Research on Medical Image Segmentation Based on Deep Learning Methods

Zhou Yuxiang

A Thesis submitted to Tokushima University in partial fulfillment of the requirements for the degree of Doctor of Philosophy

March, 2024

Department of Information Science and Intelligent Systems
Graduate School of Advanced Technology and Science
Tokushima University, Japan

Contents

List of Tables

List of Figures

Abstract

Medical imaging involves the technique and process of generating visual representations of a patient's body for clinical analysis and medical intervention. Healthcare professionals heavily depend on medical images for accurate diagnosis and treatment. In clinical practice, segmentation is typically performed manually. However, when processing a vast number of medical images, the quality of segmentation can vary based on the expertise of the medical professional. This variability underscores the need for a more consistent and efficient method to enhance the performance of segmentation tasks. Amassing such datasets is a complex task and is often beyond the capacity of a single institution within a limited timeframe. As a result, medical datasets tend to be smaller in size and can exhibit inconsistencies. Another notable characteristic of medical images is the imbalance between the foreground and background areas. Unlike natural scene images, where the foreground and background might be more balanced, medical images often have a much larger background compared to the foreground.

Extensive research has been conducted over the years to achieve fully automatic segmentation of the region of interest in medical images, aiming to improve efficiency and accuracy in comprehending such images. Based on deep learning and the server's powerful data processing capabilities, pixel-level processing and segmentation methods are usually used to process medical images, especially for brain tumor segmentation in 3D MRI scans, and for organ and lesion segmentation in 2D images. With the continuous development of deep learning, various neural network models have made remarkable achievements in semantic segmentation, stimulating research interest in medical image segmentation using deep learning.

In this work, we propose an Multi-Attention and Depthwise Separable Convolution U-Net (MDSU-Net), a variation of the U-Net, for 2D medical image segmentation. MDSU-Net incorporates both multi-attention and DSC layer for improved performance. The multi-attention module within our framework utilizes dual attention and attention gates to capture rich contextual information and fuse features of different convolutional

layers. MDSU-Net uses DSC layer to reduce model complexity without degrading model performance, which is suitable for different segmentation tasks. MDSU-Net registers a Dice score of 0.7055 on ATLAS, 0.9760 on CHAOS, and 0.8883 on NERVE, respectively. Notably, these scores outperform the state-of-the-art (SoTA) benchmarks.

Additionally, in terms of 3D medical image segmentation, we propose Dual Encoder U-Net (DEU-Net), which uses Transformer and CNN respectively to extract 3D medical image features in the encoder. Transformer is a pre-trained model in BTCV, which improves its ability to capture contextual features of medical images and increases the learning speed. Besides, we introduce CBAM with each convolutional layer in the encoder part to enhance CNN's feature extraction capabilities for 3D medical images. To fuse the two kinds of features, we proposed a Dual Feature Fusion Module (DFFM) to fuse the features extracted from the Transformer and CNN in the encoder, making full use of the feature extraction capabilities of the two extractors for datasets of different sizes. We compare DEU-Net with other SoTA methods on the BraTS 2020 and BraTS 2021 datasets, and the results show that the performance of DEU-Net has improved in 3D medical image segmentation task.

Keywords: U-Net, Multi-attention, Transformer, Convolutional neural network, 2D medical image segmentation, 3D medical image segmentation

1 Introduction

1.1 Motivation

Medical imaging involves the technique and process of generating visual representations of a patient's body for clinical analysis and medical intervention [1], the main approach involves examining a series of 2D slice images or 3D images to identify pathological conditions. This often requires the expertise of medical professionals for accurate interpretation. Utilizing computer image processing techniques to analyze and process 2D slice images allows for segmentation, extraction, 3D reconstruction, and display of human organs, soft tissues, and pathological structures. This can assist medical professionals in qualitatively and even quantitatively analyze pathological conditions and other areas of interest, significantly enhancing the accuracy and reliability of medical diagnosis.

Medical image segmentation has been widely discussed and concerned in the field of image processing, which has become a basic component and a crucial stage of image processing [2,3]. Healthcare professionals heavily depend on medical images for accurate diagnosis and treatment. However, manual interpretation and analysis of these images can be time-consuming and prone to inaccuracies, especially when the interpreter lacks proper training. Extensive research has been conducted over the years to achieve fully automatic segmentation of the region of interest in medical images, aiming to improve efficiency and accuracy in comprehending such images.

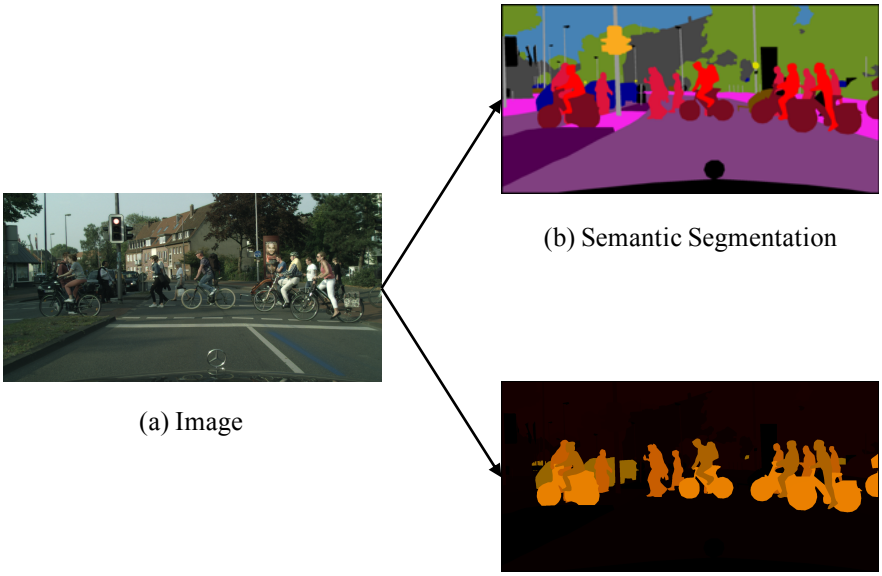
Based on deep learning and the server's powerful data processing capabilities, pixel-level processing and segmentation methods are usually used to process medical images. Recently, Convolutional Neural Networks (CNNs) have achieved advanced performance in a wide range of visual recognition tasks [4–7]. These deep models dominate medical image segmentation and achieve excellent performance in a wide range of applications, such as brain tumor segmentation [8]. U-Net [9] is a variant of CNN, which has achieved great performance in medical segmentation tasks. U-Net adapts a skip connection operation to connect downsampling layers and upsampling layers, so that the segmentation results are more accurate.

Medical image segmentation is a crucial research area within medical image processing. The advancement and application of its technology have the potential to enhance the functioning of medical systems, alleviate the workload of medical professionals, improve diagnostic efficiency for patients, and ease doctor-patient relationships. Additionally, this technology plays a crucial supporting role in medical education, surgical planning, surgical simulation, and various medical research endeavors.

1.2 Background

There are two main categories of general image segmentation tasks: semantic segmentation [10,11] and instance segmentation [12,13]. As shown in Fig. 1.1, image semantic segmentation is a pixel-level classification task that involves predicting the category for each pixel in an image. In contrast, image instance segmentation not only involves pixel-level classification but also requires distinguishing between different instances, meaning independently segmenting different objects within the same category. In the field of medical image segmentation, there are some peculiarities. Due to significant differences in appearance between different organs or tissues in medical images, the emphasis is often on semantic segmentation rather than instance segmentation. Therefore, medical image segmentation tasks usually involve semantic segmentation of medical images. Currently, the main medical image segmentation tasks include abdominal multi-organ segmentation [14], brain tumor segmentation [15], optic disc segmentation [16], cell segmentation [17], skin cancer segmentation [18], and lung nodule segmentation [19]. These tasks are crucial for precisely locating and quantifying different structures and abnormalities in medical images.

Image segmentation refers to the process of distinguishing and separating different regions in an image that hold specific meaning. These regions are non-overlapping, and each region exhibits a certain level of consistency or homogeneity. The goal is to divide the image into meaningful and distinct segments, facilitating the analysis and understanding of different structures or components within the image. This technique is fundamen-



(a) Image

(b) Semantic Segmentation

(c) Instance Segmentation

Fig. 1.1. Semantic segmentation and instance segmentation.

tal for various applications, especially in fields like medical imaging, where it is used for tasks such as identifying organs, tumors, or other relevant structures. Medical image segmentation has been widely discussed and concerned in the field of image processing, which has become a basic component and a crucial stage of image processing [2, 3]. Medical image segmentation is an indispensable technique for extracting quantitative information from medical images, enabling the identification of specific tissues or structures. It serves as a crucial preprocessing step and prerequisite for visualization. Segmented images find widespread applications in various contexts, including quantitative analysis of tissue volumes , diagnostics, localization of pathological tissues, anatomical structure learning, treatment planning, local effects correction in functional imaging data, and computer-guided surgeries.

Medical image segmentation still faces significant challenges, which lies in the inherent complexity and diversity of medical images. The intricate shapes of human tissue structures and the vast individual variations pose additional difficulties for medical image segmentation. Variances in patient age, differences in imaging equipment, and even re-

gional disparities among medical institutions can all impact the overall quality of medical image slices to varying degrees. The diversity in types and storage formats of medical images, as well as the multitude of lesions or organ locations in patients, presents a formidable challenge to the universality of segmentation methods. Inevitably, medical images exhibit features such as blurriness and unevenness, making segmentation methods less universally applicable compared to natural scene images.

1.2.1 Tasks in Medical Image Segmentation

Medical imaging techniques mainly include Magnetic Resonance Imaging (MRI), Computed Tomography (CT), and Ultrasound Imaging (UI) [20–22]. MRI images are a measure of the size of the magnetic resonance signals generated by hydrogen nuclei in human tissues, organs and lesions under the action of an external strong magnetic field. Computers perform 3D image reconstruction based on the information data received by signal detectors, offering highly detailed images of human soft tissue anatomy and abnormalities. CT images involve scanning a specific part of the human body using an X-ray beam. The computer then utilizes the received X-ray signal data to reconstruct corresponding 3D images of the human cross-section, providing clear images of human bone tissue anatomy and abnormalities. UI employs ultrasound beams to scan the human body, obtaining images of visceral organs through the reception and processing of reflected signals.

Compared to monomodal medical images, multimodal medical images can provide doctors with rich complementary information. Among these, CT images are commonly used for the diagnosis and imaging of musculoskeletal disorders, such as bone tumors and fractures. MRI images, on the other hand, offer better soft tissue contrast. Multimodal MRI [23–25] can also offer supplementary information based on differences in acquisition parameters, including T1-weighted imaging (T1), contrast-enhanced T1-weighted imaging (T1ce), T2-weighted imaging (T2), and Fluid Attenuated Inversion Recovery (FLAIR) images. Taking brain tumors as an example, T2 and FLAIR are suitable for visualizing peritumoral edema, while T1 and T1ce images are suitable for delineating the

tumor core without peritumoral edema. Therefore, the use of multimodal imaging for segmentation can reduce information uncertainty and enhance both clinical diagnosis and segmentation accuracy.

Tab. 1.1. Commonly used datasets for medical image segmentation.

Datasets	Tasks	Modalities	Subjects	Formats
LiTS [26]	Liver/Liver tumors	CT	131+7	nii
Sliver07 [27]	Liver	CT	20+10	MetaImage
MSD [28]	Liver blood vessels	CT	443	nii
MSD	Lung	CT	96	nii
MSD	Colon cancer	CT	126+64	nii
MSD	Pancreatic tumors	CT	282+139	nii
Vessel12 [29]	Pulmonary blood vessels	CT	20	Raw
Covid19-ct-scans [30]	COVID-19 infection	CT	20	nii
Chaos [31]	Liver/Kidney/Spleen	CT+MRI	40CT+120MRI	dcm
MSD	Brain tumor	MRI	484+266	nii
MSD	Hippocampus	MRI	394	nii
BraTS 2020 [23–25]	Brain tumor	MRI	369+125	nii
BraTS 2021 [32]	Brain tumor	MRI	1251+219+530	nii
ATLAS [33]	Stroke	MRI	299	nii
ISLES 2022 [34]	Stroke	MRI	250+150	nii
EchoNet [35]	Heart	MRI	10300	nii
MMWHS [36]	Heart	CT/MRI	20CT+20MRI	nii
MSD	Left atrium	MRI	20+10	nii
NERVE [37]	Nerve	UI	5635	tif
DRIVE [38]	Fundus blood vessels	Picture	40	JPEG
STARE [39]	Fundus blood vessels	Picture	400	ppm.gz

In clinical diagnosis, medical images provide doctors with main patient condition information, and medical image segmentation facilitates clinical diagnosis and treatment. In clinical practice, segmentation is typically performed manually. However, when processing a vast number of medical images, the quality of segmentation can vary based

on the expertise of the medical professional. This variability underscores the need for a more consistent and efficient method to enhance the performance of segmentation tasks. One significant challenge in this domain is the acquisition and creation of high-quality datasets. Amassing such datasets is a complex task and is often beyond the capacity of a single institution within a limited timeframe. As a result, medical datasets tend to be smaller in size and can exhibit inconsistencies. Another notable characteristic of medical images is the imbalance between the foreground and background areas. Unlike natural scene images, where the foreground and background might be more balanced, medical images often have a much larger background compared to the foreground [40]. Medical image segmentation datasets commonly used for medical image segmentation are listed in Tab. 1.1.

1.2.2 Method for Medical Image Segmentation

Based on deep learning and powerful data processing capabilities of servers, pixel-level processing and segmentation of medical images are generally performed [41]. U-Net adapts a skip connection operation to connect downsampling layers and upsampling layers, so that the segmentation results are more accurate [42]. However, the receptive field of U-Net decreases during training, which results in the inability to extract wider and richer contexts [43]. Many extensions of the U-Net have emerged to address image segmentation in practical applications. Alom et al. [44] proposed to extend the U-Net architecture with RCNN and recurrent residual CNN, but it did not improve the ability to extract long-range contextual features. Gu et al. [45] proposed a contextual encoder network (CE-Net) to extract more advanced information for 2D medical images. Sun et al. [46] proposed a Shape Attentive U-Net (SAU-Net) to improve model robustness and interpretability. Chen et al. [47] proposed a Multi-View U-Net (MV U-Net), which effectively improved the robustness of the network. Due to the small amount of data in medical images datasets, traditional U-Net is not suitable for medical image segmentation tasks.

Attention mechanisms [48] can effectively integrate local and global features [49]

and are widely used in segmentation tasks [50]. Attention mechanisms focus on the most relevant features, which avoid using multiple similar or repeated feature maps and significantly extract task-related features [51–53]. X-Net [54] and MSDF-Net [55] introduced a single attention mechanism to integrate contextual features specifically for stroke lesion segmentation. However, the combination of a solitary self-attention module with convolutional layers has shown limited success in augmenting the model's capability to model non-local features. In response to this limitation, MALUNet [56] and RemaNet [57] incorporated multi-attention mechanisms for skin lesion segmentation. Despite these advancements, achieving a universal solution for various segmentation tasks remains a challenge. Addressing this, TA-Net [58] and MAD-Net [59] employed multi-attention to bolster the model's efficacy across a spectrum of segmentation tasks. It's worth noting, however, that while multi-attention can enhance performance, it also considerably increases the model's parameters. Moreover, current implementations do not sufficiently explore feature fusion across different convolutional layers.

Inspired by the good performance of Visual Transformers (ViT) in the field of natural images [60], there have been many studies trying to combine visual Transformers with medical image segmentation and achieved performance close to or even better than CNN on different datasets. Several researchers have ventured into incorporating Transformers within the realm of medical image segmentation. Notable models such as TransUNet [61], Swin-Unet [62], and UNETR [63] have harnessed the power of self-attention to capture long-range dependencies inherent in medical images. For multi-scale medical images, DS-TransUNet [64] uses Swin Transformer to extract features in the encoder and fuses features extracted by CNN in the decoder. As shown in the Fig. 1.2, we list the main U-Net-based models for medical image segmentation according to the 2015-2023 timeline. In recent years, models using Transformer for medical image segmentation have become increasingly popular. However, it lacks feature fusion in the encoder. However, a notable challenge with Transformer-based models is their inherent complexity. They often necessitate large datasets for training to achieve optimal performance. Consequently, Transformers may not be the ideal choice for every medical segmentation task.

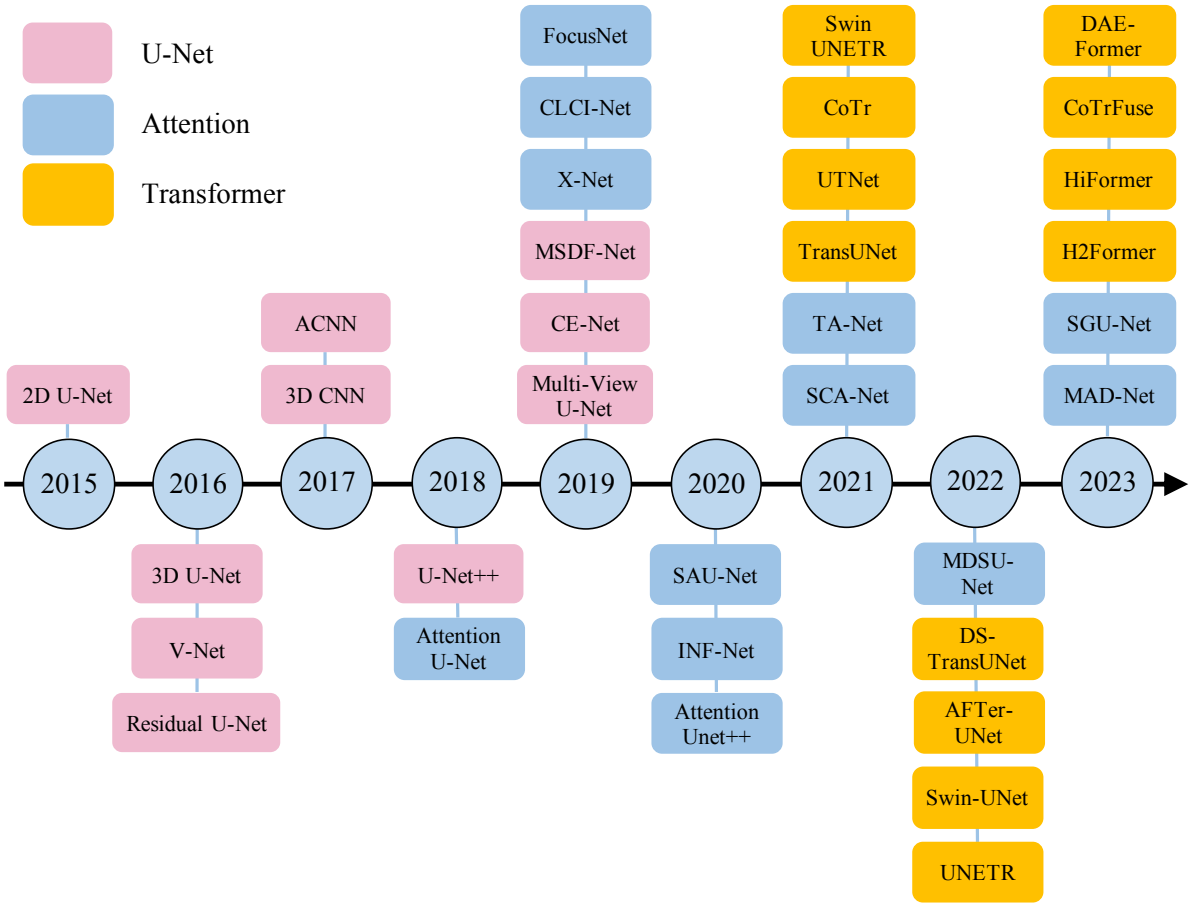


Fig. 1.2. The main U-Net-based models for medical image segmentation according to the 2015-2023 timeline.

