Semi-Supervised Medical Image Segmentation with ADMM-Based Energy Regularization

MPhil Thesis

Qian Wang

Data Science and Statistics

Department of Mathematics and Statistics

University of Strathclyde, Glasgow

April 17, 2026

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Abstract

This paper addresses the challenge of data scarcity in medical image segmentation by introducing a suitable energy functional as a loss term within a semi-supervised learning framework. Inspired by the variational model proposed by Wei et al. (2017), in which the Potts energy functional is minimized using a region force derived from the negative log-likelihood of the Bernoulli distribution, this study explores the integration of such energy formulations with a U-Net–based deep learning model. Unlike the classical Chan–Vese model, which is prone to local minima due to its non-convex nature, the convex relaxed Potts energy model enables global optimization. However, when directly employed as a loss function, the energy functional remains highly nonlinear and leads to suboptimal convergence. To overcome this limitation, an ADMM-based optimization strategy, following the approach of Wei et al., is adopted to enhance convergence stability and computational efficiency. When incorporated as a regularization term in the neural network, the proposed energy-based loss serves as a spatial prior that mitigates the limitations of deep learning in enforcing spatial coherence and boundary consistency, particularly for irregular or noisy lesion boundaries. Moreover, this integration effectively reduces the dependence on large-scale labeled data, thereby improving the applicability of deep models to semi-supervised segmentation scenarios. Comprehensive numerical experiments demonstrate that the proposed ADMM-enhanced semi-supervised method achieves accurate and stable segmentation performance under limited annotation conditions.

Acknowledgements

This acknowledgment marks the completion of my MPhil study in the Department of Mathematics and Statistics at the University of Strathclyde. Time has flown by, yet I still remember stepping into Livingstone Tower on my first day—full of excitement and anticipation. The warmth and kindness of everyone in the department made me feel truly at home, and for that, I am deeply grateful.

My heartfelt thanks go to my supervisor, Professor Ke Chen, for introducing me to the world of research. His guidance, patience, and academic rigor have shaped my thinking and inspired me to pursue excellence. What I have achieved so far is only the beginning, and his mentorship will continue to guide me in the years ahead.

I am also thankful for the wonderful friends I met here, especially Charlotte and Hannah, whose support helped me through moments of uncertainty and stress. I am grateful to everyone and everything I encountered along this journey—they have all contributed to who I am today.

As this chapter closes, a new one begins. I hope to remain passionate, brave, and able to find joy in every moment.

May your life set sail toward great ideals.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Image segmentation is a fundamental task in computer vision, aiming to partition an image into semantically meaningful regions. In medical image analysis, the primary objective is to delineate lesion areas or anatomical structures of interest accurately. AI models can learn target features from a large number of images and accurately segment the area of interest in complex content [3]. At the same time, the analysis results of the AI model provide medical workers with powerful solutions. They are often based on a large number of data sets for model training and have extremely high requirements for the accuracy of image annotation. With the growing diversity of disease types and increasing diagnostic demands, AI-assisted diagnosis has become essential in modern medical practice [3]. All of these are inseparable from accurate lesion segmentation. In addition, traditional mathematical models continue to constitute an important foundation for image segmentation research. Mathematical models classify pixels based on their common features, thereby distinguishing the texture, intensity and color of the image [4]. Region growing and emerging [5], watershed algorithms [6], minimum description length criteria [7], Mumford-Shah energy minimization [8], PDE-based active contour models for curve evolution [9], etc. are more popular mathematical variational segmentation models. Taking the classic active contour and level set model Chan-vese [10] as an example, under constraints, it drives the curve evolution by minimizing a curve-dependent energy functional under constraints, so that it converges to find the local minimum of the target boundary, thereby obtaining a segmentation result with

continuous and smooth boundaries [10] [4]. The Potts model minimizes region boundary length by penalizing inconsistent jumps in neighboring labels, thereby encouraging internal consistency and clean, smooth boundaries in the resulting segmented regions. This model is a classic smoothing prior for multi-class piecewise constant segmentation [11]. This type of model can use the prior on boundary continuity to obtain image information and identify segmentable regions. It exhibits excellent segmentation performance when faced with limited labeled data and small datasets.

AI segmentation models are primarily used for two-dimensional or three-dimensional images such as tomography or magnetic resonance imaging (MRI), which almost all share the characteristics of low contrast, blurred boundaries, and complex anatomical structures. Deep learning models typically make pixel-by-pixel predictions, which can lead to spatial inconsistencies when trained on small samples. Furthermore, preparing large datasets is relatively difficult in the field of medical image analysis. Rare diseases, the lack of public access to images due to patient privacy concerns, and the costly and time-consuming manual annotation all pose challenges in preparing good datasets.

Accurate AI models for medical image segmentation are challenging. Mathematical models are often affected by noise and lack flexibility when segmenting noisy images. To address these challenges, this paper proposes a joint segmentation model that combines the strengths of mathematical models and deep learning. This model combines a variational energy model with a deep learning neural network model to form a semi-supervised segmentation framework. Energy-based regularization is incorporated into the neural network. While data-driven feature learning is performed, input image features are obtained through feature priors and used as input for the next round of feature learning. This approach ensures spatial consistency in model learning and can be applied to both labeled and unlabeled data, ensuring robustness while reducing reliance on labeled data. Specifically, an energy-based loss function is introduced as a regularization term, inspired by the variational Potts model proposed by Wei et al. [12]. This model, a continuous convex relaxation of the Potts model [11], can be defined on a multi-class function space. The global optimal solution is found through continuous minimization, combining a weighted sum of boundary perimeters to achieve the seg-

mentation objective. This method can be optimized using the PDHG (Primal-Dual Hybrid Gradient) and ADMM (Alternating Direction Method of Multipliers) numerical solution method [12]. This method can achieve good segmentation results even with limited datasets and labels. To evaluate the effectiveness and generalization of the proposed framework, we also systematically compare it with several representative segmentation methods. These methods include purely supervised learning (using Dice or cross-entropy to train a U-Net model), traditional variational models exemplified by Chan-Vese, the proposed semi-supervised model, and a semi-supervised model optimized with ADMM. We evaluated each method based on its theoretical foundation, segmentation accuracy, and adaptability to real-world medical images. Through these comparisons, we demonstrate that the proposed energy function regularization framework maintains excellent segmentation quality and structural consistency even with limited datasets, demonstrating its significant potential for challenging medical image analysis tasks.

Chapter 2

Overview of Image Segmentation Methods

Image segmentation is the process of dividing meaningful regions within an image into distinct parts based on human perception and needs. It is a crucial component of image analysis [13]. Image segmentation results can be used for target recognition, disease diagnosis in medical devices, and subsequent image analysis. Before image segmentation, the image undergoes preprocessing to meet the requirements of subsequent complex segmentation algorithms. Segmentation algorithms primarily use image texture, pixels, and intensity as a basis to delineate regions with distinct characteristics. Zhu et al. [13] metaphorically describe the segmentation object as an "unclearly defined object." This object can exist in a variety of scenarios, from natural scenery to piles of debris to lesions within tissues and organs. It often overlaps with other image information, making it a challenge to clearly define the object itself [13]. Traditional methods consider the image's entirety, excluding the background, as global information. Objects can be decomposed into features such as color, texture, and intensity, which can be represented mathematically using curvature or convexity. These features are considered local information, typically captured through statistical calculations [13]. In recent years, researchers have also used traditional segmentation algorithms as a foundation for studying optimization problems, achieving remarkable results in achieving accurate segmentation. Traditional mathematical segmentation models, including variational

and image-based models, are generally considered unsupervised methods [13]. They typically model the pixel values, textures, or gradients of the original image and obtain segmentation results through energy functionals or minimization. These unsupervised methods generally do not rely on manual annotations or require training samples, meaning they can segment regions of interest without knowing the segmentation objective. Zhu et al. [13] broadly categorize unsupervised methods into clustering-based methods, graph-based methods, and variational models. Clustering-based methods (such as K-means, Gaussian mixture models, and mean shift) group pixels based on similarity in feature space, while graph-based methods (such as regularized cuts and graph cuts) treat the image as a weighted graph and find the optimal segmentation by minimizing a cost function on the graph. Superpixel generation methods (such as SLIC and SEEDS) can be viewed as an over-segmentation step, generating homogeneous regions for further processing [13]. This paper focuses on variational models. Variational model segmentation typically frames the problem as minimizing an energy functional. The seminal work of Mumford and Shah [8] formulated the segmentation problem as an energy functional minimization problem, striking a balance between data fidelity, region smoothness, and boundary regularization. Subsequent research has proposed a variety of simplified and numerical solutions, such as the level set-based active contour model and the piecewise constant table [4]. Among these methods, the Chan–Vese model is a simplification and generalization of the Mumford–Shah variational model, effectively handling objects with blurred boundaries or weak gradients. On the other hand, the Potts model, originally derived from spin system modeling in statistical physics, has been widely used in image segmentation and is often viewed as a discrete counterpart or multi-label extension of the piecewise constant Mumford–Shah model. The following sections will detail these two models and their mathematical derivations.

On the other hand, with the advancement of artificial intelligence (AI) technology, deep learning models are increasingly being used in image analysis. Deep learning segmentation models aim to identify and isolate regions of interest (ROIs) in images through training and learning, allowing them to be used in subsequent analysis. These models share a similar processing strategy: image training set preprocessing, model

training, and model estimation to obtain results. With the continuous optimization of AI models, the field of image segmentation has evolved from relying on handcrafted features and energy minimization models to data-driven, end-to-end trainable neural networks. As summarized in their review article by Rayed et al. [14], deep learning has become the mainstream approach for medical image segmentation. The introduction of fully convolutional networks (FCNs) marked a significant milestone in this field, demonstrating for the first time that convolutional neural networks can directly achieve pixel-level predictions, laying the theoretical and methodological foundation for deep learning semantic segmentation [15]. Subsequently, numerous studies have proposed improvements based on this approach, including encoder-decoder architectures, dilated convolutions, and attention mechanisms. These improvements have significantly improved segmentation accuracy and robustness, particularly in complex medical image analysis scenarios [14]. Among these models, the U-Net and its improved versions have become the most widely adopted architecture for medical image segmentation, thanks to their symmetrical encoder-decoder design, cross-layer skip connections for precise localization, and their ability to achieve good performance even with limited training data [1]. These characteristics make the U-Net particularly suitable for clinical applications where labeled data is scarce. The next section will provide a detailed introduction to the U-Net architecture and explain its role in the proposed method.

2.1 The Image Segmentation Problem

The primary objective of image segmentation is to partition an image into clear and semantically meaningful regions that correspond to different anatomical structures or functional components. By accurately delineating these regions, segmentation not only enables quantitative analysis and assists in diagnosis, but also provides an essential foundation for subsequent image-based modeling and interpretation.

However, due to the complexity of image intensity distributions and spatial structures, achieving high-precision segmentation remains a considerable challenge. Ideally, each region within an image should exhibit a consistent intensity pattern, allowing clear separation between objects and background. In practical scenarios, however, the inten-

sity distributions of different objects or regions often overlap significantly because of imaging noise, limited spatial resolution, and variations in acquisition conditions, leading to blurred boundaries and reduced contrast. For example, in an image containing multiple geometric shapes, if circular and triangular regions share similar intensity levels, segmentation based solely on pixel intensity would fail to accurately distinguish the target region. Similar situations commonly occur in natural scenes, where objects with comparable colors or brightness may differ in shape or spatial arrangement. Therefore, segmentation algorithms need to go beyond single intensity features and incorporate additional cues such as shape priors, spatial structure, and contextual information to achieve more robust and generalizable results. This necessity has driven the development of modern segmentation methods that jointly model intensity, geometry, and semantic features, thereby attaining higher accuracy in complex visual environments.

At the same time, deep learning models still face several challenges. The limited number of training samples and the scarcity of annotated data, particularly in the medical imaging domain where data are highly complex and heterogeneous, significantly constrain the generalization capability of segmentation models. As highlighted by Litjens et al. [16], deep learning algorithms still face substantial challenges in medical image analysis. The first is the often-mentioned problem of dataset size. A complete deep learning model requires large-scale training data, but this approach isn't feasible in every case. For example, in radiology, DICOM-based PACS has become standard operating system in many large hospitals in Europe and the United States since the late 2000s, accumulating millions of highly structured images [17]. In contrast, the progress of image archiving and standardization in other medical specialties (such as endoscopy, echocardiography, digital pathology, etc.) is relatively uneven, and it is still difficult to match radiology in terms of data scale, annotation quality and accessibility. Another challenge is that although the imaging data is huge, it lacks task-related annotation information. For example, free text reports (such as radiologist reports) often record examination results, but these reports are difficult to be directly used by deep learning algorithms and require complex text mining methods. How to convert free text or semi-structured reports into the annotations required for network training

remains a key issue in the future [16]. Secondly, annotation noise and uncertainty are common. Even experienced experts may have significant differences in lesion boundary annotation, which will directly affect the stability and generalization ability of the model [18]. Third, the class imbalance problem is particularly prominent in medical imaging tasks. For example, in breast cancer screening, normal samples far outnumber positive samples, which causes the model to easily ignore minority lesions, thereby reducing detection sensitivity [19]. In addition, how to effectively integrate clinical information with imaging features remains a challenge. Many studies have shown that relying solely on imaging features may not be sufficient to support complex diagnoses, and additional clinical data (such as patient history, age, etc.) are often not fully utilized in model training [19]. Finally, medical images generally have the characteristics of ultra-high resolution and large-scale data volume. For example, full-slice images in digital pathology can reach several gigapixels, which poses a great challenge to the storage and computing resources of deep learning models. Existing methods usually rely on cropping or downsampling, which makes it difficult to take into account both global and local information at the same time [16] [1]. In summary, the challenges of medical image segmentation are not only reflected in data acquisition and labeling, but also involve many aspects such as category distribution, clinical information fusion and large-scale image processing. These problems urgently need to be solved in future research.

2.2 Variational Models for Image Segmentation

In the field of image segmentation, variational models have established a unified and rigorous mathematical framework. Their core idea is to achieve optimal segmentation by designing an energy functional and finding its minimum. The Mumford–Shah model [8] represents an early and influential variational approach that integrates regional fidelity and boundary regularity within a unified framework. By jointly considering these two aspects for the first time, it established a rigorous foundation for later developments, inspiring a series of extensions and improved models, including the Chan–Vese and

Potts formulations. The Mumford–Shah energy functional is usually written as:

$$E(u, K) = \int_{\Omega \setminus K} (f(x) - u(x))^2 \, dx + \lambda \int_{\Omega \setminus K} |\nabla u(x)|^2 \, dx + \nu |K|, \tag{2.1}$$

The function u is employed as a piecewise smooth approximation to the input image $f(x)$, while the set K captures the discontinuity curves or edges. The parameters λ and ν, both strictly positive, regulate the contribution of the three components in the energy formulation. Following the interpretation given by Vitti [20] the three terms in the Mumford–Shah functional can be interpreted as follows:

- **Data fidelity term** $\int_{\Omega \setminus K} (f(x) - u(x))^2 dx$: Ensures that the segmentation result $u(x)$ is as close as possible to the original image $f(x)$, thereby maintaining the image's true grayscale distribution. In medical scenarios, this ensures that the segmentation does not deviate from the actual lesion or tissue signal.

- **Smoothing term** $\lambda \int_{\Omega \setminus K} |\nabla u(x)|^2 \, dx$: Encourages smoothness of $u(x)$ within each region while allowing sharp changes across the discontinuity set K. It encourages consistency within the region and reduces noise interference. It is particularly suitable for medical images with high noise levels such as MRI and CT, making the anatomical structure more coherent.

- **Boundary length term** $\nu |K|$: Penalizes the total length of the segmentation boundaries, favoring compact and simpler boundary structure. In medical image analysis, it is important to ensure concise segmentation contours and avoid jagged or unnatural boundaries.

A detailed mathematical analysis of the Mumford–Shah model and its numerical treatments can be found in classical variational literature, including the comprehensive survey by Vitti [20]. Below, we will briefly introduce the Mumford-Shah model in the context of the article. In theory, the Mumford–Shah model is described as a free discontinuity problem. Unlike traditional free boundary problems, this model allows the segmentation boundary K not to be predetermined as a smooth curve, but to emerge automatically during the energy minimization process as an unknown dis-

continuous set. Mumford and Shah [8] proposed a conjecture: minimizers exist, and the segmentation boundaries can be composed of several smooth curve segments, with only three types of singularities possibly occurring — crack tips (which are endpoints of discontinuity curves located inside the image domain), triple points (three curves meeting at 120°), and boundary points (where the segmentation boundary meets the outer boundary). Although this conjecture has not been fully proven, it is supported by a large number of results [20] [8]. For existence results, the direct method typically relies on compactness arguments and lower semicontinuity. However, the complexity of the Mumford–Shah problem lies in the coupling between the two-dimensional energy (data fidelity and smoothness terms) and the one-dimensional measure (boundary length). This coupling makes it extremely difficult to prove the *lower semicontinuity* of the last term directly, and hence obtaining rigorous analytical results for the Mumford–Shah model remains a highly challenging problem. Here, lower semicontinuity refers to the variational property that, for any sequence $u_k \to u$ in $L^1(\Omega)$,

$$\mathcal{E}(u) \leq \liminf_{k\to\infty} \mathcal{E}(u_k),$$

where $\mathcal{E}$ denotes the Mumford–Shah energy. Intuitively, this ensures that the limit of a minimizing sequence does not increase the energy and is therefore fundamental for proving the existence of minimizers.

To overcome this difficulty, theoretical developments introduced the special function space of bounded variation (SBV) and the notion of relaxation. The space $\mathrm{SBV}(\Omega)$ consists of functions whose distributional derivative can be decomposed into an absolutely continuous part and a jump part, meaning that discontinuities are allowed but can only occur on a set of finite $\mathcal{H}^1$-measure. In this framework, we allow u to exhibit jumps on sets of measure zero, use the jump set S_u to represent the unknown discontinuity set K, and measure its "length" by the one-dimensional Hausdorff measure $\mathcal{H}^1(S_u)$. The relaxed Mumford–Shah functional then reads

$$\int_\Omega (u-f)^2\,dx + \lambda \int_\Omega |\nabla u|^2\,dx + \ \alpha\,\mathcal{H}^1(S_u), \tag{2.2}$$

which involves only the function u itself and the measure of its jump set [20]. Here, λ retains the same meaning as in Eq. (2.1), weighting the smoothness term. The parameter α denotes the weight of the jump-set penalty in the relaxed formulation and corresponds to the boundary penalty parameter ν in the classical Mumford–Shah functional in Eq. (2.1).

Within the SBV framework, the works of Ambrosio [21], De Giorgi [22], and others established the *existence* of minimizers, laying the basis for rigorous analysis. Existence results further allow one to study the Euler–Lagrange conditions (in the distributional sense), where natural boundary conditions on the jump set S_u appear explicitly [20]. Consequently, the Mumford–Shah model becomes an integrated variational system that couples PDEs with free discontinuities, providing a bridge between PDE analysis and geometric measure theory and enabling transformations between these two viewpoints [20].

In summary, the relaxed Mumford–Shah formulation highlights discontinuities and unifies *data fidelity*, *regional smoothness*, and *boundary complexity* within a single framework. Its core idea is to describe discontinuities via the jump set of SBV functions, measured by the Hausdorff length $\mathcal{H}^1$. This offers a mathematically rigorous route to image segmentation: one performs variational analysis over SBV functions, and the discontinuity set emerges naturally, providing a solid foundation for subsequent numerical approximations and further variational studies [20].

2.2.1 The Chan-Vese Model

Based on the framework mentioned above, subsequent research has proposed several simplified and extended models. For example, the Chan–Vese model, by assuming that the regional grayscale is constant, transforms the Mumford–Shah model into a level set-based implementation method, which is more convenient for numerical calculations; while the Potts model can be viewed as a discrete form of the piecewise constant Mumford–Shah, providing a unified and efficient solution to the multi-label segmentation problem. These developments fully demonstrate the role of the Mumford–Shah model as a theoretical cornerstone and provide inspiration for the regularization ideas

in subsequent deep learning methods. Region-based variational methods have been widely applied in biomedical image segmentation, such as cell microscopy, organ CT, brain MR, and retinal fundus images. These methods aim to partition an image into meaningful regions by minimizing an energy functional composed of data fidelity and regularization terms. Image segmentation is a crucial part of image processing, where precise segmentation often leads to superior image processing results. In early image segmentation tasks, region force methods, represented by the Chan-Vese model (CV model), were commonly used [10]. This model treats the pixel values of a grayscale image as energy and constructs an energy function. The energy function consists of data terms and regularization terms; the data terms calculate the difference in data between the foreground and background of the image, while the regularization terms smooth the segmentation boundaries.

For a given closed region $I : \Omega \to R$, let $I(x)$ be the intensity function that represents the pixel value at point x in the image. Following the formulation of Chan and Vese [10], the Chan–Vese model can be described by the following equations:

$$\begin{aligned} E(c_1, c_2, \phi) = &\int_{\Omega} \lambda_1 \left(I(x) - c_1\right)^2 H(\phi(x))\, dx \\ &+ \int_{\Omega} \lambda_2 \left(I(x) - c_2\right)^2 \left(1 - H(\phi(x))\right) dx \\ &+ \mu \int_{\Omega} \left|\nabla H(\phi(x))\right| \, dx \end{aligned} \tag{2.3}$$

where $\phi(x)$ is a level set function whose zero level defines the contour C. The Heaviside function $H(\phi(x))$ indicates the inside region of C, while $1 - H(\phi(x))$ corresponds to the outside. The constants c_1 and c_2 denote the average intensities inside and outside the contour, respectively. Here, λ_1 and λ_2 weight the data fidelity terms inside and outside the contour, respectively, while $\mu > 0$ controls the strength of the contour-length regularization term. The first two terms measure the data fidelity to region means, and the last term regularizes the contour length, which can be equivalently expressed as

$$\mu \int_{\Omega} \delta(\phi(x)) |\nabla \phi(x)| \, dx,$$

where $\delta(\phi)$ is the Dirac delta function.

By minimizing $E(c_1, c_2, \phi)$, the model evolves the contour C toward the object boundaries, achieving a partition of the image domain into two regions with approximately piecewise-constant intensity. Specifically, alternating minimization with respect to c_1 and c_2 yields the Euler–Lagrange updates

$$c_1 = \frac{\int_\Omega I(x) H(\phi(x))\, dx}{\int_\Omega H(\phi(x))\, dx},$$

$$c_2 = \frac{\int_\Omega I(x)(1 - H(\phi(x)))\, dx}{\int_\Omega (1 - H(\phi(x)))\, dx},$$

which represent the mean intensities inside and outside the contour. Meanwhile, minimizing with respect to ϕ leads to the evolution equation

$$\frac{\partial \phi}{\partial t} = \delta(\phi) \left[\mu \operatorname{div}\left(\frac{\nabla \phi}{|\nabla \phi|} \right) - \lambda_1 (I - c_1)^2 + \lambda_2 (I - c_2)^2 \right].$$

However, the standard Chan–Vese model is limited to two-phase segmentation and assumes homogeneous intensities within each region. When multiple objects or intensity inhomogeneity exist, the model's performance tends to degrade, resulting in incomplete or over-smoothed boundaries.

