)	7.95 (14.2)	26 (43)
CATS [116]	0.873 (0.088)	**0.48 (0.63)**	2.6 **(3.6)**
CATS *v2*	**0.886 (0.076)**	**0.48** (0.79)	**2.4** (4.0)

11.3.2 CrossMoDA Results

The quantitative results for the CrossMoDA dataset are presented in Tab. 11.2. We compare the models against a 2.5D CNN model [192], which was specifically designed to segment VS from MRIs characterized by substantial discrepancies between in-plane resolution and slice thickness, which is a common feature of this dataset. We observe that this CNN-only network performs better than the transformer-based encoders [28, 64] for this task. The original CATS [116] model outperformed the 2.5D CNN. With subsequent enhancements, our updated CATS *v2* model further refined the quality of segmentation, delivering the highest performance in terms of Dice score. Fig. 11.3 shows the qualitative results of VS segmentation. While the original CATS model undersegments the VS (marked by arrow), the proposed CATS *v2* effectively compensates for this

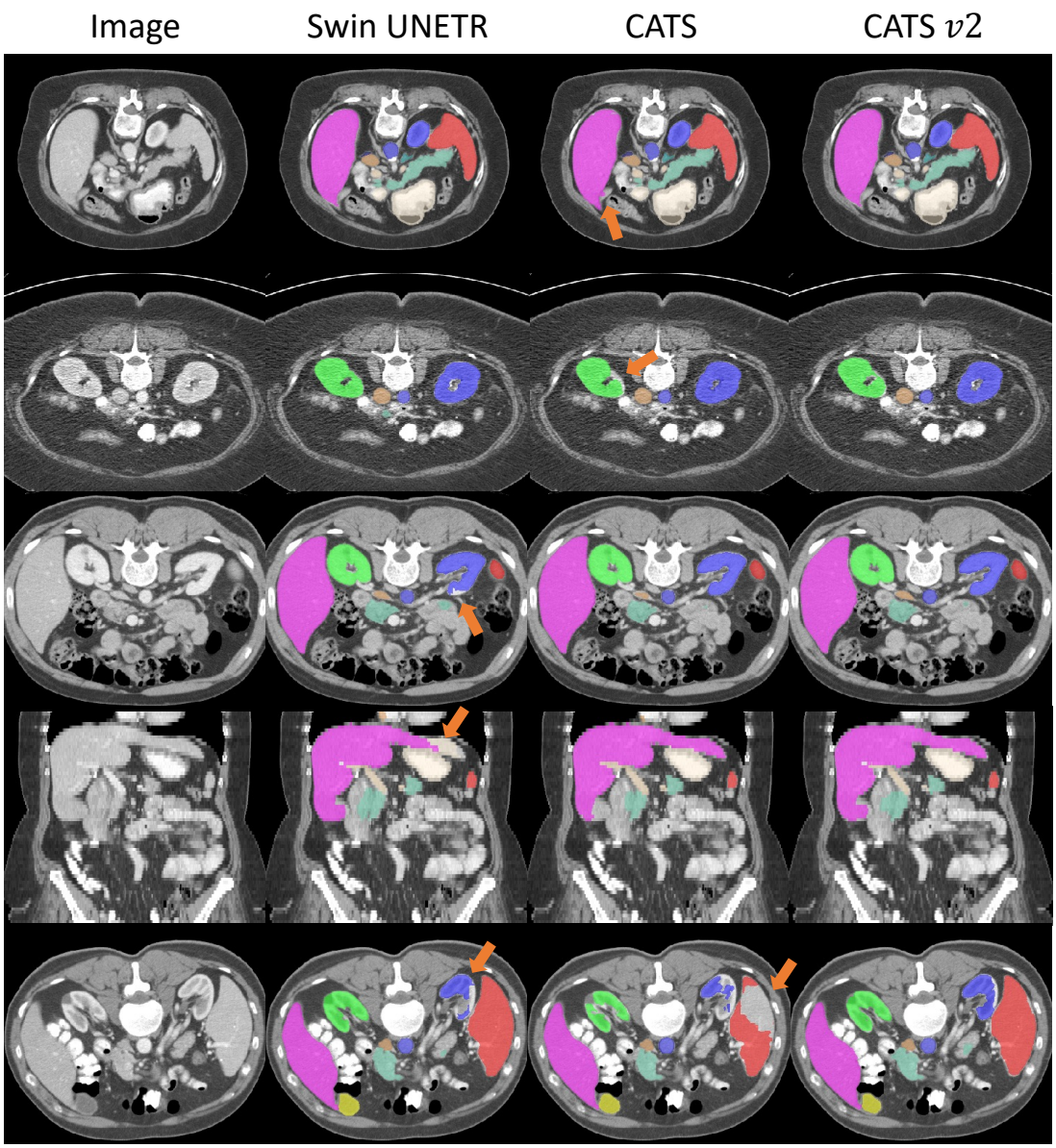


Figure 11.2: Qualitative results in BTCV. Some major differences are highlighted by orange arrows.

limitation and produces robust results that align more closely with the ground truth segmentations.

11.3.3 MSD-5 Results

We compared the nnUnet [94], TransFuse [247] and CATS [116] to our proposed method for the prostate segmentation task in Tab. 11.3. CATS *v2* has the highest Dice scores on all labels, i.e., both the peripheral zone (PZ) and the transition zone (TZ). This dataset was chosen because of the inherent challenge in segmenting two closely adjoined regions that exhibit considerable inter-subject variability. The qualitative improvements between original CATS and CATS *v2* are shown in Fig. 11.4. A more robust segmentation is produced by the

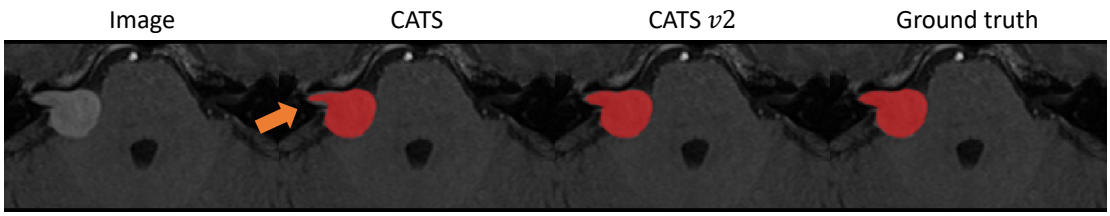


Figure 11.3: Qualitative results in CrossMoDA. Local segmentation errors are highlighted with arrows.

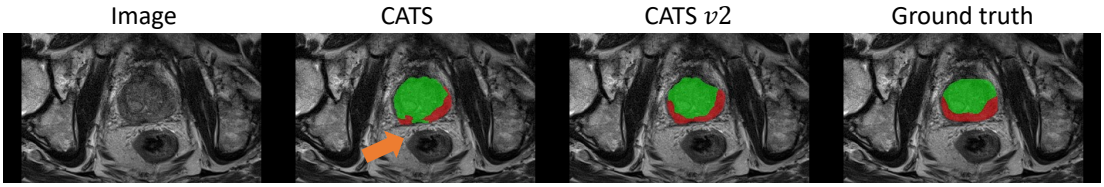


Figure 11.4: Qualitative results in MSD-5. Local segmentation errors are highlighted with arrows. Red and green labels denote the peripheral zone (PZ) and the transition zone (TZ), respectively.

proposed method by correcting the false positives.

11.4 Discussion and Conclusion

In this work, we introduce CATS *v2*, which is a segmentation network with hybrid encoders, specifically, a U-shaped CNN complemented with a Swin Transformer. We evaluated our proposed methods on three public datasets that present large inter-subject variations. Our proposed model outperforms state-of-the-art models on each task. Relative to the original CATS, the Swin Transformer is able to further enhance the segmentation ability of the encoder. However, we observe inconsistent improvements in the BTCV dataset, indicating that one encoder may dominate the results. Exploration of other fusion strategies to overcome this issue remains as future work. In addition, due to the use of hybrid encoders as well as deeper architecture design, our proposed network might require slightly more computational resources than the original CATS. In the future work, we aim to design a light-weight model for 3D medical image segmentation.

Table 11.3: Mean Dice scores in MSD-5 dataset. PZ and TZ denote the peripheral zone and the transition zone, respectively. Bold numbers indicate the best performance.

Method	PZ	TZ	Avg.
2D nnUnet [94]	0.6285	0.8380	0.7333
3D nnUnet [94]	0.6663	0.8410	0.7537
TransFuse [247]	0.6738	0.8539	0.7639
CATS [116]	0.7136	0.8618	0.7877
CATS *v2*	**0.7356**	**0.8713**	**0.8034**

Part III

Data-centric Medical Image Segmentation

CHAPTER 12

Unsupervised Cross-Modality Domain Adaptation for Segmenting Vestibular Schwannoma and Cochlea with Data Augmentation and Model Ensemble

Magnetic resonance images (MRIs) are widely used to quantify the volume of the vestibular schwannoma (VS) and cochlea. Recently, deep learning methods have shown state-of-the-art performance for segmenting these structures. However, training segmentation models may require manual labels in target domain, which is expensive and time-consuming. To overcome this problem, unsupervised domain adaptation is an effective way to leverage information from source domain to obtain accurate segmentations without requiring manual labels in target domain. In this paper, we propose an unsupervised learning framework to segment the VS and cochlea. Our framework leverages information from contrast-enhanced T1-weighted (ceT1-weighted) MRIs and its labels, and produces segmentations for T2-weighted MRIs without any labels in the target domain. We first applied a generator to achieve image-to-image translation. Next, we combined outputs from an ensemble of different models to obtain final segmentations. To cope with MRIs from different sites/scanners, we applied various 'online' data augmentations during training to better capture the geometric variability and the variability in image appearance and quality. Our method is easy to build and produces promising segmentations, with a mean Dice score of 0.7930 and 0.7432 for VS and cochlea respectively in the validation set of the cross-MoDA challenge.

12.1 Introduction

Vestibular schwannoma (VS) is a benign tumor of the human hearing system. For better understanding the disease progression, quantitative analysis of VS and cochlea from magnetic resonance images (MRIs) is important. Recently, deep learning frameworks have been dominating the medical segmentation field [192, 213, 40, 126, 191, 180] with state-of-the-art performances. However, supervised learning methods often require a high level of consistency between training and testing data. Consequently, such supervised methods often lack domain generalizability or ability to deal with images from various sites that have different intensity distributions, i.e., distribution shift or domain shift. Such a shift is usually caused by different image acquisition protocols or scanners; different image modalities could also be considered a domain shift problem.

Furthermore, in medical image analysis, lack of human delineations in one or multiple domains is another common issue, which is problematic for supervised learning. Unsupervised domain adaptation (UDA) is a

This work is published at Brainlesion 2021.
Li, Hao, et al. "Unsupervised cross-modality domain adaptation for segmenting vestibular schwannoma and cochlea with data augmentation and model ensemble." International MICCAI Brainlesion Workshop. Cham: Springer International Publishing, 2021.

solution for increasing generalizability of deep learning models to deal with new data from different domains.

In the cross-MoDA challenge[1], the ceT1-weighted and T2-weighted MRIs are provided, but only ceT1-weighted MRIs are labeled by experts. To obtain the segmentations on T2-weighted MRIs, we consider it as a UDA problem and propose an unsupervised cross-modality domain adaptation framework for segmenting the VS and cochlea. Our framework contains 2 parts: synthesis and segmentation. For synthesis, we apply a CycleGAN [258] to perform unpaired image translation between ceT1-weighted and T2-weighted MRIs. For segmentation, we use the generated T2-weighted MRIs as input and train an ensemble of models with various data augmentations, each of which yields candidate segmentations of VS and cochlea. We fuse those candidate segmentations to form the final segmentation.

12.2 Related Work

Supervised learning is an effective way for medical image segmentation when sufficient labels are available in target domain. Wang et al. [213] proposed an attention-based 2.5D convolutional neural network (CNN) to segment VS from T2-weighted MRIs with anisotropic resolution. In a following publication, Shapey et al. [192] employed this 2.5D CNN and further explored the performance of segmenting VS on both T1-weighted and T2-weighted MRIs. As we know, obtaining manual annotations is labor-intensive and time-consuming. Dorent et al. [40] introduced a novel weakly-supervised domain adaptation framework for VS segmentation on T2-weighted MRIs. In their work, only scribbles are needed as weak supervision in the target domain. They leveraged information from T1-weighted (source domain) MRIs to segment VS in the target domain based on co-segmentation and structured learning. However, in scenarios where there is no label available in target domain, UDA is a solution. Typical UDA methods try to align the image features between source domain and target domain [87, 27]. Once the features are well aligned, downstream tasks, such as segmentation, are relatively easy to accomplish. CycleGAN [258] and MUNIT [86] are popular methods to achieve unpaired image translation. Huo et al. [87] propose an end-to-end framework for unpaired synthesis between CT and MRI images, and jointly segment the spleen on CT images without any label from CT images during training. In their framework [87], CycleGAN is used for unpaired image translation. Chen et al. [27] present a method to segment the cardiac structures from late-gadolinium enhanced (LGE) images by leveraging information from balanced steady-state free precession (bSSFP) images. Similarly, no label is used from target domain (LGE images) during training. The MUNIT is used as image translation network. While our framework is similar to the approach proposed by Huo et al. [87], we use an ensemble model with various augmentation strategies for our segmentation component to improve the robustness of our results.

[1] https://crossmoda.grand-challenge.org/

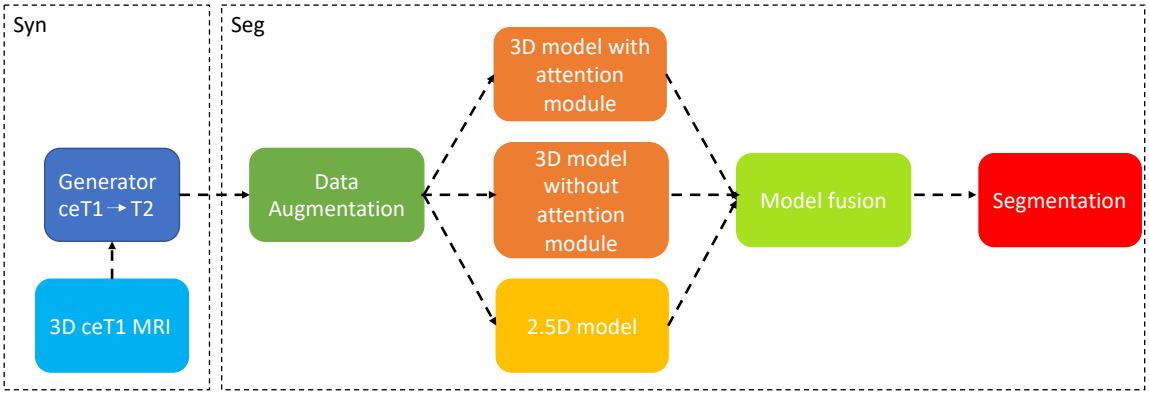


Figure 12.1: Proposed overall framework. Our framework contains 2 parts: synthesis (Syn) and segmentation (Seg). We use a CycleGAN model as generator in the synthesis part. Results from the different models (and with different augmentations) are fused to obtain the final segmentation.

12.3 Methods and Material

12.3.1 Dataset

The cross-modality domain adaptation for medical image segmentation challenge dataset (cross-MoDA[2]) contains two different MRI modalities: contrast-enhanced T1-weighted (ceT1-weighted) with an in-plane resolution of 0.4×0.4mm and slice thickness between $1 - 1.5$mm, and high-resolution T2-weighted with an in-plane resolution of 0.5×0.5mm and slice thickness between $1 - 1.5$mm. ceT1-weighted imaging was performed with an MPRAGE sequence and T2-weighted imaging with 3D CISS or FIESTA sequence. The training set contains 105 ceT1-weighted and 105 T2-weighted MRIs, and the validation set contains 32 T2-weighted MRIs. Expert manual VS and cochlea labels are available for the 105 ceT1-weighted training MRIs. More detailed information about this dataset can be found in [3].

12.3.2 Overall Framework

Fig. 12.1 displays our proposed framework. There are two parts in our framework: synthesis (Syn) and segmentation (Seg). For the Syn part, the 3D ceT1-weighted image, after undergoing pre-processing, is fed to the CycleGAN [258] pipeline to achieve image-to-image translation (ceT1-weighted to T2-weighted). Data augmentation strategies are applied on the generated T2-weighted MRIs to increase the model robustness. We send the augmented data into four different models: a 2.5D model with two different augmentation schemes, a 3D model with attention module, and a 3D model without attention module. We finally employ a union operation to fuse the outputs of each model to form the final segmentation.

[2] https://crossmoda.grand-challenge.org/CrossMoDA/
[3] https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=70229053

12.3.3 Preprocessing and Post-processing

Our preprocessing pipeline contains 5 steps: (1) non-local mean filter denoising, (2) image alignment with MNI template by rigid registration, (3) bias field correction, (4) image cropping based on region of interest (ROI), and (5) linear intensity normalization with range [0,1]. Before rigid registration, MNI template was resampled to the size of $512 \times 512 \times 128\,\text{voxel}^3$, with spatial resolution $0.377 \times 0.447 \times 1.508\,\text{mm}^3$. The ROIs are identified based on the labels of ceT1-weighted MRIs. We first make a common bounding box by taking the union of each bounding box over all labels. In order to cover all structures from all MRIs, the bounding box was then extended to the size $256 \times 128 \times 48\,\text{voxel}^3$.

In post-processing, we extracted the largest connected component for VS from the network output. Finally, we applied the inverse transformation from the rigid registration (step 2 of preprocessing) to move the segmentations back to their original space.

12.3.4 Synthesis: Image-to-image Translation

The CycleGAN [258] framework is used for image-to-image translation in 2D. We first split the 3D cropped ROIs into 2D slices. Next, we feed those 2D slices to the CycleGAN for training; in this context, we consider the ceT1-weighted and T2-weighted MRIs to be two different domains. After the training process, we stack 2D slices back together to form a 3D MRI volume. The model convergence is determined based on the best performance by visual inspection at each epoch.

12.3.5 Data Augmentation

Data augmentation is widely used in medical image segmentation to help minimize the gap between datasets/domains, producing more robust segmentations. Here, we design an 'online' augmentation strategy during training and randomly apply data transformations to input images. These transformations are in 3 different groups:

- **Spatial augmentation.** 3 types of random spatial augmentation are used: affine transformation with angle range of $[-10^\circ, 10^\circ]$, and scale factor from 0.9 to 1.2; elastic deformation with control points $= 7$, max displacement $= 6$; and a combination of affine and elastic deformation [174]. The same spatial augmentations and parameters are applied to both MRIs and the labels.

- **Image appearance augmentation.** To minimize the different image appearance between MRIs from different sites and scanners, we randomly apply multi-channel Contrast Limited Adaptive Equalization (mCLAHE) and gamma correction with γ from 0.5 to 2 to adjust image contrast.

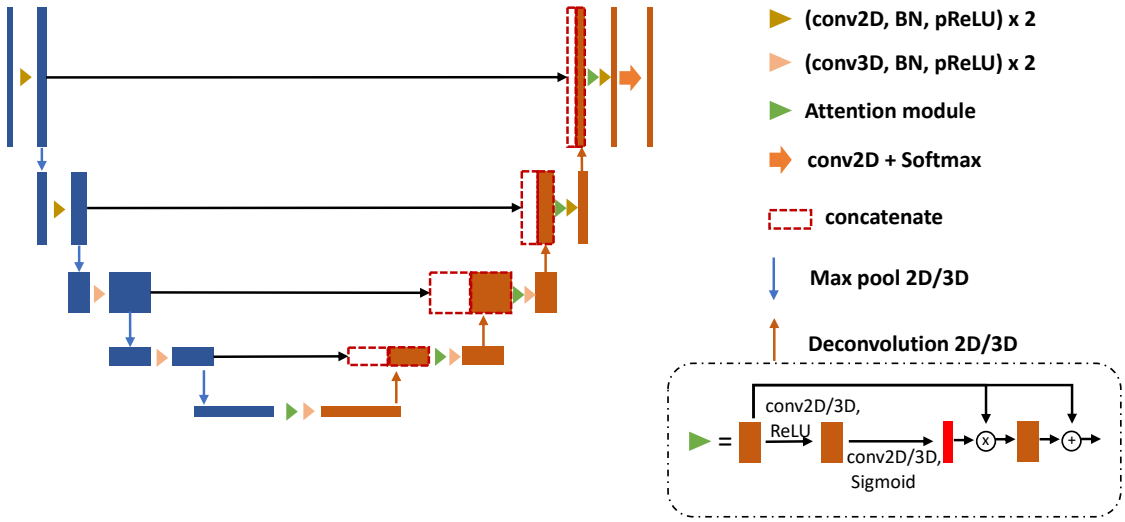


Figure 12.2: Network architecture of 2.5D model from the work [213, 192]. To deal with anisotropic image resolution, 2D convolutions are used in the first 2 levels, and 3D convolutions are used in levels 3-5.

- **Image quality augmentation.** In this context, image quality refers to resolution and noise level. We randomly blur the image using Gaussian kernel with σ_{blur} from 0.5 to 1.5, we add Gaussian noise with $\sigma_{noise} = 0.01$, and we sharpen the image by $I_s = I_b + (I_b - I_{bb}) \times \alpha$, where I_s is sharpened image, I_b is the image blurred with a Gaussian kernel (σ_{blur}), and I_{bb} is the image blurred twice with the same Gaussian kernel. In our case, we set $\alpha = 10, \sigma_{blur} = 1.5$.

12.3.6 Segmentation: 2.5D Model and its Architecture

We leverage a 2.5D CNN [213] model to alleviate the impact of the anisotropic image resolution. Network architecture details can be found in Fig. 12.2. This 2.5D network uses both 2D and 3D convolutions to capture both in-slice and global information. 2D convolutions are used in the first 2 levels and 3D convolutions for the remainder. Adapted from U-Net [180], the 2.5D model contains an encoder and a decoder, and the skip connections between encoder and decoder. The max-pooling and deconvolution operations connect the features between levels. Batch normalization and parametric rectified linear unit (pReLU) are used in the network architecture. An attention module is applied to assist in segmenting the small ROI. A second 2.5D model with identical architecture was also used in the ensemble, with the only difference being in the gamma correction parameter in the augmentation stage (with $\gamma \in [0.5, 1.5]$ rather than $[0.5, 2]$).

12.3.7 Segmentation: 3D Model and its Architecture

We used a fully convolutional neural network for our 3D models [126]. The network architecture details can be found in Fig. 12.3. Similar to a 3D U-Net, it consists of an encoder and a decoder. 3D max-pooling and

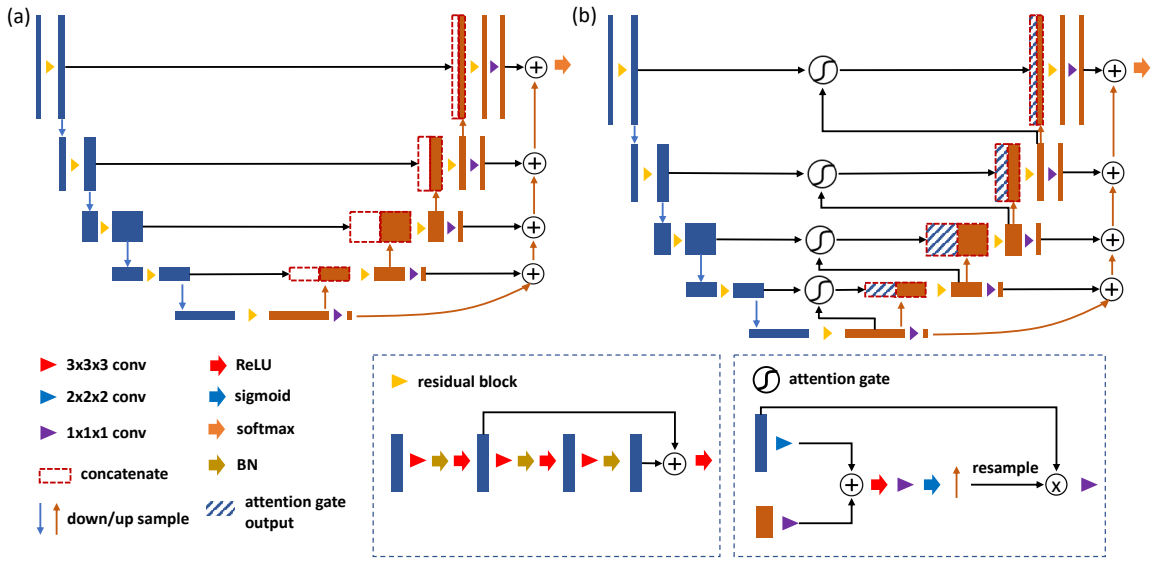


Figure 12.3: The details of 3D models. Residual block and deep supervision are used in the 3D model. (a) 3D model without attention module, (b) 3D model with attention module.

3D nearest neighbor upsampling are used in both the encoder and decoder. We further reinforce the output by adding the feature maps at the end of each level. In one of our two 3D models, we employ an attention module in the skip connections to emphasize the small ROI and preserve the information from encoder to decoder. Note that the two 3D models are identical except the attention module.

12.3.8 Implementation Details

The Adam optimizer was used with L2 penalty of 0.00001, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and an initial learning rate of 0.0002 for the CycleGAN generators [258] and 0.0001 for the segmentation models [213, 192, 126]. For the generators, the learning rate was left at the initial value for the first 100 epochs, and then dropped to 0 in following 100 epochs. For segmentation models, learning rate was decayed by a factor of 0.5 every 50 epochs. We evaluated our generator and segmentation models every epoch and selected the optimal model based on visual inspection of image quality for generator, and Dice score for segmentation. Based on the Dice loss [154], we defined our loss function as $1 - \text{mean}(\text{Dice})$ from multiple labels, with equal weight ($w_{FG} = 1$) for all foreground labels and decayed weight ($w_{BG} = 0.1$) for the background. The training, which has a batch size of 2, was conducted on NVIDIA GPUs and implemented using PyTorch.

Table 12.1: Quantitative results in validation phase. The Dice score and average symmetric surface distance (ASSD) reported as $mean(std.)$. Baseline denotes no cropping in pre-processing, followed by synthesis and segmentation. Crop denotes the cropped input with certain size. + represents the cumulative approach based on the previous method. Bold numbers indicate the best results in each column.

	VS		Cochlea	
Method	Dice	ASSD	Dice	ASSD
baseline	0.408(0.284)	8.205(11.327)	0.405(0.195)	1.697(3.684)
crop ($256 \times 128 \times 96\, voxel^3$)	0.637(0.324)	2.495(7.443)	0.603(0.219)	2.481(5.688)
crop ($256 \times 128 \times 48\, voxel^3$)	0.662(0.265)	3.008(3.916)	0.620(0.048)	0.454(0.158)
+ data augmentation	0.740(0.193)	2.481(9.592)	0.731(0.043)	**0.288(0.114)**
+ ensemble (proposed)	**0.794(0.156)**	**0.634(0.359)**	**0.741(0.041)**	0.294(0.060)

12.4 Results

12.4.1 Quantitative Results

Tab. 12.1 displays the quantitative results of our methods in the validation phase, reported as mean(stdev). The initial result of our method is shown as baseline in the Tab. 12.1, which uses the resampled MRIs without any cropping as input to the 3D model with attention module. Different crop sizes in the preprocessing step leads to different results, as evidenced by the results presented in Tab. 12.1. By adding various data augmentations, the performance of segmentation network is dramatically improved, 7.8% and 11.1% on Dice scores for VS and cochlea respectively. Lastly, using the model ensemble boosts the Dice score of VS to nearly 80%. We submitted the proposed method as our final result in the validation phase of the cross-MoDA challenge.

12.4.2 Qualitative Results

Representative qualitative results from 4 different subjects can be viewed in Fig. 12.4. In the first row of Fig. 12.4, we observe that our 2.5D method produces an accurate segmentation. Varying the augmentation parameters of the 2.5D model improves performance in some challenging cases (e.g., fourth row of Fig. 12.4-(c)). However, in some cases, both 2.5D models under-segments the VS, which can be seen in the second and third rows of Fig. 12.4-(b,c). In such cases, 3D models compensate for this tendency of the 2.5D models. Although the attention module has good ability to capture small ROIs, which can be observed in the second row of Fig. 12.4-(e), it may also lead to over-fitting (first and third rows of Fig. 12.4-(e)). Thus, we also include a 3D model without attention module (first and third rows of Fig. 12.4-(d)) in the model ensemble for best results. The model fusion balances the strengths of the individual models and is able to produce consistently good segmentations in a variety of images (Fig. 12.4-(f)).

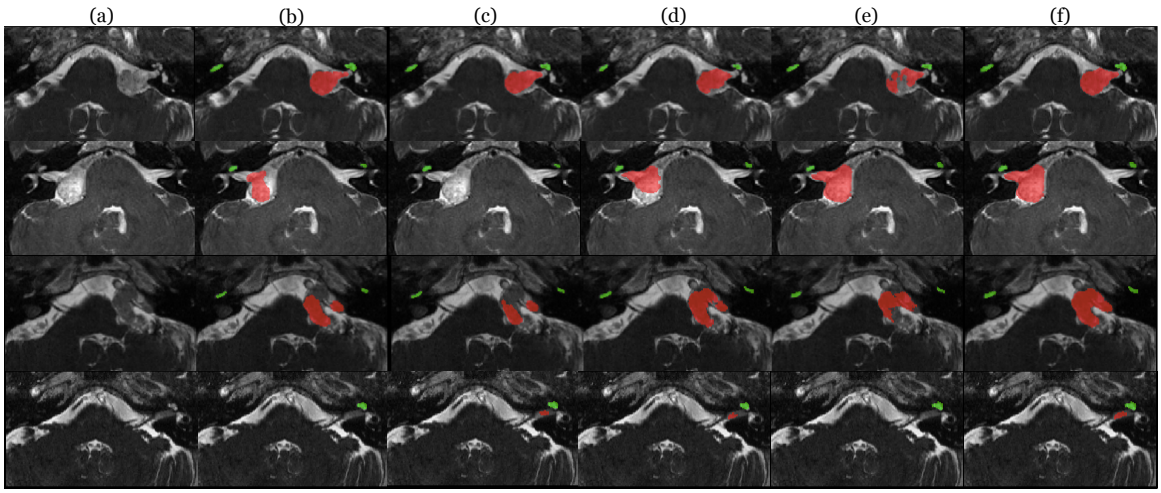


Figure 12.4: Segmentation results from 4 different subjects. Red, VS, green, cochlea. (a) Input image. (b) 2.5D model with $\gamma \in [0.5, 2]$. (c) 2.5D model with $\gamma \in [0.5, 1.5]$. (d) 3D model without attention. (e) 3D model with attention. (f) Final segmentation. While each individual model has performance issues in some rows, the final model fusion step produces consistently good segmentations for all rows.

12.5 Discussion

There are four main design choices of our method that merit discussion: (1) cropping ROI, (2) image-to-image translation, (3) segmentation network, and (4) self-training strategy.

- **Cropping ROI.** Given the small size of the segmented objects and the overall MRI size, cropping an ROI is an essential preprocessing step. Using the cropped ROI as input not only reduces redundant computation and improves the computational efficiency by letting the network focus on the the structures of interest increases accuracy. An additional benefit is the reduced GPU memory requirements. However, determining the optimal size for cropped MRIs is not straightforward. Too small ROIs might not fully contain all the structures of interest; however, too large ROIs would negatively impact the quality of the image-to-image translation and segmentation results by increasing the intensity variability in the input samples. How to balance the size of ROIs and the quality of generated MRIs from image-to-image translation remains an important problem.

- **Image-to-image translation.** Image-to-image translation is a critical step in our pipeline and generally in UDA problems [87, 27]. Moreover, generating well-aligned pseudo T2-weighted MRIs from source domain (i.e., from ceT1-weighted MRIs) could determine the accuracy of segmentations in the target domain, by minimizing the gap between the two domains. Thus, choosing a suitable image-to-image translation method is important. In our experiments, we employ CycleGAN [258], a popular method for unpaired image-to-image translation. However, synthesis artifacts (intensity shift between slices) appear along the depth direction of generated images in 2D CycleGAN. This is a common issue of

2D CycleGAN for translating volumetric MRIs, since slices are mapped independently from each other during training. We also found that 3D CycleGAN requires large amount of GPU memory, and the quality of images generated by using downsampled input MRIs is not satisfactory. Furthermore, patch-based 3D CycleGAN approaches could blur MRIs along the borders of the patches. However, modifying the architecture of the generator in CycleGAN pipeline could lead to satisfactory results [250]. Thus, improving the quality of results from image-to-image translation is another way to boost the performance. This argument is not only about the training manner of CycleGAN pipeline, but also more generally for methods such as CUT [166] and MUNIT [86].

- **Segmentation network.** In our pipeline, the segmentation performance is primarily determined by the quality of the generated MRIs in the target domain. Nevertheless, given the same input MRIs, different segmentation models produce different results. In our work, we combine results from an ensemble of models to form the final segmentation, since each model could bring different advantages, even with small changes. The design choices then become the specific models and the ensembling strategy. Additional models with different hyper-parameters could thus improve the segmentation results. The good performance of the top teams in the challenge [41] suggests that the nnU-Net [194], which is a well-known framework for biomedical image segmentation, could provide a good alternative for our ensemble models. Additionally, while we currently use a simple union operation, a more sophisticated ensembling strategy could be designed to minimize the weaknesses of each model while maximizing the strengths.

