UTS UNIVERSITY OF TECHNOLOGY SYDNEY

Learning with Imperfect Datasets in Medical Image Segmentation

by Yuhang Ding

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Prof. Yi Yang

University of Technology Sydney
Faculty of Engineering and Information Technology

March 2024

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, *Yuhang Ding* declare that this thesis, is submitted in fulfillment of the requirements for the award of Doctor of Philosophy, in the *Faculty of Engineering and Information Technology*, at the University of Technology Sydney, Australia.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

SIGNATURE: Production Note:
Signature removed prior to publication.

[Yuhang Ding]

DATE: 26th March, 2024

PLACE: Sydney, Australia

ACKNOWLEDGMENTS

Firstly, I would like to express my deepest gratitude to my principal supervisor, Professor Yi Yang, and my co-supervisor, Dr. Mingjie Li. Their guidance, counsel, and mentorship have been invaluable throughout my Ph.D. journey. They have skillfully directed me in choosing my research topic and have provided me with the necessary resources to succeed.

I also thank my collaborators and colleagues at the University of Technology Sydney. I would like to thank Prof. Wenguan Wang, Prof. Xin Yu, Prof. Hehe Fan, Dr. Yifan Sun, Prof. Xiaojun Chang, Prof. Yu Wu, Dr. Liulei Li, Dr. Jiaxu Miao, Dr. Zongxin Yang, Dr. Ruijie Quan, Dr. Fan Ma, Dr. Yunqiu Xu, Dr. Feng Zhu, Dr. Gengwei Zhang, Dr. Yu Lu, Dr. Yuanzhi Liang, Dr. Xuanmeng Zhang, and many others. Working with them and engaging in intellectual conversations was a fortunate experience.

On a personal note, I wish to thank my family, my mother, Xiaomei Jiang and my father, Yinlin Ding, for their understanding and endless love, throughout the duration of my studies.

List of Publications

Related to the Thesis :

1. **Y. Ding**, X. Yu, and Y. Yang, "Modeling the Probabilistic Distribution of Unlabeled Data for One-shot Medical Image Segmentation," in *Association for the Advancement of Artificial Intelligence (AAAI)*, 2021.

2. **Y. Ding**, X. Yu, and Y. Yang, "RFNet: Region-aware Fusion Network for Incomplete Multi-modal Brain Tumor Segmentation," in *IEEE International Conference on Computer Vision (ICCV)*, 2021.

3. **Y. Ding**, L. Li, W. Wang, and Y. Yang, "Clustering Propagation for Universal Medical Image Segmentation, " in *IEEE Computer Vision and Pattern Recognition Conference (CVPR)*, 2024.

4. **Y. Ding** and H. Liu, "Barely-supervised Brain Tumor Segmentation via Employing Segment Anything Model", in *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2024.

Others :

5. **Y. Ding**, H. Fan, M. Xu, and Y. Yang, "Adaptive Exploration for Unsupervised Person Re-Identification," in *ACM Transactions on Multimedia Computing, Communications and Applications (TOMM)*, 2020.

6. Y. He, **Y. Ding**, P. Liu, L. Zhu, H. Zhang, and Y. Yang, "Learning Filter Pruning Criteria for Deep Convolutional Neural Networks Acceleration," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

7. H. Fan, X. Yu, **Y. Ding**, Y. Yang, and M. Kankanhalli, "Pstnet: Point Spatio-Temporal Convolution on Point Cloud Sequences," in *International Conference on Learning Representations (ICLR)*, 2021.

Abstract

Medical image segmentation aims to partition medical images into distinct physiological regions, such as organs and lesions, which is crucial for disease diagnosis and treatment planning. The advent of deep neural networks has significantly advanced this field. However, the performance in real-world scenarios remains unsatisfactory due to imperfect data and annotations during both training and deployment. First, scaling up training data and annotations is challenging. This is because obtaining medical images is difficult due to privacy concerns. Furthermore, annotating medical images requires substantial expertise, making the process costly and difficult to carry out. Second, the quality of medical images cannot always be guaranteed in real-world scenarios, leading to significant performance drops in outlier cases. Third, real-world medical applications are safety-critical and demand extremely accurate predictions, a requirement that most existing models fail to meet adequately. These challenges hinder the practical deployment of medical image segmentation. Consequently, both academic and industrial communities are striving to develop highly accurate medical image segmentation algorithms that can perform well despite imperfect data. To this end, this thesis proposes deep learning methods to develop a well-performed medical image segmentation model that can be effectively trained with limited and low-quality data/annotations. Specifically, a comprehensive suite is proposed from three directions: (1) applying image registration to generate realistic and diverse training samples and adopting barely-supervised learning paradigms to enable learning with insufficient annotated data; (2) devising region-aware fusion module to address missing modality problem; (3) incorporating automatic and interactive medical image segmentation into a single model and one training session to achieve sufficient segmentation performance for practical use. Extensive experiments on several medical image segmentation tasks, such as brain tumor segmentation, brain structure segmentation and abdominal organ segmentation, demonstrate the effectiveness and efficiency of the proposed techniques.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE **Page**

List of Tables

Table **Page**

CHAPTER 1

INTRODUCTION

This dissertation presents four work regarding learning with imperfect datasets for medical image segmentation (MIS), *i.e.*, one-shot/barely-supervised MIS, incomplete multi-modal MIS, and interactive MIS. This chapter begins by introducing medical image segmentation, including background and challenges, and then outlines research objectives and the organization of the thesis.

1.1 Medical Image Segmentation

1.1.1 Background

Medical image segmentation (MIS) classifies each pixel or voxel of medical images into specific anatomical or pathological categories. This enhances the visibility of structures and helps the measurement of critical metrics, making MIS crucial for medical applications, such as disease diagnosis, clinical evaluations, and surgical preparation. MIS aims to segment various anatomical areas in different human parts, such as brain structures and tumors [32, 136], retina [164], cardiac [169], and live tumors [107, 176]. This variety requires distinct imaging techniques, such as Magnetic Resonance Imaging (MRI), Computed Tomography (CT), and Optical Coherence Tomography (OCT). For example, MRI excels at imaging soft tissue contrast, making it ideal for diagnosing conditions in the brain, spinal cord, nerves, and muscles, while CT is particularly effective for high-contrast resolution in dense structures, such as bones. Thus, research efforts are

necessitated for various anatomical regions captured from different scanners.

Early research begins by adapting conventional techniques, such as template matching techniques, edge detection, active contours, statistical shape models, and machine learning, to medical image segmentation. For example, Lalonde *et al.* [95] and Chen *et al.* [27] introduce template matching for disc inspection and ventricular segmentation, respectively. Yu *et al.* [218] propose a novel edge detection algorithm of mathematical morphology for lung CT images. Tsai *et al.* [170] develop a shape-based method for cardiac and prostate MRI segmentation. Li *et al.* [104] combine level sets and support vector machines (SVMs) for medical image segmentation, while Held *et al.* [73] introduce the use of Markov Random Fields (MRF) to segment brain MRI images. Despite numerous conventional algorithms being explored, traditional methods are hindered by the poor representation ability of hand-crafted features and still yield limited segmentation performance. Consequently, the research focus has shifted towards deep learning methods using more powerful, deeply-learned features.

Deep-learning research attempts mainly explore various network architectures and training objectives loss functions. Improving network architectures enables the extraction of more representative features. Early research efforts mainly develop fully convolutional neural networks [126], such as U-Net [33, 153], V-Net [140], and U-Net++ [241]. Recently, with the rise and emergence of a series of transformer technologies, only not pure-transformer-based architectures have been devised in the MIS field, such as Swin-Unet [18], DS-TransUNet [114], nnFormer [234], MISSFormer [76], TransDeepLab [5], but also hybrid models are explored, such as TransUNet [25], TransBTS [115], MedT [173], UNETR [64], Swin UNETR [63], Swin UNETR++ [184], Segtran [105], CoTr [203], and HiFormer [72]. Research on training losses mainly improves the optimization of networks and maximizes the capabilities of the networks and training data. For example, weighted cross-entropy loss, Dice loss, Tversky loss [156], and generalized Dice loss [165] are devised for imbalanced data and improve the network capability on small and rarely-seen areas. Based on these basic architectures and objective loss functions, Isensee *et al.* [78] propose nnUNet to handle images of various structures scanned from various imaging techniques in a fully supervised manner and achieve the SOTA accuracy on all those applications. These deeply-learned methods achieve impressive segmentation performance yet face several challenges in practice.

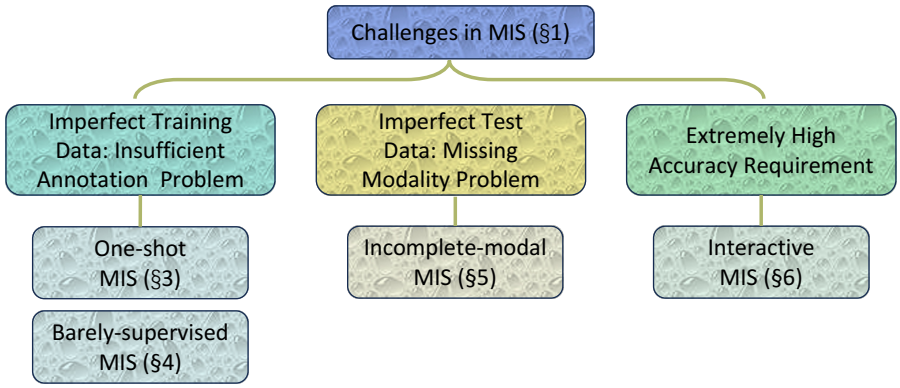

Figure 1.1: Overview of the challenges and solutions in medical image segmentation.

1.1.2 Challenges

Although deep learning has significantly improved the accuracy of medical image segmentation (MIS), it still faces several challenges (as seen in Fig. 1.1) that hinder the deployment of MIS in real-world applications. First, deep learning-based methods demand large-scale training data, yet perfectly labeled datasets are rarely available, which hinders networks from achieving optimal performance. Second, low-quality samples often exist in clinic practice, and deep learning networks are particularly vulnerable to these outliers. Third, real-world applications are safety-critical and thus often require extremely high accuracy for MIS, which is difficult to achieve with existing deep learning methods. In the following, the details of these challenges will be presented, and potential solutions will be discussed.

Imperfect Training Data. Deep learning-based methods usually require perfectly-sized and carefully-labeled training datasets to achieve optimal performance. However, these datasets are hard to collect in the real world. Firstly, collecting patient data, including medical images and metadata, into a centralized data lake is often impractical. To be specific, patient data is usually collected from various institutions and hospitals, and cannot be combined into one centralized data lake for training due to privacy regulations [82, 152]. For this challenge, decentralized and privacy-preserving training schemes, such as federated learning [134, 137], are explored to exploit the decentralized data effectively. Second, obtaining pixel-wise segmentation annotations for 3D medical images costs too much time and expertise. Even worse, medical image segmentation usually focuses on various anatomical human regions. This variety increases the difficulty of the annotation process as different regions may need different expertise. Therefore, collecting large-scale labeled medical segmentation datasets is impractical, while deep

learning-based methods often require much reliable supervision from these datasets to perform accurately. For this challenge, research efforts pay attention to one-shot/semi-supervised/barely-supervised learning, which mainly aims to exploit unlabeled data effectively. In this fashion, the demand for labeled data is reduced.

Imperfect Test Data. Due to varying patient conditions and scanning protocols, deep learning-based methods usually encounter various defective medical images in real-world scenarios. If deep learning-based networks do not encounter similar images during training, they become vulnerable to these imperfect images and cannot segment them accurately. For example, in some medical image segmentation applications, multiple modalities are employed to boost the performance of segmentation. However, the absence of certain modalities frequently occurs in practice [23, 45], leading to severe segmentation accuracy reduction. To address this challenge, researchers simulate missing modality scenarios during network training. They also propose aligning predictions from full-modal and partial-modal images to further enhance network capability. In addition to the missing modality challenge, low-quality challenges often occur, such as motion blurring, ghosting, and spike artifacts. For this challenge, multiple data augmentation techniques are employed to enhance network robustness against these artifacts.

Extremely High Accuracy Requirement. Medical applications are often safety-critical [36, 219], and minor mistakes may lead to severe consequences. For example, a misdiagnosis may lead doctors to adopt inappropriate treatment methods, delaying the timely containment of the underlying condition, which may result in the patient's death. Additionally, even slight deviations during surgery may cause the procedure to fail, potentially leading to the patient's death. Therefore, medical image segmentation in these applications typically requires extremely high segmentation accuracy. However, existing deep learning-based methods rarely meet this requirement. For this challenge, interactive segmentation has been proposed to incorporate expert interactions, allowing for the iterative prediction refinement and achieving satisfactory results.

Discussions The deployment of medical image segmentation (MIS) models in real-world practice faces significant challenges related to training, testing and strict requirements. Furthermore, these issues often occur simultaneously, compounding the difficulty of resolving them. Firstly, decentralized data increases annotation costs as it must be annotated in separate locations, potentially with different principles. For this combined challenge, the application of semi-supervised learning within the framework of federated learning can reduce the need for extensive annotations in decentralized data. Secondly, the interplay between insufficient training data and annotations exacerbates

the performance degradation. For this combined challenge, one potential solution is to generate more training augmentations using generative models equipped with image registration and various data augmentation techniques. Thirdly, the strict accuracy requirements often conflict with the challenges of imperfect test data. To be specific, low-quality test samples typically result in accuracy drops, thereby preventing networks from meeting the required standards. To overcome this, interactive segmentation models need to improve to better adapt to imperfect test data. This thesis considers the three challenges and provides the corresponding solutions, including the insufficient annotations challenge (§3 and §4), the missing modality challenge (§5) and the strict accuracy requirement challenge (§6). Additionally, the combined challenge of insufficient annotations and missing modality is also considered in §4.

1.2 Research Objectives

The objectives of the project are as follows:

- To enhance the accuracy of medical image segmentation with insufficient annotations, I conduct research on generating diverse, realistic and labeled training samples using image registration and variational autoencoder (VAE) networks.
- To improve the accuracy of barely-supervised brain tumor segmentation, I conduct research on effectively exploiting unlabeled data through not only the segmentation foundation model, *i.e.*, Segment Anything Model (SAM), but also the consistency supervision derived from full-modal and incomplete-modal images.
- To improve the accuracy of brain tumor segmentation under missing modality scenarios, I conduct research on adaptively fusing incomplete multi-modal features and designing a region-aware fusion model.
- To help medical image segmentation achieve satisfactory accuracy in real-world practice, I conduct research on developing a segmentation system that integrates automatic and interactive medical image segmentation within a unified network.

1.3 Thesis Organization

This thesis is organized as follows:

- *Chapter2*: This chapter reviews the related work on medical image segmentation (MIS), including standard MIS, atlas-based MIS, one-shot MIS, semi-supervised MIS, barely-supervised MIS, incomplete multi-modal MIS and, interactive MIS.

- *Chapter3*: This chapter addresses one-shot medical segmentation problem, where only one labeled (called atlas) and a few unlabeled images are available, by generating labeled training augmentations. The generation process begins by using image registration and VAEs to learn the probability distributions of deformations, including shapes and intensities, between the atlas and unlabeled images. Thus, VAEs can generate diverse deformations that match the distributions of the whole dataset. Then, a designed warp operation applies these deformations to the atlas as well as its segmentation mask, so that diverse, realistic and labeled training augments can be synthesized. Extensive experiments on two benchmarks prove the effectiveness of the proposed data augmentation. Its excellent generalization ability is also demonstrated via experiments conducted across different datasets.

- *Chapter4*: This chapter explores barely-supervised brain tumor segmentation where minimal supervision, *i.e.*, fewer than ten labeled samples, is available. Current methods often neglect two key problems in barely-supervised segmentation: i) the insufficient labeled data may not be able to offer enough information to networks for accurately segmenting tumor areas across various cases; ii) networks might overfit to the relation of multiple modalities of the limited labeled data, thus overly depending on certain modalities while overlooking other valuable modalities during segmentation. To tackle these two problems, this chapter introduces a barely-supervised training framework, called BarelySAM. BarelySAM first employs Segment Anything Model (SAM) during training by generating pseudo labels for unlabeled data. In this manner, pre-trained knowledge exhibited in SAM can be exploited to compensate for limited knowledge in labeled data, boosting network training and thus improving performance. For the overfitting problem, Multi-modality Dependency Minimization (MDM) is designed in BarelySAM to construct various partial combinations for full-modal samples, thus enforcing networks to exploit each modality effectively. Experiments on two benchmark datasets validate the effectiveness of the integrated SAM and the designed MDM module.

- *Chapter5*: This chapter focuses on the problem of certain modalities being missing in medical images, which often happens in clinical practice. For the missing modality problem, this chapter introduces a **R**egion-aware **F**usion **Net**work (**RFNet**)

that can adaptively and effectively utilize various combinations of multi-modal data for tumor segmentation. In light of the observation that different modalities are sensitive to different brain tumor regions, I design a Region-aware Fusion Module (RFM) in RFNet to perform feature fusion from available image modalities tailored specifically to different regions. Benefiting from RFM, RFNet can adaptively segment tumor regions from an incomplete set of multi-modal images by effectively aggregating modal features. Furthermore, a segmentation-based regularizer is developed to prevent RFNet from insufficient and unbalanced training caused by incomplete multi-modal data. Specifically, apart from obtaining segmentation results from fused modal features, RFNet also segments each modal image individually from the corresponding encoded features. In this fashion, each modal encoder is compelled to learn distinctive features, thereby enhancing the representational quality of the fused features. Remarkably, extensive experiments on three benchmarks demonstrate that RFNet outperforms the state-of-the-art significantly.

- *Chapter6*: This chapter introduces S2VNet, a universal framework that leverages **S**lice-**to**-**V**olume propagation to unify automatic/interactive segmentation within a single model and one training session. Inspired by clustering-based segmentation techniques, S2VNet makes full use of the slice-wise structure of volumetric data by initializing cluster centers from the cluster results of previous slice. This enables knowledge acquired from prior slices to assist in the segmentation of the current slice, further efficiently bridging the communication between remote slices using mere 2D networks. Moreover, such a framework readily accommodates interactive segmentation with no architectural change simply by initializing centroids from user inputs. S2VNet distinguishes itself by swift inference speeds and reduced memory consumption compared to prevailing 3D solutions. It can also handle multi-class interactions, with each of them serving to initialize different centroids. S2VNet demonstrates state-of-the-art performance that surpasses task-specified solutions on both automatic/interactive setups across three volumetric datasets.

CHAPTER 2

LITERATURE REVIEW

This chapter presents a survey of the literature on i) medical image segmentation, that is volumetric medical segmentation (§2.1); ii) medical image segmentation with limited annotations, including atlas-based medical segmentation (§2.2), one-shot medical segmentation (§2.3), semi-supervised medical segmentation (§2.4), and barely-supervised medical segmentation (§2.5); iii) incomplete multi-modal tumor segmentation (§2.6); and iv) interactive medical segmentation (§2.7).

2.1 Volumetric Medical Segmentation

Medical images are usually scanned in a volume-wise manner to better capture the 3D nature of human anatomical or pathological structures. Consequently, recent research mainly pay attention to volumetric medical segmentation and can be broadly grouped into two categories [248]: **slice-wise** and **volume-wise**. The **slice-wise** methods [96, 172, 177, 205, 250] usually split 3D images into 2D slices along the z-axis, and then segment them separately. Since the proposal of [153], there has been a research surge based on the U-shaped architecture [40, 57–59, 61, 68, 75, 102, 144, 174, 175, 193, 209, 214, 225, 236, 241, 246]. Such paradigm enjoys fast inference but makes no use of the 3D structure of images. In contrast, **volume-wise** methods [6, 33, 36, 47, 69, 79, 81, 85, 89, 98, 140, 147, 183, 192, 217, 219, 226, 247, 248] directly process 3D images by extending 2D operations to their 3D counterparts. While capturing spatial context in three dimensions, they are inefficient in establishing meaningful connections between

distant regions due to the limited receptive field of CNNs[25]. Recently, efforts have been made to leverage Transformer to capture long-range dependencies[18, 56, 63, 64, 117, 122, 145, 173, 185, 190, 194, 203, 212, 224, 235]. However, the inputs are still 3D image patches that contains only nearby slices, remaining unable to bridge remote slices.

2.2 Atlas-based Medical Segmentation

An atlas is defined as a template image and its segmentation mask, while atlas-based medical segmentation methods [71, 93] aim to segment target images with the help of only one or a few atlases. Single-atlas-based segmentation methods [46] begin by aligning the atlas to target images with image registration, and then transfer atlas labels to target image masks according to the alignment. Multi-altas-based segmentation is based on single-atlas-based segmentation but improves performance in other aspects, such as atlas selection [211] and label fusion [44, 211]. Atlas-based medical segmentation typically exploits limited samples and often relies on hand-crafted features; thus, it cannot enjoy the recent advances in data-driven deep learning techniques. Therefore, altas-based medical segmentation struggles to deal with complex and varied scenarios encountered in real-world practice.

2.3 One-shot Medical Segmentation

To remedy the deficiency of altas-based medical segmentation, one-shot medical segmentation [71, 93] extend single-atlas-based segmentation by additionally incorporating unlabeled training data. In this manner, deep neural networks can driven by these data with minimal annotation efforts to achieve satisfactory segmentation.

Wang *et al.* [188] accomplish one-shot medical segmentation also through image registration, similar to previous atlas-based segmentation studies [46]. Meanwhile, they additionally include unlabeled data so that they can achieve better image registration with data-driven learning. In particular, they develop a forward-backward consistency training scheme to exploit unlabeled data to optimize a well-performed registration network. However, employing image registration for segmentation is indirect and error-prone. Specifically, registration networks still suffer misalignment, leading to inferior segmentation results. Therefore, several recent attempts [207, 227, 244] consider to generate labeled training augments with image registration and train a network with these augments to achieve explicit segmentation. For example, Zhao *et al.* [227] leverage

image registration to learn shape and intensity deformations between the atlas and unlabeled images. Then, labeled training augments can be synthesized by applying these deformations to the atlas. With the help of these synthesized data, one segmentation network can be effectively trained for explicit segmentation. In [207, 244], networks of image registration and segmentation are jointly optimized so that the image registration networks can learn to generate more suitable training augments for segmentation network learning. These attempts regarding training data generation make great progress. However, the deterministic nature of image registration in these studies limits the variety of training augmentations, hindering the segmentation networks from reaching optimal performance. Even worse, these techniques primarily rely on image registration, which usually underperforms in irregular regions. Consequently, they are ill-suited for the abnormal area segmentation, such as brain tumor segmentation.

2.4 Semi-supervised Medical Segmentation.

This task focuses on achieving accurate medical image segmentation by leveraging knowledge from a small set of labeled images and a vast collection of unlabeled images. Recent research for semi-supervised medical segmentation have explored various techniques, such as self-training methods [7, 110, 158, 180], adversarial learning methods [51, 106, 146, 197], registration-based methods [43, 166, 189, 208, 228, 245], multi-task methods [22, 26, 87, 106, 131, 159], uncertainty-based methods [19, 110, 128, 158, 195, 201, 213], logic-induced methods [111], and consistency-based methods [10, 19, 34, 51, 55, 62, 70, 109, 128, 131, 186, 195, 202, 206, 210, 213, 220, 231]. In the following, we will only review the most popular technique, consistency-based methods, for brevity.

In general, consistency-based methods can be divided into three groups: First, consistency comes from training samples [10, 55, 109, 201, 202]. Li *et al.*[109] encourage consistent predictions from the same input with different augmentations, including rotation, flipping, and scaling augmentation, while Bortsova *et al.*[10] build consistency based on elastic transformation. Xia *et al.*[201] develop consistency among different image views to train the network. Second, consistency is drawn from networks [34, 51, 62, 70, 128, 195, 213, 231]. The methods [51, 186, 201, 210] leverage consistency between predictions from two networks with different initialization while others [70, 128] take advantage of the consistency of predictions from different stages in hierarchical architectures. Recently, exponential moving averaging (EMA) [34, 62, 195, 213, 231]

has been extensively studied to capture temporal consistency between networks during training. Third, consistency roots from different tasks [131]. Luo *et al.*[131] propose to leverage unlabeled data by enforcing consistent predictions from two distinct tasks, *i.e.*, standard pixel-wise segmentation and signed distance map-based segmentation.

2.5 Barely-supervised Medical Segmentation

Semi-supervised medical segmentation can largely reduce the demand for annotations but still necessitates a considerable amount of labeled images. For example, 16 out of 80 images need to be annotated for left atrium (LA) segmentation in [127]. To further scale down the demand for labeled images, several efforts [17, 103, 116, 198] have been made for barely-supervised medical image segmentation, where extremely limited labeled data is available. For example, Cai *et al.*[17] and Li *et al.*[103] only use slice-wise annotations for volume medical image segmentation, whereas Lin *et al.*[116] and Wu *et al.*[198] focus on scenarios with fewer than 10 labeled images. To be specific, Cai *et al.*[17] and Li *et al.*[103] begin by extending slice-wise labels to volume-wise labels with image registration, and then exploit these volume-wise labeled data as well as unlabeled data to train segmentation networks. However, incorporating slice-wise annotations often encounters challenges in selecting appropriate slices, as the areas of interest may appear at various locations in various shapes. Lin *et al.*[116] and Wu *et al.*[198] reduce the impact of limited annotations by effectively using unlabeled data. To this end, they employ consistency among multiple networks and leverage comprehensive information from these networks to generate better pseudo-labels.

2.6 Incomplete Multi-modal Tumor Segmentation

Medical images, derived from various imaging techniques and operations, often contain various modalities. For example, MRI imaging for brain tumor segmentation usually provides four modalities, including T1, T2, Flair and T1ce. These modalities can provide comprehensive information for better segmentation. However, multi-modal segmentation usually encounters the missing modality problem [123, 168, 221, 232] because of diverse patient conditions and mistaken imaging operations. Therefore, several research efforts have been made in incomplete multi-modal brain tumor segmentation, which is more practical but more challenging compared with the standard one [65, 141, 160, 230].

Shen *et al.*[161] treat various missing modalities as different domains and then leverage adversarial learning to project images from these domains into a unified feature space during segmentation. However, since it is difficult to align distinct and diverse distributions simultaneously, their method can only handle a small number of missing modalities. Zhou *et al.*[237] generate the features of missing modalities according to the correlations between different modalities. However, their method may not be suitable when few modalities are available because only one or two modalities are not enough to generate reliable features for the missing modalities.

In addition to feature alignment [161] and feature completion [237], several prior work [23, 48, 66] attempt to leverage feature fusion to solve the missing modality problem: Havaei *et al.*[66] aggregate partial modalities by calculating the mean and variance of the available features. Dorent *et al.*[48] embed all observed modalities into a shared latent representation by employing a multi-modal variational auto-encoder. Chen *et al.*[23] aggregate incomplete modalities via concatenation and leverage feature disentanglement jointly to achieve a modality-invariant and discriminative representation.

2.7 Interactive Medical Segmentation

Though achieving promising performance, automatic medical segmentation methods still face challenges in clinical applications due to the severe biological variation present in medical images[239]. In response to this, interactive medical segmentation [9, 11, 12, 38, 52, 108, 113, 132, 143, 178, 182] is emerging as a practical strategy to improve accuracy by incorporating user interactions, which includes bounding boxes[151, 240], scribbles[3, 4, 100, 182, 223], clicks[94, 130, 133, 181], and extreme points[88]. Moreover, endeavors have been striven to enhance accuracy by emphasizing the effective integration of user interactions, such as extracting informative cue maps[130, 133, 181], or adapting networks to inference images[157, 179]. Recently, alternate research[119, 120, 162, 238, 239] explores interactive segmentation in a mask propagation manner, *i.e.*, wrapping the mask of previous slice according to affinity matrix to predict the next slices.

Since interactive medical segmentation can attain high levels of segmentation accuracy, it finds utility in the annotation process. Several recent studies have endeavored to enhance the labeling process by integrating interactive segmentation with other methodologies, such as few-shot learning [52], active learning [108, 178], weakly-supervised learning [143], and reinforcement learning [113, 132]. Further, studies like [9, 38] have developed workflows for medical image labeling rooted in interactive segmentation.

CHAPTER 3

ONE-SHOT MEDICAL IMAGE SEGMENTATION WITH STATISTICAL-DISTRIBUTION-MODELING-BASED DATA GENERATION

3.1 Introduction

Medical image segmentation aims to partition medical images, such as magnetic resonance imaging (MRI) images, into different anatomic regions. It plays a crucial role in numerous medical analysis applications, *e.g.*, computer-assisted diagnosis and treatment planning. In recent years, benefiting from deep convolution neural networks (CNNs), fully supervised medical image segmentation methods [28, 242] have been extensively studied and achieved promising progress. However, labeling anatomic regions for large-scale 3D images requires a huge amount of time and expert knowledge. Hence, obtaining sufficient labelled data becomes the bottleneck of fully supervised segmentation methods.

One-shot medical image segmentation has been proposed to reduce the demand for copious labeled data. The prior arts [139, 148, 154, 155] mainly adopt hand-crafted data augmentations, such as random elastic deformations, to generate new labeled images to improve segmentation performance. However, those methods often generate non-realistic images since they do not take the distribution of real images into account. Thus, their networks usually fail to generalize well on real data. Recently research [21, 188, 207,

This chapter is based on joint work [43] with Xin Yu and Yi Yang, presented primarily as it appears in the AAAI 2021 proceedings.

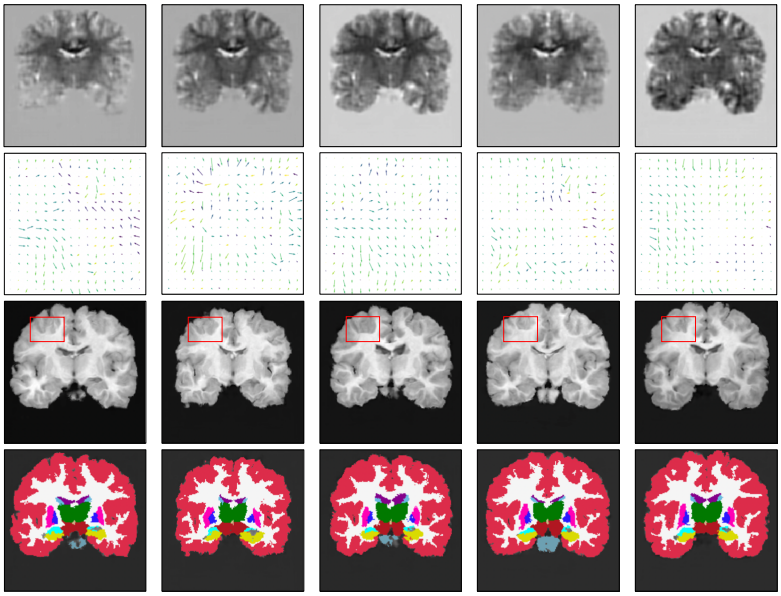

Figure 3.1: Illustration of our generated diverse deformations. From top to bottom: intensity offsets, shape deformations, synthesized images using the corresponding deformations and segmentation labels. Red frames highlight variations.

227, 244] mainly focuses on deep learning-based data augmentation. Those methods often leverage image registration to obtain shape and intensity differences between the only labeled image and other unlabeled images, and then combine the learned shape and intensity deformations to generate new images for segmentation.

Considering the domain gap and insufficient variations of synthesized data by previous methods, we aim to develop a novel medical image (*i.e.*, MRI) augmentation method to address one-shot medical image segmentation tasks. To achieve this goal, we propose a probabilistic data augmentation approach to generate sufficient training images while ensuring they follow the distribution of real MRI images in terms of brain shapes and MRI intensities, as shown in Fig 3.1. Thus, our segmentation network trained on our synthesized data will be robustly adapted to real MRI images.