When there are multiple segmentation objects and numerous segmentation boundaries, the performance of the CV model appears somewhat constrained. As shown in the following image:

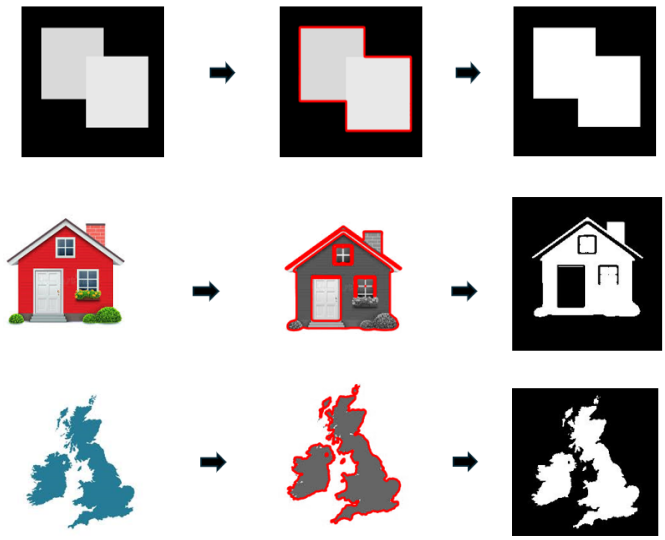


Figure 2.1: When the CV model processes an image with fewer segmentation objects, clearer boundaries, and simpler segmentation regions, it can achieve precise image segmentation. The image above shows that the CV model demonstrates relatively accurate segmentation capabilities.

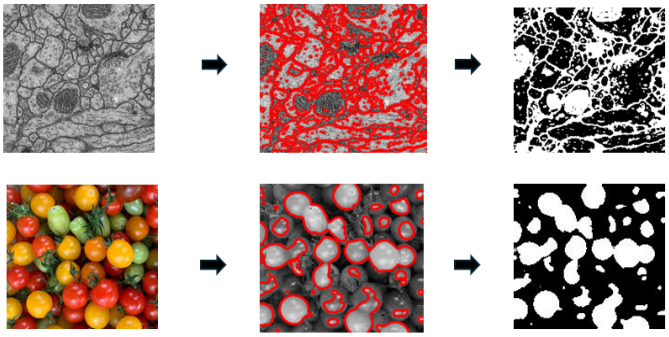


Figure 2.2: When the CV model encounters such images, it struggles to accurately compute the target boundaries, as seen in examples like cell structure images or multi-colored cherry tomatoes. The excessive number of boundaries or the similarity between foreground and background colors makes it difficult for the CV model to establish reliable boundaries.

2.2.2 Potts Model

The Potts model was first proposed by R. B. Potts in 1952 [11]. Building upon the Ising model studied by Kramers and Wannier for the zero-field square lattice, Potts extended the framework to a more general system with r possible spin states, in order to describe the order–disorder phase transition in multistate systems [11]. In the Potts model [11], a system consists of a set of lattice sites, where each site i can take one of r possible states:

$$s_i \in \{1, 2, \ldots, r\}.$$

The energy (Hamiltonian) of the system is defined as:

$$H(s) = -J \sum_{\langle i,j \rangle} \delta(s_i, s_j), \tag{2.4}$$

where

- $\langle i, j \rangle$: denotes all pairs of nearest-neighbor sites;
- $J > 0$: is the coupling constant, which determines the degree of energy reduction when neighboring sites are in the same state;
- $\delta(s_i, s_j)$: is the Kronecker delta function, defined as

$$\delta(s_i, s_j) = \begin{cases} 1, & s_i = s_j, \\ 0, & s_i \neq s_j. \end{cases}$$

Here, when two neighboring lattice sites i and j take the same spin state ($\delta = 1$), the interaction contributes a lower energy of $-J$. Conversely, when they take different states ($\delta = 0$), the energy remains unchanged. Consequently, the system tends to minimize its total energy by favoring configurations in which neighboring sites share identical states. At low temperatures, this leads to an ordered phase characterized by large homogeneous regions, whereas at high temperatures, thermal fluctuations induce a disordered phase with randomly distributed states. When $r = 2$, the Potts model

reduces to the Ising model [11]:

$$H = -J \sum_{\langle i,j \rangle} s_i s_j, \quad s_i \in \{-1, +1\}. \tag{2.5}$$

Typically, the formula can be rewritten as follows:

$$E(s) = \sum_{(i,j) \in \mathcal{N}} w_{ij}[s_i \neq s_j], \tag{2.6}$$

where w_{ij} denotes the similarity or edge weight between pixels i and j, and $[s_i \neq s_j]$ is an indicator function, when the labels of adjacent pixels are different, it is 1, otherwise it is 0. That is, if two neighbors have different labels, how much penalty is imposed. This is completely consistent with the hope that "similar regions are classified into the same category" in the image segmentation task, so the Potts model was later widely used in the field of image segmentation.

In image segmentation, if adjacent pixels belong to the same tissue or structure, we typically want them to be assigned the same label. The Potts model represents this as a spatial consistency prior, meaning that adjacent pixels share the same label. Minimizing the Potts energy corresponds to producing piecewise-constant regions that retain clear boundaries even when the data changes, forming continuous regions and reducing noise-induced segmentation errors. This property is crucial for segmenting medical images with complex tissues. The Potts model can be viewed as a discretized form of the Mumford-Shah model [23]. It is also a special case of a Markov random field (MRF) with a pairwise potential function, commonly used in probabilistic segmentation frameworks [24]. In conditional random field (CRF)-based methods, the pairwise term often takes the Potts form to enforce spatial consistency between adjacent pixels [24]. In medical image segmentation, the Potts model is widely used as a spatial regularizer to maintain tissue coherence and suppress noise. For example, the Bayesian nonparametric Potts model proposed by Johnson et al. [25] is a typical application of the Potts model in medical image segmentation. This method introduces a Potts spatial prior on the voxel label field to enforce neighborhood consistency constraints. It

also incorporates a Dirichlet process mixture model to adaptively describe the intensity distribution of each voxel. The model employs Markov Chain Monte Carlo(MCMC) inference and solves the Potts parameters using Swendsen–Wang clustering updates and path sampling techniques. Experiments demonstrate that this method effectively maintains spatial coherence and functional area boundaries in functional magnetic resonance imaging (fMRI) activation segmentation, significantly reducing false positives and over-smoothing [25]. However, its discrete and non-convex nature makes optimization challenging, motivating recent research that combines Potts regularization with continuous relaxations or deep neural networks.

2.3 Learning-based Segmentation

In recent years, learning-based segmentation methods have become increasingly important in the field of image analysis. Compared with traditional rule-based or variational approaches, learning-based methods offer several notable advantages: they can automatically learn discriminative features from data, achieve high segmentation accuracy, and produce predictions efficiently—often generating reliable results within seconds once trained. This rapid inference capability makes them particularly suitable for real-time or large-scale applications.

Generally, learning-based segmentation can be categorized into data-driven and model-driven approaches. Data-driven methods, such as deep neural networks, rely heavily on large annotated datasets to learn mappings between image appearance and semantic regions. In contrast, model-driven approaches integrate learning with prior knowledge or physical constraints, improving robustness and interpretability in scenarios with limited data.

The versatility of learning-based segmentation has led to its widespread adoption across multiple domains, including medical imaging, biological microscopy, remote sensing, geospatial analysis, and astronomical imaging. Each application domain emphasizes different aspects of segmentation — for instance, medical imaging focuses on precise boundary delineation and tissue differentiation, whereas remote sensing prioritizes large-scale region classification. Consequently, the choice of segmentation model

should be tailored to the domain characteristics: convolutional neural networks (CNNs) and transformer-based architectures are often favored for medical and biological data due to their strong feature-learning capabilities, while hybrid or graph-based models may be more suitable for complex spatial relationships encountered in geospatial and astronomical data. A review article by Litjens et al. [16] on Medical Image Analysis mentions that application-level image segmentation tasks systematically classify medical images, including neuroimaging, retinal images, lung CT, breast imaging, cardiac imaging, abdominal organ segmentation, musculoskeletal system, and digital pathology. Each type of image segmentation presents different challenges. For example, in fundus image segmentation, conditions such as diabetic retinopathy (DR), hard exudates, and microaneurysm are relatively common, but their manifestations in fundus images are not readily apparent. These images are inherently affected by noise. Furthermore, during the image capture process, fundus cameras already undergo some algorithmic processing, such as image reconstruction algorithms that attempt to capture more detail. For doctors, fundus images with numerous influencing factors still require careful analysis to confirm the condition, let alone an AI segmentation model that attempts to accurately segment even tiny hemorrhages or hemangiomas. Therefore, deep learning model features and methods vary across different fields. Litjens et al. [16] summarized similar challenges in medical image segmentation in a review article published in Medical Image Analysis. Specifically, there's the dataset size issue: the extremely high cost of labeling medical images limits the scale of deep learning model training; secondly, the generalization issue: domain differences resulting from different institutions, scanning devices, and parameters make the model unstable when applied across image types; and thirdly, the interpretability issue: while deep learning methods outperform traditional methods, artifacts can occur during training, making it difficult to fully verify the authenticity of the training results. The authors note that addressing these issues requires a combination of emerging approaches such as semi-supervised and weakly supervised learning, generative models, domain adaptation, and explainable AI [16]. Currently, most mainstream image segmentation tasks are based on CNNs and their derivative structures. CNNs have obvious advantages in multi-label segmentation, 3D image pro-

cessing, and small sample learning. Transfer learning and data enhancement have also alleviated the problem of limited medical datasets to a certain extent [16]. The most well-known is U-net [1], which is an encoder-decoder segmentation network that fuses semantic and spatial information through skip connections. During training, it inputs the entire image and outputs a pixel-level segmentation map of the same size as the input image, which can avoid the problem of patch-based CNN requiring splicing [16] [1]. Based on the U-net framework, researchers have proposed various extensions and improvements. Çiçek et al. [26] proposed the 3D U-net, which can achieve 3D segmentation using 2D slice annotations, effectively extending segmentation capabilities to volumetric data. Subsequently, Milletari et al. [27] proposed V-net, which introduced a 3D convolutional structure and adopted the Dice coefficient as the optimization objective to alleviate the class imbalance problem that is common in medical images. At the same time, Drozdzal et al. [28] combined the short skip connection mechanism in the ResNet architecture to enhance feature propagation and gradient transfer, thereby further improving the segmentation performance.

In recent years, large-scale pretrained models have gradually transformed the research landscape of medical image segmentation. Unlike traditional U-net or CNN-based methods, which rely on task-specific annotated data, these models leverage massive amounts of multimodal data and feature retrieval capabilities, enabling cross-task and cross-modality segmentation in a wider range of scenarios. At the methodological level, research mainly focuses on three directions: the first is the pretrain-finetune mode, such as Swin-UNet and MedSegDiff, which achieves better performance by pretraining on large-scale natural images or medical images and then migrating to specific tasks [29]; the second is prompt-based segmentation, represented by Segment Anything Model (SAM), which can achieve segmentation of any object through points, boxes or text prompts. Related medical improved versions (such as MedSAM) have shown superiority in organ and lesion segmentation [30] [31]; the third is vision-language models, which combine medical images with text information (such as case reports) to achieve segmentation based on language prompts, such as Contrastive Language–Image Pretraining (CLIP)-based. The visual language model of the framework aligns images

and language modalities, using a cross-modal approach to improve the content analysis capabilities of medical images [32]. These large models offer numerous advantages. They excel in few-shot or even zero-shot segmentation scenarios, helping to alleviate the problem of insufficient medical image annotation; they better support multimodal fusion (MRI, CT, ultrasound, pathology slides, etc.); and they demonstrate enhanced generalization across diverse diseases and imaging conditions. However, they also face several challenges, including high computational and storage overhead, difficulty in incorporating medical-specific priors, and insufficient clinical interpretability and credibility [33] [34]. Since the research work in this article is based on U-net, its structure and development context will be systematically introduced below.

Due to its effectiveness and simplicity, UNet has become the de facto standard architecture for segmentation tasks in medical imageing. It provides a widely adopted architecture for segmetation tasks in medical image segmentation. It was originally introduced by Ronneberger et al. [1] in 2015 as a solution for biomedical image segmentation tasks. The authors first introduced the use of a U-shaped architecture for semantic segmentation of biological images. The model primarily consists of two components: an encoder on the left and a decoder on the right. Each row of the encoder follows a standard convolutional neural network structure, commonly referred to as the contracting path. Each contracting block contains three convolutional layers, each followed by a ReLU activation function (a non-linear function defined as $f(x) = \max(0, x)$) [35] and a 2×2 max-pooling operation (which performs a local maximum operation over a spatial neighborhood to downsample feature maps) [36] for downsampling. During this process, the number of feature channels is doubled at each downsampling step.

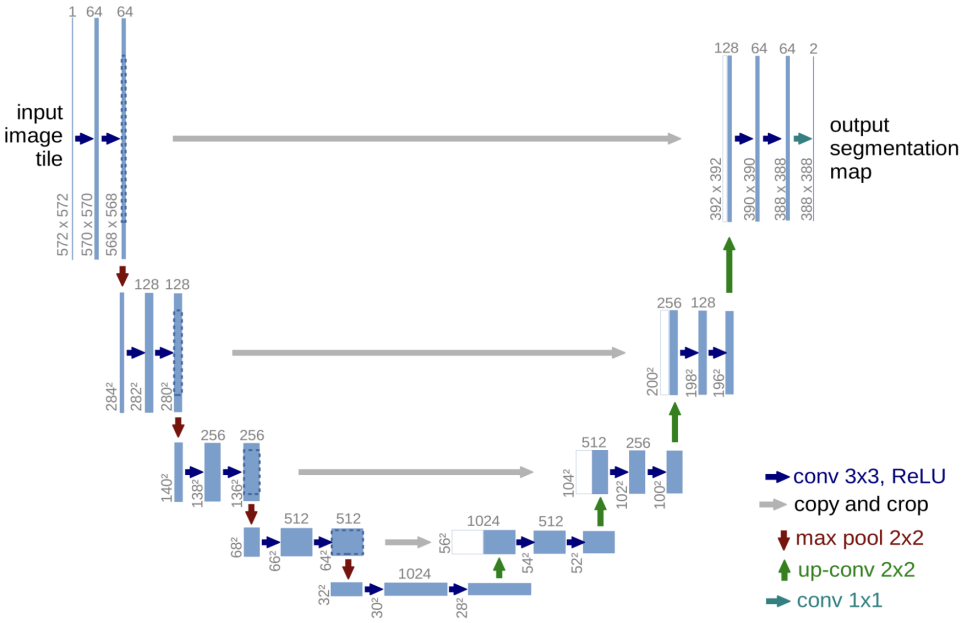


Figure 2.3: U-Net network structure (Reproduced from Ronneberger et al., [1]).The blue blocks represent 3×3 convolution layers with ReLU activations. Down-sampling is performed by 2×2 max-pooling (red arrows), while up-sampling is achieved by 2×2 transposed convolutions (green arrows). Feature maps from the encoder are copied and concatenated to the decoder via skip connections (grey arrows), and a final 1×1 convolution maps the features to the output segmentation.

In contrast, the decoder performs the opposite operations. At each stage, it applies upsampling to the feature maps, reducing the number of channels by half. These upsampled feature maps are then concatenated with the corresponding feature maps from the symmetric contracting path via skip connections (which directly pass feature maps from encoder to decoder to preserve spatial information) [1].

The use of skip connections in the UNet architecture enables the network to recover semantic information that is often lost during the downsampling process. While deeper convolutional layers capture high-level and fine-grained semantic representations, the shallower layers focus more on low-level features such as color, texture, and edges. Since downsampling inevitably leads to a loss of spatial information, the skip connections help to recover this lost information by directly linking low-level and high-level features. This mechanism is particularly advantageous in medical image analysis, where fine structures

and complex boundaries are common and must be preserved.

Moreover, the UNet architecture is relatively simple and lightweight in terms of the number of parameters, yet it is capable of capturing rich semantic features more effectively than many more complex models. The skip connections play a crucial role in enhancing segmentation performance, making UNet especially suitable for tasks in biomedical image processing.

The network architecture employed in this study is based on U-Net and incorporates a Transformer bottleneck layer (which captures long-range dependencies using self-attention mechanisms) [37] to enhance global context modeling. The overall structure follows an encoder–bottleneck–decoder paradigm, where the encoder and decoder adopt standard U-Net modules, while the bottleneck is implemented using a Vision Transformer (ViT) module.

The encoder consists of four convolutional modules, each comprising two 3×3 convolutions (which applies filters to extract local spatial features) [36] followed by batch normalization and ReLU activation, with downsampling performed via 2×2 max-pooling layers. As the feature maps progress through the encoder, the number of channels increases from 64 to 512, while the spatial resolution is halved at each stage.

At the deepest level, the conventional bottleneck in U-Net is replaced by a Transformer module that operates on the lowest-resolution feature maps to capture long-range dependencies.

The decoder mirrors the encoder structure, using transposed convolution or upsampling blocks to progressively recover spatial resolution. Skip connections between corresponding encoder and decoder layers are used to retain fine-grained spatial information. A final 1×1 convolution is applied to generate a single-channel segmentation map with the same spatial dimensions as the input.

U-Net is widely recognized as a highly effective architecture for semantic segmentation tasks, particularly in medical imaging. Its success can be attributed to the following structural advantages:

Skip Connections. U-Net employs skip connections to directly transfer high-resolution features from shallow encoder layers to corresponding decoder layers. This

design:

- Preserves spatial localization and boundary information;
- Mitigates information loss caused by deep network downsampling;
- Enhances segmentation of fine structures (e.g., organ boundaries, small lesions).

Symmetric Encoder–Decoder Structure. U-Net adopts a symmetric encoder–decoder design:

- The encoder captures contextual semantics via progressive downsampling;
- The decoder restores spatial resolution via upsampling while fusing high-resolution cues from skip connections;
- The overall structure balances global context and local detail.

Fully Convolutional with Multi-Scale Context.

- The network avoids fully connected layers, preserving spatial correspondence;
- Pixel-wise predictions are determined by localized receptive fields;
- Multi-scale contextual features are naturally aggregated through the hierarchical convolutional structure.

Parameter Efficiency and Data Efficiency.

- Performs well on small datasets, a common scenario in medical imaging;
- The architecture is modular and scalable, enabling extensions such as 3D U-Net, attention U-Net, and Transformer-based variants.

U-Net as a Function Approximator. The segmentation task can be formalized as learning a function:

$$f : \mathbb{R}^{H \times W \times C} \to \mathbb{R}^{H \times W \times K}$$

Given an input image $x \in \mathbb{R}^{H\times W\times C}$, the network outputs a mask $y = f(x) \in [0,1]^{H\times W\times K}$, where K is the number of classes (e.g., $K = 2$ for binary segmentation). The function f is expected to:

- Predict a K-dimensional categorical distribution at each pixel location;
- Integrate global semantic context and local spatial details;
- Represent a complex, nonlinear mapping from image space to label space.

Skip Connections as Residual Augmentation. A decoding layer in U-Net can be expressed as:

$$y_{\text{up}} = \mathcal{F}(x_{\text{deep}}) + x_{\text{skip}}$$

Where:

- x_{deep} is the deep encoder feature;
- x_{skip} is the corresponding shallow feature from the encoder;
- $\mathcal{F}$ denotes the upsampling and convolution operation.

This structure resembles residual learning and facilitates gradient flow while preserving high-frequency details from the input.

Approximate Solution to Variational Models. In classical variational methods such as the Mumford–Shah and Potts models, segmentation is obtained by minimizing an energy functional of the form:

$$\min_u \int_\Omega \|\nabla u(x)\| + \lambda \cdot \|u(x) - f(x)\|^2 dx$$

Where:

- $u(x)$ is the predicted segmentation mask;
- The first term encourages boundary regularity;
- The second term enforces fidelity to the input image data $f(x)$.

Rather than explicitly minimizing the above energy, U-Net implicitly learns to do so through supervised training. It serves as an *indirect energy minimizer*, driven by data and network structure.

The ReLU activation function is used after the convolutional layer. ReLU (Rectified Linear Unit) is defined as follows:

$$\mathrm{ReLU}(x) = \max(0, x)$$

That is:

- If $x > 0$, the output is x
- If $x \leq 0$, the output is 0

In the network structure, ReLU is used as follows:

$$\mathbf{y} = \mathrm{ReLU}(\mathrm{BN}(W * \mathbf{x} + b))$$

Where:

- $\mathbf{x}$: input feature map
- W: convolution kernel weights
- b: bias term of the convolution layer
- $*$: convolution operation
- BN: Batch Normalization (standardizes intermediate outputs)
- ReLU: nonlinear activation

1. Nonlinear Mapping Capability

In the absence of nonlinear activation functions such as ReLU, a deep neural network reduces to a composition of linear transformations:

$$\mathbf{y} = W_n * (W_{n-1} * \cdots * W_1 * \mathbf{x})$$

This results in a globally linear mapping, which is inherently limited in its capacity to represent complex functions, such as classification boundaries or nonlinear spatial structures in images.

To address this limitation, ReLU activations are applied after each linear transformation:

$$\mathbf{y} = W_n * \mathrm{ReLU}(W_{n-1} * \mathrm{ReLU}(\cdots \mathrm{ReLU}(W_1 * \mathbf{x})))$$

Since ReLU outputs zero for non-positive inputs ($x \leq 0$), the network effectively becomes a piecewise linear and nonlinear function, thereby enabling it to approximate highly complex and non-linear decision surfaces.

2. Sparse Activation

Mathematically, ReLU is defined as:

$$\mathrm{ReLU}(x) = 0 \quad \text{for } x \leq 0$$

This property ensures that a neuron is activated only when its input is positive, leading to sparse activations. Such sparsity provides several benefits:

- Reduced computational overhead due to fewer active neurons
- Lower risk of overfitting, as only a subset of the network contributes to each forward and backward pass

When training neural networks using backpropagation, the derivative of the ReLU activation function is given by:

$$\frac{d}{dx}\mathrm{ReLU}(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{if } x \leq 0 \end{cases}$$

This implies:

- When a neuron is activated ($x > 0$), the gradient can propagate normally.
- When $x \leq 0$, the gradient becomes zero, rendering the corresponding neuron inactive.

This property provides a clear gradient flow path for active neurons, which facilitates efficient training. While ReLU is employed in hidden layers to enhance nonlinearity and learning capacity, the final network outputs still need to be normalized into probabilistic maps for segmentation. This is typically done using the Softmax function, which also provides a suitable representation for subsequent energy-based regularization.

Let $\mathbf{z} = [z_1, z_2, \ldots, z_C]$ denote the output logits of the network, where C is the number of classes (in our case, $C = 2$). The Softmax function is defined as:

$$\text{Softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}}, \quad i = 1, 2, \ldots, C$$

The Softmax function converts the raw, unnormalized output logits into a probability distribution over classes. It satisfies the following properties:

- Each output lies in the interval $(0, 1)$, i.e., $\text{Softmax}(z_i) \in (0, 1)$
- The sum of all outputs equals 1: $\sum_{i=1}^{C} \text{Softmax}(z_i) = 1$

Therefore, the network outputs can be interpreted as pixel-wise class probabilities, which constitute the foundation of standard supervised segmentation and are widely adopted in multi-class prediction tasks [38]. In conventional supervised learning, these probability maps are optimized against ground-truth annotations using loss functions such as cross-entropy loss or Dice loss. This paradigm has achieved considerable success in a wide range of image segmentation applications, particularly in medical image analysis [39] [40].