1.3 Research Contents and Contributions

1.3.1 MDSU-Net for 2D Medical Image Segmentation

To solve the problem in the field of 2D medical image segmentation, we propose an Multi-Attention and Depthwise Separable Convolution U-Net (MDSU-Net), a variation of the U-Net, for 2D medical image segmentation. MDSU-Net incorporates both multi-attention and DSC layer for improved performance. The multi-attention module within our framework utilizes dual attention and attention gates to capture rich contextual information and fuse features of different convolutional layers. MDSU-Net uses DSC layer to reduce model complexity without degrading model performance, which is suitable for different segmentation tasks.

1.3.2 DEU-Net for 3D Medical Image Segmentation

To solve the problem in the field of 3D medical image segmentation, we propose Dual Encoder U-Net (DEU-Net), which uses Transformer and CNN respectively to extract 3D medical image features in the encoder. Transformer is a pre-trained model in BTCV, which improves its ability to capture contextual features of medical images and increases the learning speed. Besides, we introduce CBAM with each convolutional layer in the encoder part to enhance CNN's feature extraction capabilities for 3D medical images. To fuse the two kinds of features, we proposed a Dual Feature Fusion Module (DFFM) to fuse the features extracted from the Transformer and CNN in the encoder, making full use of the feature extraction capabilities of the two extractors for datasets of different sizes.

The main contributions of our work are summarized as follows:

1. We propose a novel MDSU-Net for 2D medical image segmentation, which consists of multi-attention and DSC layer. One DSC layer consists of 3 DSC Blocks and a residual connection. DSC blocks are used to reduce model complexity.

2. We propose a novel attention gate and introduce a dual attention to construct multi-attention. Dual attention captures rich contextual information, and attention gates

fuse features of different convolutional layers. Our multi-attention can accurately predict and localize the small lesion area.

3. Our MDSU-Net is versatile and can be employed for 2D medical image segmentation across various 2D image types.

4. We propose a novel DEU-Net, which uses pre-trained Transformer and CNN respectively to extract 3D medical image features in the encoder, which improves its ability to capture contextual features of medical images and increases the learning speed. We introduce CBAM with each convolutional layer in the encoder part to enhance CNN's feature extraction capabilities for 3D medical images.

5. To fuse the two kinds of features, we proposed a Dual Feature Fusion Module (DFFM) to fuse the features extracted from the Transformer and CNN in the encoder, making full use of the feature extraction capabilities of the two extractors for datasets of different sizes.

1.4 Thesis Organizations

This paper focuses on the exploration and study of medical image segmentation models, specifically investigating 2D and 3D image segmentation models in response to the diverse dimensions of medical images. This includes research on feature extraction for both 2D and 3D images, a study on the generality of segmentation models for different modal images, as well as the extraction and fusion of features from different modal images. The organizational framework of this paper is as follows:

Chapter 1: Introduction

In this chapter, we discuss the motivation, background of this research and introduce the research contents and contributions of our work.

Chapter 2: Related Work

In this chapter, we introduced the basic concepts of medical image segmentation, introduced the basic models and principles used for medical image segmentation, and

variant networks. We also introduced the principles and applications of attention mechanisms and Transformers for image segmentation.

Chapter 3: MDSU-Net for 2D Medical Image Segmentation

In this chapter, to improve the feature extraction and contextual relationship capabilities of traditional U-Net for 2D medical images, we propose an MDSU-Net, a variation of the U-Net, for medical image segmentation. MDSU-Net incorporates both multi-attention and DSC layer for improved performance. The multi-attention module within our framework utilizes dual attention and attention gates to capture rich contextual information and fuse features of different convolutional layers. MDSU-Net uses DSC layer to reduce model complexity without degrading model performance, which is suitable for different segmentation tasks.

Chapter 4: DEU-Net for 3D Medical Image Segmentation

In this chapter, to better integrate multi-modal features for 3D medical images and combine the advantages of Transformer and CNN, We propose Dual Encoder U-Net (DEU-Net), which uses pre-trained Transformer and CNN respectively to extract medical image features in the encoder. We introduce CBAM with each convolutional layer in the encoder part to enhance CNN's feature extraction capabilities for 3D medical images. To fuse the two kinds of features, we proposed a Dual Feature Fusion Module (DFFM) to fuse the features extracted from the Transformer and CNN in the encoder, making full use of the feature extraction capabilities of the two extractors for 3D medical image.

Chapter 5: Conclusion and Future work

In this chapter, we summarize the research content and discuss the future work.

2 Related Work

2.1 CNN for Medical Image Segmentation

Based on deep learning and powerful data processing capabilities of servers, pixel-level processing and segmentation of medical images are generally performed [41]. Compared with traditional CNN, fully convolutional network (FCN) [65] is composed solely of convolutional layers, eliminating any fully connected layers at the network's end. Additionally, the feature maps from intermediate layers can be resized to match the dimensions of the input image. Therefore, the predictions of FCN have a spatial one-to-one correspondence with the input image, which greatly promotes semantic segmentation research. Based on the FCN network, the U-Net [9] network was proposed, and other networks based on U-Net appeared at the same time. The structure of U-Net is illustrated in Fig. 2.1, consisting of an Encoder and a Decoder. These two components together form a U-shaped architecture, hence the name U-Net.

Encoder: The Encoder is primarily responsible for capturing contextual information in the image. It comprises multiple convolutional layers, activation functions, and max-pooling layers. The sequence involves two consecutive operations of 3×3 convolution and 2×2 max-pooling. Each convolutional layer processes the input feature map to extract features. As the network depth increases, the size of the feature maps gradually decreases while the channel count increases, aiding the model in learning higher-level abstract features.

Decoder: The Decoder is mainly tasked with precisely locating the boundaries of objects in the image. It consists of multiple upsampling layers, convolutional layers, and activation functions. The sequence involves one 2×2 upsampling and two 3×3 convolutions to restore the size of the segmentation map. During the upsampling process, the size of the feature maps gradually enlarges, recovering to the original input image's resolution, allowing the model to capture high-level features. Simultaneously, the channel count of the feature maps decreases, losing localization information. Therefore, after each upsampling, the output of each upsampling layer is fused with the corresponding level of

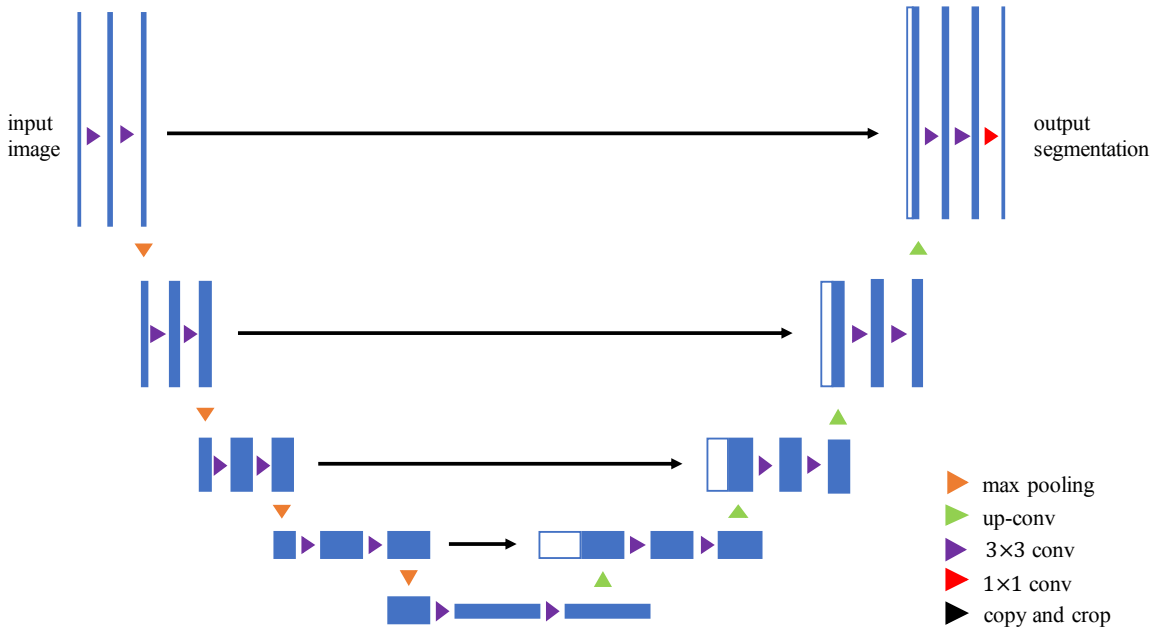


Fig. 2.1. Overview of U-Net.

feature map in the contracting path, a process known as skip connections.

Skip Connections: Skip connections are a key feature of U-Net, combining the feature maps from the Encoder with the corresponding levels in the Decoder. This helps retain more detailed information in the segmentation result, enhancing segmentation accuracy. Specifically, the feature maps from the Encoder are concatenated channel-wise with the feature maps at the corresponding level in the Decoder.

Output Layer: At the end of the expansive path, there is a convolutional layer used to transform the feature maps into per-pixel classification results. This output layer typically employs a 1×1 convolutional kernel, with the output channel count set based on the number of classes for segmentation. Finally, employing per-pixel classification strategy generates the segmentation result.

Since the introduction of U-Net, CNN-based models have set the benchmark for performance on various medical image segmentation tasks in both 2D and 3D [50, 66]. In contrast, 3D methods directly exploit full volumetric images represented by a series of 2D slices or modalities [67]. Most of the current 3D methods are based on the U-Net

framework, which uses the U-shaped structure of the encoder and decoder to better learn the features of different convolutional layers.

U-Net has exhibited strong performance for medical image segmentation [68], so that U-Net is widely adopted as fundamental framework for medical image segmentation [69, 70]. There are many variations of U-Net proposed to improve accuracy and stability of segmentation [71]. Multi-scale strategies were used to obtain different medical image features from different scales [72]. Chen et al. [73] proposed a MV U-Net for cardiac short-axis image segmentation. Fang et al. [74] proposed MIMO-FAN to employ multi-scale inputs to better exploit hierarchical information. Skip connection was also re-designed to reduce the semantic gap by merging feature maps from different encoders and decoders, like UNet++ [75, 76] and ResUNet++ [77].

For 3D medical images, professional 3D networks were proposed to deal with the feature extraction of multi-sequence 3D images. Zheng et al. [66] proposed HFA-Net, which exploited complementary information from 3D data. Chen et al. [67] proposed DMFNet, which reduced computational cost for real-time dense volume segmentation. Hatamizadeh et al. [63] proposed UNETR to capture global multi-scale information for 3D medical image. Yan et al. [78] proposed AFTer-UNet, which addressed segmentation by considering aspects within and between slices.

3D medical images are more complex and contextual features are more difficult to capture, some work achieves contextual feature extraction by adding fusion modules. Tseng et al. [79] proposed a deep convolutional encoder-decoder structure with fusion layers to segment multi-modal 3D medical images. Zhang et al. [80] proposed 3D contextual residual network (ConResNet) for accurate segmentation of 3D medical images. Sun et al. [81] effectively extracted features from multi-modal 3D MRI images by applying a multi-channel architecture to feature extraction. Despite achieving success, the limitation of these networks lies in their pixel-wise segmentation approach, which hinders their ability to capture global context, resulting in suboptimal performance on large-scale medical image datasets.

2.2 Attention Mechanism for Medical Image Segmentation

Attention refers to the direction and concentration of psychological activities towards specific objects, while attention is the measure of the degree of focus, consisting of breadth, stability, allocation, and transfer of attention. When observing or listening to something, individuals tend to selectively focus on what they consider important and ignore what they deem unimportant. Over time, an individual's visual attention shifts based on changes in the focus area, demonstrating the manifestation of attention. The process of weight allocation involves unconsciously focusing on a particular part or entity while disregarding others. Higher weights indicate a greater concentration of attention on important events, while lower weights weaken the significance of unimportant information, and adjustments to weight allocation occur continuously throughout this process.

The attention mechanism mimics biological observation behavior. It involves a comprehensive scan of the entire image to identify regions of interest, focusing on extracting feature information from those specific areas while suppressing irrelevant information. The concept of attention was initially introduced in the field of image recognition by Mnih et al. [82] in 2014, simulating the attention mechanism of the human brain.

The attention mechanism has become a common component in neural network architectures, widely applied in tasks such as image recognition and natural language processing [83, 84]. Currently, there are three main network frameworks that combine with the attention mechanism. First is the classic Encoder-Decoder framework, which is the framework most models use. Second is combining with memory networks, storing task-related information in auxiliary memory for retrieval when needed. The last is a special neural network structure that can capture long-distance dependencies without using Recurrent Neural Networks (RNNs) [85], which has found applications in certain scenarios.

The attention has been gradually introduced in medical image segmentation. Oktay et al. [86] proposed Attention Gate (AG) for medical imaging to learn objects of different sizes and shapes. Sinha et al. [87] introduced a Dual Attention to extract the feature of medical images. Henry et al. [88] introduced a Convolutional Block Attention Module

(CBAM) for brain tumor segmentation. Attention mechanism can extract contextual information and long-distance features of medical images. Below we introduce these three attention mechanisms in detail:

2.2.1 Attention Gate

A model based on AGs for medical image applications has been proposed, aiming to automatically learn to differentiate the shapes and sizes of targets. This AG-equipped model, during training, learns to suppress irrelevant regions and focus on meaningful salient features, proving effective for a specific task. This approach can be beneficial for tasks such as CNNs without explicit localization for organs and structures.The integration of AGs into standard CNN models is straightforward, requiring minimal additional computational overhead while significantly improving model sensitivity and accuracy.

The AG can extract rough image information through skip connection for image feature recovery [89]. As shown in Fig. 2.2, there are two inputs for AG. The first input is $x \in \mathbb{R}^{F\times H\times W\times D}$, where F, H, W, and D represent channel, height, width and depth, respectively. The second input is gating signal $g \in \mathbb{R}^{2F\times H\times W\times D}$. x is fed into a $1\times 1\times 1$ convolution to get $x_1 \in \mathbb{R}^{F\times H\times W\times D}$. At the same time, g is fed into an upsampling layer to get $g_1 \in \mathbb{R}^{F\times H\times W\times D}$. In order to fuse features from encoder and decoder, we add x_1 with g_1 and apply a rectified linear unit (ReLU) to get $r \in \mathbb{R}^{F\times H\times W\times D}$ by the following formula:

$$r = \max(0, x_1 + g_1). \tag{2.1}$$

r is fed into a $1\times 1\times 1$ convolution to get $r_1 \in \mathbb{R}^{2F\times H\times W\times D}$ and apply a Sigmoid to get $s \in \mathbb{R}^{2F\times W\times H\times D}$. In order to adjust the dimension of s to be unified with x, s is fed into an upsampling layer to get $s_1 \in \mathbb{R}^{F\times H\times W\times D}$. s_1 is added with x to get $\hat{x} \in \mathbb{R}^{F\times H\times W\times D}$.

$$\hat{x} = \mathrm{up}(\frac{1}{1+e^{-r_1}}). \tag{2.2}$$

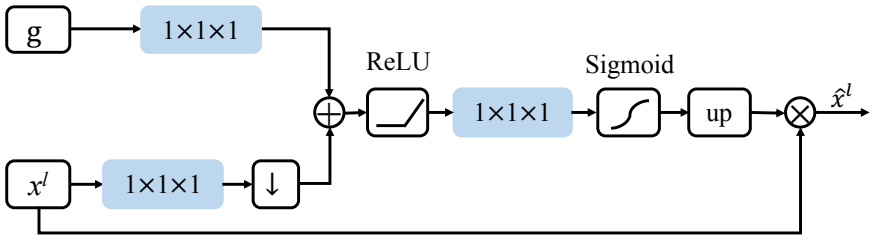


Fig. 2.2. Schematic of attention gate.

Additionally, downsampling the input feature map to the dimension of the gate signal is carried out, effectively reducing dimensionality. The corresponding linear transformation decouples the feature map and maps it to a lower-dimensional space for gate operations. The model enforces distinctiveness in the semantic representation of intermediate feature maps at each image scale, ensuring that attention units at different scales have the capability to influence responses to foreground content across a wide range.

2.2.2 Dual Attention

Dual Attention mechanism consists of position attention module and channel attention module. This model can adaptively integrate local features and their global dependencies. Specifically, the authors augment the traditional FCN with two types of attention modules that simulate semantic interdependencies in both spatial and channel dimensions. The position attention module selectively aggregates features from each position by weighting and summing features across all positions, mimicking semantic interdependencies in spatial dimensions. Similar features are considered correlated regardless of their spatial separation. Simultaneously, the channel attention module selectively emphasizes mutually dependent channel mappings by integrating correlated features across all channel maps. These attention modules collectively capture global information in the image.