In this work, we first employ image registration to obtain the shape deformations and intensity changes between an unlabeled MRI image and the atlas. However, since registration errors might occur in the registration procedure, directly classifying the registered images will lead to erroneous segmentation results. The prior art [227] combines the registered deformation fields and intensity changes to generate new labeled images and exploits them to train a segmentation network, thus mitigating registration errors. However, [227] cannot provide new deformation fields and intensity changes. Therefore, the variety of generated images is still limited.

In contrast to prior work, we propose to exploit two variational autoencoders (VAEs) to capture the probabilistic distributions of deformation fields and intensity offsets with respect to the atlas. After that, our VAEs are employed to generate various profile deformations and intensity changes. The generative deformation fields and intensity variations are used to synthesize new MRI images. In this manner, our synthesized training data is not only abundant and diverse but also authentic to real MRIs. Hence, using our augmented data, we improve the performance of our segmentation network significantly and achieve superior performance to that of SOTA.

Since different MRI machines (*i.e.*, imaging sources) may lead to different characteristics in MRI images, such as intensity changes and signal-to-noise ratio, we also conduct experiments on unseen MRI sources to evaluate the robustness of our method. Thus, we propose a more challenging benchmark with an additional unseen test set. Benefiting from our generated diverse training data, our segmentation network also performs better than the state-of-the-art on unseen MRI sources, thus demonstrating the superiority of our presented probabilistic augmentation method.

Overall, our contributions are threefold:

- We devise probabilistic data augmentation based on VAEs to generate diverse and realistic training images for the downstream segmentation task.
- We propose a new challenging segmentation benchmark to evaluate the performance of our proposed method and competing methods. It contains 3D brain MRI images from different sources. Thus, we can also test the generalization ability of the methods on unseen MRI sources.
- Taking advantage of our generated images, our method outperforms the state-of-the-art one-shot segmentation algorithms on both seen and unseen image sources.

3.2 Proposed Method

In this work, we leverage an image registration network and two VAEs to generate diverse and authentic brain MRI training samples. The generative samples are then employed to improve our segmentation network. Here, we introduce the procedure of image registration as well as modeling the probabilistic distributions of those deformations via our shape and intensity 3D VAEs, respectively. After obtaining the models of the deformations, we randomly sample from the distributions of the deformations and then

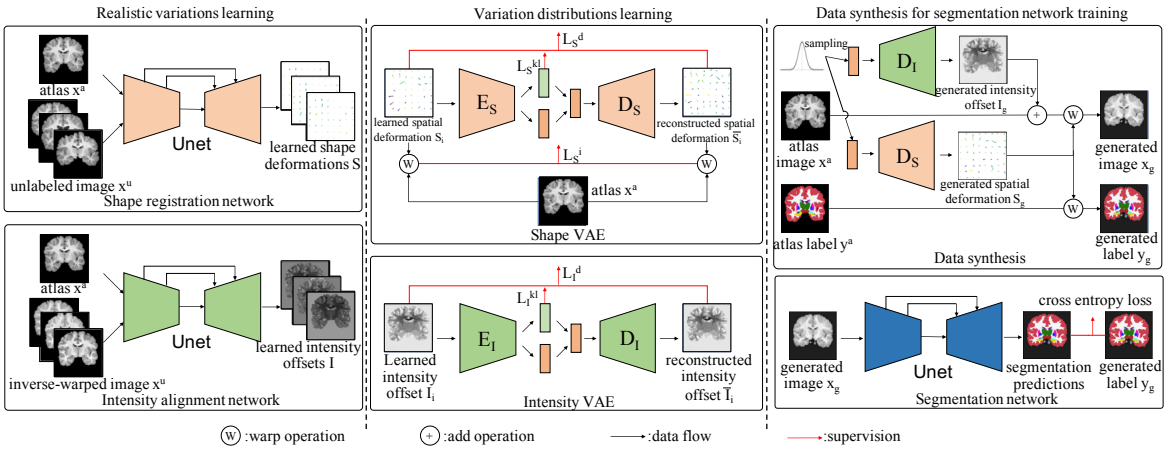

Figure 3.2: The framework of our proposed method: (i) image deformations are obtained by two Unet-based registration networks; (ii) our shape and intensity VAEs are proposed to learn the variation distributions and generate new deformations; (iii) new training samples are synthesized by applying the generated deformations to the atlas image and our segmentation network is trained on these samples.

construct new MRI images with the atlas image. The newly synthesized MRI images with their labels will be used to train our segmentation network.

3.2.1 Learning Deformations from Image Registration

Image registration [138, 249] aims to align an image to a template one, called an atlas, by learning shape deformations between them. Most existing registration-based segmentation methods [188, 207, 244] only consider the structure differences between two images. However, due to different patients, scan machines and operations, image intensities also vary. Therefore, we model both shape and intensity deformations.

First, as shown in Fig. 3.2, we leverage a Unet-based [154] registration network (named shape registration network) to learn 3D shape deformations. Denote an atlas image and its segmentation mask as (x^a, y^a) and N unlabeled images as $\{x_1^u, x_2^u, \cdots, x_N^u\}$. Taking the atlas image x^a and an unlabeled training image x_i^u as the input, the registration network is trained to propagate the atlas image x^a to an unlabeled image x_i^u by estimating a shape deformation S_i. In other words, S_i is optimized to warp x^a to x_i^u: $x_i^u \leftarrow x^a \circ S_i$, where $\circ$ represents a warping operation implemented by a differentiable bilinear interpolation-based spatial transformer layer [8]. Following the work [60], we employ a local cross-correlation (CC) loss $\mathscr{L}_{CC}$ and a deformation smoothness regularization $\mathscr{L}_S^{reg}$ to train our shape registration network in an unsupervised manner and its

objective $\mathscr{L}_{srn}$ is formulated as:

$$\begin{aligned}
\mathscr{L}_{CC} &= \sum_i \sum_{p\in\Omega} \frac{g(x_i^u, [x^a \circ S_i], p)^2}{g(x_i^u, x_i^u, p)\, g([x^a \circ S_i], [x^a \circ S_i], p)}, \\
\mathscr{L}_S^{reg} &= \textstyle\sum_i \|\nabla S_i\|_2, \\
\mathscr{L}_{srn} &= -\mathscr{L}_{CC} + \mathscr{L}_S^{reg},
\end{aligned} \tag{3.1}$$

where $g(a,b,p)$ denotes the correlation between local patches a and b on voxel p: $g(a,b,p) = \sum_{p_j}(a(p_j) - \overline{a}(p))(b(p_j) - \overline{b}(p))$, and $\overline{a}(p)$ indicates the mean of local patch intensities on p: $\overline{a}(p) = \frac{1}{\|p\|_1}\sum_{p_j} a(p_j)$. p represents a n^3 cube in a 3D image Ω and p_j denotes the pixels in the cube. We set n to 9 similar to prior methods [60]. $\mathscr{L}_{CC}$ encourages the structure similarities between two images regardless of the intensity variations while $\mathscr{L}_S^{reg}$ aims to constrain shape deformations to be smooth. ∇S_i denotes the spatial gradients of the shape variations.

Similar to learning shape deformations, we also use a Unet-based network, called intensity alignment network, to align 3D intensity deformations. As visible in Fig. 3.2, the network takes the atlas image x^a and the inverse-warped image $\hat{x}_i^u$ as input to measure the intensity deformations I_i. $\hat{x}_i^u$ is generated by aligning x_i^u to x^a, and thus $\hat{x}_i^u$ and x^a share similar profile structure. Similar to [227], we exploit a pixel-wise reconstruction loss $\mathscr{L}_{sim}$ between x^a and x_i^u and an intensity smoothness regularization $\mathscr{L}_I^{reg}$ to train our intensity alignment network. The objective function $\mathscr{L}_{irn}$ is expressed as:

$$\begin{aligned}
\mathscr{L}_{sim} &= \textstyle\sum_i \|(x^a + I_i) \circ S_i - x_i^u\|_2, \\
\mathscr{L}_I^{reg} &= \textstyle\sum_i \sum_{q_j} (1 - c^a(p_j)) |\nabla I_i(p_j)|, \\
\mathscr{L}_{irn} &= \mathscr{L}_{sim} + \lambda \mathscr{L}_I^{reg}.
\end{aligned} \tag{3.2}$$

Here, $\mathscr{L}_I^{reg}$ is designed to prevent dramatic changes of the I_i in the same brain area. $\nabla I_i(p_j)$ denotes the gradients of I_i at p_j. c^a denotes the mask of contours across different areas. λ is a trade-off weight and set to 0.02, following the work [227].

3.2.2 Diverse Image Generation via VAEs

After image registration, we obtain N shape deformations and N intensity changes from the atlas and N unlabeled images. In the work [227], these variations are directly combined to generate new labeled training images for segmentation. However, only N kinds of shape and intensity transformations are involved during training, and the diversity of the samples is not rich enough to train an accurate segmentation network.

[21] employ GANs to generate new deformations but their method requires a large number of unlabeled data to train GANs. However, we only have less than 100 unlabeled images. Thus, their method will suffer mode collapse and is not applicable in our case.

Different from previous methods, we adopt a 3D shape VAE and a 3D intensity VAE to learn the probabilistic distributions of the variations with respect to the atlas separately, since VAE does not suffer mode collapse. Furthermore, inspired by beta-VAE [13, 74], we reduce impacts of the Kullback-Leibler (KL) divergence in a conventional VAE to increase the diversity of generated samples. Doing so is also driven by the insufficiency of the training samples. After training, we sample deformations from our shape and intensity VAEs, and then generate a large number of various training images.

As illustrated in Fig. 3.2, our shape VAE first uses an encoder to project an input shape deformation into a latent vector $z = \boldsymbol{E}_S(S_i)$ and then decodes z to the image domain, *i.e.*, a reconstructed shape deformation $\overline{S}_i = \boldsymbol{D}_S(z)$. During training, three objectives, including KL divergence $\mathscr{L}_S^{kl}$ and pixel-wise reconstruction losses on the deformations $\mathscr{L}_S^d$ and image intensities $\mathscr{L}_S^i$, are employed to train our shape VAE, written as:

$$
\begin{aligned}
&\mathscr{L}_S^{kl} = \textstyle\sum_i D_{kl}(q(z|S_i)||p(z)),\\
&\mathscr{L}_S^{d} = \textstyle\sum_i \|S_i - \overline{S}_i\|_2,\\
&\mathscr{L}_S^{i} = \textstyle\sum_i \|(x^a \circ S_i) - (x^a \circ \overline{S}_i)\|_2,\\
&\mathscr{L}_S = (\mathscr{L}_S^d + \mathscr{L}_S^i) + \beta \mathscr{L}_S^{kl},
\end{aligned}
\tag{3.3}
$$

where $\mathscr{L}_S^{kl}$ forces the distribution of latent vector z to be a standard normal distribution, (*i.e.*, $z \sim \boldsymbol{N}(0,1)$), $q(z|\cdot)$ denotes the posterior distribution, $p(z)$ denotes the Gaussian prior distribution modeled by a standard normal distribution, and β is a hyper-parameter controlling rigidity of the distributions of the latent variable z and the quality of reconstruction. Here, we not only compare the decoded shape deformations with the input ones but also measure the differences between the warped images by the input shape deformations and reconstructed ones.

Smaller β indicates less attention is paid to the KL divergence loss during training and will result in a larger KL divergence between the posterior and prior distributions. As suggested by [13], larger KL divergence allows a latent vector to reside in a large space. In other words, smaller β allows our VAE to preserve variations of input images, especially when the training samples are scarce. Therefore, using a small β is preferable when training samples is insufficient. Moreover, since the latent space has been enlarged, more variations can be generated from this latent vector space via our decoder in the testing phase. Therefore, we set β to a small value (*i.e.,* 0.1) for all the experiments.

It is worth noting that we employ both $\mathscr{L}_S^d$ and $\mathscr{L}_S^i$ as the reconstruction loss for our shape VAE instead of only reconstructing network inputs by $\mathscr{L}_S^d$ as in the original VAE. When $\mathscr{L}_S^d$ is only employed, image structure information is neglected. In particular, shape deformations should pay attention to the consistency of image contour movements. However, $\mathscr{L}_S^d$ treats the movement of each pixel individually and thus may not perform consistent movements along the contour regions. On the contrary, the reconstruction loss $\mathscr{L}_S^i$ is sensitive to the movements of image contours because image intensities around contours change dramatically. In other words, small reconstruction errors in the deformations of the contours will lead to large intensity differences between two warped images. On the other hand, since $\mathscr{L}_S^i$ only measures intensity similarities, it may not preserve boundary information when two areas have similar intensities. Therefore, we leverage both $\mathscr{L}_S^i$ and $\mathscr{L}_S^d$ as the reconstruction loss in learning our shape VAE.

Similar to our shape VAE, we employ a VAE to model the distribution of the intensity variations with respect to the atlas. Here, we adopt the standard KL divergence loss and a pixel-wise reconstruction loss to train our intensity deformation VAE, expressed as:

$$\begin{aligned} \mathscr{L}_I^{kl} &= \textstyle\sum_i D_{kl}(q(z|I_i)||p(z)), \\ \mathscr{L}_I^{d} &= \textstyle\sum_i \|I_i - \overline{I}_i\|_2, \\ \mathscr{L}_I &= \mathscr{L}_I^{d} + \beta \mathscr{L}_I^{kl}, \end{aligned} \tag{3.4}$$

where $\overline{I}_i$ is the intensity deformation reconstructed from I_i.

After modeling the deformation distributions, our shape and intensity VAEs are exploited to generate diverse variations by random sampling. Specifically, in the process of the generation, the decoders $\boldsymbol{D}_S$ and $\boldsymbol{D}_I$ take random latent vectors sampled from a Gaussian distribution $\boldsymbol{N}(0,\sigma)$ as input and output various shape deformations S_g and intensity changes I_g, respectively. Then, our synthesized labeled training images are constructed as:

$$x_g = (x^a + I_g) \circ S_g, \qquad y_g = y^a \circ S_g, \tag{3.5}$$

where x_g and y_g represent the synthesized images and their corresponding segmentation masks. Note that, different from MRI images, segmentation masks are warped by a nearest-neighbor interpolation-based 3D spatial transformer layer [8].

3.2.3 Segmentation Network

Once augmented training samples are obtained, we can train our segmentation network on those samples. For fair comparisons[227], we employ the same 2D Unet with a five-

layer encoder and a five-layer decoder to segment each slice of 3D images individually. In the encoder and decoder, we use 3×3 2D convolutional operations followed by LeakyReLU layers. 2×2 Max-pooling layers are employed to decrease the resolution of features, while upsampling layers are used to increase resolution by a factor of 2.

In each training iteration, we construct a batch by randomly sampling slices from 3D images. The standard cross-entropy loss is applied as described:

$$\mathscr{L}_{CE} = -\sum_{i=1}^{W}\sum_{j=1}^{H}\frac{1}{H\cdot W}\log\frac{\exp(y_p[i,j,y_g(i,j)])}{\sum_{k=1}^{K}\exp(y_p[i,j,k])}, \tag{3.6}$$

where y_p is the predicted mask from our segmentation network g (*i.e.*, $y_p = g(x_g;\theta)$) and θ denotes the parameters of the segmentation network. W and H denote the width and height of a 2D slice, respectively. K indicates the number of anatomical components in an MRI image. Similar to the training process, every 3D image is split into 2D slices and segmented in a slice-wise fashion in the testing phase.

Although we incorporate two VAEs to generate labeled data, they are only used in the training phase. During testing, only our segmentation network is exploited. Therefore, our method does not increase the network parameters and FLOPs during inference and thus can be deployed as easily as previous work.

3.2.4 Implementation Details

We employ the same network architecture for our shape and intensity VAEs, and the VAEs are 3D VAEs since deformations should be consistent in 3D space. In the 3D VAE networks, group normalization [200] is employed. For the activation function, we use LeakyReLU and ReLU for the encoder and the decoder, respectively. The dimension of the latent vector is set to 512.

During training, Adam [91] optimizer is used to train our VAEs, where β_1 and β_2 are set to 0.5 and 0.999, respectively. Considering GPU memory limitation, we set the batch size to 1. The learning rate is fixed to $1e^{-4}$ for the whole 40k training iterations. The hyper-parameter β in both two VAEs is set to 0.1. In generating deformations, the shape VAE and the intensity VAE take latent vectors sampled from $\boldsymbol{N}(0,10)$ as input in order to achieve more diverse data.

For other networks (*i.e.*, shape registration, intensity alignment and segmentation networks), a default Adam with $1e^{-4}$ learning rate is employed. For the shape registration and intensity alignment networks, the batch size is set to 1 and the networks are trained for 500 epochs. For the segmentation network, the batch size is set to 16 and the network

is trained for 40,000 iterations. Our method is trained and tested on an Nvidia Tesla V100 GPU and achieves similar results on Keras with a TensorFlow backend.

Note that, in training the 3D VAEs and segmentation networks, images are generated on the fly, and thus, we train these networks in terms of iterations. In training registration and alignment networks, only 82 MRI images will be aligned to the atlas, and thus, we train the networks in terms of epochs.

3.3 Experiments

In this section, we begin by comparing our proposed method with the latest one-shot-based methods. Following that, we analyze the contributions of each component within our method. For fair comparisons, experiments are conducted on the same dataset as previous work [60, 188, 227]. Moreover, we propose a more challenging MRI benchmark to evaluate the generalization capability of the proposed method.

3.3.1 Experimental Setup

Dataset: CANDI dataset [86] consists 103 T1-weighted brain MRI images from 46 females and 57 males. In this dataset, four types of diagnostic groups are considered, including bipolar disorder without psychosis, bipolar disorder with psychosis, schizophrenia spectrum, and healthy controls. In the experiments, we use the same train and test splits as in [188]. To be specific, 20/82/1 images are employed as test/unlabeled training/atlas images. Following the work [188], We crop a volume of $160 \times 160 \times 128$ from the center of an original MRI image. For segmentation, similar to [188], we consider 28 primary brain anatomical areas.

Evaluation Metric: Dice coefficient [39] is employed to evaluate the segmentation performance, written by:

$$\text{Dice}(M^k_{y_p}, M^k_{y_{gth}}) = 2 \cdot \frac{M^k_{y_p} \bigcap M^k_{y_{gth}}}{|M^k_{y_p}| + |M^k_{y_{gth}}|}, \tag{3.7}$$

where $M^k_{y_p}$ and $M^k_{y_{gth}}$ denote segmentation masks of the anatomical region k with predicted labels y_p and its corresponding ground-truth y_{gth}.

Larger Dice scores indicate more overlaps between predictions and ground-truth labels and thus represent better segmentation performance. To showcase the segmentation capability of methods more effectively, we report not only a mean Dice score but also its corresponding standard variance, minimum Dice score and maximum Dice score.

Method	Mean(std)	Min	Max
Supervised learning	88.3(1.7)	83.5	90.3
VoxelMorph [60]	76.0(9.7)	61.7	80.1
DataAug [227]	80.4(4.3)	73.8	84.0
LT-Net [188]	82.3(2.5)	75.6	84.2
Ours	**85.1**(**1.9**)	**80.2**	**87.8**

Table 3.1: Quantitative segmentation results on CANDI. Fully-supervised segmentation accuracy is reported as an upper bound. Mean/Min/Max/ are reported to indicate the middle/worst/best Dice scores. “std” denotes the standard deviations.

Method	Shape	Intensity	VAE	Mean(std)	Min	Max
Registration based (VoxelMorph)				76.0(9.7)	61.7	80.1
Segmentation with data augmentation	✓			81.7(5.6)	65.4	87.4
	✓		✓	83.5(4.2)	71.1	**87.8**
	✓	✓		84.2(**1.7**)	79.7	86.5
	✓	✓	✓	**85.1**(1.9)	**80.2**	**87.8**

Table 3.2: Analysis of data augmentation. Shape and Intensity denote that the deformations from image registration. VAE indicates that the deformations are generated from our VAEs.

Method	Mean(std)	Min	Max
$\mathscr{L}_S^d$	81.3 (**2.8**)	**74.4**	85.0
$\mathscr{L}_S^i$	82.3(6.2)	63.9	87.7
$\mathscr{L}_S^d + \mathscr{L}_S^i$	**83.5**(4.2)	71.1	**87.8**

Table 3.3: Analysis of reconstruction losses in the shape VAE.

3.3.2 Comparison to State-of-the-arts

Two latest one-shot atlas based method, *i.e.*DataAug [227] and TL-Net [188], are compared. In addition, one unsupervised registration method *i.e.*, VoxelMorph [8] is applied to one-shot medical image segmentation for comparison. VoxelMorph and TL-Net leverage a registration network to align the atlas to test images, simultaneously applying it to the segmentation mask of the atlas to generate segmentation predictions for the corresponding test images. DataAug employs image registration to achieve shape and intensity transformation and then augments the atlas image with the attained transformation for segmentation network training. Note that these methods do not generate new deformations while our method does.

In Table 3.1, our method enhances the Dice score by 2.8% in comparison to the current best method LT-Net [188]. Moreover, our method also obtains the smallest variance, demonstrating that our method is more robust and effective.

Method	Seen			Unseen		
	Mean(std)	Min	Max	Mean(std)	Min	Max
Supervised learning	87.6(2.7)	79.3	91.1	85.9(1.7)	81.3	87.5
VoxelMorph	70.3(11.6)	33.1	82.5	62.9(13.2)	32.3	79.6
DataAug	69.6(9.02)	39.7	80.4	64.3(9.9)	35.0	77.2
Ours	**76.7(7.4)**	**53.2**	**86.5**	**74.8(6.6)**	**54.1**	**83.3**

Table 3.4: Quantitative segmentation results on our newly proposed ABIDE benchmark.

3.3.3 Ablation Study

To demonstrate the effectiveness of our VAEs, we compare four different types of data augmentation in Table 3.2. As simply applying intensity offsets to the atlas does not change the segmentation mask, synthesized images will have the same segmentation labels, thus leading to a trivial segmentation solution.

3.3.3.1 Effectiveness of our VAEs

As indicated in Table 3.2, compared with direct registration, data augmentation-based segmentation methods achieve better segmentation accuracy. Note that all the augmentation methods learn the shape deformations similar to VoxelMorph. Compared with the data augmentation methods using deformations from image registration, our VAEs can generate richer data for training a segmentation network, thus leading to better performance. Moreover, we observe that intensity deformations make great contributions to segmentation performance, and various intensity changes facilitate the generalization of our segmentation network. In Table 3.2, we also notice that our network employing registered shape and intensity deformations achieves better performance than DataAug. This is because DataAug pre-trains a segmentation network with an l2 loss and does not employ the atlas in training the segmentation network. Thus, using the atlas for training segmentation networks is important.

Effectiveness of the Combined Reconstruction Loss. To demonstrate the effectiveness of our combined reconstruction loss *i.e.*, $\mathscr{L}_S^d + \mathscr{L}_S^i$, we train the shape VAEs with $\mathscr{L}_S^d$, $\mathscr{L}_S^i$ and $\mathscr{L}_S^d + \mathscr{L}_S^i$, respectively, and then apply them to augment data. To avoid the influence of the intensity augmentation, we do not use intensity augmentation and the segmentation results are reported in Table 3.3. As seen, our combined reconstruction loss is more suitable for the learning and generation of shape deformations.

Hyper-parameter β in Eq. (3.3) and Eq. (3.4), and σ for sampling latent codes. As aforementioned, a small β introduces more diversity into the generated deformations, thus improving the segmentation performance. Figure 3.3 manifests that using a small

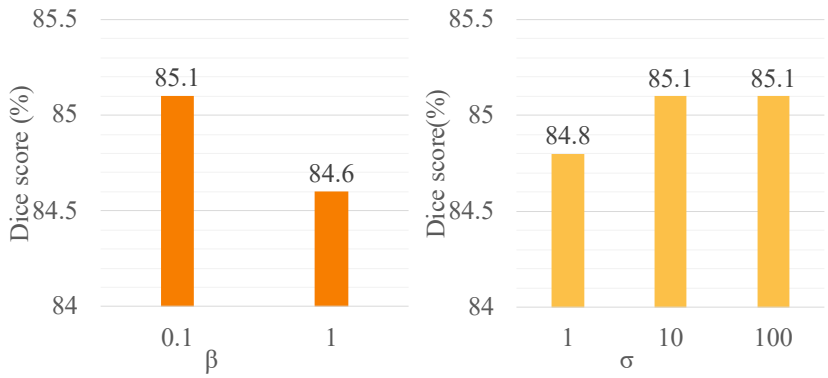

Figure 3.3: Analysis of hyper-parameter β and σ. β controls the weight of the KL divergence and σ is the standard deviation of a prior Gaussian distribution $\boldsymbol{N}(0,\sigma)$ in VAEs.

β, we achieve better segmentation accuracy. Thus, in all the experiments, β is set to 0.1. Furthermore, as illustrated in Fig. 3.3, the segmentation performance degrades when the standard deviation σ for sampling latent codes is set to 1. This is because we employ a small β to enforce the KL divergence during training, and the latent vector space would deviate from the standard normal distribution. Thus, we use a larger σ to sample latent codes. Figure 3.3 shows the segmentation accuracy is similar when σ is set to 10 and 100. Thus, σ is set to 10 for all the experiments.

3.3.4 Our Proposed ABIDE Benchmark

Since the MRI images in CANDI are collected from only one source, the variances (including shape and intensity) mainly come from different individuals. However, different MRI machines and operations may also lead to variations. Therefore, to validate the robustness of our method, we propose a new standard segmentation benchmark, called ABIDE benchmark, as visible in Fig. 3.4.

From **A**utism **B**rain **I**maging **D**ata **E**xchange (ABIDE) database, collected from 17 international sites, we sample 190 T1-weighted MRI images from ten imaging sources and split them into 100, 30, 60 volumes for training, validation and testing, respectively. These testing images form a *seen* test set. As suggested by [60], the image most similar to the average volume is chosen as the atlas. We also sample 60 images from the rest imaging sources as an *unseen* test set. All the volumes are resampled into a $256\times256\times256$ with 1mm isotropic voxels and then cropped to $160\times160\times192$. 28 anatomical regions are annotated by FreeSurfer [54].

As our benchmark collects images from multiple sites and contains an unseen test,

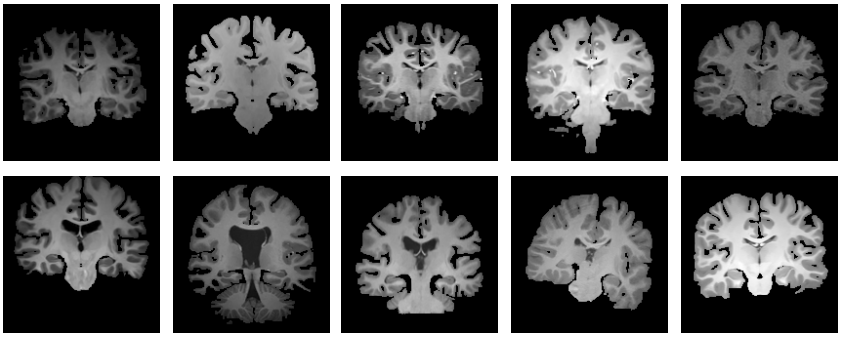

Figure 3.4: Illustration of significant variances in our ABIDE benchmark. The 96-th slices of ten 3D MRI images are shown. (Top row: images from seen datasets; Bottom row: images from unseen datasets.) More images are shown in supplementary materials.

it is more challenging and is also able to evaluate the robustness of a method. We compare our method with VoxelMorph[60] and DataAug[227] [1] in Table 3.4. The fully supervised performance is also reported as the upper bound. Compared with the prior arts, we achieve superior performance on both seen and unseen datasets, demonstrating the effectiveness of our data augmentation method. In addition, our performance only degrades 1.9% on the unseen test dataset while the performance of the competing methods decreases more than 5%. This demonstrates that our method achieves a better generalization ability with the help of our generated various deformations.

3.4 Conclusion

This chapter introduces the devised 3D VAE-based data augmentation scheme for one-shot medical image segmentation. Specifically, the shape and intensity deformation VAEs are developed to learn the deformation distributions of unlabeled real images relative to the only labeled image. Subsequently, these two VAEs generate numerous shape and intensity deformations, which can be randomly combined to further enhance diversity. Finally, the combined deformations are applied to the labeled images, generating realistic, diverse and labeled training samples that facilitate sufficient network training. Extensive experiments demonstrate the superiority of the proposed data augmentation scheme on both seen and unseen datasets. However, the effectiveness of this data augmentation scheme depends on precise image registration, which limits its applicability in lesion segmentation. The challenge emerges as the shapes and locations of lesions

[1] Since [188] do not release their code, we do not include their results.

vary, resulting in frequent registration misalignments. For this issue, barely-supervised learning approaches are being considered for lesion segmentation, as discussed in §4.

CHAPTER 4

BARELY-SUPERVISED BRAIN TUMOR SEGMENTATION VIA EMPLOYING SEGMENT ANYTHING MODEL

4.1 Introduction

Automatic brain tumor segmentation is crucial for clinical assessment and treatment planning. In the past few years, considerable research [39, 50, 65, 79, 83, 84, 141] has been devoted to exploring and advancing fully-supervised brain tumor segmentation. However, these fully-supervised approaches require a substantial number of labeled images to attain satisfactory accuracy. Unfortunately, annotating brain tumors requires significant expertise and effort. Consequently, achieving discriminative tumor segmentation networks with limited labeled data is highly desirable in clinical practice.

To mitigate the problem of high annotation costs, numerous research efforts [10, 34, 51, 55, 62, 70, 109, 128, 131, 195, 201, 202, 213, 231] have been undertaken in semi-supervised medical image segmentation where only a small portion of data requires annotation. Nonetheless, these methods still necessitate a considerable amount of annotated images. For instance, 16 out of 80 images need to be annotated for left atrium (LA) segmentation in [127]. Recently, there has been an exploration of one-shot medical image segmentation methods [43, 166, 208, 228, 245] inspired by atlas-based segmentation [1, 16], where only a single labeled image is required. Nevertheless, these

This chapter is based on joint work [42] with Hongming Liu, presented primarily as it will appear in the TCSVT 2024.

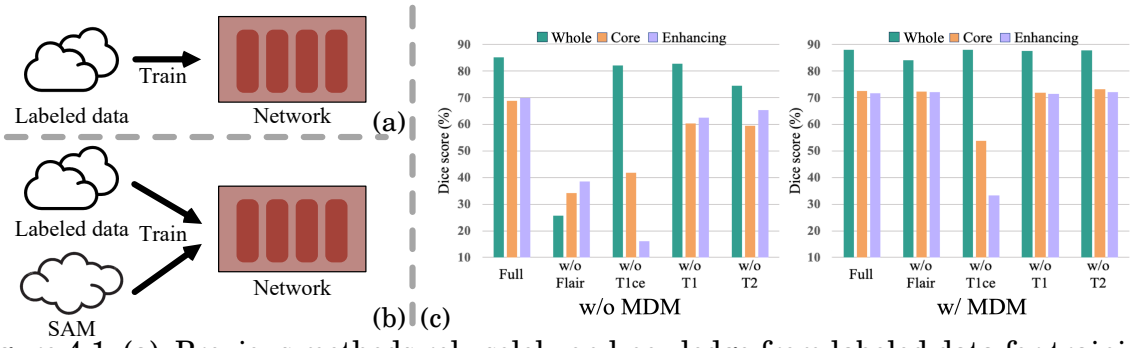

Figure 4.1: (a): Previous methods rely solely on knowledge from labeled data for training networks. (b): BarelySAM exploits the pre-trained knowledge from SAM to boost network training. (c): During inference, full and partial modalities (with one modality removed) are fed to the networks. Without the help of the proposed MDM, the network overly relies on the Flair modality and suffers obvious performance decreases in all three tumor regions, *i.e.*, whole, core, and enhancing tumors, when the Flair modality is removed.

techniques primarily rely on image registration, which usually underperforms in regions with varying shapes and locations. Consequently, they are ill-suited for abnormal area segmentation, such as brain tumor segmentation.