- **Self-training strategy.** Finally, the training strategy contributes to the accuracy of segmentation results. Compared to other teams [41], our training strategy is potentially the biggest limitation of our work, since we just use traditional supervised learning after generating T2-weighted MRIs. Following this training strategy, we only use the generated images instead of the real images in the segmentation task, which leads to our model lacking information in the original target domain. In an ideal situation, the features of generated T2-weighted MRIs should be very close or identical to the target domain, and the segmentation model could achieve a good accuracy by leveraging those features. However, given the limitations of image-to-image translation, not all of the generated T2-weighted MRIs are well-aligned to the target domain. Self-training could be a good approach to better handle such situations in the challenge [41]; this may involve directly segmenting real T2-weighted MRIs by a pre-trained model with generated T2-weighted MRIs as inputs, and designing an algorithm to find plausible segmentations among the results. Next, we could use these plausible segmentations as 'ground truths' to further finetune the pretrained segmentation model. Iterating this process could further improve the

segmentation accuracy. We believe that using such a self-training strategy would improve the final segmentation results.

12.6 Conclusion

In this work, we proposed an unsupervised cross-modality domain adaptation framework for VS and cochlear segmentation. There are two parts in our framework: synthesis and segmentation. We applied various 'online' data augmentations to deal with the MRIs from different sites and scanners. In addition, a model ensemble is used for increasing the performance. In the validation stage of the crossMoDA challenge, our method shows promising results.

CHAPTER 13

Deep Learning Based Unsupervised Domain Adaptation via Unified Model for Multi-site Prostate Lesion Detection: A Large-scale Study on Diffusion MRI with Various B-values

In Chapter 12, we proposed an unsupervised domain adaptation (UDA) method to segment vestibular schwannomas (VS) and cochleae in unlabeled T2-weighted images (target domain) by leveraging information from labeled T1-weighted images (source domain). However, in practice, multi-domain shifts in medical image segmentation present common challenges, particularly for large-scale datasets collected from multiple imaging sites where data are acquired under varied conditions from diverse populations. Typically, UDA methods, as described in Chapter 12, require training multiple generators for each domain pair, which becomes unfeasible when faced with a large number of domains. In this work, we addressed the multi-domain problem for prostate lesion detection from diffusion-weighted images with various b-value pairs, where each pair is considered a unique domain. Our hypothesis is that UDA using diffusion-weighted images, generated with a unified model, offers a promising and reliable strategy for enhancing the performance of supervised learning models in multi-site prostate lesion detection, especially when various b-values are present. We have proposed a novel UDA framework with a unified model to increase detection accuracy and tested it on a dataset comprising 5,150 cases collected from nine different clinical sites.

13.1 Introduction

Prostate cancer (PCa) is one of the most common cancer in men, and once detected early the patient can have an improved prognosis, including better treatment outcomes and lower mortality rates [167, 201]. Earlier studies have shown promising results on early PCa diagnosis by using multi-parametric magnetic resonance imaging (mp-MRI) [1, 104] or bi-parametric magnetic resonance imaging (bp-MRI) [111, 6]. Recently, deep learning-based methods have achieved high performance for PCa detection by leveraging information from bp-MRI images [183, 237, 222, 182, 46]. These methods could boost productivity of radiologists by shortening the time needed for interpreting imaging through automated lesion detection. Additionally, they have the potential to heighten diagnostic accuracy, notably for less experienced radiologists, and to enhance consistency among different readers [222]. In these techniques, DWI stands out as a crucial element, offering a pronounced distinction in signal intensities between cancerous and healthy tissues. This distinction is especially noticeable in apparent diffusion coefficient (ADC) maps and high b-value images.

Many convolutional neural networks (CNNs) from prior studies were trained and tested using supervised

This work is submitted to Radiology: Artificial Intelligence.

learning (SL) methods on datasets either from a single institution or multiple sites that adhered to similar acquisition protocols, particularly with b-value settings as recommended by the Prostate Imaging-Reporting and Data System (PI-RADS [205]) guideline. Under these conditions, test samples are tightly matched to the training set, allowing CNNs to yield reliable results for such in-distribution (ID) data. However, in real-world scenarios, clinical sites may have their own preferences for b-value selections. While variances in ADC and high b-value images due to diverse b-value choices may seem negligible to human observers, they can significantly influence deep learning models. This is largely attributed to the domain shifts observed across images from different datasets [60]. The performance of CNN drops and produces inaccurate results when encountering domain shifts or processing an out-of-distribution (OOD) test sample from target domain, whose b-values are not included in the training set [139]. Fig. 13.1(a) provides an illustrative example.

One straightforward way to overcome the domain shift problem is to retrain the generic model with a larger dataset to enlarge the b-values distribution. Yet, in practice, it is hard to transfer data between different imaging centers to get extra training data, and obtaining human labels for large datasets is time-consuming and expensive. In addition, there is no assurance that all b-values will be included.

Domain adaptation (DA) is a potential solution to address the domain shift issue, which attempts to alleviate the decrease of generalization ability caused by the distribution shift between source domain training data and target domain test data [37, 83, 84, 59]. However, such methods require human delineations from the target domain during the training process. Addressing the label availability challenge, unsupervised domain adaptation (UDA) has been introduced. While UDA methods are extensively employed across various medical image analysis tasks [17, 88, 25, 118], only a few studies have been conducted in the field of PCa detection [32]. Furthermore, most existing UDA researches are focused on single-domain mapping. When applied to multi-domain settings, these methods are required to train multiple generators for every domain pair, as well as their downstream task networks, after obtaining the generated labeled data. This approach is time-consuming and not feasible in practice (Fig. 13.1(b)).

The aim of this study is to assess whether harmonized ADC map and high b-value image improve the accuracy of predicting prostate lesion from multi-site data with various b-values. Designed for practical applications, we propose a UDA framework with a unified model for multi-domain mapping, as depicted in Fig. 13.1(c), which is computationally efficient, especially when encountering a large number of domains. Moreover, our method does not require annotations from the test data, and it can be applied to any pre-trained network for practical use without retraining. Specifically, our approach uses the ADC map and high b-value images derived from various b-value settings as inputs. It leverages meta-information to more effectively synthesize the ADC map and high b-value image at a consistent (and standard) b-value setting. Our method is evaluated using a large-scale dataset where various b-values were presented.

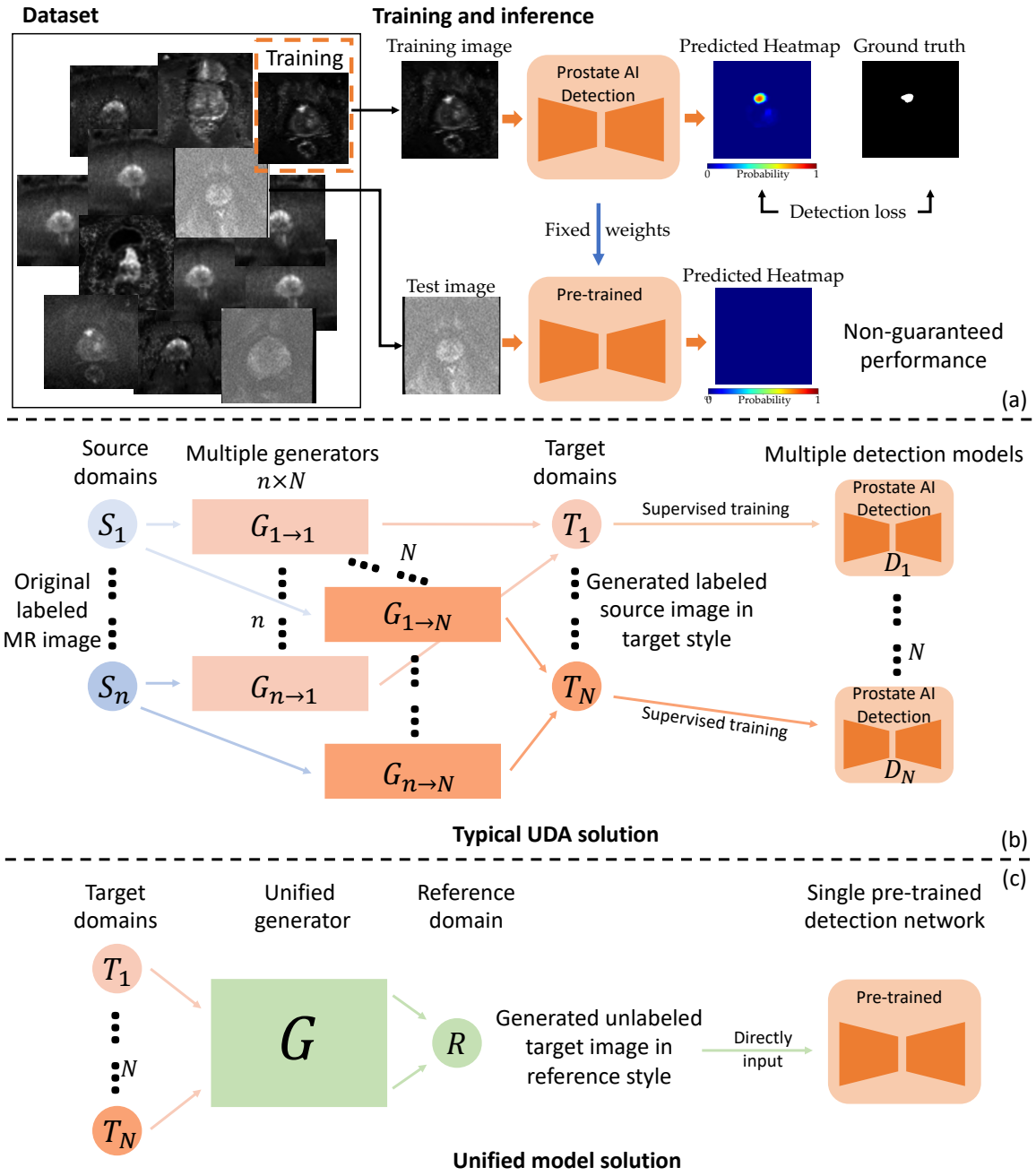


Figure 13.1: (a) shows the common domain shift problem for prostate cancer detection with supervised learning (SL) model. When MR images are collected using very different protocols the performance of the pre-trained model is not guaranteed. (b) shows the typical unsupervised domain adaptation (UDA) solution on domain shift, in which multiple generators and detection models are trained. (c) represents our solution, which only needs a single generator and pre-trained detection network. Specifically, the proposed method aims to translate image style from the target domain (unlabeled test data) to the reference domain (labeled data used to train the detection model) using a unified generator. S and T indicate source and target domains, respectively. R represents the reference domain. The training set is regarded as the source domain, with the reference domain being its subset. Best viewed in color.

The main contributions of our work are:

1. A novel unsupervised domain adaptation (UDA) method utilizes a unified generator to translate the image style from multiple target (test) domains into the reference (training) domain used for training the supervised learning (SL) detection model. The proposed method shows that UDA can reduce the dependency of the AI performance on the DWI scan protocol.

2. The dynamic filter is proposed and used to enable multi-domain mapping within a unified generator that accepts target images with arbitrary b-values as distinct domains, leveraging meta-information to differentiate between domains.

3. The comprehensive results of unseen 1,692 (2,393 samples) test cases from the area under the receiver operating characteristic curve (AUC) for PI-RADS$\geq$3 indicate that the proposed method improves the performance of the baseline SL method from 0.729 to 0.786 ($p < 0.05$). Notably, when the b-values of the test image represent out-of-distribution (OOD) samples from the training set, particularly for low-b=$\{150, 200\}$ and high-b=2000, increasing from 0.489 to 0.756 ($p < 0.05$).

13.2 Dataset and Annotation Process

13.2.1 Dataset

This retrospective analysis uses a multi-cohort dataset of 5,150 cases collected from nine different clinical sites, all of which have bp-MRI prostate examinations consisting of T2w acquisition and DWI of at least two different b-values. The details are presented in Tab. 13.1, and 2,170 of the 5,150 cases have been previously reported in [222]. The inclusion criteria for the study encompassed the following conditions: (1) patients who were treatment-naïve; (2) clear visibility of the prostate gland in the field of view (FOV) of bi-parametric MRI images; (3) acquisition of images using 1.5 T or 3 T axial MRI scans with either a body or endorectal coil. Conversely, the exclusion criteria included: (1) cases involving prostatectomy or any scenario where the prostate was partially resected; (2) cases with severe artifacts resulting from implants or motion. The prior article utilized only one pair of b-values for each case whereas in this study all qualified b-values were included to investigate their impact on the detection performance. The b-values ranging from 0 to 200 are considered as low b-values, while those between 600 and 2000 are considered as high b-values. The ADC images and DWI b=2000 images are computed based on each pair of low b-value and high b-value DWI images by performing a nonlinear least-squares fit to the equation $S(b) = S_0 \cdot \exp(-b \cdot \text{ADC})$. For each voxel, the coefficient of b was employed as its ADC value (with a scaling factor of 10^6), and the intensity of B-2000 image was calculated through extrapolation at b=2000. To maintain consistency and reduce variation in ADC computation, vendor-provided ADC maps were excluded from the study. In this way, each pair of b values

Table 13.1: Patient and imaging characteristics.

Characteristic	Training (n=3458)	Test (n=1692)
Age (y)[†]	65 (59-70)	66 (60-70)
Manufacturer		
GE	0	777 (43.8)
Siemens	3458 (100.0)	915 (56.2)
Field strength (T)		
1.5	21 (0.6)	24 (1.4)
3	3437 (99.4)	1668 (98.6)
Year of acquisition[††]	2014-2021	2014-2021
T2w		
TR (ms)[†]	5660 (4300-6500)	4960 (4000-5660)
TE (ms)[†]	101 (101-104)	104 (101-119)
In-plane spacing (mm)[†]	0.563 (0.469-0.625)	0.391 (0.391-0.5)
Slice thickness (mm)[†]	3.6 (3.6-4)	3.6 (3-4)
DWI		
TR (ms)[†]	4000 (3600-5100)	4025 (3003-4821)
TE (ms)[†]	73 (63-84)	63 (59-63)
In-plane spacing (mm)[†]	1.625 (0.877-2)	1.563 (0.938-2)
Slice thickness (mm)[†]	3.6 (3.6-4)	3.5 (3-4)
PI-RADS category		
1, 2	1389 (40.1)	804 (47.5)
3	470 (13.6)	176 (10.4)
4	894 (25.9)	419 (24.8)
5	705 (20.4)	293 (17.3)

Note.—Unless otherwise noted, data are numbers of patients, with percentages in parentheses. T2w = T2-weighted imaging, DWI = diffusion-weighted imaging, TR = repetition time, TE = echo time, PI-RADS = Prostate Imaging Reporting and Data System, PZ = peripheral zone, TZ = transition zone.

[†] Data are medians, with first quartile to third quartile ranges in parentheses.

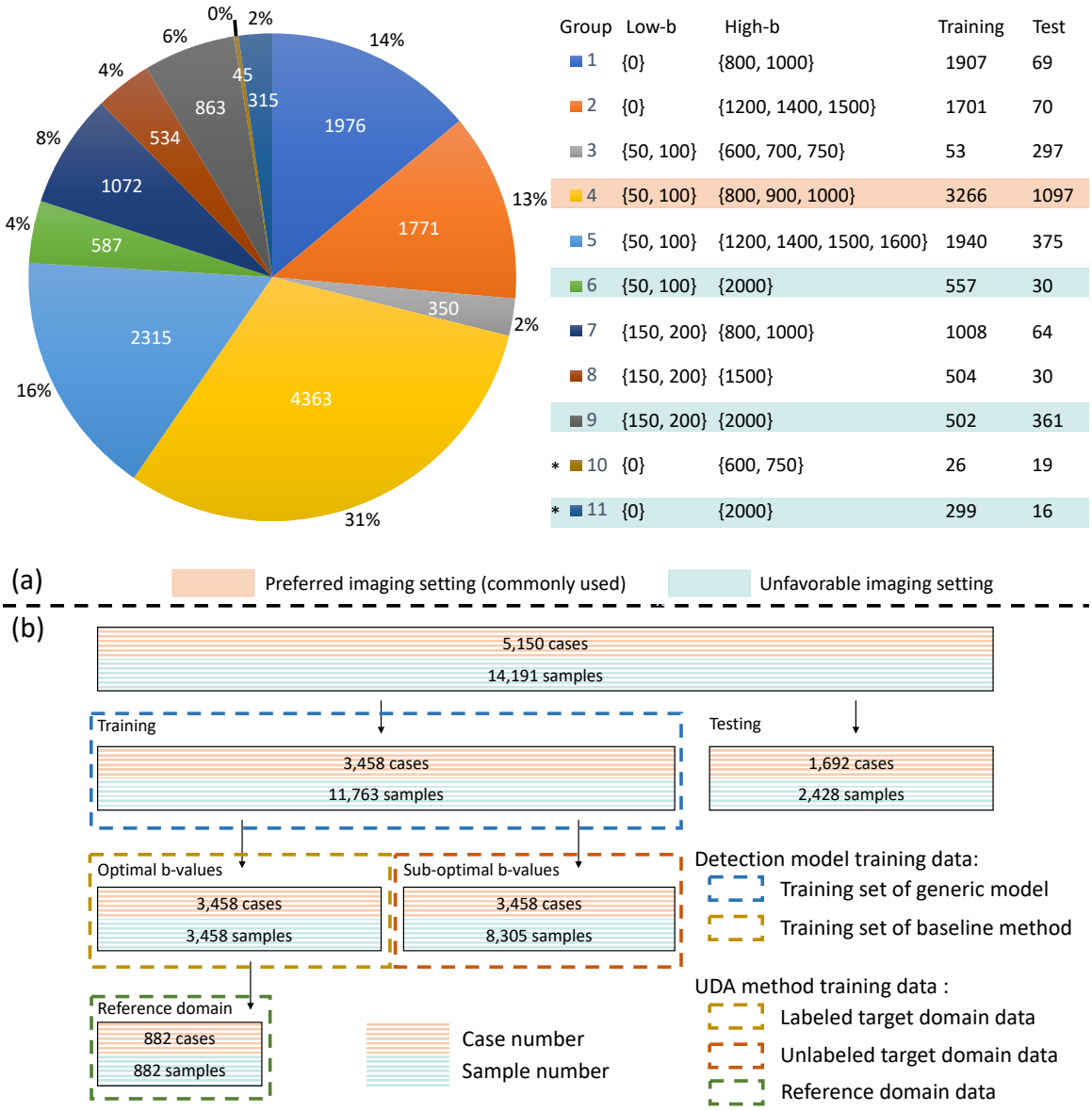
[††] Year of acquisition are based on data for which acquisition dates are available.

from the same case can be considered as a unique sample from a different domain. This yields a total of 14,191 samples of 34 different combinations of b-values from all cases. We categorized all samples into a few subgroups based on the range of b-values. The details of each subgroup can be viewed in Fig. 13.2(a).

A total of 3,458 cases were used for training, as shown in Fig. 13.2(b). For training the baseline method (i.e. the SL model from [236]), the best pair of b-values (optimal) was selected and only one single sample from each cases was utilized. The PI-RADS guideline recommends using one low b-value set at 0-100 sec/mm^2 (preferably 50-100 sec/mm^2) and one intermediate b-value set at 800-1000 sec/mm^2 for ADC computations [221]. We followed this suggestion and selected the b-values that are the closest to 50 and 1000 as low and high b-value respectively. For other methods (generic model and proposed UDA methods), additional samples with all possible b-value pairs were used, which consisted of 11,763 samples from the same training cases. In the UDA training process, 882 samples whose b-values are from the standard domain (low b-value=50, high b-value=800) were selected as reference domain data to train the unified generator of UDA methods, and the rest of the data are considered target domain samples. The independent testing set contains 1,692 cases with 2,428 samples. The results of 2,393 samples are reported in this work due to very limited sample number for some b-value subgroups, e.g. group 10 and 11 in Fig. 13.2(a). All the cases had lesion-based PI-RADS information and voxel-based annotations of the lesion boundaries. The PCa lesion annotations are obtained based on the clinical radiology reports and carefully reviewed by an expert radiologist. A positive case is identified if it contains PI-RADS ≥ 3 lesions.

13.2.2 Details of Annotation Process

The annotation process for the voxel-level prostate lesion segmentation and PI-RADS score was done as follows. Initially, we collated all clinical radiology reports, each accompanied by the respective PIRADS score for individual lesions, along with their corresponding lesion annotations. The raw clinical annotations exhibited variations, comprising landmarks indicating the lesion, bounding boxes around the lesion, or single contours on at least one slice to guide annotators regarding the location of the lesion of interest. In the subsequent step, annotators were tasked with delineating a complete 3D mask of the lesion based on the original annotations. All annotators underwent training provided by radiologists with Doctor of Medicine degrees and residency in Radiology. Supervised by radiologists, annotators received guidance throughout the process, addressing any uncertainties that arose. In the third step, 3D annotations from annotators underwent review and necessary corrections by radiologists within the annotation team. Radiologists possessed the authority to overrule annotators' annotations when deemed necessary. In the final step, all annotations and corresponding clinical reports underwent meticulous review by an expert radiologist with five years of experience in



Group	Low-b	High-b	Training	Test
1	{0}	{800, 1000}	1907	69
2	{0}	{1200, 1400, 1500}	1701	70
3	{50, 100}	{600, 700, 750}	53	297
4	{50, 100}	{800, 900, 1000}	3266	1097
5	{50, 100}	{1200, 1400, 1500, 1600}	1940	375
6	{50, 100}	{2000}	557	30
7	{150, 200}	{800, 1000}	1008	64
8	{150, 200}	{1500}	504	30
9	{150, 200}	{2000}	502	361
* 10	{0}	{600, 750}	26	19
* 11	{0}	{2000}	299	16

Figure 13.2: Overview of the dataset. The b-value distribution and data split are shown as (a) and (b), respectively. The left side (pie chart) of (a) shows the sample number and percentage of each group, whereas the details are shown on the right side. * denotes the group excluded from testing due to the limited sample number. In (b), the detailed training data for different methods are marked by different dashed rectangle boxes, best viewed in color.

radiology, specializing in prostate MRI examinations. As part of the annotation, we relied on the clinical assignment of PI-RADS as determined by the original clinical site. The primary objective of the annotation process was to harmonize and extend lesion contouring across slices, providing a 3D mask for each identified lesion. The expert radiologist made minimal changes to the original PI-RADS assignment ($< 1\%$) during case reviews. All annotations were performed using an internally customized tool where anonymized MRI

DICOM series were loaded. It's important to note that only lesions with PI-RADS scores of 3 and above were annotated in this study.

13.2.3 Image Preprocessing

We adopted the same preprocessing procedures as described in [139, 236]. Similar as the PI-RADS guideline that recommends a high b-value set at $\geq 1,400\text{sec}/\text{mm}^2$, in our study we re-computed a high b-value image at a fixed $b = 2,000\text{sec}/\text{mm}^2$ to ensure a representation in which lesions stand out. The fixed b-value was selected to further eliminate the b-value variances among datasets [181]. The preprocessing pipeline took raw bp-MRI acquisitions and generated well-formatted data volumes for all subsequent synthesis and detection models.

Initially, the original T2w and DWI series were extracted from the raw DICOM files. We employed a voxel-wise logarithmic extrapolation of the fitted signal decay curves to compute the ADC map and new DWI volumes at b-values of 0 and 2000 sec/mm^2 which are denoted as DWI b-0 and b-2000 images. The same process was repeated for each pair of low and high b-values if the case has more than two b-values. Next, whole prostate gland segmentation was performed on T2w volumes using the method presented in [230]. Subsequently, all DWI volumes were aligned to the T2w volumes using rigid registration. This alignment was meticulously verified through visual inspection. Additionally, volumes were center cropped and resampled to an image dimension of $240 \times 240 \times 30$ and a voxel spacing of $0.5 \times 0.5 \times 3\text{mm}^3$. Finally, we normalized all volumes to facilitate the training process. For T2w volumes, we linearly normalized to the range [0, 1] based on the 0.05 and 99.95 percentiles of their intensities. Given that ADC volumes represent quantitative parametric maps, they were normalized using a constant factor of 3000. For computed DWI b-2000 volumes, we first normalized them by a factor obtained as median intensity within the prostate gland region of the corresponding DWI b-0 volumes, and then normalized by a constant value to linearly map the approximate range to [0, 1].

13.2.4 Algorithm Design

Fig. 13.3 shows the proposed framework which aims to solve two common practical issues in PCa detection, i.e. domain shift and label availability for test data. The proposed framework contains two parts: synthesis and detection. To increase the generalizability of the SL detection network for OOD test samples, generators align the style of DWI B-2000 and ADC test samples from the target domain to the reference domain at the image level. Next, the detection model predicts the PCa heatmap which uses the concatenation of T2w, generated ADC, generated DWI B-2000, and prostate mask as inputs. Notably, this entire process operates without the need for test data labels. To more accurately mimic real-world scenarios, we initially trained the

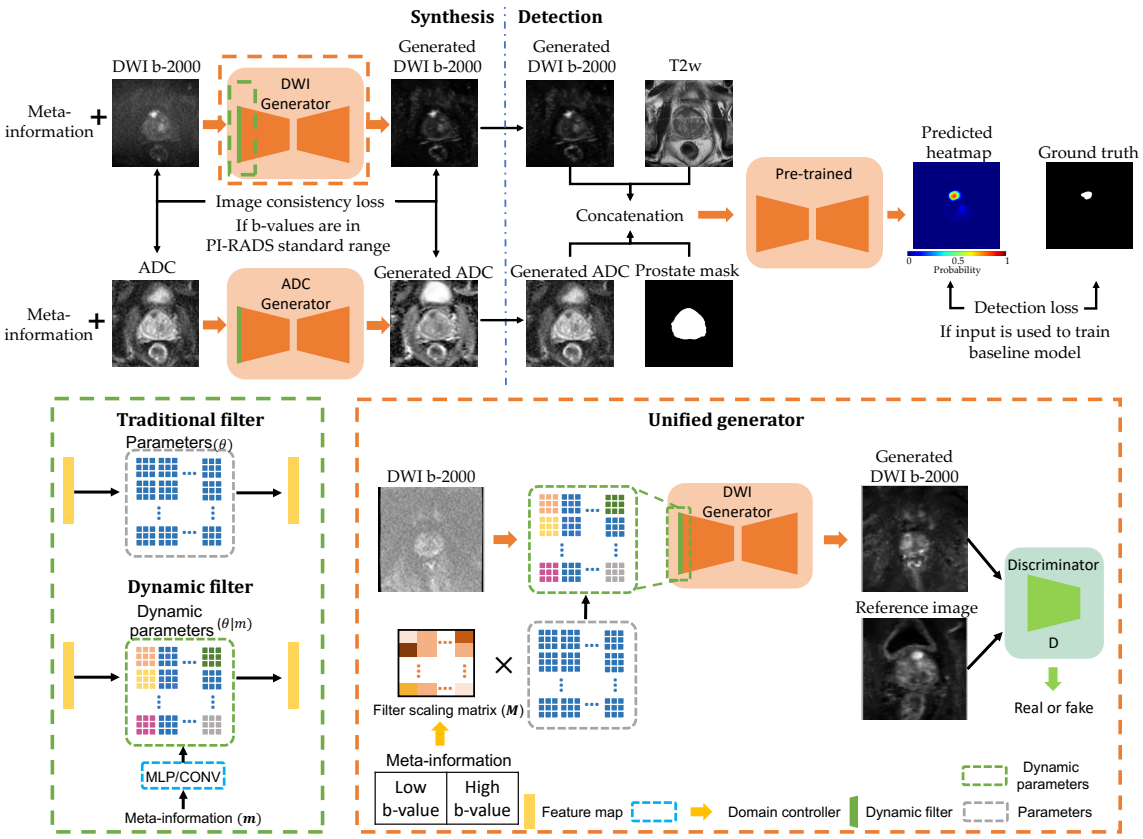


Figure 13.3: Proposed UDA framework for PCa detection, which contains two parts: synthesis (top left) and detection (top right). Specifically, the SL model from baseline method [237] is first trained to provide supervision to generator. In addition, both generators (orange) are identical but with different weights. The dynamic filter (green) is plugged into the generator for multi-domain mapping and takes meta-information as input to specify domain information. The bottom left shows the differences between traditional and dynamic filters, and the bottom right provides details of the proposed unified generator.

detection model and then utilized the trained model to educate the generators.

13.2.5 Detection Network

Our 2D detection network is a U-Net embedded with residual blocks. This network utilizes information from the 2D slice of T2w, ADC, DWI B-2000, and prostate mask to generate the corresponding PCa lesion heatmap. Further details of the architecture can be referenced in [236]. In line with the configurations of [236], a 3D heatmap is derived from all slices for each sample, with the non-zero regions considered as lesion candidates. To identify the true positives (TPs) and false positives (FPs), a threshold is used to get a set of connected components. The TPs can be identified if the connected components overlap on annotations or are less than 5mm away from the lesion center. Otherwise, such connected components are classified as FPs. Any PCa lesions that lack corresponding detections are termed false negatives (FNs).

13.2.6 Synthesis Network and Dynamic Filter

The synthesis models (generators) are adapted from CUT (Contrastive learning for Unpaired image-to-image Translation) [166] which is a U-shaped network. While the architecture for the DWI B-2000 and ADC generators is identical, they have different parameters, i.e., separate networks are trained for the DWI B-2000 and ADC images. Each generator takes a 2D image from the target domain as input and produces a 2D image styled in the reference domain, as illustrated in Fig. 13.3. The discriminator evaluates the performance of the generator during training by distinguishing between generated and real reference domain images. When both the input image and its label were used to train the baseline method, the detection loss is applied to offer supplementary guidance to the generator. To emphasize correct mapping to the reference domain, we employ an additional consistency loss ($L_{\text{consistency}}$) at the image level to preserve the original information from the input image, especially if its b-values align with the PI-RADS standard range. Specifically, the mean square error serves as the metric.

Unlike typical UDA methods, which require multiple networks for multi-domain mapping, we use only a unified model for each modality. However, the performance of the unified model may be limited in producing robust results between multiple domains due to a lack of domain information. To address this issue, we propose a dynamic filter as a domain indicator, which aims to increase domain generalizability of the unified model by leveraging meta-information. In Fig. 13.3, the traditional convolutional layer processes feature map with parameters θ which are learned during training. In contrast, the parameters of the dynamic filter are dynamically generated based on various different conditions by a scaling layer. Thus, the unified generator could achieve robust multi-domain mapping with provided meta-information. Contrary to other works use one-hot encoding for the dynamic filter [231, 134, 252], which can be complex for large-scale studies with multiple domains. In contrast, we convert low and high b-values from meta-information into a 2D tensor, which simplifies the input for the domain controller and preserves original information. As suggested in [134, 2], a filter scaling strategy generates a kernel-wise scale factor, uniformly weighting all parameters instead of individually scaling them. Specifically, the domain controller learns to generate the corresponding filter scaling matrix (M) based on the provided meta-information. Each element in M represents the scale of the corresponding kernel, and the parameters of the dynamic filter are dynamically adjusted through scalar multiplication.

13.2.7 Implementation Details

The training process of the detection model (baseline) is followed in [236], and we used binary cross-entropy as detection loss (L_{det}). To train the generator, we set the batch size to 96 and used the same loss functions as [166] which are denoted as L_{CUT}. The total epoch is set to 100. The domain controller is a simple

Algorithm 1: Training process of DWI generator$\rightarrow G(\theta)$

Input: DWI b-2000 $\rightarrow x$ and its meta-information $\rightarrow m = (low-b, high-b)$

Load pretrained detection network $\rightarrow D$ and its training set $\rightarrow \mathscr{D}_{train}$;
while $epoch \leq 100$ **do**
  if $x \in \mathscr{D}_{train}$ **then**
    $L_{syn} = L_{det} + L_{CUT}$;
    if $50 \leq low-b \leq 100$ *and* $800 \leq high-b \leq 1000$ **then**
      $L_{syn} = L_{det} + L_{CUT} + L_{consistency}$;
    end
  else
    $L_{syn} = L_{CUT}$;
  end
  Update θ;
end

convolutional layer with 7 as kernel size, an input channel is 2, and an output channel is 128, where 64 scaling factors and bias weights are included. The hyperparameters of the generators are the same as [166]. In addition, the loss selections of generators (L_{syn}) depend on three scenarios related to the input image: (1) $L_{\text{syn}} = L_{\text{CUT}}$ for unlabeled target domain data; (2) $L_{\text{syn}} = L_{\text{det}} + L_{\text{CUT}}$ for labeled target domain data; and (3) $L_{\text{syn}} = L_{\text{det}} + L_{\text{CUT}} + L_{\text{consistency}}$ for reference domain data. The detailed training process of generator can be viewed in Algorithm 1. The training was conducted on NVIDIA A100 GPUs and implemented using PyTorch.

13.2.8 Model Comparisons

We compared our proposed framework with two deep learning methods for multi-site PCa lesion detection, which are (1) baseline: a SL pre-trained detection model [236]; and (2) generic model: retrain the baseline method using a larger dataset with various b-values. We also reported the results of the ablation study to show the effectiveness of the proposed method in 13.3.5.

13.2.9 Statistical Analysis

The area under the receiver operating characteristic curve (AUC) score is computed as case-level performance, which is the primary endpoint of this work. The maximum value of the 3D heatmap is defined as the prediction score of the sample to calculate the AUC score. The confidence interval was computed based on a bootstrap approach with 2,000 resamples. We set a statistical significance threshold of 0.05. In addition, the free-response receiver operating characteristic curve (FROC) is used as a metric to evaluate the lesion-level performance as supplementary results. Moreover, peak signal-to-noise ratio (PSNR), mean square error (MSE), and structural similarity index measure (SSIM) are used as metrics to evaluate the image quality of

generated images. We used t-SNE visualization to assess the impact of the generated ADC and DWI B-2000 images on the detection network. Specifically, we randomly selected 100 samples from the unseen test set with a low b-value of 200 and a high b-value of 2000. These samples served as input for our proposed framework. For the t-SNE visualization, we extracted the feature maps of these selected cases from the bottleneck feature map of the baseline method.