However, supervised objectives mainly encourage label-wise agreement between predictions and annotations, while they do not explicitly encode structural priors such as spatial smoothness, boundary regularity, or region consistency. The obtained segmentation may still be structurally inconsistent or less reliable in challenging scenarios.

At the same time, medical image segmentation continues to face several challenges. First, the availability of annotated medical data is often limited, as producing high-quality labels requires expert knowledge and is both time-consuming and costly. Second, medical imaging data exhibit substantial heterogeneity across scanners, acquisition

protocols, and clinical centers, leading to domain shift and reduced model generalizability in real-world settings. Third, many clinical targets, such as tumors or lesions, are small and irregular in shape, resulting in severe class imbalance and difficulty in accurately segmenting tiny or low-contrast structures. Finally, deep neural networks typically lack interpretability and robustness, making their behavior difficult to scrutinize and raising concerns for reliable deployment in safety-critical clinical environments. Addressing these challenges remains essential for developing segmentation systems suitable for widespread clinical adoption.

To address these limitations, the next chapter introduces an ADMM-based energy loss that incorporates additional structural constraints into the segmentation framework to complement conventional supervised learning.

Chapter 3

An ADMM-Regularized Semi-Supervised Framework for Image Segmentation

Although deep learning–based segmentation methods have demonstrated remarkable performance across various domains, they often suffer from limited generalization and weak interpretability, particularly when labeled data are scarce. Variational models, on the other hand, provide strong regularization mechanisms and incorporate spatial smoothness priors that improve segmentation stability. To address the limitations of each individual paradigm, recent research has increasingly focused on integrating deep learning with variational formulations, aiming to leverage their complementary strengths.

Deep networks excel at data-driven feature extraction and nonlinear representation learning, while variational models contribute explicit spatial regularization and theoretical interpretability. By coupling these two approaches, it becomes possible to achieve more robust and accurate segmentation, especially in semi-supervised settings where labeled data are limited and noise levels are high. Motivated by this insight, we propose an ADMM-regularized semi-supervised framework that unifies deep learning and variational optimization within a joint formulation.

3.1 Joint Model: Formulation and Motivation

In recent years, growing attention has been devoted to combining deep learning and variational models for image segmentation. This hybrid strategy aims to exploit the complementary strengths of both paradigms. Specifically, deep networks excel at data-driven feature extraction and nonlinear representation learning, while variational models contribute spatial regularization, smoothness constraints, and improved interpretability to the segmentation process. By coupling these two approaches, researchers seek to achieve more accurate and robust segmentation results, particularly in scenarios with limited labeled data and high noise levels, which are especially common in medical imaging. Liam et al. [2] proposed a hybrid segmentation pipeline for CT images that integrates a variational model with a deep learning algorithm. Their framework was primarily applied to stented aortic aneurysms, abdominal organs, and brain lesions. In their method, image registration, i.e., the process of estimating a spatial transformation to align different images within a common coordinate system, was used to replace manual labeling within the variational model, thereby enabling semi-automatic segmentation. The registered regions of interest (ROIs) were then aligned with new image volumes and input into a pre-trained convolutional neural network (CNN) to obtain refined segmentation results. The estimated masks generated by the variational model were also incorporated into the CNN training process, completing the network's backpropagation using both labeled and pseudo-labeled data. To optimize this integration, they designed a hybrid loss function that jointly considers labeled and unlabeled data [2]. Let θ denote the parameters of the segmentation network, which are optimized by minimizing the loss function, and let $u_\theta(x_i)$ represent the output of the network for the input image x_i. Here, $\mathrm{DICE}(\cdot,\cdot)$ denotes the Dice similarity coefficient between two segmentation maps. For two masks or probability maps a and b, it is defined as

$$\mathrm{DICE}(a,b) = \frac{2\sum_x a(x)b(x) + \mathrm{smooth}}{\sum_x a(x) + \sum_x b(x) + \mathrm{smooth}},$$

where $a(x)b(x)$ denotes the pointwise product at pixel x, so that $\sum_x a(x)b(x)$ measures the probabilistic overlap between the two segmentation maps, and $\mathrm{smooth} > 0$ is a

small constant introduced for numerical stability and to avoid division by zero. The loss function is defined as

$$\mathscr{L}(\theta) = \sum_{i=1}^{L} (1 - \mathrm{DICE}(GT_i,\, u_\theta(x_i))) + \xi \sum_{i=1}^{U} (1 - \mathrm{DICE}(v_i,\, u_\theta(x_i))) \tag{3.1}$$

where L and U denote the numbers of labeled and unlabeled images, respectively, and the index i refers to the i-th training sample. For labeled data, GT_i denotes the ground-truth segmentation mask of the i-th image, i.e., a pixel-wise annotation indicating the target region to be segmented, rather than the original image itself. The function $u_\theta(x_i)$ represents the predicted segmentation result of the network. For unlabeled data, v_i denotes the segmentation result obtained from the variational model in the previous step.

The pipeline uses a simple CNN network as shown in the Figure 3.1, to perform the segmentation task of the deep learning part. In addition, two structurally optimized networks are trained, called "LightNet" (LNet) and "SimplifiedNet" (SNet). They are used to compare the experimental results of the pipeline. The feature of LNet is that it uses fewer filters, while SNet reduces the steps of Max-pooling (a spatial downsampling operation that reduces resolution and may lead to loss of fine structural details) [1] and upsampling. The network structure is shown in the figure below:

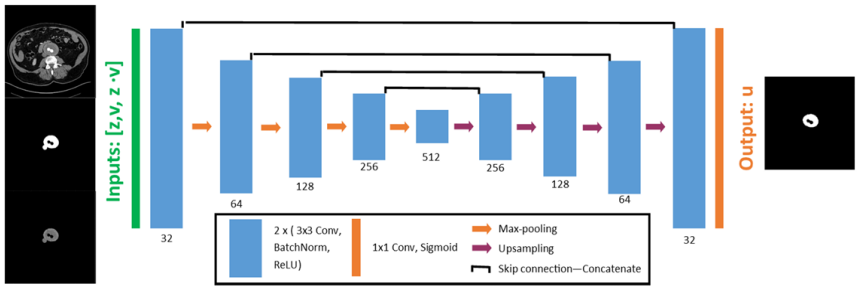


Figure 3.1: Segmentation network construction (Reproduced from Liam et al., [2]).

This pipeline selected two public sets of data, Brats and the Abdomen data from the Multi-Atlas Labeling Beyond the Cranial Vault challenge18. In addition, 70 CT images from patients who underwent endovascular aneurysm repair (EVAR) for abdominal aortic aneurysms (AAAs) at Siemens Healthcare, Frimley, UK, were included. Of these,

50 manually segmented images were used for training and 10 for testing. Unlabeled images were also used for CNN training.

Each of the three datasets is used to test the final model. The performance and accuracy are compared with other standard segmentation methods proposed recently. The DICE is used here to illustrate the comparison results. Experimental results show that Hybrid pipeline exhibits more accurate segmentation results. The paper by Liam et al. [2] presents a variational approach with CNNs algorithm to improve the automatic segmentation efficiency and accuracy of organ and tumor CT images based on a large volume datasets. Labeled images are not strongly relied on, and the number of training datasets is expanded. Liam et al. [2] also noted that future work would aim to extend the pipeline to a broader range of lesions rather than limiting it to the diseases considered in the current study. Overall, the results of their study provide a useful basis for further improvement and optimization of the pipeline.

Ke Wei and colleagues [12] proposed a new variational model based on the Potts energy formulation. To introduce the variational principle underlying this method, we first revisit the classical Potts model, a piecewise-constant segmentation framework in which each pixel is assigned to one of N regions and each region is represented by a binary indicator function. The Potts model balances a likelihood-based fitting term with a regularization term that penalizes discontinuities. Formally, the Potts energy is written as

$$E_{\text{Potts}}(\phi) = \sum_{k=1}^{N} \int_{\Omega} f_k(x)\, \phi_k(x)\, dx + \sum_{k=1}^{N} \int_{\Omega} \alpha(x)\, |\nabla \phi_k(x)|\, dx, \tag{3.2}$$

where $\phi_k(x) \in [0,1]$ denotes the membership function of pixel x to region k, with $\sum_{k=1}^{N} \phi_k(x) = 1$, and $f_k(x)$ denotes the region fitting term. The first term corresponds to data fidelity, and the second term is a spatial regularization term controlled by $\alpha(x)$.

To model region likelihoods, the probability of each pixel x belonging to region ω_k is defined as

$$p_k(x) = \frac{\exp\left(-|z(x) - c_k|/(2\sigma^2)\right)}{\sum_{\ell=1}^{N} \exp\left(-|z(x) - c_\ell|/(2\sigma^2)\right)}, \tag{3.3}$$

where $z(x)$ denotes the pixel intensity, c_k denotes the mean intensity of region k, and

$\sigma > 0$ is a scale parameter that controls the variance of the Gaussian likelihood. A smaller σ leads to sharper region assignments, while a larger σ results in smoother probabilities. This soft probabilistic assignment handles blurred boundaries and noise by avoiding a hard region partition.

Using the Bernoulli distribution, the likelihood of pixel assignment is expressed as

$$P(\phi_k(x_i)) = \big(p_k(x_i)\big)^{\phi_k(x_i)}\big(1 - p_k(x_i)\big)^{1-\phi_k(x_i)}, \tag{3.4}$$

where $\phi_k(x_i) \in \{0, 1\}$. Minimizing the negative log-likelihood yields

$$f_k(x) = -\log p_k(x). \tag{3.5}$$

The adaptive regularization weight is defined as

$$\alpha(x) = \frac{\beta}{1 + \gamma|\nabla I_\sigma(x)|^2}, \tag{3.6}$$

where $\beta > 0$ is a scaling parameter controlling the overall strength of the adaptive regularization weight, and $\gamma > 0$ determines its sensitivity to image gradients. Here, $|\nabla I_\sigma(x)|^2$ denotes the squared gradient magnitude of the smoothed input image, which is computed using the Sobel operator. As a result, $\alpha(x)$ becomes smaller near strong edges and remains relatively larger in homogeneous regions, which helps preserve edges while smoothing noise, following the principle of the Mumford–Shah model.

Given that each pixel label follows a Bernoulli distribution with success probability $p_k(x)$, the segmentation problem becomes

$$\min_{\phi_1(x_j)} \sum_{j=1}^{n^2} \left[f_1(x_j)\phi_1(x_j) + f_2(x_j)\big(1 - \phi_1(x_j)\big) \right], \qquad j = 1, \ldots, n^2. \tag{3.7}$$

Here, $\phi_1(x) = 1$ assigns pixel x to region 1 and $\phi_1(x) = 0$ assigns it to region 2. Substituting (3.5) into (3.2) yields the refined energy

$$E(\phi_i) = \int_\Omega \left[f_1(x)\phi_1(x) + f_2(x)\phi_2(x) \right] dx + \int_\Omega \alpha(x)|\nabla\phi_i(x)|dx, \tag{3.8}$$

which improves consistency between the model output and the original data, leading to more accurate and structurally coherent segmentation results.

To further enhance the Potts-based formulation, Wei et al. introduced a multiphase segmentation and semi-supervised clustering framework in which the region force is reformulated using the negative log-likelihood of the Bernoulli distribution. This modification improves the alignment between the model prediction and the observed data, thereby enhancing segmentation accuracy. The resulting energy is efficiently minimized using primal-dual (PDHG) and ADMM optimization schemes, which enable scalable optimization for both segmentation and clustering tasks. In the data clustering task, the developed Potts model can be expressed as:

$$\min_{\phi_k \in \{0,1\}} \sum_{k=1}^{K} \sum_{x_i \in V} f_k(x_i)\phi_k(x_i) + \sum_{k=1}^{K} \sum_{x_i \in V} \alpha(x_i)|\nabla \phi_k(x_i)|_1 \tag{3.9}$$

Here, let K denote the number of segmentation classes. For example, $K = 2$ corresponds to binary segmentation, while $K = 3$ corresponds to a three-class segmentation problem. And the gradient term constrains the boundaries between categories, ensuring that similar data points are grouped together. The term $f_k(x_i)$ serves as the region force term and adopts the form of the negative log-likelihood of a Bernoulli distribution.

At the moment, I am exploring deep learning-based approaches to solve the new variational model. Using the classical U-Net architecture with a symmetric encoder-decoder structure, a simple 3-phase classification task was designed with 3-channel outputs. The energy functional was used as a custom loss function, with the U-Net network's output serving as the input to the energy functional, dynamically computing the image regions. In this setup, the energy functional was applied to the optimization task following U-Net segmentation to smooth the segmentation boundaries. Simultaneously, the Dice loss was employed to enhance the distinction between categories, improving the robustness of the segmentation model and the quality of the boundaries. As defined earlier, the Dice loss used in this work is written as follows:

$$\text{Dice Loss} = 1 - \frac{2 \cdot \sum(\text{preds} \cdot \text{targets}) + \text{smooth}}{\sum \text{preds} + \sum \text{targets} + \text{smooth}} \tag{3.10}$$

Apply softmax to the output so that the probability sum of each pixel equals 1. The term preds represents the predicted probabilities, obtained from the model output using softmax. The term targets refers to the ground truth mask. The term smooth > 0 is a small constant added to both the numerator and denominator to improve numerical stability and avoid division by zero (e.g., smooth $= 10^{-6}$).

In the energy function, the class centers c_1, c_2, and c_3 are dynamically calculated based on the current segmentation probabilities.

$$c_k = \frac{\sum u_k \cdot z}{\sum u_k + \epsilon}$$

The class center values are calculated based on the pixel intensity z and the current segmentation probability u_k.The term ϵ is a smoothing factor used to avoid division by zero in the denominator. The regularization term smooths the boundaries through the total variation (TV) regularization, and its formula is given as:

$$\text{Reg Term} = \sum \alpha \cdot (\|\nabla u_1\|_1 + \|\nabla u_2\|_1) \tag{3.11}$$

The dynamic regularization weight α is computed from the gradient of the input image according to Eq. (3.6).

3.1.1 Energy Loss Design and Comparative Rationale

The total loss combines the Dice loss and energy loss. The weighting coefficients λ_{Dice} and λ_{Energy} are introduced to balance the contributions of the Dice term and the energy-based term.

Currently, the existing energy function is being adapted to process 3-phase images by classifying the mask into three categories and mapping them to the labels $[0, 1, 2]$. The specific formulation used in the experiment is as follows:

$$\mathcal{L}_{\text{energy}} = \lambda_{\text{data}} \sum (f_1 u_1 + f_2 u_2 + f_3(1 - u_1 - u_2)) + \lambda_{\text{reg}} \sum (\alpha \cdot |\nabla u_1| + \alpha \cdot |\nabla u_2|) \tag{3.12}$$

The corresponding data fidelity term is given by

$$\mathcal{L}_{\text{data}} = \sum \left(f_1 u_1 + f_2 u_2 + f_3 (1 - u_1 - u_2) \right)$$

where:

- u_1 and u_2 are the probability distributions output by the model, representing the likelihood of belonging to different categories.
- f_1, f_2, f_3 are the energy terms for different categories, computed based on z (input image) as follows:

$$f_i = -\log(p_i(x) + \epsilon)$$

Here, Likelihood of Pixel Values is used for computing the data term:

$$p_i(x) = \frac{\exp\left(-\frac{(z(x) - c_i)^2}{2\sigma^2} \right)}{\sum_{k=1}^{3} \exp\left(-\frac{(z(x) - c_k)^2}{2\sigma^2} \right)}, \quad i = 1, 2, 3.$$

Regional mean values is Representing the typical pixel values for each category:

$$c_1 = \frac{\sum u_1 \cdot z}{\sum u_1 + \epsilon}, \quad c_2 = \frac{\sum u_2 \cdot z}{\sum u_2 + \epsilon}, \quad c_3 = \frac{\sum (1 - u_1 - u_2) \cdot z}{\sum (1 - u_1 - u_2) + \epsilon}$$

During each iteration, the class centers c_k are dynamically updated as the weighted mean intensities of pixels assigned to the corresponding classes. The purpose of the regularization term is to maintain the smoothness of segmentation boundaries and avoid unnecessary noise. Here is the formula:

$$\mathcal{L}_{\text{reg}} = \sum \left(\alpha \cdot |\nabla u_1| + \alpha \cdot |\nabla u_2| \right)$$

The Gradient Calculation is:

$$|\nabla u_1| = |\partial_x u_1| + |\partial_y u_1|, \quad |\nabla u_2| = |\partial_x u_2| + |\partial_y u_2|,$$

$$\alpha = \frac{\beta}{1 + \gamma|\nabla z|^2}$$

α is Smoothness Factor. Where:

$$|\nabla z|^2 = (\partial_x z)^2 + (\partial_y z)^2$$

α controls the strength of the regularization and depends on the gradient information of the original input image z. In edge regions (where $|\nabla z|$ is large), α becomes smaller, reducing the influence of regularization and preserving the boundary.

Finally, the total loss is defined as:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{Dice}}\mathcal{L}_{\text{Dice}} + \lambda_{\text{Energy}}\mathcal{L}_{\text{Energy}}. \tag{3.13}$$

where $\mathcal{L}_{Dice}$ is the Dice loss defined in Eq. (3.10), and $\mathcal{L}_{Energy}$ denotes the energy-based loss derived from the variational model (see Eq. (3.12)).

3.1.2 Why Combine Supervised Loss with Energy Loss Function

The Potts model, originally proposed by R. B. Potts [11] in 1952 as a statistical physics model, describes local interactions that encourage similar neighboring elements to share the same state. In image segmentation, this idea is extended to promote spatial smoothness—assigning pixels with similar characteristics to the same region while maintaining sharp transitions at region boundaries.

In deep learning–based segmentation, supervised loss functions such as cross-entropy or Dice loss primarily optimize pixel-wise classification accuracy. However, they lack explicit spatial constraints, often leading to fragmented or noisy predictions, especially under limited labeled data or low-contrast conditions.

By contrast, the energy loss derived from the Potts model introduces spatial regularization that enforces consistency among neighboring pixels, thereby preserving structural coherence. Combining these two objectives allows the model to benefit from both data-driven feature learning and structure-aware regularization.

Following this principle, Ke Wei et al. [12] reformulated the Potts model in the

following variational form:

$$\min_{\{\Omega_k\}_{k=1}^K} \sum_{k=1}^{K} \int_{\Omega_k} f_k(x)\, dx + R\left(\{\Omega_k\}_{k=1}^K\right), \tag{3.14}$$

where $f_k(x)$ denotes the region force representing the energy cost for assigning pixel x to class k, and

$$R(\{\Omega_k\}) = \sum_k \int_{\partial\Omega_k} \alpha(x)\, ds$$

is the boundary regularization term that enforces smoothness along the class boundaries. Wei *et al.* introduced the following region force:

$$f_k(x) = -\log(p_k(x)) + \log(1 - p_k(x)),$$

where $p_k(x)$ is the probability that pixel x belongs to class k. This definition assumes that each pixel label $\phi_k(x)$ follows a Bernoulli distribution:

$$P(\phi_k(x)) = p_k(x)^{\phi_k(x)}(1 - p_k(x))^{1-\phi_k(x)},$$

which corresponds to the binary decision of whether pixel x belongs to class k.

The region force $f_k(x)$ thus measures the energy cost of assigning pixel x to class k. If the intensity or feature of a pixel follows a distribution similar to that of a certain class (e.g., tumor), $p_k(x)$ will be relatively large, leading to a smaller energy $f_k(x)$. During the energy minimization process, the model therefore tends to assign such pixels to the class with a lower energy, thereby reducing the overall Potts energy.

For brain tumor MR images, the model usually assumes that the intensity of each class follows a Gaussian distribution:

$$p_k(x) = \frac{\exp(-|I(x) - c_k|^2/2\sigma^2)}{\sum_{k'} \exp(-|I(x) - c_{k'}|^2/2\sigma^2)},$$

where c_k denotes the mean intensity of class k, and $\sigma > 0$ is the standard deviation parameter of the Gaussian distribution around the class mean. Consequently, a smaller σ makes $p_k(x)$ more sensitive to deviations from c_k, whereas a larger σ produces a

smoother likelihood distribution. When a voxel intensity $I(x)$ is close to the class mean c_k corresponding to the tumor region (denoted as c_{tumor}), the corresponding probability $p_k(x)$ (i.e., $p_{\text{tumor}}(x)$) increases, resulting in a smaller energy cost $f_k(x)$. Consequently, this voxel is more likely to be assigned to the tumor class during energy minimization.

Meanwhile, the boundary regularization term penalizes the gradient magnitude $|\nabla \phi_k|$, suppressing isolated misclassifications and maintaining the continuity of the tumor boundary.

In other words, regions whose intensity features are consistent with those of the tumor and exhibit coherent spatial structure are automatically identified as tumor regions, while inconsistent or irregular regions are classified as normal tissue. This mechanism enables the Potts energy to jointly capture both statistical discrimination and geometric regularization, which are essential for accurate and spatially consistent medical image segmentation.

3.2 ADMM-Based Optimization of the Energy Loss Function

Although the classical Potts model provides a principled formulation for enforcing spatial coherence, the original discrete optimization problem is not directly suitable as a loss function in deep neural networks. Specifically, the binary and piecewise-constant constraints make the optimization non-differentiable and incompatible with gradient-based training. To address this limitation, we adopt a relaxed continuous formulation of the Potts energy that can be optimized jointly with network parameters. This relaxation enables the model to encourage spatial coherence in the segmentation of unlabeled data while remaining compatible with back-propagation.

The relaxed energy functional over the predicted soft segmentation map $\phi(x) \in [0, 1]$ is defined as:

$$E(\phi) = \int_{\Omega} r(x)\phi(x)\, dx + \lambda \int_{\Omega} |\nabla \phi(x)|\, dx, \tag{3.15}$$

where the first term encourages consistency between $\phi(x)$ and the underlying image

intensity encoded by $r(x)$, and the second term promotes boundary smoothness through gradient regularization. The hyperparameter λ balances the two components.

To efficiently minimize this energy term, we adopt an existing optimization framework based on the *Augmented Lagrangian Method*, implemented using the *Alternating Direction Method of Multipliers (ADMM)*. This method originates from prior works on dual formulations of the Potts model and continuous max-flow optimization, and is integrated here as a modular component [12].

The ADMM module operates on auxiliary dual variables q_k, h_k, λ and Lagrange multipliers ϕ_k, and minimizes the following augmented Lagrangian:

$$\begin{aligned}\mathcal{L}(\lambda, [h_k], [q_k], [\phi_k]) = \int_\Omega \lambda \, dx + \sum_{k=1}^{K} \int_\Omega \phi_k (\operatorname{div} q_k - \lambda + h_k) \, dx \\ - \frac{c}{2} \sum_{k=1}^{K} \int_\Omega (\operatorname{div} q_k - \lambda + h_k)^2 \, dx.\end{aligned}$$

1. Dual Variable Updates

- **Update λ:**

$$\lambda^{l+1} = \frac{1}{K} \sum_{k=1}^{K} \left(\operatorname{div} q_k^l + h_k^l - \frac{\phi_k^l}{c} \right) + \frac{1}{Kc}$$

  Here, $c > 0$ is the penalty parameter associated with the augmented Lagrangian, which balances the constraint enforcement and affects the convergence behavior of the ADMM algorithm.