Given the constraints in semi-supervised and one-shot medical image segmentation, several studies [116, 198] have shifted focus towards barely-supervised learning, where a minimal amount (typically less than 10) of labeled samples are involved. These work aim to improve pseudo label quality by tackling the class imbalance problem [116] or using confidence maps [198]. Nonetheless, they only consider labeled data knowledge, as shown in Fig. 4.1(a), and overlook two critical issues inherent in barely-supervised segmentation. Firstly, the limited labeled samples cannot provide sufficient information for accurate tumor segmentation across different cases. The scarcity of labeled samples indicates that they cannot capture the complexity and variability of tumors across various patients. Therefore, when trained with these labeled samples, networks lack the ability to generalize and accurately segment tumors across diverse cases. Secondly, networks often overfit to labeled data in exploiting multiple modalities. To be specific, brain tumor segmentation typically employs four distinct modalities, and the correlations between these modalities in labeled data will train networks how to exploit modalities effectively. However, in barely-supervised learning, the networks cannot see sufficient labeled variations, where different modalities contribute to tumor segmentation differently. Therefore, the networks may be misled by multi-modality correlations in the labeled data to overly rely on certain modalities while neglecting other valuable modalities.

In this paper, we propose BarelySAM, a training framework for barely-supervised brain tumor segmentation. To address the problem of limited labeled data, we propose

to harness Segment Anything Model (SAM) [92], a class-agnostic foundation model capable of predicting fine-grained object masks according to the corresponding prompts, as depicted in Fig. 4.1(b). Specifically, our network first predicts the locations of each tumor area, which can be used to generate box and point prompts for SAM. Then, SAM-based pseudo-labels are leveraged as an additional supervision for network optimization. In this fashion, SAM offers its pre-trained knowledge to compensate for insufficient supervision from limited labeled data, leading to a robust and effective network. As employed only during training, our proposed SAM-based pseudo-label generation process does not produce any additional computational budget during the deployment phase.

To address the overfitting issue in handling multiple modalities, we propose a Multi-modality Dependency Minimization (MDM) module. Specifically, MDM re-arranges training samples with full modalities into a range of partial modality combinations and enforces the output of these diverse combinations to align closely with the ground truth. We further extend the application of MDM to unlabeled data by introducing an innovative consistency supervision mechanism. This involves encouraging consistent predictions between full modalities and other modality combinations of the same training sample. Specifically, we use pseudo labels derived from full-modality samples to guide the training of their partial-modality counterparts. Through the utilization of various modality combinations during training, MDM enables our network to effectively leverage each modality and to adaptively aggregate information from various modality combinations. Consequently, MDM can prevent our network from relying on specific correlations present in labeled data and mitigate the network's overdependence on specific modalities. Additionally, MDM enforces our model to train with various modality combinations, enabling it to process scenarios with missing modalities, as depicted in Fig. 4.1(c).

Benefitting from the incorporated SAM and proposed MDM module, BarelySAM attains impressive results in both full- and incomplete-modal brain tumor segmentation tasks on BRATS2015 and BRATS2020. In particular, our method, relying on merely 6 (2%) labeled samples, achieves accuracy comparable to a fully supervised approach in whole tumor segmentation on BRATS2020, with a slight 1.09% drop in Dice score.

Our contributions can be concluded in three aspects:

- We propose a novel barely-supervised training framework called BarelySAM. BarelySAM uses SAM [92] to generate pseudo labels as an additional supervision, employing its pre-trained knowledge to boost network training.

- BarelySAM develops a Multi-modality Dependency Minimization (MDM) module to

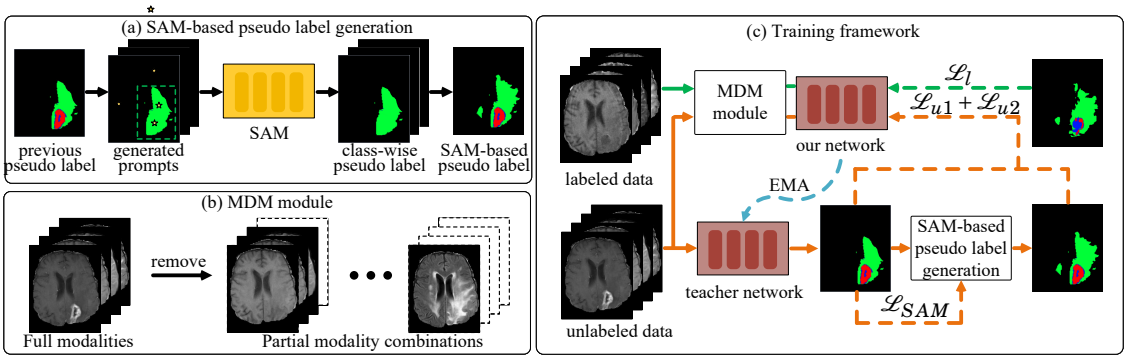

Figure 4.2: The illustration of BarelySAM. (a): SAM generates pseudo labels according to previous pseudo labels. (b): MDM re-arranges training samples with full modalities into a range of partial modality combinations. (c): Our training framework consists of our network and a teacher network. Our network is updated by gradients from $\mathscr{L}_l$, $\mathscr{L}_{u1}$, and $\mathscr{L}_{u2}$, while the teacher network is updated in an EMA manner.

prevent networks from the overfitting problem, thus exploiting modalities properly. Additionally, novel consistency supervision is devised for unlabeled data based on MDM, further facilitating network training.

- Our method inherently exhibits robustness towards incomplete modal brain tumor segmentation. Extensive experiments demonstrate that our barely-supervised approach surpasses state-of-the-art methods in both full-modality and incomplete-modality brain tumor segmentation tasks.

4.2 Proposed Method

This work tackles the problem of insufficient annotations for brain tumor segmentation. To this end, we propose BarelySAM to boost the training by incorporating SAM [92] and devising a multi-modality dependency minimization (MDM) strategy. In this section, we first introduce adapting SAM into training in §4.2.1. SAM offers its pre-trained knowledge to assist networks in distinguishing various tumor areas by generating pseudo-labels. In §4.2.2, MDM leverages diverse partial modality combinations to encourage networks to use each modality effectively, thereby reducing the risk of overfitting. In §4.2.3, the training framework with SAM and MDM is introduced.

4.2.1 SAM-based Pseudo Label Generation

During the training of barely-supervised brain tumor segmentation, only insufficient labeled data are available. These labeled data may lack enough knowledge, which enables networks to segment objects accurately across various domains or cases. For this problem, we incorporate a well-trained and generalized foundation model, *i.e.*, SAM [92], to provide additional information for better network optimization.

SAM improves network training by generating reliable pseudo labels for unlabeled data, as depicted in Fig. 4.2. To be specific, necessary box and point prompts for SAM are first generated according to the previous pseudo labels. Given that brain tumors comprise multiple regions and SAM cannot process multiple objects simultaneously, we generate class-specific prompts for SAM and produce class-wise masks. For the box prompt for each class, we use the bounding box of each tumor region. For point prompts, we randomly select 5 points per class, and then for each class, we view the corresponding 5 points as positive points and others as negative points. This method of contrastive point selection enables SAM to better differentiate between tumor areas, thus improving the quality of pseudo labels and network performance. This is evidenced in Table 4.7. Based on class-wise prompts, SAM generates predictions for each foreground class, and pseudo labels are obtained by merging these predictions. In the merging process, an area is considered as background if all class predictions identify it as such, whereas the foreground class is determined by comparing predictions across classes. The merging process is defined by,

$$\widetilde{\boldsymbol{Y}}(i,j) = \begin{cases} 0 & \text{if } \bigcup_{k=1}^{K}(\boldsymbol{P}_k(i,j) < 0.5), \\ \arg\max_k \boldsymbol{P}_k(i,j) & \text{otherwise.} \end{cases} \tag{4.1}$$

$\boldsymbol{P}_k$ denotes SAM-generated predictions (after the sigmoid function) for the k-th class and $\widetilde{\boldsymbol{Y}}$ denotes merged pseudo labels.

4.2.2 Multi-modality Dependency Minimization

Brain tumor segmentation usually exploits comprehensive information from four modalities to obtain high performance. However, in barely-supervised learning, networks cannot see sufficient labeled variations, where different modalities contribute differently in various tumor regions. Thus, networks may over-fit in the multi-modality dependency existing in the limited labeled samples, overly relying on specific modalities and neglecting some useful modalities. For this problem, we propose Multi-modality Dependency

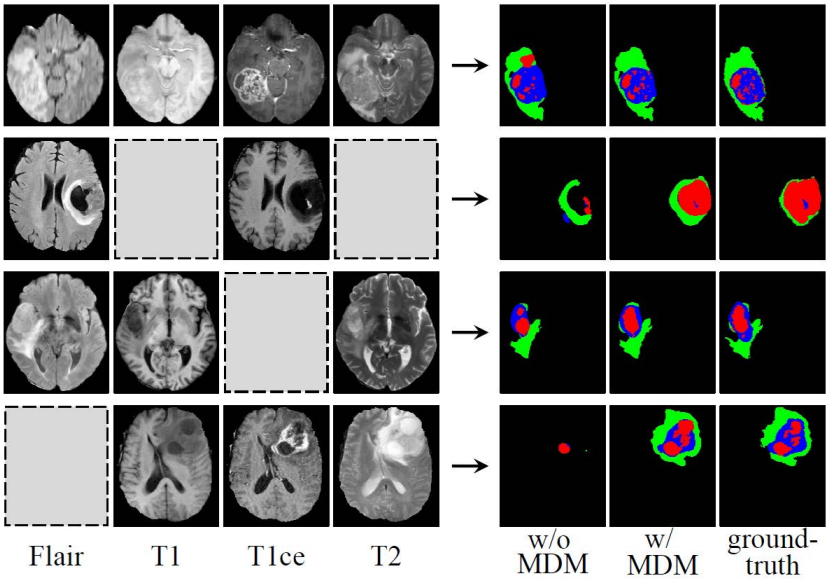

Figure 4.3: Visual predictions with and without the proposed MDM. Left: input MRI modalities, including full and partial modality combinations. Right: segmentation results produced from the networks with and without MDM and the corresponding ground-truth.

Minimization (MDM), which considers various partial modality combinations instead of paying all attention to full modalities.

As shown in Fig. 4.2, MDM firstly generates various partial modality combinations from a full-modality sample by randomly removing one or a few of its modalities and then chooses to use these combinations at probability p_s during training. In particular, fourteen partial modality combinations can be constructed by randomly erasing one (four cases), two (six cases), or three (four cases) modalities from each four-modal sample.

In this work, MDM is employed for both labeled and unlabeled data. For labeled data, MDM directly encourages predictions from various modality combinations to be similar to the corresponding ground truth. For unlabeled data, we develop novel multi-modality consistency supervision to facilitate the training. In particular, although some modalities were removed, we believe the remaining partial modality combinations could still represent the corresponding samples and thus should gain similar predictions to the full-set versions of the same samples. Therefore, as shown in Fig. 4.2, our network generates pseudo labels from full modalities of unlabeled samples first and uses these pseudo labels to supervise the predictions from corresponding partial modality combinations. We do not consider generating pseudo labels using those partial modality combinations because full modalities contain more modalities as well as information and thus are more likely to produce more accurate pseudo labels.

By additionally feeding partial modality combinations to the network, MDM is able to prevent our network from overly relying on certain modalities while enforcing to fully exploit each modality. Therefore, MDM is able to achieve better segmentation predictions by reducing multi-modal dependency and enhancing feature representations. As shown in Fig. 4.3, our network with MDM achieves more accurate segmentation predictions from full modalities. Moreover, when it comes to incomplete modalities, our network with MDM can still perform well, while the network without MDM cannot since it overly relies on specific modalities (such as the Flair modality).

4.2.3 Training Framework

As seen in Fig. 4.2, the training framework is divided into two parts: i) supervised training for the labeled data and ii) self-training with the proposed MDM-based multi-modality consistency for the unlabeled data.

For the **labeled data** $(\boldsymbol{X}^l, \boldsymbol{Y}^l)$, fully-supervised learning is exploited for network training. Besides, the deep supervision technique [49, 99] is employed to improve hierarchical representations of our model. The training loss for the labeled data, combining the Dice loss $\mathscr{L}_{dl}$ and the cross-entropy loss $\mathscr{L}_{ce}$, is defined as:

$$\mathscr{L}_l = \sum_{s}^{S} \beta_s \cdot (\mathscr{L}_{dl}(\boldsymbol{P}_s^l, \boldsymbol{Y}^l) + \mathscr{L}_{ce}(\boldsymbol{P}_s^l, \boldsymbol{Y}^l)), \tag{4.2}$$

where $\boldsymbol{P}_s^l = f(\phi(\boldsymbol{X}^l);\theta)_s$ denotes the prediction of the labeled sample $\boldsymbol{X}^l$ at the s-th stage of our network f. $\phi(\cdot)$ denotes the process of MDM (as depicted in Fig. 4.2), and θ denotes the parameters of our network. S denotes the number of stages of our UNet-based network, and β_s is the loss weighting factor for the s-th stage in deep supervision.

For the **unlabeled data** $\boldsymbol{X}^u$, we adopt a self-training scheme with the proposed MDM-based multi-modality consistency. As shown in Fig. 4.2, we firstly keep a temporal ensembling version f^{te} [167] of our network f, which is called teacher network. The teacher network is updated in an exponential moving average (EMA) manner, defined by:

$$\theta_t^{te} = \alpha \cdot \theta_{t-1}^{te} + (1-\alpha) \cdot \theta_{t-1}, \tag{4.3}$$

where θ^{te} denotes parameters of the teacher network. t denotes the t-th iteration and α is a coefficient controlling the update rate.

Then, we incorporate the multi-modality consistency based on MDM to facilitate the exploitation of unlabeled data during training. Instead of focusing on full modalities

during all training periods, the designed consistency considers partial modalities of unlabeled data and helps the network to fully exploit each modality while adaptively aggregating various modalities. To be specific, we input full modalities into the teacher network to generate pseudo labels and leverage them as targets to supervise the learning of our network with corresponding modality combinations, including partial and full modalities. Employing the teacher network to generate pseudo labels is because the teacher network can be viewed as an ensemble of our networks in previous iterations and thus is able to achieve more reliable pseudo labels. Note that, to increase the data diversity and perturbations in consistency-based learning, we additionally employ different intensity augmentation and mirror augmentation before inputting samples into our network and the teacher network. The training loss for unlabeled data is defined as:

$$
\begin{aligned}
&\boldsymbol{P}_s^{u'} = f(\phi(\boldsymbol{X}^{u'});\theta)_s, \quad \boldsymbol{P}_s^{u''} = \varphi(f^{te}(\boldsymbol{X}^{u''};\theta^{te})_s),\\
&\boldsymbol{Y}_s^{u''} = \text{argmax}(\boldsymbol{P}_s^{u''}), \quad \boldsymbol{M} = \mathbb{1}(\max(\boldsymbol{P}_s^{u''}) > \tau),\\
&\mathscr{L}_{te}(\boldsymbol{X}^{u'},\boldsymbol{X}^{u''})_s = \mathscr{L}_{dl}(\boldsymbol{M}\circ\boldsymbol{P}_s^{u'},\boldsymbol{M}\circ\boldsymbol{Y}_s^{u''}) \; + \\
&\qquad\qquad \mathscr{L}_{ce}(\boldsymbol{M}\circ\boldsymbol{P}_s^{u'},\boldsymbol{M}\circ\boldsymbol{Y}_s^{u''}) \; + \\
&\qquad\qquad \mathscr{L}_{mse}(\boldsymbol{M}\circ\boldsymbol{P}_s^{u'},\boldsymbol{M}\circ\boldsymbol{P}_s^{u''}),
\end{aligned}
\tag{4.4}
$$

where $\boldsymbol{X}^{u'}$ and $\boldsymbol{X}^{u''}$ denote the unlabeled data $\boldsymbol{X}^{u}$ with two different data augmentations, and $\boldsymbol{P}_s^{u'}$ and $\boldsymbol{P}_s^{u''}$ denote the corresponding predictions from our network and the teacher network, respectively. φ denotes a flipping operation to maintain the pixel-wise correspondence of predictions from the sample with different mirror augmentation. $\boldsymbol{Y}^{u''}$ denotes the pseudo label generated from $\boldsymbol{P}_s^{u''}$. We leverage a threshold technique with the threshold τ (set to 0.8) to select reliable pseudo labels, and $\boldsymbol{M}$ denotes the threshold mask. $\mathbb{1}$ denotes the indicator function and $\circ$ denotes the element-wise product. In addition to the Dice loss and the cross-entropy loss, we also exploit the mean squared error loss $\mathscr{L}_{mse}$ to facilitate network training by providing different constraints as well as gradients.

With $\mathscr{L}_{te}$ in Eq. (4.4), the training loss for all unlabeled data is defined by:

$$
\mathscr{L}_{u1} = \sum_{s}^{S} \beta_s \cdot (\mathscr{L}_{te}(\boldsymbol{X}^{u'},\boldsymbol{X}^{u''})_s + \mathscr{L}_{te}(\boldsymbol{X}^{u''},\boldsymbol{X}^{u'})_s),
\tag{4.5}
$$

where β_s is the loss weighting factor for the s-th stage in deep supervision.

In addition to $\mathscr{L}_{u1}$, we also use the pseudo labels generated from SAM to supervise network training. Since SAM can only process 2D images, we select 20 slices for each medical volume and exploit SAM to supervise the training of these slices. The objective

Table 4.1: Quantitative segmentation results on BRATS2020 in barely-supervised brain tumor segmentation. "Whole", "Core", and "Enhancing" denote three tumor regions, *i.e.*, the whole tumor, the tumor core, and the enhancing tumor, respectively. "Avg" denotes the average results of the three tumor regions.

Methods	Dice (%) ↑											
	2 Labels / 267 Unlabels				4 Labels / 265 Unlabels				6 Labels / 263 Unlabels			
	Whole	Core	Enhancing	Avg	Whole	Core	Enhancing	Avg	Whole	Core	Enhancing	Avg
Fully supervised	91.01	86.02	80.70	85.91	91.01	86.02	80.70	85.91	91.01	86.02	80.70	85.91
SASSNet [106]	56.71	19.94	32.19	36.28	77.29	53.98	58.28	63.18	78.80	57.15	61.13	65.69
UA-MT [213]	68.32	45.59	51.60	55.15	74.33	53.16	56.99	61.49	77.42	57.53	65.00	66.65
DTC [127]	72.81	24.60	52.09	49.83	79.62	54.32	60.23	64.72	82.21	58.23	62.27	67.57
URPC [128]	74.59	54.56	57.73	62.29	80.41	62.06	62.68	68.38	81.38	64.18	66.25	70.60
CPS[29]	79.83	56.60	65.00	67.14	85.67	68.04	69.77	74.49	87.31	71.22	71.12	76.55
CLD [116]	77.57	53.81	62.30	64.56	81.16	59.28	66.04	68.83	83.11	63.03	68.36	71.50
ComWin [198]	83.07	64.72	73.54	73.77	85.87	68.31	73.91	76.03	87.64	71.57	73.64	77.62
BarelySAM(Ours)	**88.48**	**77.58**	**74.90**	**80.32**	**89.70**	**78.38**	**76.37**	**81.48**	**89.92**	**81.80**	**77.85**	**83.19**

loss is defined by:

$$\begin{aligned}\mathscr{L}_{u2} = \ & \mathscr{L}_{dl}(P^{u'}, \widetilde{Y}^{u''}) + \mathscr{L}_{ce}(P^{u'}, \widetilde{Y}^{u''}) \\ & + \mathscr{L}_{dl}(P^{u''}, \widetilde{Y}^{u'}) + \mathscr{L}_{ce}(P^{u''}, \widetilde{Y}^{u'}),\end{aligned} \tag{4.6}$$

where P and $\widetilde{Y}$ denote the prediction of our network and the pseudo label of SAM. As SAM is pre-trained on natural images, adapting it to the appearance of various tumor areas is necessary. To this end, we also train SAM with the help of the teacher network. The loss is defined by:

$$\mathscr{L}_{SAM} = \mathscr{L}_{dl}(\widetilde{P}, Y) + \mathscr{L}_{ce}(\widetilde{P}, Y), \tag{4.7}$$

where $\widetilde{P}$ and Y denote the prediction of SAM and the pseudo label of the teacher network. Note that, $\mathscr{L}_{u2}$ and $\mathscr{L}_{SAM}$ also leverage the data augmentation and the threshold strategies introduced in Eq. (4.4).

The overall loss is written as:

$$\mathscr{L} = \mathscr{L}_l + \lambda(\mathscr{L}_{u1} + \mathscr{L}_{u2}) + \mathscr{L}_{SAM}, \tag{4.8}$$

where $\lambda = 0.1$ is a coefficient and aims to control the contributions of losses.

4.2.4 Implementation Details

Network Architecture. We employ nnUNet [79], the prevailing benchmark model in medical image segmentation, as the network architecture. The model is composed of a pixel encoder and a pixel decoder, interlinked through a series of skip connections.

Table 4.2: Quantitative segmentation results on BRATS2015 in barely-supervised brain tumor segmentation.

Methods	Dice (%) ↑											
	2 Labels / 252 Unlabels				4 Labels / 250 Unlabels				6 Labels / 248 Unlabels			
	Whole	Core	Enhancing	Avg	Whole	Core	Enhancing	Avg	Whole	Core	Enhancing	Avg
Fully supervised	91.15	83.28	79.13	84.52	91.15	83.28	79.13	84.52	91.15	83.28	79.13	84.52
SASSNet [106]	67.45	39.12	43.81	50.13	70.82	61.37	68.00	66.73	80.57	63.54	67.61	70.57
UA-MT [213]	68.14	42.67	45.44	52.08	72.18	54.48	58.62	61.76	76.83	67.32	65.39	69.85
DTC [127]	55.42	44.88	57.31	52.54	79.47	67.35	65.86	70.89	85.04	66.63	66.36	72.68
URPC [128]	76.36	64.62	62.57	67.85	78.69	64.38	61.76	68.28	81.79	67.32	66.02	71.71
CPS [29]	76.38	64.47	64.93	68.59	84.94	**69.67**	68.24	74.28	86.56	**71.58**	70.99	76.38
CLD [116]	79.34	65.29	63.08	69.23	82.28	68.77	66.90	72.65	83.40	69.55	67.90	73.62
ComWin [198]	76.99	63.88	**66.84**	69.24	84.96	67.97	**69.40**	74.11	86.50	70.62	69.36	75.49
BarelySAM(Ours)	**85.93**	**66.96**	66.42	**73.10**	**86.64**	68.97	69.19	**74.93**	**86.93**	70.50	**73.53**	**76.99**

Image Preprocessing. The pre-processing of our MRI images initiates with skull-stripping, co-registering, and re-sampling to a spatial resolution of $1mm^3$. Subsequently, non-brain regions (black background) are cut out, and the images within brain areas are normalized to a zero mean and unit variance.

Training. The training images initially undergo random scaling and rotation, followed by random cropping to create $128 \times 128 \times 128$ patches. Subsequently, we implement random mirror and intensity augmentations, *i.e.*, Gaussian blur, Gaussian noise, brightness modifications, contrast modifications, low-resolution simulation, and gamma correction. The network is optimized $25,000$ iterations using Stochastic Gradient Descent (SGD), with the momentum and weight decay set to 0.9 and $3e^{-5}$. The batch size is set to 2. We adjust the learning rate in a "poly" way, as $0.01 \times (1 - \frac{\text{iter}}{\text{max\_iter}})^{0.9}$. The selection probability P is defined as 0.2. For the loss weight β in the labeled loss equation, we use a sequence of $[1.0, 0.5, 0.25, 0.125, 0.0]$ combined with the L1 norm. The update momentum α for the TE network, as specified in the Exponential Moving Average (EMA) update equation, is dynamically set to $0.99 + 0.01(1 - \cos(\pi \cdot \frac{\text{iter}}{\text{max\_iter}} + 1)/2)$.

Testing. We utilize a sliding window technique for image prediction. Specifically, eight overlapping patches of dimensions $128 \times 128 \times 128$ are extracted, and the final predictions are achieved by amalgamating these patch-level predictions. At the testing phase, augmentation is applied through mirroring along all axes. Our methodology includes a post-processing step to minimize false alarms by suppressing minor components in the predictions. Specifically, brain tumors do not always contain enhancing areas. When the count of pixels predicted to be part of an enhancing tumor is very low (*i.e.*, fewer than 100), we treat these pixels as non-enhancing tumors, considering it as a false alarm.

Table 4.3: Quantitative segmentation results on BRATS2020 in barely-supervised incomplete brain tumor segmentation.

Methods	Dice (%) ↑											
	2 Labels / 267 Unlabels				4 Labels / 265 Unlabels				6 Labels / 263 Unlabels			
	Whole	Core	Enhancing	Avg	Whole	Core	Enhancing	Avg	Whole	Core	Enhancing	Avg
Fully supervised	86.05	76.77	60.74	74.52	86.05	76.77	60.74	74.52	86.05	76.77	60.74	74.52
SASSNet [106]	30.24	8.37	11.78	16.80	39.72	18.93	20.62	26.42	42.08	20.51	23.37	28.65
UA-MT [213]	35.80	21.63	18.91	25.45	39.59	18.34	20.67	26.20	40.50	21.76	23.47	28.58
DTC [127]	37.56	13.17	17.90	22.88	41.20	17.82	21.55	26.86	42.23	22.67	23.36	29.42
URPC [128]	48.61	26.56	22.31	32.49	52.88	36.16	27.76	38.93	53.33	35.04	29.33	39.23
CPS [29]	43.71	28.64	24.01	32.12	44.20	30.66	27.82	34.22	41.44	30.93	28.10	33.49
CLD [116]	38.65	21.49	21.94	27.36	38.38	23.40	23.41	28.40	39.37	22.51	23.75	28.54
ComWin [198]	42.26	27.12	26.00	31.79	41.04	26.64	26.92	31.53	42.30	27.35	27.47	32.37
RobustSeg [23]	47.62	26.63	26.83	33.69	59.28	35.09	29.46	41.28	61.97	39.89	33.72	45.19
RFNet [45]	61.69	36.72	31.99	43.47	70.24	46.94	37.87	51.68	72.25	49.77	41.61	54.54
BarelySAM(Ours)	**83.65**	**64.10**	**53.18**	**66.98**	**83.78**	**66.09**	**51.98**	**67.28**	**84.39**	**68.69**	**54.11**	**69.06**

Table 4.4: Quantitative segmentation results on BRATS2015 in barely-supervised incomplete brain tumor segmentation.

Methods	Dice (%) ↑											
	2 Labels / 252 Unlabels				4 Labels / 250 Unlabels				6 Labels / 248 Unlabels			
	Whole	Core	Enhancing	Avg	Whole	Core	Enhancing	Avg	Whole	Core	Enhancing	Avg
Fully supervised	84.47	77.93	55.96	72.79	84.47	77.93	55.96	72.79	84.47	77.93	55.96	72.79
SASSNet [106]	36.26	16.18	14.41	22.28	38.56	25.76	26.40	30.24	49.17	29.75	26.98	35.30
UA-MT [213]	37.02	19.59	17.06	24.56	41.24	26.20	20.88	29.44	47.66	33.26	27.83	36.25
DTC [127]	39.98	17.01	19.75	25.58	44.43	29.38	25.06	32.97	45.90	28.03	26.27	33.40
URPC [128]	50.63	35.01	27.37	37.67	50.30	37.57	30.62	39.50	52.77	40.84	31.53	41.71
CPS [29]	42.29	38.39	24.16	34.95	42.30	38.71	26.93	35.98	39.01	45.03	30.74	38.26
CLD [116]	36.68	22.66	20.38	27.24	40.95	25.61	22.45	29.67	36.13	25.20	22.20	27.85
ComWin [198]	37.17	27.06	24.48	29.57	41.32	28.70	26.19	32.07	42.14	30.70	26.67	33.17
RobustSeg [23]	46.86	30.92	27.22	35.00	53.65	34.79	24.73	37.72	53.14	36.57	30.72	40.14
RFNet [45]	61.64	38.55	44.07	48.09	67.56	43.52	34.52	48.53	71.07	47.75	40.57	53.13
BarelySAM(Ours)	**78.97**	**55.62**	**46.31**	**60.30**	**79.28**	**55.66**	**46.52**	**60.49**	**79.92**	**58.93**	**48.39**	**62.41**

4.3 Experiments

Dataset. We evaluate our method on two brain tumor segmentation benchmarks [136], *i.e.*, BRATS2020 and BRATS2015. BRATS2020[1]contains 369 training samples which are randomly split into 269/100 samples for training/testing. BRATS2015 contains 274 training samples, among which 254/20 samples are adopted for training/testing respectively.

Evaluation Metric. Dice coefficient [39], Sensitivity, Specificity, and Hausdorff Distance

[1]Since the annotations of the validation and test sets are held by the challenge organizer, following [23, 45], we leverage the training subjects only.

Table 4.5: Ablation study. SAM denotes the SAM-based pseudo label generation.

Methods		Dice (%) ↑											
		2 Labels / 267 Unlabels				4 Labels / 265 Unlabels				6 Labels / 263 Unlabels			
MDM	SAM	Whole	Core	Enhancing	Avg	Whole	Core	Enhancing	Avg	Whole	Core	Enhancing	Avg
Label only		74.69	53.52	60.20	62.80	78.44	59.80	63.98	67.41	80.94	65.26	69.80	72.00
		78.61	58.49	67.07	68.06	86.07	68.37	72.25	75.56	86.87	69.21	72.24	76.11
✓		88.02	72.54	71.69	77.42	88.99	74.68	73.82	79.16	89.37	80.59	75.45	81.80
	✓	87.56	74.37	74.64	78.86	88.82	75.77	**76.88**	80.49	89.45	80.79	77.08	82.44
✓	✓	**88.48**	**77.58**	**74.90**	**80.32**	**89.70**	**78.38**	76.37	**81.48**	**89.92**	**81.80**	**77.85**	**83.19**

Table 4.6: Quantitative segmentation results under three other testing criteria, *i.e.*, Sensitivity, Specificity and Hausdorff Distance 95% (HD95).

Methods	Sensitivity (%) ↑				Specificity (%) ↑				HD95 (mm) ↓			
	Whole	Core	Enhancing	Avg	Whole	Core	Enhancing	Avg	Whole	Core	Enhancing	Avg
Fully supervised	90.22	85.31	85.10	86.88	99.82	99.90	99.93	99.88	4.42	4.52	2.61	3.85
SASSNet [106]	55.37	24.25	50.31	43.31	99.36	99.50	99.36	99.41	38.39	42.02	51.16	43.86
UA-MT [213]	82.35	74.92	74.52	77.26	98.65	98.45	99.57	98.89	47.51	54.33	45.15	49.00
DTC [127]	66.56	24.22	55.48	48.75	99.77	**99.94**	**99.94**	**99.88**	17.94	28.28	20.55	22.26
URPC [128]	79.48	62.95	65.68	69.37	99.45	99.59	99.88	99.64	33.02	37.69	25.00	31.90
CPS [29]	76.41	52.42	67.92	65.58	99.50	99.85	99.91	99.75	32.36	30.81	22.98	28.72
CLD [116]	77.86	60.60	62.66	67.37	99.68	99.61	99.94	99.74	23.41	26.22	15.67	21.76
ComWin [198]	79.08	74.85	78.22	77.39	**99.86**	99.62	99.93	99.80	11.15	15.69	7.97	11.60
BarelySAM(Ours)	**91.60**	**77.24**	**84.35**	**84.40**	99.66	99.89	99.90	99.82	**8.36**	**9.12**	**6.75**	**8.08**

(HD) are used to measure the performance of segmentation predictions, defined as:

$$
\begin{aligned}
&\text{Dice} = 2 \cdot \frac{|\boldsymbol{P} \bigcap \boldsymbol{Y}|}{|\boldsymbol{P}| + |\boldsymbol{Y}|}, \\
&\text{Sensitivity} = \frac{|\boldsymbol{P} \bigcap \boldsymbol{Y}|}{|\boldsymbol{Y}|}, \\
&\text{Specificity} = \frac{|(1-\boldsymbol{P}) \bigcap (1-\boldsymbol{Y})|}{|1-\boldsymbol{Y}|}, \\
&\text{HD} = \max\{\sup_{p \in \partial \boldsymbol{P}} \inf_{y \in \partial \boldsymbol{Y}} \|p-y\|_2, \sup_{y \in \partial \boldsymbol{Y}} \inf_{p \in \partial \boldsymbol{P}} \|p-y\|_2\},
\end{aligned}
\tag{4.9}
$$

where $\boldsymbol{P}$ and $\boldsymbol{Y}$ denote predictions and the corresponding ground truth, respectively.