In top part of Fig. 2.3, the input is $F \in \mathbb{R}^{C\times W\times H}$, where C, W, and H represent the channel, width and height, respectively. In the first path, F is reshaped to $I_0 \in \mathbb{R}^{(W\times H)\times C/8}$. In the second path, F is convoluted and transposed to $I_1 \in \mathbb{R}^{C/8\times(W\times H)}$

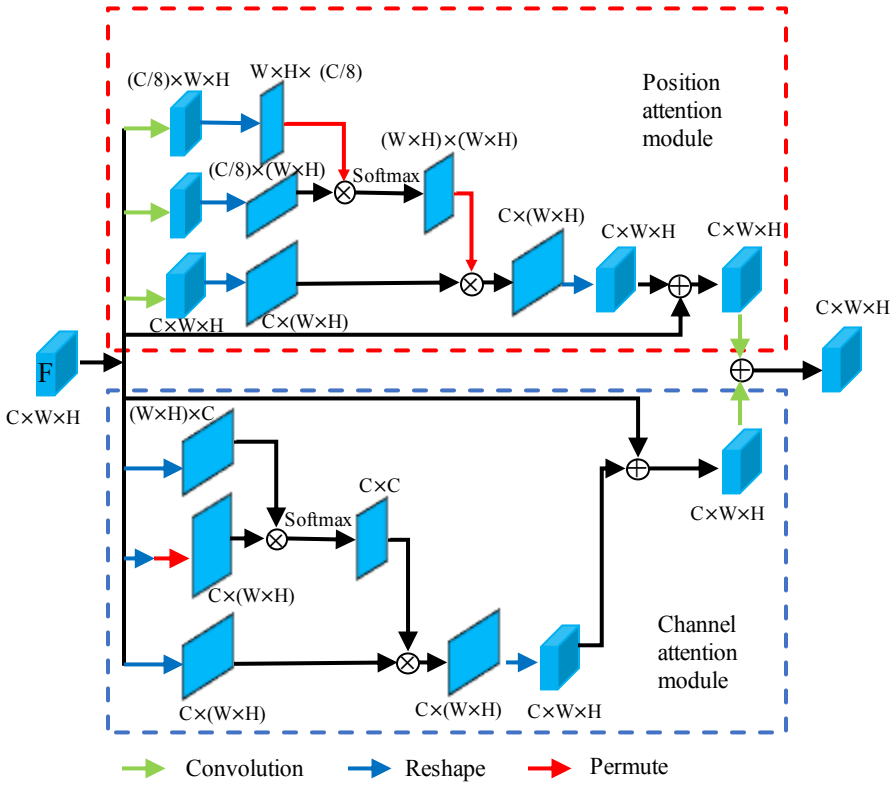


Fig. 2.3. Schematic of dual attention.

to diminish the impact of channel count on position representation of features. The multiplication operation is applied to I_0 and I_1, followed by a softmax to focus on extracting position-sensitive features, yielding $P \in \mathbb{R}^{(W\times H)\times(W\times H)}$ as follows:

$$P_{i,j} = \frac{\exp(I_{0,i} \cdot I_{1,j})}{\sum_{i=1}^{W\times H} \exp(I_{0,i} \cdot I_{1,j})}, \tag{2.3}$$

where $P_{i,j}$ measure j^{th} position's impact on i^{th} position. In the third path, F is convoluted and reshaped to $I_2 \in \mathbb{R}^{C\times(W\times H)}$ to enhance feature extraction capabilities. Finally, I_2 is multiplied by P, and the result is fused with F to get position attention map to aggregate position contextual information, which is $I_P \in \mathbb{R}^{C\times W\times H}$ as follows:

$$I_{P,j} = \varphi_1 \sum_{i=1}^{W\times H} P_{i,j} I_{2,j} + F_j, \tag{2.4}$$

where φ_1 is initialized to 0.

In bottom part of Fig. 2.3, $F \in \mathbb{R}^{C\times W\times H}$ is reshaped and permuted to get $I_0 \in \mathbb{R}^{(W\times H)\times C}$, $I_1 \in \mathbb{R}^{C\times(W\times H)}$, and $I_2 \in \mathbb{R}^{C\times(W\times H)}$, respectively. I_0 is multiplied by I_1 to eliminate dimensional effects and highlight relationships between channels. The application of softmax enables us to obtain $C \in \mathbb{R}^{C\times C}$, which is described as follows:

$$C_{i,j} = \frac{\exp(I_{0,i} \cdot I_{1,j})}{\sum_{i=1}^{C} \exp(I_{0,i} \cdot I_{1,j})}, \tag{2.5}$$

where $C_{i,j}$ measure the j^{th} channel's impact on i^{th} channel and model interdependencies between channels. Then we multiply C and I_2 and fuse with F to get channel attention map to reintroduce position information from medical images, which is $I_C \in \mathbb{R}^{C\times W\times H}$ as follows:

$$I_{C,j} = \varphi_2 \sum_{i=1}^{C} C_{i,j} I_{2,j} + F_j, \tag{2.6}$$

where φ_2 is also initialized to 0. Finally, I_P and I_C are summed to get dual attention features.

Lastly, the position and channel features are merged to leverage the Dual Attention to enhance representation of higher-level features, effectively capturing contextual information in medical images.

2.2.3 Convolutional Block Attention Module

CBAM is an attention mechanism that combines spatial attention module and channel attention module.

The channel attention module is shown in top part of Fig. 2.4. To mitigate the impact on the spatial dimensions, the input feature F undergoes global max-pooling and global average-pooling separately to compress spatial information, resulting in two feature maps. Subsequently, each of these feature maps undergoes multilayer perceptron (MLP) processing, yielding two new feature maps. The fusion of these two feature maps is then carried out, followed by a Sigmoid operation, generating channel attention features $M_c(F)$. The channel attention mechanism can be expressed as:

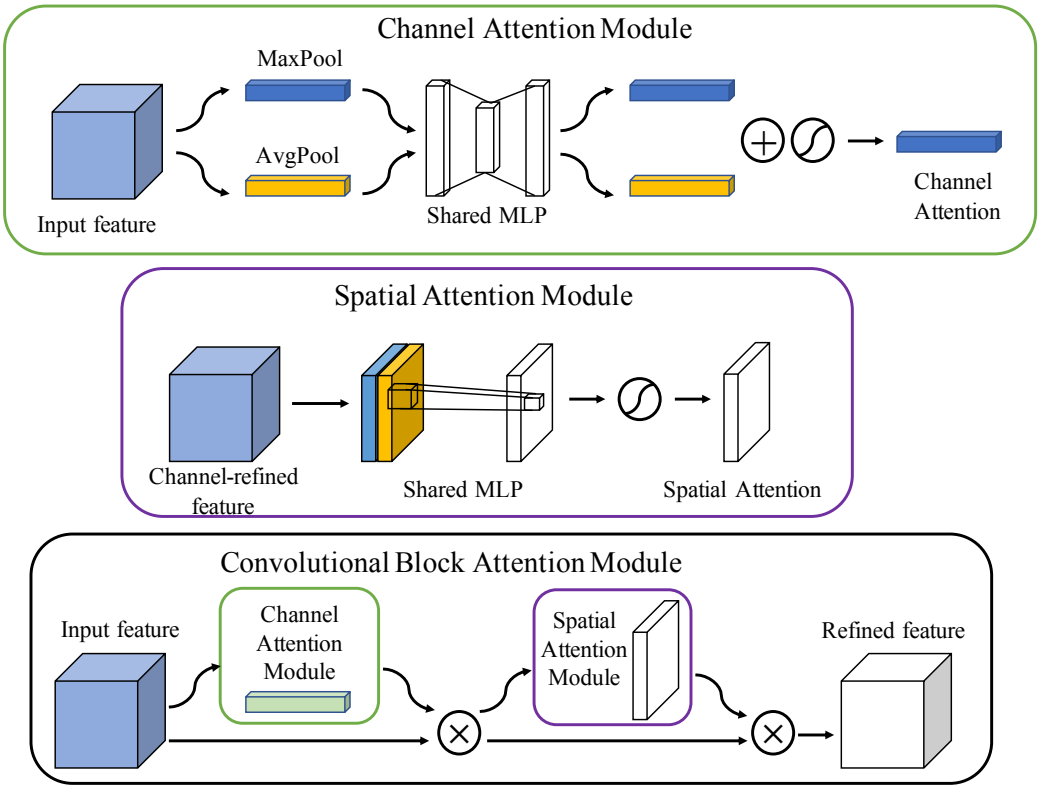


Fig. 2.4. Schematic of convolutional block attention module.

$$\begin{aligned} M_c(F) &= \sigma(\text{MLP}(\text{Avgpool}(F)) + \text{MLP}(\text{MaxPool}(F))) \\ &= \sigma(W_1(W_0(F^c_{\text{avg}})) + W_1(W_0(F^c_{\text{max}}))), \end{aligned} \tag{2.7}$$

where σ represents the Sigmoid, $W_0 \in \mathbb{R}^{C/r \times C}$, and $W_1 \in \mathbb{R}^{C \times C/r}$.

As shown in bottom part of Fig. 2.4, through residual connections, the channel attention features $M_c(F)$ and the input feature F are merged, producing the input features required for the subsequent spatial attention module.

The spatial attention module, as depicted in middle part of Fig. 2.4, is designed to mitigate the impact on channel dimensions. The input feature $F^{'}$ is fed through global max-pooling and global average-pooling to compress channel information, resulting in two feature maps. These two feature maps are then concatenated to fuse the features. Subsequent convolution operations and a Sigmoid function generate spatial attention features $M_s(F^{'})$. The spatial attention mechanism can be expressed as:

$$\begin{aligned} M_s(F^{'}) &= \sigma(f^{7\times7}([\text{Avgpool}(F); \text{MaxPool}(F)])) \\ &= \sigma(f^{7\times7}([F^s_{\text{avg}}; F^c_{\text{max}}])), \end{aligned} \tag{2.8}$$

where σ represents the Sigmoid, $F^s_{\text{avg}} \in \mathbb{R}^{1\times H\times W}$, $F^c_{\text{max}} \in \mathbb{R}^{1\times H\times W}$, and $f^{7\times7}$ is a 7×7 size convolution kernel.

Similarly, as shown in bottom part of Fig. 2.4, through residual connections, the spatial attention features $M_s(F^{'})$ and the input feature $F^{'}$ are merged to obtain the final result. In image processing, channel attention focuses on determining meaningful features in the image, while spatial attention concentrates on identifying significant features within the image. Average pooling provides feedback on each pixel in input feature, whereas max pooling only provides feedback on the most responsive areas in input feature.

In terms of introducing multi-attention modules, Li et al. [58] proposed TA-Net, which designed channel with self-attention, and spatial attention for global information capturing. Qin et al. [59] proposed multi-attention dense network for bone marrow segmentation. Ruan et al. [56] proposed MALUNet for skin lesion segmentation, which

utilized four attention modules to extract global and local feature respectively. From the above, it can be seen that advancements have been achieved in various aspects of medical image segmentation tasks. However, the generality of most of the above methods in medical image segmentation tasks is challenging, and most of model parameters are relatively large. Besides, most of the above multi-attention methods do not explore the fusion between feature of different convolutional layers.

Overall, a well-designed attention mechanism can enhance the efficiency of neural network processing, enabling systems to achieve specific goals more quickly and effectively. The application of attention is not only significant in understanding human psychological activities but also plays a crucial role in the field of artificial intelligence.

2.3 Transformer for Medical Image Segmentation

Vision Transformers have recently gained attention in computer vision tasks. In medical image segmentation, Transformer-based models designed for end-to-end tasks perform prominently on multiple benchmarks. There have been many studies trying to combine visual Transformers with medical image segmentation and achieved performance close to or even better than CNN on different datasets.

The input image is $x \in \mathbb{R}^{H\times W\times C}$, which is first divided into fixed-size patches. The size of each patches is $P \times P$, and $N = (H \times W)/(P^2)$ is the number of patches. As shown in Fig. 2.5, the flattened patches are linearly mapped. In order to retain the position information of each patch, position encoding information is added to each slice before the slice is sent to the Transformer encoder. The Transformer encoder consists of L layers standard Transformer modules. Each module is composed of layer normalization (LN) [90], multi-head self-attention module (MSA), multi-layer perceptron (MLP) and residual connection [91]. MLP consists of two linear layers with GELU activation function, and the MSA sub-layer consists of n parallel self-attention (SA) heads. The calculation process is as follows:

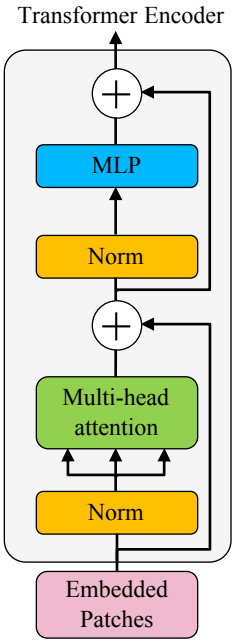


Fig. 2.5. Schematic of Transformer encoder.

$$z_0 = [x_{\text{class}}; x_p^1 E; x_p^2 E; \cdots; x_p^N E] + E_{\text{pos}}, \qquad E \in \mathbb{R}^{(P^2 \cdot C) \times D}, E_{\text{pos}} \in \mathbb{R}^{(N+1) \times D} \tag{2.9}$$

$$z_l^{'} = \text{MSA}(\text{LN}(z_{l-1})) + z_{l-1}, \qquad l = 1 \ldots L \tag{2.10}$$

$$z_l = \text{MLP}(\text{LN}(z_l^{'})) + z_l^{'}, \qquad l = 1 \ldots L \tag{2.11}$$

$$y = \text{LN}(z_L^0). \tag{2.12}$$

Where input image $x \in \mathbb{R}^{H \times W \times C}$, 2D patches $x_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$, C is the number of channel, P is the size of patch, and $N = (H \times W)/(P^2)$ is the number of patches. $z_l^{'}$ and z_l represent the output features of MHSA and MLP in the l^{th} module respectively. Finally, the last layer of sequence representation y is obtained from the Transformer encoder.

Inspired by the good performance of ViT in the field of natural images [60], there have been many studies trying to combine visual Transformers with medical image segmentation and achieved performance close to or even better than CNN on different datasets. Hatamizadeh [63] directly used Transformer as an encoder to extract abdominal organ features, but it lacks the feature extraction of CNN and the ability to capture long-distance

features. For multi-scale medical images, DS-TransUNet [64] uses Swin Transformer to extract features in the encoder and fuses features extracted by CNN in the decoder. However, it lacks feature fusion in the encoder.

Chen et al. [61] proposed TransUNet, which is the first research work to combine Transformer with medical image segmentation. It combines the respective advantages of Transformer and U-Net and embeds Transformer into the encoder. Swin UNETR [92] directly uses Swin Transformer as an encoder to extract medical image features. Xie et al. [93] proposed Cotr, which employs a CNN as the feature extraction backbone, utilizes a Transformer for encoding representation processing, and employs a CNN decoder to make predictions for the segmentation output. Yan et al. [94] proposed AFTer-UNet, which benefited from convolutional layers for detailed feature extraction and harnessed Transformers' strengths in modeling long sequences. Compared with these methods, DEU-Net fuses the features extracted from Transformer and CNN in the encoder by proposing a DFFM, making full use of the feature extraction capabilities of the two extractors for datasets of different sizes.

3 MDSU-Net for 2D Medical Image Segmentation

3.1 Introduction

Medical imaging techniques mainly include Magnetic Resonance Imaging (MRI), Computed Tomography (CT), and Ultrasound Imaging (UI) [20–22]. As shown in Fig. 3.1, there are three examples of different types of medical images from ATLAS [33], CHAOS [31], and NERVE [37], respectively. The top line are original images and the bottom line is Ground Truth (GT) from three datasets. In clinical diagnosis, medical images provide doctors with main patient condition information, and medical image segmentation facilitates clinical diagnosis and treatment [95, 96]. In clinical practice, segmentation is typically performed manually. However, when processing a vast number of medical images, the quality of segmentation can vary based on the expertise of the medical professional. This variability underscores the need for a more consistent and efficient method to enhance the performance of segmentation tasks. One significant challenge in this domain is the acquisition and creation of high-quality datasets. Amassing such datasets is a complex task and is often beyond the capacity of a single institution within a limited timeframe. As a result, medical datasets tend to be smaller in size and can exhibit inconsistencies. Another notable characteristic of medical images is the imbalance between the foreground and background areas. Unlike natural scene images, where the foreground and background might be more balanced, medical images often have a much larger background compared to the foreground.

Medical image segmentation has been widely discussed and concerned in the field of image processing, which has become a basic component and a crucial stage of image processing [2, 3]. Based on deep learning and powerful data processing capabilities of servers, pixel-level processing and segmentation of medical images are generally performed [41]. U-Net [9] is a variant of CNNs, which has achieved great performance in medical segmentation tasks [64]. U-Net adapts a skip connection operation to connect downsampling layers and upsampling layers, so that the segmentation results are more accurate [42]. However, the receptive field of U-Net decreases during training, which

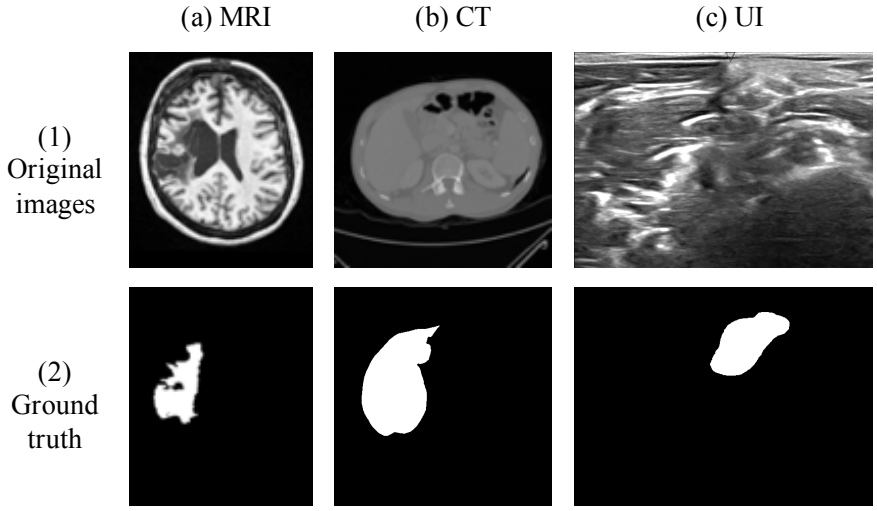


Fig. 3.1. Three examples of different types of medical images.

results in the inability to extract wider and richer contexts [43]. We introduce multi-attention with U-net to capture rich contextual information from medical images. Due to the small amount of data in medical images datasets, traditional U-Net is not suitable for medical image segmentation tasks. We adopt depthwise separable convolution (DSC) layer instead of traditional convolution to reduce model complexity without degrading model performance [97].

Several researchers have ventured into incorporating Transformers within the realm of medical image segmentation. Notable models such as TransUNet [61], Swin-Unet [62], and UNETR [63] have harnessed the power of self-attention to capture long-range dependencies inherent in medical images. However, a notable challenge with Transformer-based models is their inherent complexity. They often necessitate large datasets for training to achieve optimal performance. Consequently, Transformers may not be the ideal choice for every medical segmentation task.