Table 13.2: Case-level AUC score on unseen test sets. Bold and underline formattings indicate the best AUC scores for PIRADS≥3 and PIRADS≥4, respectively. Baseline: a SL pre-trained detection model [236]. Generic model: retrain the baseline method using dataset with all possible b-value combinations.

	b-values		Case-level AUC (PIRADS≥3 / PIRADS≥4)		
Group	Low	High	Baseline	Generic	Proposed
1	0	800, 1000	0.840 / 0.851	0.802 / 0.811	**0.865** / 0.878
2	0	1200, 1400, 1500	0.740 / 0.827	**0.872** / 0.884	0.856 / 0.918
3	50, 100	600, 700, 750	**0.712** / 0.739	0.659 / 0.677	0.660 / 0.674
4	50, 100	800, 900, 1000	**0.831** / 0.836	0.777 / 0.796	0.825 / 0.831
5	50, 100	1200, 1400, 1500, 1600	0.682 / 0.722	0.692 / 0.715	**0.730** / 0.764
6	50, 100	2000	0.708 / 0.795	0.769 / 0.775	**1.000** / 0.995
7	150, 200	800, 1000	0.897 / 0.951	0.802 / 0.846	**0.948** / 0.972
8	150, 200	1500	0.880 / 0.920	0.829 / 0.860	**0.958** / 1.000
9	150, 200	2000	0.489 / 0.502	0.612 / 0.653	**0.756** / 0.765
Total			0.729 / 0.767	0.733 / 0.756	**0.786** / 0.798

Table 13.3: The overall AUC performances (PI-RADS $\geq$ 3) across various zones are presented as AUC [lower 95% CI, higher 95% CI], with bold text indicating the best performance. n denotes the sample number, and for negative samples, $n = 1212$.

	Methods		
Zone	Baseline	Generic	Proposed method
PZ ($n = 555$)	0.694 [0.668, 0.718]	0.670 [0.643, 0.696]	**0.732 [0.706, 0.757]***
TZ ($n = 349$)	0.793 [0.766, 0.819]	0.798 [0.772, 0.824]	**0.819 [0.791, 0.848]***
Both ($n = 277$)	0.759 [0.729, 0.789]	0.733 [0.700, 0.767]	**0.802 [0.773, 0.832]***

13.3 Results

13.3.1 Case-level Performance

Tab. 13.2 displays the case-level performance. For both PI-RADS≥3 and PI-RADS≥4 labels, the overall performance of the proposed model significantly improves ($p < 0.05$) from other models. The baseline method is effective when test data b-values closely match the training set, primarily in the reference domain (group 4), but struggles with OOD samples, as seen in group 9. Even with diverse b-values added to the training data, the generic model fails to differentiate across domains, resulting in unstable performance, especially in the commonly used reference domain. Our UDA method improves the results of the baseline

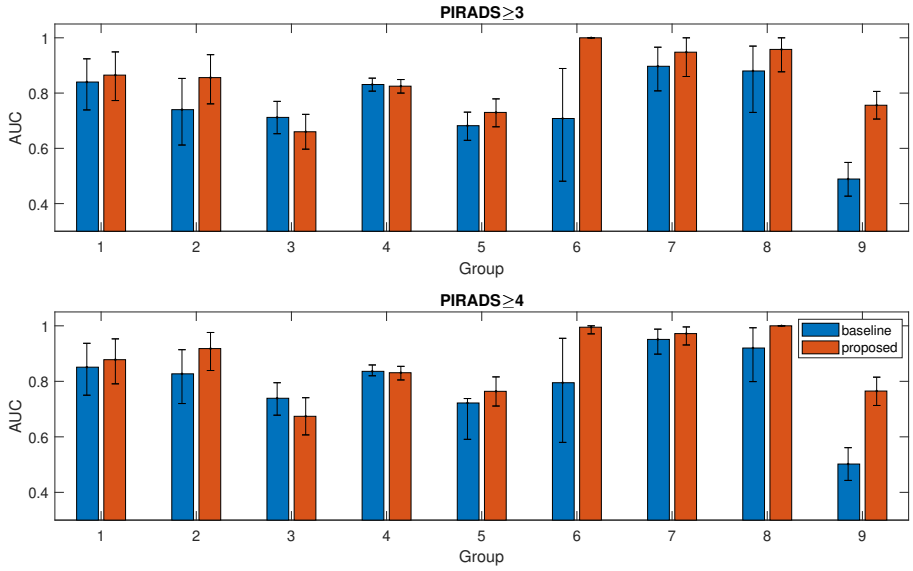


Figure 13.4: Comparison of case-level AUC between baseline and proposed methods. Error bars indicate the 95% confidence interval using a bootstrapping approach. The AUC was calculated based on bootstrap resampling of test samples for each group and repeated for 2000 times. The 2.5th and 97.5th percentile of the bootstrapped AUC distribution was used for an estimation of the 95% confidence interval.

method in all groups when b-values deviate from the standard, and maintains consistent performance for the reference domain. In addition, the overall AUC performance in different zones can be viewed in Tab. 13.3.

We present the overall AUC performance across all b-value groups for cases with lesions in different zones (Fig. 13.4). In this analysis, cases were categorized into a specific zone if 80% of their lesions were within that area; otherwise, they were labeled as "both." Given the absence of prostate gland annotations, we generated zone masks using the gland segmentation model outlined in Section S.1. The observed improvement is statistically significant ($p < 0.05$) for all zone categories when com-paring our proposed method with other approaches. Notably, cases with lesions in the peripheral zone (PZ or "both") exhibited a higher margin of performance improvement compared to cases with transitional zone lesions only. This finding is in accordance with the PI-RADS guideline, high-lighting the influential role of DWI images in determining the PI-RADS assessment for lesions in the peripheral zone.

13.3.2 Lesion-level Performance

FROC representing the FP to TP ratio is detailed in Tab. 13.4. The proposed method outperforms the baseline in most domains except for the false positive rate in group 3 and the true positive rate in group 7. In Fig. 13.5, we visualize four typical domains with the highest testing samples. For standard b-values, the baseline method excels in AUC scores. Yet, the proposed method demonstrates superior performance, particularly when b-values diverge from the standard.

Table 13.4: FROC results of baseline and proposed methods. TPR and FPp denote true positive rate and false positive per case, respectively. Bold formatting indicates better performance. In details, the true positive rate (TPR=0.75 and 1) and false positives per patient (FPp=0.65 and 0.7) are used for the evaluation.

	Methods (Baseline / Proposed method)			
Group	TPR@FPp=0.75 ↑	TPR@FPp=1 ↑	FPp@TPR=0.65 ↓	FPp@TPR=0.70 ↓
1	0.768 / **0.778**	0.803 / **0.825**	0.495 / **0.287**	0.610 / **0.320**
2	0.613 / **0.730**	0.726 / **0.777**	0.829 / **0.551**	0.842 / **0.590**
3	0.413 / **0.452**	0.471 / **0.520**	**2.316** / 2.482	**3.033** / 6.090
4	0.696 / **0.711**	0.734 / **0.746**	0.669 / **0.635**	0.769 / **0.668**
5	0.466 / **0.544**	0.500 / **0.592**	2.208 / **1.354**	9.891 / **1.683**
6	0.674 / **0.875**	0.778 / **1.000**	0.673 / **0.133**	1.683 / **0.833**
7	**0.962** / 0.830	**0.962** / 0.901	0.353 / **0.092**	0.406 / **0.112**
8	0.708 / **0.986**	0.833 / **1.000**	0.327 / **0.100**	7.667 / **0.180**
9	0.200 / **0.568**	0.231 / **0.599**	5.194 / **1.493**	5.194 / **2.153**
Total	0.452 / **0.637**	0.526 / **0.684**	1.642 / **0.778**	2.314 / **1.109**

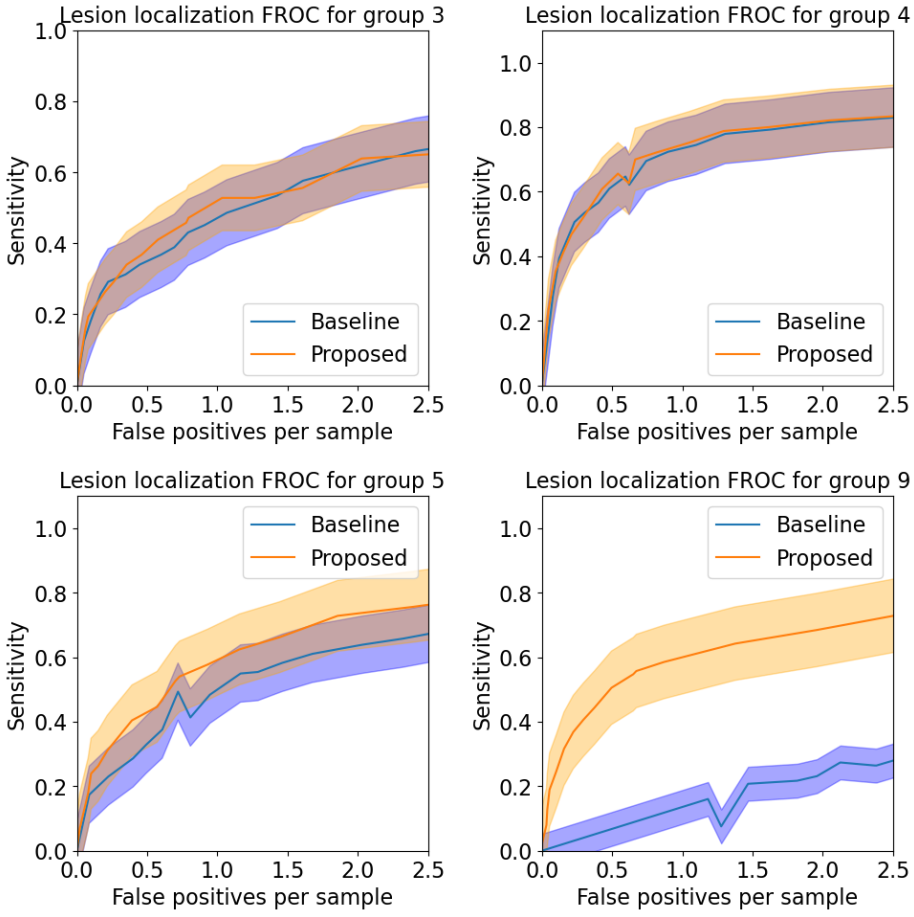


Figure 13.5: FROC curves of comparative methods. Blue and orange represent the baseline and proposed methods, respectively. The 95% confidence intervals for each are shaded. The selected groups are common actual b-value imaging settings and fall into three categories: near, within, and distant from the standard b-value range.

Table 13.5: Image similarity comparison of computed DWI b-2000 images, presented as PSNR / MSE ($\times 10^{-}3$) / SSIM. The results are calculated by comparing original and generated images with their corresponding images in the reference domain.

b-values		DWI b-2000	
Low	High	Original	Generated
150	1500	27.65 / 1.85 / 0.849	32.89 / 0.74 / 0.899
200	2000	20.21 / 10.1 / 0.625	31.21 / 1.11 / 0.840

13.3.3 Quality of Generated Image

We demonstrate the superiority of our proposed framework in terms of the image quality of generated DWI B-2000 images by comparing them with paired DWI B-2000 images in the reference domain. This requires the same cases to have DWI images in both reference domain and other target domains. To achieve this, we selected all cases from the testing dataset that had six different b-values. The DWI B-2000 images were computed using naturally acquired DWI images of three different b-value pairs: (50, 800), (150, 1500) and (200, 2000). The proposed method was applied to the B-2000 images computed using b-values of (150, 1500) and (200, 2000) to generate new B-2000 images. The original and generated B-2000 images were compared with the corresponding one computed by using (50, 800) b-values. Three different metrics, peak signal-to-noise ratio (PSNR), mean-square error (MSE) and structural similarity index measure (SSIM), which are commonly used as quality measurement between the original and a compressed image, were adopted here to assess the similarity to the reference domain images. The results are provided in Tab. 13.5. In the analysis of both b-value pairs, the generated images have higher PSNR, lower MSE and higher SSIM. Our findings reveal that the DWI B-2000 images generated by our method are more similar to the reference domain images than the original target domain images.

Fig. 13.6 displays original bp-MRI images, generated DWI images and detection heatmaps of 4 example cases (2 positive and 2 negative). In each group one case is from the reference domain and one case is from the target domain with b-values far from PI-RADS guideline recommendation. For cases in the reference domain, no obvious changes can be observed in the generated images and the predictions are similar by using original or generated images. For cases not in the reference domain, we can observe a significant improvement of the image quality. The predictions are also more accurate comparing with the ground truth annotations.

13.3.4 t-SNE Visualization

To better visualize the relationship between reference domain images, original target domain images and generated images, we applied t-SNE algorithm to reduce the bottleneck layer features of the PCa detection

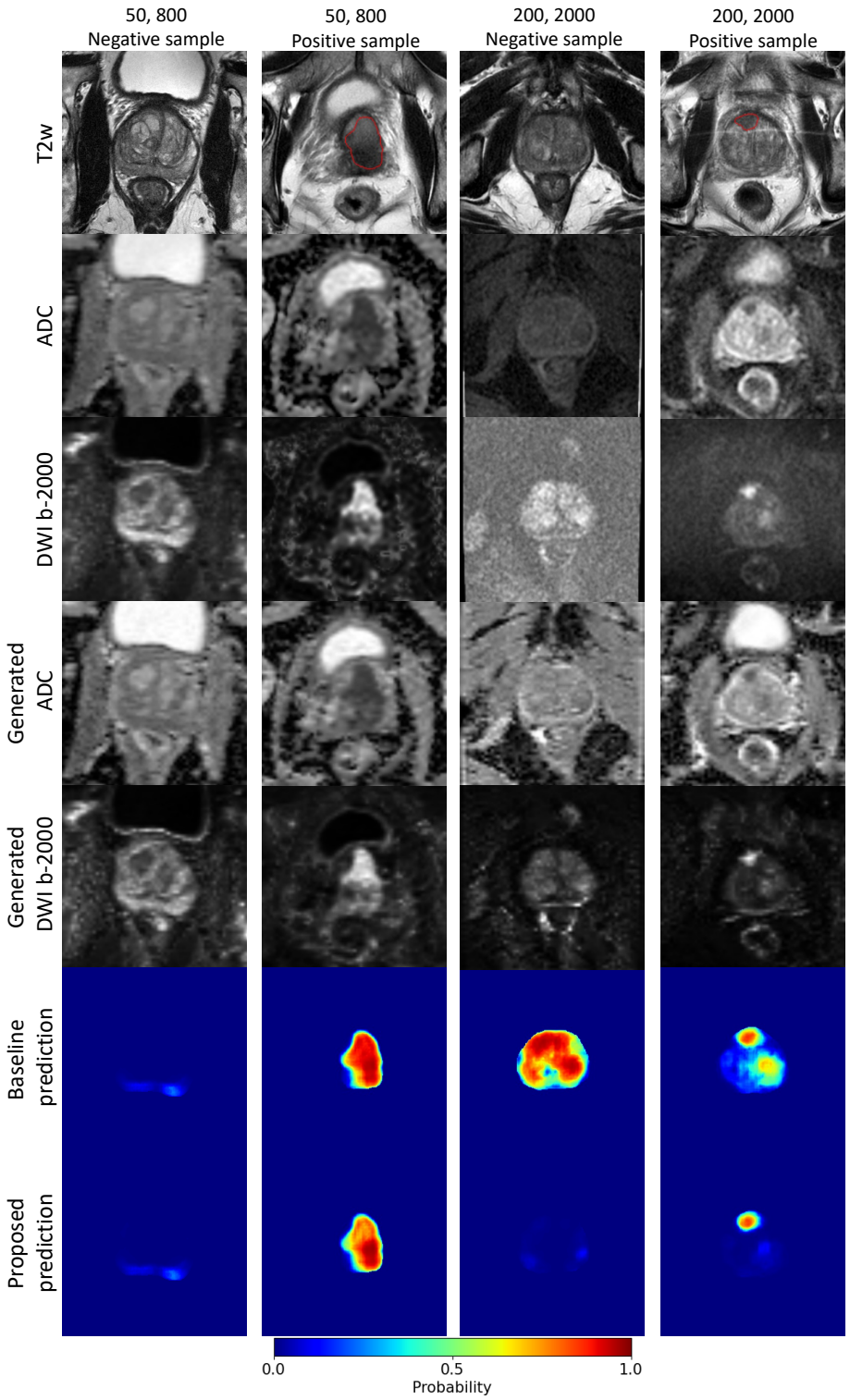


Figure 13.6: Qualitative results from four example samples are presented based on their ground truth labels and b-values, which are indicated at the top of each column and represented as “low b-value, high b-value”. The type of image for each sample is labeled on the left side of the figure. Red contours outline the ground truth lesions.

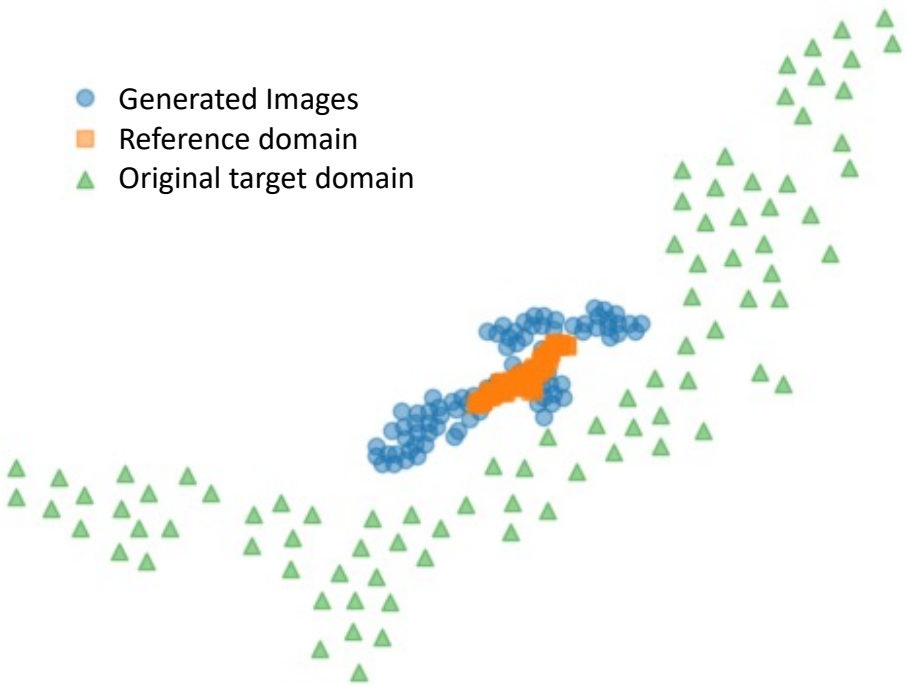


Figure 13.7: t-SNE visualization based on the feature map of pre-trained detection network at bottleneck layer. Orange squares represent inputs from reference domain images, green triangles denote inputs using original target domain images, and blue dots correspond to inputs of images generated by the proposed method.

network to a 2-dimensional representation. For neat visualization, 100 cases whose original DWI acquisitions are low b-value=200 and high b-value=2000 were selected for this analysis. The original and generated DWI images were used as input of the detection network. Another 100 cases from the PROSTATEx Challenge dataset [132] were selected as the reference domain (i.e. low b-value=50 and high b-value=800) cases for comparison. The 2-dimensional plot of the bottleneck layer features are displayed in Fig. 13.7. The generated DWI images using the proposed method form a tighter cluster compared to the original target domain image. Moreover, the generated images align more closely with the reference domain data, which indicates a higher similarity in the latent space of the detection network.

13.3.5 Ablation Study

The compared methods are: (1) CUT: our solution (see Fig. 13.1(c)) to domain shift that utilizes CUT to achieve unpaired image-to-image (I2I) translation; (2) CUTe: Form CUT into an end-to-end workflow; (3) CUTd: CUT with dynamic filter; (4) CUTd+e: Form CUTd into an end-to-end workflow; and (5) the proposed method: Additional consistency loss added on the CUTd+e approach. The details of the compared methods can be viewed in Fig. 13.8. For a fair comparison, the same network architecture is used for the detection model and generator, except for the additional controller for the methods with dynamic filters. All

UDA results are produced from the baseline method [236].

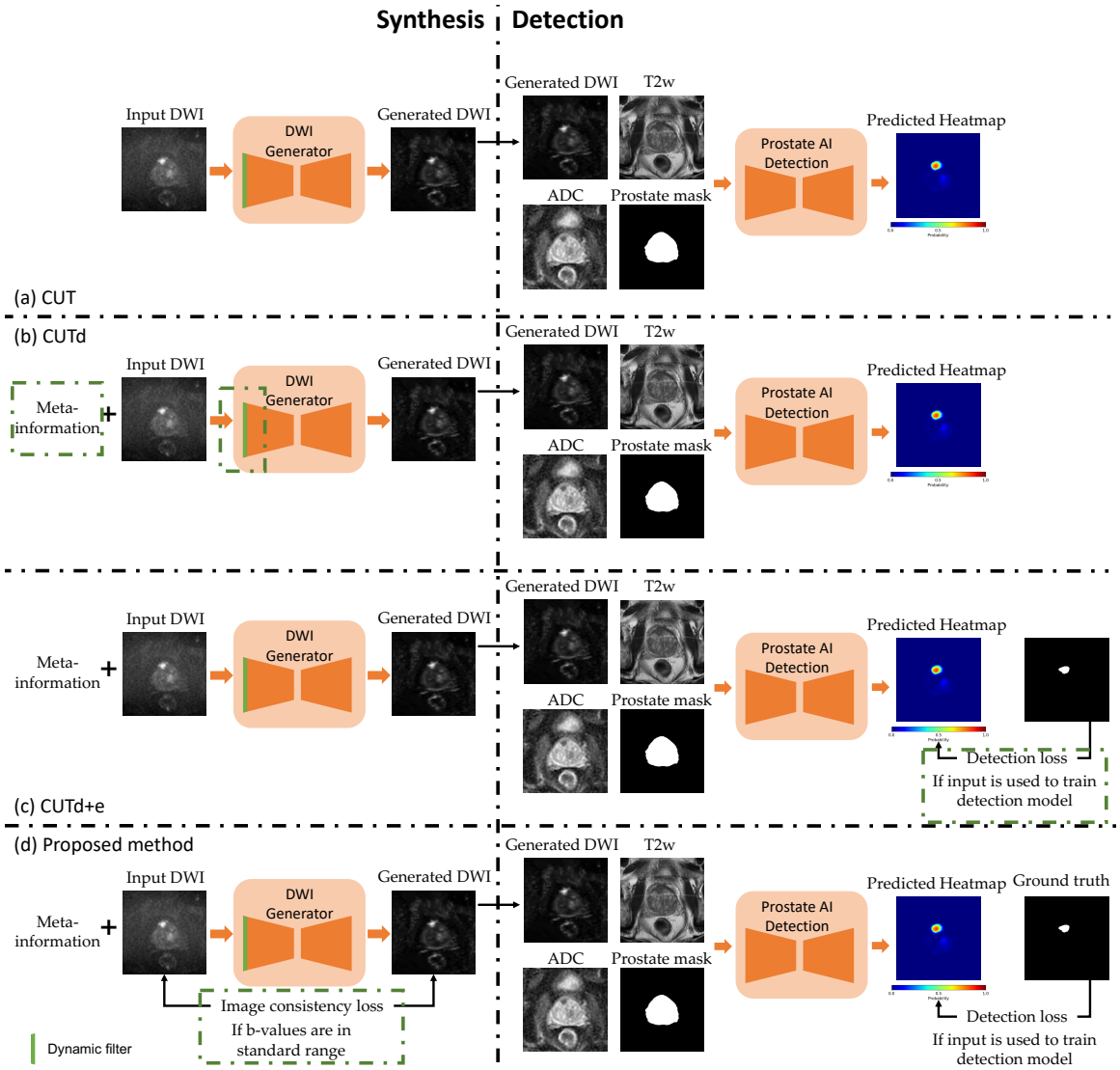


Figure 13.8: Illustrations of DWI generators among the compared methods are as follows: (a) CUT, (b) CUTd, (c) CUTd+e, and (d) proposed methods. The green boxes highlight the improvements made from the previous step. CUTe is excluded for brevity, and its improvements can be observed between (b) and (c).

Tab. 13.6 details the ablation study on the DWI generator, showing that UDA with I2I translation improves PCa detection for some groups. Performance in CUT and CUTe was unstable, particularly in group 4. By considering domain information with CUTd, improvements were observed in groups 1, 2, 6, 7, 8, and 9, highlighting the efficacy of the dynamic filter in varied domains. Building on CUTd, an end-to-end framework (CUTd+e) was tested. However, the I2I mapping of the generator might be influenced by the pre-trained detection model from various domains, potentially causing misalignment of generated images and leading to reduced accuracy. To address this, we introduced a consistency loss in our proposed method to retain details

of input samples with b-values from the standard PI-RADS range to counter biases from the pre-trained SL model. The qualitative results can be viewed in Fig. 13.9.

Table 13.6: Quantitative AUC score of ablation studies on DWI b-2000 images only. The **bold** and underline formattings indicate the best AUC scores among the UDA methods for PIRADS ≥ 3 and PIRADS ≥ 4, respectively. "e" notes the end-to-end training with supervision by the pre-trained detection network. "d" represents the usage of the dynamic filter.

		Ablation study of DWI generator (PIRADS≥3 / PIRADS≥4)				
Group	Baseline	CUT	CUTe	CUTd	CUTd+e	Proposed
1	0.840 / 0.851	0.804 / 0.833	0.760 / 0.779	0.833 / 0.849	0.820 / 0.834	**0.851** / 0.872
2	0.740 / 0.827	0.784 / 0.872	0.771 / 0.872	0.741 / 0.826	0.775 / 0.876	**0.834** / 0.923
3	0.659 / 0.739	0.670 / 0.697	0.643 / 0.680	0.587 / 0.593	**0.685** / 0.718	0.660 / 0.669
4	0.831 / 0.836	0.814 / 0.812	0.801 / 0.801	0.790 / 0.784	0.812 / 0.816	**0.828** / 0.834
5	0.682 / 0.722	0.686 / 0.731	0.674 / 0.713	0.635 / 0.650	0.684 / 0.728	**0.705** / 0.745
6	0.708 / 0.795	0.819 / 0.915	0.880 / 0.940	0.833 / 0.835	**0.903** / 0.980	0.847 / 0.795
7	0.897 / 0.951	0.896 / 0.992	0.885 / 0.863	0.917 / 0.895	0.887 / 0.931	**0.924** / 0.946
8	0.880 / 0.920	0.829 / 0.910	0.903 / 0.915	**0.940** / 0.935	0.833 / 0.920	0.912 / 0.960
9	0.489 / 0.502	0.670 / 0.697	0.633 / 0.666	0.680 / 0.697	0.673 / 0.706	**0.701** / 0.725

13.4 Discussion and Conclusion

In this paper, we proposed a novel UDA method with a unified model to solve practical common issues, domain shift and label availability, for PCa lesion detection. Unlike typical UDA methods, only a unified model is used in our framework for multi-domain mapping instead of multiple networks being trained. In order to achieve better performance of a unified model in multi-domain scenarios, we proposed and employed a dynamic filter to leverage domain information. When benchmarked against other methods using a large-scale, multi-site dataset comprising 5,150 cases (14,191 samples), our approach consistently demonstrates an enhanced capability to perform a more accurate PCa detection.

To demonstrate the feasibility for practical use, this study was conducted on a large-scale dataset with different imaging protocols, where the heterogeneous domain shifts are present and pose a challenge to achieve a consistent performance. The proposed method leverages information from the entire dataset, notably unlabeled data, to reduce the annotation effort, which is usually a burden for large datasets. Importantly, the proposed method can seamlessly be integrated into any pre-trained PCa detection framework and can be used as an image adapter at the upstream level to reduce discrepancies between domains. The method overall improves the generalizability of downstream PCa detection models. Importantly, there is no need to retrain or modify the network for new target data, making the method suitable for a variety of medical image applications.

No prior study has explored the domain shift in PCa detection using ADC and DWI high b-value images,

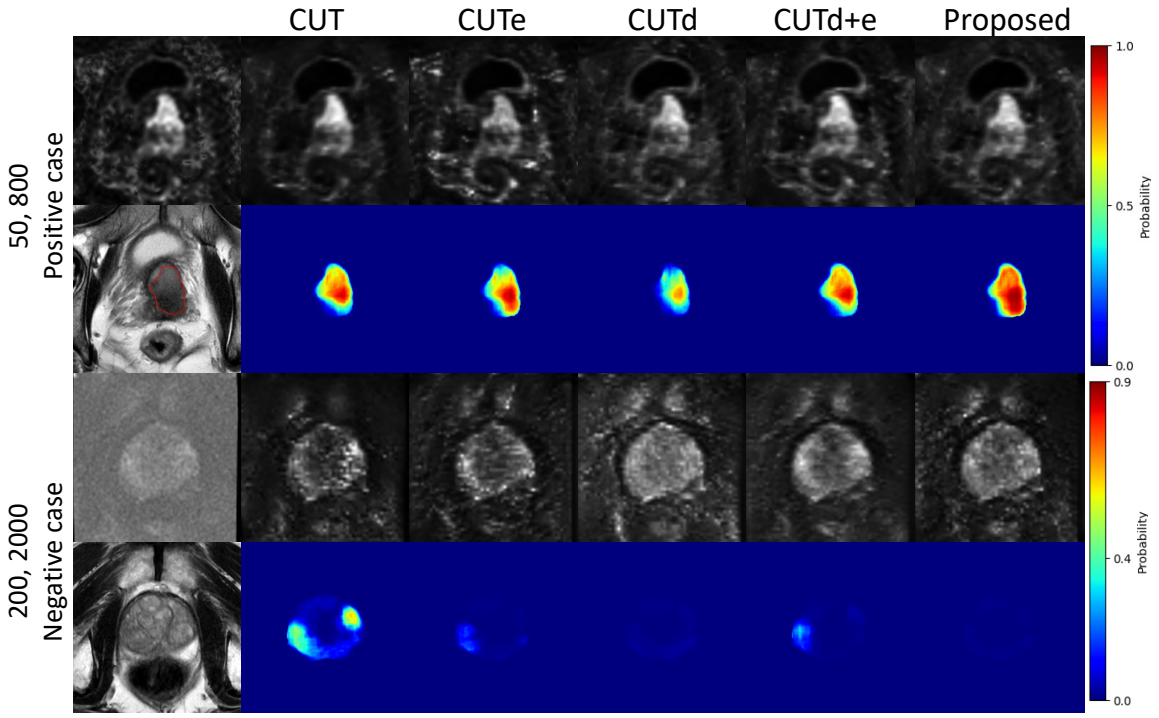


Figure 13.9: Qualitative results among compared UDA methods of ablation studies on DWI b-2000 images only, where high b-value images and the predictions are shown.

especially given the various domains in our study. We validated the common practical solutions as part of our experimental contributions. Although such methods are not the optimal solutions, important findings emerged: (1) using the original test image is preferable if its b-values closely align with training set, and (2) retraining a generic model could produce unpredictable results due to its broad adaptability and limitations in specific learning. These findings offer valuable insights for future studies, especially those targeting clinical applications.

Several existing studies have tried to address these prevalent practical challenges by producing more consistent DWI images. For instance, in [222], the authors suggested recalculating ADC maps and high b-value images at a fixed 2000 $\mathrm{sec}/\mathrm{mm}^2$ rather than utilizing the originally acquired images. However, this method cannot avoid the diffusion kurtosis effect if the acquired DWI uses high b-value over 1500 $\mathrm{sec}/\mathrm{mm}^2$. The proposed method with an I2I technique effectively translates OOD target samples into the style of reference domain. Resulting generated ADC and DWI b-2000 images are similar to the real reference domain image both at image-level and latent-level. The most pronounced improvements occur when the high b-value deviates farther from the standard range. Additionally, when high b-values are within the reference domain, our method boosts performance for low b-values that are OOD samples. Comprehensive detection results indicate that high b-values influence domain discrepancy more than low b-values.

We introduce the dynamic filter which can be treated as a domain indicator and plug it into any generator to leverage meta-information. The proposed dynamic filter generates conditional parameters according to the corresponding meta-information to differentiate domains. This is unlike encoding meta-information as one-hot vectors, which often requires a codebook to rigidly encode the corresponding relationship. An encoding approach like that might not be ideal for large-scale studies involving multiple domains. Additionally, the codebook would need adjustments whenever a new combination of b-values emerges. In contrast, we designed an effective yet straightforward strategy that directly uses b-values as input. This not only retains the original meta-information but also simplifies the process, making it adaptable to arbitrary b-value combinations.

In this study, our approach is limited by utilizing only b-values from meta-information to provide domain information. This constraint may contribute to suboptimal performance, especially when high b-values closely resemble those of the reference domain. It is crucial to acknowledge this limitation while also recognizing the potential expansion of our proposed domain adaptation method to include T2w images in future iterations. Subsequent efforts will involve incorporating additional meta-information, such as field strength, sequence selection, and the number of aver-ages, to more comprehensively quantify domain shifts resulting from diverse acquisition protocols.

Another contributing factor to occasional unsatisfactory performance is the restricted number of training samples available for specific b-value settings, as evident in the comparatively inferior performance of group 3, comprising only 53 training samples. Future endeavors will concentrate on manifold learning within the latent space of meta-information, offering a continuous represen-tation of b-value variations and enhancing generalizability to unseen b-value settings. Moreover, as our proposed UDA framework demonstrates its feasibility, subsequent work will involve developing additional image synthesis models based on our current approach. A comparative analysis of widely used synthesis models will be a crucial step forward. Additionally, recent developments in network architectures like nnU-Net and its variants have proven effective for PCa detection tasks [92, 13]. The detection network could be redesigned using similar architectures to better adapt to synthesized images, potentially enhancing detection accuracy.