- **Update h_k:**

$$h_k^{l+1} = \min \left\{ \frac{\phi_k^l}{c} + \lambda^{l+1} - \operatorname{div} q_k^l, \, f_k \right\}$$

- **Update q_k via projected gradient descent:**

$$q_k^{l+1} = \Pi_{|q_k| \le \alpha(x)} \left(q_k^l + \beta \nabla \left(\operatorname{div} q_k^l - \lambda^{l+1} + h_k^{l+1} - \frac{\phi_k^l}{c} \right) \right)$$

The function $\alpha(x)$ denotes a spatially varying regularization weight, as defined in Eq. (3.11), which adapts to image gradients and controls the strength of the total

variation regularization. Here, ϕ_k^l denotes the current estimate of the segmentation variable for class k at iteration l in the ADMM optimization.

2. Lagrangian Multiplier Update

$$\phi_k^{l+1} = \phi_k^l - c\left(h_k^{l+1} + \operatorname{div} q_k^{l+1} - \lambda^{l+1}\right)$$

Integration of Softmax-Based Foreground Probability

Let the output of the segmentation network be a 4D tensor of shape $[B, 2, H, W]$, where each spatial location contains logits representing the likelihood of belonging to background or foreground.

Applying the Softmax function, which normalizes the output logits into class probabilities, yields a pixel-wise probability map:

$$u_k(x) = \frac{\exp(z_k(x))}{\exp(z_0(x)) + \exp(z_1(x))}, \quad k \in \{0, 1\}.$$

Thus, the network output can be written as

$$u(x) = \text{Softmax}(z(x)) = [u_0(x), u_1(x)],$$

where $z_k(x)$ denotes the logit corresponding to class k at pixel x, and $u_k(x) \in [0, 1]$ represents the predicted probability of pixel x belonging to class k.

In the binary segmentation setting, $u_0(x)$ and $u_1(x)$ correspond to the background and foreground probabilities, respectively. In particular, we denote the foreground probability as

$$u_{\text{fg}}(x) = u_1(x). \tag{3.16}$$

In our implementation, the foreground probability defined in Eq. (3.16) is used as the soft segmentation variable in the ADMM optimization module. Therefore, the Softmax probability map serves as the continuous optimization target for energy minimization:

$$\min_{u\in[0,1]} \int_{\Omega} \left(|\nabla u(x)| + \lambda\, r(x)u(x)\right)\, dx, \tag{3.17}$$

where:

- $u(x) \in [0, 1]$ is the foreground probability,
- ∇u denotes the gradient, encouraging boundary smoothness,
- $r(x)$ is the region force term encoding the data fidelity.

This formulation provides a bridge between the probabilistic CNN output and the continuous variational optimization process, allowing the foreground probability map to be refined through ADMM-based energy minimization.

Final Loss Integration

The classical Potts model provides a spatial regularization prior but cannot be directly used as a loss due to its discrete, non-differentiable form. By adopting the continuous relaxation described above, we obtain an energy functional that can be optimized numerically. However, direct gradient-based minimization of this energy remains challenging; therefore, we employ a fixed number of ADMM iterations to approximate its minimization.

At each training step, the network prediction is treated as the initial soft segmentation map, and ADMM is applied to refine it. The resulting energy value $\mathcal{E}(u)$ serves as a differentiable regularization term, encouraging spatial coherence.

The total training objective is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sup}}(\phi(\theta), y) + \alpha \cdot \mathcal{E}(u), \tag{3.18}$$

where $\phi(\theta)$ denotes the segmentation output of the network parameterized by θ, and u denotes the soft segmentation variable initialized from $\phi(\theta)$ and refined through the ADMM optimization module.. $\mathcal{L}_{\text{sup}}$ is the supervised loss computed on labeled data, and $\mathcal{E}(u)$ is the energy term minimized via the ADMM optimization module. The weight α controls the regularization strength.

By combining the ADMM-optimized Potts energy with supervised loss, our framework achieves improved segmentation accuracy and spatial smoothness, especially under limited annotation settings. This hybrid formulation effectively couples supervised learning and unsupervised variational regularization, allowing the network to leverage both annotated and unlabeled data. The ADMM-optimized energy term enforces spatial coherence and anatomical consistency.

3.3 Implementation Details and Inference

This section presents the implementation details and inference procedure of the proposed ADMM-regularized semi-supervised segmentation framework. Building upon the joint model introduced in the previous section, the optimization is formulated based on the Potts energy model, which enforces both regional consistency and boundary smoothness. Following the formulation in Wei et al. [12], the Potts energy is expressed as:

$$E_{\text{Potts}}(\phi) = \sum_{k=1}^{N} \int_{\Omega} f_k \phi_k \, dx + \sum_{k=1}^{N} \int_{\Omega} \alpha(x) |\nabla \phi_k| \, dx \tag{3.19}$$

The first term is the region force. The second term is the regularization term. This type of energy model achieves unsupervised or semi-supervised segmentation by minimizing the Potts energy, thereby promoting regional consistency and boundary smoothness. The Potts energy model, based on the number of regions N in Eq. (3.13), can be used for classification or multi-phase segmentation problems. When N=2, the target region is divided into foreground and background, and the model is a two-phase segmentation task. In this case, the energy function can be solved using convex relaxation or ADMM methods for efficient computation. When $N > 2$, the segmentation task involves multiple regions, and the model needs to process multiple region functions $\{\phi_k\}_{k=1}^{N}$ simultaneously, and satisfy the mutual exclusivity constraint $\sum_k \phi_k(x) = 1$. The multi-phase segmentation problem then becomes a non-convex optimization problem, and the solution is more complex than the two-phase problem. The coupling of different regions leads to a high number of local minima in the energy function. Fur-

thermore, medical images can exhibit weak contrast and blurred boundaries between different tissues and lesions. These issues make it difficult to find a unified energy weighting between the region force and boundary regularization, ensuring good performance in all region categories and smooth boundaries. This manifests itself in blurred segmentation boundaries and even inaccurate segmentation results. Multi-phase segmentation tasks still faces significant challenges.

3.3.1 Two phase Segmentation

To simplify the implementation and validate the feasibility of the proposed loss formulation, the segmentation problem is first considered in a two-phase setting, which represents a special case of the multi-phase model. In this case, the image domain consists of two regions, foreground and background. The function $\phi(x)$ is the segmentation function to be optimized, representing the assignment of each pixel to either the foreground or background. Specifically, $\phi(x) = 1$ indicates that pixel x belongs to the foreground, while $\phi(x) = 0$ indicates that it belongs to the background. The corresponding energy formula can be written as:

$$E_{\text{two-phase}}(\phi) = \int_{\Omega} f_1(x)\phi(x)\,dx + \int_{\Omega} f_0(x)\big(1-\phi(x)\big)\,dx + \int_{\Omega} \alpha(x)\,|\nabla\phi(x)|\,dx \quad (3.20)$$

Here, $f_1(x)$ and $f_0(x)$ denote the region forces for the foreground and background, respectively, as defined in Eq. (3.3), where each term penalizes assigning a pixel to a region inconsistent with the underlying likelihood. $\alpha(x)$ is the boundary regularization weight used to constrain the smoothness of the boundary. The first two terms in the formula jointly determine the region to which a pixel belongs. The model determines the region to which it belongs based on the pixel's grayscale value and feature information, achieving regional consistency. The third term is used to smooth the segmentation boundary and suppress noise. The two-phase model involves only one continuous variable, ϕ, making the optimization problem relatively simple and capable of being efficiently solved using the ADMM method proposed by Ke Wei et al [12].

In their method, the segmentation boundary can also be formulated as a closed curve that minimizes the overall energy. Its geometric significance lies in finding the optimal boundary between foreground and background within the image region. In the proposed framework, the Potts model serves as a loss function, integrated with the training process of the deep learning network. This allows for the introduction of spatial prior constraints on the supervised learning process, achieving structural constraints on the foreground region. The output of the Potts energy loss not only provides a mask reference for subsequent training but also acts as a regularizer, further enhancing the spatial consistency of network learning.

3.3.2 Multi-phase Segmentation Problem

The multi-phase segmentation problem aims to partition an image into several mutually exclusive regions, each representing a distinct phase or semantic class. Unlike binary segmentation, which only separates the foreground from the background, multi-phase segmentation further distinguishes multiple regions within the image domain. This formulation naturally arises in various applications, including medical imaging, material analysis, and remote sensing, where different regions correspond to distinct tissue types, material phases, or land-cover classes. Mathematically, the multi-phase segmentation problem can be formulated as a multi-region energy minimization task, where the objective is to partition the image domain into multiple disjoint subregions that are spatially coherent and consistent with image intensity or texture features.

These subregions are spatially correlated and often share adjacent boundaries, meaning their interactions cannot be modeled merely by pixel-wise classification or intensity thresholding. Instead, such problems can be formulated as a *multi-region energy minimization* task, where the goal is to partition the image domain $\Omega \subset \mathbb{R}^n$ into K disjoint regions $\{\Omega_k\}_{k=1}^K$, satisfying

$$\Omega = \bigcup_{k=1}^{K} \Omega_k, \quad \Omega_i \cap \Omega_j = \emptyset, \text{ for } i \neq j.$$

The overall energy functional can then be expressed as

$$E(\{\Omega_k\}) = \sum_{k=1}^{K} \int_{\Omega_k} f_k(x)\, dx + \lambda \sum_{k=1}^{K} \mathrm{Per}(\Omega_k),$$

where $f_k(x)$ represents the regional fitting term (for instance, an intensity-based likelihood or probability map), and $\mathrm{Per}(\Omega_k)$ denotes the perimeter length of each region, acting as a regularization term to balance data fidelity and spatial smoothness. Minimizing the perimeter term encourages smooth and regular boundaries, penalizes irregular or fragmented regions, and suppresses small isolated components caused by noise, thereby promoting spatial coherence in the segmentation.

This formulation represents the fundamental form of the *multi-phase Potts model*. In traditional variational frameworks, the Potts model is often expressed using multiple level-set functions or label fields, such as in the extended formulations of the Chan–Vese or Boykov–Jolly models.

However, as the number of segmentation classes increases, the inter-class coupling within the energy terms becomes stronger. The model must maintain *mutual exclusivity* between regions while ensuring *boundary consistency*, making the global optimization problem increasingly difficult. When applied to neural networks, this constraint becomes even more challenging, as each class channel must maintain balanced smoothness and separation simultaneously.

Directly optimizing a multi-phase Potts model over multiple channels leads to high computational costs and potential convergence issues, as the energy minimization process may become unstable or fail to converge.

For the ADMM optimization framework, multi-phase segmentation implies that multiple flux variables q_k and Lagrange multipliers λ_k must be updated simultaneously, with the energy terms of all classes being interdependent. These variables must satisfy the following constraints:

$$\sum_{k=1}^{K} u_k(x) = 1, \quad u_k(x) \geq 0,$$

where $u_k(x)$ represents the probability map of class k.

This constraint leads to an optimization space whose dimensionality increases linearly with the number of classes, while the coupling between the energy terms introduces strong nonlinearity. Furthermore, in the ADMM algorithm, the splitting variable updates require handling multiple gradient–divergence operators for all channels, resulting in a significant increase in computational cost and memory usage per iteration.

Consequently, directly constructing a multi-phase ADMM optimization model within the current neural network training framework becomes highly challenging to achieve stable convergence.

To address the aforementioned challenges, this study proposes a two-phase hierarchical segmentation strategy. In many medical imaging tasks, pathological regions exhibit a nested structure — for example, brain tumors typically consist of an outer edema region that surrounds the inner cores of the tumor. Performing fine-grained segmentation directly on the entire image may lead to ambiguous boundaries and increased misclassification due to interference from normal tissues. Therefore, we first perform a coarse global localization to detect the overall tumor region, followed by a refined segmentation focused within this region. This progressive scheme enables more reliable regional partitioning and improves boundary discrimination, while maintaining a unified network architecture (U-Net + ADMM Energy).

Phase 1 — Global Tumor Detection

In the first phase, the model performs a binary classification task on multi-modal MRI inputs (T1/T2/FLAIR) to distinguish the "tumor region" from the "background." The energy optimization function for this stage is defined as:

$$\mathcal{L}_1 = \beta \cdot \mathcal{L}_{\text{Dice}}(u, y) + \alpha \cdot \mathcal{L}_{\text{ADMM}}(u),$$

where $\mathcal{L}_{\text{ADMM}}$ enforces global structural smoothness and boundary continuity within the overall tumor region. The resulting mask M_{tumor} serves as a coarse localization map indicating the tumor's global spatial extent.

Phase 2 — Sub-region Refinement

In the second phase, the tumor mask from Phase 1 is projected back onto the original MRI, and the masked region is cropped or extracted to obtain the local subregion image I_{sub}. The same network structure and loss formulation (Dice + ADMM Energy) are then applied to refine the segmentation within the cropped area, focusing on the tumor's internal substructures such as necrotic core, enhancing region, and edema. The energy optimization objective for this phase is more locally constrained and defined as:

$$\mathcal{L}_2 = \beta' \cdot \mathcal{L}_{\text{Dice}}(u_{\text{sub}}, y_{\text{sub}}) + \alpha' \cdot \mathcal{L}_{\text{ADMM}}(u_{\text{sub}}),$$

where $\mathcal{L}_{\text{ADMM}}$ imposes stronger boundary-consistency and smoothness constraints within the subregion, ensuring precise structural refinement at the local level.

From the perspective of energy optimization, the proposed two-phase framework can be interpreted as a form of *hierarchical energy decomposition*. If we define the complete multi-phase energy as:

$$E_{\text{multi}} = \sum_{k=1}^{K} \mathcal{E}(u_k) + \gamma \sum_{i \neq j} \text{Reg}(u_i, u_j),$$

where $\mathcal{E}(u_k)$ denotes the single-phase energy term and $\text{Reg}(u_i, u_j)$ represents the pairwise regularization between different classes. For instance, a typical single-phase energy can be written as

$$\mathcal{E}(u_k) = \int_\Omega f_k(x)\, u_k(x)\, dx + \lambda \int_\Omega |\nabla u_k(x)|\, dx, \tag{3.21}$$

where $f_k(x)$ is the region fidelity term defined previously (see Eq. (3.5)), derived from the input image intensity $I(x)$ via a likelihood model, which measures how well pixel x fits region k. And the second term enforces spatial smoothness via total-variation regularization. The inter-class regularization term $\text{Reg}(u_i, u_j)$ can take the form

$$Reg(u_i, u_j) = \int_\Omega u_i(x) u_j(x)\, dx \tag{3.22}$$

which penalizes overlap between different phase indicators to encourage mutually ex-

clusive region assignments.

The proposed two-phase strategy can be regarded as an approximate decomposition of this energy formulation. Specifically, the model first optimizes the global structural energy term $\sum_k \mathcal{E}(u_k)$ (corresponding to Phase 1: global tumor detection), and then minimizes the inter-class regularization terms $\mathrm{Reg}(u_i, u_j)$ within the subspace (corresponding to Phase 2: sub-region refinement).

This hierarchical optimization scheme effectively reduces the complexity of energy coupling while maintaining convergence stability and structural consistency during training.

Compared with end-to-end methods that directly solve the multi-channel Potts model, the proposed two-phase framework exhibits clear advantages in terms of computational efficiency and numerical stability. By isolating the foreground region in the first phase, the optimization domain in the second phase is significantly reduced, allowing the energy gradients to be concentrated within the region of interest (ROI). This enhances the effectiveness of structural regularization and accelerates convergence during training. Moreover, since both phases adopt the same network architecture and loss configuration, the framework maintains a modular design that facilitates extension to additional classes or other organ segmentation tasks. In summary, this section introduced our two-phase hierarchical segmentation framework and detailed how the proposed ADMM-regularized energy formulation is integrated with a deep network to enhance spatial coherence and reduce inter-class interference. The design maintains modularity and reusability, making the framework extensible to other segmentation tasks.

In the following section, we empirically evaluate the proposed approach on multimodal brain tumor datasets, demonstrating its effectiveness compared to state-of-the-art baselines and analyzing the contribution of each component.

Chapter 4

Experiments

This section presents the experimental evaluation of the proposed ADMM-regularized semi-supervised segmentation framework. The experiments are designed to validate the effectiveness of the proposed loss formulation, demonstrate the contribution of ADMM-based energy regularization, and assess the model's generalization capability on both labeled and unlabeled data.

To provide a comprehensive evaluation, we compare four representative approaches: (1) the traditional variational model Chan–Vese, (2) a fully supervised segmentation method, (3) the proposed semi-supervised framework, and (4) the ADMM-regularized semi-supervised model. These comparisons allow us to analyze the impact of both the supervision strategy and the ADMM optimization on segmentation performance and boundary consistency.

The experiments are primarily conducted on the BraTS 2021 dataset, which contains multimodal MRI scans of brain tumors, while the OASIS-1 dataset is employed as a supplementary testing dataset to evaluate the generalization ability of the proposed framework on healthy brain structures.

4.1 Experimental Setup

The experimental setup is designed to validate the effectiveness of the proposed ADMM-based energy regularization in enhancing segmentation consistency and boundary smooth-

ness under limited labeled data. All experiments are conducted under a semi-supervised learning paradigm, combining both labeled and unlabeled MRI samples.

In the following subsections, we first describe the datasets used for training and testing, then detail the preprocessing and data augmentation strategies, and finally outline the implementation details, including the network configuration and optimization settings.

4.1.1 Datasets Description

This study used two publicly available brain MRI datasets: the BraTS 2021 dataset [41] and the OASIS-1 (Open Access Series of Imaging Studies) dataset [42]. These two datasets provide MRI images of pathological (brain tumors) and healthy brain structures, respectively, and can be used together to evaluate the performance of the model in different scenarios. Among them, the BraTS 2021 dataset is used for labeled tumor segmentation training and validation, as well as energy constraints for unlabeled tumor samples to improve the generalization ability and structural consistency of the model, while the OASIS-1 dataset serves as a supplement to the experimental results for testing the robustness of the model. The BraTS 2021 dataset was jointly released by the American Society of Radiology (RSNA), the North American Society of Neuroradiology (ASNR), and the International Society for Medical Image Computing and Computer-Assisted Intervention (MICCAI) [43]. It is currently one of the largest and most well-annotated brain tumor segmentation benchmarks. The dataset contains preoperative multimodal MRI scans of approximately 2,040 adult glioma patients, with each subject including four modalities: T1, T1-enhanced (T1Gd), T2, and T2-FLAIR. In the present study, only three modalities, namely T1, T2, and T2-FLAIR, were used as network inputs. These modalities were stacked channel-wise to form a three-channel input tensor. This was a practical implementation choice made to match the three-channel input setting of the proposed framework. Since the four modalities provide partially overlapping visual contrast information, the T1-enhanced modality was not included in the present implementation. This choice was made for implementation consistency rather than based on a dedicated comparative study. All data underwent a

unified preprocessing pipeline, including skull removal, registration to the SRI24 standard space, resampling to 1 mm^3 isotropic voxels, and conversion to NIfTI format. The BraTS 2021 dataset contains two tasks: (1) brain tumor segmentation task (Task 1), and (2) MGMT promoter methylation status prediction task (Task 2). This study only uses Task 1, which contains expert-manually annotated tumor region masks. Its labels define three main tumor subregions: ET (Enhancing Tumor): enhancing tumor region; NCR/NET (Necrotic/Non-Enhancing Tumor Core): necrotic or non-enhancing core; ED (Peritumoral Edema): tumor-associated edema region [43]. OASIS-1 (Open Access Series of Imaging Studies I) was released by Washington University School of Medicine in St. Louis to promote open research in brain imaging. This dataset contains T1-weighted structural MRI scans from 416 subjects (aged 18–96 years), including over 100 individuals with Alzheimer's disease and the remainder from healthy individuals. Each subject provides three to four repeatedly acquired high-resolution T1 images to improve signal-to-noise ratio and data consistency. The OASIS data has been processed for skull removal, registration, and intensity normalization, resulting in high image quality. In this study, we selected only T1 MRI images from OASIS-1 that were free of pathological changes (i.e., healthy individuals) as the dataset for evaluating model generalization during the experimental process [44]. These images were not used for actual semi-supervised training with or without labeled samples. These images provide rich prior information about normal brain structure, further demonstrating that the joint loss function helps improve the segmentation model's robustness in identifying tumor regions [44].

In this experiment, multi-modal medical images from the BraTS 2021 dataset are used as input. The three MRI modalities, T1, T2, and FLAIR are concatenated along the channel dimension to form a three-channel input, allowing the model to learn complementary information across modalities. The overall data construction process is as follows:

1. **Multi-modal concatenation principle:** Each sample contains three grayscale images corresponding to the T1, T2, and FLAIR modalities. During preprocessing, the three images are concatenated along the channel dimension to form a

tensor of shape $[3, H, W]$. This three-channel image is then fed into the network. This method enables the model to jointly utilize tissue information from multiple modalities, improving the accuracy of lesion or tumor segmentation and increasing data diversity.

2. **Data loading and label reading:** A dataset class is implemented to load the three modalities and their corresponding segmentation masks. Each sample directory contains T1, T2, and FLAIR images, as well as a label map. The label map uses pixel values of 0–127 to indicate tumor regions, with other regions representing the background.

3. **Image resizing:** All three modalities and the label map are loaded using the Python Imaging Library (PIL), a widely used Python package for image processing tasks [45], and resized to a consistent spatial size (e.g., 256×256), resulting in an input tensor of shape $[3, 256, 256]$.

Before training, a portion of the dataset is randomly selected as the validation set. This setup allows for flexible model tuning and testing, improving the overall generalization performance of the model.

4.1.2 Reprocessing and Augmentation

During the training process of this study, all MRI multi-modal images were uniformly preprocessed and standardized before being fed into the network. The preprocessing ensured consistent spatial dimensions, intensity ranges, and label formats across samples. The complete data preprocessing workflow includes modality concatenation, spatial resizing, intensity normalization, label remapping, and tensor conversion, as detailed below:

First, each sample contains three MRI modalities: T1, T2, and FLAIR. During data loading, each modality's grayscale image is read separately. Then, the three modality images are concatenated along the channel dimension to form a three-channel input tensor for the network, enabling the model to learn cross-modal complementary features. The final input tensor for each sample is of shape $[3, 256, 256]$. To ensure

consistent spatial dimensions, all images are resized using bilinear interpolation (for MRI images) to maintain structural smoothness while avoiding edge artifacts, and nearest-neighbor interpolation (for label maps) to preserve discrete boundaries. This guarantees that input dimensions remain identical between training and validation phases.

During the intensity normalization stage, the mean and standard deviation of voxel intensities are computed separately for each modality to ensure consistent normalization across channels. Given an input image I, the normalized image I' is computed as:

$$I' = \frac{I - \mu}{\sigma + 10^{-8}},$$

where μ and σ denote the mean and standard deviation of voxel intensities for that modality. Normalization is computed only on nonzero regions to reduce the statistical influence of background areas. This process helps mitigate intensity bias caused by differences in scanners, acquisition parameters, or subjects, thus improving the stability of multi-modal feature learning.