In this work, we propose an Multi-Attention and Depthwise Separable Convolution U-Net (MDSU-Net), a variation of the U-Net, for medical image segmentation. MDSU-Net incorporates both multi-attention and DSC layer for improved performance. The multi-attention module within our framework utilizes dual attention and attention gates

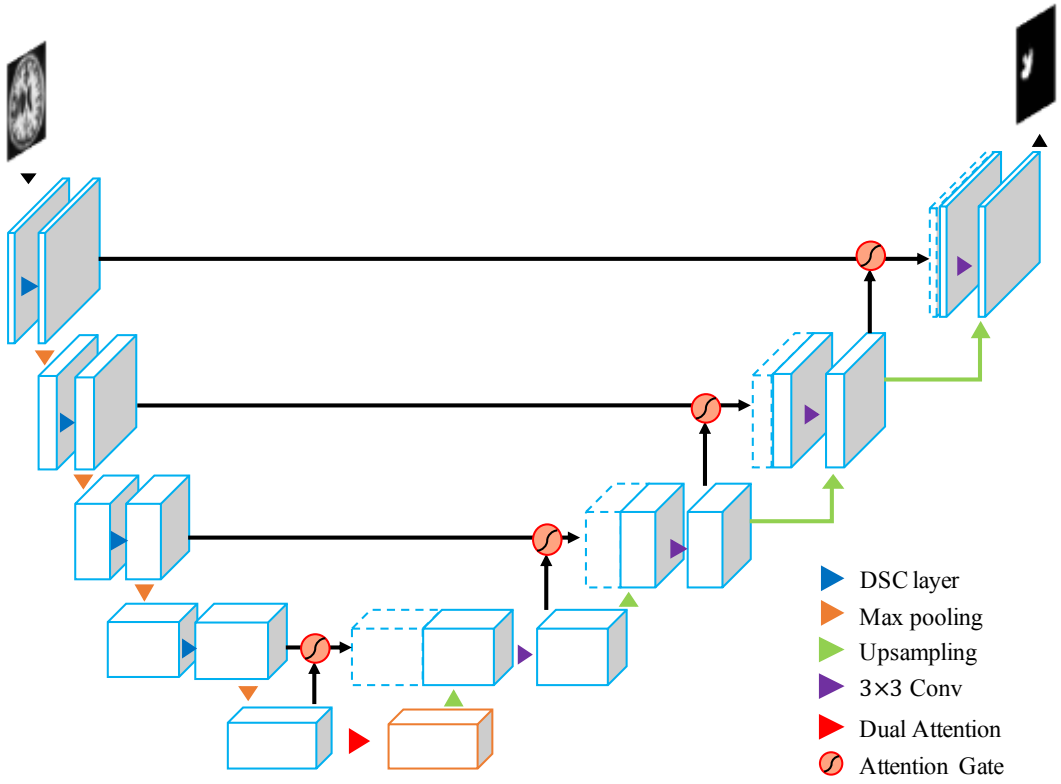


Fig. 3.2. Overview of MDSU-Net.

to capture rich contextual information and fuse features of different convolutional layers. MDSU-Net uses DSC layer to reduce model complexity without degrading model performance, which is suitable for different segmentation tasks.

3.2 Methodology

In this section, we first elaborate on the proposed model, MDSU-Net, as shown in Fig. 3.2. MDSU-Net combines multi-attention and DSC layer. Afterward, we introduce the DSC layer and multi-attention in detail.

We propose an MDSU-Net for medical image segmentation, which is a variation of U-Net architecture. Our MDSU-Net incorporates both multi-attention and DSC layer to improve performance. As shown in Fig. 3.3, we randomly select segmentation examples of small lesion area in stroke from ATLAS [33], where the yellow part is lesion area. The

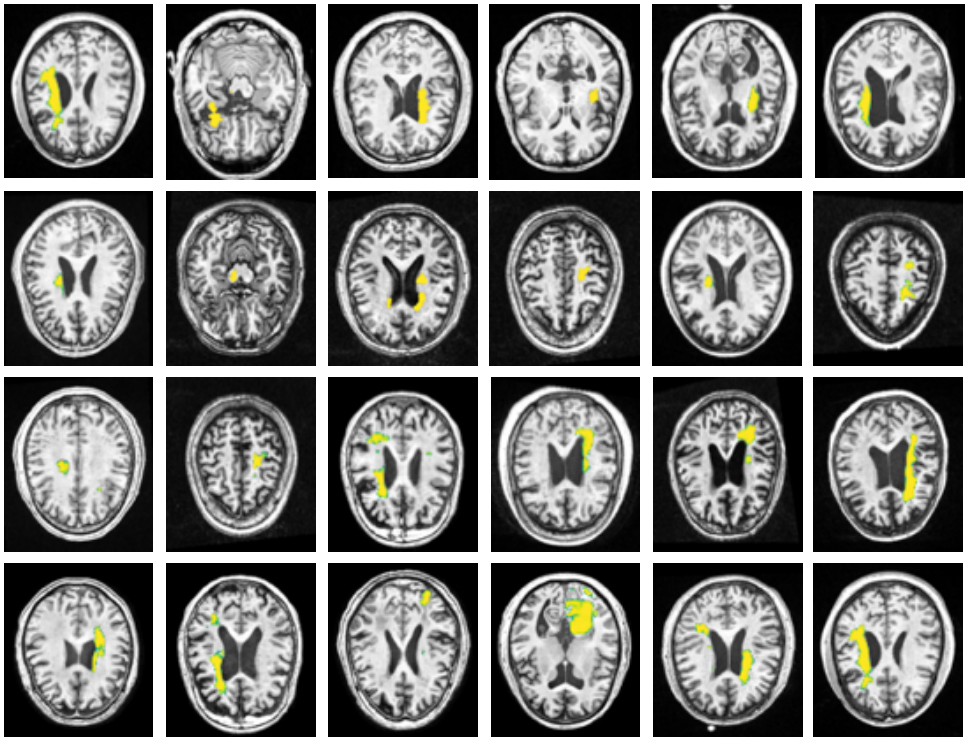


Fig. 3.3. Segmentation examples of small lesion area from ATLAS.

size of lesion is variable, the boundary of lesion area is not smooth and obvious, and the context features are relatively dense in medical images. In the encoder part, DSC layer is used to reduce model complexity. The multi-attention within our framework utilizes dual attention and attention gates to fuse feature of different convolutional layers, which enables better extraction of dense contextual features of medical images.

3.2.1 Depthwise Separable Convolution Module

In actual clinical application, the computing power of device is limited. However, U-Net [9] is heavy with traditional convolution and is prone to overfitting. Inspired by X-Net [54], we introduce DSC layer to reduce trainable parameters [97]. DSC decomposes a complete convolution operation into two steps, namely Depthwise Convolution (DC) and Pointwise Convolution (PC). Unlike convolution, a kernel of DC is responsible for one channel. The operation of PC is similar to the convolution operation, and its convolution

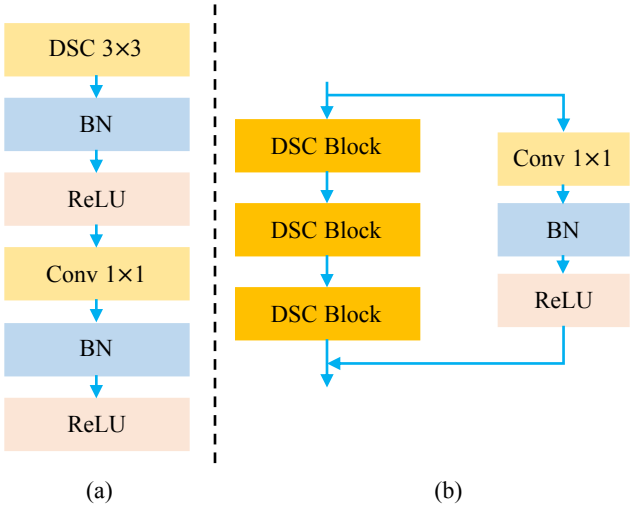


Fig. 3.4. Schematic of depthwise separable convolution module.

kernel has a size of 1×1. Therefore, model parameters are greatly reduced.

As shown in Fig. 3.4, (a) is DSC Block and (b) is DSC layer, where we use DSC to build a DSC block. Different from DSC of MobileNet series [98–100], we utilize three DSC blocks and one residual connection [91] to form a DSC layer. The design of combination of three DSC blocks can improve feature extraction from medical images in the encoder, and addition of residual connections can support the deep construction of network and prevent overfitting. Compared with the traditional convolution, the DSC layer reduces the complexity of the model without sacrificing the performance of model, which is suitable for various types of medical image segmentation.

3.2.2 Dual Attention Module

To avoid the insufficiency in extracting advanced features of original networks, we introduce a dual attention. The function of dual attention is to integrate different representations of features in both position and channel dimensions to better learn long-range dependencies, enhance ability to capture contextual information, and focus on extracting position-sensitive features in medical images. The information capture ability of the self-attention mechanism models are typically complex, and its performance in small datasets

is not as good as CNN. Different from the self-attention mechanism, dual attention can better capture rich contextual information.

Position attention module: Position attention module is used to aggregate position contextual information. There are four branches, with top two branches initially reducing channels to 1/8 of the original to diminish the impact of channel count on position representation of features. After fusion of two branches, the use of a softmax activation function allows the model to focus on extracting position-sensitive features in medical images. The outcome is then combined with the features from the third branch to reintroduce channel information, enhancing its feature extraction capabilities. Finally, it is directly fused with the input feature to enhance position feature extraction.

The input is $F \in \mathbb{R}^{C\times W\times H}$, where C, W, and H represent the channel, width and height, respectively. In the first path, F is reshaped to $I_0 \in \mathbb{R}^{(W\times H)\times C/8}$. In the second path, F is convoluted and transposed to $I_1 \in \mathbb{R}^{C/8\times(W\times H)}$ to diminish the impact of channel count on position representation of features. The multiplication operation is applied to I_0 and I_1, followed by a softmax to focus on extracting position-sensitive features, yielding $P \in \mathbb{R}^{(W\times H)\times(W\times H)}$ as follows:

$$P_{i,j} = \frac{\exp(I_{0,i} \cdot I_{1,j})}{\sum_{i=1}^{W\times H} \exp(I_{0,i} \cdot I_{1,j})}, \tag{3.1}$$

where $P_{i,j}$ measure j^{th} position's impact on i^{th} position. In the third path, F is convoluted and reshaped to $I_2 \in \mathbb{R}^{C\times(W\times H)}$ to enhance feature extraction capabilities. Finally, I_2 is multiplied by P, and the result is fused with F to get position attention map to aggregate position contextual information, which is $I_P \in \mathbb{R}^{C\times W\times H}$ as follows:

$$I_{P,j} = \varphi_1 \sum_{i=1}^{W\times H} P_{i,j} I_{2,j} + F_j, \tag{3.2}$$

where φ_1 is initialized to 0.

Channel attention module: Channel attention is used to capture interdependencies between channels. Similarly, there are four branches in total. The middle two branches initially conceal position information of medical images to highlight relationships be-

tween channels in features. Then, the results are fused with the feature from the fourth branch, reintroducing position information from medical images, enhancing its feature representation capability, and increasing the receptive field. Finally, it is also directly fused with the input feature to enhance channel feature extraction.

$F \in \mathbb{R}^{C\times W\times H}$ is reshaped and permuted to get $I_0 \in \mathbb{R}^{(W\times H)\times C}$, $I_1 \in \mathbb{R}^{C\times(W\times H)}$, and $I_2 \in \mathbb{R}^{C\times(W\times H)}$, respectively. I_0 is multiplied by I_1 to eliminate dimensional effects and highlight relationships between channels. The application of softmax enables us to obtain $C \in \mathbb{R}^{C\times C}$, which is described as follows:

$$C_{i,j} = \frac{\exp(I_{0,i} \cdot I_{1,j})}{\sum_{i=1}^{C} \exp(I_{0,i} \cdot I_{1,j})}, \tag{3.3}$$

where $C_{i,j}$ measure the j^{th} channel's impact on i^{th} channel and model interdependencies between channels. Then we multiply C and I_2 and fuse with F to get channel attention map to reintroduce position information from medical images, which is $I_C \in \mathbb{R}^{C\times W\times H}$ as follows:

$$I_{C,j} = \varphi_2 \sum_{i=1}^{C} C_{i,j} I_{2,j} + F_j, \tag{3.4}$$

where φ_2 is also initialized to 0. Finally, I_P and I_C are summed to get dual attention features.

Lastly, the position and channel features are merged to leverage the dual attention to enhance representation of higher-level features, effectively capturing contextual information in medical images.

3.2.3 Attention Gate Module

The AG further improves the ability to fuse features of different convolutional layers in the encoder and decoder. Among them, i is the output feature on the encoder, and g is the output feature on the decoder. Through fusion, the feature representation ability of the skip connection can be enriched. Subsequently, the skip connection and convolution

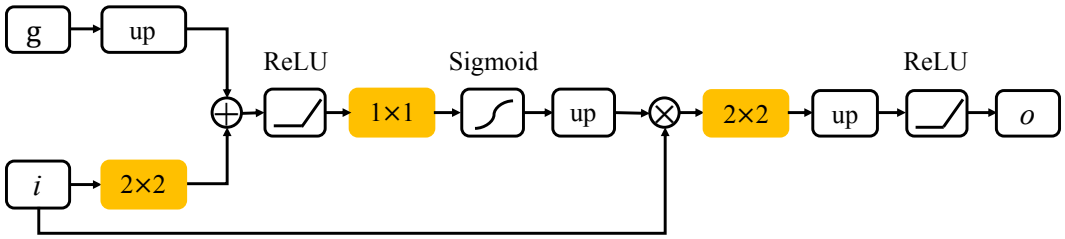


Fig. 3.5. Schematic of attention gate module in MDSU-Net.

calculation inside the AG are used to improve the decoder's ability to handle long-range dependencies.

The AG can extract rough image information through skip connection for image feature recovery [89]. As shown in Fig. 3.5, there are two inputs for AG. The first input is $i \in \mathbb{R}^{C\times W\times H}$, where C, W, and H represent channel, width and height, respectively. The second input is gating signal $g \in \mathbb{R}^{2C\times W\times H}$.

We feed i into a 2×2 convolution to get $i_1 \in \mathbb{R}^{C\times W\times H}$. At the same time, g is fed into an upsampling layer to get $g_1 \in \mathbb{R}^{C\times W\times H}$. In order to fuse features from encoder and decoder, we add i_1 with g_1 and apply a ReLU to get $r \in \mathbb{R}^{C\times W\times H}$ by the following formula:

$$r = \max(0, i_1 + g_1). \tag{3.5}$$

To enhance feature expression capabilities of MDSU-Net, we feed r into a 1×1 convolution to get $r_1 \in \mathbb{R}^{2C\times W\times H}$ and apply a Sigmoid to get $s \in \mathbb{R}^{2C\times W\times H}$ as follows:

$$s = \frac{1}{1+e^{-r_1}}. \tag{3.6}$$

In order to adjust the dimension of s to be unified with i, s is fed into an upsampling layer to get $s_1 \in \mathbb{R}^{C\times W\times H}$. We add s_1 and i to get $s_2 \in \mathbb{R}^{C\times W\times H}$. To enhance the ability to extract image feature, we feed s_2 into a 3×3 convolution and a batch normalization to get $s_3 \in \mathbb{R}^{C\times W\times H}$. Finally, a ReLU is applied to get $o \in \mathbb{R}^{C\times W\times H}$ as follows:

$$o = \max(0, s_3). \tag{3.7}$$

3.2.4 Loss Function

The cross-entropy (CE) loss function evaluates predictions for all pixels, but in cases of extremely imbalanced datasets, it can lead the model into local optima, causing predictions to strongly favor the background. On the other hand, the dice loss function [101] directly optimizes the dice coefficient. When calculating the intersection and ratio, dice loss function ignores a substantial number of background pixels, which effectively addresses the issue of foreground-background imbalance while also enhancing convergence speed. Therefore, we fuse the two loss functions to take advantage of both, which is defined as follows:

$$\mathscr{L} = \mathscr{L}_{\text{ce}} + \lambda_{\text{dice}} \mathscr{L}_{\text{dice}}, \tag{3.8}$$

where $\lambda_{\text{dice}} = 1.0$. The ablation experiments regarding λ_{dice} weight selection can be found in Section 3.4.3.

3.3 Experiment Setup

3.3.1 Datasets

We employ three medical segmentation datasets to validate our method, which are the ATLAS [33] dataset, the CHAOS [31] challenge, and the Ultrasound Nerve Segmentation challenge [37], as shown in Tab. 3.1.

In ATLAS [33], there are 229 MRI segmentation subjects. The size of each image is 233×197, and each 3D image is sliced into 189 slices, which corresponds 43,281 slices. We crop the image size to 224×192. We randomly select subjects for training, validation, and testing in a ratio of 6:2:3.

CHAOS [31] contains two datasets (abdominal CT and MRI), and we choose the dataset for the CT part, that is, the liver segmentation. The dataset contains CT images from 40 different patients, where each of size is 512×512. In total, 2,874 slices are used for training, 2,874 slices are used for validation, and 1,408 slices are used for testing.

Ultrasound Nerve Segmentation [37] is an dataset for the brachial plexus. There are 5,635 images for training and 5,508 images for testing with a size of 580×420. We randomly select 20% of training images for validation.

Tab. 3.1. Three different types of medical image segmentation datasets.

Tasks	Datasets	Modalities	Subjects	Size
Stroke	ATLAS	MRI	299	233×197×189
Liver	CHAOS	CT	40	512×512
Nerve	NERVE	UI	5,635	580×420

3.3.2 Evaluation Metrics

We choose a series of evaluation metrics to quantify the segmentation networks. The evaluation index formula is as follows:

$$\text{Dice} = \frac{2TP}{2TP+FP+FN}, \tag{3.9}$$

$$\text{IoU} = \frac{TP}{TP+FP+FN}, \tag{3.10}$$

$$\text{Precision} = \frac{TP}{TP+FP}, \tag{3.11}$$

$$\text{Recall} = \frac{TP}{TP+FN}. \tag{3.12}$$

where TP represent true positive, FP is false positive, and FN is false negative.

3.3.3 Implementation Details

All models are constructed using the PyTorch framework and trained on an NVIDIA TITAN X GPU. During model training, we utilize Adam optimizer [102] with an initial learning rate of 0.001 and weight decay of 0.0005. We choose resolutions of 224×192

for the ATLAS and training epochs are set to 100, 512×512 for the CHAOS and training epochs are set to 800, and 580×420 for the Ultrasound Nerve and training epochs are set to 400, respectively. The batch size is set to 8. Additionally, an early stopping strategy is employed.

Attention gate: We use 4 AGs to apply in the skip connections with resolution of 1, $1/2$, $1/4$, and $1/8$, respectively. We select features from decoder layers as input of gating signal.

Dual attention: We apply dual attention after bottom layer with resolution $1/16$.

3.4 Experimental Results

3.4.1 Comparison with State-of-the-arts

ATLAS: The training of network converges over 100 epochs on ATLAS. In Tab. 3.2, we get validation results of our network, U-Net [9], CLCI-Net [103], SAN-Net [104], and 3D-ResU-Net [105]. The results indicate that our network achieves a superior Dice of 0.0545 compared to 3D-ResU-Net. Our network sets a state-of-the-art (SoTA) benchmark with IoU of 0.5450, Precision of 0.7789, and Recall of 0.6447, respectively. Compared with other U shape networks, the parameters of our network are reduced over 10M, which improves the efficiency of network training.

CHAOS: The training of network converges over 800 epochs on CHAOS. In Tab. 3.3, we get the validation results of our network, U-Net [9], FC-Densenet [106], SGU-Net [107], and Nas-UNet [108]. The results indicate that our network achieves a superior Dice of 0.0020 compared to Nas-UNet. Our network sets a SoTA benchmark with IoU of 0.9932. Furthermore, the parameters of our network are 6.7M less than those of Nas-UNet.

ULTRASOUND NERVE: The training of network converges over 400 epochs on NERVE. In Tab. 3.4, we get the validation results of our network, U-Net [9], FC-Densenet [106], EHA-Net [53], and Nas-UNet [108]. Our network achieves a superior Dice of 0.0073 compared to Nas-UNet. In terms of IoU, our network achieves score of 0.9892.

Tab. 3.2. Comparison results on the ATLAS dataset.