As an early-stage study, we observed a significant improvement in overall performance with the proposed method, particularly when the high b-value deviates farther from the standard range. These improvements indicate that the proposed method may provide higher quality images and more accurate detection results, facilitating the interpretation time for radiologists, particularly for less experienced readers, and potentially increasing inter-reader agreement. However, the performance is constrained by the imbalanced cases of b-value distribution, affecting not only between training and test sets for certain group but also different groups in the training set, which limits the performance for certain groups. In practice, the test sample from such

group may directly fed into detection network. Future work will aim to improve the accuracy across diverse data distribution using cost-sensitive learning.

In conclusion, our unified model-based UDA method for multi-site PCa detection showed marked improvement on a large dataset, especially outside the reference domain. To the best of our knowledge, this is the first large-scale study exploring the impact of b-value properties on ADC and DWI b-2000 images with the aim of improving detection outcomes for a multi-domain scenario. It also benefits for future research, potentially inspiring further studies in this field.

CHAPTER 14

Self-supervised Test-time Adaptation for Medical Image Segmentation

The performance of convolutional neural networks (CNNs) often drop when they encounter a domain shift. Recently, unsupervised domain adaptation (UDA) and domain generalization (DG) techniques have been proposed to solve this problem. However, access to source domain data is required for UDA and DG approaches, which may not always be available in practice due to data privacy. In this paper, we propose a novel test-time adaptation framework for volumetric medical image segmentation without any source domain data for adaptation and target domain data for offline training. Specifically, our proposed framework only needs pre-trained CNNs in the source domain, and the target image itself. Our method aligns the target image on both image and latent feature levels to source domain during the test-time. There are three parts in our proposed framework: (1) multi-task segmentation network (Seg), (2) autoencorders (AEs) and (3) translation network (T). Seg and AEs are pre-trained with source domain data. At test-time, the weights of these pre-trained CNNs (decoders of Seg and AEs) are fixed, and T is trained to align the target image to source domain at image-level by the autoencoders which optimize the similarity between input and reconstructed output. The encoder of Seg is also updated to increase the domain generalizability of the model towards the source domain at the feature level with self-supervised tasks. We evaluate our method on healthy controls, adult Huntington's disease (HD) patients and pediatric Aicardi Goutières Syndrome (AGS) patients, with different scanners and MRI protocols. The results indicate that our proposed method improves the performance of CNNs in the presence of domain shift at test-time.

14.1 Introduction

Convolutional neural networks (CNNs) show excellent performance in supervised medical image segmentation tasks if the distribution of the training set (source domain) is tightly matched to the test set (target domain). However, for multi-site studies, domain shift is often present among different imaging sites due to different scanners and MRI protocols. In such scenarios, data from the target domain can be considered as out-of-distribution for the source domain, and the CNN performance can significantly drop during testing due to this domain shift.

Unsupervised domain adaptation (UDA) is a solution to minimize the gap between source and target domains. [27, 87, 45, 164]. However, the UDA normally requires data from both source and target domains to

This work is published at MLCN 2022.
Li, Hao, et al. "Self-supervised test-time adaptation for medical image segmentation." International Workshop on Machine Learning in Clinical Neuroimaging. Cham: Springer Nature Switzerland, 2022.

train. Moreover, for multiple target domains, UDA needs to train a separate model for each target domain, which is time-consuming. Another solution is domain generalization (DG), which tries to increase the model generalizability to unseen target domain data [44, 244, 11]. DG might need large amounts of source domain data or augmented data for training, and it may not adequately represent the data in the unseen target domains to produce robust segmentations. Furthermore, source domain data could be unavailable to researchers/clinicians between sites due to privacy issues. In contrast, a pre-trained model from the source domain is often easier to obtain, but domain shifts could lead to unreliable segmentations when such pre-trained models are directly applied on unseen target domain data.

To produce robust results with access to only the pre-trained models from the source domain and unseen test data, test-time adaptation (TTA) could reduce the effects of domain shift by adapting the target data to source data at either the image level or the latent feature level. Wang et al. [212] proposed an image-specific fine tuning pipeline in the testing phase for interactive segmentation by adapting the pre-trained CNN to the unseen target data, and the priors on the predicted segmentations were used for adaptation. Sun et al. proposed a test-time training approach for improving the model performance when domain shift is present between training and test data [200]. They adapt part of the model using a self-supervised rotation task on target data. Furthermore, Wang et al. proposed test-time entropy minimization for adaptation [210]. He et al. proposed a TTA network which is based on autoencoders trained on source domain [72]. During inference, the adaptation is applied on each target data by minimizing the reconstruction loss of autoencoders with fixed weights. Similarly, Karani et al. proposed an adaptable network for TTA [103]. In their work, the weights of the pre-trained segmentation network and the denoising autoencoder are fixed while updating the parameters of the normalized network to achieve adaptation during test-time. However, most TTA methods adapt target data either in image-space or fully/partially in feature-space, and may not have the ability to deal with images with bigger domain shifts, such as anatomical content shifts in addition to image intensity or contrast shifts. In addition, user interaction is needed in [212], which is problematic for large studies. Only feature-level adaptation [200, 210] may fail on some cases without image-level adaptation. Finally, a good alignment between target and source domains may not be possible when only partially features are adapted during test-time [72] or without feature-level adaptation [103].

In this work, inspired by previous works [200, 72], we propose a test-time adaptation framework for volumetric medical image segmentation, by adapting the target image at both image and feature levels. Our network has three components: (1) a multi-task segmentation network (Seg) with segmentation and reconstruction tasks, (2) autoencoders (AEs) optimizing the similarity between their input and output, and (3) an image translation network (T) to translate the image from target domain to source domain. The Seg and AEs are trained offline on labeled source data. At test-time, these pre-trained CNNs are fixed, and only the T and

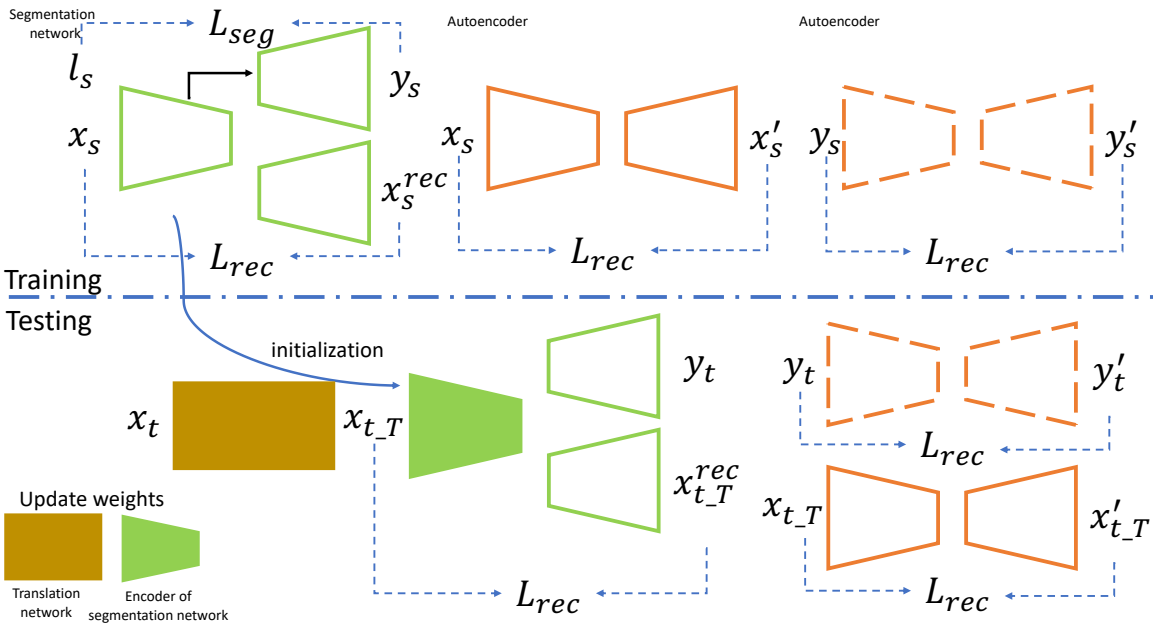


Figure 14.1: Proposed test-time adaptation framework. During offline training (top), the multi-task segmentation network (green) is supervised by segmentation and reconstruction losses, and autoencoders (orange) are trained for measuring the similarities. During test-time (bottom), the target image first goes through the translation network (brown), then fed to the pre-trained segmentation network. Only the translation network and the encoder of segmentation network (painted) are updating parameters during testing. $\mathscr{L}_{seg}$ and $\mathscr{L}_{rec}$ denote the Dice loss and MSE loss, respectively.

the encoder of Seg update weights with target data to achieve test-time adaptation by minimizing the reconstruction losses from self-supervised tasks. We evaluate our method with healthy adults, adult Huntington's disease patients [143] and pediatric Aicardi Goutières Syndrome (AGS) patients [207] for the brain extraction task. The data thus includes different brain sizes, shapes and tissue contrast, and ranges from healthy to severely atrophied anatomy.

14.2 Methods

Fig. 14.1 shows our proposed test-time adaptation framework, which consists of three parts: a multi-task segmentation network (Seg), autoencoders (AEs) and a translation network (T). In the offline training phase (top row of Fig. 14.1), the Seg (green) is trained in a supervised manner with a dataset from source domain $D_s = \{x_s, l_s\}_{s=1}^{N}$ consisting of input MRIs x_s and corresponding labels l_s. In addition, similar to [72], two AEs (orange) are trained after fixing the weights of Seg to measure the similarities between inputs and reconstructed outputs. At test-time (bottom row), the weights of the AEs and the Seg decoder are fixed. For a given image x_t from target domain, the T (brown) is trained to translate x_t to source domain as image x_{t_T}. Then the translated image x_{t_T} is fed to Seg to obtain the segmentation mask y_t and the reconstructed image $x_{t\_T}^{rec}$. The

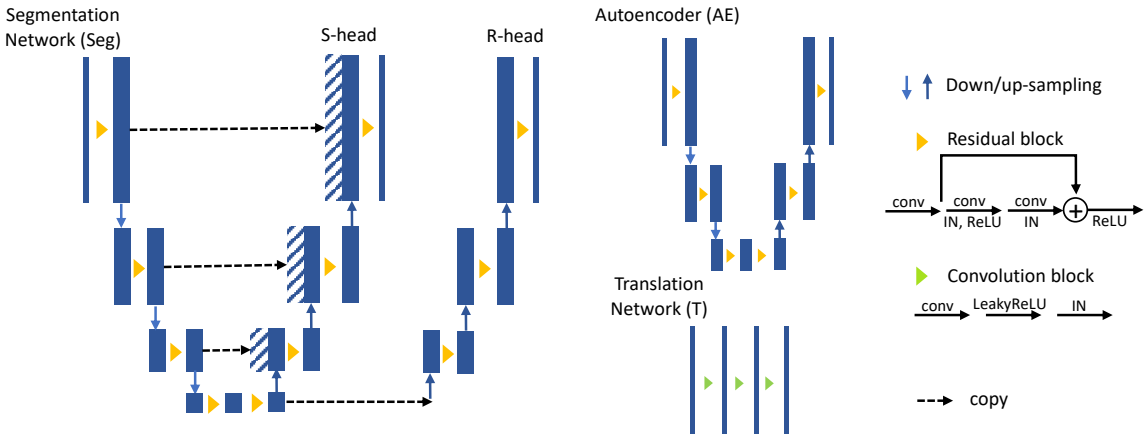


Figure 14.2: Network architecture of segmentation network (Seg), Autoencoder (AE) and translation network (T). The S-head and R-head of Seg are for segmentation and reconstruction task, respectively.

T and the encoder of Seg are optimized with self-supervised learning, which is the key step in our proposed TTA framework during inference; this is achieved via self-supervised tasks for the reconstruction path of Seg and the AEs. For image-level alignment, the pre-trained AEs control the quality of the translated image x_{t_T}. In other words, the AE loss ($\mathscr{L}_{AE}$) indicates the gap between x_{t_T} and source domain data. Specifically, smaller loss represents the target image x_t has been well translated into source domain. On the other hand, using the fixed decoders, the Seg aims to align the features in feature space, especially for the latent code, to source domain by updating the encoder weights that are self-supervised by the reconstruction path. Thus, the proposed method aligns the target image to source domain on both image and feature levels by updating the weights of T and the encoder of Seg, respectively.

14.2.1 Networks

Segmentation network. The segmentation network Seg (Fig. 14.2) is a multi-task network with segmentation and reconstruction tasks, which is adopted from the 3D U-Net [35] with residual blocks [67]. There are 64 feature maps at each level, except the input and output channel. The whole network takes 3D MRIs as input and outputs the 3D segmentation mask and the reconstructed image. In the offline training phase on $D_s = \{x_s, l_s\}_{s=1}^{N}$, the network is supervised by a segmentation loss and a reconstruction loss: $\mathscr{L}_S = \mathscr{L}_{seg} + \lambda \mathscr{L}_{rec}$. During test-time, the decoder weights are fixed, and only the encoder weights are updated with self-supervised tasks for adapting the target images at the feature level.

Autoencoders. The AE architecture (Fig. 14.2) is a U-Net without skip connections. The AEs are trained offline and designed for optimizing the similarities, and they can be used during test-time to self-supervise the adaptation. There are 32, 16, and 8 feature maps at the three levels. We design two AEs to reconstruct

x_s and y_s at image-level; these are trained by $\mathscr{L}_{rec}(x_s, x'_s)$ and $\mathscr{L}_{rec}(y_s, y'_s)$, respectively, where y_s is the output logits of pre-trained Seg on input image x_s.

Translation network. The translation network T is used to translate a given test image from target domain to source domain, and its architecture can be viewed in Fig. 14.2. However, we found that a complicated translation network would lead to blurry images and geometry shifts, as also discussed by [72, 103]. We found that convolution with kernel size 3 also caused similar problems in our experiments. Thus, to preserve the image quality and information, we build T as a shallow network, which consists of three conv-norm-act layers with $1 \times 1 \times 1$ convolution, IN and LeakyReLU activation function for each layer. The channel numbers are 64, 64, and 1, respectively. In this design, the translation network is able to mimic the intensity and contrast for different scanners or imaging protocols without any major changes of geometry. The T takes images from target domain as inputs and produces translated images which are closer to the source domain. During testing, the translation network is optimized by self-supervised tasks for each target image.

14.2.2 Test-time Optimization

At test-time, two components have updated weights in our framework: the encoder of Seg and T. To increase the generalizability of Seg to target images in feature space, at test-time, the encoder is initialized with the pre-trained weights and updated for all test images instead of reinitialization after each subject. This allows the encoder of Seg to take advantage of the distributional information of the target dataset. For T, the weights are initialized and updated for adapting each target image. In addition, the translation is self-supervised by AEs during test-time. In this way, the target image is aligned to source domain at the image level. For our experiments, we used a single optimizer to update the weights of encoder of Seg and T rather than updating them separately.

14.2.3 Datasets and Implementation Details

We evaluate our proposed method on the scenario of moderate domain shift (inside same multi-site dataset) and big domain shift (across two different multi-site datasets of different age groups and diseases) using T1-w MRIs for segmenting whole brain masks (i.e., skullstripping).

Adult dataset. We use a subset of the multi-site PREDICT-HD database [143], with 3D T1-w 3T MRIs of 16 healthy control subjects (multiple visits per subject, total of 26 MRIs) and 10 Huntington's disease (HD) patients (19 MRIs). The training/validation sets consist of 14/2 healthy control subjects with 22/4 MRIs respectively, and all HD subjects are used for testing. The training and validation MRIs are from a single type of scanner, and the testing set are from several other scanners.

Pediatric AGS dataset. We use a multi-site dataset of 3D T1-w MRIs (1.5T and 3T) from 58 Aicardi

Table 14.1: Quantitative results for HD and AGS datasets. Bold numbers indicate best performance. Significant improvements between the proposed method and all compared methods (2-tailed paired t-test, $p < 0.005$) are denoted via $*$.

	HD dataset			AGS dataset		
Method	Dice	ASD	HD95	Dice	ASD	HD95
NA	96.67(.010)	.442(.385)	1.93(2.25)	90.12(.057)	2.600(1.675)	9.82(4.72)
HM	96.77(.008)	.434(.246)	1.63(1.35)	90.54(.055)	2.222(1.595)	8.08(4.74)
CycleGAN	96.52(.012)	.504(.500)	2.12(2.60)	85.16(.056)	3.630(1.715)	11.4(4.34)
m-NA	96.68(.011)	.509(.451)	2.02(2.53)	90.21(.050)	1.916(1.304)	6.67(3.84)
m-adp	96.13(.012)	.690(.523)	3.82(4.08)	90.63(.048)	1.840(1.245)	6.58(3.85)
SDA-Net [72]	96.55(.009)	.419(.212)	**1.54(0.48)**	90.69(.046)	1.950(1.356)	6.92(4.76)
DAE [103]	96.54(.011)	.462(.372)	1.89(1.99)	90.85(.045)	2.225(1.489)	8.28(5.10)
Proposed	**96.78(.008)**	**.363(.168)**	1.56(0.40)	**92.07(.039)**$*$	**1.154(0.890)**$*$	**4.35(3.06)**$*$

Goutières Syndrome (AGS) subjects [207]. These patients range from infants to teenagers. We again use 16/2 MRIs from adult healthy controls (first dataset) for training/validation, and all AGS subjects are used during testing. The preprocessing steps can be found in [127]. Additionally, images were resampled to $96 \times 96 \times 96$.

Implementation details. The Dice loss [154] is used for segmentation ($\mathscr{L}_{seg}$) during training. In addition, MSE loss was used for every $\mathscr{L}_{rec}$ in both training and inference (test-time). The Adam optimizer with L2 penalty 0.00001, $\beta_1 = 0.9$, $\beta_2 = 0.999$ was used for both training and testing. For offline training, a constant learning rate of 0.0001 was used for the segmentation network and autoencoders, and the learning rate was set to 0.00001 in test-time for updating the weights of T and the encoder of Seg. We validated the performance every epoch during offline training, and we used early stopping if the average validation result did not increase for 20 epochs. Additionally, for each target image, the test-time training was stopped if the loss was greater than the previous iteration. The total epochs were set to 200/10 for training and testing. The training/testing batch size was 1. We implemented the models using an NVIDIA TITAN RTX and PyTorch.

14.3 Results

We compared the proposed method to the following methods: **(1)** no adaptation (NA), i.e., directly apply the pre-trained model on test set, **(2)** histogram matching (HM) between train and test set [179], **(3)** a typical UDA method that employs CycleGAN [258] to translate the test image to source domain, **(4)** a multi-task network only, without adaptation (m-NA), **(5)** a segmentation network only with adaptation of encoder (m-adp), **(6)** TTA with autoencoders (SDA-Net) [72], and **(7)** TTA with denoised autoencoder (DAE) [103]. For a fair comparison, we use the same network architectures (except the reconstruction path of Seg) and apply identical data augmentations for all methods, except for CycleGAN. Additionally, we modified the SDA-Net to 3D version based on the authors' source code. We use the Dice score, average surface distance (ASD) and

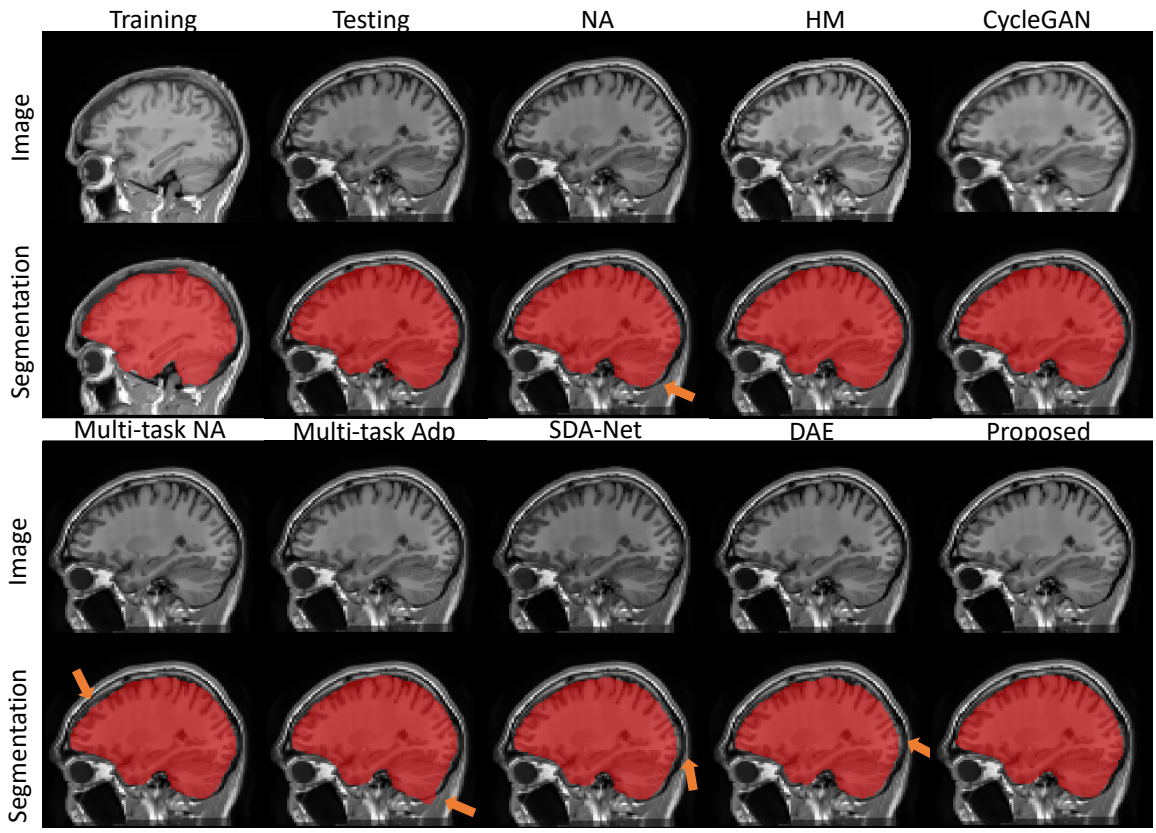


Figure 14.3: Adult HD results. $1^{st}/3^{rd}$ rows show the images/adapted images, and $2^{nd}/4^{th}$ rows show segmentation results. Local segmentation defects are highlighted by arrows.

95-percent Hausdorff distance (HD95) for evaluation.

Adult HD dataset. The quantitative results of the adult HD dataset are shown in the left panel of Tab. 14.1. While it uses different scanners than the healthy adult source domain, this dataset has only a moderate domain shift, and may be treated as a supervised segmentation task (NA) in practice. Thus, all methods work relatively well even without adaptation. Nevertheless, our method has the best performance in all metrics except the HD95.

The qualitative results are shown in Fig. 14.3, where we again observe that all compared methods are able to produce reliable segmentations. However, local defects are present in the baseline methods, as highlighted by orange arrows. Although HM and CycleGAN have better visual quality of adapted image, we note that these methods require access to the source domain data. Among the TTA methods, the proposed method has the best segmentation performance with good quality adapted image.

Pediatric AGS dataset. The quantitative results are shown in the right panel of Tab. 14.1. This dataset has more pronounced domain shift from source domain, and our proposed method has the highest Dice score. In addition, while Dice score is not sensitive to local errors, the ASD and HD95 distance metrics demonstrate

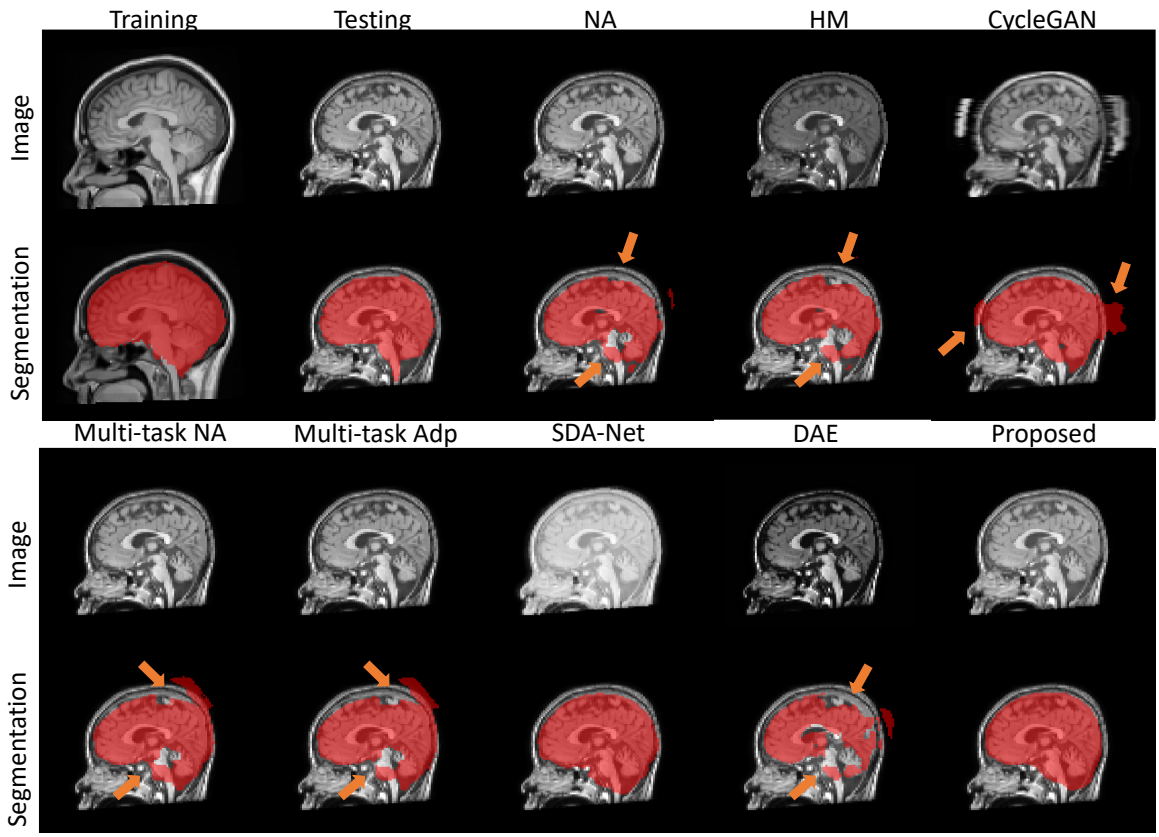


Figure 14.4: Pediatric AGS results. Local segmentation defects are highlighted by arrows.

that our method produces superior segmentations with fewer local defects. 2-tailed paired t-tests show that the proposed method has significantly better ($p < 0.005$) performance for all metrics and compared to each baseline.

Fig. 14.4 shows the qualitative results of adapted images and segmentations. We observe that each baseline method presents either over-segmented or under-segmented regions due to intensity and shape shifts between source and target domains. Specifically, NA and HM under-segment the brain stem area. CycleGAN produces an over-segmented result. For the methods based on multi-task segmentation network, both under-segmented and over-segmented areas appear. Although the SDA-Net produces a plausible segmentation with only small errors in the neck area, it nevertheless achieves worse quantitative results than our proposed method. DAE has good average Dice but poor qualitative results locally (as also evidenced by high ASD and HD95 scores). In contrast, our proposed method produces superior segmentation and is visually closest to the ground truth. We also present the adapted images for each method with image-level adaptation. Even though HM and CycleGAN require training data for image translation, HM changes the contrast in the wrong direction and geometry shifts appear in the CycleGAN since the brain sizes are different between the

domains. All TTA methods translate the target image without any major changes and preserve the details of the target image. However, the adapted images from SDA-NET and DAE have visible biases compared to the training image. Our proposed method produces not only a superior segmentation, but also adapted images visually similar to the training data. In addition, compared to the original image, the contrast between cerebrospinal fluid and other tissues is softened in the proposed adapted image, similar to the adult subjects.

14.4 Conclusion

In this paper, we propose a novel test-time adaptation framework for medical image segmentation in the presence of domain shift. Our proposed framework aligns the target data to source domain at both image and feature levels. We evaluated our method on two datasets with moderate and severe domain shifts. Specifically, intensity and geometry shifts appear between source and target domains for the pediatric AGS dataset. Compared to the baseline methods, our proposed method produced the best segmentations. Quantitative evaluation of the adapted images remains as future work. In future work, we will also apply our method to more datasets with different structures of interest as well as a wider range of image modalities.

CHAPTER 15

Promise: Prompt-driven 3D Medical Image Segmentation Using Pretrained Image Foundation Models

To address prevalent issues in medical imaging, such as data acquisition challenges and label availability, transfer learning from natural to medical image domains serves as a viable strategy to produce reliable segmentation results. However, several existing barriers between domains need to be broken down, including addressing contrast discrepancies, managing anatomical variability, and adapting 2D pretrained models for 3D segmentation tasks. In this paper, we propose ProMISe, a prompt-driven 3D medical image segmentation model using only a single point prompt to leverage knowledge from a pretrained 2D image foundation model. In particular, we use the pretrained vision transformer from the Segment Anything Model (SAM) and integrate lightweight adapters to extract depth-related (3D) spatial context without updating the pretrained weights. For robust results, a hybrid network with complementary encoders is designed, and a boundary-aware loss is proposed to achieve precise boundaries. We evaluate our model on two public datasets for colon and pancreas tumor segmentations, respectively. Compared to the state-of-the-art segmentation methods with and without prompt engineering, our proposed method achieves superior performance. The code is publicly available at https://github.com/MedICL-VU/ProMISe

15.1 Introduction

Recently, image segmentation foundation models [108, 260] have revolutionized the field of image segmentation, demonstrating wide generalizability and impressive performance by training on massive amounts of data to learn general representations. Prompt engineering further improves the segmentation capability of these models. Given proper prompts as additional inputs, these models can handle various zero-shot tasks across domains and produce reliable segmentations during inference. Unlike these broad successes, medical image segmentation is often limited by issues such as expensive data acquisition and time-consuming annotation processing, resulting in a lack of massive public datasets available for training. Thus it is desirable to leverage transfer learning from the natural image domain for robust medical image segmentation [152].

However, directly leveraging pretrained 2D natural image foundation models for 3D medical image segmentation often leads to sub-optimal results [69]. This is primarily because: (1) medical images have their own unique contrast and texture characteristics; (2) anatomical differences among individuals make medical image segmentation challenging; and (3) slice-wise (2D) segmentation with transfer learning discards impor-

This work is accepted by ISBI 2024.

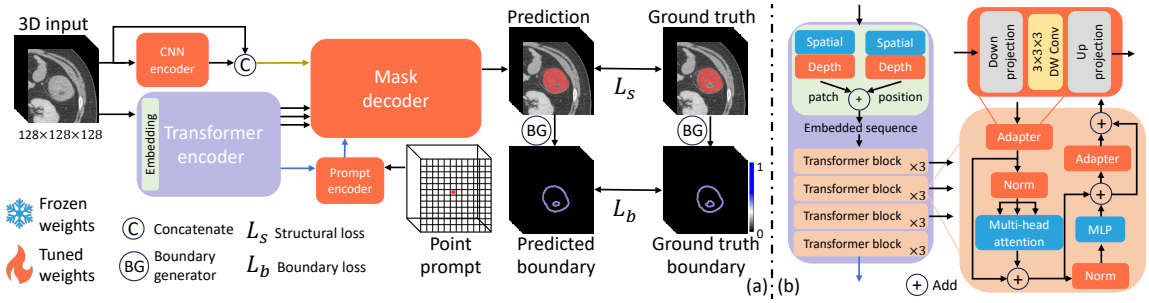


Figure 15.1: The proposed framework (ProMISe) and details of transformer encoder are shown as (a) and (b), respectively.

tant depth-related spatial context in 3D medical data. Given these challenges, can we effectively adapt the pretrained models to achieve robust 3D medical segmentations?

In this paper, we propose ProMISe, **pro**mpt-driven 3D **m**edical **i**mage **se**gmentation using pretrained image foundation models (see Fig. 15.1). Specifically, ProMISe takes a 3D input image and a single point prompt as inputs, and uses image and prompt encoders to produce segmentation. Unlike most promptable models, a shallow convolutional neural network (CNN) is used as complementary path alongside the pretrained transformer image encoder [108], with adapters employed within the transformer to capture 3D depth context. During training, most weights of the adapted transformer encoder remain static; the other components in the proposed method are designed in a lightweight manner for efficiency and trained from scratch. We use a structural loss and a novel boundary-aware loss for precise decisions. The main novel contributions are:

- We propose a method for 3D medical segmentation that adapts pretrained image foundation models. Plug-and-play lightweight adapters are used to better optimize knowledge transfer across domains and more effectively capture fine-grained features. Our method is compatible with various pretrained image models, easy to implement, and cost-effective to train.
- We present a simple yet efficient boundary-aware loss for ambiguous edges. This ready-to-use loss can be seamlessly integrated into any training process without the need for offline edge map generation from ground truth.
- We validate the performance on two public datasets for challenging tumor segmentations. Our method outperforms state-of-the-art segmentation methods consistently.

Related works. Fully finetuning image foundation models for a task requires a large amount computational resources and is not training-efficient. In contrast, partially finetuning [150] or introducing and training new

shallow layers, such as lightweight adapters [165, 29, 56, 232, 224] and the Low-Rank Adaptation (LoRA) module [82, 243], have demonstrated robust performance as parameter-efficient finetuning methods. Recent works use SAM [108] for 3D medical image segmentation in a 2D slice-wise manner, which discard important depth-wise (3D) information and may require additional efforts to create prompts [150, 243]. Other models use adapters; this approach has proven effective for adapting a pretrained model from 2D images to 3D (2D+time) videos [165, 232], and it has subsequently been utilized in 3D medical image segmentation [224] with the use of adapters in the pretrained transformer block [29]. Although these models can segment 3D medical images, the image encoder still operates in a slice-wise (2D) manner with an additional branch for depth information. The weights for this branch that are replicated from the spatial branch demand more computational resources. In contrast, a holistic adaptation of SAM for 3D medical segmentation was proposed in [56], which avoids a depth branch by including an adaptor with depth-wise convolution [165]. However, a single adapter in each transformer block may not fully achieve accurate adaptation due to the notable discrepancies between natural and medical images. Moreover, this method struggles to adequately capture details and can lead to sub-optimal results, especially for tumor segmentation. These challenges and the critical importance of precise segmentation in medical applications motivate our proposed model as a more robust solution.