Barely-supervised Setting. Our experiments are conducted under three varying situations, with 2, 4, and 6 labeled samples. In each situation, we use three distinct training and testing splits to ensure robustness. We choose to annotate an even number of images because brain tumors usually have two stages, *i.e.*, HGG and LGG, which have different characteristics and thus need to be considered simultaneously.

4.3.1 Comparisons with the State-of-the-art

Barely-supervised Brain Tumor Segmentation. Table 4.1 and Table 4.2 compare our method with five semi-supervised state-of-the-art methods, *i.e.*, SASSNet [106], UA-MT [213], DTC [127], URPC [128], and CPS[29], and two barely-supervised state-of-the-art methods, *i.e.*, CLD [116] and ComWin [198]. As shown in Table 4.1, our method yields remarkable performance on barely-supervised segmentation on BRATS2020, *e.g.*, surpassing the previous SOTA, *i.e.*, ComWin [198], by 6.55%/5.45%/5.57% in terms of average Dice score under all three barely-supervised scenarios. Table 4.2 further confirms the exceptional performance of our method on BRATS2015. This demonstrates the effectiveness of our method in exploiting multiple modalities and unlabeled data. Besides, we also observe that, as the number of labeled samples increases, the segmentation accuracy for three tumor regions improves, though the "Whole" tumor region sees a lesser improvement. The lesser improvement might be because "Whole" tumors are simpler to segment, requiring only a minimal amount of labeled images for high accuracy. Moreover, as shown in Table 4.1, our method using as few as 6 labeled images nearly matches the performance of fully-supervised segmentation, showing a minimal Dice score reduction of 1.09%/4.22%/2.85% across three tumor areas.

Barely-supervised Incomplete Brain Tumor Segmentation. To validate that our method can also deal with missing modality situations in the barely-supervised regime, we also compare our method with the state-of-the-art methods in regard to Barely-supervised Incomplete Brain Tumor Segmentation (BIBTS), as illustrated in Table 4.3 and Table 4.4. In BIBTS, a network needs to consider all fifteen modality combinations from four-modality subjects according to limited labeled training samples, and the averaged accuracy is leveraged to compare. In this fashion, BIBTS methods are able to handle both the missing modality problem and the insufficient annotation problem and thus are more easily deployed in clinical practice. In Table 4.3 and Table 4.4, two incomplete brain tumor segmentation methods, *i.e.*, RobustSeg [23] and RFNet [45], are used for better comparison. As seen in the tables, our method significantly outperforms current state-of-the-art techniques. For instance, on BRATS2020, our method exceeds RFNet by 23.57%/15.60%/14.52% in all three barely-supervised scenarios, showcasing its effectiveness in handling incomplete brain tumor segmentation under limited supervision.

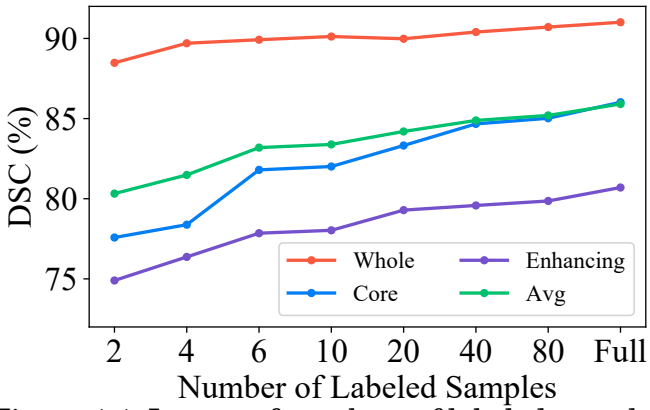

Figure 4.4: Impact of numbers of labeled samples.

Table 4.7: Analysis of the impact of SAM prompts.

Method	Dice (%) ↑			
	Whole	Core	Enhancing	Avg
w/o SAM	88.02	72.54	71.69	77.42
w/o point prompt	87.95	74.98	73.45	78.79
w/o box prompt	88.21	75.60	74.10	79.30
w/o contrastive point selection	88.35	76.20	74.13	79.56
Ours	**88.48**	**77.58**	**74.90**	**80.32**

4.3.2 Diagnostic Experiments

To rigorously test our key components, we perform a series of ablative studies on BRATS2020. The reported results are built under "2 Labels/267 Unlabels" situation unless otherwise specified.

Effectiveness of SAM and MDM. As shown in Table 4.5, employing SAM and MDM individually enhances segmentation accuracy in all barely-supervised scenarios. For example, incorporating SAM into the baseline improves the average Dice score by 10.80%, 4.93%, and 6.33% across three scenarios. Additionally, we observe that combining SAM and MDM further increases accuracy. This demonstrates the effectiveness of SAM and MDM in boosting tumor segmentation under limited annotation situations.

Impact of Number of Labeled Samples. In Fig. 4.4, we evaluate our method under the scenarios with 2, 4, 6, 10, 20, 40, 80, and, 269 (full-supervised) labeled samples. As seen, the performance exhibits large improvement as the labeled number increases from 2 to 6, but the gain becomes negligible when exceeding 6. We also observe that our method with 6 labeled images can already achieve performance nearly on par with that of the fully-supervised situation.

Impact of SAM Prompts. SAM in this work generates predictions according to the

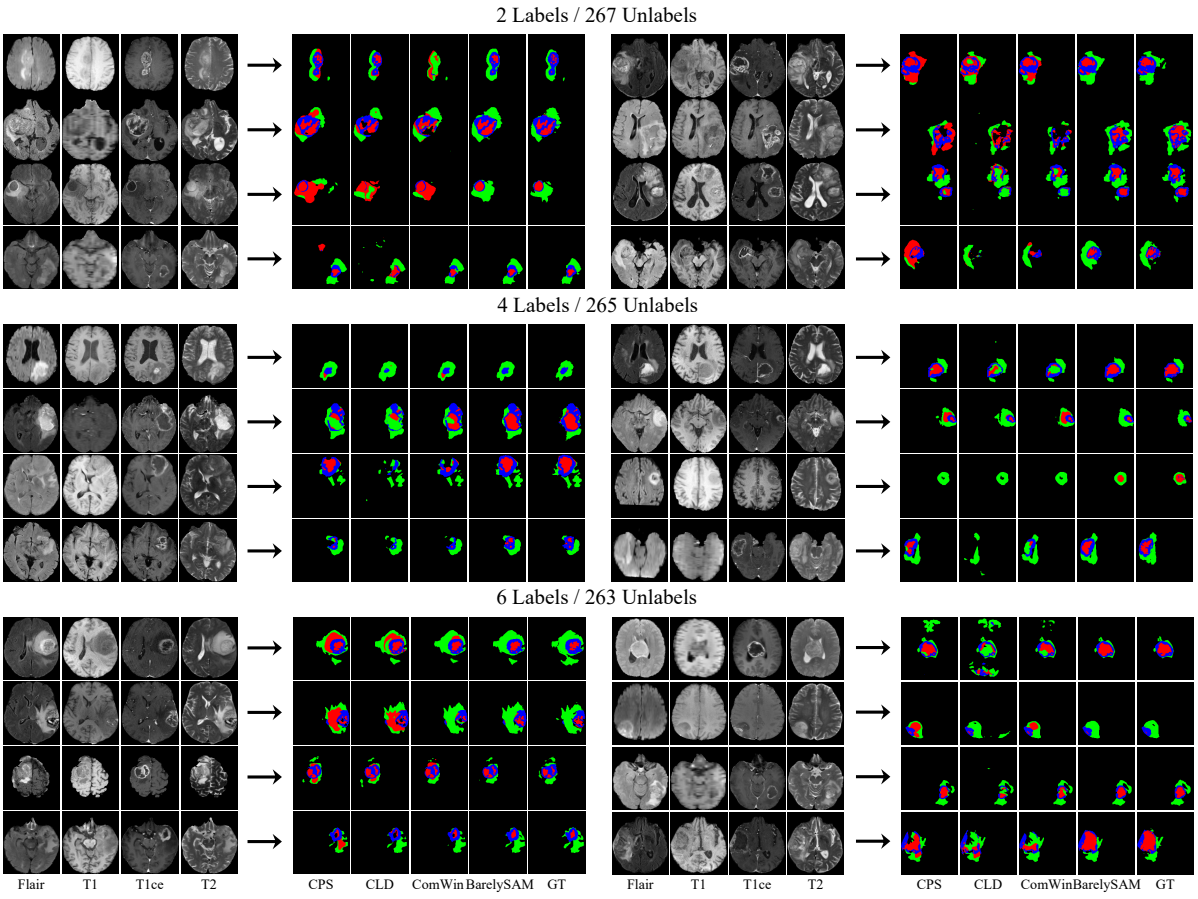

Figure 4.5: Qualitative comparisons on BRATS2020 with 2, 4 and 6 labeled samples. Multi-modal samples and the corresponding segmentation results from state-of-the-art methods are shown.

input prompts, *i.e.*, point prompt and box prompt. Table 4.7 analyzes the impact of these two prompts by applying them separately. As seen, applying point or box prompts solely leads to performance improvement. This demonstrates that incorporating SAM could provide additional knowledge for network training. An additional observation is using point prompts can bring more performance improvement than box prompts. This may be because point prompts would provide more detailed information and help SAM generate more accurate pseudo labels. Besides, Table 4.7 validates the effectiveness of the contrastive point selection, as the method without it achieves worse segmentation accuracy. Finally, Table 4.7 illustrates that combining box and point prompts will lead to further improvement as these two prompts would bring more information.

Impact of the Number of Points in SAM. For point prompts in SAM, we randomly select 5 points for each class. For each class, we assign the corresponding 5 points as

Table 4.8: Analysis of the impact of number of points in SAM.

# Points	Dice (%) ↑			
	Whole	Core	Enhancing	Avg
0	88.21	75.60	74.10	79.30
1	88.42	75.52	75.64	79.86
3	88.10	76.82	74.75	79.89
5	88.48	77.58	74.90	80.32
8	**88.67**	76.78	**75.68**	**80.38**
10	88.59	**77.61**	74.69	80.30

Table 4.9: Analysis of the probability of partial modal combinations p_s.

p_s	Dice (%) ↑			
	Whole	Core	Enhancing	Avg
0.0	87.56	74.37	74.64	78.86
0.2	88.48	77.58	**74.90**	**80.32**
0.4	**88.95**	**78.21**	73.33	80.16
0.6	88.28	74.76	72.73	78.59
0.8	88.57	74.24	72.49	78.43
1.0	81.08	47.51	47.90	58.83

positive points and points of other classes as negative points. We have conducted an ablation study on the number of selected points in Table 4.8. As seen, the performance exhibits large improvement as the point number increases from 0 to 5, but the gain becomes negligible when exceeding 5.

Impact of Partial Modality Combination Probability p_s. We investigate the impact of p_s (described in Sec. 4.2.2) in Table 4.9 by reporting results with various p_s (from 0.0 to 1.0). Table 4.9 shows that our model attains comparable segmentation accuracy across different values of P, namely 0.2 and 0.4. This observation highlights the robustness of MDM to variations in p_s. Consequently, we set p_s to 0.2 in our experimental setup. Furthermore, it is worth noting that our model exhibits inferior performance when P is set to 0.0 (that is, without MDM), indicating the effectiveness of the MDM module. Additionally, in Table 4.9, we observe a substantial decrease in accuracy when p_s is set to 1.0. The decline can be attributed to the potential limitation of leveraging partial modality combinations throughout the entire training process, hindering the network's effective utilization of information from the complete set of modalities.

Comparisons under Other Criteria. In Table 4.6, we additionally use three testing criteria, *i.e.*, Sensitivity, Specificity, and Hausdorff Distance 95% (HD95), to strengthen the comparison. As shown in Table 4.6, our method achieves better accuracy in Sensitivity (77.39% → 84.40% in $\text{Sensitivity}_{\text{avg}}$) and HD95 (11.60% → 8.08% in HD95_{avg}). This

demonstrates our method is better at locating tumor regions. While for Specificity, several previous methods tend to predict foreground as background and thus achieve better Specificity (even better than fully supervised training), their Sensitivity is lower.
Qualitative Comparisons. Figure 4.5 illustrates segmentation predictions from our method and the state-of-the-art methods on BRATS2020 under all three barely-supervised settings, *i.e.*, 2 Labels, 4 Labels, and 6 Labels. As seen in the figure, our method predicts more accurate segmentation maps.

4.4 Conclusion

This chapter introduces the proposed training framework, named BarelySAM, for barely-supervised brain tumor segmentation. BarelySAM first incorporates Segment Anything Model (SAM) into training to compensate for the lack of labeled images. Specifically, BarelySAM exploits the pre-trained knowledge from SAM via the pseudo-labeling technique, thus assisting network training and improving segmentation performance. In addition, BarelySAM devises Multi-modality Dependency Minimization (MDM), which considers partial modality combinations from multi-modal samples. In this fashion, MDM prevents networks from overly relying on specific modalities and encourages the effective exploitation of each modality. Furthermore, BarelySAM introduces consistency supervision based on MDM to fully use unlabeled data, thereby further enhancing network performance. Extensive experiments reveal the superiority of BarelySAM.

This chapter (§4) introduces BarelySAM and provides another way for the insufficient annotation challenge. Therefore, this chapter supplements §3. To be specific, BarelySAM addresses lesion segmentation, such as brain tumor segmentation, thus compensates for the inability of §3 to process lesion areas. Furthermore, §3 and §4 facilitate the full exploitation of unlabeled data and achieve improved segmentation performance, thus helping the deployment of the medical image segmentation system (§6).

CHAPTER 5

INCOMPLETE MULTI-MODAL SEGMENTATION WITH REGION-AWARE FUSION NETWORK

5.1 Introduction

Brain tumor segmentation, aiming to segment different brain tumor regions, is vital for clinical assessment and surgical planning. In order to improve the segmentation accuracy, most existing methods [24, 53, 79, 84, 171, 222, 233] use four modalities simultaneously, namely T1-weighted (T1), Fluid Attenuation Inversion Recovery (Flair), T2-weighted (T2), and contrast-enhanced T1-weighted (T1c). However, the missing modality problem often happens in clinical practice because of various patient conditions and scanning protocols. Therefore, these standard methods fail to perform well in practice.

Incomplete multi-modal brain tumor segmentation approaches [23, 48, 66, 237] have been proposed to deal with various missing situations. Havaei *et al.*[66] and Dorent *et al.*[48] compute the mean and variance across accessible multi-modal features as fused features. However, this fusion treats each modality equally regardless of different missing scenarios and thus may fail to aggregate features effectively. Later, Chen *et al.*[23] and Zhou *et al.*[237] leverage attention mechanisms to emphasize contributions from different accessible modalities. However, they do not fully exploit the relations between tumor regions and image modalities. In particular, different modalities exhibit

This chapter is based on joint work [45] with Xin Yu and Yi Yang, presented primarily as it appears in the ICCV 2021 proceedings.

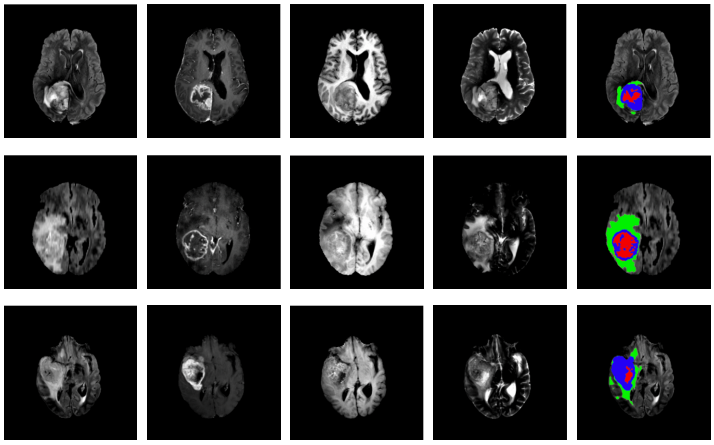

Figure 5.1: Illustration of different sensitivities of modalities to different brain tumor regions. From left to right: Images of four modalities, *i.e.*, Flair, T1c, T1 and T2, and the corresponding labels of three patients are shown. In the segmentation results, different colors denote different brain tumor regions.

unique appearances and, as a result, have varying sensitivities to different tumor regions. For example, as visible in Fig. 5.1, T1c is more sensitive to the red and blue tumor areas, while Flair and T2 provide more information for the green tumor area. This observation suggests that we should focus differently on various modalities and regions to achieve accurate brain tumor segmentation,

Taking the relations between modalities and regions into account, we propose a Region-aware Fusion Network (RFNet) to aggregate various accessible multi-modal features from different regions adaptively. Our RFNet is constructed by an encoder-decoder architecture, where four encoders are employed to extract features from different modal images. In order to establish the relations between image modalities and tumor regions, we introduce a Region-aware Fusion Module (RFM) into our RFNet. RFM first divides modal features into different regions (*i.e.*, tumor sub-structure) via a learned probability map. The probability map indicates the probabilities of tumor regions at each pixel. Then, RFM generates corresponding attention weights in each region to adaptively control the contributions of different image modalities.

Since brain tumors usually occupy a small part of the brain, we introduce a region-norm pooling operation to obtain a normalized global feature from each region. Thereby, we prevent the global feature from being numerically too small. Then, we use two fully connected layers followed by a sigmoid activation to obtain attention weights from the global feature for image modalities and tumor regions. In this fashion, RFM will generate

larger weights for the modalities that are more sensitive to certain tumor regions, thus leading to discriminative fused features for accurate segmentation.

Due to the missing hetero-modal data, RFNet will face the problem of unbalanced training. To be specific, RFNet might try to seek the easiest way to segment brain tumors from the multi-modal data. In other words, the network segments each region mainly by exploiting the modalities that are sensitive to the region rather than all the modality information. However, this will lead to poor segmentation accuracy when some modalities are missing. To tackle this problem, we develop a segmentation-based regularizer. In particular, a weight-shared decoder is employed to segment each modality individually. This approach compels each modal encoder to learn distinctive features for all tumor regions. Therefore, RFNet can segment different regions well even when some modalities are missing. Benefiting from the proposed fusion module and regularizer, RFNet achieves superior segmentation performance on BRATS2020, BRATS2018 and BRATS2015. This demonstrates the superiority of our method.

Overall, our contributions are threefold:

- We propose a novel Region-aware Fusion Network (RFNet) to solve the missing modality problem. In particular, we devise a novel Region-aware Fusion Module (RFM) by explicitly taking the relations between regions and modalities into account. With the help of RFM, RFNet effectively aggregates diverse combinations of modal features and produces discriminative fused features for segmentation.
- To address the unbalanced training problem of RFNet, we propose a segmentation-based regularizer. The proposed regularizer enforces each modal encoder to produce discriminative features for segmenting all the tumor regions, thus further improving the discriminativeness of the fused features.
- Taking advantage of the proposed fusion module and regularizer, RFNet achieves superior segmentation accuracy on the widely used benchmarks.

5.2 Proposed Method

In this work, we design RFNet for incomplete multi-modal brain tumor segmentation. In particular, we develop an RFM module to take advantage of available modalities effectively during feature fusion. In addition, we propose a segmentation-based regularizer to improve the feature representations of each modal encoder further, thus facilitating the final segmentation performance. In this section, we will introduce our designed RFNet as well as the proposed regularization term in detail.

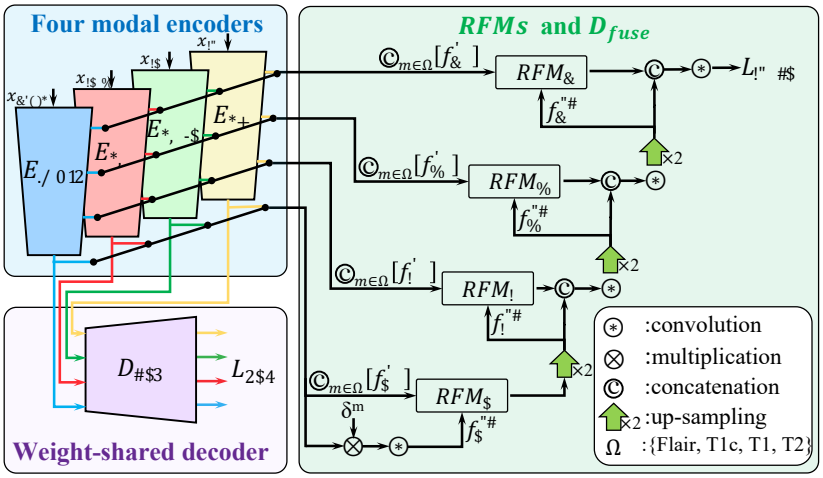

Figure 5.2: Illustration of our proposed RFNet. Four encoders, *i.e.*, $\boldsymbol{E}_{\text{Flair}}$, $\boldsymbol{E}_{\text{T1c}}$, $\boldsymbol{E}_{\text{T1}}$ and $\boldsymbol{E}_{\text{T2}}$, are employed to extract features from four modalities individually. $\boldsymbol{D}_{\text{sep}}$ is our segmentation-based regularizer network, while $\boldsymbol{D}_{\text{fuse}}$ with the designed RFM is used to attain the final segmentation predictions. δ^m simulates different missing scenarios.

5.2.1 Task Definition

Incomplete multi-modal brain tumor segmentation focuses on three brain tumor areas segmentation, *i.e.*, the whole, core and enhancing tumor, from various combinations of multi-modal MRI images, including Flair, T1c, T1 and T2. The whole tumor is composed of all three tumor sub-regions, i.e., the necrotic and non-enhancing tumor core (NCR/NET), the peritumoral edema (ED), and the GD-enhancing tumor (ET). The tumor core consists of NCR/NET and ET, while the enhancing tumor involves ET. Figure 5.1 illustrates NCR/NET, ED and ET in red, green and blue, respectively.

To assess our method's resilience in different situations of missing data, we evaluate its segmentation results on all combinations of modalities, and the average score is reported for comparison. During training, all modalities and labels are available, and we simulate missing scenarios by setting missing modal features to zero.

5.2.2 Architecture Overview

We adopt a 3D U-Net [33] architecture with a late fusion strategy to construct our RFNet. As shown in Fig. 5.2, four encoders, *i.e.*, $\{\boldsymbol{E}_m\}_{m\in\{\text{Flair,T1c,T1,T2}\}}$, are employed to extract features from four modalities separately. The decoder $\boldsymbol{D}_{\text{sep}}$ is designed to segment each modality separately, thus assisting our four encoders in learning representative region features. Furthermore, $\boldsymbol{D}_{\text{sep}}$ shares weights for the four image modalities, so that four modal features can be projected into a shared latent space. This also significantly

facilitates the later feature aggregation and fusion.

The decoder $\boldsymbol{D}_{\text{fuse}}$ is designed to obtain the final segmentation results from the aggregated features, as visible in Fig. 5.2. In each stage, the encoder features are fused by the designed RFM. Note that, RFM takes not only four encoder features but also the features from the prior layer as input. This is because that the previous layer features can be used to embed semantic information of tumor regions, thus making RFM region-aware. In the bottleneck (*i.e.*, the fourth stage S_4), there are no previous layer features available for RFM. Therefore, we leverage an additional convolutional layer to embed the encoder features into semantic features for the fusion module in Fig. 5.2.

5.2.3 Region-aware Fusion Module

Considering different sensitivities of image modalities to different regions, as shown in Fig. 5.1, our RFNet aims to pay different attention to different modalities in each region. In this fashion, discriminative features for tumor regions can be obtained, leading to the improvement of segmentation accuracy. To this end, we develop an RFM module that is designed to fuse available modal features in a region-aware fashion, as visible in Fig. 5.3. RFM mainly consists of two parts: probability map learning and region-aware multi-modal feature fusion.

Probability Map Learning: To achieve the region-aware characteristics, our RFM first learns a probability map that indicates the probabilities of brain tumor structure (including healthy brain regions) at each location. As shown in Fig. 5.3, the probability map is obtained from the decoder feature of the previous layer f^{de} and the available encoder features $\copyright_{m\in\Omega}[f^m\cdot\delta^m]$. Employing the encoder features in RFM is because they offer more detailed spatial information and can improve the accuracy of the probability maps. © denotes the concatenation operation while Ω denotes the modality set, including Flair, T1c, T1 and T2. δ^m is set to either 0 or 1, indicating whether the m modality is missing or not. The probability map learning procedure is defined as:

$$\hat{y}_{i,j}^{pm} = \frac{\exp(\phi_j(f_{i,j}^{pm};\theta_j))}{\sum_{k\in K}\exp(\phi_j(f_{i,j}^{pm};\theta_j)_k)}, \tag{5.1}$$

where $f_{i,j}^{pm}$ represents the features from $f_{i,j}^{de}$ and $\copyright_{m\in\Omega}\left[f_{i,j}^{m}\cdot\delta^m\right]$. i and j denote the i-th subject and the j-th stage/level of the network, respectively. $\hat{y}_{i,j}^{pm}$ is the learned probability map. ϕ_j denotes the region classifier in the j-th stage and θ_j is the corresponding parameters. K denotes four brain tumor regions which need to be segmented, including BG (background), NCR/NET, ED and ET.

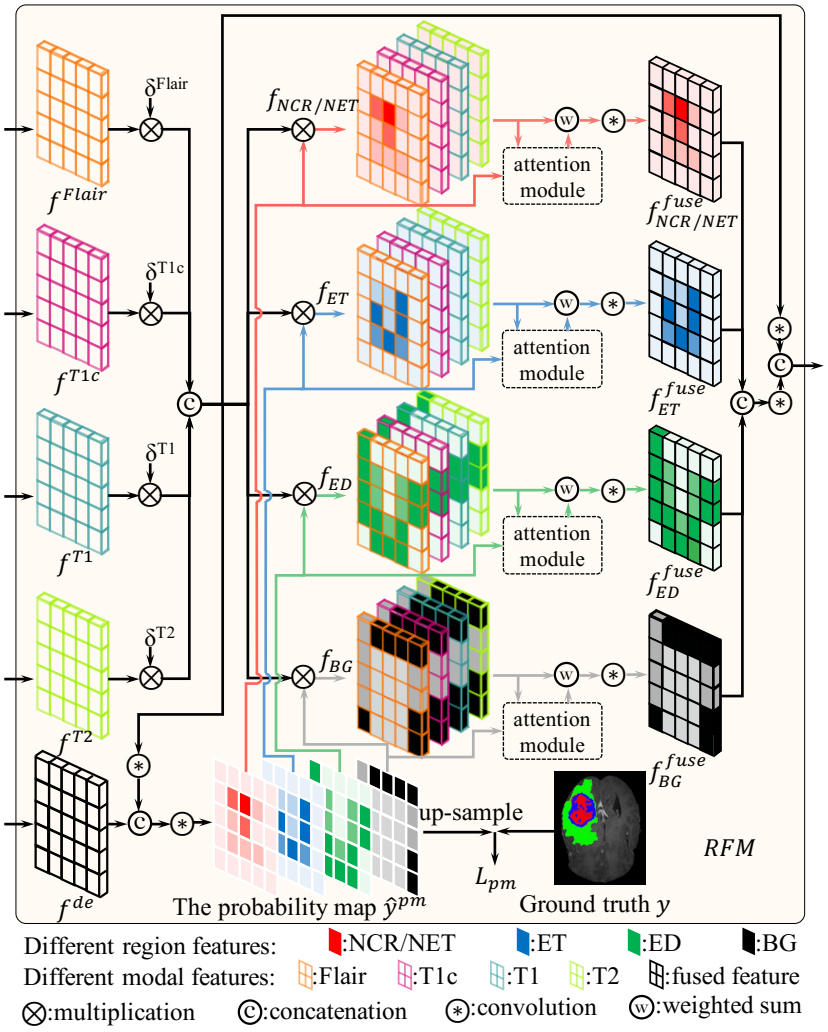

Figure 5.3: Illustration of our region-aware fusion module (RFM). The probability map is first learned to divide multi-modal features into different regions. Then, an attention mechanism is designed to aggregate features in a region-aware manner.

The probability map (shown in Fig. 5.4) is learned under the supervision of the ground truth by a weighted cross-entropy loss $\mathscr{L}_{WCE}$ [23] and a Dice loss $\mathscr{L}_{DL}$, expressed as:

$$\mathscr{L}_{pm} = \sum_{i=1}^{N} \sum_{j=1}^{S_{num}} \left(\mathscr{L}_{WCE}(\psi_j(\hat{y}^{pm}_{i,j}), y_i) + \mathscr{L}_{DL}(\psi_j(\hat{y}^{pm}_{i,j}), y_i) \right), \tag{5.2}$$

where N and S_{num} denote the number of training data and stages. ψ_j denotes the up-sampling operation in the j-th stage, aiming to match the resolution of the probability map $\hat{y}^{pm}_{i,j}$ and the ground-truth mask y_i. $\mathscr{L}_{WCE}$ is formulated as:

$$\mathscr{L}_{WCE}(\hat{y}, y) = \sum_{k \in K} \frac{\| -\alpha_k \cdot y_k \cdot \log(\hat{y}_k) \|_1}{H \cdot W \cdot Z}, \tag{5.3}$$

where $\|\cdot\|_1$ denotes the L1 norm, and W, H and Z denote the length, width, and height of the 3D volumes, respectively. α_k is the weight for the region k and $\alpha_k = 1 - \frac{\|y_k\|_1}{\sum_{k' \in K} \|y_{k'}\|_1}$. $\mathscr{L}_{DL}$ is formulated as:

$$\mathscr{L}_{DL}(\hat{y}, y) = 1 - \sum_{k \in K} \frac{2 \cdot \|\hat{y}_k \bigcap y_k\|_1}{K_{num} \cdot (\|\hat{y}_k\|_1 + \|y_k\|_1)}, \tag{5.4}$$

where $\bigcap$ denotes the overlap between predictions and ground-truth masks, and K_{num} denotes the number of regions in K.

Region-aware Multi-modal Fusion: With the help of the probability map, RFM has managed to divide multi-modal features into different regions. Thus, the region-aware fusion is conducted on the divided features in each region.

The feature division is implemented by multiplying features with the probability map, written as:

$$f_k = ©_{m \in \Omega} \left[f^m \cdot \delta^m \right] \cdot \hat{y}_k^{pm}, \tag{5.5}$$

where f_k[1] denotes the divided features of the available modalities in the tumor region k and f^m denotes the encoder feature of the modality m.