Methods	Dice	IoU	Precision	Recall	Parameters
U-Net [9]	0.4606	0.2992	0.5993	0.4449	34.5M
CLCI-Net [103]	0.5810	0.4094	0.6490	0.5810	36.8M
SAN-Net [104]	0.5711	0.3997	-	0.5977	29.64M
3D-ResU-Net [105]	0.6400	0.4706	0.6200	-	-
MDSU-Net(ours)	**0.7055**	**0.5450**	**0.7789**	**0.6447**	**23.5M**

Tab. 3.3. Comparison results on the CHAOS dataset.

Methods	Dice	IoU	Parameters
U-Net [9]	0.9370	0.9820	34.5M
FC-Densenet [106]	0.9650	0.9830	9.4M
SGU-Net [107]	0.9574	0.9183	**5.0M**
Nas-UNet [108]	0.9740	0.9850	30.2M
MDSU-Net(ours)	**0.9760**	**0.9932**	23.5M

Tab. 3.4. Comparison results on the NERVE dataset.

Methods	Dice	IoU	Parameters
U-Net [9]	0.7400	0.9890	34.5M
FC-Densenet [106]	0.8440	0.9890	**9.4M**
EHA-Net [53]	0.8009	0.6679	-
Nas-UNet [108]	0.8810	**0.9920**	30.2M
MDSU-Net(ours)	**0.8883**	0.9892	23.5M

3.4.2 Statistical Evaluation

By introducing multi-attention mechanism, MDSU-Net exhibits significant performance improvements across three diverse datasets. DSC layers are used to reduce trainable parameters, which improves the efficiency of network training.

In order to validate the effectiveness and stability of MDSU-Net for medical image segmentation, we test Dice coefficient comparison between MDSU-Net and other models on three datasets respectively. As shown in Fig. 3.6, orange line within each box represents the median and green triangle represents the mean. MDSU-Net performs better than other models in quartiles, mean, and median. In both CHAOS and NERVE datasets, the discrete distribution of MDSU-Net outperforms other models. Our test results on the three datasets have the highest mean and median of Dice coefficient. At the same time, the median is biased towards the third quartile, indicating that result distribution is more concentrated and deviation is smaller, indicating that our method is more efficient and stable.

3.4.3 Ablation Study

To thoroughly evaluate the proposed MDSU-Net framework under different settings, we conduct various ablation study on the ATLAS, CHAOS, and NERVE datasets, respectively, including: 1) the number of AGs, 2) the number of dual-attention, 3) the number of DSC blocks in residual DSC layer, and 4) the effect of λ_{dice} in loss function.

The number of AGs: We add AGs in the skip connections from resolution of 1/8 to resolution of 1 in the encoder part. The 1 AG means that only one AG is added in the skip connections with resolution of 1/8, the number of AGs increases with resolution. We sequentially increase AGs on skip connections starting from the lowest level image feature to validate the effect from different number of AGs.

As shown in Tab. 3.5, in terms of Dice, IoU, and Recall, scores increase with number of AGs. When the number of AGs is 4, network performs best, which shows that setting AGs with every skip connection can fully extract the context information of image

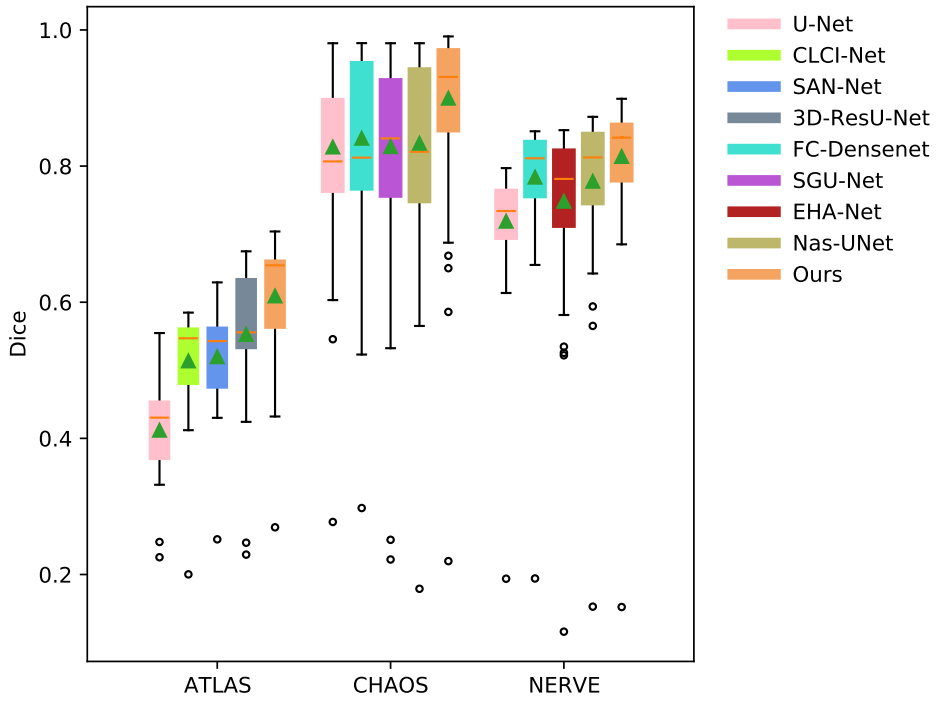


Fig. 3.6. Boxplot of Dice coefficient for all test samples from ATLAS, CHAOS, and NERVE.

Tab. 3.5. Results of different number of AGs on the ATLAS, CHAOS, and NERVE datasets.

Datasets	Methods	Dice	IoU	Precision	Recall	Parameters
ATLAS	1 AG	0.6494	0.4809	0.7960	0.5484	22.0M
	2 AGs	0.6364	0.4668	**0.8178**	0.5209	23.1M
	3 AGs	0.6674	0.5008	0.8085	0.5682	23.4M
	4 AGs	**0.7055**	**0.5450**	0.7789	**0.6447**	23.5M
CHAOS	1 AG	0.9707	0.9483	0.9904	0.9665	22.0M
	2 AGs	0.9710	0.9488	**0.9963**	0.9634	23.1M
	3 AGs	0.9738	0.9648	0.9850	**0.9717**	23.4M
	4 AGs	**0.9760**	**0.9688**	0.9884	0.9681	23.5M
NERVE	1 AG	0.8806	0.8329	0.9908	0.8942	22.0M
	2 AGs	0.8825	0.8345	0.9907	0.8857	23.1M
	3 AGs	0.8853	0.8443	0.9906	0.8973	23.4M
	4 AGs	**0.8883**	**0.8502**	**0.9933**	**0.8992**	23.5M

features.

Tab. 3.6. Results of different number of dual attention on the ATLAS, CHAOS, and NERVE datasets.

Datasets	Methods	Dice	IoU	Precision	Recall	Parameters
ATLAS	No dual attention	0.6424	0.4732	**0.8060**	0.5341	18.6M
	With dual attention	**0.7055**	**0.5450**	0.7789	**0.6447**	23.5M
CHAOS	No dual attention	0.9695	0.9445	**0.9977**	0.9618	18.6M
	With dual attention	**0.9760**	**0.9688**	0.9884	**0.9681**	23.5M
NERVE	No dual attention	0.8678	0.7863	0.9896	0.8960	18.6M
	With dual attention	**0.8883**	**0.8502**	**0.9933**	**0.8992**	23.5M

The number of dual attention: We start by pruning dual attention to validate the effect of dual attention on networks. In Tab. 3.6, the Dice of network with dual attention outperform network without dual attention by 0.0631, 0.0065, and 0.0205 in three datasets, respectively. As well as Dice, IoU and Recall of network with dual attention also outperform network without dual attention. In terms of Precision, network with dual attention is about the same as network without dual attention. When the highest-level image features of network are connected to dual attention, network performs best, which shows that dual attention can improve network's ability to focus on segmenting partial features in advanced image features, thereby improving the decoder's image segmentation performance.

The number of DSC blocks: We sequentially reduce the number of DSC blocks in the DSC layers to validate the effect from different number of DSC blocks on network performance. As shown in Tab. 3.7, in terms of Dice, IoU, and Recall, scores increase with number of DSC blocks. When the number of DSC blocks is 3, network performs best, which shows that residual network block composed of three DSC blocks can better extract the image features in the encoder and avoid overfitting phenomenon during network training.

The effect of λ_{dice}: In order to validate of weight of dice loss in the loss function

Tab. 3.7. Results of different number of DSC blocks on the ATLAS, CHAOS, and NERVE datasets.

Datasets	Methods	Dice	IoU	Precision	Recall	Parameters
ATLAS	DSC 1	0.5907	0.4191	**0.8195**	0.4617	19.9M
	DSC 2	0.7029	0.5419	0.7709	**0.6459**	21.6M
	DSC 3	**0.7055**	**0.5450**	0.7789	0.6447	23.5M
CHAOS	DSC 1	0.9705	0.9455	**0.9932**	0.9656	19.9M
	DSC 2	0.9694	0.9493	0.9931	0.9655	21.6M
	DSC 3	**0.9760**	**0.9688**	0.9884	**0.9681**	23.5M
NERVE	DSC 1	0.8792	0.8337	0.9896	0.8861	19.9M
	DSC 2	0.8810	0.8470	0.9906	0.8833	21.6M
	DSC 3	**0.8883**	**0.8502**	**0.9933**	**0.8992**	23.5M

of MDSU-Net, we selected λ_{dice} as 0, 0.2, 0.4, 0.6, 0.8, 1, 2, 3, 4, and 5, which are validated on CHAOS and NERVE datasets. As shown in Tab. 3.8, when the value of λ_{dice} is too small or large, it does not improve the performance of MDSU-Net. When λ_{dice} is 1, the model performs best, where MDSU-Net can effectively extract foreground features without ignoring background information. It achieves the balance of foreground and background feature extraction in the process of model learning medical images.

3.4.4 Efficiency Analysis

Tab. 3.9 shows the Dice, FLOPs, parameters, training time, and inference time of MDSU-Net and other SoTA methods on CHAOS dataset. FLOPs and inference time are tested on an input size of 256×256. According to our tests, MDSU-Net is a model with 23.5M parameters and 102.01G FLOPs. Compared with other models of the same scale, such as U-Net [9] and Nas-UNet [108], which have 34.5M and 30.2M parameters, and 113.44G and 111.80G FLOPs respectively. Compared with models of same scale, MDSU-Net has smaller parameters and FLOPs and achieves higher performance. Other lightweight models, such as SGU-Net [107] and FC-Densenet [106], have 5.0M and 9.4M parameters, and 5.0G and 74.09G FLOPs, respectively. Although the parame-

Tab. 3.8. Results of different λ_{dice} in loss function on CHAOS and NERVE datasets.

Datasets	λ_{dice}	Dice	IoU	Precision	Recall
CHAOS	0	0.9224	0.8560	0.9922	0.9384
	0.2	0.9605	0.9240	0.9921	0.9409
	0.4	0.9574	0.9183	0.9922	0.9482
	0.6	0.9586	0.9205	0.9897	0.9547
	0.8	0.9641	0.9307	0.9930	0.9665
	1	**0.9760**	**0.9688**	0.9884	**0.9681**
	2	0.9682	0.9384	**0.9932**	0.9617
	3	0.9657	0.9337	0.9823	0.9589
	4	0.9622	0.9272	0.9925	0.9576
	5	0.9645	0.9314	0.9929	0.9513
NERVE	0	0.7577	0.6099	0.9714	0.8743
	0.2	0.8788	0.7838	0.9890	0.8843
	0.4	0.8539	0.7450	0.9876	0.8848
	0.6	0.8707	0.7710	0.9886	0.8819
	0.8	0.8795	0.7849	0.9892	0.8955
	1	**0.8883**	**0.8502**	**0.9933**	**0.8992**
	2	0.8782	0.7828	0.9895	0.8858
	3	0.8332	0.7141	0.9857	0.8382
	4	0.8588	0.7525	0.9872	0.8871
	5	0.8721	0.7732	0.9878	0.8711

Tab. 3.9. Results of Dice, FLOPs, parameters, training time, and inference time in CHAOS.

Datasets	Methods	Dice	FLOPs	Parameters	Training Time	Inference Time
CHAOS	U-Net [9]	0.9370	113.44G	34.5M	7.30h	**27s**
	FC-Densenet [106]	0.9650	74.09G	9.4M	**6.28h**	129s
	SGU-Net [107]	0.9574	**5.0G**	**5.0M**	6.55h	46s
	Nas-UNet [108]	0.9740	111.80G	30.2M	7.45h	38s
	MDSU-Net(ours)	**0.9760**	102.01G	23.5M	7.13h	36s

ters and FLOPs of MDSU-Net are larger compared with light models, Dice obtained by MDSU-Net is better than these light models. Furthermore, MDSU-Net takes the least inference time compared to models of the same size and less than FC-Densenet. Overall, our method achieves the highest segmentation performance with reduced computational complexity.

3.4.5 Case Study

We conduct case studies of our methods with the SoTA through cases from five distinct patients in the ATLAS dataset, where the yellow part is the lesion area after segmentation. For the convenience of observation and comparison, we enlarge the lesion area in Fig. 3.7. We randomly select five patient cases where the size and location of lesions exhibit variations. The first vertical column is the original MRI, the second vertical column is GT from ATLAS, the third vertical column is the zoom of GT, and the last four vertical columns are the zoom of visualized results from our network, 3D-ResU-Net, CLCI-Net and U-Net, respectively.

In the cases of lesions with smaller sizes in the top three rows, the visualized results from our network are closest to GT. The errors of other three methods are relatively large, while lesion edges predicted by our network are smoother in small lesions. In the cases of larger size lesions in the bottom two rows, the visualized results from our network and 3D-ResU-Net outperform CLCI-Net and U-Net in segmentation accuracy with smoother predicted lesion edges. MDSU-Net performs well in different size lesions, and the margins of lesion are closest to the GT.

3.5 Discussion

In order to improve the ability of the model to capture the contextual information of medical images, we explore the importance of dual attention and AGs for extracting position-sensitive features of medical images, and by introducing a DSC layer in the encoder, we propose a multi-attention network called MDSU-Net. Among them, dual at-

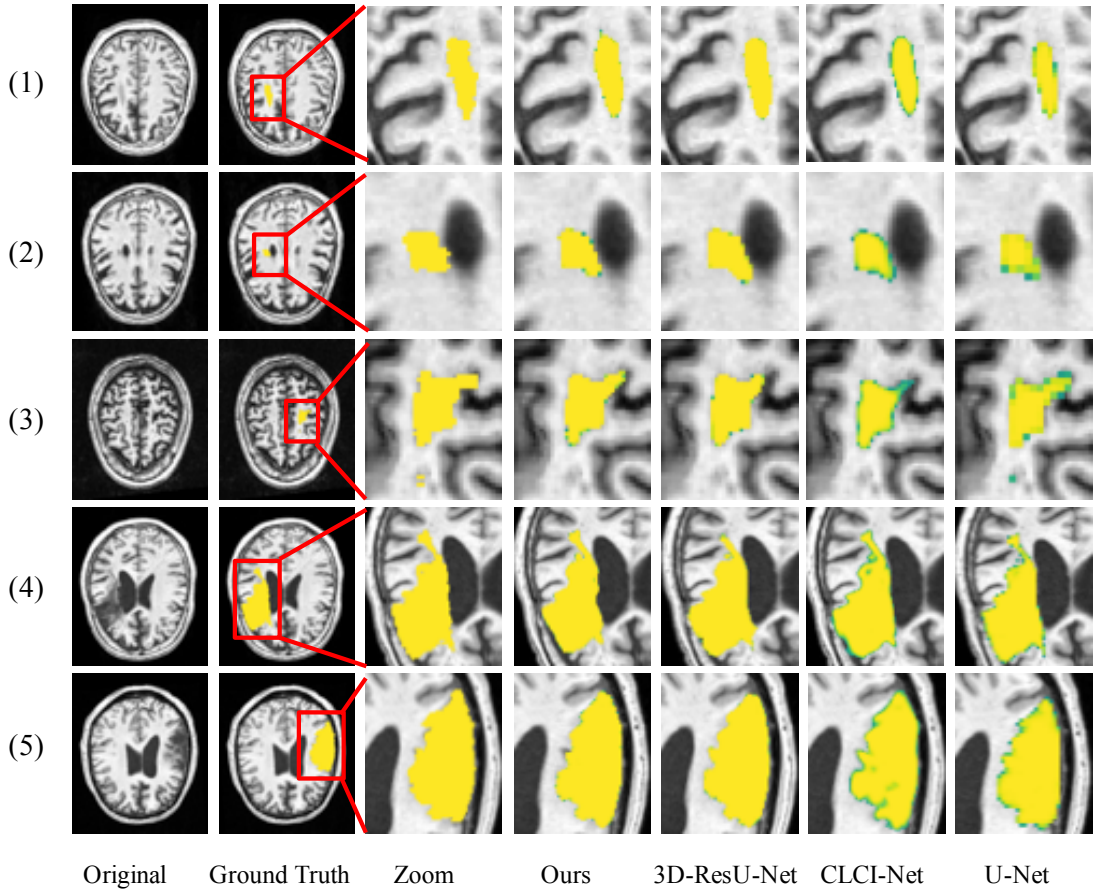


Fig. 3.7. Visualization of segmentation results from our method, 3D-ResU-Net, CLCI-Net and U-Net.

tention can fuse the channel features and position features in the encoder, AG can encode rough information and fuse features of different convolutional layers, and the DSC layer can reduce complexity of the encoder. Experiments are conducted in three different types of medical image datasets, namely ATLAS, CHAOS, and NERVE segmentation datasets. Experimental results show that MDSU-Net outperforms SoTA, especially in terms of Dice from Tab. 3.2, Tab. 3.3, and Tab. 3.4. As shown in Fig. 3.6, test results of MDSU-Net in the three datasets have the highest average and median Dice coefficients. At the same time, the median is biased towards the third quartile, and results are more concentrated and deviation is smaller, which means that our method has better effectiveness and stability.

We conduct ablation experiments in three datasets to evaluate the impact of these three modules on MDSU-Net. The results are shown in Tab. 3.5, Tab. 3.6, and Tab. 3.7, from which we can see that the introduction of dual attention and AGs have improved the performance of MDSU-Net, all of which have improved in Dice, IoU, and Recall. By introduction of DSC layer modules, number of parameters of MDSU-Net is reduced by 11.0M compared with U-Net. These ablation experiments show that the multi-attention can well fuse channel features and position features, and can well fuse features from encoder and decoder. In addition, the introduction of DSC layer modules can achieve better performance on three different datasets with fewer parameters than U-Net, which means that MDSU-Net has better accuracy and lower redundancy.

In actual clinical application, the computing power of device is limited. Our proposed MDSU-Net can reduce the operating pressure of medical equipment by reducing network complexity and improving computing efficiency. From Tab. 3.9, compared with models of same scale, MDSU-Net has smaller parameters and FLOPs. However, compared with lightweight models, MDSU-Net has larger parameters and FLOPs, which is not light enough. As shown in Tab. 3.7, in ATLAS, whether segmenting small or large areas, MDSU-Net has achieved better segmentation results, and performance has been significantly improved. However, compared with GT, segmentation edges of MDSU-Net are too smooth and details of segmentation edges are not depicted. In all, MDSU-Net is

adaptable to the segmentation of medical images of different types and sizes, increasing the practical applicability of our model.