For the segmentation mask, the original data represent different tissue regions using various grayscale values. To simplify the task into a binary segmentation (tumor vs. background), a fixed mapping rule is applied:

$$\{0 \to 0,\ 64 \to 1,\ 127 \to 1,\ 255 \to 1\},$$

meaning all tumor-related regions (necrotic, enhancing, and edema areas) are unified into the foreground class, while non-tumor regions are mapped to the background. After remapping, the mask is converted into an integer tensor for subsequent Dice loss computation and foreground probability extraction.

During the tensorization stage, all images are converted into PyTorch tensors using torchvision.transforms.ToTensor(), a standard transformation provided by the PyTorch library [45] that normalizes image intensity values to the range [0,1]. This normalization ensures consistent input scaling and improves numerical stability during network training. Finally, each preprocessed image and its corresponding label mask form a

complete input pair.

Additionally, random rotation, flipping, and brightness adjustment are introduced as data augmentation, while preserving spatial consistency across modalities to ensure stability of boundary regularization and model convergence. This controlled augmentation improves generalization without relying on large-scale random transformations.

In summary, the complete data preprocessing and loading pipeline consists of the following key steps:

1. Modality loading and concatenation: Read T1, T2, and FLAIR modalities and concatenate them along the channel dimension.

2. Spatial normalization: Resize images and masks to 256×256 using bilinear and nearest-neighbor interpolation.

3. Intensity normalization: Compute mean and standard deviation per modality and perform z-score normalization.

4. Label remapping: Convert multi-class labels into binary classes (foreground vs. background).

5. Tensor conversion: Transform images and masks into PyTorch tensors for model input.

This preprocessing pipeline preserves fine anatomical details and tumor boundaries while maintaining modality consistency, providing a stable and unified data foundation for subsequent U-Net and ADMM-based segmentation network training.

4.1.3 Implementation Deatails

This study is implemented in the PyTorch framework environment, which mainly includes model architecture design, loss function setup, optimization strategy, and training process control.

1. Network architecture

The experiment adopts an improved U-Net architecture, TransUNet, as the main segmentation network. This model combines a convolutional encoder with a Trans-

former bottleneck module, allowing it to capture both local spatial texture features and global contextual dependencies. An overview of the proposed TransUNet architecture with the integrated ADMM energy loss is illustrated in Figure 4.1.

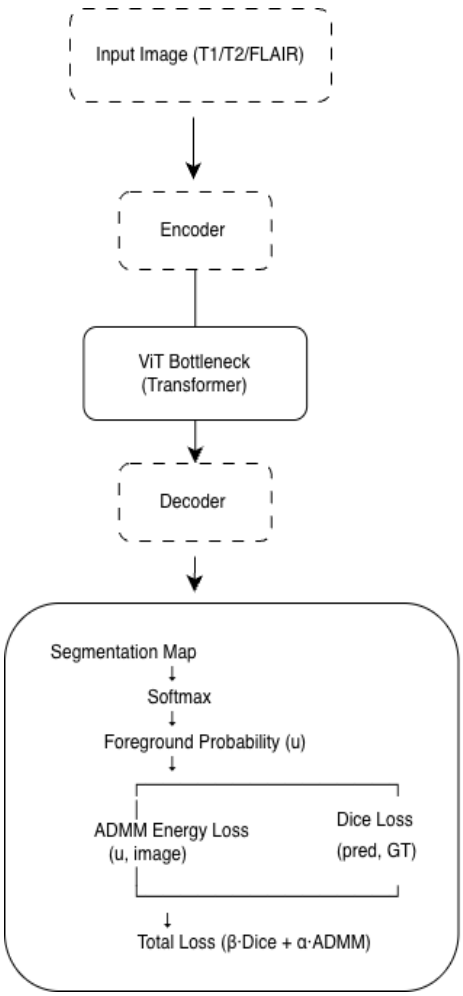


Figure 4.1: Network Architecture.

The encoder part consists of four ConvBlocks, each composed of two 3×3 convolution layers, Batch Normalization, and ReLU activation. The channel dimensions are progressively increased to 64, 128, 256, and 512, respectively. Downsampling is achieved via max pooling operations.

In the bottleneck layer of U-Net, a lightweight Vision Transformer (ViT) Bottleneck is embedded. This module performs linear patch embedding, passes through Transformer Encoder layers, and applies patch projection to re-encode convolutional

features into global representations, thereby enhancing the model's semantic aggregation capability.

The decoder part adopts hierarchical UpBlocks, each consisting of up-convolutions with skip connections to gradually reconstruct spatial details. Finally, a 1×1 convolution is applied to output a binary segmentation map (foreground vs. background).

2. Loss function and energy regularization term

During training, the total loss combines Dice Loss and ADMM Energy Loss, forming a joint optimization strategy. Dice Loss directly measures the overlap between the predicted segmentation and the ground truth, optimizing the model's pixel-wise segmentation accuracy. The ADMM Energy Loss is derived from the Potts model's energy minimization principle, introducing spatial gradient and boundary constraints based on the foreground probability map. This term promotes intra-region smoothness and inter-region boundary consistency.

The total loss is defined as:

$$L_{\text{total}} = \beta \cdot L_{\text{Dice}} + \alpha \cdot L_{\text{ADMM}},$$

where α and β represent the weighting coefficients for the ADMM Energy Loss and Dice Loss, respectively. During training, α is gradually increases from 0.1 to 0.3 while $\beta = 1 - \alpha$, meaning that Dice Loss dominates during early training, whereas the energy regularization term progressively enhances structure consistency and boundary stability in later stages. The coefficients α and β are adjustable hyperparameters rather than fixed constants.

3. Optimization and training strategy

The model is trained for 150 epochs with a batch size of 4. The optimizer is initialized with Adam (Adaptive Moment Estimation), a widely used stochastic gradient-based optimization algorithm provided by the PyTorch library [46], with a learning rate of 1×10^{-4}. When the validation Dice score plateaus for 5 consecutive epochs, the optimizer switches to SGD (learning rate 5×10^{-3}, momentum 0.9) to improve generalization stability.

A ReduceLROnPlateau learning rate scheduler (provided by the PyTorch library)

is employed, which automatically reduces the learning rate when the validation loss stops improving for a predefined number of epochs. For example, reducing the learning rate by a factor of 0.5 when no improvement is observed for 5 consecutive epochs. To prevent overfitting, an Early Stopping mechanism halts training if validation loss fails to improve for 10 consecutive epochs. The best model checkpoint (with the lowest validation loss) is saved for evaluation.

4. Hardware and implementation environment

All experiments are conducted on NVIDIA GPUs using the PyTorch framework. Training and validation are performed under identical hardware conditions to ensure result reproducibility. Throughout the process, loss variation curves are recorded, and both validation and test sets are evaluated visually using metrics including Accuracy and Dice Score. This verifies the model's spatial consistency and energy optimization effectiveness.

4.2 Compared Methods

To comprehensively evaluate the effectiveness of the proposed ADMM-regularized semi-supervised framework, several representative segmentation methods are selected for comparison. These methods include traditional variational models, fully supervised deep learning approaches, and the proposed semi-supervised variants.

Specifically, the classical Chan–Vese model is adopted as the representative of traditional variational segmentation methods, which rely on intensity homogeneity assumptions and level set evolution. For learning-based baselines, we include a supervised U-Net trained solely on labeled data, a semi-supervised U-Net incorporating unlabeled samples without ADMM regularization, and finally, our ADMM-regularized semi-supervised model that jointly optimizes the deep feature representation and energy regularization terms.

This comparative setup allows us to analyze the contribution of each component, ranging from purely variational formulations to fully data-driven learning, and to evaluate how the ADMM regularization enhances segmentation consistency and generalization performance.

Before detailing each method, Table 4.1 summarizes the overall quantitative comparison results on the BraTS 2021 dataset. Table 4.1 reports the quantitative results computed on the test set. After training, the model is applied to unseen test images to generate predicted segmentation results, which are then compared with the corresponding ground truth masks at the pixel level to compute the evaluation metrics. No training or parameter tuning is performed on the test set. These metrics respectively measure region overlap, overall segmentation accuracy, and boundary precision.

Model performance is evaluated using two metrics: Dice coefficient (Dice), Accuracy (Acc). The Dice and accuracy scores vary across test images due to inherent differences in image characteristics such as quality, contrast, structural complexity, and segmentation difficulty. The final results are reported as the mean Dice coefficient across all test images, providing a more objective and comprehensive evaluation of overall model performance. Due to the high instability of the semi-supervised model without ADMM optimization and its inability to produce reliable Dice scores, including these results in quantitative comparison is deemed inappropriate. Therefore, they are excluded from Table 4.1. Instead, only qualitative visualizations are provided to illustrate the necessity of ADMM-based structural regularization.

Specifically, the Dice coefficient measures the overlap between the prediction result P and ground truth G, and is defined as

$$\text{Dice} = \frac{2|P \cap G|}{|P| + |G|}$$

where $|P|$ and $|G|$ denote the sizes of the predicted and ground-truth foreground regions, respectively, and $|P \cap G|$ denotes the size of their overlap.

Accuracy (Acc) is computed at the pixel level as the proportion of correctly classified pixels among all pixels in the image. It is defined as

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN},$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively, with respect to the foreground class. However, in medical

image segmentation, Accuracy may be less informative when the foreground region is much smaller than the background, since the background class can dominate the overall score.

Table 4.1: Quantitative comparison of different segmentation models on the BraTS 2021 dataset. The best results are highlighted in bold.

Model	Acc	Dice ↑	Description
2-phase ADMM (Semi-supervised)	**0.992**	**0.958**	Achieves the highest boundary precision due to spatial regularization.
3-phase ADMM (Semi-supervised)	0.984	0.595	Shows unstable performance and class confusion in multi-class settings.
Two-step 3-phase ADMM (Semi-supervised, Step 1)	0.992	0.958	Initial tumor contour segmentation.
Two-step 3-phase ADMM (Semi-supervised, Step 2)	0.993	0.877	Provides stable segmentation and better generalization with hierarchical optimization.
2-phase U-Net (Supervised)	0.990	0.947	Baseline model.

As shown in the table, the 2-phase ADMM semi-supervised model outperforms the supervised baseline (U-Net) across all evaluation metrics. Compared with U-Net, the Dice coefficient increases from 0.947 to 0.958 (an improvement of approximately 1%). These results indicate that the ADMM energy regularization effectively constrains boundary morphology during optimization, enabling the model to produce smoother and more structurally consistent tumor regions during prediction.

In contrast, the 3-phase ADMM model exhibits a clear performance degradation in multi-class segmentation tasks, with the Dice dropping to 0.595. This degradation mainly arises from inter-class competition among regional forces under multi-class energy constraints, leading to boundary blurring and spatial instability, which significantly reduces overall performance.

The Two-step 3-phase ADMM model mitigates inter-class interference by decoupling global tumor detection from sub-region classification. The Dice score in Step 1 is 0.958, which is identical to that of the 2-phase ADMM model, while the Dice score in Step 2 is 0.877. Although the Step 2 Dice score is slightly lower, even falling below 0.9, the segmentation results remain more stable than those of the 3-phase ADMM model,

demonstrating the effectiveness of the two-step segmentation approach.

In summary, the ADMM energy regularization significantly improves boundary consistency and segmentation accuracy in binary scenarios, while the introduction of a stepwise structure further enhances prediction stability and generalization in complex multi-class tasks.

4.2.1 Traditional Variational Method-The Chan-Vese

In traditional image segmentation methods, variational models have been widely applied due to their strong theoretical foundations and flexibility. Among them, the Chan–Vese model is one of the most representative approaches. It is a region-based variational segmentation model first proposed by Chan and Vese in 2001 [10]. Unlike edge-based methods that rely on image gradients, the Chan–Vese model formulates segmentation as an energy minimization problem driven by region statistics. Specifically, it differentiates objects based on the average intensity within regions, rather than depending on potentially ambiguous or weak edge information. As a result, it demonstrates strong robustness in segmenting medical images that suffer from noise, intensity inhomogeneity, or blurred boundaries—particularly in modalities such as magnetic resonance imaging (MRI) and ultrasound, where intensity transitions are typically smooth and edge contrast is low.

Moreover, the model adopts a level set method to implicitly represent and evolve contours, enabling it to handle topological changes during the segmentation process. This property makes the model highly effective for segmenting images with multiple disconnected regions, such as multifocal brain lesions or complex vascular structures. In such cases, the number of target regions or the intricacy of anatomical structures does not significantly impact segmentation accuracy.

Additionally, as a mathematical model based on variational principles, the Chan–Vese approach offers explicit interpretability and parameter controllability. Its energy functional is composed of a data fidelity term and a boundary regularization term, allowing users to adjust the smoothness and accuracy of the segmentation boundaries through weighted parameters. As an unsupervised method, the model does not require large-

scale annotated datasets for training. Instead, it leverages intrinsic image information and adapts the contour evolution according to regional features. This makes it particularly suitable for medical image segmentation in data-scarce or privacy-sensitive environments, where access to manually labeled data is limited. Consequently, it also reduces the burden on medical professionals for manual annotation.

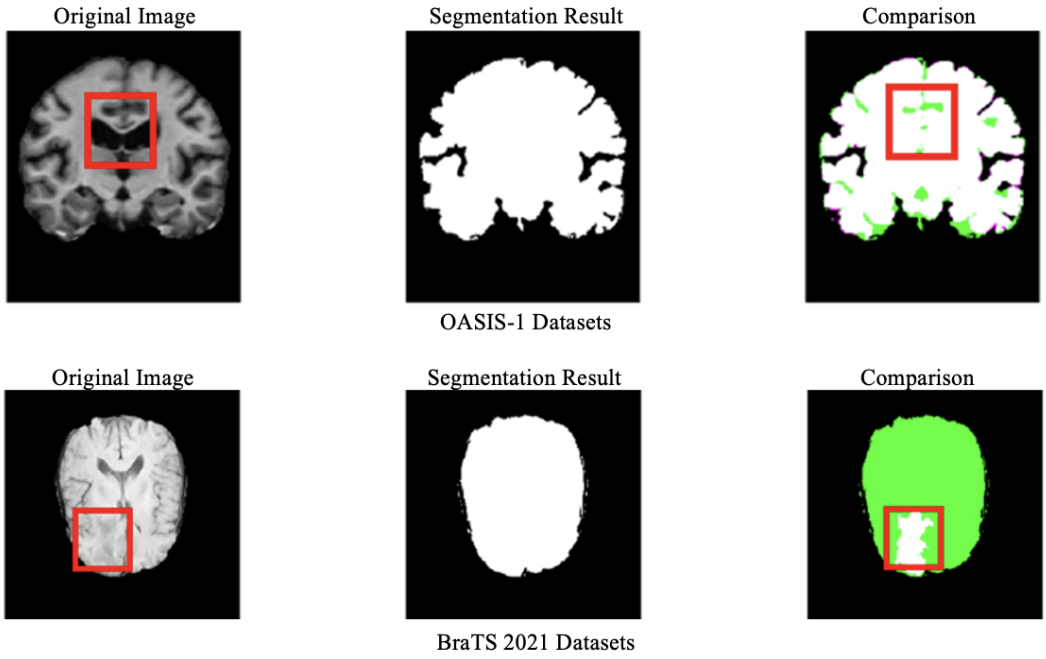


Figure 4.2: Chan-vese Model Segmentation Result. These three columns respectively show the original images, segmentation results, and comparisons of the segmentation results with the ground truth values for the OASIS-1 and BraTS 2021 datasets. The red box represents the expected segmentation result. The green area shows the difference between the expected segmentation result (white area) and the real segmentation result.

Figure 4.2 presents the segmentation results obtained using the Chan–Vese model on the OASIS-1 (Row 1) and BraTS 2021 datasets (Row 2). As shown in the figure, the model is able to accurately capture the overall brain contour (Row 1) in healthy brain images and produce smooth, continuous boundaries, as indicated by the white regions. This demonstrates that the model performs well in regions with relatively homogeneous intensity distributions, ensuring spatial consistency and contour smoothness in segmentation. However, in pathological images (Row 2), i.e., tumor cases, the model is only able to outline the external brain surface and fails to identify internal pathological

regions, resulting in boundary leakage, under-segmentation, and loss of fine structural details. These results indicate that the Chan–Vese model exhibits markedly different segmentation behavior between whole-brain tissue and tumor regions.

The root cause of this discrepancy lies in the model's energy formulation and regional constraints. The Chan–Vese model is based on the assumption of regional intensity homogeneity and achieves segmentation by minimizing a region-based energy functional. In the OASIS-1 dataset, where tissue intensities are relatively stable, the model effectively enforces regional consistency, leading to smooth and coherent segmentation boundaries. In contrast, in the BraTS 2021 dataset, tumor regions exhibit complex intensity variations and ambiguous boundaries, with significant overlap between lesion and background intensities. Consequently, the model fails to distinguish between target and background regions, especially in cases of low contrast or irregular tumor shapes, where it often converges to a local minimum, producing overly smoothed or partially inaccurate segmentations. Furthermore, the model's convergence process is highly sensitive to the initialization of the contour—an improper initialization may prevent the evolving curve from capturing the true structure, resulting in unstable outcomes.

From a structural perspective, the design of the Chan–Vese model inherently limits its flexibility. Its original formulation is binary, distinguishing only between foreground and background. Although multi-region extensions have been proposed, these modifications typically require complex mathematical reformulations and significantly increase computational complexity. In multi-class or multi-modal segmentation tasks (e.g., combining T1- and FLAIR-weighted MRI), the model cannot simultaneously represent distinct tissue or lesion distributions, thus restricting its adaptability in clinical applications. Moreover, the model relies solely on low-level intensity features and lacks the capacity to learn high-level semantic representations from data. When different tissues exhibit overlapping intensity profiles, segmentation accuracy and generalization degrade significantly, and the model performs well only when its assumptions align with the underlying image characteristics.

From a computational perspective, the Chan–Vese model leads to a single non-

linear partial differential equation derived from variational principles. This equation is solved numerically via an iterative level set evolution scheme, in which the contour is progressively updated to minimize the underlying energy functional. This iterative optimization is computationally expensive for high-resolution or volumetric medical images, leading to slow convergence and limited applicability in real-time or large-scale processing scenarios. In addition, the model's segmentation performance is sensitive to the choice of hyperparameters in the energy function, such as the relative weights of the smoothness and data fidelity terms. Because the model lacks an automated parameter selection mechanism, these hyperparameters must typically be tuned empirically or through multiple experiments to balance contour regularity against boundary fidelity. Such sensitivity to parameter settings further limits the model's transferability across heterogeneous datasets.

In summary, the Chan–Vese model performs effectively for brain tissue segmentation tasks characterized by clear structure and uniform intensity, offering strong smoothness and noise robustness. It remains a valuable unsupervised baseline in medical image segmentation. However, its performance degrades significantly in scenarios involving complex pathological structures, non-uniform intensity distributions, or multi-modal image fusion, due to limited adaptability, high computational cost, and sensitivity to manually tuned parameters. These limitations highlight the necessity of developing more adaptive and learning-driven segmentation frameworks that integrate the interpretability of variational models with the representational power of modern learning-based approaches, thereby improving accuracy, robustness, and generalization in medical image analysis.

4.2.2 Supervised Method

As a supervised baseline, we employ the U-Net architecture [1], which has been widely adopted in medical image segmentation tasks. The network follows a typical encoder–decoder design with skip connections to preserve spatial information across different scales.

The model takes three-channel multi-modal MRI images (T1, T2, and FLAIR) as

input and predicts a binary segmentation mask representing tumor regions. Training is conducted using only labeled data, and the Dice loss is used to directly optimize segmentation overlap.

Prior to training, all input images are normalized and spatially aligned. The model is optimized using the Adam optimizer with an initial learning rate of 1e-4. This supervised setting provides a baseline for evaluating the improvements introduced by the proposed semi-supervised framework with ADMM-based energy regularization.

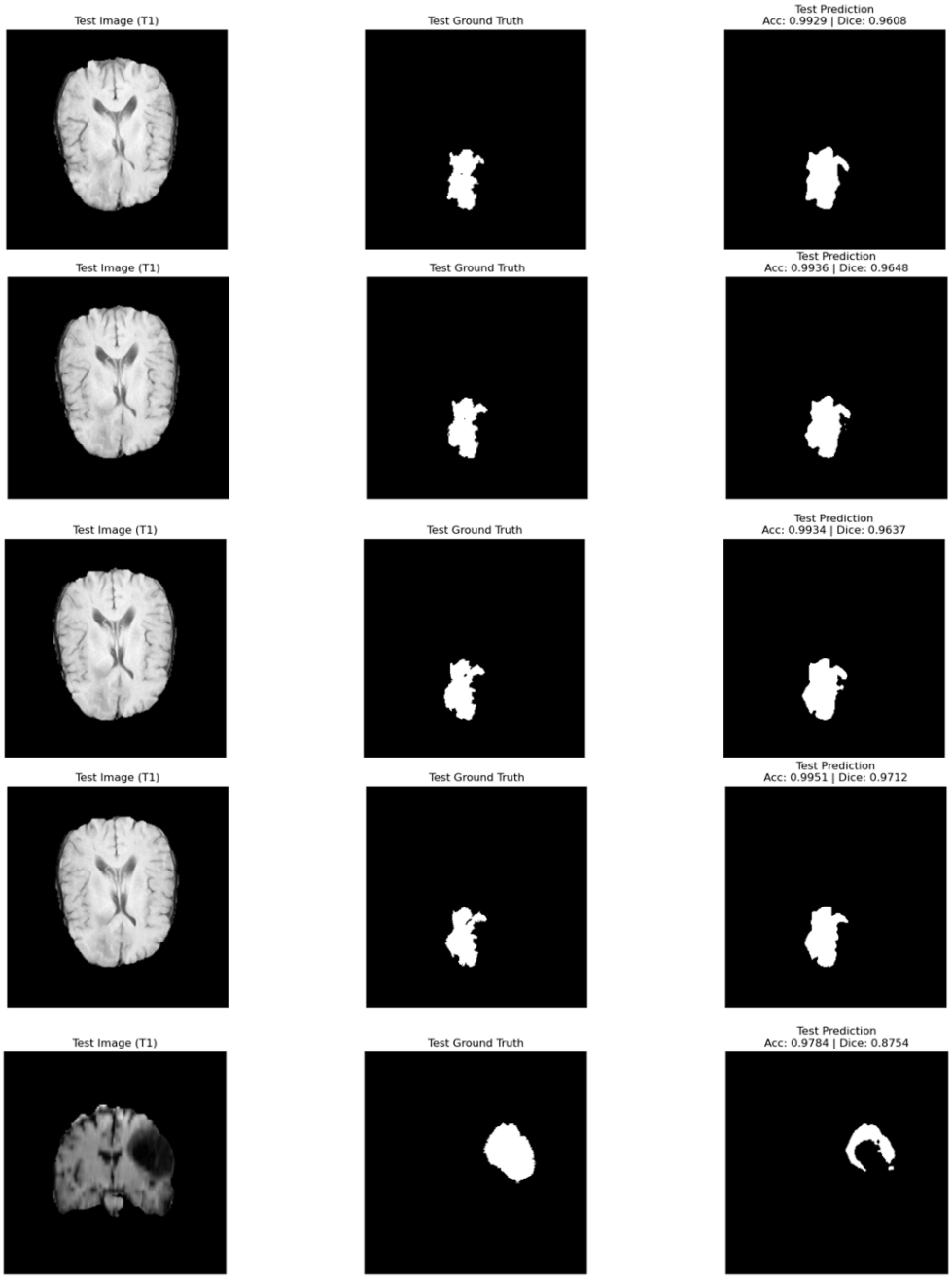


Figure 4.3: The supervised U-Net segmentation. Each column consists of the original T1 image, ground truth, and prediction result with Dice and ACC (left to right). Accurate tumor segmentation but reduced boundary precision in complex regions.