As shown in Fig. 5.3, after feature division, modal-wise attention weights are learned individually in different regions to aggregate the corresponding features. Figure 5.5 illustrates the generation procedure of the attention weight in the region k. Specifically, the global feature of the region k is obtained via an average pooling operation and is then normalized by the probability map $\hat{y}_k^{pm}$. Employing this region-norm pooling can prevent the averaged global feature from being numerically too small, given the fact that brain tumors usually occupy only a small area in a brain. Then, two fully connected layers, along with a Leaky ReLU layer and a sigmoid activation, are adopted to embed the normalized feature into a modal-wise attention weight. As shown in Fig 5.5, the generated attention weights are then applied to the divided feature f_k to adjust the contributions from available modalities to obtain discriminative fused features.

Considering the distinct sensitivities of different modalities in various regions, RFM employs separate attention modules for each region to generate corresponding attention weights, as shown in Fig. 5.3. By paying more attention to more sensitive modalities, RFM is able to generate more representative features for each region. To feed these region features to the decoder, in Fig. 5.3, RFM adopts a concatenation operation followed by a convolutional bottleneck. A shortcut connection is also employed, similar to the residual learning [67].

[1]For simplicity, we omit the subscripts i and j without causing any confusion.

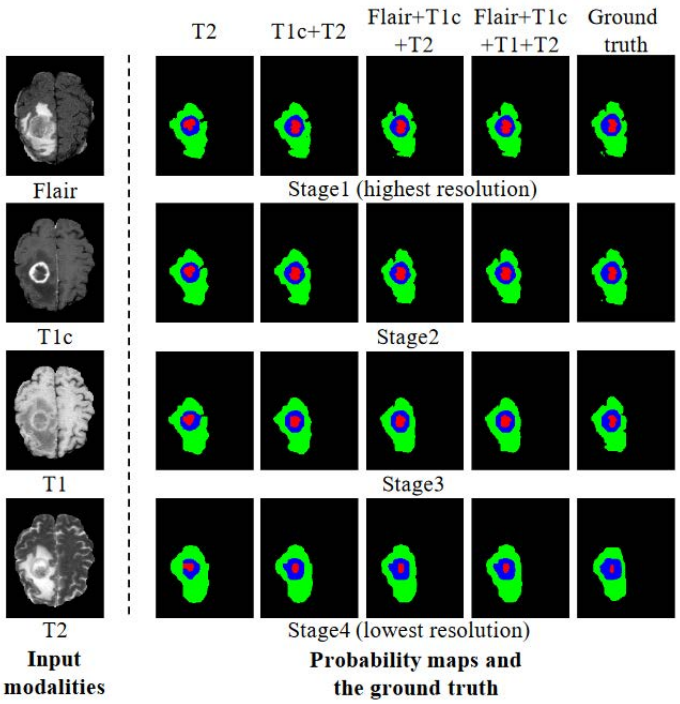

Figure 5.4: Visualization of the probability maps in four stages. Left: four image modalities. Right: Estimated probability maps from different combinations of image modalities in different stages/levels of our network and the corresponding ground truth.

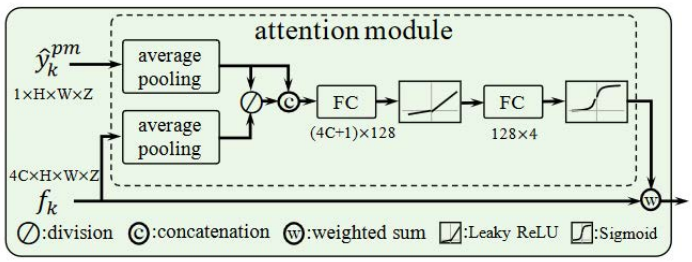

Figure 5.5: Illustration of the attention module. The region-norm pooling normalizes the global feature of f_k by the average probability of $\hat{y}_k$ to obtain the features to generate the attention weights.

5.2.4 Segmentation-based Regularizer

The phenomenon of missing multi-modal data usually introduces unbalanced training issues [191]. To be specific, deep neural networks usually opt to segment tumor regions mainly based on discriminative modalities. Therefore, some modal encoders are well-trained to be able to identify the corresponding tumor regions, while other encoders are not. This would lead to severe accuracy degradation in tumor segmentation when the discriminative modalities are missing.

To solve this problem, we propose a segmentation-based regularizer. As illustrated in Fig. 5.2, RFNet adopts a weight-shared decoder $\boldsymbol{D}_{\text{sep}}$ to segment each modal image

Table 5.1: Quantitative segmentation results on BRATS2020. "Complete": the whole tumor, "Core": the tumor core , and "Enhancing": the enhancing tumor. All the results are reproduced by using the authors' codes.

Modalities				Dice scores (%)											
				Complete				Core				Enhancing			
F	T1	T1c	T2	[66]	[48]	[23]	Ours	[66]	[48]	[23]	Ours	[66]	[48]	[23]	Ours
○	○	○	●	79.85	80.75	82.20	**86.05**	54.22	57.43	61.88	**71.02**	31.43	28.70	36.46	**46.29**
○	○	●	○	64.58	68.54	71.39	**76.77**	69.41	73.01	76.68	**81.51**	63.24	66.59	67.91	**74.85**
○	●	○	○	63.01	54.93	71.41	**77.16**	42.42	36.73	54.30	**66.02**	16.53	12.33	28.99	**37.30**
●	○	○	○	52.29	82.69	82.87	**87.32**	24.97	51.15	60.72	**69.19**	9.00	20.87	34.68	**38.15**
○	○	●	●	84.45	83.37	85.97	**87.74**	77.60	77.85	82.44	**83.45**	70.30	68.74	71.42	**75.93**
○	●	●	○	72.50	71.58	76.84	**81.12**	75.59	76.49	80.28	**83.40**	70.71	67.82	70.11	**78.01**
●	●	○	○	65.29	85.01	88.10	**89.73**	41.58	55.10	68.18	**73.07**	13.99	22.53	39.67	**40.98**
○	●	○	●	82.31	81.58	85.53	**87.73**	56.38	59.29	66.46	**73.13**	28.58	28.73	39.92	**45.65**
●	○	○	●	81.56	87.40	88.09	**89.87**	55.89	61.87	68.20	**74.14**	28.91	30.48	42.19	**49.32**
●	○	●	○	69.37	86.13	87.33	**89.89**	70.86	76.86	81.85	**84.65**	68.31	69.53	70.78	**76.67**
●	●	●	○	73.31	87.10	88.87	**90.69**	75.07	79.51	82.76	**85.07**	70.80	71.32	71.77	**76.81**
●	●	○	●	83.03	88.07	89.24	**90.60**	57.40	63.46	70.46	**75.19**	29.53	30.60	43.90	**49.92**
●	○	●	●	84.64	88.33	88.68	**90.68**	77.69	78.68	81.89	**84.97**	71.36	69.84	71.17	**77.12**
○	●	●	●	85.19	84.27	86.63	**88.25**	79.05	79.99	82.85	**83.47**	71.67	69.74	71.87	**76.99**
●	●	●	●	85.19	88.81	89.47	**91.11**	78.58	80.40	82.87	**85.21**	71.49	70.50	71.52	**78.00**
Average				75.10	81.24	84.17	**86.98**	65.45	67.19	73.45	**78.23**	47.73	48.55	55.49	**61.47**

separately. The corresponding weighted cross-entropy loss and Dice loss are employed as the regularization term, written as:

$$\mathscr{L}_{reg} = \sum_{i=1}^{N} \sum_{m \in \Omega} \left(\mathscr{L}_{WCE}(\hat{y}_{i,m}^{sep}, y_i) + \mathscr{L}_{DL}(\hat{y}_{i,m}^{sep}, y_i) \right), \tag{5.6}$$

where $\hat{y}_{i,m}^{sep}$ denotes the predicted segmentation mask of the i-th subject from the modality m. The segmentation-based regularizer enforces each modal encoder to be discriminative to each tumor region. In this manner, RFNet is able to obtain representative encoder features, thus improving the segmentation performance.

5.2.5 Overall Loss

As shown in Fig. 5.2, $\boldsymbol{D}_{\text{fuse}}$ is employed to predict the final segmentation mask from the fused features. The weighted cross-entropy loss and Dice loss are used to align the predictions to the corresponding ground-truth segmentation maps, expressed as:

$$\mathscr{L}_{fuse} = \sum_{i=1}^{N} \left(\mathscr{L}_{WCE}(\hat{y}_{i}^{fuse}, y_i) + \mathscr{L}_{DL}(\hat{y}_{i}^{fuse}, y_i) \right), \tag{5.7}$$

where $\hat{y}_i^{fuse}$ is the predicted segmentation mask from the i-th subject. Therefore, the overall loss of our RFNet is defined as:

$$\mathscr{L} = \mathscr{L}_{pm} + \mathscr{L}_{reg} + \mathscr{L}_{fuse}. \tag{5.8}$$

5.3 Experiments

5.3.1 Implementation Details

RFNet adopts 3D-Unet [33] with four-stage encoders ($\{\boldsymbol{E}_m\}_{m\in\Omega}$) and decoders ($\boldsymbol{D}_{sep}$ and $\boldsymbol{D}_{fuse}$). In the data preprocessing phase, images undergo skull stripping, co-registration, and adjusting the resolution to $1mm^3$ per voxel. Building upon the methods of previous studies [23, 48], we additionally eliminate the surrounding black areas of brains and normalize the brain area with a mean of zero and variance of one for all MRI modalities.

In the training phase, images undergo augmentation, including random crop ($80 \times 80 \times 80$), random rotations, brightness adjustments, and mirror flips. The network is trained for 300 epochs, and the batch size is set to 2. Besides, we exploit Adam [90] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ to optimized networks, and set weight decay to $1e^{-5}$. The "poly" learning rate strategy is employed, adjusting learning rate as $2e^{-4} \times (1 - \frac{epoch}{max\_epoch})^{0.9}$.

Following the approach [23], we segment volumes in a patch-wise way. To be specific, we segment $80 \times 80 \times 80$ patches, which slide over test volumes, and merge these patch predictions to obtain final predictions. During the sliding process, 50% overlap is ensured between adjacent patches. After prediction, we perform a post-processing step to minimize false alarms by suppressing minor components in the predictions. To be specific, brain tumors do not always have enhancing tumors. When the count of pixels predicted as enhancing areas is below 500, we treat these pixels as non-enhancing tumors, considering it as a false alarm.

5.3.2 Datasets and Evaluation Metric

Datasets: We assess RFNet using three datasets from BRATS [136], namely BRATS2015, BRATS2018, and BRATS2020. Samples in these three datasets contain usually contain four modalities, including T2, T1, T1c, and Flair.

BRATS2020 contains 219/50/100 for `train`/`val`/`test`. **BRATS2018** comprises 199/29/57 for `train`/`val`/`test`. Additionally, for BRATS2018, we adopt a three-fold validation using the same division lists as referenced in [48]. **BTATS2015** has 242/12/20 for

Table 5.2: Ablation study on RFNet. The average Dice scores of fifteen multi-modal combinations are reported. "Reg": the proposed segmentation-based regularizer, "RFM": the developed region-aware fusion module, "PostPro": the post-processing technique.

Methods	Average Dice scores (%)		
	Complete	Core	Enhancing
Baseline	83.20	71.72	53.73
+RFM	85.07	75.91	56.78
+Reg	86.07	76.89	57.96
+Reg+RFM	**86.98**	**78.23**	59.05
+Reg+RFM+PostPro	**86.98**	**78.23**	**61.47**

Table 5.3: The necessity of our regularizer and RFM. "wi rec regularizer': employing a reconstruction-based regularizer rather than the segmentation-based regularizer. "modal-wise" and "channel-wise": applying modal-wise and channel-wise attention to the feature maps instead of in a region-aware manner.

Methods	Average Dice scores (%)		
	Complete	Core	Enhancing
wi rec regularizer	85.38	75.50	59.64
channel-wise	85.81	76.36	60.11
modal-wise	85.87	77.02	61.01
RFNet	**86.98**	**78.23**	**61.47**

`train/val/test`. Given that BRATS2020 is the most recent and largest dataset, our primary focus in this work is on BRATS2020.

Evaluation Metric: Dice coefficient [39] is employed in this work, as defined by:

$$\text{Dice}_{\bar{k}}(\hat{y}, y) = \frac{2 \cdot \|\hat{y}_{\bar{k}} \bigcap y_{\bar{k}}\|_1}{\|\hat{y}_{\bar{k}}\|_1 + \|y_{\bar{k}}\|_1}, \tag{5.9}$$

where $\bar{k}$ represents various tumor classes. $\text{Dice}_{\bar{k}}$ refers to the Dice score for the tumor class $\bar{k}$. Higher Dice scores indicate greater similarity between predictions and the ground truth, reflecting improved segmentation accuracy.

5.3.3 Comparisons to State-of-the-arts

In Table 5.1 and Fig. 5.6, RFNet is compared with three SOTA methods, including HeMIS [66], U-HVED [48], and RobustSeg [23]. HeMIS [66] leverages the mean and variance of available modal features as the aggregated feature for segmentation. U-HVED [48] introduces multi-modal variational auto-encoders (MVAE) [199] to project different incomplete multi-modal images into a shared latent space. RobustSeg [23] disentangles content codes from appearance for segmentation and introduces a gated feature fusion to

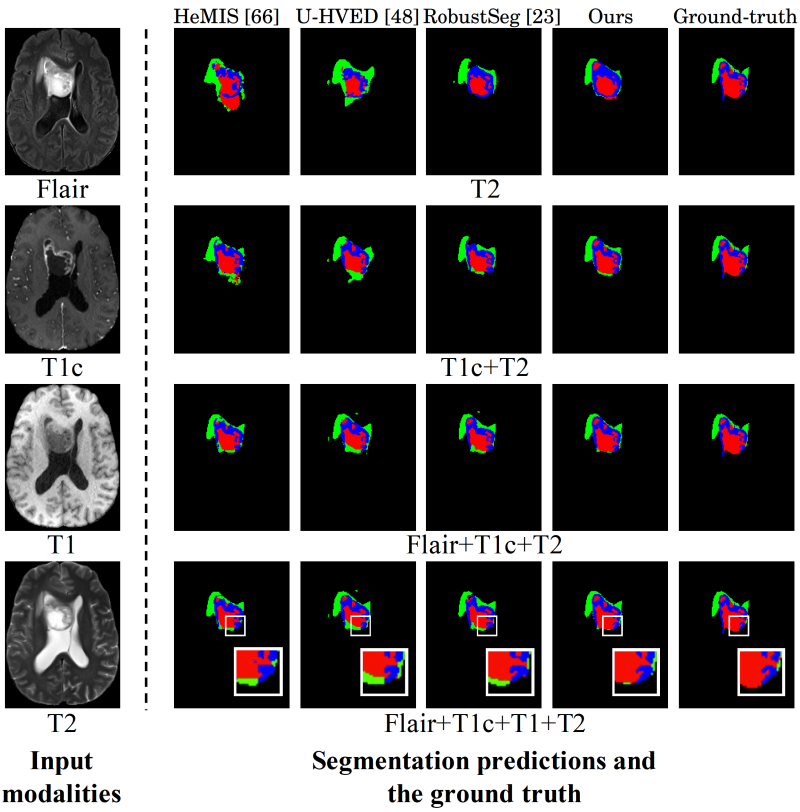

Figure 5.6: Visual comparison results. On the left are four image modalities. On the right, segmentation masks from various methods under four different missing situations.

aggregate multi-modal features. These methods do not explicitly take advantage of the relations between modalities and regions and neglect the unbalanced training problem.

As shown in Table 5.1, RFNet achieves superior segmentation performance. For instance, compared with the second best method, *i.e.*, RobustSeg [23], our RFNet improves the average Dice scores by 2.81%, 4.78% and 5.98% in the whole, core, and enhancing tumor, respectively. Moreover, RFNet outperforms the state-of-the-art methods on all fifteen multi-modal combinations. This demonstrates the superiority of RFNet.

5.3.4 Ablation Study

Table 5.2 reports the ablation study of RFNet. The baseline model leverages a $3 \times 3 \times 3$ convolutional layer to aggregate encoder features. As seen, the proposed region-aware fusion module and the segmentation-based regularizer can both improve the network significantly. For example, employing RFM increases the average Dice scores of three tumor areas by 1.87%, 4.19% and 3.05%, respectively. This is because RFM manages to effectively aggregate features and thus provides representative information for segmen-

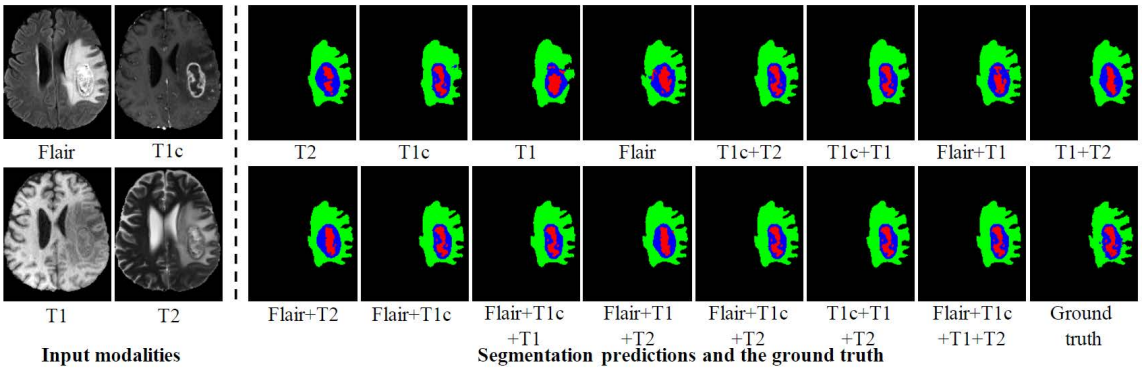

Figure 5.7: Visual results of RFNet. On the left are four image modalities. On the right, segmentation maps predicted by our RFNet under all missing situations.

Table 5.4: Quantitative segmentation results on BRATS2015. "†": reproduced based on the authors' code.

Methods	Average Dice scores (%)		
	Complete	Core	Enhancing
HeMIS [66]	68.22	54.07	43.86
U-HVED† [48]	81.57	64.68	56.76
RobustSeg [23]	84.45	69.19	57.33
Ours	**86.13**	**71.93**	58.98
Ours+PostPro	**86.13**	**71.93**	**64.13**

tation. Moreover, since the proposed regularizer helps the modal encoders discriminative to each region, applying the regularizer with RFM further improves segmentation results, as visible in Table 5.2. The post-processing technique is introduced to reduce false alarms of enhancing tumors, thus improving the segmentation performance of enhancing tumors. To demonstrate the effectiveness of the region-aware operation, we apply modal-wise attention to each modal feature (*i.e.*, a scalar for each modality) and channel-wise attention to all the concatenated features. As shown in Table 5.3, the model without the proposed region-aware operation yields inferior segmentation accuracy. This is because applying the same attention weights, either modal-wise or channel-wise attention, to the entire image does not enable a network to focus on the tumor regions. In Table 5.3, a reconstruction-based regularizer is adopted to replace the proposed segmentation-based regularizer but achieves inferior performance. This is because the reconstruction-based regularizer mainly focuses on restoring brain appearances rather than learning discriminative representations for tumor segmentation.

Table 5.5: Quantitative segmentation results on BRATS2018. "∗": provided by the authors.

Methods	Average Dice scores (%)		
	Complete	Core	Enhancing
HeMIS [66]	78.60	59.70	48.10
U-HVED [48]	80.10	64.00	50.00
RobustSeg* [23]	84.37	69.78	51.02
Ours	**85.67**	**76.53**	54.15
Ours+PostPro	**85.67**	**76.53**	**57.12**

5.3.5 Comparisons in BRATS2015 and BRATS2018

In addition to BRATS2020, we also validate the superiority of RFNet on BRATS2015 and BRATS2018 in Table 5.4 and Table 5.5, respectively. Note that, U-HVED [48] and RobustSeg [23] conduct experiments on only one dataset, *e.g.*, BRATS2018 or BRATS2015. Therefore, we obtain the BRATS2015 accuracy of U-HVED [48] with their official code and attain the BRATS2018 results of RobustSeg [23] from the authors. As shown in Table 5.4 and Table 5.5, our method improves the segmentation accuracy significantly on both two datasets. For instance, the average Dice scores of the three tumor areas on BRATS2018 are boosted by 1.30%, 6.75% and 6.10% by our RFNet. This validates the superiority of our method.

5.3.6 Visualization

Visualization of the Segmentation Results: In Fig. 5.7, we illustrate the segmentation masks from RFNet with all fifteen multi-modal combinations. Figure 5.7 demonstrates RFNet's capability to effectively segment brain tumors under various scenarios of missing data. For instance, RFNet produces an accurate segmentation map using only Flair and T1c modal images.

Visualization of the Attention Weights: In Fig. 5.8, we illustrate our generated attention weights, which are employed to fuse available modalities adaptively in each region. Since the deeper stage in RFNet encodes high-level semantic information, which is vital for segmentation, we opt to visualize the attention weights at the fourth stage. During inference, since missing modal features (zero tensors) provide no information, we set the corresponding attention weights to zero. As shown in Fig. 5.8, T1c modality (in red) receives more attention in NCR/NET and ET, while in ED, more attention is paid to Flair (in blue) and T2 (in yellow) modalities. This finding aligns with the observation in Fig. 5.1, which shows that the T1c modality is particularly sensitive to NCR/NET

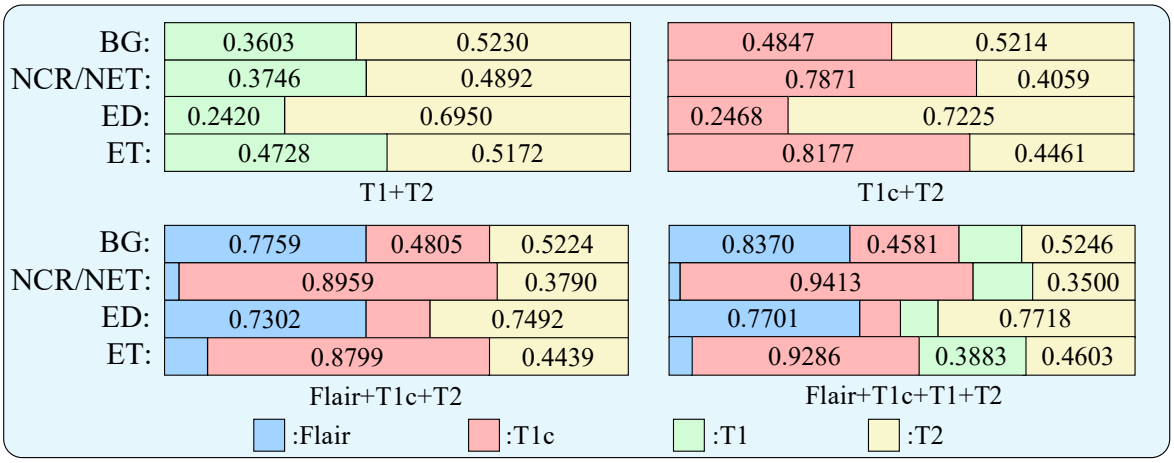

Figure 5.8: Visualization of the generated attention weights by our RFM at the fourth stage. The four panels demonstrate different cases of missing modalities. In each panel, attention weights (in numbers) are used to aggregate available modalities (in colors) adaptively in diverse regions (in rows). Larger colored boxes denote larger attention weights for the corresponding modality.

and ET regions, whereas the Flair and T2 modalities are more responsive to ED areas. Therefore, RFNet is able to provide larger attention weights for the sensitive modalities and thus obtains discriminative features for each region.

5.4 Conclusion

This chapter devises a region-aware fusion network (RFNet) for incomplete multi-modal brain tumor segmentation. RFNet develops a region-aware fusion module (RFM) to aggregate various available modalities effectively in a region-aware manner, thereby achieving more representative features and better segmentation performance. Besides, RFNet devises a segmentation-based regularizer, which not only improves each modal encoder but also expedites network training. Extensive experiments show that RFNet markedly surpasses the current state-of-the-art in performance. This chapter considers the missing modality challenges in real-world practice, which can largely improve the robustness of the medical image segmentation system (§6) to multi-modal images.

CHAPTER 6

CLUSTERING PROPAGATION FOR UNIVERSAL MEDICAL IMAGE SEGMENTATION

6.1 Introduction

In the realm of medical imaging, the practice of precisely revealing anatomical or pathological structure changes in a pixel observation holds the promise to substantially advance diagnostic efficiency[187]. Depending on the presence of user interactions, it can be categorized into automatic or interactive medical image segmentation (AMIS/IMIS)[149], with the latter involving active user engagement (*e.g.*, click, scribble) throughout the segmentation process[179, 239].

Benefiting from the rapid development of deep learning techniques, both AMIS and IMIS have witnessed great progress in their respective field. For AMIS, the emerging of seminal work [153] leads the research efforts towards developing stronger backbones [33, 140, 172], harnessing multi-scale features [84, 163, 243] or incorporating attention mechanism [40, 144, 192, 196], *etc*. Conversely, IMIS centers its primary focus on effectively integrating user inputs into segmentation models[130, 181], yielding remarkable performance. Nevertheless, such a tailored paradigm for each task greatly diffuses the research endeavors, impeding the seamless transfer of advancement made from one task to another due to the fundamental differences in model architecture and

This chapter is based on joint work [41] with LiuLei Li, Wenguan Wang, and Yi Yang, presented primarily as it appears in the CVPR 2024 proceedings.

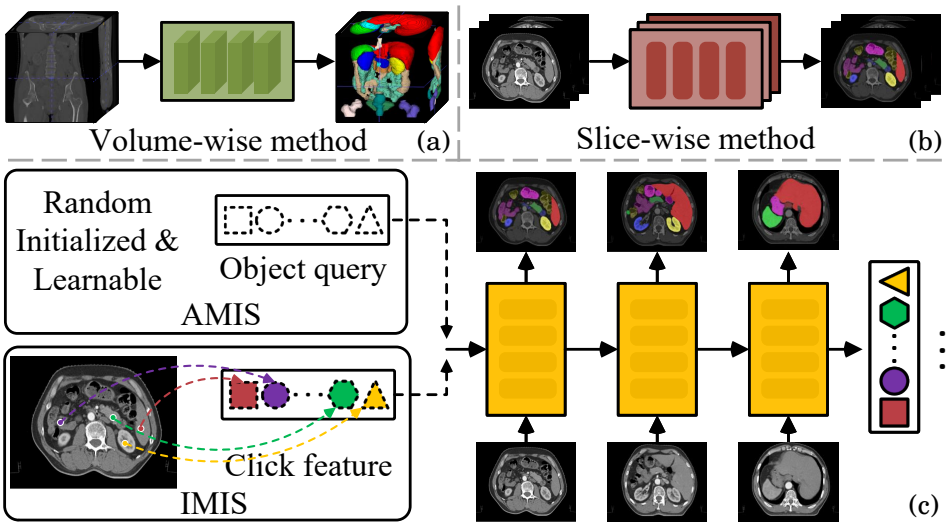

Figure 6.1: (a-b) Existing *volume-wise* and *slice-wise* solutions. (c) Our slice-to-volume solution that bridges distant slices by cluster center propagation and further unifies automatic/interactive segmentation under the same model with 2D segmentation networks.

training strategy. Moreover, when working with the same dataset, current solutions necessitate the development of two separate models for AMIS and IMIS, respectively. This results in a duplication in terms of both training time and network parameters.

In this work, we aim to formulate a ***universal*** segmentation framework capable of addressing both AMIS and IMIS within ***one unified model*** and a ***single training session***. Towards this, we first initiate a thorough exploration of the limitations commonly observed in current AMIS and IMIS solutions: **i)** the top-leading approaches for volume segmentation rely heavily on 3D networks which suffer from slow inference [84] and present significant challenges in deploying on hospital devices that usually exhibit limited parallel computation capabilities, **ii)** they prove inefficient in bridging remote slices due to the usage of sliding window inference to handle large memory consumption, which further hinders the broadcast of user inputs to entire volumes, **iii)** current interactive solutions are limited to handling single foreground class, in contrast to automatic approaches, which develop rapidly and excel in multi-class segmentation.

To solve the aforementioned limitations as well as reconcile AMIS and IMIS, we proposed S2VNet. It draws inspiration from the clustering-based image segmentation methods [112, 142, 215, 216] that utilize a set of learnable queries as cluster centers to aggregate pixel features associated with target objects and update in an iterative manner. This insightful approach prompts us to reformulate volumetric segmentation by utilizing mere 2D segmentation models. Specifically, it is observed that objects in a volume usually manifest identical representations across different slices. This inherent consistency forms the basis for a novel slice-to-volume propagation method that centroids,

after comprehensive updates in one slice, can be passed forward and serve as the initial values for cluster centers in successive slices, facilitating effortless transfer of knowledge retrieved in the prior segmentation process to the next round. This paradigm is simple yet powerful, harnessing both the key principle of clustering-based methodologies and the slice-wise structure of volumetric data. Moreover, this framework is readily adapted to IMIS without architectural changes by initializing centroids from backbone features at the position of user inputs, which clearly signify intended objects. Since there would be multiple clicks for an identical object, we further design an adaptive sampling strategy to reweight feature points when given new interactions. Finally, as the current pipeline may be affected by outliers and face decaying awareness of prior cues after rounds of propagation, we devise a recurrent centroid aggregation strategy to collect historic centroids and fuse them into a single vector, which introduces nearly no additional cost to deliver a more robust network inference.

Taking advantage of such a slice-to-volume propagation paradigm, S2VNet unveils several compelling facets: **First**, it seamlessly accommodates AMIS and IMIS into a unified model through a single training process, accomplished by initializing a subset of cluster centers from user inputs while the others are left as random, enabling both automatic and interactive segmentation learning. **Second**, in leveraging of reusing centroids, S2VNet extends user inputs or slice cues throughout the entire volume with 2D networks, contributing to a significant alleviation in computational resource (*i.e.*, 15 times faster inference speed and 48.2% memory reduction compared to 3D counterparts). **Third**, S2VNet can simultaneously accept multiple classes of user inputs, with each of them initializing one cluster center. This facilitates parallel refinement for multiple instances of ***different classes*** in a ***single network forward pass***, while prior work could only handle one foreground class [121, 181, 238]. **Fourth**, given the universal characteristic of S2VNet, we could build a diagnosis system that meets rigorous clinical requirements. Concretely, S2VNet is able to provide coarse segmentation results for multiple lesion/organ classes via AMIS. Physicians can then choose instances of interest and conduct refinement with precise feedback, saving considerable time in the initial search for lesions/organ areas.

We open a new avenue for medical image segmentation from the universal perceptive and further provide a feasible solution via clustering-based slice-to-volume propagation. To comprehensively evaluate our method, we experiment with S2VNet on three volumetric datasets, *i.e.*, WORD [129], BTCV [97], and AMOS [80]. Our empirical findings substantiate that S2VNet could consistently yield superior performance even compared

to the specified solutions for each task through the utilization of only one single model. Our implementation will be released upon acceptance.