Although MDSU-Net performs well on three different types of datasets, there are some shortcomings in this work. (1) The segmentation results of MDSU-Net in the 3D MRI dataset are far inferior to those in the CT and UI datasets. MDSU-Net is based on the characteristics of two-dimensional U-Net, which is mainly used for 2D medical image segmentation tasks. In 3D medical image segmentation tasks, MDSU-Net lacks the ability to correlate information between slices. (2) While MDSU-Net employs the DSC layer in lieu of the traditional convolutional layer to mitigate model complexity and alleviate strain on medical equipment, challenges persist. The extensive convolutions within the DSC layer, combined with the introduction of multi-attention, result in a model that still has a relatively heavy parameter load. (3) MDSU-Net performs well on small and medium-sized datasets, but its performance on large datasets requires further experimental validation.

3.6 Summary

In order to enhance reliability of medical image segmentation, we propose a novel MDSU-Net, which introduces multi-attention and DSC layers. MDSU-Net registers a Dice score of 0.7055 on ATLAS, 0.9760 on CHAOS, and 0.8883 on NERVE. Notably, these scores surpass the current state-of-the-art (SoTA) benchmarks. The coarse information extracted is effectively applied to the decoder through AGs, which can fuse features of different convolutional layers. By incorporating dual attention, the network's capability to extract abundant contextual information is significantly improved, thereby enhancing its feature extraction capacity. DSC layer reduces parameters of MDSU-Net and improves network training efficiency. By introducing multi-attention and DSC layer, MDSU-Net can perform medical image segmentation tasks efficiently and reliably. However, in 3D medical images, the segmentation result is easily affected by the size of lesion in the slice, so in future work we need to propose a 3D network to correlate information between slices

and extract whole 3D lesion features. In addition, we need to reduce model complexity further to reduce computational pressure of device.

4 DEU-Net for 3D Medical Image Segmentation

4.1 Introduction

Medical image segmentation is a crucial task in medical image processing [2, 3], especially 3D medical image segmentation. Its significance lies not only in the precise segmentation of lesions to better guide medical image classification but also in the importance of extracting features such as the size and morphology of segmented lesions for determining the malignancy of tumors and preoperative analysis [109]. In clinical practice, 3D medical image segmentation is often performed manually [96]. However, errors can easily occur when dealing with a large number of medical images, and the accuracy of segmentation relies on the experience of medical professionals [110]. With the advancement of medical imaging technology, Magnetic Resonance Imaging (MRI) has taken a leading role in the field of radiological imaging [62]. MRI typically consists of multiple image sequences, forming its 3D structure [58]. Manual segmentation of 3D images requires extensive segmentation expertise and skillful operation [111]. Therefore, an effective automated method for 3D image segmentation is needed to achieve efficient segmentation tasks. Collecting MRI data is particularly challenging compared to other natural scene segmentation data, resulting in a scarcity of large MRI datasets.

Recently, CNNs have demonstrated SoTA performance in a wide range of visual recognition tasks [56, 59, 64]. With the introduction of deep learning networks, medical image segmentation has made significant progress [41, 96]. Currently, the mainstream framework for medical image segmentation is the encoder-decoder structure known as U-Net [9], which has shown excellent performance in various segmentation tasks [42, 112]. However, with the emergence of large medical image datasets, pixel-wise image segmentation methods alone are insufficient to support the network's ability to process a massive amount of data [68].

Inspired by the good performance of ViT in the field of natural images [60], there have been many studies trying to combine visual Transformers with medical image segmentation and achieved performance close to or even better than CNN on different datasets.

Hatamizadeh [63] directly used Transformer as an encoder to extract abdominal organ features, but it lacks the feature extraction of CNN and the ability to capture long-distance features. For multi-scale medical images, DS-TransUNet [64] uses Swin Transformer to extract features in the encoder and fuses features extracted by CNN in the decoder. However, it lacks feature fusion in the encoder.

We propose Dual Encoder U-Net (DEU-Net), which uses Transformer and CNN respectively to extract medical image features in the encoder. Transformer is a pre-trained model in Beyond the Cranial Vault (BTCV) [113], which improves its ability to capture contextual features of medical images and increases the learning speed. Besides, we introduce CBAM with each convolutional layer in the encoder part to enhance CNN's feature extraction capabilities for 3D medical images. To fuse the two kinds of features, we proposed a Dual Feature Fusion Module (DFFM) to fuse the features extracted from the Transformer and CNN in the encoder, making full use of the feature extraction capabilities of the two extractors for 3D medical image.

4.2 Methodology

In this work, we propose DEU-Net shown in Fig. 4.1, which uses Transformer and CNN respectively to extract medical image features in the encoder. Among them, CBAM can extract 3D medical image feature in the encoder, pre-trained Transformer can improves its ability to capture contextual features of medical images and increases the learning speed in the encoder, and the DFFM can fuse the features extracted from the Transformer and CNN in the encoder, making full use of the feature extraction capabilities of the two extractors for 3D medical image.

4.2.1 Transformer Encoder

As shown in Fig. 4.2, the input image is a 3D volume $x \in \mathbb{R}^{H\times W\times D\times C}$, which is first divided into fixed-size slices in Linear Projection module. The size of each slice is $P\times P\times P$, and $N=(H\times W\times D)/(P^3)$ is the number of slices. The flattened slices are

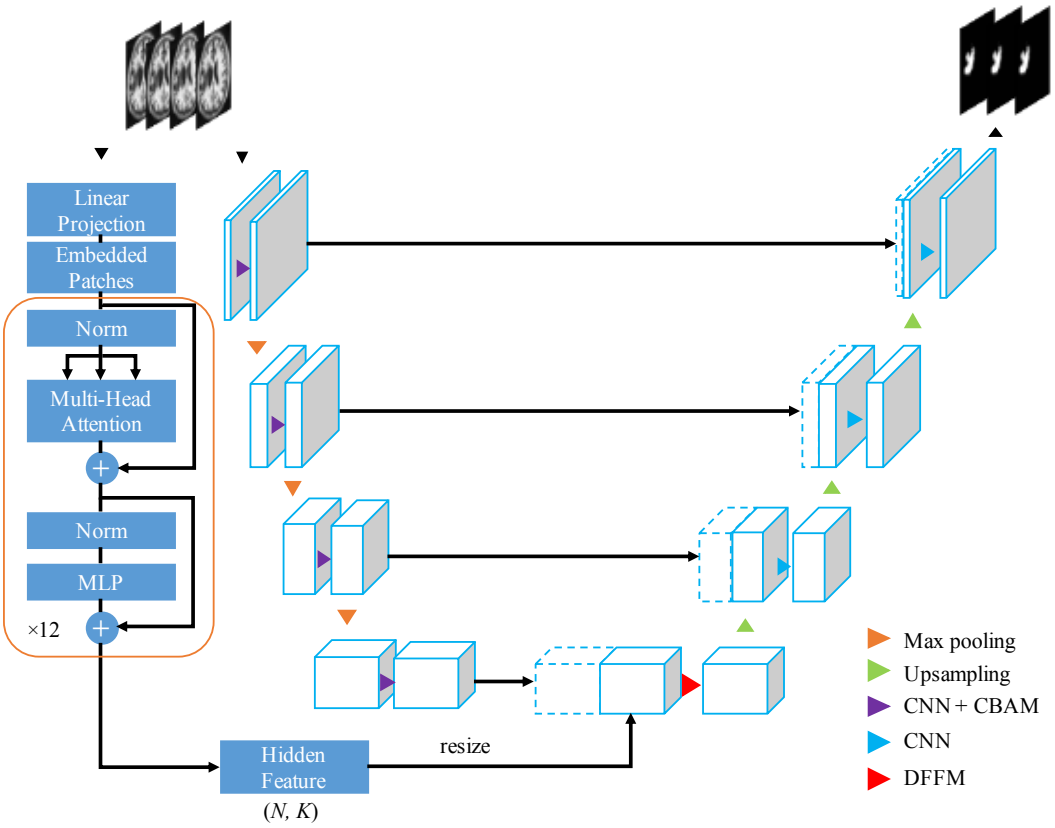


Fig. 4.1. Overview of DEU-Net architecture.

then linearly mapped. In order to retain the position information of each slice, position encoding information is added to each slice before the slice is sent to the Transformer encoder. Then, we project the patches into a K dimensional embedding space using linear layers, which remains constant throughout the entire transformer layer.

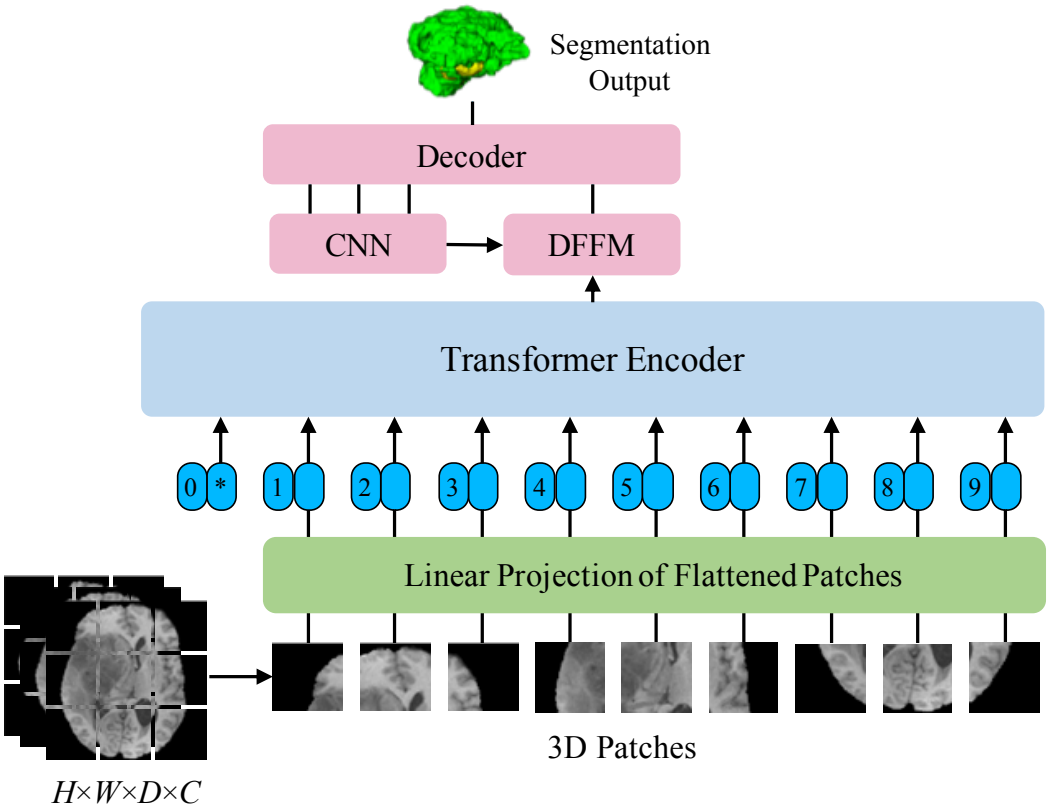


Fig. 4.2. Schematic of Transformer encoder in DEU-Net.

We extract the hidden feature from the last layer of sequence representation in Transformer, with dimension $N \times K$. In order to obtain more accurate image features and denser contextual information, we fuse the hidden features with the features extracted from the last layer of the CNN Encoder. To align their dimensions, we reshape the hidden feature into an $H/P \times W/P \times D/P \times K$ tensor. Additionally, we employ a $3 \times 3 \times 3$ convolutional layer and normalization layer to map the reshaped features from the embedding space back to the input space. At the bottleneck layer of the U-shaped network, we concatenate

the features extracted by the Transformer after reshape and convolution with the features obtained by the CNN. Finally, the fused feature pass through the DFFM and undergo upsampling before entering the Decoder.

4.2.2 CNN Encoder with CBAM

The encoder takes feature representation of the image as input and encodes the input features layer by layer through multiple encoder layers. This process is designed to capture contextual information and feature representations within the input sequence. Generally, the repeated stacking of encoder layers contributes to the deep feature representation of the input sequence. There are four convolutional layers in the CNN Encoder. As shown in Fig. 4.3, each of which contains two 3D convolutional blocks and one CBAM. Each 3D convolutional block contains a 3D convolution, followed by BN and ReLU activation functions. Finally, the output features and input features are fused through a residual connection [91] to enhance the feature extraction capability of the encoder.

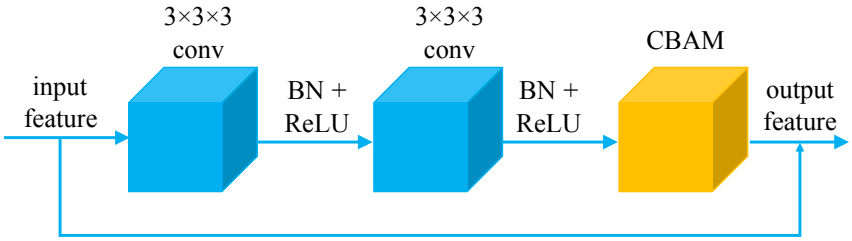


Fig. 4.3. Schematic of CNN encoder with CBAM.

4.2.3 Dual Feature Fusion Module

While employing Transformers alone to extract medical image features has shown satisfactory performance, this approach may not fully leverage the potential of Transformers in the context of segmentation. After patching, the resolution of each slice is much lower than that of the original image, which can lead to the loss of fine segmentation details during the encoding process. To address segmentation information loss, DEU-Net

adopts a hybrid CNN-Transformer architecture as a parallel encoder to achieve long-range contextual feature capture.

After implementing parallel encoders to extract medical image features, it is important to fuse two features from different encoders. As shown in Fig. 4.4, we propose a DFFM to fuse features extracted from Transformer and CNN respectively. At the output of the last layer of Transformer, we reconstruct the output feature $I_t \in \mathbb{R}^{(H/16\times W/16\times D/16)\times 768}$ into a 3D feature map $I_t^1 \in \mathbb{R}^{H/16\times W/16\times D/16\times 768}$, and then obtain the feature map $I_t^2 \in \mathbb{R}^{H/8\times W/8\times D/8\times 512}$ through a deconvolution to increase its resolution by 2 times. Then, we fuse the adjusted feature with output feature by the CNN encoder and feed them to a deconvolution layer to upsample the output to get the feature map $I_h \in \mathbb{R}^{H/8\times W/8\times D/8\times 512}$. I_h passes through a 3D convolution layer to obtain $I_h^1 \in \mathbb{R}^{H/8\times W/8\times D/8\times 512}$, and through a residual connection [91] to get $I_h^2 \in \mathbb{R}^{H/8\times W/8\times D/8\times 512}$ to prevent the network from overfitting and improve the feature expression ability. The module finally obtains $I_h^3 \in \mathbb{R}^{H/8\times W/8\times D/8\times 216}$ through upsampling. The final output is fed to the CNN Decoder.

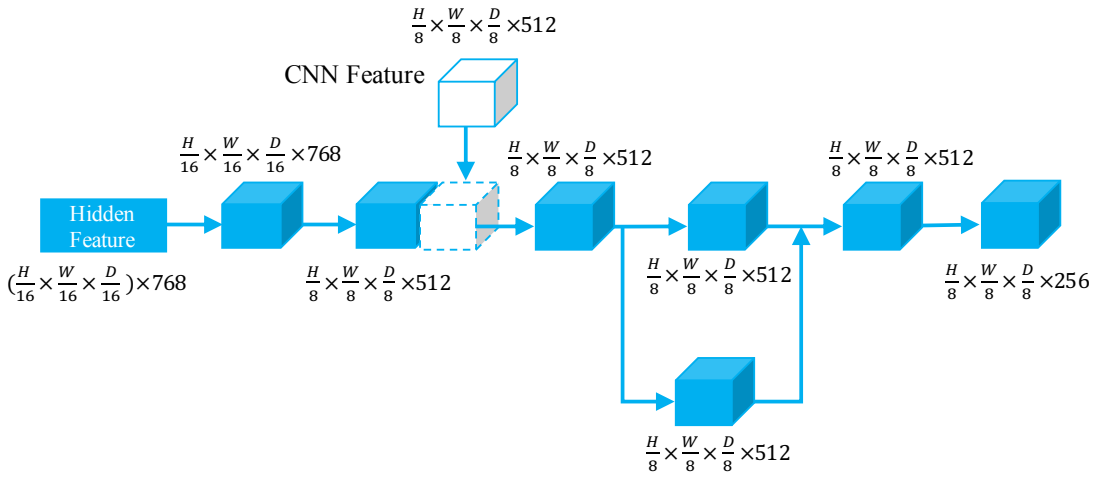


Fig. 4.4. Schematic of dual feature fusion module.

4.2.4 CNN Decoder

The decoder receives the encoded input sequence (output from the encoder) and progressively generates the input sequence through multiple decoder layers. There are three

convolutional layers in the CNN Decoder, each of which contains two 3D convolutional blocks. As shown in Fig. 4.3, each 3D convolutional block contains a 3D convolution, followed by BN, ReLU and upsampling layer.

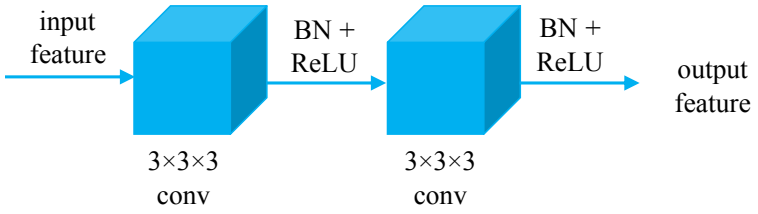


Fig. 4.5. Schematic of CNN decoder.

4.2.5 Loss Function

We use Dice Loss to train the DEU-Net. The loss $\mathscr{L}$ is calculated per batch and channel:

$$\mathscr{L} = \frac{1}{n}\sum_{n}\frac{S \cdot T + \varphi}{S^2 + T^2 + \varphi}, \tag{4.1}$$

where n is the number of channels, S is the output of the neural network after Sigmoid activation, T is the true label, and φ is a constant to prevent division by zero.

4.3 Experiment Setup

4.3.1 Datasets

BraTS 2020 [23–25] and BraTS 2021 [32] is a series of datasets for the MICCAI brain tumor segmentation competition. As shown in Tab. 4.1, on BraTS 2020, the training set has 369 cases, the validation set has 125 cases, and the testing set has 125 cases. On BraTS 2021, the training set has 1251 cases, the validation set has 219 cases, and the testing set has 530 cases. Each case contains 4 modalities and 3 segmentation tasks. The four modalities are T1, T2, FLAIR, and T1ce. Multi-modality is a commonly used way

of reflecting images in medical images. The three segmentation tasks are whole tumor (WT), enhance tumor (ET), and tumor core (TC).

Tab. 4.1. BraTS datasets for 3D medical image segmentation.