15.2 Methods

Fig. 15.1 illustrates ProMISe, our proposed framework for 3D medical image segmentation, which employs prompt engineering and a pretrained image foundation model. Specifically, a 3D patch is taken as input and is fed through complementary CNN and transformer encoders. The prompt encoder utilizes the deepest feature from the transformer encoder (blue arrow in Fig. 15.1) as input together with the point prompt. Subsequently, all features, including the original input, are used to predict the segmentation mask via a lightweight CNN decoder. During training, the transformer encoder is partially tuned, while the rest are trained from scratch.

Image encoders. Our model is designed to effectively capture both global and local information using complementary transformer and CNN encoders, respectively.

For the transformer encoder (Fig. 15.1(b)), the input 3D image patch first passes through an embedding layer to create tokens with their positional information. Specifically, the pretrained weights from SAM [108] are employed for spatial patch embedding, and we introduced a trainable depth embedding layer for 3D data. The same approach is applied for positional encoding. Furthermore, we adapted the pretrained weights from SAM and finetuned the normalization layer in every transformer block. Unlike other works that employ a single lightweight adapter at the beginning of the transformer block [165, 56], an additional adapter is used before the output to optimize knowledge transfer across domains and further refine the image features.

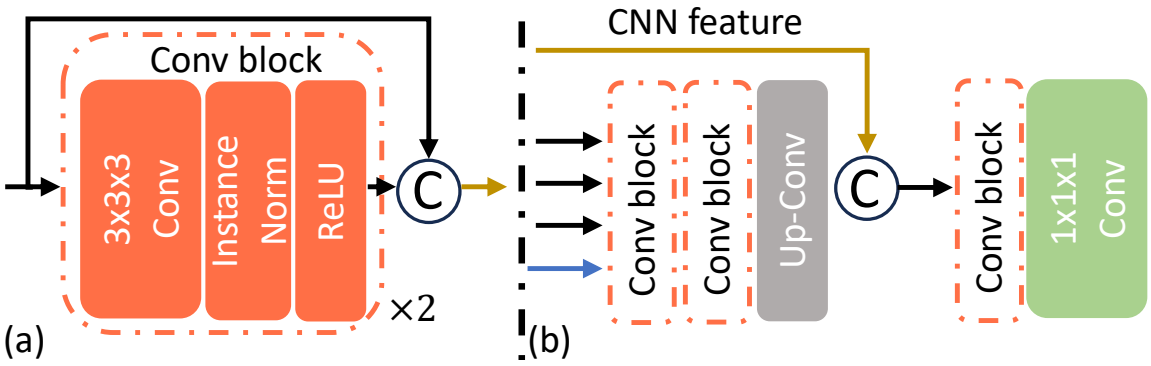


Figure 15.2: The details of (a) CNN encoder, and (b) mask decoder.

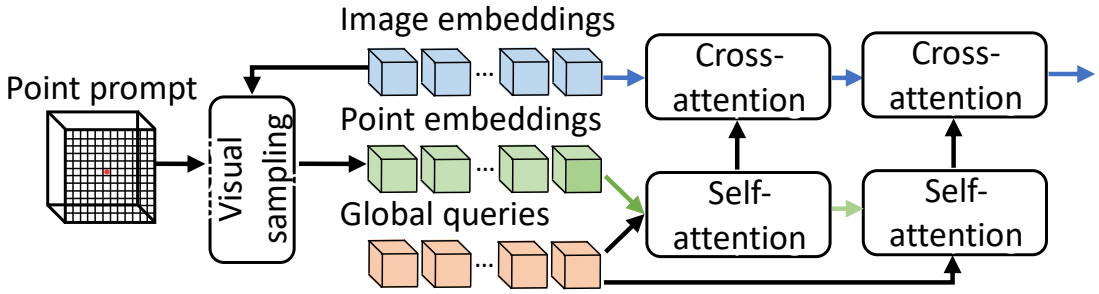


Figure 15.3: The details of the proposed prompt encoder.

Notably, the two adapters employ depth-wise convolution to handle 3D images, and they are identical.

Inspired by the hybrid network design [116], a CNN encoder is used to capture detailed information to complement the transformer. This is particularly desirable for tumor segmentation, as the boundaries are often ambiguous. It is designed as a shallow network for efficiency (Fig. 15.2(a)).

Prompt encoder. We adapt the visual prompt encoder based on [260] (Fig. 15.3). Unlike the prompt encoder proposed in SAM [108], we incorporate image embeddings from the transformer encoder as an additional input. Point embeddings are derived from the given point prompt and image embedding using visual sampling (e.g., grid sampling) to ensure that their semantic features are aligned with image embeddings. Subsequently, the self-attention layer is applied to the point embeddings and learnable global queries. Afterwards, the image embeddings are applied to these queries via cross-attention. The output of the prompt encoder is fed to the mask decoder.

During training, 10 random points from background are provided for each input patch to increase the generlizability to noisy prompts. In contrast to previous work that utilized 40 points from target region as prompts [56], we randomly select 10 point prompts during each iteration if the input patch contains foreground. For prompt engineering [120], our goal is a single click with minimal prior knowledge, but more prompts are supported if desired during inference.

Mask decoder. Instead of directly adapting the mask decoder from the foundation model in a 2D manner, we designed a shallow network to efficiently capture features in 3D and trained it from scratch (Fig. 15.2(b)). The multi-level features from the transformer encoder (Fig. 15.2(b)) are refined by two successive convolutional blocks. These are followed by a transposed convolution to ensure the features remain the same size. The fused features are processed through another convolutional block and a segmentation head for final results.

Boundary-aware loss. In medical image segmentation, accurately delineating the boundaries of objects is important, especially for irregularly shaped objects such as tumors [137]. Besides popular structural segmentation losses, such as the combined Dice loss and cross-entropy loss (denoted as $L_{structural}$), we further propose a boundary loss ($L_{boundary}$) to preserve fine details and produce robust segmentations. Moreover, by emphasizing edge accuracy, the model might generalize better to unseen data for tumor segmentation. As shown in Fig. 15.1, we extract a smooth boundary map rather than a binary boundary for a more robust representation, and because learning from a binary boundary is a challenging task. Specifically, we use average-pooling operation P_{ave} with kernel size 5 as boundary generator. Given either a logits map or a binary mask M, the smooth boundary is derived: $B(M) = |M - P_{ave}(M)|$. The total objective function is:

$$L(S,G) = \lambda_1 L_{structural}(S,G) + \lambda_2 L_{boundary}(B(S),B(G))$$

where S and G represent segmentation and ground truth. $L_{structural} = L_{Dice} + L_{CE}$ is used to capture the structural information and $L_{boundary} = L_{MSE}$ recovers the detailed contours. Unlike other methods [105] that require complicated offline computation of edge or distance maps to avoid iterative generation, our proposed ready-to-use boundary loss is computationally efficient and can be easily adapted to any segmentation task, and is independent of any augmentation.

15.3 Results

Datasets. We evaluated our proposed method on two public datasets from the Medical Segmentation Decathlon (http://medicaldecathlon.com/) for challenging tumor tasks from pancreas and colon applications, where ambiguous edges are present. These consist of 281 ($0.61 \times 0.61 \times 0.7$ to $0.98 \times 0.98 \times 7.5 mm^3$) and 126 ($0.54 \times 0.54 \times 1.25$ to $0.98 \times 0.98 \times 7.5 mm^3$) 3D CT volumes, respectively. Following the setup from the prior study [56], we used the same data split for each task with a training/validation/testing split of 0.7/0.1/0.2 and only use tumor labels to focus on binary segmentation.

Preprocessing. We resample to $1mm$ isotropic resolution, intensity clip based on foreground 0.5 and 99.5 percentiles, and Z-score normalize based on all foreground voxels. Four data augmentations were used:

Table 15.1: Dice and normalized surface Dice (NSD) for colon and pancreas tumor. Bold indicates best performance. Significant improvements (2-tailed paired t-test, $p < 0.05$) are denoted via *. The promptable models use 1 point prompt per 3D volume.

Dataset	Metric	nnU-Net [92]	3D UX-Net [114]	nnFormer [255]
Colon	Dice	45.60	23.07	21.36
	NSD	53.01	32.84	32.05
Pancreas	Dice	39.12	37.57	35.98
	NSD	57.66	55.25	53.45
		Swin-UNETR [202]	3DSAM-adapter [56]	ProMISe
Colon	Dice	37.23	57.32	**66.81***
	NSD	51.16	73.65	**81.24***
Pancreas	Dice	37.98	54.41	**57.46***
	NSD	56.42	77.88	**79.76***

random flip, rotation, zoom, and intensity shift. During training, an input patch of size $128 \times 128 \times 128$ was randomly selected such that its center pixel is equally likely to be foreground or background. Subsequently, each dimension was upsampled to 512.

Implementation details. We utilized pretrained ViT-B from SAM [108] as transformer encoder, and set $\lambda_1 : \lambda_2 = 1 : 10$ during training. The batch size was 1, and initial learning rate was 0.0004 with decreased amount $2e^{-}6$ every epoch. The AdamW optimizer was used with a maximum of 200 epochs. We used PyTorch, MONAI and an NVIDIA A6000 GPU for our experiments. The Dice score and normalized surface Dice (NSD) are used for evaluation. Compared state-of-the-art methods include: CNN (nnU-Net [92]), CNN with large kernel (3D UX-Net [114]), Swin-encoder with CNN decoder (Swin-unetr [202]), pure transformer (nnFormer [255]), and adaptation method with adapters (3DSAM-adapter [56]). We retrained using their official codes, and the pretrained weights are also employed if publicly available.

Quantitative results. Tab. 15.1 presents a detailed comparison of results for colon and pancreas tumor segmentation. Notably, while CNN-based networks segment these tumors more effectively than transformers, prompt-driven methods outperform others when provided with only a single point in the entire volume. Our proposed method consistently outperforms all in terms of both Dice and boundary (NSD) metrics.

Ablation study. We also investigated the efficiency variations of the proposed ProMISe (Tab. 15.2). The use of two adapters and the boundary-aware loss mostly improved the results. Interestingly, switching from trilinear upsampling to up-convolution improved the performance for the colon, but showed a decline for the pancreas. This implies that trilinear upsampling may be more appropriate for pancreas tumors, which are typically round in shape. Using concatenation (-C) in the CNN encoder offers better Dice scores than residual connections (-R), though the latter improves surface quality more. While the performance of ProMISe

Table 15.2: Quantitative results of ablation study with single point prompt unless noted. R and C represent residual and concatenate fusions, and B indicates boundary loss. + shows the cumulative variants. Best viewed by individual sections.

	Colon		Pancreas	
Method	Dice	NSD	Dice	NSD
baseline [56]	57.32	73.65	54.41	77.88
+ two adapters	61.61	73.88	56.08	77.89
+ up-Conv	62.92	77.62	55.37	77.38
ProMISe-R	63.67	79.96	55.15	79.02
ProMISe-R-B	64.75	79.77	56.57	79.46
ProMISe-C	64.76	77.59	56.35	78.01
ProMISe-C-B (**proposed**)	66.81	81.24	57.46	79.76
baseline [56] (10 prompts)	63.09	79.97	55.94	79.18
ProMISe-C-B (10 prompts)	67.28	81.63	58.05	80.36

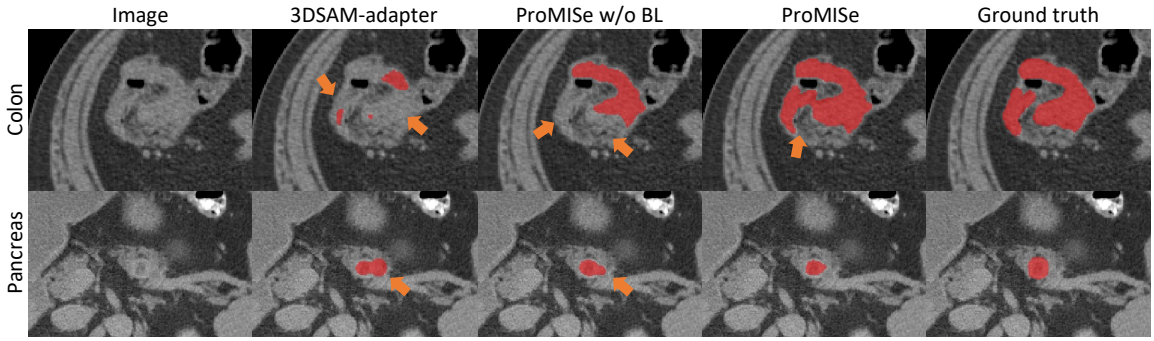


Figure 15.4: Qualitative results. BL denotes boundary-aware loss. The major differences are highlighted by orange arrows.

improves with 10 prompts, the improvement is limited over a single prompt. Furthermore, it is challenging to identify the tumor area due to ambiguous boundaries, making the use of a single click preferable in practice, as it requires less expert knowledge.

Qualitative results. Fig. 15.4 shows qualitative visualizations from top-performing promptable methods. ProMISe yields results that closely align with the ground truth. 3DSAM-adapter [56] fails to detect certain regions that ProMISe captures, even without the boundary-aware loss. This indicates the improved generalizability of the model through our proposed modifications. Moreover, the use of the boundary-aware loss yields robust segmentations, alleviating issues of both under-segmentation for colon and over-segmentation for pancreas tumors, respectively. Notably, the boundary-aware loss improves segmentation not just for the irregularly shaped colon tumors but also for the pancreas tumors, which typically have a more regular, rounded shape. However, slight under-segmented areas are found in pancreas segmentation.

15.4 Conclusion

In this paper, we propose a promptable network, named ProMISe, designed for robust 3D tumor segmentation using pretrained weights from image foundation models. We evaluate on two public datasets, where our model consistently outperforms state-of-the-art methods across all tasks. Moreover, the critical role of the two adapters and boundary-aware loss techniques are demonstrated. Future work will aim to improve the efficiency through knowledge distillation and further explore different point sampling strategies.

CHAPTER 16

Assessing Test-time Variability for Interactive 3D Medical Image Segmentation with Diverse Point Prompts

Interactive segmentation model leverages prompts from users to produce robust segmentation. This advancement is facilitated by prompt engineering, where interactive prompts serve as strong priors during test-time. However, this is an inherently subjective and hard-to-reproduce process. The variability in user expertise and inherently ambiguous boundaries in medical images can lead to inconsistent prompt selections, potentially affecting segmentation accuracy. This issue has not yet been extensively explored for medical imaging. In this paper, we assess the test-time variability for interactive medical image segmentation with diverse point prompts. For a given target region, the point is classified into three sub-regions: boundary, margin, and center. Our goal is to identify a straightforward and efficient approach for optimal prompt selection during test-time based on three considerations: (1) benefits of additional prompts, (2) effects of prompt placement, and (3) strategies for optimal prompt selection. We conduct extensive experiments on the public Medical Segmentation Decathlon dataset for challenging colon tumor segmentation task. We suggest an optimal strategy for prompt selection during test-time, supported by comprehensive results. The code is publicly available at https://github.com/MedICL-VU/variability

16.1 Introduction

To date, deep learning methods have demonstrated superior performance in various medical image segmentation tasks [137]. However, the generalizability and effectiveness of these fully automated methods may be limited by the amount of available labeled medical data [218]. Instead, interactive segmentation methods that integrate user knowledge and application requirements have been proposed [212].

To tackle the data and label availability issue in medical imaging, it is desirable to utilize knowledge from the natural images domain, wherein large public datasets are more readily accessible [152]. Recent studies [150, 224, 56, 121, 36, 249, 246] have leveraged insights from pretrained natural image foundation models, such as the Segment Anything Model (SAM) [108], which is trained on massive datasets, to facilitate robust medical image segmentation through parameter-efficient transfer learning techniques. These methods, which use diverse visual prompts, achieve robust performance due to the strong prior provided by the interactive prompt during test-time.

Compared to other interaction formats, such as boxes or scribbles, point prompts are preferable in prac-

This work is accepted by ISBI 2024.

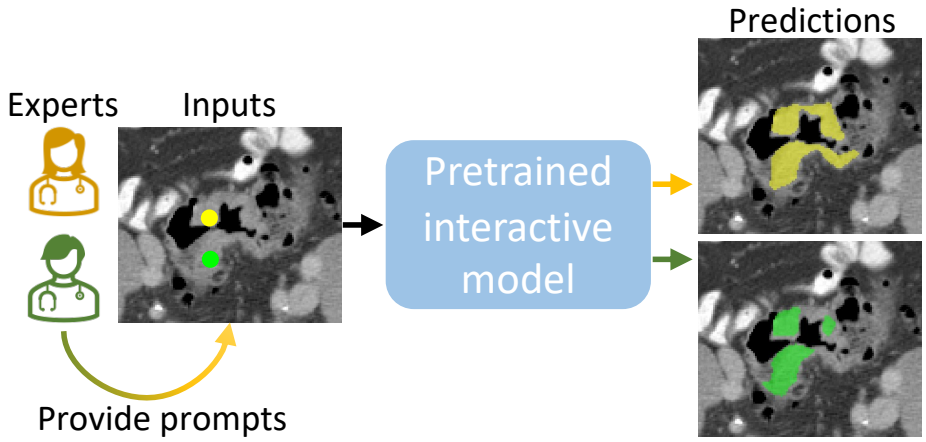


Figure 16.1: A 3D interactive segmentation model, which takes an image and a point prompt as inputs. The segmentation result can vary with different prompts provided by experts.

tice, as they require less effort, especially for 3D medical images. However, in the context of interactive segmentation models, determining precise key points during inference can be elusive, especially in medical images characterized by low quality, poor contrast, and ambiguous boundaries. Furthermore, subjectivity leads to variability in prompt choice as seen in Fig. 16.1, and this can be exacerbated by different user expertise levels, resulting in different segmentation outcomes at test-time [185, 233, 31]. A random selection strategy is widely used, but random points might not represent the key features of a medical image effectively. To the best of our knowledge, no existing work has explicitly addressed prompt selection in the medical field. Previous works have highlighted the importance of key points and have leveraged these during training to produce robust segmentations [130, 115]. In this paper, we aim to investigate the optimal strategy for point prompt selection for pretrained interactive segmentation models at test-time.

We first assess the test-time variability caused by point prompt selection of interactive 3D medical image segmentation model. Specifically, we use the ProMISe [121] model as the backbone, which leverages the pretrained weights from SAM [108]. It takes image and point prompts as inputs to produce a segmentation. We evaluate the segmentation performance variability caused by different number, region, and selection strategy for prompts during inference to quantify how such interactive models respond to diverse point prompts (Fig. 16.2). To provide a pragmatic solution, our investigation focuses on three aspects: (1) determining the necessary number of prompts, (2) identifying effective prompt placement locations, and (3) formulating a strategy for prompt selection.

We conduct our evaluation on colon tumor segmentation from the public Medical Segmentation Decathlon dataset [4], which is characterized by irregular shapes and ambiguous boundaries. Our findings suggest a straightforward strategy for test-time prompt selection without requiring additional effort, leading

to significantly improved segmentation results over random prompt selection.

16.2 Methods

Dataset. We used the Medical Segmentation Decathlon [4] for our experiments. We select the challenging colon tumor segmentation task where ambiguous edges and irregular shape are present. The dataset contains 126 3D CT volumes with resolution ranging from $0.54 \times 0.54 \times 1.25$ to $0.98 \times 0.98 \times 7.5 mm^3$, resampled to $1mm$ isotropic resolution. The training, validation and test sets contain 88, 12 and 26 subjects, respectively. The experimental settings are same as Chapter 15.

Interactive segmentation model. We employ ProMISe [121], a SAM-based interactive segmentation model which takes an input image and point prompts as inputs. During training, an input patch of size $128 \times 128 \times 128$ was randomly selected such that its center voxel is equally likely to be foreground or background. In addition, random flip, rotation, zoom and intensity shift are used as data augmentation strategies. For point prompts, 10 positive points from the foreground and 20 negative points from the background are randomly selected for each input patch, respectively. We use negative points only if the number of positive points in an input patch is fewer than 10. During inference, ProMISe takes the input image and only a single random point prompt within the whole tumor area to produce the segmentation (Fig. 16.1). We consider this random prompt selection as the baseline method and evaluate various other point prompt selection strategies, aiming to identify a better strategy to optimize the segmentation performance without requiring extensive additional user effort.

Sub-region generation. As shown in Fig. 16.2(a), we divide the ground truth mask into three distinct sub-regions: boundary, margin, and center. We generate these regions using an average-pooling operation, denoted as P^k_{ave}, where k represents the kernel size. For a given a binary segmentation S, the binary boundary sub-region is derived: $B(S) = thres(S \odot |S - P^3_{ave}(S)|)$, where $|\bullet|$ denote the absolute value function, $\odot$ represents element-wise multiplication, and *thres* is binary thresholding with a threshold of 0. Similarly, we obtain the margin sub-region as: $M(S) = thres(S \odot |S - P^7_{ave}(S)|) - B(S)$, and the center sub-region is simply $C(S) = S - M(S) - B(S)$. This sub-region generation can be seamlessly integrated into the inference without reducing computation speed, and may be used to facilitate pseudo-label learning.

Random selection. To ensure objectivity and fair comparison, random selection serves as the underlying method in our approach for every selection strategy, instead of actual user prompts. Additionally, we use randomly selecting single point prompts within the whole tumor area as the baseline method for comparison, as this represents the most common selection strategy. We specify the random seed to control variability and assess the impact of different settings within these selection strategies. It is important to note that the location of point prompts depends on the seed and the selected region. In other words, for a specific region and seed,

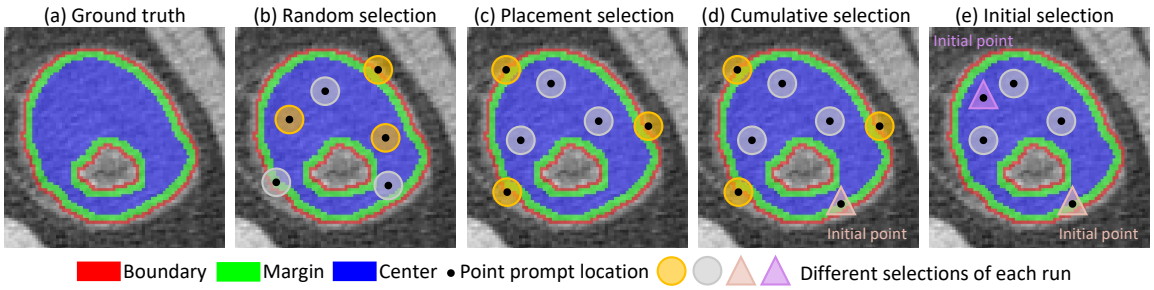


Figure 16.2: Various point prompt selection strategies. Different prompt colors represent different runs. **(a)** Ground truth, highlighting the entire tumor and its three different sub-regions. **(b)** Random selection throughout the entire tumor. **(c)** Prompts confined to a specific sub-region. **(d)** Cumulative selection, where the initial point (triangle) is fixed while cumulative points (circles) vary between runs. **(e)** Initial selection, where initial points vary while cumulative points remain fixed.

the point locations remain the same if the number of point prompts is unchanged, and differences only arise with additional points. Moreover, we investigate the benefits of using more than one prompt. Specifically, we use 1, 5, 10, 20, and 100 prompts.

Placement selection. To identify the effectiveness of point placement, the point prompts are randomly selected within the given regional constraints. Fig. 16.2(c) depicts a situation with a same seed setting but different selection region.

Strategy selection. The initial selection is important in prompt selection [130]. To further examine its test-time variability, we assess two strategy selections, cumulative and initial selections, as shown in Figs. 16.2(d) and (e), respectively. For cumulative selection, we randomly set an initial point (triangle in Fig. 16.2(d)) in a specific region as the fixed point, and then compare the segmentation performance for different placements of the cumulative points (circles). All selected prompts, including initial and cumulative points, are used for each run. Similarly, in Fig. 16.2(e), we select different initial points while keeping the cumulative points fixed. These approaches are practical since selecting a single point prompt is preferred in practice for its minimal effort, compared to other prompt types such as boxes, scribbles or multiple point prompts. Furthermore, a preliminary segmentation or pseudo label may first be generated from the initial point(s) to serve as a reference for the selection of the cumulative points.

Implementation details. We trained the ProMISe [121] model for a maximum of 200 epochs with batch size of 1. The initial learning rate was 0.0004, decreasing by $2e^{-6}$ every epoch, and the AdamW optimizer was employed. During inference, we randomly selected a set of 50 random seeds and used it for each selection strategy. The Dice score and normalized surface Dice [186] were used for evaluation, and we report mean and standard deviations over the 50 runs (seeds). We used PyTorch, MONAI and an NVIDIA A6000 for our experiments.

Table 16.1: Quantitative results of random selection, presented as $mean \pm std.\,dev.$ ○ and ● denote absent and present for the point prompt select region. **Bold** and underline denote the best performances for each **column (region-wise)** and row (prompt-wise), respectively. Blue indicates the baseline method.

Region			Dice score				
B	*M*	*C*	1*P*	5*P*	10*P*	20*P*	100*P*
●	○	○	.622±.014	.650±.009	.652±.008	.650±.007	**.655±.006**
○	●	○	.630±.015	.652±.010	.654±.009	.652±.008	.654±.006
○	○	●	**.642±.012**	**.654±.006**	.654±.006	.652±.009	.653±.005
●	●	○	.632±.014	.653±.010	.652±.008	**.654±.007**	.652±.005
●	○	●	.634±.016	.652±.010	.653±.009	.653±.008	.652±.005
○	●	●	.634±.016	.654±.009	.654±.008	.653±.009	.654±.007
●	●	●	.637±.014	.653±.008	**.655±.008**	.653±.007	.652±.007
Region			Normalized surface Dice				
B	*M*	*C*	1*P*	5*P*	10*P*	20*P*	100*P*
●	○	○	.768±.015	.798±.010	.803±.009	.801±.009	.806±.007
○	●	○	.776±.016	.802±.011	.805±.009	.805±.009	**.807±.008**
○	○	●	**.788±.015**	.802±.006	.804±.007	.803±.010	.805±.007
●	●	○	.777±.016	.803±.011	.802±.010	**.806±.007**	.804±.007
●	○	●	.779±.017	.801±.010	.803±.010	.804±.009	.805±.007
○	●	●	.779±.017	.803±.011	.804±.009	.805±.010	.807±.009
●	●	●	.783±.016	**.803±.008**	**.806±.009**	**.806±.007**	.804±.008

B=boundary, *M*=margin, *C*=center, *P*=point prompt(s) per volume.

16.3 Results

Quantitative results. Tab. 16.1 presents a detailed comparison of random selections, focusing mainly on two practical questions: whether more points are needed, and where to select them. The results clearly show that using more than a single prompt benefits segmentation in both metrics compared to a single point prompt. However, a distinct cutoff in improvement is observed at five point prompts per 3D volume, which suggests diminishing returns past that. Compared to the baseline method, choosing a single point prompt focusing on the center region often yields superior segmentation. As the number of points increases, the impact of choosing different regions becomes less pronounced.

Tab. 16.2 compares the different cumulative strategies with different cumulative point placement and different number of prompts. We found that randomly selecting a single point prompt in the whole tumor area, with cumulative points picked from the center region achieves the best results. This improves the Dice score of baseline method by 2%, with statistical significance confirmed ($p < 0.001$) through a 2-tailed paired t-test. Therefore, we suggest this straightforward selection strategy as the optimal solution during test-time. We note that the cumulative points from the easily identifiable center area can either be derived from a pseudo label or require minimal extra effort by experts to improve segmentation performance. Thus, in practice, selecting a single initial point is enough for the recommended method, resulting in good performance for minimal prompt input effort.

Table 16.2: Quantitative results of cumulative selection. **Bold** and underline denote the best performances for each **column (region-wise)** and row (prompt-wise), respectively. Orange indicates the suggested selection strategy.

Region				(Initial + cumulative) points			
Init.	*Cumu.*			Dice score			
W	*B*	*M*	*C*	$(1+4)P$	$(5+5)P$	$(10+10)P$	$(20+80)P$
●	●	○	○	.651±.011	.652±.008	.652±.007	**.655±.005**
●	○	●	○	.654±.011	.654±.008	.653±.008	.653±.005
●	○	○	●	**.657±.008**	**.654±.007**	**.655±.007**	.653±.007
●	●	●	●	.654±.010	**.654±.007**	.653±.007	.652±.007
Init.	*Cumu.*			Normalized surface Dice			
W	*B*	*M*	*C*	$(1+4)P$	$(5+5)P$	$(10+10)P$	$(20+80)P$
●	●	○	○	.799±.009	.803±.009	.804±.008	**.807±.007**
●	○	●	○	.804±.009	**.806±.009**	.806±.009	.806±.007
●	○	○	●	**.805±.008**	.804±.008	**.808±.009**	.806±.008
●	●	●	●	.804±.010	.805±.008	.805±.008	.804±.008

W=whole, *B*=boundary, *M*=margin, *C*=center,
P=point prompt(s) per volume, *Init.*=initial, *Cumu.*=cumulative.

Table 16.3: Quantitative results of initial selection. The **bold** and underline denoted best performances for each **column (region-wise)** and row (prompt-wise), respectively. Orange indicates the suggested selection strategy. W refers to the whole region, i.e., $W = B + M + C$.

		Region (Dice)			
Init.	*Cumu.*	*B*	*M*	*C*	*W*
$1(W)$	$4P$	.651±.011	.654±.011	**.657±.008**	.654±.010
$1(W)$	$9P$	.652±.008	.655±.008	**.655±.006**	.653±.010
$1(C)$	$4P$	.653±.008	.651±.007	**.654±.006**	.652±.008
$1(C)$	$9P$	.654±.007	.654±.007	**.654±.006**	.653±.009

Tab. 16.3 shows the impact of initial selection region. The results indicate that the whole and center areas are the best regions for initial and cumulative points, respectively. However, the differences among the methods are minor.

Fig. 16.3 presents the impact of suggested method on different subjects grouped by Dice. Compared to baseline method, the suggested method has higher Dice scores for most groups. In addition, the suggested method has more contribution to the subjects with high-quality segmentations produced by baseline method, as evidenced by Dice scores and number of cases for the two highest Dice groups.

Qualitative results. Fig. 16.4 shows qualitative results for the random selection with a single point prompt. Both boundary and margin selections produce under-segmented results near the ambiguous areas, while center selection captures the missing areas. In Fig. 16.5, the suggested method dramatically improves segmentation. The results match the ground truth well, with the exception of slightly over-segmented areas.

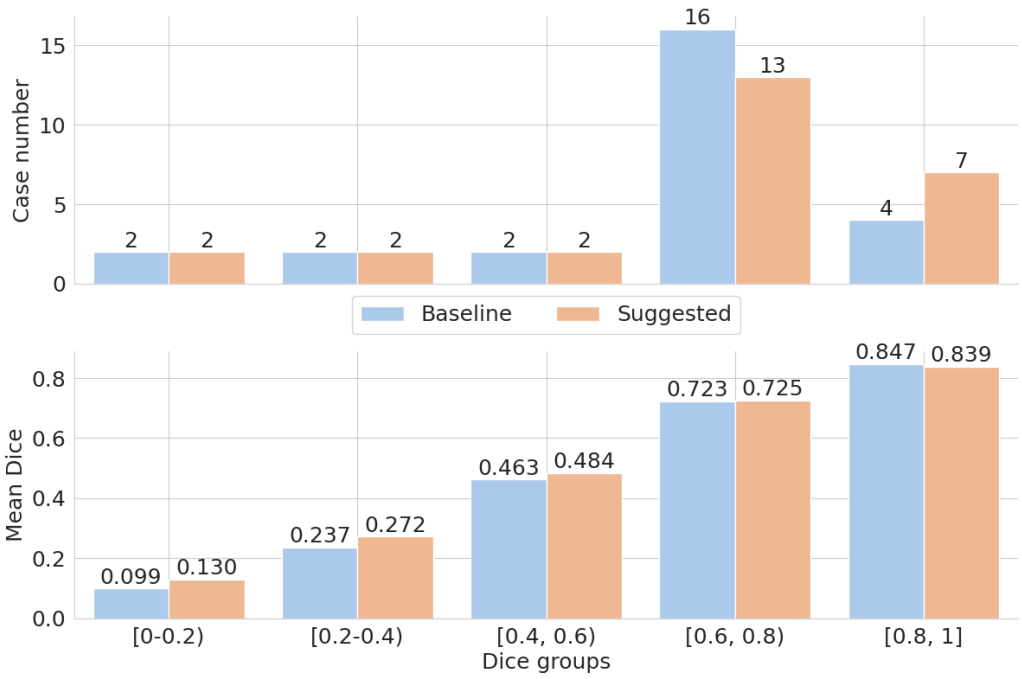


Figure 16.3: Comparison of performance on Dice distribution.

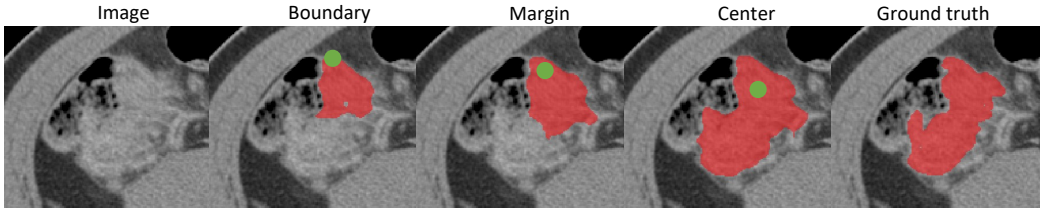


Figure 16.4: Qualitative results for random selection with a single point prompt (green). Labels show the selection regions.