6.2 Method

6.2.1 Preliminary: K-Means Cross-Attention

Inspired by DETR[20], contemporary query-based image segmentation methods[30, 31] typically introduce a set of learnable embeddings as queries to collect pixel features associated with specific objects via *cross-attention*:

$$\hat{\boldsymbol{C}} = \boldsymbol{C} + \text{softmax}_{HW}(\boldsymbol{Q}(\boldsymbol{K})^{\top})\boldsymbol{V}, \tag{6.1}$$

where $\boldsymbol{C} \in \mathbb{R}^{N \times D}$ represents N object queries with dimension size D, $\hat{\boldsymbol{C}}$ denotes the updated queries, $\boldsymbol{Q} \in \mathbb{R}^{N \times D}$, $\boldsymbol{K} \in \mathbb{R}^{HW \times D}$, $\boldsymbol{V} \in \mathbb{R}^{HW \times D}$ stand for the features for query, key, and value. Here softmax_{HW} means to conduct softmax along the spatial dimension of image features, *i.e.*, computing the probability of affiliated to a unique query across all pixels. It is crucial to note that this mechanism involves attending to a substantial number of pixels. In contrast to above, [14] devise the *k-means* cross attention:

$$\hat{\boldsymbol{C}} = \boldsymbol{C} + \text{argmax}_{N}(\boldsymbol{Q}(\boldsymbol{K})^{\top})\boldsymbol{V}. \tag{6.2}$$

Here, Eq. 6.2 compels $\boldsymbol{Q}$ to query pixel features belonging to a specific object, and subsequently inspect which query embedding within $\boldsymbol{C}$ these features correspond to by applying argmax along the query dimension N. Such process is similar to the *k-means*[125] algorithm which proceeds by alternating between the *assignment* and *update* two steps:

$$\begin{aligned} \text{Assignment Step:} \quad & \hat{\boldsymbol{C}} = \boldsymbol{A}\boldsymbol{V}, \\ \text{Update Step:} \quad & \hat{\boldsymbol{C}} = \boldsymbol{C} + \hat{\boldsymbol{C}}, \end{aligned} \tag{6.3}$$

where $\boldsymbol{A} = \text{argmax}_{N}(\boldsymbol{Q}(\boldsymbol{K})^{\top})$ is the assignment matrix (*i.e.*, attention map) where each element indicates whether a pixel feature is assigned to a particular cluster. As a results, following the execution of a succession of Transformer decoder layers composed by *k-means* cross attention, the query embeddings $\boldsymbol{C}$ can be regarded as the cluster centers, which adeptly captures the representation of target objects.

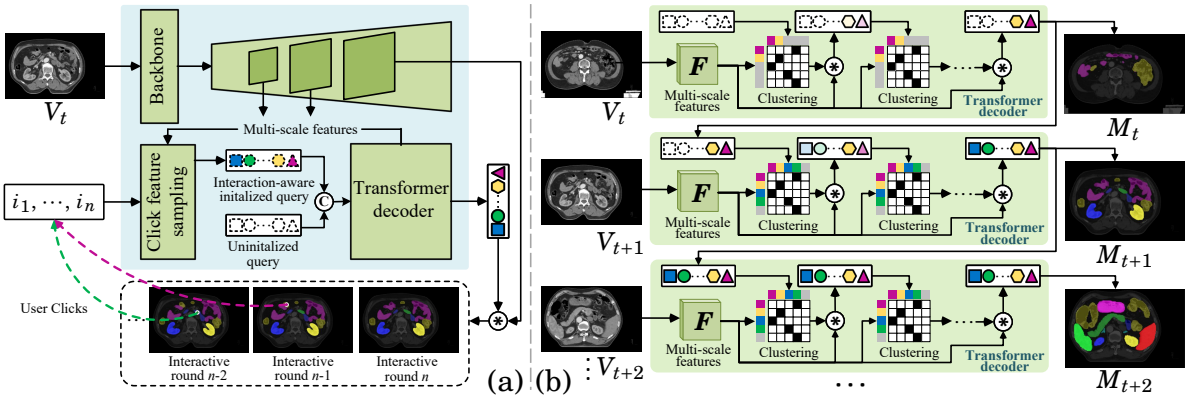

Figure 6.2: Our centroid propagation-driven universal segmentation framework (§6.2.2). (a) S2VNet adapts multi-class interactive segmentation and refinement by iteratively initializing cluster centers from user clicks and propagating to the entire volume. (b) Our proposed clustering-based slice-to-volume propagation pipeline where the centroids are evolved during slice-level segmentation and passed to the next slices.

6.2.2 Centroid Propagation-Driven Universal Segmentation Framework

Motivation. Given a volume $V \in \mathbb{R}^{C\times H\times W}$ with a spatial size of $H\times W$ for C slices, volumetric image segmentation aims to group it into a set of segments with corresponding semantic labels. This task is distinguished by the inherent structural property of volumetric image data, *i.e.*, anatomical or pathological regions of interest often spanning across multiple consecutive slices and exhibiting consistent visual patterns. This property allows the same class of targets in distinct slices to be compressed within a shared object-centric representation. Given this context, we introduce the clustering-based methodologies into volume segmentation. Specifically, our approach involves extending the dynamic evolution of cluster centers $\boldsymbol{C}$, which is originally conducted within the image-level mask decoding process to volume-level by using the same collection of $\boldsymbol{C}$ throughout the segmentation for all slices in V. As such, the separate *slice-wise* segmentation for each individual slice is seamlessly integrated into a coherent segmentation process, and iteratively delivering intermediate output for each slice.

Slice-to-Volume Cluster Center Propagation. Denoting $\mathscr{F}$ as the feature encoder, N cluster centers $\{\boldsymbol{C}_n^t\}_{n=1}^N$ are employed to extract the object-centric representation for each class within the given slice V_t by:

$$\{\hat{\boldsymbol{C}}_n^t\}_{n=1}^N = \mathscr{D}(\mathscr{F}(V_t), \{\boldsymbol{C}_n^t\}_{n=1}^N), \tag{6.4}$$

where $\mathscr{D}$ is the Transformer decoder composed of *k-means* cross attention [216]. In the context of automatic volumetric image segmentation, the segmentation often begins from the first slice along the z-axis of the volume, which typically contains no foreground

objects. It is common for these foreground objects to appear in the middle part of the volume. To address the challenge that all cluster centers collect features of the background class and further impose negative impact to the segmentation of subsequent slices, only cluster centers matched with foreground classes will be propagated to the next slice. To achieve this, we perform one-to-one bipartite matching between the mask predictions $\{\hat{Y}_n^t\}_{n=1}^N$ and the ground truth $\{Y_k^t\}_{k=1}^K$ by:

$$\hat{\theta} = \arg\min_{\theta \in \Theta_N} \sum_{n=1}^{N} \mathscr{L}_{\text{match}}(Y_n, \hat{Y}_{\sigma(n)}). \tag{6.5}$$

Here $\hat{\theta}$ represents the optimal assignment among a permutations of N elements $\theta \in \Theta_N$. Based on $\hat{\theta}$, we select cluster centers $\{\hat{\boldsymbol{C}}_k^t\}_{k=1}^K$ associated with foreground classes and pass them to the next slice V_{t+1} as the initial values:

$$\{\boldsymbol{C}_k^{t+1}\}_{k=1}^K = \{\hat{\boldsymbol{C}}_k^t\}_{k=1}^K. \tag{6.6}$$

As such, these object-centric representations could encapsulate the coherent appearances of regions across different slices, fostering a more compact and informative representation for subsequent segmentation and analysis. Note that during the inference stage, we keep elements in $\{\hat{\boldsymbol{C}}_n\}_{n=1}^N$ only if the corresponding class $\{\hat{c}_n^t\}_{n=1}^N$ is not identified as the background class, and pass them to subsequent slices.

Interaction-Aware Cluster Center Initialization. In prior research [52, 178, 181], the user input is conventionally represented as an binary mask $M \in \{0,1\}^{H \times W}$ where the foreground region signifies user guidance. Subsequently, M is combined with gray-scale images as inputs to segmentation networks. Though achieving promising results, such a paradigm suffers from several limitation: **i)** concatenating user inputs with images introduces architectural modifications and disrupts the integration with automatic segmentation into a unified framework, and **ii)** prior methods encounter challenges when accommodating multiple semantic classes, thereby limiting the application to more complex scenarios. To tackle the above limitations, instead of explicitly incorporating user guidance as the input to networks, we harness the clustering-based property of S2VNet. Specifically, denoting $\{Q_k\}_{k=1}^K = \{(P_k, c_k, t_k)\}_{k=1}^K$ as a set of user inputs where each element Q_k represents a click P_k associated for one semantic class c_k annotated on the slice V_{t_k}, we initialize the cluster center $\boldsymbol{C}$ from user input by:

$$\begin{aligned} \hat{\boldsymbol{C}}_k &= \texttt{FFN}(\boldsymbol{O}_k), \\ \boldsymbol{O}_k &= \texttt{Sample}(\boldsymbol{F}_{t_k}, P_k), \end{aligned} \tag{6.7}$$

where `Sample` indicates retrieving the point feature $\boldsymbol{O}_k$ from backbone features $\boldsymbol{F}_{t_k}$ of slice V_{t_k} according to the click position P_k, and `FFN` is a simple feed forward network

to project $\boldsymbol{O}_k$ to the same size as $\boldsymbol{C}$. $\hat{\boldsymbol{C}}_k$ further serves to aggregate pixel features similar to the user-indicated regions and will be passed to subsequent slices. This realizes user-guided segmentation across the whole volume by leveraging the above automatic segmentation pipeline while introducing no modification to the network architecture. Moreover, it can accommodate an **arbitrary number of classes** with each of them serving to initialize one cluster center, perfectly addressing all aforementioned limitations. Notably, extending beyond these benefits, such a centroid initialization-based interactive segmentation strategy offers several additional advantages: **first**, in contrast to prior work treating user interactions and images equally by concatenating them as inputs, which can not exercise the guidance ability of interactions to the fullest extent, our interaction-aware centroid initialization implicitly guarantees predictions always conforming to user highlighted regions and enhances interpretability. **Second**, our method enables unified learning for interactive/automatic segmentation, as the only difference lies in the initial states of centroids. The input data, network architecture, and training objectives remain consistent.

Adaptive Pixel Feature Sampling. Interactive segmentation commonly involves multiple rounds of refinement to improve the precision of previously segmentation results by incorporating newly provided user inputs. These iterative refinements yield multiple instances of Q_k associated with the same category label, thus calling for an adaptive strategy to initialize cluster centers for a specific semantic category from multiple user inputs. As the latest user input should play a more important role in refinement compared to prior clicks, we adopt a weighted sum to combine the pixel feature $\boldsymbol{O}_k^r$ sampled from the user input at the current refinement round r with those sampled from prior rounds by:

$$\begin{aligned}\hat{\boldsymbol{O}}_k^r &= \boldsymbol{O}_k^r + \beta^1 \cdot \boldsymbol{O}_k^{r-1} + \cdots + \beta^n \cdot \boldsymbol{O}_k^1, \\ &= \boldsymbol{O}_k^r + \beta \cdot \hat{\boldsymbol{O}}_k^{r-1},\end{aligned} \tag{6.8}$$

where $\hat{\boldsymbol{O}}_k^r$ is the weighted output controlled by the factor $\beta \in [0,1]$. Then $\hat{\boldsymbol{O}}_k^r$ at each round of refinement is utilized to initialize a new cluster center, delivering a pair of prediction $\{\hat{\boldsymbol{M}}_k^r, \hat{\boldsymbol{c}}_k^r\}$ where $\hat{\boldsymbol{M}}_k^r \in \mathbb{R}^{C\times H\times W}$ is the binary mask score for all C slices in volume V and $\hat{\boldsymbol{c}}_k^r \in \mathbb{R}^C$ is the score for class c_k. Consequently, multiple predictions are delivered for each semantic class. To obtain the ultimate output, we first multiply $\hat{\boldsymbol{c}}_k$ with corresponding $\hat{\boldsymbol{M}}_k$ and then retrieve the maximum value across all R rounds of predictions by:

$$\hat{M}_k = \max_R(\hat{\boldsymbol{M}}_k^0 \cdot \hat{\boldsymbol{c}}_k^0, \cdots, \hat{\boldsymbol{M}}_k^R \cdot \hat{\boldsymbol{c}}_k^R). \tag{6.9}$$

It is crucial to emphasize that for all refinement rounds in S2VNet, the pixel features associated with user inputs are sampled from the ***same backbone features*** which only

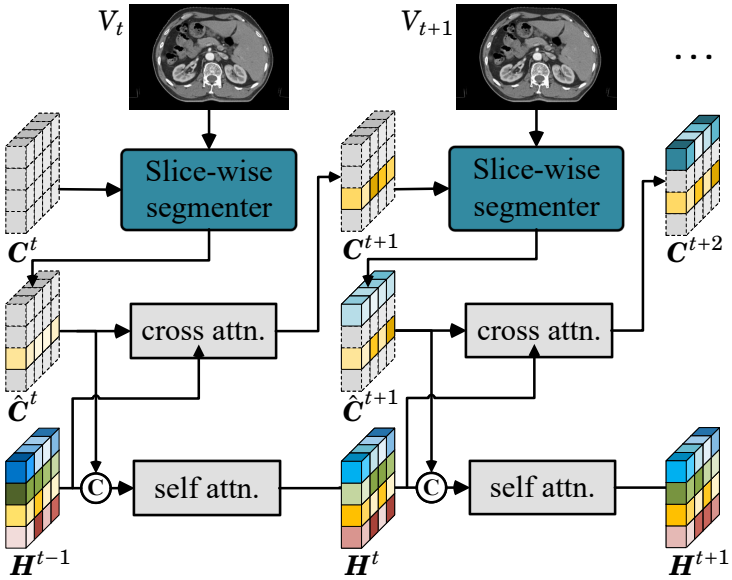

Figure 6.3: Illustration of recurrent centroid aggregation (§6.2.2). After clustering within the slice-wise segmentation for each slice, the centroids are recurrently merged with the historical ones to assist in the initialization of centroids belonging to the subsequent slice.

need to be computed once. This stands in stark contrast to prior work [52, 179, 181] that repetitively combines prior results with image data and conducts a full network pass at each refinement round. This also contributes to accelerated inference and enhances the efficiency of computer-aided diagnosis.

Recurrent Centroid Aggregation. Though the cluster centers undergo continuous evolution during the mask decoding so as to effectively associate successive slices, they tend to be drifted by outliers such as foreign objects and artifacts commonly encountered in clinical practice [150] and lose track of distant structural cues with the slice-wise segmentation process iterates. To deliver a robust inference and retain enduring cues of remote slices, we propose to accumulate historic centroids of each slice and fuse them into a consolidated entity in a recurrent manner. Specifically, denoting $\boldsymbol{H}_k^{t-1}$ as the fused vector for cluster center $\boldsymbol{C}_k$ covering its value from slice V_0 to V_{t-1}. When given new centroid $\hat{\boldsymbol{C}}_k^t$ after mask decoding for slice V_t, we fuse it with $\boldsymbol{H}_k^{t-1}$:

$$\boldsymbol{H}_k^t = \texttt{FFN}(\texttt{SelfAttn}([\boldsymbol{H}_k^{t-1}; \hat{\boldsymbol{C}}_k^t])), \tag{6.10}$$

where $[;]$ means concatenation. Here `SelfAttn` is employed to identify the most relevant information within the concatenated vector $[\boldsymbol{H}_k^{t-1}; \hat{\boldsymbol{C}}_k^t]$, and `FFN` is subsequently used to project it into the same dimension as $\hat{\boldsymbol{C}}_k^t$. In this way, rather than introducing a memory bank that would impose additional GPU memory and computational time overhead, we efficiently store historic structural cues by recurrently merging new centroids into the existing one. Then, when initializing the centroid $\boldsymbol{C}_k^{t+2}$ for slice V_{t+2}, we incorporate not

only the cluster center obtained after mask decoding at slice V_{t+1} (*i.e.*, $\hat{\boldsymbol{C}}_k^{t+1}$), but also query the centroids from the previous t slices stored in $\boldsymbol{H}_k^t$ by:

$$\boldsymbol{C}_k^{t+2} = \hat{\boldsymbol{C}}_k^{t+1} + \texttt{CrossAttn}(\hat{\boldsymbol{C}}_k^{t+1}, \boldsymbol{H}_k^t). \tag{6.11}$$

We adopt standard cross-attention here as $\boldsymbol{H}_k^t$ can benefit multiple elements in $\hat{\boldsymbol{C}}_k^{t+1}$ and there is no need to enforce an exclusive relation via argmax in *k-means* cross attention.

6.2.3 Implementation Details

Network Configuration. S2VNet is constructed upon the clustering-based image segmenter. Specifically, for the slice-wise segmentation, we adopt Mask2Former[30] and integrate *k-means* cross attention[216] to replace the standard ones in the Transformer decoder. Other setups remain consistent with the default configuration. In order to align S2VNet with the most recent top-leading solutions[18, 64, 235] for medical image segmentation that favor Transformer-based backbones, we employ Swin-B [124] for feature extraction. For the weighted factor β utilized in adaptive pixel feature sampling, we empirically set it as 0.8.

Interaction Simulation. To evaluate S2VNet under the interactive setup, we opt for click as the primary mode of user interaction, which is generally more accessible and can accommodate various input devices like mice, touchscreens, and styluses. Following conventions[181, 204, 238], we adhere to the automatic evaluation pipeline wherein the clicks are simulated based on ground truth and current segmentation results. Specifically, the initial click is sampled near the center of the target object, while subsequent clicks aimed at refinement are generated iteratively from the most significant error regions by comparing the current prediction mask with the ground truth. The user clicks comprise both positive and negative ones, with the former targeting foreground objects and the latter being applied to the background.

Unified Segmentation Learning. To facilitate the slice-to-volume propagation learning, we randomly sample three slices from each volume and use clustering results obtained in the previous slice to initialize centroids for the next slice. We designate 20 cluster centers for each semantic class, with each click serving as the trigger to initialize one of them, *i.e.*, allowing up to 20 clicks. Notably, for classes presenting in the inputs, there exists a 50% probability that the cluster centers are initialized from simulated user clicks, while the left are randomized initialized from empty, so as to enable both automatic and interactive segmentation learning. Following prior work[64, 79, 181, 235], the final learning target is the combination of the Cross Entropy loss and Dice loss.

6.3 Experiments

6.3.1 Experimental Setup

Datasets. Our experiments are conducted on three datasets:

- **WORD** [129] is a large-scale real clinical abdomen benchmark, providing high-quality annotations for up to 16 organs in the abdominal region. It contains 100/20/30 CT images for `train/val/test`, respectively.
- **BTCV** [97] consists of 30 CT volumes which is divided into 24 and 6 volumes for `train` and `val`. This dataset provides careful annotation for 13 organs, including 8 of them from Synapse. Following existing work[64, 235], We report the DSC score on all 13 abdominal organs.
- **AMOS** [80] is a large-scale, diverse dataset collected from multiple centers and provides voxel-level annotations for 15 abdominal organs. It covers CT and MRI two modalities, with each of them containing 200/100/200 and 40/20/40 scans for `train/val/test`.

Training. We train S2VNet for 20k iterations and set the batch size to 8. The AdamW [90] optimizer with an initial learning rate 0.0002 and weight decay 0.02 is adopted. The learning rate is scheduled following the step policy, *i.e.*, decaying by 10 at 14K and 18K steps, respectively. A learning rate multiplier of 0.1 is applied to the backbone, which is initialized with ImageNet[37] pre-trained weights. After adapting the volumetric data into 2D slices, we employ z-score normalization to rescale image intensities within the range of 0 to 255. The remaining setups are determined following [25, 64, 79, 181, 238] for fair comparison. Specifically, for data augmentation, we use standard large-scale jittering (LSJ) augmentation with a random scaling sampled from range 0.5 to 1.75, followed by a fixed-size crop of 512×512 for WORD[129], 256×256 for BTCV[97], 256×256 for AMOS[80]. Random horizontal flipping is also applied to enhance diversity.

Testing. The inference steps are tailored to optimize the usage of user inputs. Please note that we adopt an identical network architecture and model weight for both two tasks:

- ***Automatic.*** Inference starts from the first slice along the z-axis, proceeding sequentially till the final slice.
- ***Interactive.*** Inference is initiated from the slice with user inputs and broadcast bidirectionally throughout the entire volume, emphasizing the significance of user interactions.

Methods	Average		Liv	Spl	Kid L	Kid R	Sto	Gal	Eso	Pan	Duo	Col	Int	Adr	Rec	Bla	Fem L	Fem R
	HD95 ↓	DSC ↑																
Automatic Setup																		
UNETR [64]	17.34	79.77	94.67	92.85	91.49	91.72	85.56	65.08	67.71	74.79	57.56	74.62	80.40	60.76	74.06	85.42	89.47	90.17
CoTr [203]	12.83	84.66	95.58	94.90	93.26	93.63	89.99	76.40	74.37	81.02	63.58	84.14	86.39	69.06	80.00	89.27	91.03	91.87
Swin UNETR [63]	14.24	84.34	96.08	95.32	94.20	94.00	90.32	74.86	76.57	82.60	65.37	84.56	87.37	66.84	79.66	92.05	86.40	83.31
ESPNet [135]	15.02	79.92	95.64	93.90	92.24	94.39	87.37	67.19	67.91	75.78	62.03	78.77	72.80	60.55	74.32	78.58	88.24	89.04
DMFNet [24]	7.52	85.10	95.96	94.64	94.70	94.96	89.88	79.84	74.10	81.66	66.66	83.51	86.95	66.73	79.62	88.18	91.99	92.55
LCOVNet [229]	9.11	85.82	95.89	95.40	95.17	95.78	90.86	78.87	74.55	82.59	**68.23**	84.22	87.19	69.82	79.99	88.18	92.48	93.23
SwinMM [194]	9.35	86.18	96.30	95.46	93.83	94.47	91.43	80.08	76.59	83.60	67.38	**86.42**	**88.58**	69.12	80.48	90.56	92.16	92.40
S2VNet (Ours)	**4.64**	**87.36**	**96.72**	**96.01**	**95.84**	**95.93**	**91.80**	**82.96**	**77.28**	**85.10**	67.07	86.19	88.46	**72.40**	**83.27**	**91.73**	**93.30**	**93.75**
Interactive Setup																		
iSegFormer† [119]	-	-	-	92.14†	91.07†	93.86†	-	72.01†	73.37†	-	69.52†	-	-	69.91†	48.13†	-	-	-
Mem3D† [239]	-	-	-	94.88†	93.55†	93.96†	-	77.38†	80.61†	-	76.29†	-	-	74.57†	73.37†	-	-	-
SwinMM† [194]	-	-	-	95.78†	94.27†	95.11†	-	82.26†	80.33†	-	78.54†	-	-	72.96†	85.12†	-	-	-
S2VNet (Ours)	**3.28**	**91.41**	**96.91**	**96.37**	**96.15**	**96.22**	**94.79**	**87.23**	**86.32**	**88.51**	**83.91**	**90.50**	**91.17**	**77.73**	**90.73**	**94.35**	**95.85**	**95.82**

†: An independent model is trained for each class as prior work can only handle binary segmentation.

Table 6.1: Quantitative segmentation results with comprehensive scoring for each organ on WORD[129] `test`.

For fair comparison, we follow prior work[194, 235] to use the input resolution of 512×512 for all datasets[80, 97, 129].

Evaluation Metric. Following the standard evaluation protocol[2, 129, 235], We adopt Dice Similarity Coefficient (DSC)[39], Hausdorff Distance (HD)[77] and normalized surface dice (NSD)[80] to assess the performance under both automatic and interactive setups. DSC quantifies the overlap between predictions and ground-truths, whereas HD functions for measuring the 3D surface distance between them. To eliminate the impact of outliers, we employ HD95, which captures the 95% distance of all points from one surface to the other. For NSD, it scores the category-wise segmentation quality for evaluating the precision of boundaries.

Reproducibility. S2VNet is implemented in PyTorch and trained on four NVIDIA Tesla A100 GPUs. Evaluation for all methods is conducted on the same machine. Our full implementation shall be related to guarantee reproducibility.

IMIS Comparison. As existing interactive approaches[100, 113, 119, 239] are limited to binary segmentation with a single foreground class, we train an independent model for each target class while considering remaining classes as background. To render a more comprehensive comparison, we adapt the top-leading automatic work into the interactive setup by concatenating user clicks and prior round predictions with image data. Given this substantial workload, we only report performance for several representative classes with relatively lower performance across each dataset.

6.3.2 Comparison to State-of-the-arts

WORD[129]. As shown in Table 6.1, S2VNet yields remarkable performance on the automatic setup, *i.e.*, surpassing SwinMM[194] by **1.18%** in terms of DSC and outperforming

Method	Avg DSC	Gal	Eso	IVC	PSV	RAG	LAG
Automatic Setup							
TransUNet [25]	76.72	59.84	70.96	77.23	71.47	65.24	64.06
TransBTS [190]	81.31	68.38	75.61	82.48	74.21	67.23	67.03
UNETR [64]	76.00	58.23	71.21	76.51	70.37	66.25	63.04
Swin-UNETR [63]	80.44	65.37	75.43	81.61	76.30	68.23	66.02
nnFormer [235]	81.62	65.29	76.22	80.80	75.97	**70.20**	66.05
3D-UX-Net [101]	80.76	64.32	75.17	80.42	75.39	69.52	65.77
S2VNet (Ours)	**83.81**	**65.63**	**78.29**	**84.41**	**79.77**	68.38	**72.28**
Interactive Setup							
iSegFormer† [119]	-	-	69.37†	72.78†	-	64.40†	66.89†
Mem3D† [239]	-	-	74.84†	79.52†	-	68.45†	67.88†
nnFormer† [235]	-	-	82.47†	83.65†	-	70.41†	67.34†
S2VNet (Ours)	**86.11**	**69.94**	**87.92**	**89.96**	**81.64**	**72.23**	**73.22**

†: An independent model is trained for each target class.

Table 6.2: Quantitative segmentation results on BTCV[97] `val`.

all 3D solutions in terms of HD95 which emphasizes on the coherence of predictions across slices. This demonstrates the effectiveness of our 2D slice-to-volume propagation strategy in bridging distance cues. Under the interactive setup, S2VNet achieves a **4.05%** average improvement in DSC compared to the automatic setup, verifying the superiority of our interaction-aware centroid initialization strategy. Especially, our approach boosts the performance up to **83.91%** for the class 'Duo.', surpassing both existing interactive and adapted automatic approaches by a large margin.

BTCV [97]. Table 6.2 compares our method against several top-leading approaches on BTCV[97] `val`. As seen, S2VNet achieves the best performance on both automatic and interactive setups. In particular, compared with nnFormer[235], which is the previous SOTA, our approach earns **2.19%** improvement in terms of averaged DSC score for the automatic setup. This indicates that S2VNet can generalize well to different datasets with various challenging scenarios. We also provide detailed scores for six representative organs with poor performance, where S2VNet gives **2%**~**6%** performance gain compared to prior work.

AMOS [80]. Table 6.3 confirms again the exceptional performance of S2VNet in the segmentation of both CT and MRI images. Specifically, our algorithm achieves an improvement of **0.52%**/**6.41%** over 3D-UX-Net[101] in terms of DSC/NSD. Moreover, with the incorporation of interaction-aware query initialization, S2VNet consistently surpasses existing methods across all modalities and metrics.

Method	Average		CT		MRI	
	DSC↑	NSD↑	DSC↑	NSD↑	DSC↑	NSD↑
Automatic Setup						
CoTr [203]	77.31	67.12	77.13	64.15	77.50	70.10
UNETR [64]	76.81	63.40	78.33	61.49	75.30	65.3
TransUNet [25]	-	-	85.05	73.86	-	-
TransBTS [190]	-	-	86.52	75.49	-	-
nnFormer [235]	83.12	74.07	85.63	74.15	80.60	**74.00**
Swin UNETR [63]	81.04	70.60	86.37	75.32	75.70	65.80
3D-UX-Net [101]	-	-	87.28	76.48	-	-
S2VNet (Ours)	**86.22**	**77.23**	**87.80**	**82.89**	**84.64**	71.57
Interactive Setup						
S2VNet (Ours)	**88.75**	**80.94**	**89.65**	**85.27**	**87.84**	**76.61**

Table 6.3: Quantitative segmentation results on AMOS[80] `val`.

#	S2V Propagation	Interaction Initialization	Adaptive Sampling	Recurrent Aggregation	HD95 ↓	DSC ↑
1					16.63	78.67
2	✓				5.03	86.19
3	✓			✓	4.64	87.36
4	✓	✓			4.30	89.70
5	✓	✓	✓		3.79	90.64
6	✓	✓	✓	✓	**3.28**	**91.41**

Table 6.4: Analysis of essential component on WORD[129] `test`.

6.3.3 Qualitative Comparison Result

Fig. 6.4 depicts visual comparison on WORD[129] `test`. As seen, S2VNet yields more accurate results compared to SwinMM[194], and the interactive mode can handle various challenging cases with small objects or distortions.

6.3.4 Diagnostic Experiments

To evaluate the core designs and gain further insights, we conduct a series of ablative studies on WORD[129] `test`.

Key Component Analysis. We first examine the efficacy of each component in Table 6.4, where the *row* #1 indicates directly segmenting each slice using 2D networks without any form of association. Upon the integration of clustering-based slice-to-volume propagation (*i.e.*, *row* #2), both DSC and HD95 exhibit noteworthy improvement, which demonstrates the effectiveness of our design. For interactive segmentation, as seen in

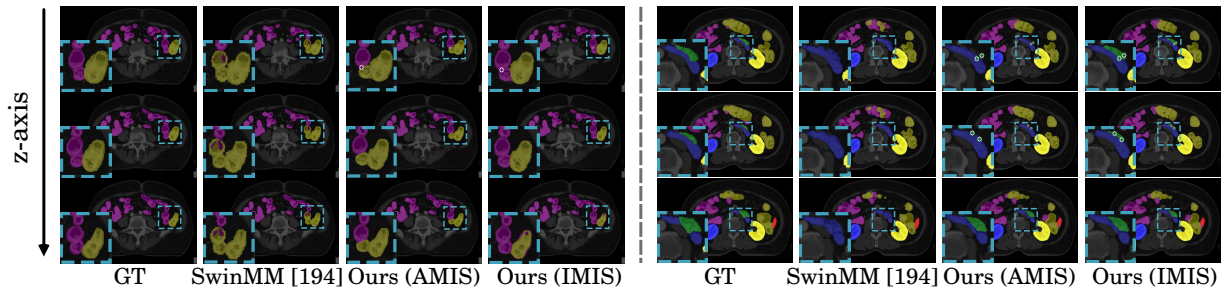

Figure 6.4: Visual comparison results on WORD[129] `test`. See §6.3.3 for detailed analysis.

Method	Memory (G) ↓	Volume Per Minute ↑	HD95 ↓	DSC ↑
CoTr[203]	26	0.18	12.83	84.66
Swin UNTER[63]	23	0.21	14.24	84.34
SwinMM[194]	27	0.15	9.35	86.18
Baseline	11	2.69	16.63	78.67
S2VNet	**14**	**2.33**	**4.64**	**87.36**

Table 6.5: Comparison of running efficiency on WORD[129] `test`.

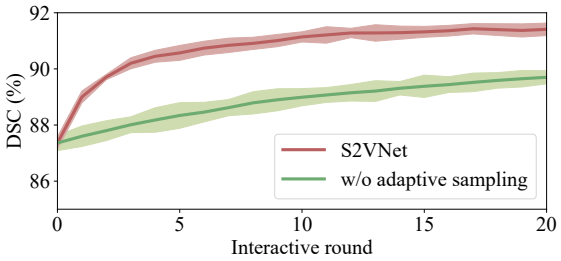

Figure 6.5: Convergence analysis on WORD[129] `test`. We report the DSC score with different round of user interactions.

row #4, our interaction-aware centroid initialization strategy can bring up to **3.51%** performance gains in DSC. With adaptive pixel-feature sampling (*i.e.*, *row* #5) to fuse different rounds of user interactions, the performance further boosts to **90.64%**. Finally, after incorporating recurrent centroid aggregation, S2VNet obtains the best performance on both setups (*i.e.*, *row* #3 and #6), underscoring the general compatibility of this module.