In the baseline setting, the U-Net model was trained purely under supervised learning using labeled data only. As shown in Figure 4.3, the model successfully identifies tumor regions and produces predictions that are visually consistent with the ground

truth masks. This demonstrates that U-Net is capable of capturing the global tumor structure in binary segmentation tasks. Based on the prediction results of different segmentation models on the BraTS 2021 dataset summarized in Table 4.1, the supervised U-Net baseline model achieved an overall accuracy of 0.990 and a Dice coefficient of 0.947, making it an excellent baseline model.

However, subtle differences can still be observed along tumor boundaries. While the predicted regions exhibit similar overall shapes to the ground truth, the contours appear smoother and less detailed, failing to capture fine-grained boundary variations. Consequently, U-Net tends to produce coarse or slightly misaligned edges, leading to minor under- or over-segmentation near complex tumor borders. This limitation reflects the model's difficulty in accurately reconstructing detailed tumor morphology under low-contrast or noisy conditions, despite maintaining high overall pixel-level accuracy.

U-Net has become a widely used architecture in medical image segmentation due to its relatively simple structure and strong performance, it still faces several fundamental challenges when applied in practice.

One key issue lies in its limited ability to precisely segment fine boundaries, especially in cases with low contrast, overlapping structures, or image noise. While the skip connections help to preserve spatial details by combining low-level features from the encoder with high-level semantic features in the decoder, they are not always sufficient to restore lost boundary information. As a result, the network often produces smooth, blurry contours that lack the accuracy required in clinical applications.

In addition, U-Net may fail to detect or properly segment small structures or subtle lesions. These regions are often lost during repeated downsampling in the encoder path, and the network may struggle to recover them fully during upsampling.

Another significant limitation is the model's reliance on large-scale annotated datasets. Although U-Net can extract meaningful features from medical images, it still depends heavily on pixel-level supervision to learn effectively. In many real-world medical scenarios, acquiring sufficient annotated data is difficult—due to patient privacy concerns, the rarity of certain diseases, or the high cost of expert annotations. This makes it challenging to train reliable models across different medical tasks or institutions.

Furthermore, U-Net lacks mechanisms to enforce anatomical plausibility or incorporate domain knowledge, such as expected organ shapes or spatial relationships. This can lead to inconsistent segmentations, particularly in ambiguous or pathological cases where boundaries are unclear.

These challenges suggest that while U-Net remains a powerful tool, its performance can degrade in data-limited or structurally complex situations. This motivates the exploration of hybrid methods that integrate data-driven learning with explicit structural regularization to improve segmentation quality and generalizability.

To address these limitations, our proposed semi-supervised model, by introducing ADMM-based spatial regularization onto the original data, effectively guarantees the spatial consistency and smoothness of the predicted segmentation map. Even under challenging conditions, the predicted boundaries are clearer, more consistent, and more aligned with the ground truth boundaries. Specific prediction results will be presented in detail in Section 4.2.4.

4.2.3 Semi-Supervised Segmentation(Before ADMM Optimization)

To evaluate the effect of the energy term alone, we first train a semi-supervised U-Net with the energy-based loss but without applying the ADMM optimization. In this setting, the energy term is only optimized through standard back-propagation. The goal is to observe how the model behaves when the energy regularization is applied, but the structured ADMM refinement is not yet used.

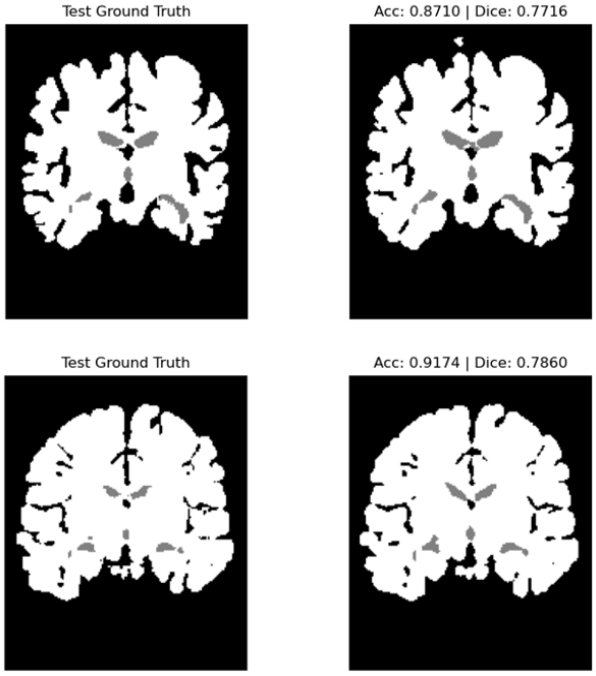


Figure 4.4: Segmentation results on the OASIS dataset using the energy loss without ADMM optimization. The first column shows the ground truth, and the second column presents the segmentation results predicted by the model. In the ground truth and predicted masks, black denotes background, white represents the primary brain tissue, and grey corresponds to additional anatomical structures. The model captures global brain contours but fails to preserve fine internal structures.

As shown in Figure 4.4, the model trained with energy loss but without ADMM optimization successfully captures the global brain contours; however, the internal anatomical boundaries remain over-smoothed, lose details, and are spatially unstable.

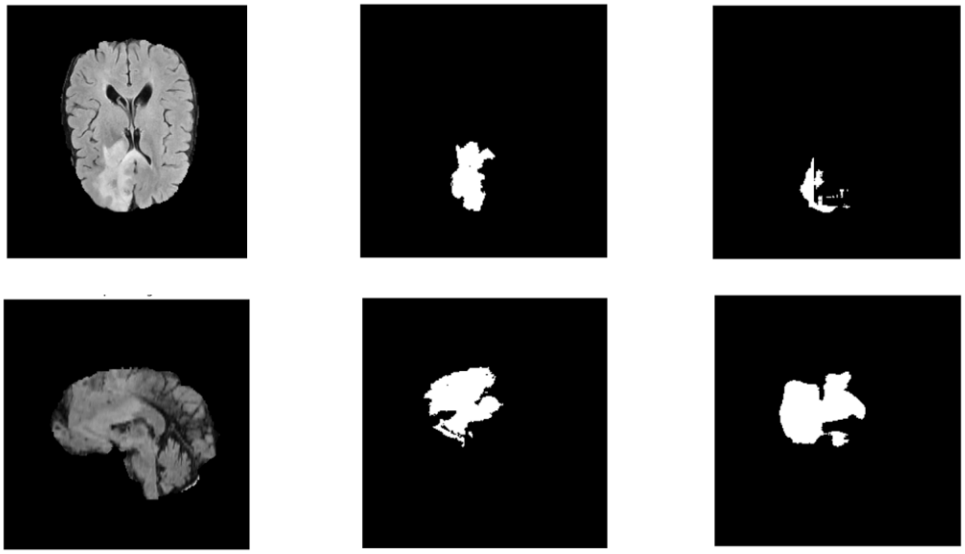


Figure 4.5: Segmentation results on the BraTS dataset using the energy loss without ADMM optimization. Each column consists of the original T1 image, ground truth, and prediction result (left to right). The model fails to capture the full tumor extent, producing fragmented and incomplete masks.

As shown in Figure 4.5, when applied to the BraTS dataset, the energy-based model without ADMM optimization exhibits a notable performance drop compared to the OASIS case. The model often fails to capture the complete tumor region, producing fragmented and unstable masks, and sometimes detecting only partial lesions.

Figure 4.4 and Figure 4.5 both illustrate the semi-supervised segmentation results obtained using the energy loss without ADMM optimization. On the OASIS dataset, the model is able to delineate the global brain outline reasonably well, producing clean and coherent outer boundaries. However, the internal anatomical regions are not clearly distinguished. The segmentation tends to over-smooth deeper brain structures, causing fine anatomical details to disappear and adjacent regions to merge. Moreover, the results show instability across slices, with noticeable variations in the sharpness and consistency of internal regions.

When applied to tumor segmentation on the BraTS dataset, these limitations become more pronounced. Although the model can roughly localize the tumor area, the segmented masks are often fragmented, incomplete, or missing thin and small tumor components. In certain cases, the predictions collapse entirely, highlighting only por-

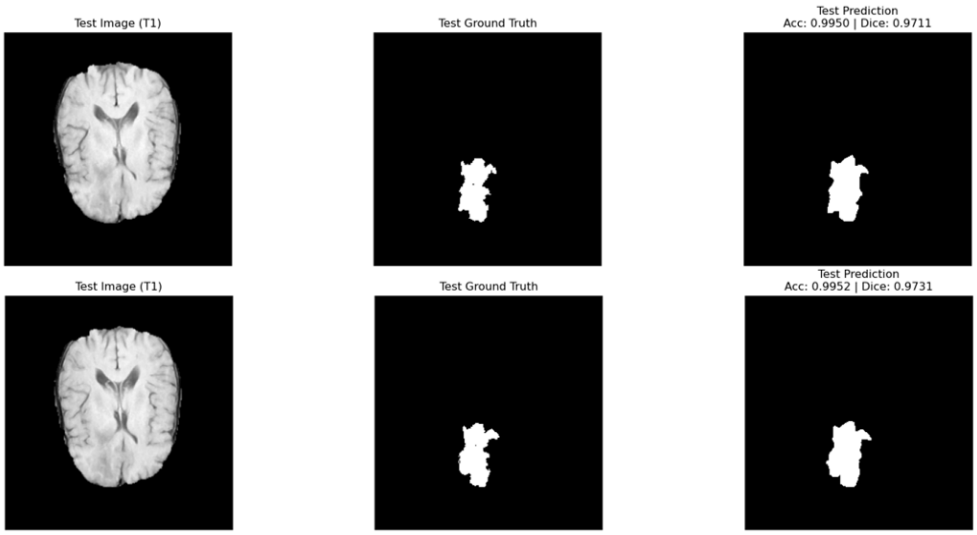
tions of the lesion. This demonstrates weak structural consistency and the inability to maintain accurate tumor boundaries under variations in intensity and texture.

Overall, these observations indicate that relying solely on gradient-based optimization of the energy loss results in over-smoothed segmentations and unstable performance, particularly for heterogeneous tumor structures. This emphasizes the necessity of the subsequent ADMM optimization, which reinforces spatial regularity and substantially improves segmentation completeness and stability.

4.2.4 Our ADMM-Based Semi-Supervised Segmentation Method

To further illustrate the effectiveness of the proposed method, representative visual segmentation results are presented below.

2-phase ADMM-Based Semi-Supervised Segmentation. Figure 4.6 shows the segmentation performance of brain tumor images in the case of 2-phase segmentation.



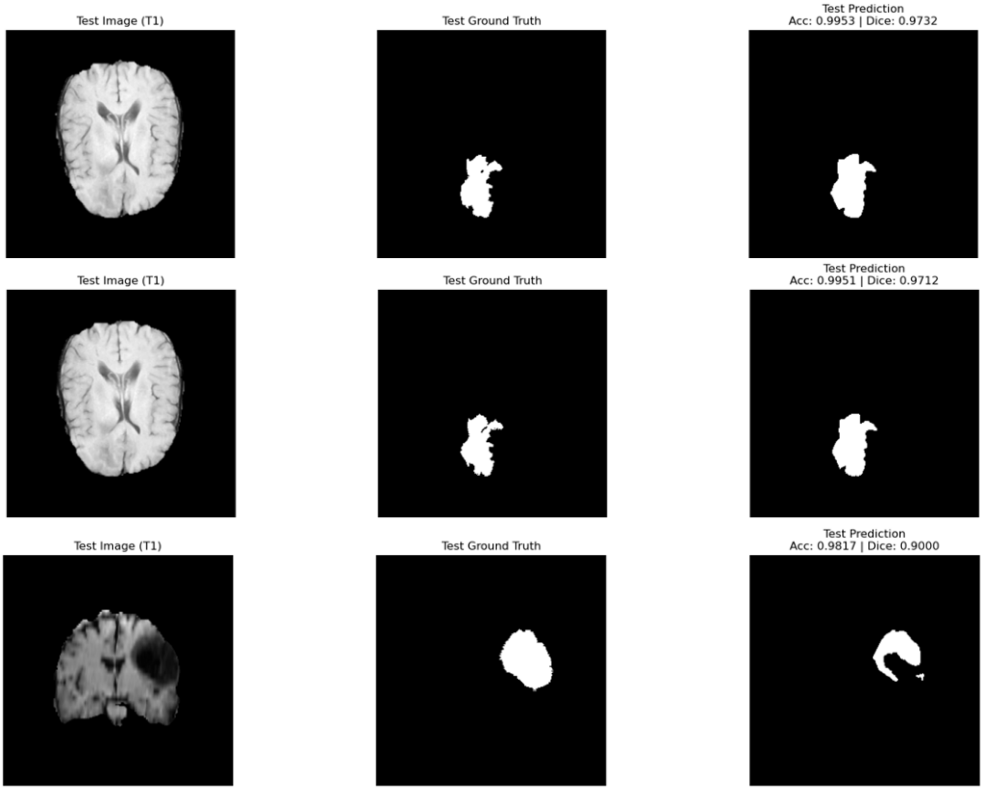


Figure 4.6: 2-phase ADMM-Based Semi-Supervised Segmentation. The figure presents representative visual results of tumor segmentation. Each column consists of the original T1 image, ground truth, and prediction with Dice and Acc (left to right).

Figure 4.6 illustrates the segmentation results of the proposed 2-phase ADMM semi-supervised model on the BraTS 2021 testing samples. Each row displays the original T1-weighted MRI image, the corresponding ground truth mask, and the model prediction. It introduces the model prediction. The predicted results are annotated with the overall accuracy (Acc) and Dice coefficient.

As shown in the results, the proposed model achieves high segmentation accuracy across different subjects (Acc ≈ 0.992, Dice ≈ 0.958). The predicted tumor masks exhibit strong consistency with the ground truth in terms of shape, area, and boundary localization, indicating that the proposed ADMM-based semi-supervised framework effectively identifies tumor regions and accurately reconstructs their structural boundaries. Compared with purely supervised models, the inclusion of the ADMM energy constraint introduces spatial smoothness and regional consistency during training, re-

sulting in smoother boundaries, more complete internal regions, and significantly fewer boundary discontinuities or isolated false detections.

- Boundary smoothness and structural completeness: The model generates smoother and more coherent boundaries along the tumor contours, demonstrating that the ADMM energy term effectively suppresses high-frequency noise and abrupt prediction changes.
- Small-region consistency: In regions with small lesions or low contrast, the predicted segmentation remains continuous and complete, avoiding false negatives or missed detections.
- Cross-sample stability: The segmentation results across different subjects exhibit consistent accuracy and morphology, suggesting strong generalization ability of the model under semi-supervised training conditions.

These improvements can be attributed to the hierarchical energy minimization mechanism introduced by the ADMM optimization process. During training, it jointly constrains both labeled and unlabeled samples through spatial energy regularization, suppressing overfitting and allowing the network to learn reliable structural priors even with limited annotations. Overall, the proposed 2-phase ADMM semi-supervised framework achieves precise tumor region delineation in the global detection stage, providing a stable and accurate foreground mask foundation for subsequent fine-grained subregion segmentation (e.g., necrotic and enhancing regions).

3-phase ADMM-Based Semi-Supervised Segmentation. Figure 4.7 shows the segmentation performance of brain tumor images in the case of 3-phase segmentation.

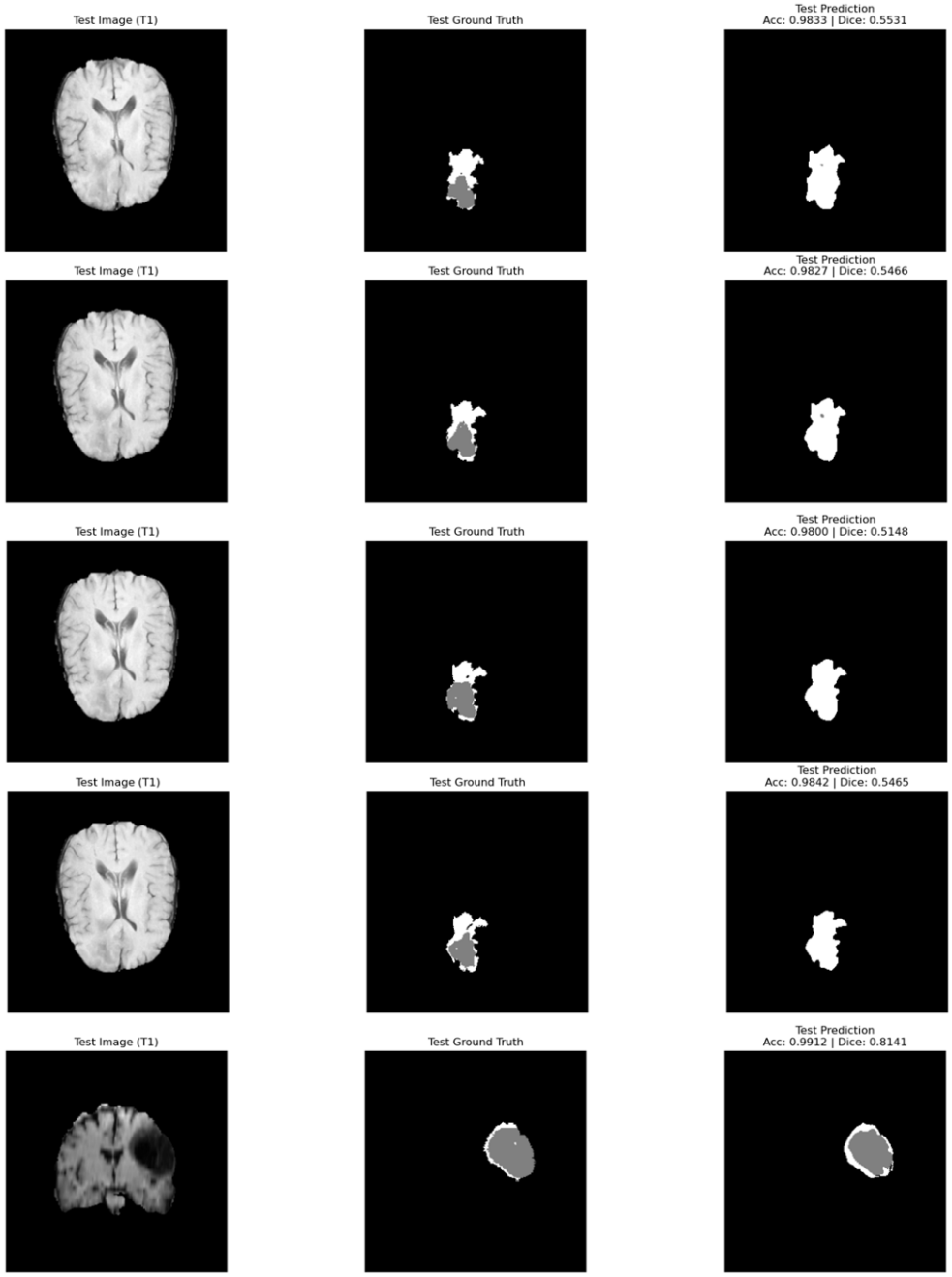


Figure 4.7: 3-phase ADMM-Based Semi-Supervised Segmentation. The figure presents representative visual results of tumor segmentation. Each column consists of the original T1 image, ground truth, and prediction with Dice and Acc (left to right). In the ground truth, black denotes background, white represents the primary tumor region, and grey corresponds to additional tumor subregions (e.g., edema or necrotic regions). The model fails to capture certain tumor subregions, particularly in ambiguous regions, leading to degraded performance in multi-class segmentation and revealing limitations in three-class settings.

Figure 4.7 shows the segmentation results of the three-stage ADMM semi-supervised model on the BraTS 2021 dataset. Unlike the previous two-stage model, the three-stage framework segments multiple tumor subregions (e.g., necrotic core, enhancement region, and edema) simultaneously during the prediction process, aiming to achieve more refined region division.

The results show that, although the overall segmentation accuracy remains high (Acc ≈ 0.98), the Dice coefficient drops significantly to approximately 0.59, indicating a decline in segmentation performance in multi-class scenarios. In other words, the model fails to effectively distinguish between different tumor subregions; it tends to ignore certain classes (e.g., grey regions), revealing a fundamental limitation in its multi-class representation capability.

Specifically, the model exhibits clear instability in boundary localization among tumor subregions.This can be summarized as follows:

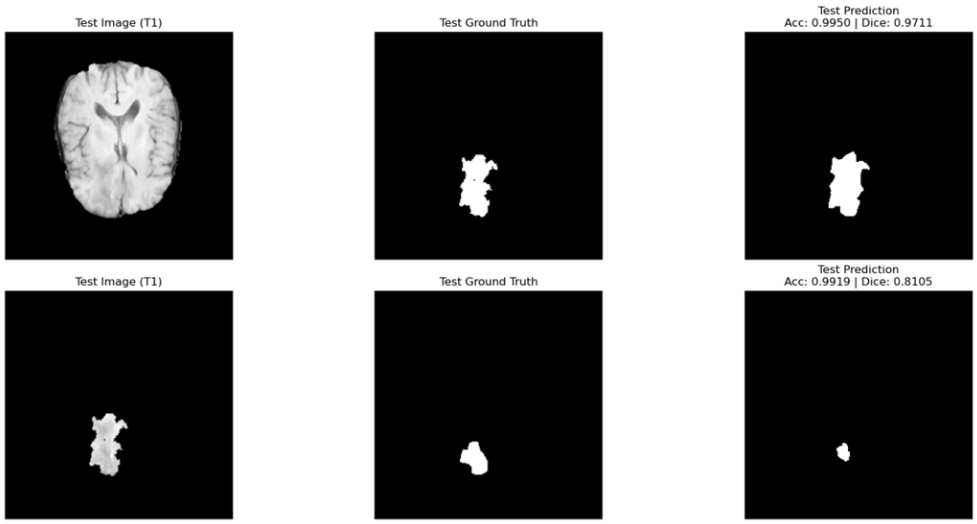
- Class confusion: some enhancing and necrotic regions are incorrectly merged or exhibit blurred boundaries, resulting in spatial overlap between labels.
- Boundary discontinuity and over-smoothing: during multi-class optimization, the ADMM energy constraint introduces competitive interactions between adjacent regions, leading to excessive boundary smoothing.
- Difficulty in small-region recognition: since the 3-phase model predicts multiple categories simultaneously under limited labeled data, the network tends to miss or incompletely segment small or low-contrast lesions such as necrotic areas.

These observations indicate that, in complex multi-class segmentation tasks, a single-stage ADMM semi-supervised structure struggles to balance class discriminability and spatial consistency. In other words, although the energy regularization effectively constrains boundary morphology during binary foreground detection, it weakens inter-class separability in the multi-label setting, resulting in degraded fine-grained segmentation accuracy.