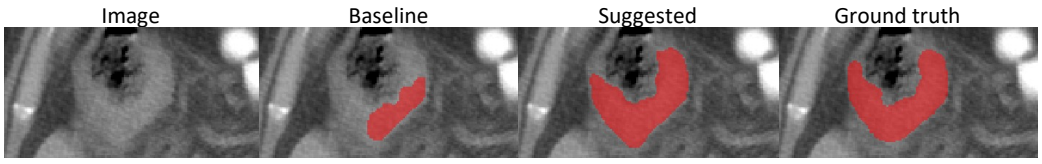


Figure 16.5: Qualitative results of suggested and baseline methods.

16.4 Conclusion

In this paper, we suggest a straightforward prompt strategy for interactive test-time segmentation, without requiring extensive additional effort. We evaluate on a public dataset for the challenging colon tumor segmentation task, with a significant improvement over the baseline method. Future work will validate the suggested method on more models and public datasets, as well as negative point selection strategy.

CHAPTER 17

PRISM: A Promptable and Robust Interactive Segmentation Model with Visual Prompts

In this paper, we present PRISM, a **P**romptable and **R**obust **I**nteractive **S**egmentation **M**odel, aiming for precise segmentation of 3D medical images. PRISM accepts various visual inputs, including points, boxes, and scribbles as sparse prompts, as well as masks as dense prompts. Specifically, PRISM is designed with four principles to achieve robustness: (1) Iterative learning. The model produces segmentations by using visual prompts from previous iterations to achieve progressive improvement. (2) Confidence learning. PRISM employs multiple segmentation heads per input image, each generating a continuous map and a confidence score to optimize predictions. (3) Corrective learning. Following each segmentation iteration, PRISM employs a shallow corrective refinement network to reassign mislabeled voxels. (4) Hybrid design. PRISM integrates hybrid encoders to better capture both the local and global information. Comprehensive validation of PRISM is conducted using four public datasets for tumor segmentation in the colon, pancreas, liver, and kidney, highlighting challenges caused by anatomical variations and ambiguous boundaries in accurate tumor identification. Compared to state-of-the-art methods, both with and without prompt engineering, PRISM significantly improves performance, achieving results that are close to human levels. The code is publicly available at https://github.com/MedICL-VU/PRISM.

17.1 Introduction

In recent years, deep learning-based methods have achieved state-of-the-art performance in various segmentation tasks [137, 92]. However, achieving robust outcomes remains a challenge due to significant anatomical variations among individuals and ambiguous boundaries in medical images. Alternatively, interactive segmentation models, such as the Segment Anything Model (SAM) [108], offer a solution by involving humans in the loop and utilizing their expertise to achieve precise segmentation. This requires users to indicate the target region by providing visual prompts, such as points [140, 108, 147, 56, 121, 120, 214, 52], boxes [220, 246, 36, 212, 108, 149, 26, 233], scribbles [214, 30, 223, 34], and masks [108, 199, 215, 31, 140].

Importantly, a robust interactive segmentation model should be effective in responding to visual prompts given by users with minimal interactions. Inspired by SAM [108], many interactive segmentation methods have been proposed in medical imaging. However, most of these SAM-based methods are limited to using a single type of prompt [220, 56, 233, 52, 121, 36, 149, 26, 246]. This limits the available information, impacting the effectiveness of the model. Due to the heterogeneity in anatomy and appearance of medical

This work is submitted to MICCAI 2024.

images, the models should leverage the advantages of different prompts to achieve optimal performance across various medical applications.

Moreover, these interactive methods do not involve humans in the loop [56, 52, 246, 121, 36, 220, 149, 26, 233], i.e., the visual prompts are applied to the automatic segmentation model only once, which may not be sufficient to achieve robust outcomes with the initial prompts provided. In contrast, practical applications often necessitate iterative corrections based on new prompts from the user until the outcome meets their criteria [197]. Unfortunately, few studies have explored such human-in-loop approach for medical interactive segmentation [31, 215], and their performance has been suboptimal.

An additional concern is the efficiency of the model with respect to training and user interaction requirements. Training a model from scratch with extensive datasets [31, 223, 149, 215] improves overall representational capabilities, but this is time-consuming and typically requires substantial computational resources for optimal performance. Furthermore, models may lack specificity for particular applications due to existing domain gaps in medical imaging [215]. This requires an increased number of user prompts to reach adequate performance. Moreover, 2D models [31, 223, 149] are not considered "efficient" for 3D images as they require significant user effort to provide visual prompts for each slice.

In this work, we propose a robust method for interactive segmentation in medical imaging. We strive for human-level performance, as a human-in-loop interactive segmentation model with prompts should gradually refine its outcomes until they closely match inter-rater variability [61]. We present PRISM, a **P**romptable and **R**obust **I**nteractive **S**egmentation **M**odel for 3D medical images, which accepts both sparse (points, boxes, scribbles) and dense (masks) visual prompts. We leverage an iterative learning [108, 199] and sampling strategy [197] to train PRISM to achieve continual improvements across iterations. The sparse visual prompts are sampled based on the erroneous region of the prediction from previous iteration, which simulates the human correcting behavior. For each iteration, multiple segmentation masks are generated [129] along with regressed confidence scores. The output with the highest score is selected, increasing the robustness of the model. This approach is similar to model ensembling [131]. Moreover, inspired from [158, 97], the output is fed into a shallow corrective refinement network to correct the mislabeled voxels and refine the final segmentation, adhering to the goal of being effectiveness. A hybrid image encoder with parallel convolutional and transformer paths [116] is used as the backbone to better capture the local and global information from a medical image. Our main contributions are summarized as:

- Interaction: PRISM accepts various visual prompts, offering versatility that effectively addresses the diverse challenges of medical imaging segmentation. This approach ensures precise outcomes through user-friendly interaction.

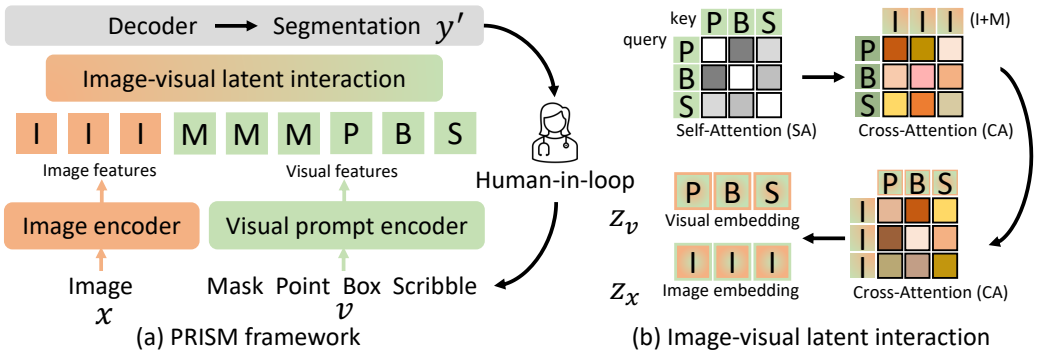


Figure 17.1: **(a)** PRISM takes an image (x) and visual prompts (v) to produce a segmentation (y'). The user then provides prompts for next iteration. **(b)** Interaction between image and visual features in the latent space to produce image (Z_x) and visual (Z_v) embeddings with self- and cross-attention mechanisms.

- Human-in-loop: PRISM is trained with iterative and confidence learning, allowing for continuous improvements and robust performance. With each user input, the performance progressively approaches expert-level precision.
- Network architecture: PRISM employs a corrective refinement network and a hybrid encoder to produce precise segmentation for challenging conditions.

We evaluate PRISM on four public tumor datasets, including tumors in colon [4], pancreas [4], liver [10] and kidney [74], where anatomical differences among individuals and ambiguous boundaries are present. Comprehensive validation is performed against state-of-the-art automatic and interactive methods, and PRISM significantly outperforms all of them with substantial improvements.

17.2 Methods

PRISM employs a generic encoder-decoder architecture as displayed in Fig. 17.1(a). For a given input image x, along with visual prompts v, the image and prompt encoders generate the image and visual features, respectively. These latent features then interact through self- and cross-attention mechanisms (Fig. 17.1(b)). The resulting embeddings, Z_x and Z_v, are fed into the decoder to produce the final segmentation y'. The expert then gives sparse prompt based on the erroneous regions in y' for next iteration, in human-in-loop manner.

Fig. 17.2(a) details PRISM, where a hybrid encoder combines parallel CNN and vision transformer paths to better extract image features. The last feature map of decoder f_d is integrated with the visual embedding to generate multiple mask predictions and associated confidence scores (Fig. 17.2(b)). A selector is employed to pick the candidate prediction with the highest confidence score, which is subsequently fed into the corrective refinement network to generate the y'. In our experiments, we use a sampler to mimic expert corrective actions by identifying point and scribble prompts from the false positive and negative areas within y'.

Iterative learning. PRISM is designed for clinical applications, which require improvement through successive iterations. We define the loss for a single iteration as L_i, and the total loss as $L_{total} = \sum_{i=1}^{N} L_i$, where N denotes the total number of iterations. The input visual prompts come only from the current iteration, i.e., they are not cumulative. Since the dense prompt y'_i is generated from the previous iteration and retains the gradient, PRISM can learn the relationship between iterations. For efficiency, the image features are generated only once at the initial iteration. The following prompt types are allowed (Fig. 17.2(b)):

- At each iteration i, point prompts are randomly sampled with uniform distribution from the false negative (FN) and positive (FP) regions of y'_{i-1}.
- The scribbles are generated similar to Scribbleprompt [223] by first extracting the skeleton of the FN and FP regions of y'_{i-1} from the previous iteration. Next, a random mask is created by pixels randomly sampled from a normal distribution, applying Gaussian blur, and thresholding the image based on its mean value. This random mask is applied to divide the skeleton of y'_{i-1} into separate, smaller parts. Finally, a random deformation field and random Gaussian filtering are used to change the scribble curvature and thickness.
- The 3D bounding box (BB) is determined based on the ground truth y at the initial iteration ($i = 1$) and remains unchanged for the subsequent iterations.
- The logits map at iteration $i-1$ is used to provide additional information instead of a binary mask as the dense prompt for iteration i, for $i > 1$.

Confidence learning. Unlike traditional methods that output a single mask, PRISM increases robustness by generating multiple outputs, each with a confidence score. The last feature map of the decoder, denoted as $f_d = Decoder(Z_x)$, interacts with the visual embedding to produce continuous maps m_j and confidence scores s_j via $m_j = f_d \times MLP_j^m(Z_v)$ and $s_j = MLP_j^s(Z_v)$, where MLP indicates the multi-layer perceptrons, and j is an index over the multiple outputs. Thus, m_j interacts with visual prompts at both latent- and image-level. The process of confidence supervision for each iteration is defined by: $L_{con} = \sum_{j=1}^{M}(\lambda_s L_s(m_j, y) + \lambda_b L_b(m_j, y) + \lambda_r L_r(s_j, 1 - L_{Dice}(m_j, y)))$, where L_s, L_b, L_r and L_{Dice} indicate structural, boundary, regression and Dice loss, respectively. More precisely, the structural information is captured through a combination of Dice loss and cross-entropy loss: $L_s = L_{Dice} + L_{CE}$. For a given logits map or binary mask, the smooth boundary is derived as: $B(m) = |m - P_{ave}(m)|$, where P_{ave} denotes the average pooling layer. The boundary loss (L_b) is evaluated by $L_{MSE}(B(m), B(y))$, with L_{MSE} representing the mean square error loss. We use L_{MSE} for the regression loss.

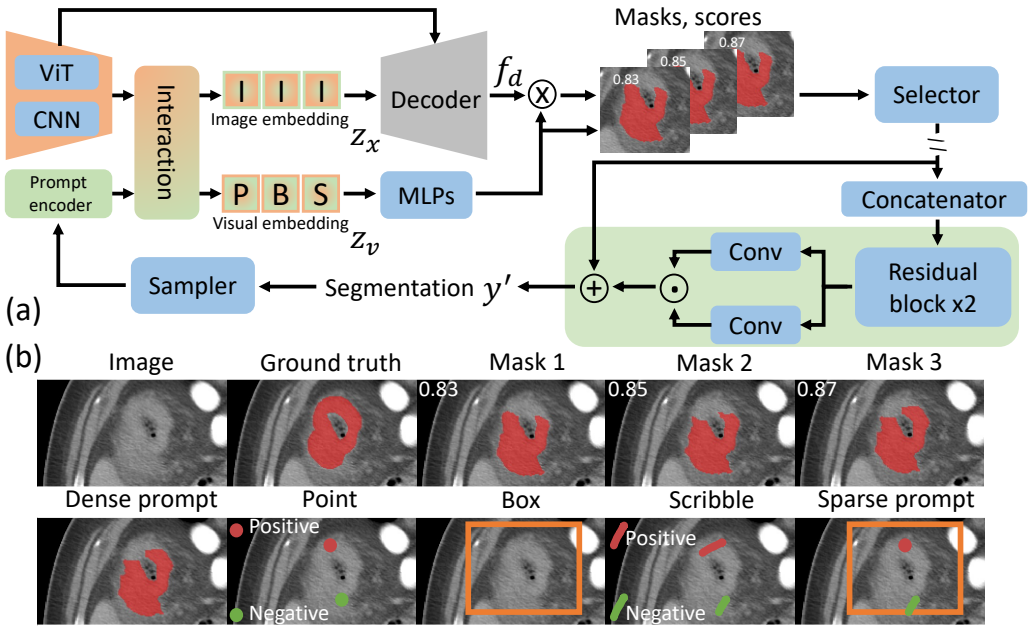


Figure 17.2: **(a)** Details of PRISM. Green highlights the corrective refinement network. **(b)** Top row shows the multi-mask prediction with labeled confidence scores. The selector would then pick Mask 3 as the dense prompt. Possible visual prompts given this dense prompt are shown in the bottom row.

Corrective learning. As shown in Fig. 17.2(a), a selector is employed to choose the candidate mask $\hat{m}$ corresponding to the highest confidence score s_j, which is then fed into the corrective refinement network. We design this network to be both shallow and efficient, consisting of two cascaded residual blocks and two convolution operations in two different branches. It takes the input x_c at original image resolution. Importantly, x_c is the concatenation of 4 images: (1) the input image x, (2) the binary mask $\hat{b}$ derived from $\hat{m}$, (3) the binary positive prompt maps accumulated from iteration 0 to i, and (4) the similar binary cumulative negative prompt map. For efficiency, the x_c is downsampled and then upsampled back for y'_i. The final segmentation y'_i is supervised by: $L_{cor} = \lambda_s L_s(y'_i, y) + \lambda_b L_b(y'_i, y)$. Notably, the corrective network does not backpropagate to the network weights for candidate mask generation and focuses exclusively on refining any given logit map. The y'_i is fed to sampler for next iteration $i+1$.

Implementation details. For training, the total objective function is: $L_{total} = \sum_{i=1}^{N}(L_{con_i} + L_{cor_i})$, where the ratio of $\lambda_s : \lambda_b : \lambda_r$ as $1 : 10 : 1$. The batch size was set to 2, and the initial learning rate was $4e^{-5}$, which was reduced by $2e^{-6}$ after every epoch. We employed the AdamW optimizer across a maximum of 200 epochs. Our study was conducted on an NVIDIA A6000. The input image was cropped to a 3D $128 \times 128 \times 128$ patch centered on a foreground pixel. The number of iterations (N) was fixed at 11, with 1 to 50 point prompts randomly chosen per iteration during training. The Dice and normalized surface Dice (NSD) are used as evaluation metrics. We compared to both fully automated [92, 202, 114] and interactive

Table 17.1: Quantitative results. Bold indicates best performance. PRISM-plain is a simplified model for a fair comparison to methods not allowing BB prompts [56, 121, 215], and methods[108, 56, 121, 215] only using 1 point prompt per input. For liver tumor, 10 points are used for all methods to accommodate multiple tumors.

	Public datastes (Dice % / NSD %)			
Methods	Colon tumor	Pancreas tumor	Liver tumor	Kidney tumor
nnU-Net [92]	43.91 / 52.52	41.65 / 62.54	60.10 / 75.41	73.07 / 77.47
3D UX-Net [114]	28.50 / 32.73	34.83 / 52.56	45.54 / 60.67	57.59 / 58.55
Swin-UNETR [202]	35.21 / 42.94	40.57 / 60.05	50.26 / 64.32	65.54 / 72.04
SAM [108]	28.83 / 33.63	24.01 / 26.74	8.56 / 5.97	36.30 / 29.86
3DSAM-adapter [56]	57.32 / 73.65	54.41 / 77.88	56.61 / 69.52	73.78 / 83.86
ProMISe [121]	66.81 / 81.24	57.46 / 79.76	58.78 / 71.52	75.70 / 80.08
SAM-Med3D [215]	54.34 / 78.58	65.61 / 92.40	23.64 / 26.97	76.50 / 88.41
SAM-Med3D-organ	70.75 / 91.03	76.40 / 97.75	66.52 / 77.97	88.20 / 97.80
SAM-Med3D-turbo	73.77 / 94.95	74.87 / 96.43	69.36 / 81.70	89.26 / 98.40
PRISM-plain	67.18 / 85.28	65.73 / 89.51	79.70 / 91.60	85.29 / 93.55
PRISM-ultra	**93.79 / 99.96**	**94.48 / 99.99**	**94.18 / 99.99**	**96.58 / 99.80**

"PRISM-plain" only uses 1 point and no BB. "-ultra" adds the BB, and scribbles.

[56, 121, 215, 108] methods. We used $y'_{i=11}$ as final segmentation for the iterative methods [215] and PRISM. We retrained using the published code of each model, and the pretrained weights were also employed if publicly available. We present two versions of our method. PRISM-plain only uses point prompts, while PRISM-ultra can handle other sparse visual prompts such as boxes and scribbles. All results are generated with seeds.

17.3 Results

Datasets. We use four public 3D CT datasets, including colon (N=126) [4], pancreas (N=281) [4], liver (N=118) [10] and kidney (N=300) [74] tumors. Challenges include large anatomical differences among individuals, varying shapes of target structures, and ambiguous boundaries. Additionally, multiple tumors [74, 10] may occur in a single subject. We adopted the same data split as a prior study [56] for each task with a training/validation/testing split of 0.7/0.1/0.2, and we only use tumor labels to focus on binary segmentation. We resampled to a $1mm$ isotropic resolution, performed intensity clipping based on the 0.5 and 99.5 percentiles of the foreground, and applied Z-score normalization based on foreground voxels. Random zoom and intensity shift are used as data augmentations.

Comparison with state-of-the-art. Table 17.1 compares quantitative segmentation results (See qualitative results in Fig. 17.3). As expected, fully automated methods [92, 202, 114] are unable to deliver satisfactory outcomes for these challenging datasets. Even with visual prompts, SAM [108] struggles with the substantial domain shift between natural and medical images. Furthermore, the interactive segmentation models

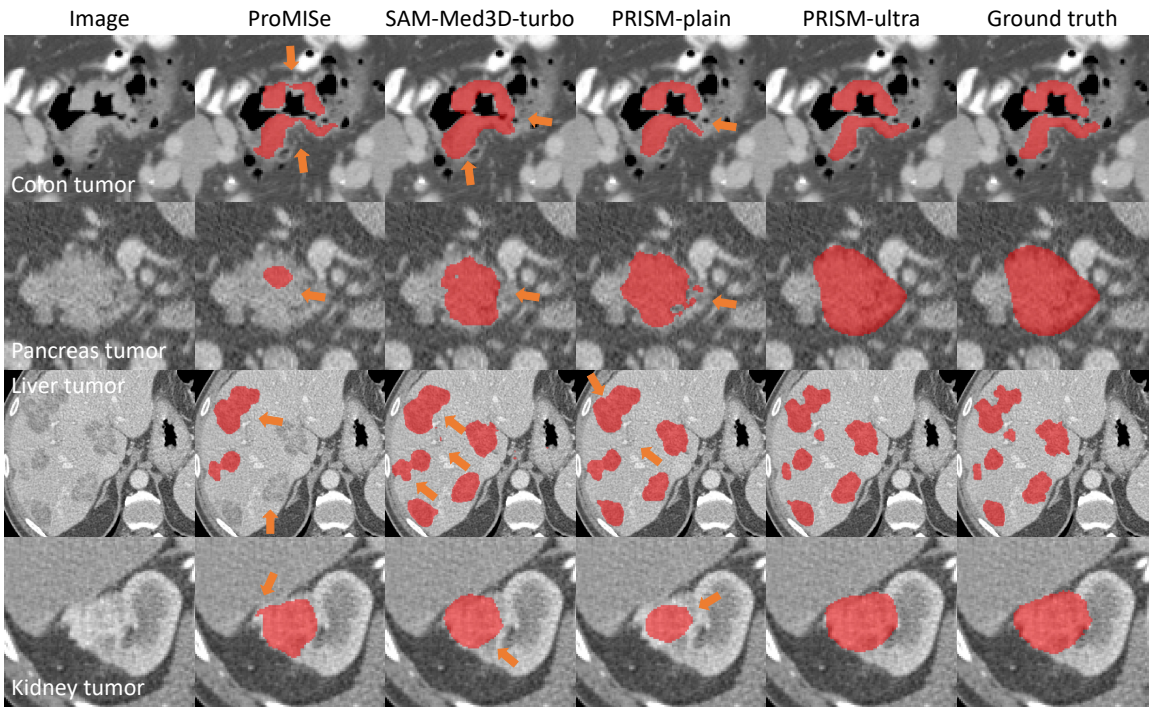


Figure 17.3: Qualitative results of four different tumor segmentation tasks. The orange arrows indicate the major defects.

[56, 121, 215] trained from scratch fail to produce adequate results. Notably, "-organ" and "-turbo" denote the finetuning of the pretrained SAM-Med3D model [215]. "-organ" focuses on organ-specific datasets, while "-turbo" includes these organ datasets (including, in some cases, the test data in their pretraining) alongside a broader range of other medical imaging data. In contrast, excluding the pretrained models, the proposed PRISM-plain demonstrates superior overall Dice performance with only a single point as sparse prompt. Moreover, as more visual prompts (points, BB, and scribbles) are provided, PRISM-ultra delivers outcomes close to human level performance, demonstrating its strength as an effective human-in-loop algorithm.

Analysis of iterative learning performance. Fig. 17.4 shows the performance of PRISM across the iterative learning process. Although PRISM-plain delivers superior outcomes (Table 17.1), it does not have monotonically increasing performance across iterations and can suffer from a decrease in performance at later iterations, except for liver tumor. In practice this would be frustrating to the human user, as more input prompts paradoxically lowers the segmentation accuracy. However, PRISM-ultra presents not just substantial improvements in overall performance, but also a monotonically increasing accuracy and narrower 95% confidence intervals. This highlights the PRISM-ultra as a robust model, and its qualitative results in Fig. 17.5 and Fig. 17.6 illustrate the iterative correction. Initially, the results may not meet expectations due to ambiguous boundaries and anatomical variations. Yet, as iterations advance, the outcomes progressively

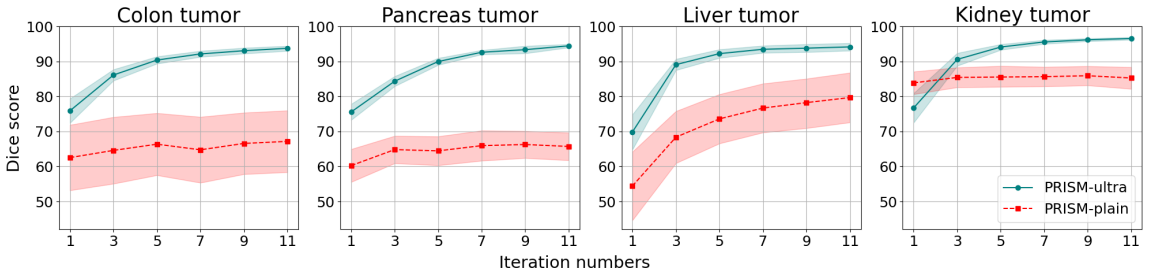


Figure 17.4: Dice score of proposed PRISM on four tumor datasets, where the mean values (lines) and their 95% confidence intervals (shades) are presented.

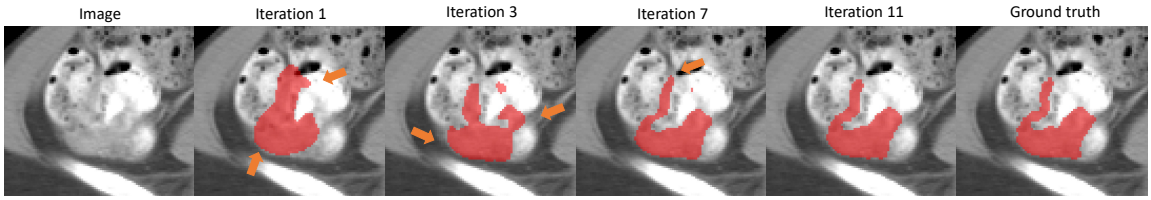


Figure 17.5: Qualitative results of PRISM-ultra for colon tumors characterized by irregular shapes and ambiguous boundaries. The orange arrows indicate the major defects which are corrected in the subsequent iteration. The initial output has noticeable errors that rapidly get corrected in the first few iterations.

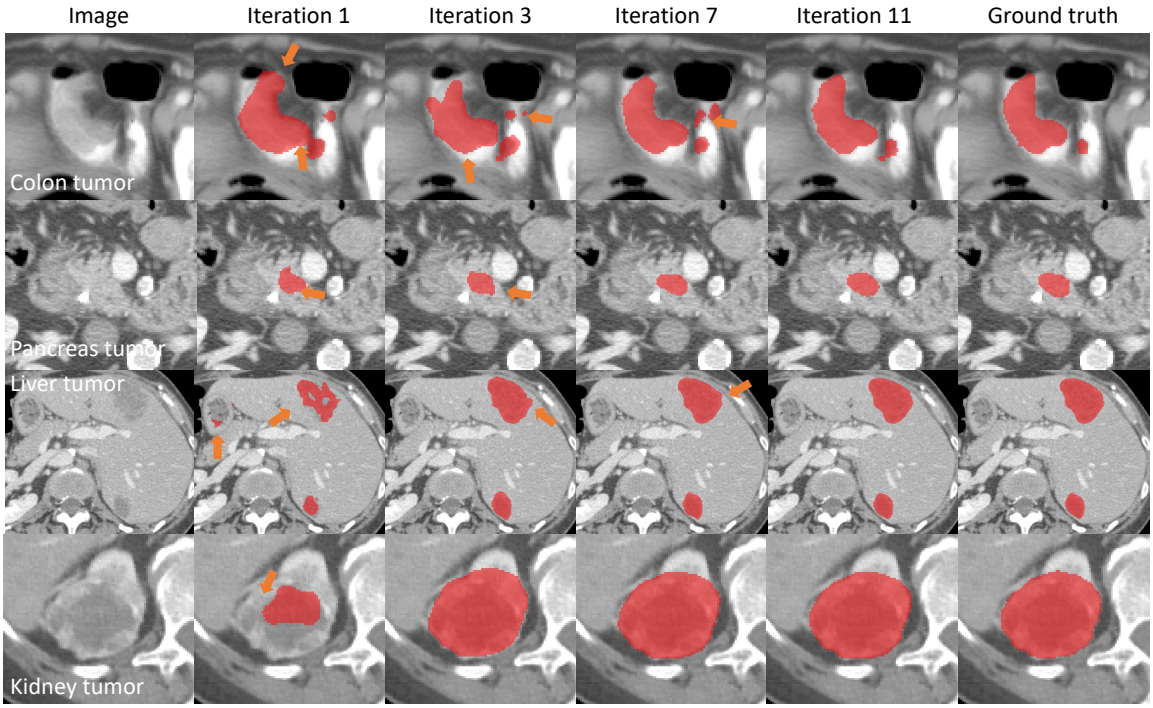


Figure 17.6: Qualitative results of PRISM-ultra across iterations. The orange arrows indicate the major defects which are corrected in the subsequent iteration.

align more closely with the human label. Major corrections occur at the early stages which indicate the efficiency and effectiveness of PRISM-ultra.

Table 17.2: Ablation study on the colon tumor. ConL: confidence learning. PRISM-plain-b: PRISM-plain plus BB. Model configurations are detailed in Table 17.4.

	ViT encoder	CNN encoder	Hybrid	PRISM-plain	PRISM-plain-b	no ConL
Dice	58.38	63.58	66.40	67.18	76.40	73.65

Table 17.3: Prompt analysis for colon tumors. "-1", "-50": number of points used during training. PRISM-basic: PRISM-plain-b with random number of points. "-erode" and "-dilate": modified BB by 5 voxel radius. CorL: corrective learning. PRISM-ultra: PRISM-basic plus test-time scribbles. PRISM-ultra+: PRISM-ultra plus train-time scribbles. Prompt configurations are detailed in Table 17.5.

Test Points	-plain-b-1	-plain-b-50	-basic	no CorL	-erode	-dilate	-ultra	-ultra+
1	76.40	61.33	76.66	77.65	72.99	75.47	93.79	96.47
10	77.71	82.42	83.85	80.12	82.85	83.72	93.95	96.54
50	72.52	89.91	89.36	81.60	89.99	89.42	94.07	96.73
100	69.60	91.30	90.26	81.20	91.06	90.70	94.28	96.93

Analysis of framework. An ablation study for various frameworks on the colon dataset are presented in Table 17.2, and the details of model configurations in Table 17.4. The CNN encoder outperforms the use of a Vision Transformer (ViT) alone. Using a hybrid encoder improves results, while proposed confidence and corrective learning further increase PRISM-plain performance. PRISM-plain is advanced to PRISM-plain-b through the incorporation of a 3D BB, a favored prompt in medical segmentation due to its strong image prior, leading to improved outcomes. For efficiency, drawing a box does not require significantly more effort than point prompt. For effectiveness, points can address the limited flexibility of boxes in making corrections in subsequent iterations. The proposed confidence learning approach also contributes to the performance of the model.

Analysis of number of prompts. We analyze the performance with respect to the number of point prompts in Table 17.3. The compared model configurations are detailed in Table 17.5. We first evaluate the number of prompts used for training. The model plain-b-1 uses 1 prompt for training, plain-b-50 uses 50, and basic uses a random number between 1 and 50. Too few or too many training points lowers performance, while the random selection is the most robust. We also note that corrective learning has a more pronounced contribution when more prompts are used. We next test the sensitivity to BB precision by dilating and eroding the true BB. Either eroding or dilating BB decreases performance compared to PRISM-basic when few points are used, but using more points resolves this issue. Adding test-time scribbles (ultra) and train-time scribbles (ultra+) both improves the performance of PRISM-basic. However, this improvement increases training time, making scribbles less practical for large datasets.

Table 17.4: Detailed settings for architecture and learning strategy in the proposed ablation studies.

	Encoder type			Learning strategy	
Variants	ViT	CNN	ViT + CNN	Confidence learning	Corrective learning
ViT encoder	✓				
CNN encoder		✓			
Hybrid			✓		
PRISM-plain			✓	✓	✓
PRISM-plain-b			✓	✓	✓
no ConL			✓		✓
PRISM-basic			✓	✓	✓
no CorL			✓	✓	
PRISM-ultra			✓	✓	✓

Table 17.5: Detailed prompt setting for the proposed ablation studies. Tr. and T. represent training and test, respectively. The notation [1, 50] indicates the range from which the number of point prompts is randomly sampled during training. The term "varies" refers to the number of points used in tests, which include 1, 10, 50, and 100. "-erode" and "-dilate" denote box sizes that are 5 voxels smaller or larger in each dimension.

	Detailed prompt setting				
Variants	point	point num. (Tr. / T.)	box	box type in T.	scribble usage
ViT encoder	✓	1 / 1			
CNN encoder	✓	1 / 1			
PRISM-plain	✓	1 / 1			
PRISM-plain-b	✓	1 / 1	✓	tight	
no ConL	✓	1 / 1	✓	tight	
-plain-b-1	✓	1 / 1	✓	tight	
-plain-b-50	✓	50 / varies	✓	tight	
PRISM-basic	✓	[1, 50] / 1	✓	tight	
no CorL	✓	[1, 50] / 1	✓	tight	
-erode	✓	[1, 50] / varies	✓	undersized	
-dilate	✓	[1, 50] / varies	✓	oversized	
PRISM-ultra	✓	[1, 50] / varies	✓	tight	T.
PRISM-ultra+	✓	[1, 50] / varies	✓	tight	Tr. and T.

17.4 Conclusion

In this study, we propose PRISM, a **P**romptable **R**obust **I**nteractive **S**egmentation **M**odel for 3D medical images. It supports various visual prompts and applies diverse learning strategies to achieve performance comparable to that of human raters. This is demonstrated across four challenging tumor segmentation tasks, where existing state-of-the-art automatic and interactive segmentation approaches do not meet the expected standards. Future work will include adaptation of pretrained models, and learning from datasets with sparse annotations.

Part IV

Conclusion and Future work

CHAPTER 18

Conclusion

18.1 Summary

In this dissertation, I focused on developing data-driven and data-centric deep learning methods aimed at accurate and robust medical image segmentation, which is crucial for improving the accuracy of clinicians and radiologists in detecting and diagnosing diseases through medical image analysis. Specifically, medical image segmentation facilitates critical tasks such as tumor segmentation and quantification of target regions, thereby directly influencing treatment planning and monitoring. Additionally, subcortical or whole-brain segmentation plays an important role in understanding disease progression, quantifying various biomarkers, and contributing to neuroanatomical and pathological studies.