Run-Time Analysis. Next, we probe the running efficiency of S2VNet during inference. Here, 'Baseline' represents a 2D segmentation network without association. As evidenced in Table 6.5, S2VNet achieves nearly **15** times faster inference speed in terms of FPS and saves **48.2%** memory usage compared to the previous state-of-the-art (*i.e.*, SwinMM[194]). Moreover, our association strategy incurs minor additional costs compared to the baseline method while elevating the performance by an impressive **8.69%** in DSC scores. All of

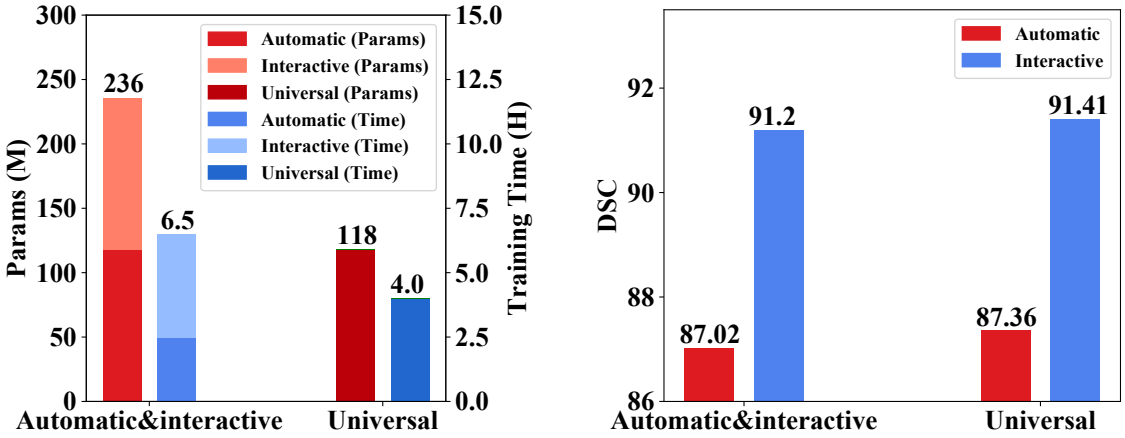

Figure 6.6: Analysis of unified training on WORD[129] `test`.

the above confirms the urgency of shifting the traditional 3D segmentation paradigm to a more efficient one, with S2VNet providing a pragmatic and effective answer.

Convergence Analysis. We study the correlation between the number of refinement rounds and resulting DSC scores on WORD[129] `val`. As seen in Fig. 6.5, the performance of S2VNet exhibits a stable improvement as the rounds of refinement increase, and consistently outperforms the variant without adaptive feature sampling to consider interactions in prior rounds. To strike a balance between accuracy and efficiency, we constrain the average class rounds to 15 from which there is no significant gain in performance.

Unified Training. We provide the network parameters and training time comparison between task-specific models for automatic/interactive segmentation and the universal model in Fig. 6.6. As seen, our universal model requires only half of parameters and training times. Furthermore, the performance under such a unified training paradigm even enjoys improvement compared to task-specific training strategies.

6.4 Conclusion

This chapter presents S2VNet, unifying automatic/interactive medical image segmentation in one system via a slice-to-volume propagation manner. To be specific, S2VNet makes use of clustering-based methods, wherein the knowledge pertaining to targets is compressed within centroids and passed to the next slices to produce coherent and robust predictions with merely 2D segmentation networks. On this basis, S2VNet realizes automatic segmentation via learnable cluster centers while achieving interactive segmentation by initializing the cluster centers with respect to user guidance. This also facilitates concurrent interaction across multiple classes, which overcomes the limitation

of prior work confined to binary setups. Finally, to eliminate the impact of outliers and enhance the awareness of preceding slice cues, a recurrent aggregation approach is proposed to collect historic centroids. All of the above contributes to a flexible solution for volumetric image segmentation characterized by remarkable speed and state-of-the-art accuracy. This chapter presents a flexible medical image segmentation system. This system enables basic predictions from automatic segmentation and manages to achieve real-world accuracy requirements with the help of interactive segmentation. By adopting techniques in §3, §4, and §5, this system can be robust to real-world data and requires fewer annotations.

CHAPTER 7

FUTURE WORK

The thesis tackles major challenges in deploying medical image segmentation in three directions: training with imperfect data, handling imperfect test data, and meeting extremely high accuracy requirements. To be specific, §3 and §4 address the challenge of insufficient annotations through one-shot and barely-supervised segmentation, respectively. §5 focuses on the issue of missing modalities during testing and proposes incomplete-modal medical image segmentation. §6 combines automatic and interactive segmentation to create a system that meets strict accuracy requirements. Together, these four components facilitate the effective deployment of medical image segmentation. However, beyond these above challenges, several additional issues require attention and will form the basis of my feature research.

7.1 Decentralized Data

As detailed in §1.1.2, the decentralization of data storage [134, 137], caused by privacy regulations, impedes the ability of networks to achieve optimal accuracy. This challenge contains several sub-problems that need to be solved. Decentralized data sets are typically biased and unbalanced, influenced by the variable patient conditions at different collection sites. Data calibration may be one potential solution for this problem. Furthermore, the separate annotation processes for decentralized data can lead to label inconsistency. For instance, annotations may be present for one specific anatomical region in datasets from one site but absent in others. For this inconsistency problem, techniques

for partial label learning need to be developed.

7.2 Universal Segmentation

The current landscape of medical image segmentation generally targets specific anatomical regions. This specificity can lead to inefficiencies, particularly in terms of training expenditure and model storage costs. Therefore, there is a clear necessity to develop a universal image segmentation model [15, 118]. Such a model would be capable of processing various anatomical regions across different imaging scanners, potentially improving resource utilization and operational efficiency.

7.3 Inefficient Training and Inference

Existing medical image segmentation methods often rely heavily on 3D networks which suffer from slow inference and present significant challenges in the deployment on hospital devices that usually exhibit limited parallel computation capabilities. As a potential solution, §6 suggests the integration of 2D networks with slice-wise propagation to reduce inference costs. Nonetheless, this method remains constrained by the transformer architecture, which is particularly resource-intensive during training, especially for universal segmentation involving large-scale training images. Considering these constraints, future work may explore more efficient architectures, such as the Mamba framework [35], to reduce both training and inference costs effectively.

BIBLIOGRAPHY

[1] P. ALJABAR, R. A. HECKEMANN, A. HAMMERS, J. V. HAJNAL, AND D. RUECKERT, *Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy*, Neuroimage, 46 (2009), pp. 726–738.

[2] M. ANTONELLI, A. REINKE, S. BAKAS, K. FARAHANI, A. KOPP-SCHNEIDER, B. A. LANDMAN, G. LITJENS, B. MENZE, O. RONNEBERGER, R. M. SUMMERS, ET AL., *The medical segmentation decathlon*, Nature communications, 13 (2022), p. 4128.

[3] M. ASAD, L. FIDON, AND T. VERCAUTEREN, *Econet: Efficient convolutional online likelihood network for scribble-based interactive segmentation*, in International Conference on Medical Imaging with Deep Learning, 2022.

[4] M. ASAD, H. WILLIAMS, I. MANDAL, S. ATHER, J. DEPREST, J. D'HOOGE, AND T. VERCAUTEREN, *Adaptive multi-scale online likelihood network for ai-assisted interactive segmentation*, arXiv preprint arXiv:2303.13696, (2023).

[5] R. AZAD, M. HEIDARI, M. SHARIATNIA, E. K. AGHDAM, S. KARIMIJAFARBIGLOO, E. ADELI, AND D. MERHOF, *Transdeeplab: Convolution-free transformer-based deeplab v3+ for medical image segmentation*, in International Workshop on PRedictive Intelligence In MEdicine, 2022, pp. 91–102.

[6] W. BAE, S. LEE, Y. LEE, B. PARK, M. CHUNG, AND K.-H. JUNG, *Resource optimized neural architecture search for 3d medical image segmentation*, in MICCAI, 2019.

[7] W. BAI, O. OKTAY, M. SINCLAIR, H. SUZUKI, M. RAJCHL, G. TARRONI, B. GLOCKER, A. KING, P. M. MATTHEWS, AND D. RUECKERT, *Semi-supervised learning for network-based cardiac mr image segmentation*, in MICCAI, 2017, pp. 253–260.

[8] G. BALAKRISHNAN, A. ZHAO, M. R. SABUNCU, J. V. GUTTAG, AND A. V. DALCA, *An unsupervised learning model for deformable medical image registration*, in CVPR, 2018, pp. 9252–9260.

[9] S. BERG, D. KUTRA, T. KROEGER, C. N. STRAEHLE, B. X. KAUSLER, C. HAUBOLD, M. SCHIEGG, J. ALES, T. BEIER, M. RUDY, ET AL., *Ilastik: interactive machine learning for (bio) image analysis*, Nature Methods, 16 (2019), pp. 1226–1232.

[10] G. BORTSOVA, F. DUBOST, L. HOGEWEG, I. KATRAMADOS, AND M. DE BRUIJNE, *Semi-supervised medical image segmentation via learning consistency under transformations*, in MICCAI, 2019, pp. 810–818.

[11] Y. Y. BOYKOV AND M.-P. JOLLY, *Interactive graph cuts for optimal boundary & region segmentation of objects in nd images*, in ICCV, 2001.

[12] S. BUDD, E. C. ROBINSON, AND B. KAINZ, *A survey on active learning and human-in-the-loop deep learning for medical image analysis*, Medical Image Analysis, 71 (2021), p. 102062.

[13] C. P. BURGESS, I. HIGGINS, A. PAL, L. MATTHEY, N. WATTERS, G. DESJARDINS, AND A. LERCHNER, *Understanding disentangling in β-vae*, Arxiv, (2018).

[14] S. A. BURNEY AND H. TARIQ, *K-means cluster analysis for image segmentation*, International Journal of Computer Applications, 96 (2014).

[15] V. I. BUTOI, J. J. G. ORTIZ, T. MA, M. R. SABUNCU, J. GUTTAG, AND A. V. DALCA, *Universeg: Universal medical image segmentation*, in ICCV, 2023, pp. 21438–21451.

[16] M. CABEZAS, A. OLIVER, X. LLADÓ, J. FREIXENET, AND M. B. CUADRA, *A review of atlas-based segmentation for magnetic resonance brain images*, Computer Methods and Programs in Biomedicine, 104 (2011), pp. e158–e177.

[17] H. CAI, S. LI, L. QI, Q. YU, Y. SHI, AND Y. GAO, *Orthogonal annotation benefits barely-supervised medical image segmentation*, in CVPR, 2023, pp. 3302–3311.

[18] H. CAO, Y. WANG, J. CHEN, D. JIANG, X. ZHANG, Q. TIAN, AND M. WANG, *Swin-unet: Unet-like pure transformer for medical image segmentation*, in ECCV, 2022.

[19] X. CAO, H. CHEN, Y. LI, Y. PENG, S. WANG, AND L. CHENG, *Uncertainty aware temporal-ensembling model for semi-supervised abus mass segmentation*, IEEE TMI, 40 (2020), pp. 431–443.

[20] N. CARION, F. MASSA, G. SYNNAEVE, N. USUNIER, A. KIRILLOV, AND S. ZAGORUYKO, *End-to-end object detection with transformers*, in ECCV, 2020.

[21] K. CHAITANYA, N. KARANI, C. F. BAUMGARTNER, A. S. BECKER, O. DONATI, AND E. KONUKOGLU, *Semi-supervised and task-driven data augmentation*, in IPMI, 2019, pp. 29–41.

[22] A. CHARTSIAS, T. JOYCE, G. PAPANASTASIOU, S. SEMPLE, M. WILLIAMS, D. NEWBY, R. DHARMAKUMAR, AND S. A. TSAFTARIS, *Factorised spatial representation learning: Application in semi-supervised myocardial segmentation*, in MICCAI, 2018, pp. 490–498.

[23] C. CHEN, Q. DOU, Y. JIN, H. CHEN, J. QIN, AND P.-A. HENG, *Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion*, in MICCAI, 2019, pp. 447–456.

[24] C. CHEN, X. LIU, M. DING, J. ZHENG, AND J. LI, *3d dilated multi-fiber network for real-time brain tumor segmentation in mri*, in International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 184–192.

[25] J. CHEN, Y. LU, Q. YU, X. LUO, E. ADELI, Y. WANG, L. LU, A. L. YUILLE, AND Y. ZHOU, *Transunet: Transformers make strong encoders for medical image segmentation*, arXiv preprint arXiv:2102.04306, (2021).

[26] S. CHEN, G. BORTSOVA, A. G.-U. JUÁREZ, G. VAN TULDER, AND M. DE BRUIJNE, *Multi-task attention-based semi-supervised learning for medical image segmentation*, in MICCAI, 2019, pp. 457–465.

[27] W. CHEN, R. SMITH, S.-Y. JI, K. R. WARD, AND K. NAJARIAN, *Automated ventricular systems segmentation in brain ct images by combining low-level segmentation and high-level template matching*, BMC medical informatics and decision making, 9 (2009), pp. 1–14.

[28] X. CHEN, B. M. WILLIAMS, S. R. VALLABHANENI, G. CZANNER, R. WILLIAMS, AND Y. ZHENG, *Learning active contour models for medical image segmentation*, in CVPR, 2019, pp. 11632–11640.

[29] X. CHEN, Y. YUAN, G. ZENG, AND J. WANG, *Semi-supervised semantic segmentation with cross pseudo supervision*, in CVPR, 2021, pp. 2613–2622.

[30] B. CHENG, I. MISRA, A. G. SCHWING, A. KIRILLOV, AND R. GIRDHAR, *Masked-attention mask transformer for universal image segmentation*, in CVPR, 2022.

[31] B. CHENG, A. SCHWING, AND A. KIRILLOV, *Per-pixel classification is not all you need for semantic segmentation*, in NuerIPS, 2021.

[32] V. CHERUKURI, P. SSENYONGA, B. C. WARF, A. V. KULKARNI, V. MONGA, AND S. J. SCHIFF, *Learning based segmentation of ct brain images: application to postoperative hydrocephalic scans*, IEEE Transactions on Biomedical Engineering, 65 (2017), pp. 1871–1884.

[33] Ö. ÇIÇEK, A. ABDULKADIR, S. S. LIENKAMP, T. BROX, AND O. RONNEBERGER, *3d u-net: learning dense volumetric segmentation from sparse annotation*, in International Conference on Medical Image Computing and Computer-assisted Intervention, 2016, pp. 424–432.

[34] W. CUI, Y. LIU, Y. LI, M. GUO, Y. LI, X. LI, T. WANG, X. ZENG, AND C. YE, *Semi-supervised brain lesion segmentation with an adapted mean teacher model*, in International Conference on Information Processing in Medical Imaging, 2019, pp. 554–565.

[35] T. DAO AND A. GU, *Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality*, in ICML, 2024.

[36] L. DAZA, J. C. PÉREZ, AND P. ARBELÁEZ, *Towards robust general medical image segmentation*, in MICCAI, 2021, pp. 3–13.

[37] J. DENG, W. DONG, R. SOCHER, L.-J. LI, K. LI, AND L. FEI-FEI, *Imagenet: A large-scale hierarchical image database*, in CVPR, 2009, pp. 248–255.

[38] A. DIAZ-PINTO, S. ALLE, V. NATH, Y. TANG, A. IHSANI, M. ASAD, F. PÉREZ-GARCÍA, P. MEHTA, W. LI, M. FLORES, ET AL., *Monai label: A framework for ai-assisted interactive labeling of 3d medical images*, arXiv preprint arXiv:2203.12362, (2022).

[39] L. R. DICE, *Measures of the amount of ecologic association between species*, Ecology, 26 (1945), pp. 297–302.

[40] F. DING, G. YANG, J. LIU, J. WU, D. DING, J. XV, G. CHENG, AND X. LI, *Hierarchical attention networks for medical image segmentation*, arXiv preprint arXiv:1911.08777, (2019).

[41] Y. DING, L. LI, W. WANG, AND Y. YANG, *S2vnet: universal multi-class medical image segmentation via clustering-based slice-to-volume propagation*, in CVPR, 2024.

[42] Y. DING AND H. LIU, *Barely-supervised brain tumor segmentation via employing segment anything model*, IEEE TCSVT, (2024).

[43] Y. DING, X. YU, AND Y. YANG, *Modeling the probabilistic distribution of unlabeled data for one-shot medical image segmentation*, in AAAI, vol. 35, 2021, pp. 1246–1254.

[44] Z. DING, X. HAN, AND M. NIETHAMMER, *Votenet: A deep learning label fusion method for multi-atlas segmentation*, in MICCAI, 2019, pp. 202–210.

[45] DING, YUHANG AND YU, XIN AND YANG, YI, *Rfnet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation*, in ICCV, 2021, pp. 3975–3984.

[46] N. K. DINSDALE, M. JENKINSON, AND A. I. NAMBURETE, *Spatial warping network for 3d segmentation of the hippocampus in mr images*, in MICCAI, 2019, pp. 284–291.

[47] J. DOLZ, K. GOPINATH, J. YUAN, H. LOMBAERT, C. DESROSIERS, AND I. B. AYED, *Hyperdense-net: a hyper-densely connected cnn for multi-modal image segmentation*, IEEE Transactions on Medical Imaging, 38 (2018), pp. 1116–1126.

[48] R. DORENT, S. JOUTARD, M. MODAT, S. OURSELIN, AND T. VERCAUTEREN, *Hetero-modal variational encoder-decoder for joint modality completion and segmentation*, in MICCAI, 2019, pp. 74–82.

[49] Q. DOU, H. CHEN, Y. JIN, L. YU, J. QIN, AND P.-A. HENG, *3d deeply supervised network for automatic liver segmentation from ct volumes*, in MICCAI, 2016, pp. 149–157.

[50] P. DVOŘÁK AND B. MENZE, *Local structure prediction with convolutional neural networks for multimodal brain tumor segmentation*, in International MICCAI workshop on Medical Computer Vision, 2015, pp. 59–71.

[51] K. FANG AND W.-J. LI, *Dmnet: Difference minimization network for semi-supervised segmentation in medical images*, in MICCAI, 2020, pp. 532–541.

[52] R. FENG, X. ZHENG, T. GAO, J. CHEN, W. WANG, D. Z. CHEN, AND J. WU, *Interactive few-shot learning: Limited supervision, better medical image segmentation*, IEEE TMI, 40 (2021), pp. 2575–2588.

[53] L. FIDON, W. LI, L. C. GARCIA-PERAZA-HERRERA, J. EKANAYAKE, N. KITCHEN, S. OURSELIN, AND T. VERCAUTEREN, *Scalable multimodal convolutional networks for brain tumour segmentation*, in International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2017, pp. 285–293.

[54] B. FISCHL, *Freesurfer*, Neuroimage, 62 (2012), pp. 774–781.

[55] G. FOTEDAR, N. TAJBAKHSH, S. ANANTH, AND X. DING, *Extreme consistency: Overcoming annotation scarcity and domain shifts*, in MICCAI, 2020, pp. 699–709.

[56] Y. GAO, M. ZHOU, AND D. N. METAXAS, *Utnet: a hybrid transformer architecture for medical image segmentation*, in MICCAI, 2021.

[57] K. B. GIRUM, G. CRÉHANGE, AND A. LALANDE, *Learning with context feedback loop for robust medical image segmentation*, IEEE TMI, 40 (2021), pp. 1542–1554.

[58] R. GU, G. WANG, T. SONG, R. HUANG, M. AERTSEN, J. DEPREST, S. OURSELIN, T. VERCAUTEREN, AND S. ZHANG, *Ca-net: Comprehensive attention convolutional neural networks for explainable medical image segmentation*, IEEE TMI, 40 (2020), pp. 699–711.

[59] Z. GU, J. CHENG, H. FU, K. ZHOU, H. HAO, Y. ZHAO, T. ZHANG, S. GAO, AND J. LIU, *Ce-net: Context encoder network for 2d medical image segmentation*, IEEE TMI, 38 (2019), pp. 2281–2292.

[60] GUHA BALAKRISHNAN AND AMY ZHAO AND MERT R. SABUNCU AND JOHN V. GUTTAG AND ADRIAN V. DALCA, *Voxelmorph: A learning framework for deformable medical image registration*, IEEE TMI, 38 (2019), pp. 1788–1800.

[61] Y. GUO, L. BI, E. AHN, D. FENG, Q. WANG, AND J. KIM, *A spatiotemporal volumetric interpolation network for 4d dynamic medical image*, in CVPR, 2020.

[62] W. HANG, W. FENG, S. LIANG, L. YU, Q. WANG, K.-S. CHOI, AND J. QIN, *Local and global structure-aware entropy regularized mean teacher model for 3d left atrium segmentation*, in MICCAI, 2020, pp. 562–571.

[63] A. HATAMIZADEH, V. NATH, Y. TANG, D. YANG, H. R. ROTH, AND D. XU, *Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images*, in International MICCAI Brainlesion Workshop, 2021.

[64] A. HATAMIZADEH, Y. TANG, V. NATH, D. YANG, A. MYRONENKO, B. LANDMAN, H. R. ROTH, AND D. XU, *Unetr: Transformers for 3d medical image segmentation*, in WACV, 2022.

[65] M. HAVAEI, A. DAVY, D. WARDE-FARLEY, A. BIARD, A. COURVILLE, Y. BENGIO, C. PAL, P.-M. JODOIN, AND H. LAROCHELLE, *Brain tumor segmentation with deep neural networks*, Medical Image Analysis, 35 (2017), pp. 18–31.

[66] M. HAVAEI, N. GUIZARD, N. CHAPADOS, AND Y. BENGIO, *Hemis: Hetero-modal image segmentation*, in International Conference on Medical Image Computing and Computer-Assisted Intervention, 2016, pp. 469–477.

[67] K. HE, X. ZHANG, S. REN, AND J. SUN, *Deep residual learning for image recognition*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[68] S. HE, Y. FENG, P. E. GRANT, AND Y. OU, *Segmentation ability map: Interpret deep features for medical image segmentation*, Medical Image Analysis, 84 (2023), p. 102726.

[69] Y. HE, D. YANG, H. ROTH, C. ZHAO, AND D. XU, *Dints: Differentiable neural network topology search for 3d medical image segmentation*, in CVPR, 2021.

[70] Y. He, G. Yang, Y. Chen, Y. Kong, J. Wu, L. Tang, X. Zhu, J.-L. Dillenseger, P. Shao, S. Zhang, et al., *Dpa-densebiasnet: Semi-supervised 3d fine renal artery segmentation with dense biased network and deep priori anatomy*, in MICCAI, 2019, pp. 139–147.

[71] R. A. Heckemann, J. V. Hajnal, P. Aljabar, D. Rueckert, and A. Hammers, *Automatic anatomical brain MRI segmentation combining label propagation and decision fusion*, NeuroImage, 33 (2006), pp. 115–126.

[72] M. Heidari, A. Kazerouni, M. Soltany, R. Azad, E. K. Aghdam, J. Cohen-Adad, and D. Merhof, *Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation*, in WACV, 2023, pp. 6202–6212.

[73] K. Held, E. R. Kops, B. J. Krause, W. M. Wells, R. Kikinis, and H.-W. Muller-Gartner, *Markov random field segmentation of brain mr images*, IEEE TMI, 16 (1997), pp. 878–886.

[74] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, *beta-vae: Learning basic visual concepts with a constrained variational framework*, in ICLR, 2017.

[75] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, *Unet 3+: A full-scale connected unet for medical image segmentation*, in IEEE ICASSP, 2020.

[76] X. Huang, Z. Deng, D. Li, X. Yuan, and Y. Fu, *Missformer: an effective transformer for 2d medical image segmentation*, IEEE TMI, (2022).

[77] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, *Comparing images using the hausdorff distance*, IEEE TPAMI, 15 (1993), pp. 850–863.

[78] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, *nnu-net: a self-configuring method for deep learning-based biomedical image segmentation*, Nature Methods, (2020), pp. 1–9.

[79] Isensee, Fabian and Jaeger, Paul F and Kohl, Simon AA and Petersen, Jens and Maier-Hein, Klaus H, *nnu-net: a self-configuring method for deep learning-based biomedical image segmentation*, Nature Methods, 18 (2021), pp. 203–211.

[80] Y. JI, H. BAI, C. GE, J. YANG, Y. ZHU, R. ZHANG, Z. LI, L. ZHANNG, W. MA, X. WAN, ET AL., *Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation*, in NuerIPS, 2022.

[81] Y. JI, R. ZHANG, Z. LI, J. REN, S. ZHANG, AND P. LUO, *Uxnet: Searching multi-level feature aggregation for 3d medical image segmentation*, in MICCAI, 2020.

[82] G. A. KAISSIS, M. R. MAKOWSKI, D. RÜCKERT, AND R. F. BRAREN, *Secure, privacy-preserving and federated machine learning in medical imaging*, Nature Machine Intelligence, 2 (2020), pp. 305–311.

[83] K. KAMNITSAS, W. BAI, E. FERRANTE, S. MCDONAGH, M. SINCLAIR, N. PAWLOWSKI, M. RAJCHL, M. LEE, B. KAINZ, D. RUECKERT, ET AL., *Ensembles of multiple models and architectures for robust brain tumour segmentation*, in International MICCAI Brainlesion Workshop, 2017, pp. 450–462.

[84] K. KAMNITSAS, C. LEDIG, V. F. NEWCOMBE, J. P. SIMPSON, A. D. KANE, D. K. MENON, D. RUECKERT, AND B. GLOCKER, *Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation*, Medical Image Analysis, 36 (2017), pp. 61–78.

[85] D. KARIMI, S. D. VASYLECHKO, AND A. GHOLIPOUR, *Convolution-free medical image segmentation using transformers*, in MICCAI, 2021.

[86] D. N. KENNEDY, C. HASELGROVE, S. M. HODGE, P. S. RANE, N. MAKRIS, AND J. A. FRAZIER, *Candishare: a resource for pediatric neuroimaging data*, Neuroinformatics, 10 (2011), p. 319,Äì322.

[87] H. KERVADEC, J. DOLZ, E. GRANGER, AND I. B. AYED, *Curriculum semi-supervised segmentation*, in MICCAI, 2019, pp. 568–576.

[88] S. KHAN, A. H. SHAHIN, J. VILLAFRUELA, J. SHEN, AND L. SHAO, *Extreme points derived confidence map as a cue for class-agnostic interactive segmentation using deep neural network*, in MICCAI, 2019.

[89] S. KIM, I. KIM, S. LIM, W. BAEK, C. KIM, H. CHO, B. YOON, AND T. KIM, *Scalable neural architecture search for 3d medical image segmentation*, in MICCAI, 2019.

[90] D. P. KINGMA AND J. BA, *Adam: A method for stochastic optimization*, International Conference on Learning Representations, (2014).

[91] D. P. KINGMA AND J. BA, *Adam: A method for stochastic optimization*, in ICLR, 2015.

[92] A. KIRILLOV, E. MINTUN, N. RAVI, H. MAO, C. ROLLAND, L. GUSTAFSON, T. XIAO, S. WHITEHEAD, A. C. BERG, W.-Y. LO, P. DOLLÁR, AND R. GIRSHICK, *Segment anything*, arXiv:2304.02643, (2023).

[93] A. KLEIN, B. MENSH, S. S. GHOSH, J. A. TOURVILLE, AND J. HIRSCH, *Mindboggle: Automated brain labeling with multiple atlases*, BMC Medical Imaging, 5 (2005), p. 7.

[94] N. A. KOOHBANANI, M. JAHANIFAR, N. Z. TAJADIN, AND N. RAJPOOT, *Nuclick: a deep learning framework for interactive segmentation of microscopic images*, Medical Image Analysis, 65 (2020), p. 101771.

[95] M. LALONDE, M. BEAULIEU, AND L. GAGNON, *Fast and robust optic disc detection using pyramidal decomposition and hausdorff-based template matching*, IEEE TMI, 20 (2001), pp. 1193–1200.

[96] R. LALONDE, Z. XU, I. IRMAKCI, S. JAIN, AND U. BAGCI, *Capsules for biomedical image segmentation*, Medical Image Analysis, 68 (2021), p. 101889.

[97] B. LANDMAN, Z. XU, J. IGELSIAS, M. STYNER, T. LANGERAK, AND A. KLEIN, *Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge*, in MICCAI Multi-Atlas Labeling Beyond Cranial Vault‚ÄîWorkshop Challenge, 2015.

[98] A. J. LARRAZABAL, C. MARTÍNEZ, J. DOLZ, AND E. FERRANTE, *Orthogonal ensemble networks for biomedical image segmentation*, in MICCAI, 2021.

[99] C.-Y. LEE, S. XIE, P. GALLAGHER, Z. ZHANG, AND Z. TU, *Deeply-supervised nets*, in Artificial Intelligence and Statistics, 2015, pp. 562–570.

[100] H. LEE AND W.-K. JEONG, *Scribble2label: Scribble-supervised cell segmentation via self-generating pseudo-labels with consistency*, in MICCAI, 2020.

[101] H. H. LEE, S. BAO, Y. HUO, AND B. A. LANDMAN, *3d ux-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation*, in ICLR, 2023.

[102] H. J. LEE, J. U. KIM, S. LEE, H. G. KIM, AND Y. M. RO, *Structure boundary preserving segmentation for medical image with ambiguous boundary*, in CVPR, 2020.

[103] S. LI, H. CAI, L. QI, Q. YU, Y. SHI, AND Y. GAO, *Pln: Parasitic-like network for barely supervised medical image segmentation*, IEEE TMI, 42 (2022), pp. 582–593.

[104] S. LI, T. FEVENS, AND A. KRZYŻAK, *A svm-based framework for autonomous volumetric medical image segmentation using hierarchical and coupled level sets*, in International Congress Series, vol. 1268, Elsevier, 2004, pp. 207–212.

[105] S. LI, X. SUI, X. LUO, X. XU, Y. LIU, AND R. GOH, *Medical image segmentation using squeeze-and-expansion transformers*, arXiv preprint arXiv:2105.09511, (2021).

[106] S. LI, C. ZHANG, AND X. HE, *Shape-aware semi-supervised 3d semantic segmentation for medical images*, in MICCAI, 2020, pp. 552–561.

[107] W. LI ET AL., *Automatic segmentation of liver tumor in ct images with deep convolutional neural networks*, Journal of Computer and Communications, 3 (2015), p. 146.

[108] X. LI, M. XIA, J. JIAO, S. ZHOU, C. CHANG, Y. WANG, AND Y. GUO, *Hal-ia: A hybrid active learning framework using interactive annotation for medical image segmentation*, Medical Image Analysis, (2023), p. 102862.

[109] X. LI, L. YU, H. CHEN, C.-W. FU, L. XING, AND P.-A. HENG, *Transformation-consistent self-ensembling model for semisupervised medical image segmentation*, IEEE Transactions on Neural Networks and Learning Systems, 32 (2020), pp. 523–534.

[110] Y. LI, J. CHEN, X. XIE, K. MA, AND Y. ZHENG, *Self-loop uncertainty: A novel pseudo-label for semi-supervised medical image segmentation*, in MICCAI, 2020, pp. 614–623.

[111] C. LIANG, W. WANG, J. MIAO, AND Y. YANG, *Logic-induced diagnostic reasoning for semi-supervised semantic segmentation*, in ICCV, 2023, pp. 16197–16208.

[112] J. LIANG, T. ZHOU, D. LIU, AND W. WANG, *Clustseg: Clustering for universal segmentation*, in ICLR, 2023.

[113] X. LIAO, W. LI, Q. XU, X. WANG, B. JIN, X. ZHANG, Y. WANG, AND Y. ZHANG, *Iteratively-refined interactive 3d medical image segmentation with multi-agent reinforcement learning*, in CVPR, 2020.

[114] A. LIN, B. CHEN, J. XU, Z. ZHANG, G. LU, AND D. ZHANG, *Ds-transunet: Dual swin transformer u-net for medical image segmentation*, IEEE Transactions on Instrumentation and Measurement, 71 (2022), pp. 1–15.

[115] J. LIN, J. LIN, C. LU, H. CHEN, H. LIN, B. ZHAO, Z. SHI, B. QIU, X. PAN, Z. XU, ET AL., *Ckd-transbts: clinical knowledge-driven hybrid transformer with modality-correlated cross-attention for brain tumor segmentation*, IEEE TMI, (2023).