Based on these findings, this study further proposes a two-step segmentation strategy. In this framework, the first step employs the 2-phase ADMM semi-supervised

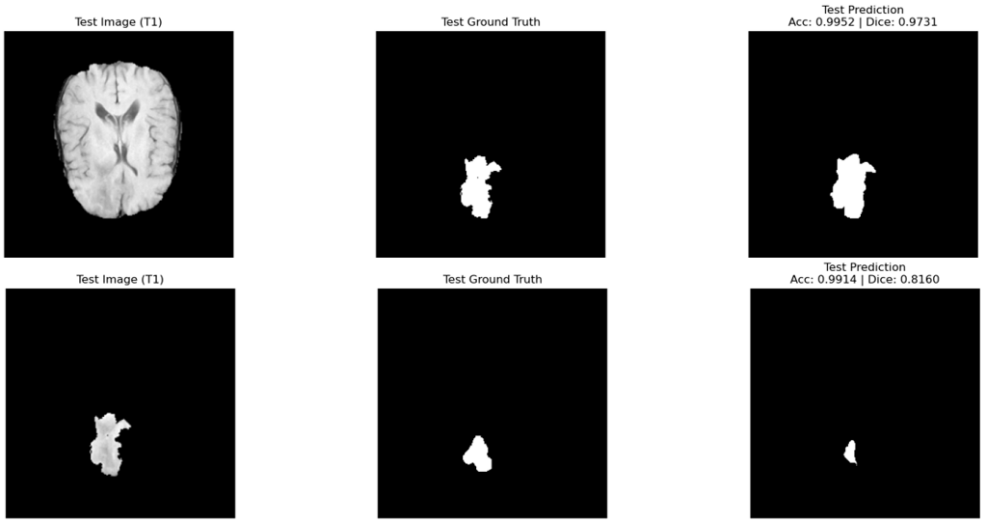
model to detect the global tumor region (foreground), and the second step refines sub-regional classification (e.g., distinguishing necrotic and enhancing regions) within the predicted foreground mask. Both steps share the same network architecture. This hierarchical segmentation framework significantly reduces inter-class interference while preserving the spatial consistency advantages of ADMM regularization, achieving more accurate and stable multi-class tumor delineation.

Two-step 3-phase ADMM-Based Semi-Supervised Segmentation. The following figure presents the segmentation results of the Two-step 3-phase method on five test images, with each case displayed in two rows corresponding to Step 1 and Step 2.

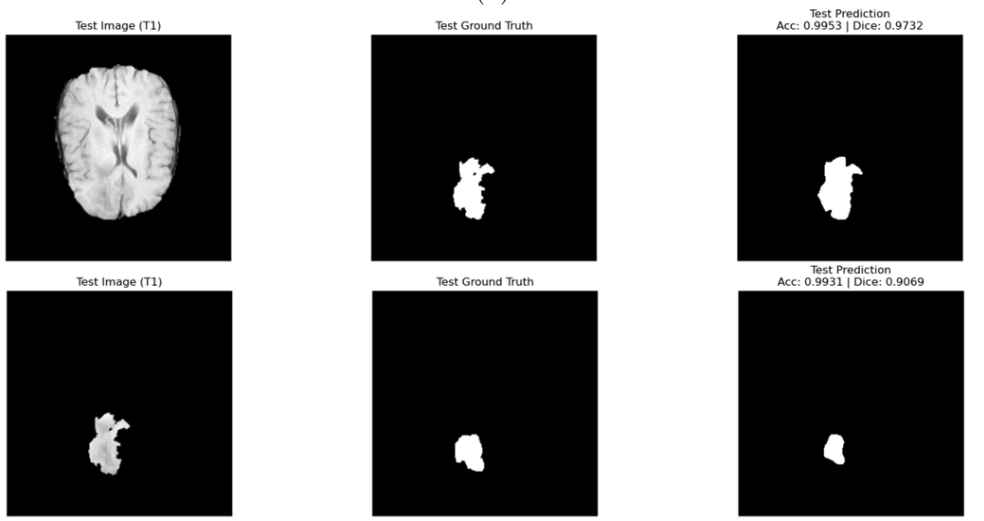


(a)

Figure 4.8: Two-step three-class segmentation pipeline with the proposed ADMM-based semi-supervised segmentation framework. Each column consists of the original T1 image, ground truth, and prediction with Dice and Acc (left to right).

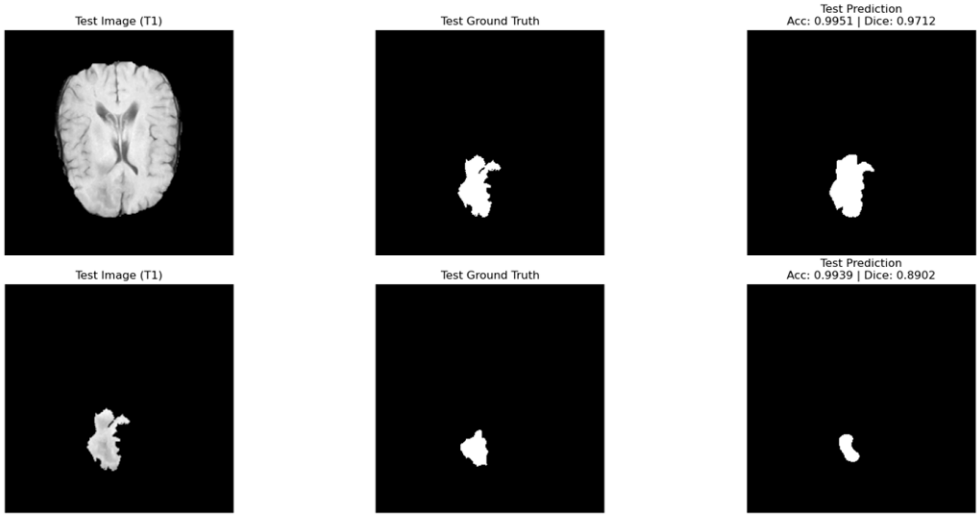


(b)

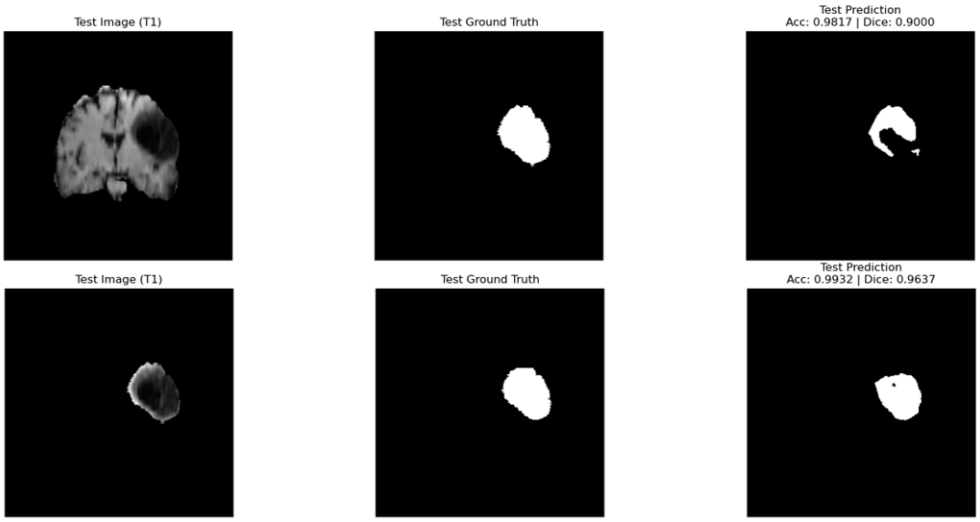


(c)

Figure 4.8: Two-step three-class segmentation pipeline with the proposed ADMM-based semi-supervised segmentation framework. Each column consists of the original T1 image, ground truth, and prediction with Dice and Acc (left to right).



(d)



(e)

Figure 4.8: Two-step three-class segmentation pipeline with the proposed ADMM-based semi-supervised segmentation framework. Each column consists of the original T1 image, ground truth, and prediction with Dice and Acc (left to right).

Figures 4.8 (a) (b) (c) (d) (e) present the experimental results of the proposed two-step ADMM semi-supervised segmentation framework. This approach decomposes the brain tumor segmentation task into two sequential stages: the first stage performs global tumor (foreground) detection using the 2-phase ADMM model, while the second stage

further refines subregion classification (e.g., necrotic core, enhancing region) within the predicted foreground mask.

The results demonstrate that the stepwise framework exhibits notable advantages in terms of stability and visual consistency.

While the Dice coefficient is slightly reduced (approximately 0.877) in the second step of sub-tumor segmentation, the proposed method still exhibits advantages over the single-stage three-phase model in several key aspects:

- Higher prediction stability: segmentation results across all samples show highly consistent morphological structures without the inter-class interference or boundary fluctuations observed in the 3-phase model, indicating that staged optimization effectively mitigates instability caused by multi-class competition.

- Clearer structures and smooth boundaries: the first-stage foreground segmentation provides reliable spatial constraints for subsequent classification, allowing the second-stage model to distinguish different tissue types within a smaller and more constrained search space. The predicted masks exhibit smooth edges and strong connectivity.

- Improved recognition of small regions: for small or low-contrast lesions, the stepwise model achieves more accurate localization within the constrained foreground area, reducing missed detections and misclassifications commonly observed in the 3-phase model.

- More stable semi-supervised regularization: the ADMM energy term contributes at both stages—the first stage constrains global structural consistency, while the second stage refines local details. Compared to direct joint optimization in a multi-label space, this hierarchical formulation aligns better with the physical interpretation of energy regularization, resulting in more stable convergence during training.

The two-step ADMM semi-supervised framework demonstrates superior performance in structural consistency, prediction robustness, and generalization stability

compared to the single-stage multi-class model. Although the Dice scores for some fine subregions are slightly reduced, the framework achieves a more balanced trade-off between visual quality and model stability. These findings suggest that the stepwise strategy effectively alleviates the interference of multi-class competition in ADMM optimization, providing a more structured and reliable path for multi-stage brain tumor segmentation.

Overall, Figure 4.8 presents representative visual results of tumor segmentation, highlighting the differences among the models in tumor region identification and boundary delineation. To further illustrate these differences, two representative cases are provided as detailed side-by-side comparisons:

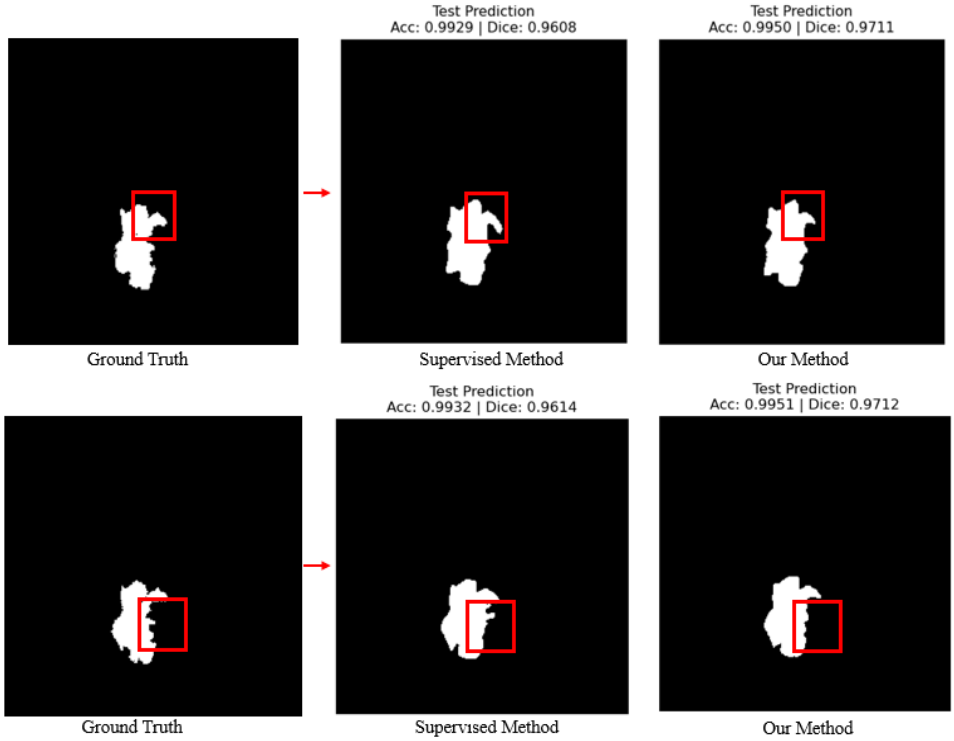


Figure 4.9: Detailed comparison between the 2-phase proposed method and the 2-phase supervised method on two representative test cases. From left to right, each column shows the ground truth, the result of the supervised method with Dice and Acc, and the result of the proposed method with Dice and Acc. The red boxes highlight the regions of significant difference.

Figure 4.9 provides a detailed comparison of the segmentation results obtained by the 2-phase supervised model and the proposed semi-supervised model on two rep-

resentative test cases. Compared with the supervised method, the proposed method produces boundaries that more closely match the ground truth, with smoother contours, fewer irregular artifacts, a higher Dice score, and more accurate local delineation in the highlighted regions. Based on these observations, the displayed results can be summarized as follows:

(1) 2-phase ADMM model: This model achieves remarkably high segmentation accuracy in the binary (foreground/background) task (Dice ≈ 0.958, Acc ≈ 0.992). The predicted masks closely match the ground truth in both shape and boundary, showing smooth outer contours and complete internal filling of the tumor region. This indicates that the ADMM energy regularization effectively suppresses false predictions and boundary discontinuities. Compared with the conventional supervised model, the 2-phase ADMM maintains strong consistency even in low-contrast regions, demonstrating the spatial constraint advantage of the semi-supervised framework when utilizing unlabeled samples.

(2) 3-phase ADMM model: When extended to multi-class segmentation, the model exhibits noticeably reduced stability. Although it can still accurately detect the overall tumor region, considerable confusion and overlap occur at the boundaries of subregions such as necrotic and enhancing areas. Some regions suffer from excessive smoothing and class mixing, with the Dice dropping to approximately 0.595 and Acc ≈ 0.984. These observations suggest that the single-stage ADMM energy constraint encounters inter-class competition in multi-class tasks, making it difficult to achieve precise boundary separation between different tumor components.

(3) Two-step ADMM model: Under the stepwise strategy, the model first employs the 2-phase ADMM to accurately identify the overall tumor foreground, followed by fine-grained classification within the predicted mask. This approach significantly improves structural stability, producing morphologically consistent and smooth boundaries across all samples. Although the overall Dice score is slightly lower, the visual segmentation results align more closely with actual tumor morphology. This indicates that the stepwise structure effectively reduces inter-class competition, allowing the ADMM energy constraint to operate independently in each phase and leading to more

robust segmentation performance.

In summary, the qualitative results further validate the effectiveness of the proposed ADMM-based semi-supervised framework. While the single-stage multi-class configuration theoretically allows simultaneous optimization of multiple regions, it tends to introduce instability. In contrast, the stepwise hierarchical optimization strategy achieves more interpretable and stable tumor segmentation results while maintaining spatial consistency.

4.2.5 Discussion

First, the ADMM energy regularization significantly enhances the structural consistency and spatial stability of the model in semi-supervised learning. In the 2-phase model, the ADMM term introduces regional smoothness and boundary continuity constraints during optimization, enabling the model to maintain reasonable segmentation structures even for unlabeled samples. Essentially, this regularization mechanism is equivalent to introducing a "physical consistency" prior into the training process, allowing the model to learn spatial topology information even under limited annotations.

Second, the single-stage multi-class (3-phase) configuration exhibits instability in complex segmentation scenarios. Although the ADMM regularization theoretically allows for simultaneous energy constraints across multiple classes, experiments show that such joint optimization leads to competition among regional forces, resulting in over-smoothed or ambiguous boundaries between different classes. Consequently, the Dice coefficient drops sharply, suggesting that a single energy term cannot simultaneously preserve both inter-class separability and spatial continuity in multi-class segmentation.

To address these issues, this study further proposes a stepwise (Two-step) ADMM semi-supervised structure. By decoupling global tumor detection (Phase 1) from sub-region refinement (Phase 2), the model performs energy optimization at both global and local scales, leading to more stable boundary reconstruction. Experimental results show that although the average Dice is slightly lower, the model exhibits more consistent performance across samples. This indicates that the hierarchical optimization strategy provides higher robustness in spatial structure and morphological prediction, offering a

more interpretable and stable training paradigm for complex structural segmentation tasks.

Finally, several limitations remain in this study. On one hand, the performance of the ADMM regularization is sensitive to the parameter λ, and the energy weights require manual adjustment across different classes or spatial scales. On the other hand, the instability observed in the 3-phase configuration reflects the lack of class-specific representation in a single energy formulation for multi-label spaces. Future research can be extended in the following directions:

1. Designing adaptive energy weighting strategies to dynamically adjust regularization strength based on sample complexity or loss behavior;

2. Integrating the ADMM framework with attention mechanisms or graph-based regularization to improve inter-class discrimination and structural consistency;

3. Validating the generalization ability of the proposed framework on large-scale unlabeled medical datasets to further assess its scalability for clinical applications.

In conclusion, the proposed ADMM-based semi-supervised segmentation framework introduces an energy optimization perspective that enables stable structural prediction under limited supervision, providing an interpretable and extensible optimization pathway for multi-stage medical image segmentation tasks.

Chapter 5

Discussion

5.1 Strengths of the proposed Method

To ensure consistency between regional accuracy and structural integrity, this study proposes a hybrid image segmentation framework that jointly optimizes Dice loss and a variational Potts-based energy term. The method integrates a U-Net with Transformer modules and incorporates a variational regularization component, which is efficiently solved using ADMM during training. This enables the model to learn from limited labels while simultaneously enforcing spatial smoothness and structural consistency in the predicted regions, achieving complementary benefits between data-driven learning and variational modeling.

Specifically, Dice loss, denoted $\mathcal{L}_{\text{Dice}}$, measures the overlap between predicted segmentation and ground truth, serving as voxel-wise surveillance for region accuracy. However, Dice loss alone is insensitive to spatial structure and boundary smoothness, which may lead to ambiguous boundaries, noisy small regions, and irregular local predictions. To address this limitation, our study incorporates an energy regularization term based on the Alternating Direction Method of Multipliers (ADMM), denoted as $\mathcal{L}_{\text{ADMM}}$, on top of Dice supervision.

The ADMM-based energy term directly acts on the predicted foreground probability map. It iteratively computes spatial smoothness and sparsity while enforcing numerical projection constraints on auxiliary variables through alternating optimization. By

minimizing the residual difference between dual variables and Lagrange multipliers at each iteration, the model progressively approaches the Potts energy's minimum configuration, thereby achieving region smoothness and boundary refinement in the predicted segmentation.

Unlike traditional Total Variation or Conditional Random Field (CRF) regularization methods, the proposed ADMM energy term is seamlessly embedded into the network's computational graph, enabling end-to-end differentiable gradient propagation. This design allows the energy regularization to be optimized jointly with the main segmentation parameters in a unified training framework, achieving a coordinated optimization between supervised learning and structural regularity. The combined loss function of this study is defined as:

$$\mathcal{L}_{\text{total}} = \beta \cdot \mathcal{L}_{\text{Dice}} + \alpha \cdot \mathcal{L}_{\text{ADMM}},$$

where α increases linearly during training, realizing a dynamic optimization strategy that gradually transitions from discriminative supervision to structural regularization. In the early training phase, the Dice loss dominates, guiding the network to learn stable region-level discriminative features. As training progresses, the weight of the ADMM energy term increases gradually, enabling the model to further optimize regional continuity and boundary smoothness based on the already established segmentation results, thus achieving more stable structural representations overall.

This joint optimization mechanism achieves a complementary effect between region discrimination and structural regularization: the Dice loss ensures voxel-level consistency between the predicted segmentation and the annotated regions, while the ADMM energy term enforces spatial gradient constraints on the network's output geometry. Consequently, the predicted segmentation becomes smoother inside each region and more continuous at the boundaries. Experimental results demonstrate that using this combined loss maintains a high Dice coefficient while effectively reducing isolated noise and boundary artifacts, yielding segmentation results that better correspond to the true anatomical structures.

Furthermore, since the ADMM energy computation depends only on the model's

predicted output rather than labeled data, the proposed design naturally possesses semi-supervised scalability. For unlabeled samples, the same energy term can act as an unsupervised constraint, encouraging the model to learn latent consistency and boundary structural cues across regions. This provides both theoretical and practical foundations for extending energy-based optimization to semi-supervised medical image segmentation.

In summary, the proposed Dice–ADMM combined loss establishes a unified and differentiable framework that bridges supervised learning and structural regularization. It enables the network to simultaneously optimize discriminative ability and spatial coherence within a single training process, thereby enhancing segmentation stability, continuity, and generalization.

5.2 Limitations and Challenges

Although the ADMM-based energy regularization method proposed in this study achieves good results in binary brain tumor segmentation tasks, several theoretical and practical challenges remain to be further investigated regarding model formulation, numerical stability, and scalability to multi-class segmentation tasks.

First, the method shows potential for extension to multi-phase or multi-class segmentation tasks. Classical Potts-type energies rely on piecewise-constant assumptions and geometric smoothness priors, which are effective for capturing global region boundaries but struggle to represent intra-region heterogeneity, fine anatomical structures, and subtle intensity transitions. Moreover, such variational models lack semantic awareness and cannot explicitly encode anatomical relationships or hierarchical tissue organization, making them sensitive to low-contrast boundaries and prone to over-smoothing or fragmented predictions in heterogeneous regions. Although multi-phase extensions are theoretically feasible, the resulting high-dimensional coupled optimization increases computational burden and training instability, especially under limited supervision. Therefore, while variational priors provide valuable structural regularization, they require complementary learning-based features and supervision signals to achieve robust and semantically consistent medical image segmentation.

The energy model in this study is based on the two-phase Potts formulation, which enforces spatial consistency and smoothness only between foreground and background regions. When the segmentation task involves more complex multi-class or hierarchical structures (such as necrosis, enhancement, and edema), the energy function must describe interactions and boundary constraints among multiple regions. Its mathematical representation requires extending the scalar gradient field to a vector-coupled system. In this case, during each ADMM iteration, the auxiliary variable q and Lagrange multiplier λ must be updated in higher-dimensional space, significantly increasing computational complexity. Moreover, the boundaries between different classes may overlap or blur in the spatial domain, making it difficult for the traditional Potts-based formulation to guarantee global convexity. As a result, three-phase or multi-class segmentation often needs to be implemented through stepwise pipelines rather than within a single unified end-to-end model.

Second, there is still room for improvement in terms of mathematical formulation and optimization stability. Although the ADMM energy term can be embedded into neural network training as a differentiable component, the optimization process still depends on fixed hyperparameters such as penalty coefficient c, step size parameter σ, and projection radius ϵ. These parameters may vary across datasets or training stages, and improper selection can affect convergence stability and learning dynamics. If chosen poorly, the energy term may converge prematurely and hinder structural learning, or fail to converge efficiently due to overly small step sizes. In addition, since ADMM only approximates the minimization under limited iteration steps, its energy convergence often remains suboptimal. When the theoretical constraint condition $\mathrm{div}(q) \approx 1 - h$ is not fully satisfied, where h denotes the predicted segmentation function (or indicator function of the foreground), the resulting energy minimization may exhibit numerical instability.