However, there are two common practical limitations in deep learning for medical image segmentation:

1. Label availability: Medical image datasets often require expert annotation, which can be both time-consuming and expensive. The deep learning model requires more data to achieve superior performance. Moreover, the scarcity of labeled data in less common or more complex conditions limits the ability of models to learn diverse patterns, particularly in rare diseases or unique patient populations.

2. Domain shift: Deep learning models trained on data from specific imaging scanners and protocols may not perform well when applied to data from different sources, which is known as domain shift. This presents a major challenge in deploying models in real-world clinical settings, where differences in scanners and protocols are inevitable. Domain shift can occur within the same modality or across modalities, caused by discrepancies between the data distributions in the training set (source domain) and the test set (target domain).

To tackle these practical issues, I began with the development of a generalizable model for a data-driven method, aiming for superior segmentation performance. Specifically, for rare diseases with limited labeled datasets, I developed a single-path network to segment subcortical structures with enhanced generalizability from healthy control subjects to Huntington's disease (HD) patients (Chapter 5). Following this, I proposed a U-shaped network to further improve segmentation performance (Chapter 6), and conducted a comprehensive extension study (Chapter 7) using large-scale datasets with more comparative methods and detailed experimental results. Moreover, I designed a longitudinal framework to achieve accurate subcortical segmentation of HD patients and better model the relationship between segmentation and clinical measurements (Chapter 8). Chapter 9 presents a variant of the proposed U-shaped network for skull-stripping, validated on patients

with rare Aicardi-Goutières syndrome (AGS) and HD. To overcome the limited receptive field of convolutional neural networks (CNNs), I developed a hybrid network, CATS, which combines vision transformer and CNN paths to better capture both global and local information (Chapter 10). CATS was evaluated on three public datasets with large individual variations, showing superior performance. Furthermore, I improved CATS by replacing the vision transformer with a Swin-transformer to improve its performance even further (Chapter 11).

Additionally, I developed data-centric methods to improve the unreliable performance on unseen data. In Chapter 12, I developed an unsupervised domain adaptation (UDA) method using only labeled source T1-weighted MRI data to segment vestibular schwannoma (VS) and cochlea from T2-weighted MRI scans. Unlike most existing UDA research that focuses on single-domain mapping, I developed a unified model across multiple imaging sites, as detailed in Chapter 13. This multi-domain mapping method was evaluated using a large-scale dataset with 5,150 cases (14,191 samples) under various imaging protocols with 34 different b-values, where each b-value is considered a unique domain, to segment and detect prostate lesions from diffusion-weighted MRIs (DWI). Practically, source domain data may not be accessible due to privacy issues across different imaging sites, which limits the feasibility of UDA methods requiring access to source domain data. For this scenario, I developed a test-time adaptation method that does not require any source domain data for adaptation (Chapter 14). Instead, pretrained models from the source domain are used, which are easier to transfer. This study was validated on healthy controls, adult HD patients, and pediatric AGS patients, where domain shifts caused by multiple scanners and severe geometric variations are present.

Furthermore, I developed data-centric methods for interactive segmentation models with prompt engineering to improve performance and find optimal prompt configurations for inference. Specifically, the provided prompts could improve the robustness of the model to domain shifts between different medical imaging modalities. Through continuous learning from expert feedback, the model adapts to the unique characteristics of each modality, ensuring reliable and accurate segmentation across various medical imaging applications. The method also effectively tackles challenging segmentation tasks within the same modality. In Chapter 15, I proposed ProMISe, a Prompt-driven 3D Medical Image Segmentation model using only a single point prompt to leverage knowledge from the pretrained Segment Anything Model (SAM). This parameter-efficient model adaptation employs lightweight adapters to effectively transfer the pretrained weights from natural image domains to medical image domains. ProMISe was validated on two tumor segmentation tasks with superior performance. Additionally, I assessed the test-time variability of ProMISe with diverse point prompts used to segment colon tumors (Chapter 16), where the variability can be caused by expertise and ambiguous boundaries. Experimental findings suggest a simple strategy for test-time prompt selection that significantly improves segmentation results over random prompt placement. Lastly, in Chapter 17, I devel-

oped PRISM, a Promptable Robust Interactive Segmentation Model for 3D medical images. PRISM mimics a "human-in-the-loop" iterative correction process with iterative and confidence learning, allowing for continuous improvements and robust performance. Specifically, PRISM uses various types of visual prompts (points, boxes, scribbles, masks) to progressively refine outcomes to match inter-rater variability. The effectiveness of the proposed PRISM has been demonstrated on four public datasets for the segmentation of the colon, pancreas, liver, and kidney, significantly outperforming all compared methods.

18.2 Summary of Contributions of Data-driven Medical Image Segmentation

- Technical:
  - I developed a single-path network and explored several of the modifications that have indeed significantly improved the segmentation accuracy of subcortical structures in diagnosed HD patients, most prominently in the caudate, one of the most affected structures in HD (Chapter 5).
  - I proposed a 2-stage cascaded framework for subcortical segmentation, with a 3D U-shaped network at each stage. This framework has a dramatic improvement of accuracy for the caudate and the putamen, which are the most atrophied subcortical structures in HD (Chapter 6). In addition, the proposed framework also benefits other segmentation models for subcortical segmentation (Chapter 7).
  - I developed a longitudinal segmentation method for HD subjects. With the longitudinal information, the model learns the relationship between the scans and achieves superior segmentation performance with higher accuracy and better consistency for the considered subcortical structures (Chapter 8).
  - I proposed a deep learning framework for human brain extraction with superior performance on HD and AGS subjects (Chapter 9).
  - I proposed CATS and CATS *v2*, a convolutional neural network with a transformer as an independent encoder. The transformer can complement the CNN by modeling long-range dependencies and capturing low-level details (Chapter 10 and 11).
- Clinical:
  - Experimental findings from Chapter 5 and Chapter 6 indicate that my methods have better generalizability not only to unseen healthy subjects, but also from healthy controls to an HD population.
  - The proposed approach in Chapter 6 is designed for practical clinical use among a large HD

population with robust segmentation. It is evaluated on a large-scale dataset with superior segmentation performance (Chapter 7).

  - The longitudinal segmentation method produced stronger correlations between volume loss and TMS decline in HD patients for most of the comparisons in the caudate and putamen, the most affected structures in HD (Chapter 8).

  - I presented the first automated method for reliably segmenting the brain in AGS patients, which allows the total brain volume to be automatically estimated and is an enabling step for further MRI analysis in AGS cohorts (Chapter 9).

18.3 Future Work of Data-driven Medical Image Segmentation

Medical images can vary widely due to differences in imaging protocols, patient demographics, and disease characteristics. A highly generalizable model can adapt to these variations and maintain high performance across different types of images and clinical tasks. Additionally, collecting and annotating medical images for training can be expensive and time-consuming. Generalizable models can learn effectively from a smaller amount of data, making them more data-efficient. As a result, I have proposed generalizable models for robust medical image segmentation. For future directions, I will develop a parameter-efficient network that maintains similar segmentation performance without significant loss. This network is intended to serve as the backbone for various segmentation tasks. Furthermore, this parameter-efficient network will be tested on various imaging modalities to demonstrate its generalizability. Additionally, this approach supports the future of real-time segmentation by ensuring that our models can be deployed on devices with limited computational resources, thus enhancing accessibility and functionality in diverse clinical environments.

18.4 Summary of Contributions in Data-Centric Medical Image Segmentation for Unsupervised Domain Adaptation

- Technical:

  - I developed a UDA framework for VS and cochlear segmentation. There are two parts in our framework: synthesis and segmentation, where the segmentation model is trained on the image generated by the synthesis model (Chapter 12).

  - I further proposed a novel UDA framework wth a unified model for multi-domain mapping instead of multiple networks being trained. I proposed and employed a dynamic filter to leverage domain information for robust performance on prostate detection and segmentation (Chapter 13).

  - I developed a test-time adaptation (TTA) method in the presence of domain shift. The proposed framework only needs pre-trained CNNs in the source domain, and the target image itself. It aligns the target image on both image and latent feature levels to the source domain during the test-time (Chapter 14).

- Clinical:

  - The proposed UDA framework addressed the lack of labeled medical data in the target domain (Chapter 13).
  - The proposed UDA with unified model is the first large-scale study exploring the impact of b-value properties on ADC and DWI b-2000 images with the aim of improving detection outcomes for a multi-domain scenario. The feasibility for practical use is successfully validated with a dataset collected from 5,150 cases (Chapter 14).
  - The proposed TTA framework tackles the data practical privacy issue. Moreover, it has been evaluated to handle the domain shift, where intensity and geometry shifts appear between source and target domains for the pediatric AGS dataset. (Chapter 14).

18.5 Future Work of Data-centric Medical Image Segmentation for Unsupervised Domain Adaptation

In my work on unsupervised multi-domain adaptation with a unified model, I introduced the dynamic filter, which can be treated as a domain indicator and plugged into any generator to leverage meta-information. The proposed dynamic filter generates conditional parameters based on the corresponding meta-information to differentiate domains. However, it is currently limited by using only b-values from meta-information to provide domain information. This constraint may contribute to suboptimal performance, especially when high b-values closely resemble those of the reference domain. Future work will involve incorporating additional meta-information, such as field strength, sequence selection, and the type of scanner, to more comprehensively quantify domain shifts resulting from diverse acquisition protocols.

Clinically, accurate detection of prostate lesions is a challenging task even for experienced experts, and this is a common practical issue in clinical settings. As one of the early studies in this area, I demonstrated the feasibility of using generated images for prostate lesion detection with a large-scale dataset. Reducing inter-rater variability is indeed an important evaluation step for this proposed method. Currently, we have not examined the agreement of multiple experts on our proposed method. This is a valuable direction for future work. Additionally, another aspect of future work will involve testing whether the proposed method can correct the predictions of radiologists, particularly those with less experience.

In addition, I will utilize current publicly available datasets and pretrained models for multi-domain mapping from a single source domain. This approach will enhance the diversity of data to achieve robust segmentation. Various imaging modalities, including 3D MRI, CT, and 2D endoscopy images, will be used for development.

18.6 Summary of Contributions in Data-Centric Medical Image Segmentation for Interactive Segmentation

- Technical:
  - I proposed ProMISe for interactive medical segmentation by using the pretrained foundation model, SAM, from the image domain. ProMISe leverages plug-and-play light-weight adapters to achieve parameter-efficient model adaptation. The role of proposed boundary-aware loss for robust segmentation is demonstrated (Chapter 15).
  - I proposed PRISM, a Promptable and Robust Interactive Segmentation Model with visual prompts. It accepts various visual prompts (points, boxes, scribbles, and masks). PRISM is trained with iterative and confidence learning, allowing for continuous improvements and robust performance. The corrective refinement network in PRISM further improves the segmentation performance (Chapter 17).
- Clinical:
  - I proposed ProMISe, an interactive segmentation method for challenging tumor segmentation tasks. Compared to the automatic deep learning segmentation methods, it delivers robust segmentation by providing prompts at inference (Chapter 15). In addition, I assessed the test-time variability for the proposed ProMISe with diverse point prompts. The assessments are based on three considerations: (1) benefits of additional prompts, (2) effects of prompt placement, and (3) strategies for optimal prompt selection. Experiments suggest an optimal strategy for prompt selection during test-time (Chapter 16).
  - PRISM aims for human-level performance as an interactive segmentation model with human-in-loop capabilities. It uses prompts to progressively refine results until they closely align with inter-rater variability. Offering versatility, PRISM effectively tackles the diverse challenges of medical imaging segmentation, ensuring precision through intuitive user interaction. With each input, its performance incrementally advances toward expert-level accuracy. (Chapter 17). PRISM is designed as a labeling tool. It accelerates data labeling while maintaining high accuracy, which

is crucial for building robust datasets. Its efficiency and precision directly benefit clinicians by providing more accurate segmentation results for diagnostics and treatment planning.

18.7 Future Work of Data-centric Medical Image Segmentation for Interactive Segmentation

Importantly, a robust interactive segmentation model should respond effectively and efficiently to visual prompts provided by users with minimal interactions. This is vital in clinical settings, where time and precision are of the essence for patient diagnosis and treatment. The proposed PRISM model has already demonstrated its effectiveness in four challenging tumor segmentation tasks on CT images, highlighting its potential in practical applications. However, for the model to be more widely applicable, it is crucial to expand its capabilities to cover a broader range of imaging modalities. Incorporating MRI and ultrasound is essential to make PRISM versatile across different imaging modalities, increasing its generalizability and utility in diverse medical scenarios. By testing PRISM on these additional imaging modalities, the model can be developed into a foundational image segmentation tool that clinicians can rely on for various tasks.

PRISM is designed as a hybrid network, utilizing complementary CNN and transformer encoders for segmentation. This hybrid architecture allows the model to leverage the strengths of both approaches, improving its segmentation accuracy. The flexibility of this design ensures that PRISM can adapt to existing foundational image models in the medical domain. Future work will focus on adapting publicly available pretrained models to improve segmentation performance further. Leveraging pretrained models allows PRISM to incorporate knowledge from other data or domains, accelerating its learning process and improving its ability to handle more complex tasks.

Given the scarcity of labeled datasets in the medical domain, future work will involve learning from datasets with sparse annotations. This approach is crucial because acquiring labeled data is expensive and time-consuming, often requiring expert input. By learning from datasets with sparse annotations, PRISM can reduce its dependence on fully labeled data and remain effective even with limited resources. Furthermore, it is essential to ensure that PRISM remains efficient, as efficiency directly impacts its practicality in real-world clinical environments. Additionally, further work will focus on finding the optimal solution for prompt selection, simplifying user interaction, and making the model more accessible for clinicians to use effectively.

18.8 Conclusion

In this dissertation, I focus on developing and innovating deep learning methods for medical image segmentation, transitioning from data-driven to data-centric approaches. The shift highlights the crucial role of data quality in enhancing the effectiveness of segmentation techniques. While data-driven strategies seek to optimize convolutional neural networks (CNNs) using existing datasets, data-centric methods emphasize the

importance of data quality and diversity for improving segmentation model performance.

The dissertation presents both data-driven and data-centric medical image segmentation methods with superior performance. Furthermore, it emphasizes the importance of this transition towards accurate medical image segmentation in practical applications. The dissertation encompasses a range of segmentation tasks across various imaging modalities and populations, demonstrating robust performance. Importantly, it suggests that the proposed methods not only represent a significant advancement but also establish a promising direction for future research. This approach has the potential to tackle new challenges in medical image segmentation.

References

[1] Ahmed, H. U., Bosaily, A. E.-S., Brown, L. C., Gabe, R., Kaplan, R., Parmar, M. K., Collaco-Moraes, Y., Ward, K., Hindley, R. G., Freeman, A., et al. (2017). Diagnostic accuracy of multi-parametric mri and trus biopsy in prostate cancer (promis): a paired validating confirmatory study. *The Lancet*, 389(10071):815–822.

[2] Alharbi, Y., Smith, N., and Wonka, P. (2019). Latent filter scaling for multimodal unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1458–1466.

[3] Andrew, S. E., Paul Goldberg, Y., Kremer, B., Telenius, H., Theilmann, J., Adam, S., Starr, E., Squitieri, F., Lin, B., Kalchman, M. A., et al. (1993). The relationship between trinucleotide (cag) repeat length and clinical features of huntington's disease. *Nature genetics*, 4(4):398–403.

[4] Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B. A., Litjens, G., Menze, B., Ronneberger, O., Summers, R. M., et al. (2022). The medical segmentation decathlon. *Nature communications*, 13(1):4128.

[5] Bai, W., Suzuki, H., Qin, C., Tarroni, G., Oktay, O., Matthews, P. M., and Rueckert, D. (2018). Recurrent neural networks for aortic image sequence segmentation with sparse annotations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 586–594. Springer.

[6] Barth, B. K., De Visschere, P. J., Cornelius, A., Nicolau, C., Vargas, H. A., Eberli, D., and Donati, O. F. (2017). Detection of clinically significant prostate cancer: short dual–pulse sequence versus standard multiparametric mr imaging—a multireader study. *Radiology*, 284(3):725.

[7] Basnet, R., Ahmad, M. O., and Swamy, M. (2021). A deep dense residual network with reduced parameters for volumetric brain tissue segmentation from mr images. *Biomedical Signal Processing and Control*, 70:103063.

[8] Bates, G. P., Dorsey, R., Gusella, J. F., Hayden, M. R., Kay, C., Leavitt, B. R., Nance, M., Ross, C. A., Scahill, R. I., Wetzel, R., et al. (2015). Huntington disease. *Nature reviews Disease primers*, 1(1):1–21.

[9] Beers, A., Chang, K., Brown, J., Sartor, E., Mammen, C., Gerstner, E., Rosen, B., and Kalpathy-Cramer, J. (2017). Sequential 3d u-nets for biologically-informed brain tumor segmentation. *arXiv preprint arXiv:1709.02967*.

[10] Bilic, P., Christ, P., Li, H. B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G. E. H., Chartrand, G., et al. (2023). The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84:102680.

[11] Billot, B., Greve, D., Van Leemput, K., Fischl, B., Iglesias, J. E., and Dalca, A. V. (2020). A learning strategy for contrast-agnostic mri segmentation. *arXiv preprint arXiv:2003.01995*.

[12] Birenbaum, A. and Greenspan, H. (2016). Longitudinal multiple sclerosis lesion segmentation using multi-view convolutional neural networks. In *Deep Learning and Data Labeling for Medical Applications*, pages 58–67. Springer.

[13] Bosma, J. S., Saha, A., Hosseinzadeh, M., Slootweg, I., de Rooij, M., and Huisman, H. (2023). Semisupervised learning with report-guided pseudo labels for deep learning–based prostate cancer detection using biparametric mri. *Radiology: Artificial Intelligence*, 5(5):e230031.

[14] Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., and Krishnan, D. (2017). Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[15] Boykov, Y. and Funka-Lea, G. (2006). Graph cuts and efficient nd image segmentation. *International journal of computer vision*, 70(2):109–131.

[16] Bruzelius, E., Scarpa, J., Zhao, Y., Basu, S., Faghmous, J. H., and Baum, A. (2019). Huntington's disease in the united states: variation by demographic and socioeconomic factors. *Movement Disorders*, 34(6):858–865.

[17] Cai, J., Zhang, Z., Cui, L., Zheng, Y., and Yang, L. (2019). Towards cross-modal organ translation and segmentation: A cycle-and shape-consistent generative adversarial network. *Medical image analysis*, 52:174–184.

[18] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., and Wang, M. (2021). Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*.

[19] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., and Wang, M. (2022). Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer.

[20] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., and Wang, M. (2023). Swin-unet: Unet-like pure transformer for medical image segmentation. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, pages 205–218. Springer.

[21] Cass, N. D., Lindquist, N. R., Zhu, Q., Li, H., Oguz, I., and Tawfik, K. O. (2022). Machine learning for automated calculation of vestibular schwannoma volumes. *Otology & Neurotology*, 43(10):1252–1256.

[22] Castillo, L. S., Daza, L. A., Rivera, L. C., and Arbeláez, P. (2017). Volumetric multimodality neural network for brain tumor segmentation. In *13th international conference on medical information processing and analysis*, volume 10572, page 105720E. International Society for Optics and Photonics.

[23] Chalana, V. and Kim, Y. (1997). A methodology for evaluation of boundary detection algorithms on medical images. *IEEE Transactions on medical imaging*, 16(5):642–652.

[24] Chang, P. D. (2016). Fully convolutional deep residual neural networks for brain tumor segmentation. In *International workshop on Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries*, pages 108–118. Springer.

[25] Chen, C., Dou, Q., Chen, H., Qin, J., and Heng, P. A. (2020). Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation. *IEEE transactions on medical imaging*, 39(7):2494–2505.

[26] Chen, C., Miao, J., Wu, D., Yan, Z., Kim, S., Hu, J., Zhong, A., Liu, Z., Sun, L., Li, X., et al. (2023a). Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation. *arXiv preprint arXiv:2309.08842*.

[27] Chen, C., Ouyang, C., Tarroni, G., Schlemper, J., Qiu, H., Bai, W., and Rueckert, D. (2019). Unsupervised multi-modal style transfer for cardiac mr segmentation. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 209–219. Springer.

[28] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., and Zhou, Y. (2021). Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*.

[29] Chen, S., Ge, C., Tong, Z., Wang, J., Song, Y., Wang, J., and Luo, P. (2022). Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678.

[30] Chen, X., Cheung, Y. S. J., Lim, S.-N., and Zhao, H. (2023b). Scribbleseg: Scribble-based interactive image segmentation. *arXiv preprint arXiv:2303.11320*.

[31] Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., et al. (2023). Sam-med2d. *arXiv preprint arXiv:2308.16184*.

[32] Chiou, E., Giganti, F., Punwani, S., Kokkinos, I., and Panagiotaki, E. (2021). Unsupervised domain adaptation with semantic consistency across heterogeneous modalities for mri prostate lesion segmentation. In *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*, pages 90–100. Springer.

[33] Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., and Haworth, A. (2021). A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, 65(5):545–563.

[34] Cho, S., Jang, H., Tan, J. W., and Jeong, W.-K. (2021). Deepscribble: interactive pathology image segmentation using deep neural networks with scribbles. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 761–765. IEEE.

[35] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer.

[36] Deng, G., Zou, K., Ren, K., Wang, M., Yuan, X., Ying, S., and Fu, H. (2023). Sam-u: Multi-box prompts triggered uncertainty estimation for reliable sam in medical image. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 368–377. Springer.

[37] DeSilvio, T., Moroianu, S., Bhattacharya, I., Seetharaman, A., Sonn, G., and Rusu, M. (2021). Intensity normalization of prostate mris using conditional generative adversarial networks for cancer detection. In *Medical Imaging 2021: Computer-Aided Diagnosis*, volume 11597, pages 121–126. SPIE.

[38] Dolz, J., Desrosiers, C., and Ayed, I. B. (2018). 3d fully convolutional networks for subcortical segmentation in mri: A large-scale study. *NeuroImage*, 170:456–470.

[39] Dong, H., Yang, G., Liu, F., Mo, Y., and Guo, Y. (2017). Automatic brain tumor detection and segmentation using u-net based fully convolutional networks. In *annual conference on medical image understanding and analysis*, pages 506–517. Springer.

[40] Dorent, R., Joutard, S., Shapey, J., Bisdas, S., Kitchen, N., Bradford, R., Saeed, S., Modat, M., Ourselin, S., and Vercauteren, T. (2020). Scribble-based domain adaptation via co-segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 479–489. Springer.

[41] Dorent, R., Kujawa, A., Ivory, M., Bakas, S., Rieke, N., Joutard, S., Glocker, B., Cardoso, J., Modat, M., Batmanghelich, K., et al. (2022). Crossmoda 2021 challenge: Benchmark of cross-modality domain adaptation techniques for vestibular schwnannoma and cochlea segmentation. *arXiv preprint arXiv:2201.02831*.

[42] Dorent, R., Kujawa, A., Ivory, M., Bakas, S., Rieke, N., Joutard, S., Glocker, B., Cardoso, J., Modat, M., Batmanghelich, K., et al. (2023). Crossmoda 2021 challenge: Benchmark of cross-modality domain adaptation techniques for vestibular schwannoma and cochlea segmentation. *Medical Image Analysis*, 83:102628.

[43] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[44] Dou, Q., Coelho de Castro, D., Kamnitsas, K., and Glocker, B. (2019). Domain generalization via model-agnostic learning of semantic features. *Advances in Neural Information Processing Systems*, 32.

[45] Dou, Q., Ouyang, C., Chen, C., Chen, H., and Heng, P.-A. (2018). Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss. *arXiv preprint arXiv:1804.10916*.

[46] Duran, A., Dussert, G., Rouvière, O., Jaouen, T., Jodoin, P.-M., and Lartizien, C. (2022). Prostattentionnet: A deep attention model for prostate cancer segmentation by aggressiveness in mri scans. *Medical Image Analysis*, 77:102347.

[47] Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., Van Der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., et al. (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3):341–355.

[48] Galdran, A., Carneiro, G., and González Ballester, M. A. (2021). Balanced-mixup for highly imbalanced medical image classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pages 323–333. Springer.

[49] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., and Lempitsky, V. (2016). Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35.

[50] Gao, Y., Li, J., Xu, H., Wang, M., Liu, C., Cheng, Y., Li, M., Yang, J., and Li, X. (2019). A multi-view pyramid network for skull stripping on neonatal t1-weighted mri. *Magnetic resonance imaging*, 63:70–79.

[51] Gao, Y., Phillips, J. M., Zheng, Y., Min, R., Fletcher, P. T., and Gerig, G. (2018). Fully convolutional structured lstm networks for joint 4d medical image segmentation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1104–1108. IEEE.

[52] Gao, Y., Xia, W., Hu, D., and Gao, X. (2023). Desam: Decoupling segment anything model for generalizable medical image segmentation. *arXiv preprint arXiv:2306.00499*.

[53] Gao, Y., Zhou, M., and Metaxas, D. N. (2021). Utnet: a hybrid transformer architecture for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 61–71. Springer.

[54] Georgiou-Karistianis, N., Scahill, R., Tabrizi, S. J., Squitieri, F., and Aylward, E. (2013). Structural mri in huntington's disease and recommendations for its potential use in clinical trials. *Neuroscience & Biobehavioral Reviews*, 37(3):480–490.

[55] Ghaffari, M., Sowmya, A., and Oliver, R. (2019). Automated brain tumor segmentation using multimodal brain scans: a survey based on models submitted to the brats 2012–2018 challenges. *IEEE reviews in biomedical engineering*, 13:156–168.

[56] Gong, S., Zhong, Y., Ma, W., Li, J., Wang, Z., Zhang, J., Heng, P.-A., and Dou, Q. (2023). 3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable medical image segmentation. *arXiv preprint arXiv:2306.13465*.

[57] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.

[58] Graves, A. (2013). Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*.

[59] Grebenisan, A., Sedghi, A., Izard, J., Siemens, R., Menard, A., and Mousavi, P. (2021). Spatial decomposition for robust domain adaptation in prostate cancer detection. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1218–1222. IEEE.

[60] Guan, H. and Liu, M. (2021). Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185.

[61] Haarburger, C., Müller-Franzes, G., Weninger, L., Kuhl, C., Truhn, D., and Merhof, D. (2020). Radiomics feature reproducibility under inter-rater variability in segmentations of ct images. *Scientific reports*, 10(1):12688.

[62] Han, S., Carass, A., He, Y., and Prince, J. L. (2020). Automatic cerebellum anatomical parcellation using u-net with locally constrained optimization. *NeuroImage*, page 116819.

[63] Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H. R., and Xu, D. (2022a). Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part I*, pages 272–284. Springer.

[64] Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H. R., and Xu, D. (2022b). Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584.

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

[66] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034.

[67] He, K., Zhang, X., Ren, S., and Sun, J. (2016a). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[68] He, K., Zhang, X., Ren, S., and Sun, J. (2016b). Deep residual learning for image recognition. In *CVPR*.

[69] He, S., Bao, R., Li, J., Stout, J., Bjornerud, A., Grant, P. E., and Ou, Y. (2023). Computer-vision benchmark segment-anything model (sam) in medical images: Accuracy in 12 datasets.

[70] He, Y., Carass, A., Liu, Y., Filippatou, A., Jedynak, B. M., Solomon, S. D., Saidha, S., Calabresi, P. A., and Prince, J. L. (2020). Segmenting retinal oct images with inter-b-scan and longitudinal information. In *Medical Imaging 2020: Image Processing*, volume 11313, page 113133C. International Society for Optics and Photonics.

[71] He, Y., Carass, A., Liu, Y., Jedynak, B. M., Solomon, S. D., Saidha, S., Calabresi, P. A., and Prince, J. L. (2019). Fully convolutional boundary regression for retina oct segmentation. In *MICCAI*, pages 120–128. Springer.

[72] He, Y., Carass, A., Zuo, L., Dewey, B. E., and Prince, J. L. (2021). Autoencoder based self-supervised test-time adaptation for medical image analysis. *Medical image analysis*, 72:102136.

[73] Heimann, T. and Meinzer, H.-P. (2009). Statistical shape models for 3d medical image segmentation: a review. *Medical image analysis*, 13(4):543–563.

[74] Heller, N., Isensee, F., Maier-Hein, K. H., Hou, X., Xie, C., Li, F., Nan, Y., Mu, G., Lin, Z., Han, M., et al. (2021). The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical image analysis*, 67:101821.

[75] Henschel, L., Conjeti, S., Estrada, S., Diers, K., Fischl, B., and Reuter, M. (2020). Fastsurfer-a fast and accurate deep learning based neuroimaging pipeline. *NeuroImage*, 219:117012.

[76] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

[77] Hou, B., Kang, G., Xu, X., and Hu, C. (2019). Cross attention densely connected networks for multiple sclerosis lesion segmentation. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2356–2361. IEEE.

[78] Hu, D., Cui, C., Li, H., Larson, K. E., Tao, Y. K., and Oguz, I. (2021a). Life: a generalizable autodidactic pipeline for 3d oct-a vessel segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 514–524. Springer.

[79] Hu, D., Li, H., Liu, H., and Oguz, I. (2022). Domain generalization for retinal vessel segmentation with vector field transformer. In *International Conference on Medical Imaging with Deep Learning*, pages 552–564. PMLR.

[80] Hu, D., Li, H., Liu, H., and Oguz, I. (2024). Domain generalization for retinal vessel segmentation via hessian-based vector field. *Medical Image Analysis*, page 103164.

[81] Hu, D., Li, H., Liu, H., Yao, X., Wang, J., and Oguz, I. (2023). Map: Domain generalization via m eta-learning on a natomy-consistent p seudo-modalities. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 182–192. Springer.

[82] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021b). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

[83] Hu, L., Zhou, D. W., Fu, C. X., Benkert, T., Xiao, Y. F., Wei, L. M., and Zhao, J. G. (2021c). Calculation of apparent diffusion coefficients in prostate cancer using deep learning algorithms: a pilot study. *Frontiers in Oncology*, page 3404.

[84] Hu, L., Zhou, D.-w., Zha, Y.-f., Li, L., He, H., Xu, W.-h., Qian, L., Zhang, Y.-k., Fu, C.-x., Hu, H., et al. (2021d). Synthesizing high-b-value diffusion–weighted imaging of the prostate using generative adversarial networks. *Radiology: Artificial Intelligence*, 3(5).

[85] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

[86] Huang, X., Liu, M.-Y., Belongie, S., and Kautz, J. (2018). Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189.

[87] Huo, Y., Xu, Z., Bao, S., Assad, A., Abramson, R. G., and Landman, B. A. (2018a). Adversarial synthesis learning enables segmentation without target modality ground truth. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 1217–1220. IEEE.

[88] Huo, Y., Xu, Z., Moon, H., Bao, S., Assad, A., Moyo, T. K., Savona, M. R., Abramson, R. G., and Landman, B. A. (2018b). Synseg-net: Synthetic segmentation without target modality ground truth. *IEEE transactions on medical imaging*, 38(4):1016–1025.

[89] Huo, Y., Xu, Z., Xiong, Y., Aboud, K., Parvathaneni, P., Bao, S., Bermudez, C., Resnick, S. M., Cutting, L. E., and Landman, B. A. (2019). 3d whole brain segmentation using spatially localized atlas network tiles. *NeuroImage*, 194:105–119.

[90] Iglesias, J. E., Liu, C.-Y., Thompson, P. M., and Tu, Z. (2011). Robust brain extraction across datasets and comparison with publicly available methods. *IEEE transactions on medical imaging*, 30(9):1617–1634.

[91] Iglesias, J. E. and Sabuncu, M. R. (2015). Multi-atlas segmentation of biomedical images: a survey. *Medical image analysis*, 24(1):205–219.

[92] Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., and Maier-Hein, K. H. (2021). nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211.

[93] Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., and Maier-Hein, K. H. (2017). Brain tumor segmentation and radiomics survival prediction: Contribution to the brats 2017 challenge. In *International MICCAI Brainlesion Workshop*, pages 287–297. Springer.

[94] Isensee, F., Schell, M., Pflueger, I., Brugnara, G., Bonekamp, D., Neuberger, U., Wick, A., Schlemmer, H.-P., Heiland, S., Wick, W., et al. (2019). Automated brain extraction of multisequence mri using artificial neural networks. *Human brain mapping*, 40(17):4952–4964.

[95] Islam, M., Vibashan, V., Jose, V. J. M., Wijethilake, N., Utkarsh, U., and Ren, H. (2019). Brain tumor segmentation and survival prediction using 3d attention unet. In *International MICCAI Brainlesion Workshop*, pages 262–272. Springer.

[96] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134.

[97] Jang, W.-D. and Kim, C.-S. (2019). Interactive image segmentation via backpropagating refinement scheme. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5297–5306.

[98] Jégou, S., Drozdzal, M., Vazquez, D., Romero, A., and Bengio, Y. (2017). The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 11–19.

[99] Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S., Sinclair, M., Pawlowski, N., Rajchl, M., Lee, M., Kainz, B., Rueckert, D., et al. (2017a). Ensembles of multiple models and architectures for robust brain tumour segmentation. In *International MICCAI brainlesion workshop*, pages 450–462. Springer.

[100] Kamnitsas, K., Ferrante, E., Parisot, S., Ledig, C., Nori, A. V., Criminisi, A., Rueckert, D., and Glocker, B. (2016). Deepmedic for brain tumor segmentation. In *International workshop on Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries*, pages 138–149. Springer.

[101] Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., Rueckert, D., and Glocker, B. (2017b). Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78.

[102] Kao, P.-Y., Ngo, T., Zhang, A., Chen, J. W., and Manjunath, B. (2018). Brain tumor segmentation and tractographic feature extraction from structural mr images for overall survival prediction. In *International MICCAI Brainlesion Workshop*, pages 128–141. Springer.