[116] Y. LIN, H. YAO, Z. LI, G. ZHENG, AND X. LI, *Calibrating label distribution for class-imbalanced barely-supervised knee segmentation*, in MICCAI, Springer, 2022, pp. 109–118.

[117] D. LIU, Y. GAO, Q. ZHANGLI, L. HAN, X. HE, Z. XIA, S. WEN, Q. CHANG, Z. YAN, M. ZHOU, ET AL., *Transfusion: multi-view divergent fusion for medical image segmentation with transformers*, in MICCAI, 2022.

[118] J. LIU, Y. ZHANG, J.-N. CHEN, J. XIAO, Y. LU, B. A LANDMAN, Y. YUAN, A. YUILLE, Y. TANG, AND Z. ZHOU, *Clip-driven universal model for organ segmentation and tumor detection*, in ICCV, 2023, pp. 21152–21164.

[119] Q. LIU, Z. XU, Y. JIAO, AND M. NIETHAMMER, *isegformer: Interactive segmentation via transformers with application to 3d knee mr images*, in MICCAI, 2022.

[120] Q. LIU, M. ZHENG, B. PLANCHE, Z. GAO, T. CHEN, M. NIETHAMMER, AND Z. WU, *Exploring cycle consistency learning in interactive volume segmentation*, arXiv preprint arXiv:2303.06493, (2023).

[121] W. LIU, C. MA, Y. YANG, W. XIE, AND Y. ZHANG, *Transforming the interactive segmentation for medical imaging*, in MICCAI, 2022.

[122] W. LIU, T. TIAN, W. XU, H. YANG, X. PAN, S. YAN, AND L. WANG, *Phtrans: Parallelly aggregating global and local representations for medical image segmentation*, in MICCAI, 2022.

[123] Y. LIU, L. FAN, C. ZHANG, T. ZHOU, Z. XIAO, L. GENG, AND D. SHEN, *Incomplete multi-modal representation learning for alzheimer‚Äôs disease diagnosis*, Medical Image Analysis, 69 (2021), p. 101953.

[124] Z. LIU, Y. LIN, Y. CAO, H. HU, Y. WEI, Z. ZHANG, S. LIN, AND B. GUO, *Swin transformer: Hierarchical vision transformer using shifted windows*, in ICCV, 2021.

[125] S. LLOYD, *Least squares quantization in pcm*, IEEE Transactions on Information Theory, 28 (1982), pp. 129–137.

[126] J. LONG, E. SHELHAMER, AND T. DARRELL, *Fully convolutional networks for semantic segmentation*, in CVPR, 2015, pp. 3431–3440.

[127] X. LUO, J. CHEN, T. SONG, AND G. WANG, *Semi-supervised medical image segmentation through dual-task consistency*, in AAAI, vol. 35, 2021, pp. 8801–8809.

[128] X. LUO, W. LIAO, J. CHEN, T. SONG, Y. CHEN, S. ZHANG, N. CHEN, G. WANG, AND S. ZHANG, *Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency*, in MICCAI, 2021, pp. 318–329.

[129] X. LUO, W. LIAO, J. XIAO, J. CHEN, T. SONG, X. ZHANG, K. LI, D. N. METAXAS, G. WANG, AND S. ZHANG, *Word: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image*, Medical Image Analysis, 82 (2022), p. 102642.

[130] X. LUO, G. WANG, T. SONG, J. ZHANG, M. AERTSEN, J. DEPREST, S. OURSELIN, T. VERCAUTEREN, AND S. ZHANG, *Mideepseg: Minimally interactive segmentation of unseen objects from medical images using deep learning*, Medical Image Analysis, 72 (2021), p. 102102.

[131] LUO, XIANGDE AND CHEN, JIENENG AND SONG, TAO AND WANG, GUOTAI, *Semi-supervised medical image segmentation through dual-task consistency*, in AAAI, vol. 35, 2021, pp. 8801–8809.

[132] C. MA, Q. XU, X. WANG, B. JIN, X. ZHANG, Y. WANG, AND Y. ZHANG, *Boundary-aware supervoxel-level iteratively refined interactive 3d image segmentation with multi-agent reinforcement learning*, IEEE TMI, 40 (2020), pp. 2563–2574.

[133] Z. MARINOV, R. STIEFELHAGEN, AND J. KLEESIEK, *Guiding the guidance: A comparative analysis of user guidance signals for interactive segmentation of volumetric images*, arXiv preprint arXiv:2303.06942, (2023).

[134] B. MCMAHAN, E. MOORE, D. RAMAGE, S. HAMPSON, AND B. A. Y ARCAS, *Communication-efficient learning of deep networks from decentralized data*, in Artificial intelligence and statistics, PMLR, 2017, pp. 1273–1282.

[135] S. MEHTA, M. RASTEGARI, A. CASPI, L. SHAPIRO, AND H. HAJISHIRZI, *Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation*, in ECCV, 2018.

[136] B. H. MENZE, A. JAKAB, S. BAUER, J. KALPATHY-CRAMER, K. FARAHANI, J. KIRBY, Y. BURREN, N. PORZ, J. SLOTBOOM, R. WIEST, ET AL., *The multimodal brain tumor image segmentation benchmark (brats)*, IEEE TMI, 34 (2014), pp. 1993–2024.

[137] J. MIAO, Z. YANG, L. FAN, AND Y. YANG, *Fedseg: Class-heterogeneous federated learning for semantic segmentation*, in CVPR, 2023, pp. 8042–8052.

[138] S. MIAO, S. PIAT, P. W. FISCHER, A. TUYSUZOGLU, P. W. MEWES, T. MANSI, AND R. LIAO, *Dilated FCN for multi-agent 2d/3d medical image registration*, in AAAI, 2018, pp. 4694–4701.

[139] F. MILLETARI, N. NAVAB, AND S. AHMADI, *V-net: Fully convolutional neural networks for volumetric medical image segmentation*, in 3DV, 2016, pp. 565–571.

[140] F. MILLETARI, N. NAVAB, AND S.-A. AHMADI, *V-net: Fully convolutional neural networks for volumetric medical image segmentation*, in 3DV, IEEE, 2016, pp. 565–571.

[141] A. MYRONENKO, *3d mri brain tumor segmentation using autoencoder regularization*, in International MICCAI Brainlesion Workshop, 2018, pp. 311–320.

[142] D. NEVEN, B. D. BRABANDERE, M. PROESMANS, AND L. V. GOOL, *Instance segmentation by jointly optimizing spatial embeddings and clustering bandwidth*, in CVPR, 2019.

[143] X. NIE, L. LIU, L. HE, L. ZHAO, H. LU, S. LOU, R. XIONG, AND Y. WANG, *Weakly-interactive-mixed learning: Less labelling cost for better medical image segmentation*, IEEE Journal of Biomedical and Health Informatics, (2023).

[144] O. OKTAY, J. SCHLEMPER, L. L. FOLGOC, M. LEE, M. HEINRICH, K. MISAWA, K. MORI, S. MCDONAGH, N. Y. HAMMERLA, B. KAINZ, ET AL., *Attention u-net: Learning where to look for the pancreas. arxiv 2018*, arXiv preprint arXiv:1804.03999, (2018).

[145] Y. OU, Y. YUAN, X. HUANG, S. T. WONG, J. VOLPI, J. Z. WANG, AND K. WONG, *Patcher: Patch transformers with mixture of experts for precise medical image segmentation*, in MICCAI, 2022.

[146] H. PEIRIS, Z. CHEN, G. EGAN, AND M. HARANDI, *Duo-segnet: Adversarial dual-views for semi-supervised medical image segmentation*, in MICCAI, 2021, pp. 428–438.

[147] C. PENG, A. MYRONENKO, A. HATAMIZADEH, V. NATH, M. M. R. SIDDIQUEE, Y. HE, D. XU, R. CHELLAPPA, AND D. YANG, *Hypersegnas: Bridging one-shot neural architecture search with 3d medical image segmentation using hypernet*, in CVPR, 2022.

[148] S. PEREIRA, A. PINTO, V. ALVES, AND C. A. SILVA, *Brain tumor segmentation using convolutional neural networks in MRI images*, TMI, 35 (2016), pp. 1240–1251.

[149] D. L. PHAM, C. XU, AND J. L. PRINCE, *Current methods in medical image segmentation*, Annual Review of Biomedical Engineering, 2 (2000), pp. 315–337.

[150] M. PRASTAWA, E. BULLITT, S. HO, AND G. GERIG, *A brain tumor segmentation framework based on outlier detection*, Medical Image Analysis, 8 (2004), pp. 275–283.

[151] M. RAJCHL, M. C. LEE, O. OKTAY, K. KAMNITSAS, J. PASSERAT-PALMBACH, W. BAI, M. DAMODARAM, M. A. RUTHERFORD, J. V. HAJNAL, B. KAINZ, ET AL., *Deepcut: Object segmentation from bounding box annotations using convolutional neural networks*, IEEE TMI, 36 (2016), pp. 674–683.

[152] N. RIEKE, J. HANCOX, W. LI, F. MILLETARI, H. R. ROTH, S. ALBARQOUNI, S. BAKAS, M. N. GALTIER, B. A. LANDMAN, K. MAIER-HEIN, ET AL., *The*

future of digital health with federated learning, NPJ digital medicine, 3 (2020), pp. 1–7.

[153] O. RONNEBERGER, P. FISCHER, AND T. BROX, *U-net: Convolutional networks for biomedical image segmentation*, in MICCAI, 2015, pp. 234–241.

[154] O. RONNEBERGER, P. FISCHER, AND T. BROX, *U-net: Convolutional networks for biomedical image segmentation*, in MICCAI, 2015, pp. 234–241.

[155] H. R. ROTH, L. LU, A. FARAG, H. SHIN, J. LIU, E. B. TURKBEY, AND R. M. SUMMERS, *Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation*, in MICCAI, 2015, pp. 556–564.

[156] S. S. M. SALEHI, D. ERDOGMUS, AND A. GHOLIPOUR, *Tversky loss function for image segmentation using 3d fully convolutional deep networks*, in International workshop on machine learning in medical imaging, Springer, 2017, pp. 379–387.

[157] B. SAMBATURU, A. GUPTA, C. JAWAHAR, AND C. ARORA, *Efficient and generic interactive segmentation framework to correct mispredictions during clinical evaluation of medical images*, in MICCAI, 2021.

[158] S. SEDAI, B. ANTONY, R. RAI, K. JONES, H. ISHIKAWA, J. SCHUMAN, W. GADI, AND R. GARNAVI, *Uncertainty guided semi-supervised segmentation of retinal layers in oct images*, in MICCAI, 2019, pp. 282–290.

[159] S. SEDAI, D. MAHAPATRA, S. HEWAVITHARANAGE, S. MAETSCHKE, AND R. GARNAVI, *Semi-supervised segmentation of optic cup in retinal fundus images using variational autoencoder*, in MICCAI, 2017, pp. 75–82.

[160] H. SHEN, R. WANG, J. ZHANG, AND S. J. MCKENNA, *Boundary-aware fully convolutional network for brain tumor segmentation*, in International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2017, pp. 433–441.

[161] Y. SHEN AND M. GAO, *Brain tumor segmentation on mri with missing modalities*, in International Conference on Information Processing in Medical Imaging, 2019, pp. 417–428.

[162] L. SHI, X. ZHANG, Y. LIU, AND X. HAN, *A hybrid propagation network for interactive volumetric image segmentation*, in MICCAI, 2022.

[163] A. SRIVASTAVA, D. JHA, S. CHANDA, U. PAL, H. D. JOHANSEN, D. JOHANSEN, M. A. RIEGLER, S. ALI, AND P. HALVORSEN, *Msrf-net: a multi-scale residual fusion network for biomedical image segmentation*, IEEE Journal of Biomedical and Health Informatics, 26 (2021), pp. 2252–2263.

[164] J. STAAL, M. D. ABRÀMOFF, M. NIEMEIJER, M. A. VIERGEVER, AND B. VAN GINNEKEN, *Ridge-based vessel segmentation in color images of the retina*, IEEE TMI, 23 (2004), pp. 501–509.

[165] C. H. SUDRE, W. LI, T. VERCAUTEREN, S. OURSELIN, AND M. JORGE CARDOSO, *Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations*, in Deep learning in medical image analysis and multimodal learning for clinical decision support, Springer, 2017, pp. 240–248.

[166] K. TA, S. S. AHN, J. C. STENDAHL, A. J. SINUSAS, AND J. S. DUNCAN, *A semi-supervised joint network for simultaneous left ventricular motion tracking and segmentation in 4d echocardiography*, in MICCAI, 2020, pp. 468–477.

[167] A. TARVAINEN AND H. VALPOLA, *Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results*, in NeurIPS, 2017, pp. 1195–1204.

[168] L. TRAN, X. LIU, J. ZHOU, AND R. JIN, *Missing modalities imputation via cascaded residual autoencoder*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1405–1414.

[169] P. V. TRAN, *A fully convolutional neural network for cardiac segmentation in short-axis mri*, arXiv preprint arXiv:1604.00494, (2016).

[170] A. TSAI, A. YEZZI, W. WELLS, C. TEMPANY, D. TUCKER, A. FAN, W. E. GRIMSON, AND A. WILLSKY, *A shape-based approach to the segmentation of medical imagery using level sets*, IEEE TMI, 22 (2003), pp. 137–154.

[171] K.-L. TSENG, Y.-L. LIN, W. HSU, AND C.-Y. HUANG, *Joint sequence learning and cross-modality convolution for 3d biomedical segmentation*, in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2017, pp. 6393–6400.

[172] J. VALANARASU AND V. PATEL, *Unext: Mlp-based rapid medical image segmentation network. arxiv 2022*, arXiv preprint arXiv:2203.04967.

[173] J. M. J. VALANARASU, P. OZA, I. HACIHALILOGLU, AND V. M. PATEL, *Medical transformer: Gated axial-attention for medical image segmentation*, in MICCAI, 2021, pp. 36–46.

[174] J. M. J. VALANARASU, V. A. SINDAGI, I. HACIHALILOGLU, AND V. M. PATEL, *Kiu-net: Towards accurate segmentation of biomedical images using over-complete representations*, in MICCAI, 2020.

[175] ——, *Kiu-net: Overcomplete convolutional architectures for biomedical image and volumetric segmentation*, IEEE TMI, 41 (2021), pp. 965–976.

[176] R. VIVANTI, A. EPHRAT, L. JOSKOWICZ, O. KARAASLAN, N. LEV-COHAIN, AND J. SOSNA, *Automatic liver tumor segmentation in follow-up ct studies using convolutional neural networks*, in Proc. Patch-Based Methods in Medical Image Processing Workshop, vol. 2, 2015.

[177] D. WANG, M. LI, N. BEN-SHLOMO, C. E. CORRALES, Y. CHENG, T. ZHANG, AND J. JAYENDER, *Mixed-supervised dual-network for medical image segmentation*, in MICCAI, 2019.

[178] G. WANG, M. AERTSEN, J. DEPREST, S. OURSELIN, T. VERCAUTEREN, AND S. ZHANG, *Uncertainty-guided efficient interactive refinement of fetal brain segmentation from stacks of mri slices*, in MICCAI, 2020.

[179] G. WANG, W. LI, M. A. ZULUAGA, R. PRATT, P. A. PATEL, M. AERTSEN, T. DOEL, A. L. DAVID, J. DEPREST, S. OURSELIN, ET AL., *Interactive medical image segmentation using deep learning with image-specific fine tuning*, IEEE TMI, 37 (2018), pp. 1562–1573.

[180] G. WANG, S. ZHAI, G. LASIO, B. ZHANG, B. YI, S. CHEN, T. J. MACVITTIE, D. METAXAS, J. ZHOU, AND S. ZHANG, *Semi-supervised segmentation of radiation-induced pulmonary fibrosis from lung ct scans with multi-scale guided dense attention*, IEEE TMI, (2021).

[181] G. WANG, M. A. ZULUAGA, W. LI, R. PRATT, P. A. PATEL, M. AERTSEN, T. DOEL, A. L. DAVID, J. DEPREST, S. OURSELIN, ET AL., *Deepigeos: a deep interactive geodesic framework for medical image segmentation*, IEEE TPAMI, 41 (2018), pp. 1559–1572.

[182] G. WANG, M. A. ZULUAGA, R. PRATT, M. AERTSEN, T. DOEL, M. KLUSMANN, A. L. DAVID, J. DEPREST, T. VERCAUTEREN, AND S. OURSELIN, *Slic-seg: A minimally interactive segmentation of the placenta from sparse and motion-corrupted fetal mri in multiple views*, Medical Image Analysis, 34 (2016), pp. 137–147.

[183] H. WANG, L. LIN, H. HU, Q. CHEN, Y. LI, Y. IWAMOTO, X.-H. HAN, Y.-W. CHEN, AND R. TONG, *Super-resolution based patch-free 3d medical image segmentation with self-supervised guidance*, arXiv preprint arXiv:2210.14645, (2022).

[184] K. WANG, H. S. TAN, AND R. MCBETH, *Swin unetr++: Advancing transformer-based dense dose prediction towards fully automated radiation oncology treatments*, arXiv preprint arXiv:2311.06572, (2023).

[185] N. WANG, S. LIN, X. LI, K. LI, Y. SHEN, Y. GAO, AND L. MA, *Missu: 3d medical image segmentation via self-distilling transunet*, IEEE TMI, (2023).

[186] P. WANG, J. PENG, M. PEDERSOLI, Y. ZHOU, C. ZHANG, AND C. DESROSIERS, *Self-paced and self-consistent co-training for semi-supervised image segmentation*, Medical Image Analysis, 73 (2021), p. 102146.

[187] R. WANG, T. LEI, R. CUI, B. ZHANG, H. MENG, AND A. K. NANDI, *Medical image segmentation using deep learning: A survey*, IET Image Processing, 16 (2022), pp. 1243–1267.

[188] S. WANG, S. CAO, D. WEI, R. WANG, K. MA, L. WANG, D. MENG, AND Y. ZHENG, *Lt-net: Label transfer by learning reversible voxel-wise correspondence for one-shot medical image segmentation*, in CVPR, 2020, pp. 9159–9168.

[189] S. WANG, S. CAO, D. WEI, R. WANG, K. MA, L. WANG, D. MENG, AND Y. ZHENG, *Lt-net: label transfer by learning reversible voxel-wise correspondence for one-shot medical image segmentation*, in CVPR, 2020.

[190] W. WANG, C. CHEN, M. DING, H. YU, S. ZHA, AND J. LI, *Transbts: Multimodal brain tumor segmentation using transformer*, in MICCAI, 2021.

[191] W. WANG, D. TRAN, AND M. FEISZLI, *What makes training multi-modal classification networks hard?*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12695–12705.

[192] X. WANG, S. HAN, Y. CHEN, D. GAO, AND N. VASCONCELOS, *Volumetric attention for 3d medical image segmentation and detection*, in MICCAI, 2019.

[193] X. WANG, T. XIANG, C. ZHANG, Y. SONG, D. LIU, H. HUANG, AND W. CAI, *Bix-nas: Searching efficient bi-directional architecture for medical image segmentation*, in MICCAI, 2021.

[194] Y. WANG, Z. LI, J. MEI, Z. WEI, L. LIU, C. WANG, S. SANG, A. L. YUILLE, C. XIE, AND Y. ZHOU, *Swinmm: masked multi-view with swin transformers for 3d medical image segmentation*, in MICCAI, 2023.

[195] Y. WANG, Y. ZHANG, J. TIAN, C. ZHONG, Z. SHI, Y. ZHANG, AND Z. HE, *Double-uncertainty weighted method for semi-supervised learning*, in MICCAI, 2020, pp. 542–551.

[196] Y. WANG, Y. ZHOU, W. SHEN, S. PARK, E. K. FISHMAN, AND A. L. YUILLE, *Abdominal multi-organ segmentation with organ-attention networks and statistical fusion*, Medical Image Analysis, 55 (2019), pp. 88–102.

[197] H. WU, G. CHEN, Z. WEN, AND J. QIN, *Collaborative and adversarial learning of focused and dispersive representations for semi-supervised polyp segmentation*, in ICCV, 2021, pp. 3489–3498.

[198] H. WU, X. LI, Y. LIN, AND K.-T. CHENG, *Compete to win: Enhancing pseudo labels for barely-supervised medical image segmentation*, IEEE TMI, (2023).

[199] M. WU AND N. GOODMAN, *Multimodal generative models for scalable weakly-supervised learning*, Advances in Neural Information Processing Systems, (2018).

[200] Y. WU AND K. HE, *Group normalization*, Int. J. Comput. Vis., 128 (2020), pp. 742–755.

[201] Y. XIA, D. YANG, Z. YU, F. LIU, J. CAI, L. YU, Z. ZHU, D. XU, A. YUILLE, AND H. ROTH, *Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation*, Medical Image Analysis, 65 (2020), p. 101766.

[202] Y. XIE, J. ZHANG, Z. LIAO, J. VERJANS, C. SHEN, AND Y. XIA, *Pairwise relation learning for semi-supervised gland segmentation*, in MICCAI, 2020, pp. 417–427.

[203] Y. XIE, J. ZHANG, C. SHEN, AND Y. XIA, *Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation*, in MICCAI, 2021, pp. 171–180.

[204] N. XU, B. PRICE, S. COHEN, J. YANG, AND T. S. HUANG, *Deep interactive object selection*, in CVPR, 2016.

[205] X. XU, Q. LU, L. YANG, S. HU, D. CHEN, Y. HU, AND Y. SHI, *Quantization of fully convolutional networks for accurate biomedical image segmentation*, in CVPR, 2018.

[206] Y. XU, X. YU, J. ZHANG, L. ZHU, AND D. WANG, *Weakly supervised rgb-d salient object detection with prediction consistency training and active scribble boosting*, IEEE TIP, 31 (2022), pp. 2148–2161.

[207] Z. XU AND M. NIETHAMMER, *Deepatlas: Joint semi-supervised learning of image registration and segmentation*, in MICCAI, 2019, pp. 420–429.

[208] Z. XU AND M. NIETHAMMER, *Deepatlas: Joint semi-supervised learning of image registration and segmentation*, in MICCAI, 2019, pp. 420–429.

[209] X. YAN, W. JIANG, Y. SHI, AND C. ZHUO, *Ms-nas: Multi-scale neural architecture search for medical image segmentation*, in MICCAI, 2020.

[210] H. YANG, C. SHAN, A. F. KOLEN, ET AL., *Deep q-network-driven catheter segmentation in 3d us by hybrid constrained semi-supervised learning and dual-unet*, in MICCAI, 2020, pp. 646–655.

[211] H. YANG, J. SUN, H. LI, L. WANG, AND Z. XU, *Neural multi-atlas label fusion: Application to cardiac MR images*, Medical Image Anal., 49 (2018), pp. 60–75.

[212] J. YANG, L. JIAO, R. SHANG, X. LIU, R. LI, AND L. XU, *Ept-net: Edge perception transformer for 3d medical image segmentation*, IEEE TMI, (2023).

[213] L. YU, S. WANG, X. LI, C.-W. FU, AND P.-A. HENG, *Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation*, in MICCAI, 2019, pp. 605–613.

[214] Q. YU, L. QI, Y. GAO, W. WANG, AND Y. SHI, *Crosslink-net: double-branch encoder network via fusing vertical and horizontal convolutions for medical image segmentation*, IEEE TIP, 31 (2022), pp. 5893–5908.

[215] Q. YU, H. WANG, D. KIM, S. QIAO, M. COLLINS, Y. ZHU, H. ADAM, A. YUILLE, AND L.-C. CHEN, *Cmt-deeplab: Clustering mask transformers for panoptic segmentation*, in CVPR, 2022.

[216] Q. YU, H. WANG, S. QIAO, M. COLLINS, Y. ZHU, H. ADAM, A. YUILLE, AND L.-C. CHEN, *k-means mask transformer*, in ECCV, 2022.

[217] Q. YU, D. YANG, H. ROTH, Y. BAI, Y. ZHANG, A. L. YUILLE, AND D. XU, *C2fnas: Coarse-to-fine neural architecture search for 3d medical image segmentation*, in CVPR, 2020.

[218] Z. YU-QIAN, G. WEI-HUA, C. ZHEN-CHENG, T. JING-TIAN, AND L. LING-YUN, *Medical images edge detection based on mathematical morphology*, in 2005 IEEE engineering in medicine and biology 27th annual conference, IEEE, 2006, pp. 6492–6495.

[219] M. YUAN, Y. XIA, H. DONG, Z. CHEN, J. YAO, M. QIU, K. YAN, X. YIN, Y. SHI, X. CHEN, ET AL., *Devil is in the queries: advancing mask transformers for real-world medical image segmentation and out-of-distribution localization*, in CVPR, 2023, pp. 23879–23889.

[220] X. ZENG, R. HUANG, Y. ZHONG, D. SUN, C. HAN, D. LIN, D. NI, AND Y. WANG, *Reciprocal learning for semi-supervised segmentation*, in MICCAI, 2021, pp. 352–361.

[221] C. ZHANG, Y. CUI, Z. HAN, J. T. ZHOU, H. FU, AND Q. HU, *Deep partial multi-view learning*, IEEE Transactions on Pattern Analysis and Machine Intelligence, (2020).

[222] D. ZHANG, G. HUANG, Q. ZHANG, J. HAN, J. HAN, Y. WANG, AND Y. YU, *Exploring task structure for brain tumor segmentation from multi-modality mr images*, IEEE TIP, 29 (2020), pp. 9032–9043.

[223] P. ZHANG, X. CHEN, Z. YIN, X. ZHOU, Q. JIANG, W. ZHU, D. XIANG, Y. TANG, AND F. SHI, *Interactive skin wound segmentation based on feature augment networks*, IEEE Journal of Biomedical and Health Informatics, (2023).

[224] Y. ZHANG, H. LIU, AND Q. HU, *Transfuse: Fusing transformers and cnns for medical image segmentation*, in MICCAI, 2021.

[225] Z. ZHANG, H. FU, H. DAI, J. SHEN, Y. PANG, AND L. SHAO, *Et-net: A generic edge-attention guidance network for medical image segmentation*, in MICCAI, 2019.

[226] Z. ZHANG, L. YANG, AND Y. ZHENG, *Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network*, in CVPR, 2018.

[227] A. ZHAO, G. BALAKRISHNAN, F. DURAND, J. V. GUTTAG, AND A. V. DALCA, *Data augmentation using learned transformations for one-shot medical image segmentation*, in CVPR, 2019, pp. 8543–8553.

[228] A. ZHAO, G. BALAKRISHNAN, F. DURAND, J. V. GUTTAG, AND A. V. DALCA, *Data augmentation using learned transformations for one-shot medical image segmentation*, in CVPR, 2019, pp. 8543–8553.

[229] Q. ZHAO, H. WANG, AND G. WANG, *Lcov-net: A lightweight neural network for covid-19 pneumonia lesion segmentation from 3d ct images*, in International Symposium on Biomedical Imaging, 2021.

[230] X. ZHAO, Y. WU, G. SONG, Z. LI, Y. ZHANG, AND Y. FAN, *A deep learning model integrating fcnns and crfs for brain tumor segmentation*, Medical Image Analysis, 43 (2018), pp. 98–111.

[231] Y.-X. ZHAO, Y.-M. ZHANG, M. SONG, AND C.-L. LIU, *Multi-view semi-supervised 3d whole brain segmentation with a self-ensemble network*, in MICCAI, 2019, pp. 256–265.

[232] T. ZHI-XUAN, H. SOH, AND D. ONG, *Factorized inference in deep markov models for incomplete multimodal time series*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 10334–10341.

[233] C. ZHOU, C. DING, Z. LU, X. WANG, AND D. TAO, *One-pass multi-task convolutional neural networks for efficient brain tumor segmentation*, in International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2018, pp. 637–645.

[234] H.-Y. ZHOU, J. GUO, Y. ZHANG, X. HAN, L. YU, L. WANG, AND Y. YU, *nnformer: Volumetric medical image segmentation via a 3d transformer*, IEEE TIP, (2023).

[235] H.-Y. ZHOU, J. GUO, Y. ZHANG, L. YU, L. WANG, AND Y. YU, *nnformer: Interleaved transformer for volumetric segmentation*, arXiv preprint arXiv:2109.03201, (2021).

[236] S. ZHOU, D. NIE, E. ADELI, J. YIN, J. LIAN, AND D. SHEN, *High-resolution encoder–decoder networks for low-contrast medical image segmentation*, IEEE TIP, 29 (2019), pp. 461–475.

[237] T. ZHOU, S. CANU, P. VERA, AND S. RUAN, *Brain tumor segmentation with missing modalities via latent multi-source correlation representation*, in MICCAI, 2020.

[238] T. ZHOU, L. LI, G. BREDELL, J. LI, AND E. KONUKOGLU, *Quality-aware memory network for interactive volumetric image segmentation*, in MICCAI, 2021.

[239] T. ZHOU, L. LI, G. BREDELL, J. LI, J. UNKELBACH, AND E. KONUKOGLU, *Volumetric memory network for interactive medical image segmentation*, Medical Image Analysis, 83 (2023), p. 102599.

[240] Y. ZHOU, L. XIE, W. SHEN, Y. WANG, E. K. FISHMAN, AND A. L. YUILLE, *A fixed-point model for pancreas segmentation in abdominal ct scans*, in MICCAI, 2017.

[241] Z. ZHOU, M. M. RAHMAN SIDDIQUEE, N. TAJBAKHSH, AND J. LIANG, *Unet++: A nested u-net architecture for medical image segmentation*, in Deep learning in medical image analysis and multimodal learning for clinical decision support, Springer, 2018, pp. 3–11.

[242] Z. ZHOU, M. M. R. SIDDIQUEE, N. TAJBAKHSH, AND J. LIANG, *Unet++: A nested u-net architecture for medical image segmentation*, in MICCAI, 2018, pp. 3–11.

[243] Z. ZHOU, M. M. R. SIDDIQUEE, N. TAJBAKHSH, AND J. LIANG, *Unet++: Redesigning skip connections to exploit multiscale features in image segmentation*, IEEE TMI, 39 (2019), pp. 1856–1867.

[244] W. ZHU, A. MYRONENKO, Z. XU, W. LI, H. ROTH, Y. HUANG, F. MILLETARI, AND D. XU, *Neurreg: Neural registration and its application to image segmentation*, in WACV, 2020, pp. 3606–3615.

[245] W. ZHU, A. MYRONENKO, Z. XU, W. LI, H. ROTH, Y. HUANG, F. MILLETARI, AND D. XU, *Neurreg: Neural registration and its application to image segmentation*, in WACV, 2020, pp. 3617–3626.

[246] Y. ZHU, Z. CHEN, S. ZHAO, H. XIE, W. GUO, AND Y. ZHANG, *Ace-net: biomedical image segmentation with augmented contracting and expansive paths*, in MICCAI, 2019.

[247] Z. ZHU, C. LIU, D. YANG, A. YUILLE, AND D. XU, *V-nas: Neural architecture search for volumetric medical image segmentation*, in IEEE 3DV, 2019.

[248] Z. ZHU, Y. XIA, W. SHEN, E. FISHMAN, AND A. YUILLE, *A 3d coarse-to-fine framework for volumetric medical image segmentation*, in IEEE 3DV, 2018.

[249] B. ZITOVA AND J. FLUSSER, *Image registration methods: a survey*, Image and vision computing, 21 (2003), pp. 977–1000.

[250] W. ZOU, X. QI, W. ZHOU, M. SUN, Z. SUN, AND C. SHAN, *Graph flow: Cross-layer graph flow distillation for dual efficient medical image segmentation*, IEEE TMI, 42 (2022), pp. 1159–1171.