Tasks	Datasets	Modalities	Training	Validation	Testing	Size
Brain tumor	BraTS 2020	MRI	369	125	125	240×240×155
	BraTS 2021	MRI	1251	219	530	240×240×155

4.3.2 Evaluation Metrics

The segmentation challenge lies in accurately delineating the ET, TC, and WT portions of the tumor. The main evaluation metric is the Dice Coefficient, which can be defined as follows:

$$\text{Dice} = \frac{2TP}{2TP+FP+FN}, \tag{4.2}$$

where TP represent true positive, FP is false positive, and FN is false negative.

4.3.3 Implementation Details

We implement DEU-Net in PyTorch. The model is trained using an NVIDIA A100 server. Because the 3D image size in the BraTS datasets is large, we resize the 3D image to $128 \times 128 \times 128$ and set the batchsize to 2 to relieve GPU computing pressure. The initial learning rate is set to 10^{-4}, the Adam optimizer [102] is used, and the training epoch of the network is set to 200. Our Transformer-based encoder follows the ViT architecture with $L = 12$ layers and embedding size $K = 768$. We used a patch resolution of $16 \times 16 \times 16$. φ is set to 1 in loss function. For all experiments, we used five-fold cross-validation.

Tab. 4.2. Comparison results with SoTA methods on the BraTS 2020 dataset.

Methods	**Dice**			
	ET	***TC***	***WT***	***Average***
3D U-Net [9]	0.7345	0.8164	0.9029	0.8179
UNETR [63]	0.7473	0.8140	0.9012	0.8208
3D attention U-Net [114]	0.7782	0.8364	0.8893	0.8346
DEU-Net (Ours)	**0.7836**	**0.8384**	**0.9047**	**0.8422**

Tab. 4.3. Comparison results with SoTA methods on the BraTS 2021 dataset.

Methods	**Dice**			
	ET	***TC***	***WT***	***Average***
3D U-Net [9]	0.8747	0.9257	0.9352	0.9119
UNETR [63]	0.8676	0.9097	0.9250	0.9008
3D attention U-Net [114]	0.8763	0.9281	0.9387	0.9144
DEU-Net (Ours)	**0.8831**	**0.9317**	**0.9408**	**0.9185**

4.4 Experiment Results

4.4.1 Comparison with State-of-the-arts

BraTS 2020 dataset: We compare performance of DEU-Net with other SoTA methods on BraTS 2020, as shown in Tab. 4.2. The training and validation are all performed on the Brats 2020 dataset. It can be observed that in the ET segmentation task, the Dice coefficient of our proposed network DEU-Net is 0.0054 higher than that of 3D attention U-Net. In the TC segmentation task, DEU-Net improves 0.0020 compared to 3D attention U-Net. In the WT segmentation task, DEU-Net improves 0.0018 compared to 3D U-Net. Taken together, in terms of the average Dice coefficient of the three segmentation tasks, DEU-Net is 0.0076 higher than 3D attention U-Net. It can be concluded that by fusing Transformer and CNN features, we have improved the network's ability to capture contextual features.

BraTS 2021 dataset: As shown in Tab. 4.3, we compare performance of DEU-Net with other SoTA methods on BraTS 2021. The training and validation are all performed on the Brats 2021 dataset. It can be observed that in the ET segmentation task, the Dice coefficient of our proposed network DEU-Net is 0.0068 higher than that of 3D attention U-Net. In the TC segmentation task, DEU-Net improves 0.0036 compared to 3D attention U-Net. In the WT segmentation task, DEU-Net improves 0.0021 compared to 3D attention U-Net. Taken together, in terms of the average Dice coefficient of the three segmentation tasks, DEU-Net is 0.0041 higher than 3D attention U-Net. It can be concluded that by fusing Transformer and CNN features, we have improved the network's ability to capture contextual features.

4.4.2 Ablation Study

To thoroughly evaluate the proposed DEU-Net framework under different settings, we conduct various ablation studies on the BraTS datasets, including: 1) effect of CBAM, 2) effect of pre-trained Transformer, 3) effect of DFFM, and 4) effect of feature size in CNN.

Tab. 4.4. Results of CBAM on the BraTS 2020 and BraTS 2021 datasets.

Datasets	**Methods**	**Dice**			
		ET	***TC***	***WT***	***Average***
BraTS 2020	- CBAM	0.7795	0.8257	0.9034	0.8362
	+ CBAM	**0.7836**	**0.8384**	**0.9047**	**0.8422**
BraTS 2021	- CBAM	0.8771	0.9234	0.9283	0.9096
	+ CBAM	**0.8831**	**0.9317**	**0.9408**	**0.9185**

Tab. 4.5. Results of Transformer with and without pre-trained on the BraTS 2020 and BraTS 2021 datasets.

Datasets	**Methods**	**Dice**			
		ET	***TC***	***WT***	***Average***
BraTS 2020	Backbone	0.7345	0.8164	0.9029	0.8179
	Transformer	0.7758	0.8336	0.9037	0.8377
	Pre-trained	**0.7836**	**0.8384**	**0.9047**	**0.8422**
BraTS 2021	Backbone	0.8747	0.9257	0.9352	0.9119
	Transformer	0.8809	0.9299	0.9380	0.9163
	Pre-trained	**0.8831**	**0.9317**	**0.9408**	**0.9185**

Effect of CBAM: We add CBAM after each layer of convolutional layer from resolution of 1/8 to resolution of 1 in the encoder part. As shown in Tab. 4.4, on the BraTS 2020 dataset, the performance of the pre-trained network is improved by 0.0041, 0.0127 and 0.0013 on the three tasks compared with the network without pre-trained, with an average improvement of 0.0080. We can know that on the BraTS 2021 dataset, the performance of the pre-trained network is improved by 0.0060, 0.0083 and 0.0025 on the three tasks compared with the network without pre-trained, with an average improvement of 0.089. When we add the CBAM with all convolutional layers, network performs best, which shows that CBAM can fully extract the context information of 3D medical image features.

Effect of Pre-trained Transformer: We train and validate the Transformer without pre-trained and the Transformer with pre-trained on the BraTS 2020 and BraTS 2021 datasets, and the results are shown in Tab. 4.5. We can know that on the BraTS 2020 dataset, the performance of the pre-trained network is improved by 0.0078, 0.0048 and 0.0010 on the three tasks compared with the network without pre-trained, with an average improvement of 0.0045. We can know that on the BraTS 2021 dataset, the performance of the pre-trained network is improved by 0.0022, 0.0018 and 0.0028 on the three tasks compared with the network without pre-trained, with an average improvement of 0.0022. That means pre-trained model can capture contextual information very well. As shown in Fig. 4.6 and Fig. 4.7, the performance of the pre-trained model is relatively stable during the training and validation processes, which shows that pre-training can significantly improve the stability of network segmentation.

Effect of DFFM: DFFM is a module we proposed that integrates Transformer and CNN features to achieve efficient and effective interaction between multi-scale features. In order to validate the contribution of DFFM to network, we do training and validation on the BraTS 2020 and BraTS 2021 datasets. As shown in Tab. 4.6, we can know that the performance of the network with DFFM is improved by 0.0350, 0.0305 and 0.0120 on the three tasks compared to the network without DFFM, with an average improvement of 0.0258 on BraTS 2020. The performance of the network with DFFM is improved by

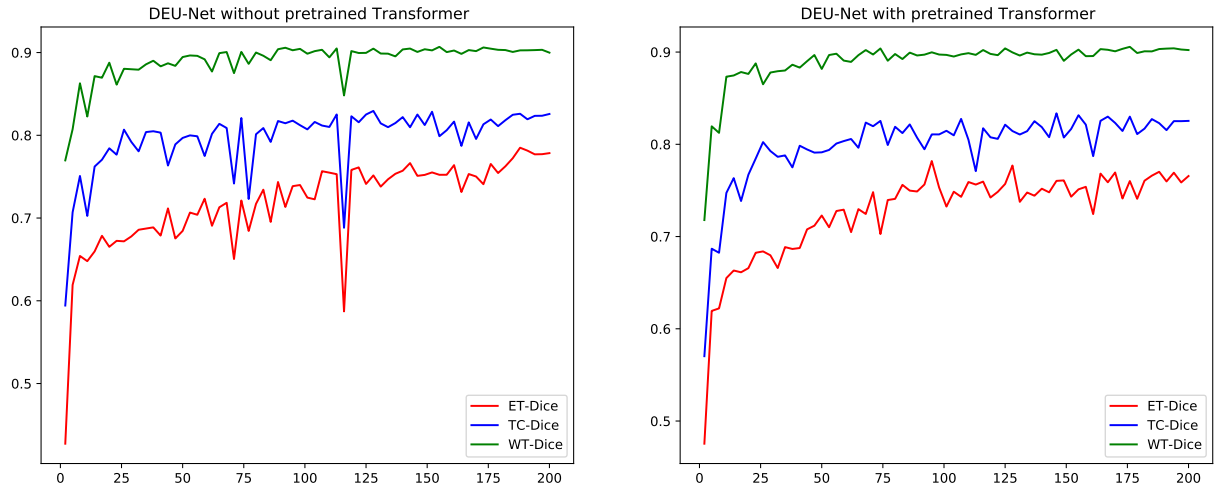


Fig. 4.6. Validation curve of Transformer with and without pre-trained on the BraTS 2020 dataset.

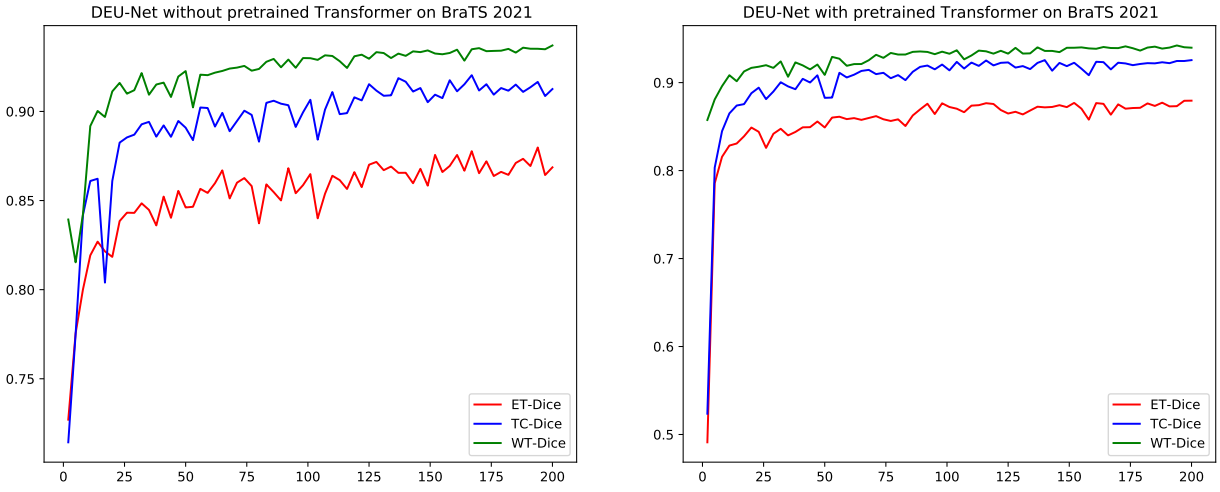


Fig. 4.7. Validation curve of Transformer with and without pre-trained on the BraTS 2021 dataset.

Tab. 4.6. Results of DFFM on the BraTS 2020 and BraTS 2021 datasets.

Datasets	Methods	Dice			
		ET	***TC***	***WT***	***Average***
BraTS 2020	- DFFM	0.7486	0.8079	0.8927	0.8164
	+ DFFM	**0.7836**	**0.8384**	**0.9047**	**0.8422**
BraTS 2021	- DFFM	0.8769	0.9188	0.9386	0.9114
	+ DFFM	**0.8831**	**0.9317**	**0.9408**	**0.9185**

0.0062, 0.0129 and 0.0022 on the three tasks compared to the network without DFFM, with an average improvement of 0.0071 on BraTS 2021. This shows that DFFM can well integrate Transformer and CNN features and improve the network's ability to extract medical image features.

Effect of feature size in CNN: In order to validate effect of input feature size of CNN in the encoder of DEU-Net, we select the input feature size as 8, 16, 32, 48, and 64. We train and validate the networks on the BraTS 2020 and BraTS 2021 datasets, and the results are shown in Tab. 4.7. We can know that both on the BraTS 2020 and BraTS 2021 datasets, as the feature size becomes larger, the performance of the network improves, but the improvement trend of network segmentation accuracy is not very obvious. When the input feature size of CNN is 8, the performance of the network has reached an advanced level. At the same time, as the feature size increases, the performance of the network is slowly increasing. However, the number of parameters of the network also increases accordingly. Compared with the input feature size of 8, when it is 64, the number of parameters of the network increases by 47.72%. In order to improve segmentation accuracy, we can increase the input feature size of CNN, but this will increase the number of parameters of the network, thereby increasing the amount of calculation and increasing the computational pressure.

Tab. 4.7. Results of feature size in CNN on the BraTS 2020 and BraTS 2021 datasets.

Datasets	**Feature Size**	**Dice**				**Parameters**
		ET	***TC***	***WT***	***Average***	
BraTS 2020	8	0.7756	0.8212	0.8842	0.8270	99.1M
	16	0.7766	0.8341	0.8939	0.8349	101.6M
	32	0.7770	0.8367	0.8961	0.8366	110.9M
	48	0.7818	0.8317	0.8953	0.8363	125.8M
	64	**0.7836**	**0.8384**	**0.9047**	**0.8422**	146.4M
BraTS 2021	8	0.8642	0.9169	0.9343	0.9051	99.1M
	16	0.8729	0.9240	0.9362	0.9110	101.6M
	32	0.8775	0.9202	0.9397	0.9125	110.9M
	48	0.8825	0.9307	0.9405	0.9179	125.8M
	64	**0.8831**	**0.9317**	**0.9408**	**0.9185**	146.4M

4.4.3 Case Study

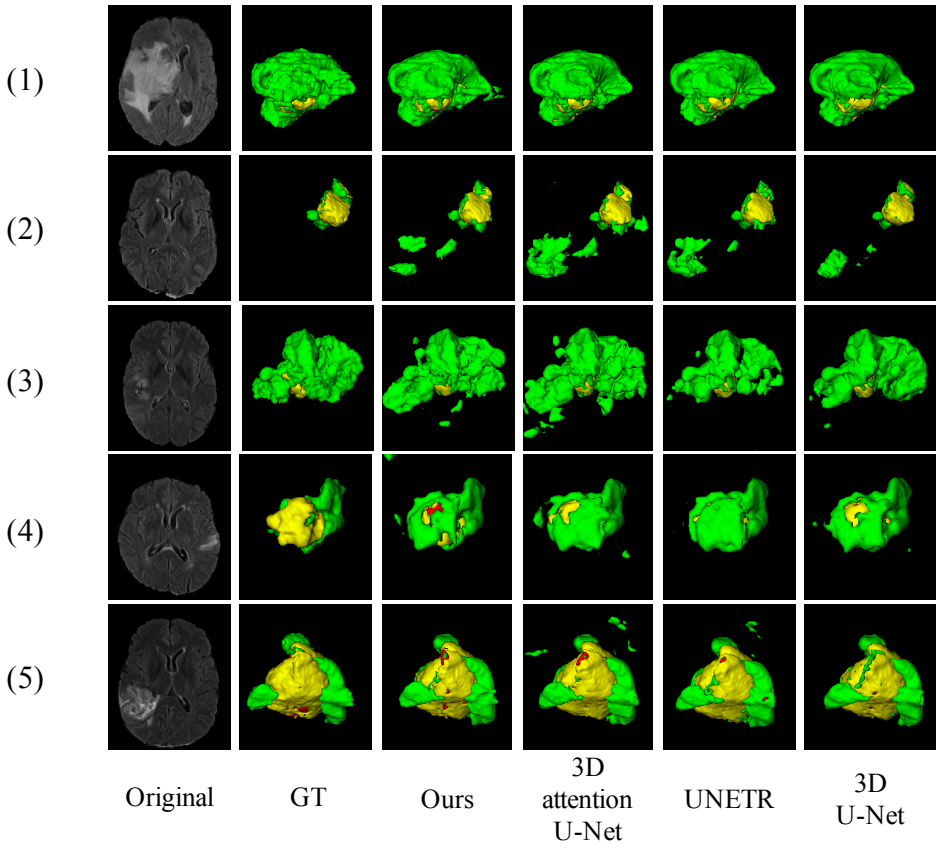


Fig. 4.8. Visualization of segmentation results from our method, 3D attention U-Net, UNETR and 3D U-Net.

In order to observe more clearly the segmentation effect of each network on brain tumors, we performed 3D visualization of the segmentation results. As shown in Fig. 4.8, we visualize segmentation results of our method (DEU-Net), 3D attention U-Net, UNETR and 3D U-Net on five different cases from BraTS 2020 dataset. The green area is peritumoral edema, the yellow area is enhancing tumor, and the red area is non-enhancing tumor. The red area is usually obscured by green and yellow areas. Due to the small size of the tumor in some cases, we enlarged the lesion. We randomly selected 5 cases from the BraTS 2020 dataset. Due to the small size of the tumor in some cases, we enlarged the lesion. Among them, the first column is the original MR image, the second column is the

GT after expert segmentation in the dataset, and the 3rd to 6th columns are our methods, 3D attention U-Net, UNETR and 3D U-Net segmentation results.

After comparative analysis, it can be found that the segmentation results obtained by the automatic segmentation method lack the segmentation details of GT, especially the surface of the lesion is relatively smooth and does not have too many details. Although our method has some over-segmentation in the WT segmentation task compared with GT, compared with the other three methods, the over-segmentation rate of our method is smaller. Moreover, automatic segmentation methods are prone to under-segmentation in ET segmentation tasks. Compared with the other three methods, the under-segmentation rate of our method is also smaller. These results show that we effectively fuse features extracted by Transformer and CNN and improve the network's ability to capture contextual features.

4.5 Discussion

To improve the ability of the 3D model to capture the contextual information of 3D medical images, we explore the importance of CBAM, pre-trained Transformer, and DFFM for extracting 3D medical image features, we propose a dual encoder network called DEU-Net. Among them, CBAM can extract 3D medical image feature in the encoder, pre-trained Transformer can improves its ability to capture contextual features of medical images and increases the learning speed in the encoder, and the DFFM can fuse the features extracted from the Transformer and CNN in the encoder, making full use of the feature extraction capabilities of the two extractors for 3D medical image. Experiments are conducted in three medical image datasets, namely BraTS 2020 and BraTS 2021 brain tumor segmentation datasets. Experimental results show that DEU-Net outperforms SoTA, in terms of Dice from Tab. 4.2 and Tab. 4.3.

Ablation experiments are conductedv on BraTS datasets to evaluate the impact of these three modules on DEU-Net. The results are shown in Tab. 4.4, Tab. 4.5, Tab. 4.6, and Tab. 4.7, from which we can see that the introduction of CBAM, pre-trained Transformer,

and DFFM have improved the performance of DEU-Net in Dice coefficients. These ablation experiments show that CBAM can extract the context information of 3D medical image features very well, DFFM can well fuse the features extracted from the Transformer and CNN in the encoder, making full use of the feature extraction capabilities of the two extractors for 3D medical image from encoder. In addition, the introduction of pre-trained Transformer can achieve better performance on segmentation tasks with higher learning speed, which means that DEU-Net has better accuracy and lower redundancy.