[103] Karani, N., Erdil, E., Chaitanya, K., and Konukoglu, E. (2021). Test-time adaptable neural networks for robust medical image segmentation. *Medical Image Analysis*, 68:101907.

[104] Kasivisvanathan, V., Rannikko, A. S., Borghi, M., Panebianco, V., Mynderse, L. A., Vaarala, M. H., Briganti, A., Budäus, L., Hellawell, G., Hindley, R. G., et al. (2018). Mri-targeted or standard biopsy for prostate-cancer diagnosis. *New England Journal of Medicine*, 378(19):1767–1777.

[105] Kervadec, H., Bouchtiba, J., Desrosiers, C., Granger, E., Dolz, J., and Ayed, I. B. (2019). Boundary loss for highly unbalanced segmentation. In *International conference on medical imaging with deep learning*, pages 285–296. PMLR.

[106] Kim, E. Y. and Johnson, H. (2013). Robust multi-site mr data processing: iterative optimization of bias correction, tissue classification, and registration. *Frontiers in Neuroinformatics*, 7:29.

[107] Kim, E. Y., Lourens, S., Long, J., Paulsen, J., and Johnson, H. (2015). Preliminary analysis using multi-atlas labeling algorithms for tracing longitudinal change. *Frontiers in Neuroscience*, 9:242.

[108] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. (2023). Segment anything. *arXiv preprint arXiv:2304.02643*.

[109] Kleesiek, J., Urban, G., Hubert, A., Schwarz, D., Maier-Hein, K., Bendszus, M., and Biller, A. (2016). Deep mri brain extraction: A 3d convolutional neural network for skull stripping. *NeuroImage*, 129:460–469.

[110] Krishnan, R., Rajpurkar, P., and Topol, E. J. (2022). Self-supervised learning in medicine and healthcare. *Nature Biomedical Engineering*, 6(12):1346–1352.

[111] Kuhl, C. K., Bruhn, R., Krämer, N., Nebelung, S., Heidenreich, A., and Schrading, S. (2017). Abbreviated biparametric prostate mr imaging in men with elevated prostate-specific antigen. *Radiology*, 285(2):493–505.

[112] Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., and Klein, A. (2015). Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, volume 5, page 12.

[113] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

[114] Lee, H. H., Bao, S., Huo, Y., and Landman, B. A. (2023). 3d UX-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. In *The Eleventh International Conference on Learning Representations*.

[115] Lee, H. J., Kim, J. U., Lee, S., Kim, H. G., and Ro, Y. M. (2020). Structure boundary preserving segmentation for medical image with ambiguous boundary. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4817–4826.

[116] Li, H., Hu, D., Liu, H., Wang, J., and Oguz, I. (2022a). Cats: Complementary cnn and transformer encoders for segmentation. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE.

[117] Li, H., Hu, D., Zhu, Q., Larson, K. E., Zhang, H., and Oguz, I. (2021a). Unsupervised cross-modality domain adaptation for segmenting vestibular schwannoma and cochlea with data augmentation and model ensemble. In *International MICCAI Brainlesion Workshop*, pages 518–528. Springer.

[118] Li, H., Hu, D., Zhu, Q., Larson, K. E., Zhang, H., and Oguz, I. (2022b). Unsupervised cross-modality domain adaptation for segmenting vestibular schwannoma and cochlea with data augmentation and model ensemble. In *International MICCAI Brainlesion Workshop*, pages 518–528. Springer.

[119] Li, H., Liu, H., Hu, D., Wang, J., Johnson, H., Sherbini, O., Gavazzi, F., D'Aiello, R., Vanderver, A., Long, J., et al. (2022c). Self-supervised test-time adaptation for medical image segmentation. In *International Workshop on Machine Learning in Clinical Neuroimaging*, pages 32–41. Springer.

[120] Li, H., Liu, H., Hu, D., Wang, J., and Oguz, I. (2023a). Assessing test-time variability for interactive 3d medical image segmentation with diverse point prompts.

[121] Li, H., Liu, H., Hu, D., Wang, J., and Oguz, I. (2023b). Promise: Prompt-driven 3d medical image segmentation using pretrained image foundation models. *arXiv preprint arXiv:2310.19721*.

[122] Li, H., Liu, H., Hu, D., Wang, J., and Oguz, I. (2024). Prism: A promptable and robust interactive segmentation model with visual prompts. *arXiv preprint arXiv:2404.15028*.

[123] Li, H., Liu, H., Hu, D., Yao, X., Wang, J., and Oguz, I. (2023c). Cats v2: Hybrid encoders for robust medical segmentation. *arXiv preprint arXiv:2308.06377*.

[124] Li, H., Zhang, H., Hu, D., Johnson, H., Long, J. D., Paulsen, J. S., and Oguz, I. (2020). Generalizing mri subcortical segmentation to neurodegeneration. In *Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-oncology: Third International Workshop, MLCN 2020, and Second International Workshop, RNO-AI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 3*, pages 139–147. Springer.

[125] Li, H., Zhang, H., Johnson, H., Long, J. D., Paulsen, J. S., and Oguz, I. (2021b). Longitudinal subcortical segmentation with deep learning. In *Medical Imaging 2021: Image Processing*, volume 11596, pages 73–81. SPIE.

[126] Li, H., Zhang, H., Johnson, H., Long, J. D., Paulsen, J. S., and Oguz, I. (2021c). Mri subcortical segmentation in neurodegeneration with cascaded 3d cnns. In *Medical Imaging 2021: Image Processing*, volume 11596, pages 236–243. SPIE.

[127] Li, H., Zhu, Q., Hu, D., Gunnala, M. R., Johnson, H., Sherbini, O., Gavazzi, F., D'Aiello, R., Vanderver, A., Long, J. D., et al. (2022d). Human brain extraction with deep learning. In *Medical Imaging 2022: Image Processing*, volume 12032, pages 369–375. SPIE.

[128] Li, J., Chen, J., Tang, Y., Wang, C., Landman, B. A., and Zhou, S. K. (2023d). Transforming medical imaging with transformers? a comparative review of key properties, current progresses, and future perspectives. *Medical image analysis*, page 102762.

[129] Li, Z., Chen, Q., and Koltun, V. (2018). Interactive image segmentation with latent diversity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 577–585.

[130] Lin, Z., Zhang, Z., Chen, L.-Z., Cheng, M.-M., and Lu, S.-P. (2020). Interactive image segmentation with first click attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13339–13348.

[131] Linmans, J., Elfwing, S., van der Laak, J., and Litjens, G. (2023). Predictive uncertainty estimation for out-of-distribution detection in digital pathology. *Medical Image Analysis*, 83:102655.

[132] Litjens, G., Debats, O., Barentsz, J., Karssemeijer, N., and Huisman, H. (2014). Computer-aided detection of prostate cancer in mri. *IEEE Transactions on Medical Imaging*, 33(5):1083–1092.

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

[134] Liu, H., Fan, Y., Li, H., Wang, J., Hu, D., Cui, C., Lee, H. H., Zhang, H., and Oguz, I. (2022a). Moddrop++: A dynamic filter network with intra-subject co-training for multiple sclerosis lesion segmentation with missing modalities. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 444–453. Springer.

[135] Liu, H., Fan, Y., Oguz, I., and Dawant, B. M. (2022b). Enhancing data diversity for self-training based unsupervised cross-modality vestibular schwannoma and cochlea segmentation. In *International MICCAI Brainlesion Workshop*, pages 109–118. Springer.

[136] Liu, H., Fan, Y., Xu, Z., Dawant, B. M., and Oguz, I. (2023a). Learning site-specific styles for multi-institutional unsupervised cross-modality domain adaptation. *arXiv preprint arXiv:2311.12437*.

[137] Liu, H., Hu, D., Li, H., and Oguz, I. (2023b). Medical image segmentation using deep learning. *Machine Learning for Brain Disorders*, pages 391–434.

[138] Liu, H., Xu, Z., Gao, R., Li, H., Wang, J., Chabin, G., Oguz, I., and Grbic, S. (2024). Cosst: Multi-organ segmentation with partially labeled datasets using comprehensive supervisions and self-training. *IEEE Transactions on Medical Imaging*.

[139] Liu, J., Lou, B., Diallo, M., Meng, T., von Busch, H., Grimm, R., Tian, Y., Comaniciu, D., Kamen, A., Winkel, D., et al. (2021a). Detecting out-of-distribution via an unsupervised uncertainty estimation for prostate cancer diagnosis. In *Medical Imaging with Deep Learning*.

[140] Liu, Q., Xu, Z., Bertasius, G., and Niethammer, M. (2023c). Simpleclick: Interactive image segmentation with simple vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22290–22300.

[141] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021b). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

[142] Long, J., Shelhamer, E., and Darrell, T. (2015a). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440.

[143] Long, J. D., Paulsen, J. S., et al. (2015b). Multivariate prediction of motor diagnosis in Huntington's disease: 12 years of PREDICT-HD. *Mov. Disorders*, 30(12):1664–1672.

[144] Long, J. D., Paulsen, J. S., Marder, K., Zhang, Y., Kim, J.-I., Mills, J. A., and of the PREDICT-HD Huntington's Study Group, R. (2014). Tracking motor impairments in the progression of huntington's disease. *Movement Disorders*, 29(3):311–319.

[145] Long, J. D., Paulsen, J. S., Marder, K., Zhang, Y., Kim, J.-I., Mills, J. A., and the Researchers of the PREDICT-HD Huntington's Study Group (2013). Tracking motor impairments in the progression of Huntington's disease. *Movement Disorders*, 29(3):311–319.

[146] Long, J. D., Paulsen, J. S., and PREDICT-HD Investigators and Coordinators of the Huntington Study Group (2015c). Multivariate prediction of motor diagnosis in Huntington's disease: 12 years of PREDICT-HD. *Movement Disorders*, 30(12):1664–1672.

[147] Luo, X., Wang, G., Song, T., Zhang, J., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., and Zhang, S. (2021). Mideepseg: Minimally interactive segmentation of unseen objects from medical images using deep learning. *Medical image analysis*.

[148] Ma, J., Chen, J., Ng, M., Huang, R., Li, Y., Li, C., Yang, X., and Martel, A. L. (2021). Loss odyssey in medical image segmentation. *Medical Image Analysis*, 71:102035.

[149] Ma, J., He, Y., Li, F., Han, L., You, C., and Wang, B. (2024). Segment anything in medical images. *Nature Communications*, 15(1):654.

[150] Ma, J. and Wang, B. (2023). Segment anything in medical images. *arXiv preprint arXiv:2304.12306*.

[151] Maier, A., Steidl, S., Christlein, V., and Hornegger, J. (2018). Medical imaging systems: An introductory guide.

[152] Matsoukas, C., Haslum, J. F., Sorkhei, M., Söderberg, M., and Smith, K. (2022). What makes transfer learning work for medical images: Feature reuse & other factors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9225–9234.

[153] McKinley, R., Wepfer, R., Aschwanden, F., Grunder, L., Muri, R., Rummel, C., Verma, R., Weisstanner, C., Reyes, M., Salmen, A., et al. (2019). Simultaneous lesion and neuroanatomy segmentation in multiple sclerosis using deep neural networks. *arXiv preprint arXiv:1901.07419*.

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

[155] Myronenko, A. (2018). 3d mri brain tumor segmentation using autoencoder regularization. In *International MICCAI Brainlesion Workshop*, pages 311–320. Springer.

[156] Nugent, A. C., Luckenbaugh, D. A., Wood, S. E., Bogers, W., Zarate Jr, C. A., and Drevets, W. C. (2013). Automated subcortical segmentation using first: test–retest reliability, interscanner reliability, and comparison to manual segmentation. *Human brain mapping*, 34(9):2313–2329.

[157] Nyúl, L. G. and Udupa, J. K. (1999). On standardizing the mr image intensity scale. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 42(6):1072–1081.

[158] Oguz, B. U., Shinohara, R. T., Yushkevich, P. A., and Oguz, I. (2017). Gradient boosted trees for corrective learning. In *International Workshop on Machine Learning in Medical Imaging*, pages 203–211. Springer.

[159] Oguz, I., Kashyap, S., Wang, H., Yushkevich, P., and Sonka, M. (2016). Globally optimal label fusion with shape priors. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 538–546. Springer.

[160] Oguz, I., Yushkevich, N., Pouch, A., Oguz, B. U., Wang, J., Parameshwaran, S., Gee, J., Yushkevich, P. A., and Schwartz, N. (2020). Minimally interactive placenta segmentation from three-dimensional ultrasound images. *Journal of Medical Imaging*, 7(1):014004–014004.

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

[162] Otsu, N. et al. (1975). A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27.

[163] Ouyang, C., Chen, C., Li, S., Li, Z., Qin, C., Bai, W., and Rueckert, D. (2022). Causality-inspired single-source domain generalization for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(4):1095–1106.

[164] Ouyang, C., Kamnitsas, K., Biffi, C., Duan, J., and Rueckert, D. (2019). Data efficient unsupervised domain adaptation for cross-modality image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 669–677. Springer.

[165] Pan, J., Lin, Z., Zhu, X., Shao, J., and Li, H. (2022). St-adapter: Parameter-efficient image-to-video transfer learning. *Advances in Neural Information Processing Systems*, 35:26462–26477.

[166] Park, T., Efros, A. A., Zhang, R., and Zhu, J.-Y. (2020). Contrastive learning for unpaired image-to-image translation. In *European conference on computer vision*, pages 319–345. Springer.

[167] Paschalis, A. and de Bono, J. S. (2020). Prostate cancer 2020:"the times they are a'changing". *Cancer Cell*, 38(1):25–27.

[168] Patenaude, B., Smith, S. M., Kennedy, D. N., and Jenkinson, M. (2011). A bayesian model of shape and appearance for subcortical brain segmentation. *Neuroimage*, 56(3):907–922.

[169] Paulsen, J. S., Long, J. D., Johnson, H. J., Aylward, E. H., Ross, C. A., Williams, J. K., Nance, M. A., Erwin, C. J., Westervelt, H. J., Harrington, D. L., et al. (2014a). Clinical and biomarker changes in premanifest huntington disease show trial feasibility: a decade of the predict-hd study. *Frontiers in aging neuroscience*, 6:78.

[170] Paulsen, J. S., Long, J. D., Ross, C. A., Harrington, D. L., Erwin, C. J., Williams, J. K., Westervelt, H. J., Johnson, H. J., Aylward, E. H., Zhang, Y., Bockholt, H. J., Barker, R. A., and PREDICT-HD Investigators and Coordinators of the Huntington Study Group (2014b). Prediction of manifest Huntington's disease with clinical and imaging measures: a prospective observational study. *The Lancet Neurology*, 13(12):1193–1201.

[171] Paulsen, J. S., Long, J. D., Ross, C. A., Harrington, D. L., Erwin, C. J., Williams, J. K., Westervelt, H. J., Johnson, H. J., Aylward, E. H., Zhang, Y., et al. (2014c). Prediction of manifest huntington's disease with clinical and imaging measures: a prospective observational study. *The Lancet Neurology*, 13(12):1193–1201.

[172] Peiris, H., Hayat, M., Chen, Z., Egan, G., and Harandi, M. (2021). A volumetric transformer for accurate 3d tumor segmentation. *arXiv preprint arXiv:2111.13300*.

[173] Peiris, H., Hayat, M., Chen, Z., Egan, G., and Harandi, M. (2022). A robust volumetric transformer for accurate 3d tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 162–172. Springer.

[174] Pérez-García, F., Sparks, R., and Ourselin, S. (2021). Torchio: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Computer Methods and Programs in Biomedicine*, page 106236.

[175] Pham, D. L., Xu, C., and Prince, J. L. (2000). Current methods in medical image segmentation. *Annual review of biomedical engineering*, 2(1):315–337.

[176] Pierson, R., Johnson, H., Harris, G., Keefe, H., Paulsen, J. S., Andreasen, N. C., and Magnotta, V. A. (2011). Fully automated analysis using brains: Autoworkup. *NeuroImage*, 54(1):328–336.

[177] Prince, J. L. and Links, J. M. (2006). *Medical imaging signals and systems*, volume 37. Pearson Prentice Hall Upper Saddle River.

[178] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.

[179] Reinhold, J. C., Dewey, B. E., Carass, A., and Prince, J. L. (2019). Evaluating the impact of intensity normalization on mr image synthesis. In *Medical Imaging 2019: Image Processing*, volume 10949, pages 890–898. SPIE.

[180] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer.

[181] Rosenkrantz, A. B., Parikh, N., Kierans, A. S., Kong, M. X., Babb, J. S., Taneja, S. S., and Ream, J. M. (2016). Prostate cancer detection using computed very high b-value diffusion-weighted imaging: how high should we go? *Academic Radiology*, 23(6):704–711.

[182] Saha, A., Hosseinzadeh, M., and Huisman, H. (2021). End-to-end prostate cancer detection in bpmri via 3d cnns: effects of attention mechanisms, clinical priori and decoupled false positive reduction. *Medical image analysis*, 73:102155.

[183] Schelb, P., Kohl, S., Radtke, J. P., Wiesenfarth, M., Kickingereder, P., Bickelhaupt, S., Kuder, T. A., Stenzinger, A., Hohenfellner, M., Schlemmer, H.-P., et al. (2019). Classification of cancer at prostate mri: deep learning versus clinical pi-rads assessment. *Radiology*, 293(3):607–617.

[184] Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., and Rueckert, D. (2019). Attention gated networks: Learning to leverage salient regions in medical images. *MedIA*, 53:197–207.

[185] Schmidt, A., Morales-Álvarez, P., and Molina, R. (2023). Probabilistic modeling of inter-and intra-observer variability in medical image segmentation. In *CVPR*, pages 21097–21106.

[186] Seidlitz, S., Sellner, J., Odenthal, J., Özdemir, B., Studier-Fischer, A., Knödler, S., Ayala, L., Adler, T., Kenngott, H., Tizabi, M., Wagner, M., Nickel, F., Müller-Stich, B., and Maier-Hein, L. (2022). Robust deep learning-based semantic organ segmentation in hyperspectral images. *Medical Image Analysis*, 80:102488.

[187] Sezgin, M. and Sankur, B. l. (2004). Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic imaging*, 13(1):146–168.

[188] Shah, M., Xiao, Y., Subbanna, N., Francis, S., Arnold, D. L., Collins, D. L., and Arbel, T. (2011). Evaluating intensity normalization on mris of human brain with multiple sclerosis. *Medical image analysis*, 15(2):267–282.

[189] Shamshad, F., Khan, S., Zamir, S. W., Khan, M. H., Hayat, M., Khan, F. S., and Fu, H. (2022). Transformers in medical imaging: A survey. *arXiv preprint arXiv:2201.09873*.

[190] Shamshad, F., Khan, S., Zamir, S. W., Khan, M. H., Hayat, M., Khan, F. S., and Fu, H. (2023). Transformers in medical imaging: A survey. *Medical Image Analysis*, page 102802.

[191] Shapey, J., Kujawa, A., Dorent, R., Wang, G., Dimitriadis, A., Grishchuk, D., Paddick, I., Kitchen, N., Bradford, R., Saeed, S. R., Bisdas, S., Ourselin, S., and Vercauteren, T. (2021). Segmentation of vestibular schwannoma from mri — an open annotated dataset and baseline algorithm. *Scientific Data*. In press. Preprint available at https://doi.org/10.1101/2021.08.04.21261588 medRXiv:10.1101/2021.08.04.21261588.

[192] Shapey, J., Wang, G., Dorent, R., Dimitriadis, A., Li, W., Paddick, I., Kitchen, N., Bisdas, S., Saeed, S. R., Ourselin, S., et al. (2019). An artificial intelligence framework for automatic segmentation and volumetry of vestibular schwannomas from contrast-enhanced t1-weighted and high-resolution t2-weighted mri. *Journal of neurosurgery*, 134(1):171–179.

[193] Shattuck, D. W., Sandor-Leahy, S. R., Schaper, K. A., Rottenberg, D. A., and Leahy, R. M. (2001). Magnetic resonance image tissue classification using a partial volume model. *NeuroImage*, 13(5):856–876.

[194] Simpson, A. L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Van Ginneken, B., Kopp-Schneider, A., Landman, B. A., Litjens, G., Menze, B., et al. (2019). A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*.

[195] Sinha, A. and Dolz, J. (2020). Multi-scale self-guided attention for medical image segmentation. *IEEE journal of biomedical and health informatics*, 25(1):121–130.

[196] Smith, S. M. (2002). Fast robust automated brain extraction. *Human brain mapping*, 17(3):143–155.

[197] Sofiiuk, K., Petrov, I. A., and Konushin, A. (2021). Reviving iterative training with mask guidance for interactive segmentation.

[198] Stoebner, Z. A., Lu, D., Hong, S. H., Kavoussi, N. L., and Oguz, I. (2022). Segmentation of kidney stones in endoscopic video feeds. In *Medical Imaging 2022: Image Processing*, volume 12032, pages 900–908. SPIE.

[199] Sun, S., Xian, M., Xu, F., Yao, T., and Capriotti, L. (2023). Cfr-icl: Cascade-forward refinement with iterative click loss for interactive image segmentation. *arXiv preprint arXiv:2303.05620*.

[200] Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., and Hardt, M. (2020). Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning*, pages 9229–9248. PMLR.

[201] Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249.

[202] Tang, Y., Yang, D., Li, W., Roth, H., Landman, B., Xu, D., Nath, V., and Hatamizadeh, A. (2022a). Self-supervised pre-training of swin transformers for 3d medical image analysis.

[203] Tang, Y., Yang, D., Li, W., Roth, H. R., Landman, B., Xu, D., Nath, V., and Hatamizadeh, A. (2022b). Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20730–20740.

[204] Tang, Y., Yang, D., Li, W., Roth, H. R., Landman, B., Xu, D., Nath, V., and Hatamizadeh, A. (2022c). Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20730–20740.

[205] Turkbey, B., Rosenkrantz, A. B., Haider, M. A., Padhani, A. R., Villeirs, G., Macura, K. J., Tempany, C. M., Choyke, P. L., Cornud, F., Margolis, D. J., et al. (2019). Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2. *European urology*, 76(3):340–351.

[206] Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I., and Patel, V. M. (2021). Medical transformer: Gated axial-attention for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 36–46. Springer.

[207] Vanderver, A., Prust, M., Kadom, N., Demarest, S., Crow, Y. J., Helman, G., Orcesi, S., Piana, R. L., Uggetti, C., Wang, J., et al. (2015). Early-onset aicardi-goutieres syndrome: magnetic resonance imaging (mri) pattern recognition. *Journal of child neurology*, 30(10):1343–1348.

[208] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

[209] Walker, F. O. (2007). Huntington's disease. *The Lancet*, 369(9557):218–228.

[210] Wang, D., Shelhamer, E., Liu, S., Olshausen, B., and Darrell, T. (2020). Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*.

[211] Wang, G., Li, W., Ourselin, S., and Vercauteren, T. (2017). Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks. In *International MICCAI brainlesion workshop*, pages 178–190. Springer.

[212] Wang, G., Li, W., Zuluaga, M. A., Pratt, R., Patel, P. A., Aertsen, M., Doel, T., David, A. L., Deprest, J., Ourselin, S., et al. (2018a). Interactive medical image segmentation using deep learning with image-specific fine tuning. *IEEE transactions on medical imaging*, 37(7):1562–1573.

[213] Wang, G., Shapey, J., Li, W., Dorent, R., Demitriadis, A., Bisdas, S., Paddick, I., Bradford, R., Zhang, S., Ourselin, S., et al. (2019). Automatic segmentation of vestibular schwannoma from t2-weighted mri by deep spatial attention with hardness-weighted loss. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 264–272. Springer.

[214] Wang, G., Zuluaga, M. A., Li, W., Pratt, R., Patel, P. A., Aertsen, M., Doel, T., David, A. L., Deprest, J., Ourselin, S., et al. (2018b). Deepigeos: a deep interactive geodesic framework for medical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1559–1572.

[215] Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su, Y., Huang, Z., Shen, Y., et al. (2023a). Sam-med3d. *arXiv preprint arXiv:2310.15161*.

[216] Wang, H., Suh, J. W., Das, S. R., Pluta, J. B., Craige, C., and Yushkevich, P. A. (2012). Multi-atlas segmentation with joint label fusion. *IEEE transactions on pattern analysis and machine intelligence*, 35(3):611–623.

[217] Wang, J., Li, H., Hu, D., Tao, Y. K., and Oguz, I. (2023b). Novel oct mosaicking pipeline with feature-and pixel-based registration. *arXiv preprint arXiv:2311.13052*.

[218] Wang, J., Li, H., Liu, H., Hu, D., Lu, D., Yoon, K., Barter, K., Bagnato, F., and Oguz, I. (2023c). Ssl2 self-supervised learning meets semi-supervised learning: multiple clerosis segmentation in 7t-mri from large-scale 3t-mri. In *Medical Imaging 2023: Image Processing*, volume 12464, pages 126–136. SPIE.

[219] Wang, J., Vachet, C., Rumple, A., Gouttard, S., Ouziel, C., Perrot, E., Du, G., Huang, X., Gerig, G., and Styner, M. A. (2014). Multi-atlas segmentation of subcortical brain structures via the autoseg software pipeline. *Frontiers in neuroinformatics*, 8:7.

[220] Wei, X., Cao, J., Jin, Y., Lu, M., Wang, G., and Zhang, S. (2023). I-medsam: Implicit medical image segmentation with segment anything. *arXiv preprint*.

[221] Weinreb, J. C., Barentsz, J. O., Choyke, P. L., Cornud, F., Haider, M. A., Macura, K. J., Margolis, D., Schnall, M. D., Shtern, F., Tempany, C. M., et al. (2016). Pi-rads prostate imaging–reporting and data system: 2015, version 2. *European urology*, 69(1):16–40.

[222] Winkel, D. J., Tong, A., Lou, B., Kamen, A., Comaniciu, D., Disselhorst, J. A., Rodríguez-Ruiz, A., Huisman, H., Szolar, D., Shabunin, I., et al. (2021). A novel deep learning based computer-aided diagnosis system improves the accuracy and efficiency of radiologists in reading biparametric magnetic resonance images of the prostate: results of a multireader, multicase study. *Investigative radiology*, 56(10):605–613.

[223] Wong, H. E., Rakic, M., Guttag, J., and Dalca, A. V. (2023). Scribbleprompt: Fast and flexible interactive segmentation for any medical image. *arXiv preprint*.

[224] Wu, J., Fu, R., Fang, H., Liu, Y., Wang, Z., Xu, Y., Jin, Y., and Arbel, T. (2023). Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*.

[225] Wu, J., Zhang, Y., and Tang, X. (2019). A joint 3d+ 2d fully convolutional framework for subcortical segmentation. In *MICCAI*, pages 301–309.

[226] Xiao, H., Li, L., Liu, Q., Zhu, X., and Zhang, Q. (2023). Transformers in medical image segmentation: A review. *Biomedical Signal Processing and Control*, 84:104791.

[227] Xie, Y., Zhang, J., Shen, C., and Xia, Y. (2021). Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 171–180. Springer.

[228] Xingjian, S., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., and Woo, W.-c. (2015). Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810.

[229] Xu, C. and Prince, J. L. (1998). Snakes, shapes, and gradient vector flow. *IEEE Transactions on image processing*, 7(3):359–369.

[230] Yang, D., Xu, D., Zhou, S. K., Georgescu, B., Chen, M., Grbic, S., Metaxas, D., and Comaniciu, D. (2017). Automatic liver segmentation using an adversarial image-to-image network. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*, pages 507–515. Springer.

[231] Yang, H., Sun, J., Yang, L., and Xu, Z. (2021). A unified hyper-gan model for unpaired multi-contrast mr image translation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 127–137. Springer.

[232] Yang, T., Zhu, Y., Xie, Y., Zhang, A., Chen, C., and Li, M. (2023). Aim: Adapting image models for efficient video action recognition. *arXiv preprint arXiv:2302.03024*.

[233] Yao, X., Liu, H., Hu, D., Lu, D., Lou, A., Li, H., Deng, R., Arenas, G., Oguz, B., Schwartz, N., et al. (2023a). False negative/positive control for sam on noisy medical images. *arXiv preprint*.

[234] Yao, X., Lou, A., Li, H., Hu, D., Lu, D., Liu, H., Wang, J., Stoebner, Z., Johnson, H., Long, J. D., et al. (2023b). Novel application of the attention mechanism on medical image harmonization. In *Medical Imaging 2023: Image Processing*, volume 12464, pages 184–194. SPIE.

[235] Yi, F. and Moon, I. (2012). Image segmentation: A survey of graph-cut methods. In *2012 international conference on systems and informatics (ICSAI2012)*, pages 1936–1941. IEEE.

[236] Yu, X., Lou, B., Shi, B., Winkel, D., Arrahmane, N., Diallo, M., Meng, T., von Busch, H., Grimm, R., Kiefer, B., et al. (2020a). False positive reduction using multiscale contextual features for prostate cancer detection in multi-parametric mri scans. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1355–1359. IEEE.

[237] Yu, X., Lou, B., Zhang, D., Winkel, D., Arrahmane, N., Diallo, M., Meng, T., Busch, H. v., Grimm, R., Kiefer, B., et al. (2020b). Deep attentive panoptic model for prostate cancer detection using biparametric mri scans. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 594–604. Springer.

[238] Zhang, D., Huang, G., Zhang, Q., Han, J., Han, J., and Yu, Y. (2021a). Cross-modality deep feature learning for brain tumor segmentation. *Pattern Recognition*, 110:107562.

[239] Zhang, H., Li, H., Larson, K., Hett, K., and Oguz, I. (2023a). Domain generalization for robust ms lesion segmentation. In *Medical Imaging 2023: Image Processing*, volume 12464, pages 195–202. SPIE.

[240] Zhang, H., Li, H., and Oguz, I. (2021b). Segmentation of new ms lesions with tiramisu and 2.5 d stacked slices. *MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure*, 61.

[241] Zhang, H., Valcarcel, A. M., Bakshi, R., Chu, R., Bagnato, F., Shinohara, R. T., Hett, K., and Oguz, I. (2019a). Multiple sclerosis lesion segmentation with tiramisu and 2.5 d stacked slices. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22*, pages 338–346. Springer.

[242] Zhang, H., Zhang, J., Zhang, Q., Kim, J., Zhang, S., Gauthier, S. A., Spincemaille, P., Nguyen, T. D., Sabuncu, M., and Wang, Y. (2019b). Rsanet: Recurrent slice-wise attention network for multiple sclerosis lesion segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 411–419. Springer.

[243] Zhang, K. and Liu, D. (2023). Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*.

[244] Zhang, L., Wang, X., Yang, D., Sanford, T., Harmon, S., Turkbey, B., Wood, B. J., Roth, H., Myronenko, A., Xu, D., et al. (2020). Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE transactions on medical imaging*, 39(7):2531–2540.

[245] Zhang, Q., Wang, L., Zong, X., Lin, W., Li, G., and Shen, D. (2019c). Frnet: Flattened residual network for infant mri skull stripping. In *ISBI*, pages 999–1002. IEEE.

[246] Zhang, Y., Hu, S., Jiang, C., Cheng, Y., and Qi, Y. (2023b). Segment anything model with uncertainty rectification for auto-prompting medical image segmentation. *arXiv preprint arXiv:2311.10529*.

[247] Zhang, Y., Liu, H., and Hu, Q. (2021c). Transfuse: Fusing transformers and cnns for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 14–24. Springer.

[248] Zhang, Y., Long, J. D., Mills, J. A., Warner, J. H., Lu, W., Paulsen, J. S., Investigators, P.-H., and of the Huntington Study Group, C. (2011). Indexing disease progression at study entry with individuals at-risk for huntington disease. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 156(7):751–763.

[249] Zhang, Y., Zhou, T., Liang, P., and Chen, D. Z. (2023c). Input augmentation with sam: Boosting medical image segmentation with segmentation foundation model. *arXiv preprint arXiv:2304.11332*.

[250] Zhang, Z., Yang, L., and Zheng, Y. (2018). Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern Recognition*, pages 9242–9251.

[251] Zhao, X., Wu, Y., Song, G., Li, Z., Zhang, Y., and Fan, Y. (2017). 3d brain tumor segmentation through integrating multiple 2d fcnns. In *International MICCAI Brainlesion Workshop*, pages 191–203. Springer.

[252] Zhao, Z., Yang, H., and Sun, J. (2022). Modality-adaptive feature interaction for brain tumor segmentation with missing modalities. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 183–192. Springer.

[253] Zhou, H.-Y., Guo, J., Zhang, Y., Han, X., Yu, L., Wang, L., and Yu, Y. (2023). nnformer: Volumetric medical image segmentation via a 3d transformer. *IEEE Transactions on Image Processing*.

[254] Zhou, H.-Y., Guo, J., Zhang, Y., Yu, L., Wang, L., and Yu, Y. (2021). nnformer: Interleaved transformer for volumetric segmentation. *arXiv preprint arXiv:2109.03201*.

[255] Zhou, H.-Y., Guo, J., Zhang, Y., Yu, L., Wang, L., and Yu, Y. (2022). nnformer: Interleaved transformer for volumetric segmentation.

[256] Zhou, T., Ruan, S., Guo, Y., and Canu, S. (2020). A multi-modality fusion network based on attention mechanism for brain tumor segmentation. In *2020 IEEE 17th international symposium on biomedical imaging (ISBI)*, pages 377–380. IEEE.

[257] Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., and Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer.

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

[259] Zhu, Q., Li, H., Cass, N. D., Lindquist, N. R., Tawfik, K. O., and Oguz, I. (2022). Acoustic neuroma segmentation using ensembled convolutional neural networks. In *Medical Imaging 2022: Biomedical Applications in Molecular, Structural, and Functional Imaging*, volume 12036, pages 228–234. SPIE.

[260] Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Gao, J., and Lee, Y. J. (2023). Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*.