Third, there remain theoretical limitations regarding the differentiability and stability assumptions in gradient propagation. Although the ADMM process is differentiable, the propagation of the gradient field may become numerically unstable during the gradient–divergence computation, especially in regions with large boundary curvature or

high local contrast between the input and the prediction. Because ADMM involves multiple convolutional and gradient computations, numerical accumulation errors can affect the stability of propagated gradients, leading to local discontinuities or gradient explosions under extreme conditions. This issue is particularly evident in high-resolution MRI images and requires further mathematical analysis to ensure differentiability and numerical stability.

In addition, the current ADMM energy term primarily focuses on spatial structural regularization and lacks explicit interaction with Transformer feature representations in the semantic space. In other words, while spatial regularization and semantic encoding remain relatively independent, the model has not yet achieved an integrated optimization that jointly constrains structural consistency and semantic understanding. This structural–semantic decoupling performs well in maintaining smoothness and continuity at region boundaries but may limit the model's ability to capture multi-scale or semantically diverse patterns (such as different tumor subregions or lesion variations). Consequently, the model's ability to represent spatial–semantic consistency remains limited.

Moreover, the supervision design for semantic-level learning also has inherent limitations. Since training mainly relies on publicly available datasets, the ground truth typically provides region-level annotations rather than pixel-wise supervision across hierarchical lesion structures. As a result, the semantic learning objectives of the model are weakly correlated with fine-grained clinical definitions of tumor subtypes. In other words, while the model's segmentation results align well with morphological boundaries, its semantic understanding of the segmented region remains limited. The network may correctly identify where the lesion is located, but it does not necessarily understand what the segmented structure represents in a pathological or clinical sense. This reveals a gap between region-level supervision and deeper medical interpretability.

Finally, there are still limitations in semi-supervised learning and cross-domain generalization. Although the ADMM energy term theoretically possesses unsupervised regularization capability independent of label information, the current training framework does not introduce explicit unlabeled consistency losses (e.g., strong–weak augmenta-

tion consistency or pseudo-label regularization). Therefore, the model's utilization of unlabeled data remains confined to structural regularization rather than genuine semi-supervised learning. Furthermore, the smoothness constraints of the ADMM energy term may behave inconsistently across centers or imaging protocols, potentially introducing domain bias and reducing generalization ability in cross-center or cross-dataset applications.

In summary, the current method still leaves room for improvement in mathematical formulation, numerical stability, and structural–semantic integration. Future work should focus on developing multi-class energy models, adaptive parameter learning, theoretical convergence analysis, and cross-domain consistency regularization to advance ADMM-based structural regularization toward more complex and generalizable medical image segmentation tasks.

5.3 Future Work and Improvements

Future research will primarily focus on four directions: data generation, model structural optimization, energy regularization extension, and theoretical refinement.

First, at the data level, the ADMM-based energy regularization framework proposed in this study has demonstrated good generalization and robustness under limited annotated samples. However, the high annotation cost and limited scale of medical imaging data remain the main bottlenecks restricting the model's stability and generalization. To further alleviate the problem of data scarcity, future work could explore high-order data generation strategies based on random diffeomorphism generation.

Inspired by the theory of quasi-conformal mapping in traditional image registration, this approach constructs a spatial transformation in the form of a complex function [47]:

$$\varphi(\mathbf{x}) = u(\mathbf{x}) + iv(\mathbf{x}), \quad \mathbf{x} \in \Omega,$$

where $u(x)$ and $v(x)$ denote the real and imaginary components of the transformation, respectively. Equivalently, the mapping $\varphi : \Omega \subset \mathbb{R}^2 \to \mathbb{R}^2$ can be interpreted as a two-dimensional deformation field $\varphi(x) = (u(x), v(x))$, which describes the spatial

displacement of each point in the image domain. Defines the field of the corresponding complex coefficient [47]:

$$\mu(\varphi(\mathbf{x})) = \frac{(\partial_x u - \partial_y v) + i(\partial_x v + \partial_y u)}{(\partial_x u + \partial_y v) - i(\partial_x v - \partial_y u)},$$

which describes the local nonlinearity and rotational distortion of the deformation. By randomly sampling the complex field μ, a series of diffeomorphic deformations can be generated that preserve topology while exhibiting diverse geometric distortions. These synthetic deformations can be applied to the original training data for spatially structured augmentation, enhancing data diversity while maintaining anatomical consistency. Such diffeomorphism-based data augmentation not only expands the data distribution but also improves the model's robustness to geometric variations and spatial heterogeneity.

Moreover, with the rapid advancement of artificial intelligence, future research may further incorporate generative model-driven data augmentation strategies, such as diffusion models or Generative Adversarial Networks (GANs). By combining the ADMM energy constraint with generative priors, these models could form a "generation–regularization–optimization" loop, where the generated data assist in learning smoother and more interpretable deformation representations. This strategy allows the model to dynamically learn energy regularization in adaptive geometric spaces, providing a more unified and robust framework for semi-supervised medical image segmentation under limited data conditions.

Second, at the model architecture level, future work can explore the deep fusion of ADMM optimization with Transformer-based feature extraction. Currently, the energy term mainly focuses on spatial regularization, while feature learning and semantic representation remain relatively independent. Introducing multi-dimensional attention mechanisms within the Transformer could enhance the joint modeling of energy and semantic consistency, thereby achieving cross-domain and multi-scale optimization. For multi-class segmentation tasks, a potential solution is to design multi-phase ADMM blocks that support region-wise energy decomposition, allowing the model to capture hierarchical structures across multiple semantic levels.

Third, at the theoretical and algorithmic level, the ADMM framework can be further refined from several perspectives:

1. Adaptive parameter learning mechanism: introducing learnable parameters (c, σ, α) that adjust dynamically with the training stage and data characteristics to maintain energy balance.

2. Convergence analysis of discrete ADMM: establishing theoretical guarantees for gradient stability and convergence in discretized ADMM formulations, providing interpretability for embedded energy optimization.

3. Multi-scale energy constraints: applying layered energy formulations to simultaneously enforce global smoothness and boundary continuity, improving adaptability across different spatial resolutions.

And, in future research, extending the ADMM-based energy regularization mechanism to handle color and volumetric medical images (3D volumetric data) will be an essential step toward enhancing the model's generalization and scalability. Current methods are primarily designed for two-dimensional grayscale segmentation, where energy regularization only enforces spatial consistency within a single-plane representation. However, in modern medical imaging, natural volumetric data often contain rich multi-channel and 3D structural information, which introduces new opportunities and challenges for integrating ADMM-based regularization into higher-dimensional models.

For color or multimodal medical images, future work can explicitly incorporate inter-channel structural consistency within the energy formulation to establish geometric coherence across different channels or modalities. For instance, in RGB images, the structural variations between color channels are often correlated through the structural flow field. Under the ADMM framework, this can be modeled by introducing joint gradient constraints across multiple channels,

$$\nabla u = [\nabla u_R, \nabla u_G, \nabla u_B],$$

combined with coupled regularization terms such as group norm or tensor norm, to op-

timize inter-channel gradient fields. Such constraints can effectively prevent structural misalignment or false separation between channels. Similarly, for multi-modal medical imaging (e.g., T1/T2/FLAIR), modality-adaptive energy terms can be designed to maintain inter-modality coherence while dynamically balancing modality-specific intensity variations, thereby ensuring robust and interpretable multi-modal feature fusion.

In terms of 3D volumetric data, ADMM optimization can be further extended into voxel space to realize true 3D energy regularization. Specifically, the two-dimensional differential operators (∂_x, ∂_y) can be generalized to three-dimensional gradients $(\partial_x, \partial_y, \partial_z)$, thereby constructing spatial smoothness constraints in volumetric form. Correspondingly, the energy field can be defined as a 3D vector field

$$q = (q_x, q_y, q_z),$$

and the constraint term can be formulated as $\|q\| \leq \alpha$ to ensure stability and boundedness in the optimization process. This 3D extension allows the ADMM-based model to maintain global smoothness and boundary consistency in volumetric segmentation tasks such as tumor delineation or organ surface reconstruction.

Nevertheless, the optimization of 3D ADMM energy regularization will inevitably lead to increased computational complexity and memory cost. The gradient and divergence computations over dense 3D grids scale quadratically with GPU memory and time consumption. Therefore, future studies may investigate strategies based on sparse voxel representation or multi-scale hierarchical optimization to reduce computational overhead while preserving the effectiveness of energy regularization. Furthermore, integrating ADMM-based energy terms with 3D convolutional neural networks (3D CNNs) or Transformer-based 3D architectures may further enhance the model's ability to capture spatially deep and structurally coherent representations, improving both segmentation consistency and generalization across complex volumetric structures.

Overall, extending ADMM-based energy optimization to color and 3D volumetric domains not only increases model complexity but also represents a higher-dimensional extension of the structural regularization principle. This development provides a natural transition from two-dimensional plane segmentation to fully volumetric space, laying

both theoretical and technical foundations for a unified, multi-scale, and interpretable energy-driven medical image segmentation framework.

Finally, from an integrated perspective, future work aims to construct a unified medical image segmentation framework driven by "data generation–energy regularization–semantic learning". Through limited annotated data and a small number of diffeomorphic transformations or generative models (e.g., diffusion models or GANs), coupled with ADMM-based structural regularization, the model will achieve consistent learning across geometry, semantics, and structure. This direction is expected to provide both theoretical and practical foundations for unified semi-supervised segmentation frameworks that are dynamically regularized and interpretable.

Chapter 6

Conclusion

This thesis addresses the problems of structural instability, boundary discontinuity, and insufficient regional discrimination that often arise in traditional supervised medical image segmentation methods under limited annotated data conditions. To overcome these challenges, we propose a semi-supervised medical image segmentation framework that incorporates an ADMM-based energy regularization term. Compared with the baseline U-Net, the proposed ADMM-regularized models demonstrate notable improvements across accuracy, Dice score, and boundary precision. Specifically, the two-phase ADMM model achieves an Accuracy of 0.995 and a Dice of 0.97, outperforming the U-Net baseline (0.993 Accuracy, 0.96 Dice). This corresponds to a 0.2% improvement in Accuracy and a 1% increase in Dice, indicating enhanced segmentation fidelity. Built upon the variational modeling theory, the proposed method introduces the energy minimization principle of the Potts model into deep learning architectures. Through the joint optimization of Dice loss and ADMM energy loss, the framework achieves collaborative training between discriminative supervision and geometric regularization, thereby enhancing the structural consistency and robustness of segmentation results while maintaining regional accuracy.

The core idea of this work lies in embedding variational-based structural regularization into the training process of neural networks, enabling the network to learn not only class-level discrimination but also spatial smoothness and boundary continuity. The Dice loss measures the regional overlap between predicted segmentation and ground

truth from a supervised perspective, while the ADMM energy term constrains the spatial distribution of the predicted probability maps from a structural perspective, thus maintaining a dynamic balance between discriminability and regularity. By employing the energy decomposition and iterative optimization mechanism of the Alternating Direction Method of Multipliers (ADMM), the framework realizes a differentiable energy optimization process within backpropagation, offering both theoretical interpretability and computational feasibility.

Experimental results on the BraTS 2021 multi-modal brain tumor dataset demonstrate that the proposed method significantly outperforms traditional variational models and purely supervised segmentation methods in terms of Dice coefficient, accuracy, and structural completeness. With the introduction of ADMM energy regularization, the model exhibits improved stability and generalization under semi-supervised training, maintaining structurally coherent segmentation results even with limited labeled data. Particularly in cases of ambiguous tumor boundaries or complex regional morphologies, the model demonstrates strong discriminative capability. Further experiments combining T1, T2, and FLAIR modalities validate the positive impact of multi-modal complementarity on structural representation, confirming that the proposed energy constraint mechanism effectively captures both local textures and global shape information.

Overall, this research achieves meaningful advances in both theoretical formulation and practical implementation. On the theoretical side, the proposed joint optimization framework unifies variational modeling and deep neural networks, allowing the energy regularization term to be seamlessly embedded into the differentiable training process. On the practical side, the framework substantially improves segmentation robustness and spatial consistency under limited supervision, providing a feasible and interpretable approach for semi-supervised medical image segmentation.

Despite its promising performance in enhancing structural consistency and discriminative accuracy, the proposed method still has several aspects that warrant further investigation. The mathematical convergence of the ADMM optimization process requires more rigorous analysis, and the model's scalability to multi-phase or multi-class

segmentation tasks needs to be further explored. Moreover, the integration between the energy constraint and semantic features is not yet fully unified, calling for deeper theoretical interpretation of the structural regularization mechanism within semantic feature space.

Future work will focus on three directions: data generation, energy-constrained optimization, and semantic-structural fusion. Specifically, we plan to employ advanced data augmentation strategies based on random diffeomorphism generation and generative models such as Diffusion or GANs, combined with differentiable ADMM optimization and adaptive energy parameter learning, to establish a unified "data generation–energy regularization–semantic learning" framework. This will further enhance model stability and interpretability in cross-domain and multi-modal medical image segmentation tasks. In summary, the ADMM-based energy regularization proposed in this thesis provides a novel perspective for structural optimization in deep segmentation networks and lays a solid foundation for developing semi-supervised medical image segmentation methods that are both theoretically interpretable and practically scalable.

Bibliography

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. Lecture Notes in Computer Science, vol. 9351. Springer, 2015, pp. 234–241.

[2] L. Burrows, K. Chen, W. Guo, M. Hossack, R. G. McWilliams, and F. Torella, "Evaluation of a hybrid pipeline for automated segmentation of solid lesions based on mathematical algorithms and deep learning," *Scientific Reports*, vol. 12, p. 14216, 2022. [Online]. Available: https://doi.org/10.1038/s41598-022-18173-0

[3] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021. [Online]. Available: https://doi.org/10.1038/s41592-020-01008-z

[4] K. Noo, "Image selective segmentation under geometrical constraints using an active contour approach," *Communications in Computational Physics*, vol. 7, no. 4, pp. 759–778, 2010. [Online]. Available: https://doi.org/10.4208/cicp.2009.09.026

[5] R. Adams and L. Bischof, "Seeded region growing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 6, pp. 641–647, Jun. 1994. [Online]. Available: https://doi.org/10.1109/34.295913

[6] L. M. Vincent and P. Soille, "Watersheds in digital spaces: An efficient algorithm based on immersion simulations," *IEEE Trans. Pattern Anal.*

Mach. Intell., vol. 13, pp. 583–598, 1991. [Online]. Available: https://api.semanticscholar.org/CorpusID:15436061

[7] P. D. Grunwald, "A tutorial introduction to the minimum description length principle," 2005, arXiv:math/0406077 [math.ST]. [Online]. Available: https://doi.org/10.48550/arXiv.math/0406077

[8] D. Mumford and J. Shah, "Optimal approximation by piecewise smooth functions and associated variational problems," *Communications on Pure and Applied Mathematics*, vol. 42, pp. 577–685, 1989. [Online]. Available: https://doi.org/10.1002/cpa.3160420503

[9] W. Tao and X. cheng Tai, "Multiple piecewise constant active contours for image segmentation using graph cuts optimization," UCLA, Center for Applied Mathematics (CAM), Los Angeles, CA, CAM Report 09-13, Feb. 2009. [Online]. Available: https://ww3.math.ucla.edu/camreport/cam09-13.pdf

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

[11] R. B. Potts, "Some generalized order–disorder transformations," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 48, no. 1, pp. 106–109, 1952.

[12] K. Wei, K. Yin, X.-C. Tai, and T. F. Chan, "A new region force for variational models in image segmentation and high-dimensional data clustering," *arXiv preprint arXiv:1704.08218*, 2017.

[13] H. Zhu, F. Meng, J. Cai, and S. Lu, "Beyond pixels: A comprehensive survey from bottom-up to semantic image segmentation and cosegmentation," *Journal of Visual Communication and Image Representation*, vol. 34, pp. 12–27, 2016.

[14] M. E. Rayed, S. M. S. Islam, S. I. Niha, J. R. Jim, M. M. Kabir, and M. F. Mridha, "Deep learning for medical image segmentation: State-of-the-art advancements and challenges," *Informatics in Medicine Unlocked*, vol. 47, p.

101504, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352914824000601

[15] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015, pp. 3431–3440.

[16] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, December 2017. [Online]. Available: https://doi.org/10.1016/j.media.2017.07.005

[17] N. H. Strickland, "Pacs (picture archiving and communication systems): filmless radiology," *Archives of Disease in Childhood*, vol. 83, no. 1, pp. 82–86, July 2000. [Online]. Available: https://doi.org/10.1136/adc.83.1.82

[18] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, E. A. Kazerooni, H. MacMahon, E. J. R. Van Beek, D. Yankelevitz, A. M. Biancardi, P. H. Bland, M. S. Brown, R. M. Engelmann, G. E. Laderach, D. Max, R. C. Pais, D. P. Y. Qing, R. Y. Roberts, A. R. Smith, A. Starkey, P. Batra, P. Caligiuri, A. Farooqi, G. W. Gladish, C. M. Jude, R. F. Munden, I. Petkovska, L. E. Quint, L. H. Schwartz, B. Sundaram, L. E. Dodd, C. Fenimore, D. Gur, N. Petrick, J. Freymann, J. Kirby, B. Hughes, D. Cavanaugh, K. S. Berbaum, S. Cappabianca, V. Lombardo, F. Macri, J. Guo, M. Salganicoff, L. Hadjiiski, G. D. Tourassi, H.-P. Chan, P. N. Cascade, F. W. Samuelson, B. Helba, M. D. Heath, M. H. Kuhn, E. Dharaiya, R. Burns, D. S. Fryd, M. Salganicoff, and L. P. Clarke, "The lung image database consortium (lidc) and image database resource initiative (idri): A completed reference database of lung nodules on ct scans," *Medical Physics*, vol. 38, no. 2, pp. 915–931, 2011.

[19] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in mri images," *IEEE Transactions on Medical*

Imaging, vol. 35, no. 5, pp. 1240–1251, 2016.

[20] A. Vitti, "The mumford–shah variational model for image segmentation: An overview of the theory, implementation and use," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 69, pp. 50–64, 2012. [Online]. Available: https://doi.org/10.1016/j.isprsjprs.2012.02.005

[21] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems*, ser. Oxford Mathematical Monographs. Oxford: Oxford University Press, 2000.

[22] E. D. Giorgi, M. Carriero, and A. Leaci, "Existence theorem for a minimum problem with free discontinuity set," *Archive for Rational Mechanics and Analysis*, vol. 108, pp. 195–218, 1989.

[23] E. S. Brown, T. F. Chan, and X. Bresson, "A convex approach for multi-phase piecewise constant mumford-shah image segmentation," UCLA Department of Mathematics, Tech. Rep. UCLA CAM Report 09-67, 2009, supported in part by NSF IIS-0914580 and ONR N00014-09-1-0105.

[24] O. Veksler and Y. Boykov, "Sparse non-local crf," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018, pp. 5125–5133.

[25] T. D. Johnson, Z. Liu, A. J. Bartsch, and T. E. Nichols, "A bayesian non-parametric potts model with application to pre-surgical fmri data," *Statistical Methods in Medical Research*, vol. 22, no. 4, pp. 364–381, 2013.

[26] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, ser. Lecture Notes in Computer Science, vol. 9901. Springer, 2016, pp. 424–432.

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

[28] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation," in *Deep Learning and Data Labeling for Medical Applications (DLMIA 2016)*, ser. Lecture Notes in Computer Science. Springer, 2016, vol. 10008, pp. 179–187.

[29] 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 *Computer Vision – ECCV 2022 Workshops*, ser. Lecture Notes in Computer Science, L. Karlinsky, T. Michaeli, and K. Nishino, Eds., vol. 13803. Springer, Cham, 2023, pp. 272–288. [Online]. Available: https://doi.org/10.1007/978-3-031-25066-8_9

[30] 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 preprint arXiv:2304.02643*, 2023. [Online]. Available: https://arxiv.org/abs/2304.02643

[31] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 654, 2024. [Online]. Available: https://doi.org/10.1038/s41467-024-44824-z

[32] S. Rui, L. Chen, Z. Tang, L. Wang, M. Liu, S. Zhang, and X. Wang, "Multi-modal vision pre-training for medical image analysis," *arXiv preprint arXiv:2303.11435*, 2023. [Online]. Available: https://github.com/openmedlab/BrainMVP

[33] R. J. Chen, C. Chen, Y. Li, T. Y. Chen, A. D. Trister, R. G. Krishnan, and F. Mahmood, "Scaling vision transformers to gigapixel images via hierarchical self-supervised learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022, pp. 16 144–16 155. [Online]. Available: https://openaccess.thecvf.com/content/

CVPR2022/html/Chen_Scaling_Vision_Transformers_to_Gigapixel_Images_via_Hierarchical_Self-Supervised_Learning_CVPR_2022_paper.html

[34] J. Wang, Z. Liu, L. Zhao, Z. Wu, C. Ma, S. Yu, H. Dai, Q. Yang, Y. Liu, S. Zhang, E. Shi, Y. Pan, T. Zhang, D. Zhu, X. Li, X. Jiang, B. Ge, Y. Yuan, D. Shen, T. Liu, and S. Zhang, "Review of large vision models and visual prompt engineering," *arXiv preprint arXiv:2307.00855*, 2023. [Online]. Available: https://arxiv.org/abs/2307.00855

[35] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[36] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 2002.

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[38] J. S. Bridle, "Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition," in *Neurocomputing*. Springer Berlin Heidelberg, 1990, pp. 227–236.

[39] Y. Gao, Y. Jiang, Y. Peng, F. Yuan, X. Zhang, and J. Wang, "Medical image segmentation: A comprehensive review of deep learning-based methods," *Tomography*, vol. 11, no. 5, p. 52, Apr. 2025.

[40] Y. Xu, R. Quan, W. Xu, Y. Huang, X. Chen, and F. Liu, "Advances in medical image segmentation: A comprehensive review of traditional, deep learning and hybrid approaches," *Bioengineering*, vol. 11, no. 10, 2024.

[41] M. et al., "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE Transactions on Medical Imaging*, 2015.

[42] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open access series of imaging studies (oasis): cross-sectional mri data in young, middle-aged, nondemented, and demented older adults," *Journal of Cognitive Neuroscience*, vol. 19, no. 9, pp. 1498–1507, 2007.

[43] U. Baid, S. Ghodasara, S. Mohan, M. Bilello, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer, F. C. Kitamura, S. Pati, L. M. Prevedello, J. D. Rudie, C. Sako, R. T. Shinohara, T. Bergquist, R. Chai, J. Eddy, J. Elliott, W. Reade, T. Schaffter, M. Yu, J. Zheng, A. W. Moawad, L. O. Coelho, O. McDonnell, E. Miller, F. E. Moron, M. C. Oswood, R. Y. Shih, L. Siakallis, Y. Bronstein, J. R. Mason, A. F. Miller *et al.*, "The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification," *arXiv preprint arXiv:2107.02314*, 2021.

[44] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open access series of imaging studies (oasis): Cross-sectional mri data in young, middle aged, nondemented, and demented older adults," *Journal of Cognitive Neuroscience*, vol. 19, no. 9, pp. 1498–1507, 2007.

[45] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," 2019. [Online]. Available: https://arxiv.org/abs/1912.01703

[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017. [Online]. Available: https://arxiv.org/abs/1412.6980

[47] K. Chen, H. Han, J. Wei, and Y. Zhang, "A novel few-shot learning framework for supervised diffeomorphic image registration network," *IEEE Transactions on Medical Imaging*, 2020. [Online]. Available: https://github.com/weijunping111/RDG-TMI.git