Although DEU-Net performs well on BraTS datasets, there are some shortcomings in this work. (1) DEU-Net does not fully utilize the Transformer to extract medical image features and does not fully integrate the feature maps output by the 12 layers of Transformer. (2) The types of 3D medical image segmentation tasks are diverse, we do not train, validate, and test on other medical image segmentation tasks. (3) The network combines Transformer and CNN as the encoder of the network. Because the standard Transformer is oriented to large data sets, its parameters are large, which greatly increases the parameter load of the network and increases the amount of calculations.

4.6 Summary

For 3D medical segmentation task, we propose DEU-Net, which uses Transformer and CNN to extract 3D medical image features in the encoder. Transformer is a pre-trained model in BTCV, which improves its ability to capture contextual features of medical images and increases the learning speed. We introduce CBAM with each convolutional layer in the encoder part to enhance CNN's feature extraction capabilities for 3D medical images. To fuse the two kinds of features, we proposed DFFM to fuse the features extracted from the Transformer and CNN in the encoder, making full use of the feature extraction capabilities of the two extractors for 3D medical image. DEU-Net registers a Dice score of 0.7836 for ET segmentation task, 0.8384 for TC segmentation task, and 0.9047 for WT segmentation task, respectively. The results demonstrate that DEU-Net outperforms state-of-the-art for three segmentation tasks on the BraTS 2020 dataset.

However, the proposed network does not fully utilize the Transformer to extract medical image features and does not fully integrate the feature maps output by the 12 layers of Transformer. In addition, the types of medical image segmentation tasks are diverse, and how our method performs in other segmentation tasks requires further validation.

5 Conclusion and Future Work

5.1 Conclusion

Medical image segmentation has been widely discussed and concerned in the field of image processing, which has become a basic component and a crucial stage of image processing. In clinical diagnosis, medical images provide doctors with main patient condition information, and medical image segmentation facilitates clinical diagnosis and treatment. In clinical practice, segmentation is typically performed manually. However, when processing a vast number of medical images, the quality of segmentation can vary based on the expertise of the medical professional. This variability underscores the need for a more consistent and efficient method to enhance the performance of segmentation tasks.

Currently, medical image segmentation still faces significant challenges. Firstly, the acquisition and creation of high-quality datasets. Amassing such datasets is a complex task and is often beyond the capacity of a single institution within a limited timeframe. As a result, medical datasets tend to be smaller in size and can exhibit inconsistencies. Secondly, notable characteristic of medical images is the imbalance between the foreground and background areas. Unlike natural scene images, where the foreground and background might be more balanced, medical images often have a much larger background compared to the foreground. The third challenge lies in the inherent complexity and diversity of medical images. The intricate shapes of human tissue structures and the vast individual variations pose additional difficulties for medical image segmentation. Variances in patient age, differences in imaging equipment, and even regional disparities among medical institutions can all impact the overall quality of medical image slices to varying degrees. The diversity in types and storage formats of medical images, as well as the multitude of lesions or organ locations in patients, presents a formidable challenge to the universality of segmentation methods. Inevitably, medical images exhibit features such as blurriness and unevenness, making segmentation methods less universally applicable compared to natural scene images.

The currently popular medical image segmentation method is still a network with an encoder-decoder structure. However, current method implementations do not fully explore feature fusion between different features. Furthermore, a notable challenge with Transformer-based models is their inherent complexity. They often necessitate large datasets for training to achieve optimal performance. Consequently, Transformers may not be the ideal choice for every medical segmentation task.

To address these two challenges in field of medical image segmentation, we proposed two models to segment 2D and 3D medical image respectively.

- For 2D medical image segmentation, we propose an MDSU-Net, a variation of the U-Net, for 2D medical image segmentation. MDSU-Net incorporates both multi-attention and DSC layer for improved performance. The multi-attention module within our framework utilizes dual attention and attention gates to capture rich contextual information and fuse features of different convolutional layers. MDSU-Net uses DSC layer to reduce model complexity without degrading model performance, which is suitable for different segmentation tasks.
- In the field of 3D medical image segmentation, we propose DEU-Net, which uses Transformer and CNN respectively to extract 3D medical image features in the encoder. Transformer is a pre-trained model in BTCV, which improves its ability to capture contextual features of medical images and increases the learning speed. We introduce CBAM with each convolutional layer in the encoder part to enhance CNN's feature extraction capabilities for 3D medical images. To fuse the two kinds of features, we proposed DFFM to fuse the features extracted from the Transformer and CNN in the encoder, making full use of the feature extraction capabilities of the two extractors for datasets of different sizes.

5.2 Future Work

There is a wide variety of tasks in 3D medical image segmentation, and the 3D network we proposed has undergone validation experiments specifically for brain tumor

segmentation. Further verification is needed for other segmentation tasks and different types of images. Given the high complexity of medical images, we will explore new model structures to enhance feature extraction and feature fusion.

We have constructed two segmentation networks for medical images in different dimensions. In the future, we will further explore a universal model that combines tasks from both dimensions. Additionally, regardless of whether it is 2D or 3D medical images, the parameter magnitudes of the two models we proposed are relatively large. For datasets with smaller volumes, we will further reduce model complexity to alleviate computational pressure and enhance the practicality of the model in real clinical environments.

Acknowledgement

Firstly, I would like to take this opportunity to express my sincere gratitude to all those who have assisted and supported me throughout the completion of my doctoral thesis.

I would like to express my gratitude and respect to Prof. Ren Fuji and Prof. Terada for providing the learning environment, resource platform, research guidance and encouragement. Your expertise and insightful perspectives have been crucial to the successful completion of my thesis. Your meticulous guidance enabled me to overcome challenges and continually improve myself during the research process. I would like to thank the Prof. Kang Xin for his valuable suggestions on research details, which has greatly improved my academic ability. The completion of experiments and the thesis would not have been possible without your careful guidance.

I appreciate the rigorous and pragmatic spirit of Prof. Fuketa and Prof. Shishibori who reviewed my doctoral thesis. Your meticulous and thorough review greatly contributed to the quality of my thesis.

I extend my gratitude to my parents for their unwavering support and understanding throughout my entire doctoral period. I am also thankful to my wife and daughter. Despite her own work and my pursuit of a doctorate, my wife provided me with limitless support in both life and research. I express heartfelt thanks and gratitude. During my doctoral period, the birth of my daughter enriched my life and spiritual world.

I also want to thank my colleagues in the research lab. Your suggestions and discussions enriched and deepened my research and thesis. I highly value the time spent working together and the friendships we formed.

Finally, I want to express my gratitude to all the scholars and researchers who provided references and datasets for this thesis. Your work has provided a solid foundation for my research and enriched the content of my thesis.

Once again, thank you to everyone who has supported and helped me. Your contributions have been a significant factor in the successful completion of my doctoral thesis.

References

[1] R. Aggarwal, V. Sounderajah, G. Martin, D. S. Ting, A. Karthikesalingam, D. King, H. Ashrafian, and A. Darzi, "Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis," *NPJ digital medicine*, vol. 4, no. 1, p. 65, 2021.

[2] A. Rahman, J. M. J. Valanarasu, I. Hacihaliloglu, and V. M. Patel, "Ambiguous medical image segmentation using diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 536–11 546.

[3] Y. Wang, T. Bai, T. Li, and L. Huang, "Osteoporotic vertebral fracture classification in x-rays based on a multi-modal semantic consistency network," *Journal of Bionic Engineering*, vol. 19, no. 6, pp. 1816–1829, 2022.

[4] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai *et al.*, "Recent advances in convolutional neural networks," *Pattern recognition*, vol. 77, pp. 354–377, 2018.

[5] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015.

[6] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *2017 international conference on engineering and technology (ICET)*. Ieee, 2017, pp. 1–6.

[7] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: analysis, applications, and prospects," *IEEE transactions on neural networks and learning systems*, 2021.

[8] Z. Liu, L. Tong, L. Chen, Z. Jiang, F. Zhou, Q. Zhang, X. Zhang, Y. Jin, and H. Zhou, "Deep learning based brain tumor segmentation: a survey," *Complex & intelligent systems*, vol. 9, no. 1, pp. 1001–1026, 2023.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[10] S. Asgari Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh, "Deep semantic segmentation of natural and medical images: a review," *Artificial Intelligence Review*, vol. 54, pp. 137–178, 2021.

[11] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 7262–7272.

[12] Z. Tian, C. Shen, and H. Chen, "Conditional convolutions for instance segmentation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 282–298.

[13] Y. Wang, Z. Xu, X. Wang, C. Shen, B. Cheng, H. Shen, and H. Xia, "End-to-end video instance segmentation with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8741–8750.

[14] P.-H. Conze, A. E. Kavur, E. Cornec-Le Gall, N. S. Gezer, Y. Le Meur, M. A. Selver, and F. Rousseau, "Abdominal multi-organ segmentation with cascaded convolutional and adversarial deep networks," *Artificial Intelligence in Medicine*, vol. 117, p. 102109, 2021.

[15] R. Ranjbarzadeh, A. Caputo, E. B. Tirkolaee, S. J. Ghoushchi, and M. Bendechache, "Brain tumor segmentation of mri images: A comprehensive review on the application of artificial intelligence tools," *Computers in biology and medicine*, vol. 152, p. 106405, 2023.

[16] L. Zhang and C. P. Lim, "Intelligent optic disc segmentation using improved particle swarm optimization and evolving ensemble models," *Applied Soft Computing*, vol. 92, p. 106328, 2020.

[17] Y. Lu, X. Qin, H. Fan, T. Lai, and Z. Li, "Wbc-net: A white blood cell segmentation network based on unet++ and resnet," *Applied Soft Computing*, vol. 101, p. 107006, 2021.

[18] H. C. Reis, V. Turk, K. Khoshelham, and S. Kaya, "Insinet: a deep convolutional approach to skin cancer detection and segmentation," *Medical & Biological Engineering & Computing*, pp. 1–20, 2022.

[19] H. Cao, H. Liu, E. Song, C.-C. Hung, G. Ma, X. Xu, R. Jin, and J. Lu, "Dual-branch residual network for lung nodule segmentation," *Applied Soft Computing*, vol. 86, p. 105934, 2020.

[20] N. Altini, B. Prencipe, G. D. Cascarano, A. Brunetti, G. Brunetti, V. Triggiani, L. Carnimeo, F. Marino, A. Guerriero, L. Villani *et al.*, "Liver, kidney and spleen segmentation from ct scans and mri with deep learning: A survey," *Neurocomputing*, vol. 490, pp. 30–53, 2022.

[21] D.-P. Fan, T. Zhou, G.-P. Ji, Y. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "Inf-net: Automatic covid-19 lung infection segmentation from ct images," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2626–2637, 2020.

[22] A. Qi, D. Zhao, F. Yu, A. A. Heidari, Z. Wu, Z. Cai, F. Alenezi, R. F. Mansour, H. Chen, and M. Chen, "Directional mutation and crossover boosted ant colony optimization with application to covid-19 x-ray image segmentation," *Computers in Biology and Medicine*, vol. 148, p. 105810, 2022.

[23] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.

[24] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos, "Advancing the cancer genome atlas glioma

mri collections with expert segmentation labels and radiomic features," *Scientific data*, vol. 4, no. 1, pp. 1–13, 2017.

[25] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge," *arXiv preprint arXiv:1811.02629*, 2018.

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

[27] T. Heimann, B. Van Ginneken, M. A. Styner, Y. Arzhaeva, V. Aurich, C. Bauer, A. Beck, C. Becker, R. Beichel, G. Bekes *et al.*, "Comparison and evaluation of methods for liver segmentation from ct datasets," *IEEE transactions on medical imaging*, vol. 28, no. 8, pp. 1251–1265, 2009.

[28] A. L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. Van Ginneken, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze *et al.*, "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," *arXiv preprint arXiv:1902.09063*, 2019.

[29] R. D. Rudyanto, S. Kerkstra, E. M. Van Rikxoort, C. Fetita, P.-Y. Brillet, C. Lefevre, W. Xue, X. Zhu, J. Liang, I. Öksüz *et al.*, "Comparing algorithms for automated vessel segmentation in computed tomography scans of the lung: the vessel12 study," *Medical image analysis*, vol. 18, no. 7, pp. 1217–1232, 2014.

[30] "Covid-19 ct scans," 2020. [Online]. Available: https://www.kaggle.com/datasets/andrewmvd/covid19-ct-scans

[31] A. E. Kavur, N. S. Gezer, M. Barış, S. Aslan, P.-H. Conze, V. Groza, D. D. Pham, S. Chatterjee, P. Ernst, S. Özkan *et al.*, "Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation," *Medical Image Analysis*, vol. 69, p. 101950, 2021.

[32] U. Baid, S. Ghodasara, S. Mohan, M. Bilello, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer, F. C. Kitamura, S. Pati *et al.*, "The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification," *arXiv preprint arXiv:2107.02314*, 2021.

[33] S.-L. Liew, J. M. Anglin, N. W. Banks, M. Sondag, K. L. Ito, H. Kim, J. Chan, J. Ito, C. Jung, N. Khoshab *et al.*, "A large, open source dataset of stroke anatomical brain images and manual lesion segmentations," *Scientific data*, vol. 5, no. 1, pp. 1–11, 2018.

[34] M. R. Hernandez Petzsche, E. de la Rosa, U. Hanning, R. Wiest, W. Valenzuela, M. Reyes, M. Meyer, S.-L. Liew, F. Kofler, I. Ezhov *et al.*, "Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset," *Scientific data*, vol. 9, no. 1, p. 762, 2022.

[35] D. Ouyang, B. He, A. Ghorbani, M. P. Lungren, E. A. Ashley, D. H. Liang, and J. Y. Zou, "Echonet-dynamic: a large new cardiac motion video data resource for medical machine learning," in *NeurIPS ML4H Workshop: Vancouver, BC, Canada*, 2019.

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

[37] k. s. W. C. y. Anna Montoya, Hasnin, "Ultrasound nerve segmentation," 2016. [Online]. Available: https://kaggle.com/competitions/ultrasound-nerve-segmentation

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

[39] A. Hoover and M. Goldbaum, "Locating the optic nerve in a retinal image using the fuzzy convergence of the blood vessels," *IEEE transactions on medical imaging*, vol. 22, no. 8, pp. 951–958, 2003.

[40] Y. Zhou, X. Kang, F. Ren, H. Lu, S. Nakagawa, and X. Shan, "A multi-attention and depthwise separable convolution network for medical image segmentation," *Neurocomputing*, vol. 564, p. 126970, 2024.

[41] C. Ouyang, C. Chen, S. Li, Z. Li, C. Qin, W. Bai, and D. Rueckert, "Causality-inspired single-source domain generalization for medical image segmentation," *IEEE Transactions on Medical Imaging*, 2022.

[42] H. Hui, X. Zhang, F. Li, X. Mei, and Y. Guo, "A partitioning-stacking prediction fusion network based on an improved attention u-net for stroke lesion segmentation," *IEEE Access*, vol. 8, pp. 47 419–47 432, 2020.

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

[44] M. Z. Alom, C. Yakopcic, M. Hasan, T. M. Taha, and V. K. Asari, "Recurrent residual u-net for medical image segmentation," *Journal of Medical Imaging*, vol. 6, no. 1, pp. 014 006–014 006, 2019.

[45] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, and J. Liu, "Ce-net: Context encoder network for 2d medical image segmentation," *IEEE transactions on medical imaging*, vol. 38, no. 10, pp. 2281–2292, 2019.

[46] J. Sun, F. Darbehani, M. Zaidi, and B. Wang, "Saunet: Shape attentive u-net for interpretable medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23*. Springer, 2020, pp. 797–806.

[47] C. Chen, C. Biffi, G. Tarroni, S. Petersen, W. Bai, and D. Rueckert, "Learning shape priors for robust cardiac mr segmentation from multi-view images," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd*

International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22. Springer, 2019, pp. 523–531.

[48] 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.

[49] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021.

[50] Y. Zhou, X. Kang, and F. Ren, "Mdsu-net: A multi-attention and depthwise separable convolution network for stroke lesion segmentation," in *Proceedings of the 2022 9th International Conference on Biomedical and Bioinformatics Engineering*, 2022, pp. 11–16.

[51] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*. Springer, 2021, pp. 36–46.

[52] W. Ji, S. Yu, J. Wu, K. Ma, C. Bian, Q. Bi, J. Li, H. Liu, L. Cheng, and Y. Zheng, "Learning calibrated medical image segmentation via multi-rater agreement modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 341–12 351.

[53] X. Huang, Q. Wang, J. Chen, L. Chen, and Z. Chen, "Effective hybrid attention network based on pseudo-color enhancement in ultrasound image segmentation," *Image and Vision Computing*, p. 104742, 2023.

[54] K. Qi, H. Yang, C. Li, Z. Liu, M. Wang, Q. Liu, and S. Wang, "X-net: Brain stroke lesion segmentation based on depthwise separable convolution and long-range dependencies," in *Medical Image Computing and Computer Assisted Intervention–*

MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22. Springer, 2019, pp. 247–255.

[55] X. Liu, H. Yang, K. Qi, P. Dong, Q. Liu, X. Liu, R. Wang, and S. Wang, "Msdf-net: Multi-scale deep fusion network for stroke lesion segmentation," *IEEE Access*, vol. 7, pp. 178 486–178 495, 2019.

[56] J. Ruan, S. Xiang, M. Xie, T. Liu, and Y. Fu, "Malunet: A multi-attention and light-weight unet for skin lesion segmentation," in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2022, pp. 1150–1156.

[57] L. Yang, C. Fan, H. Lin, and Y. Qiu, "Rema-net: An efficient multi-attention convolutional neural network for rapid skin lesion segmentation," *Computers in Biology and Medicine*, vol. 159, p. 106952, 2023.

[58] Y. Li, J. Yang, J. Ni, A. Elazab, and J. Wu, "Ta-net: Triple attention network for medical image segmentation," *Computers in Biology and Medicine*, vol. 137, p. 104836, 2021.

[59] C. Qin, B. Zheng, W. Li, H. Chen, J. Zeng, C. Wu, S. Liang, J. Luo, S. Zhou, and L. Xiao, "Mad-net: Multi-attention dense network for functional bone marrow segmentation," *Computers in Biology and Medicine*, vol. 154, p. 106428, 2023.

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

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

[62] 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 *European Conference on Computer Vision*. Springer, 2022, pp. 205–218.

[63] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 574–584.

[64] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang, "Ds-transunet: Dual swin transformer u-net for medical image segmentation," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–15, 2022.

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

[66] H. Zheng, L. Yang, J. Han, Y. Zhang, P. Liang, Z. Zhao, C. Wang, and D. Z. Chen, "Hfa-net: 3d cardiovascular image segmentation with asymmetrical pooling and content-aware fusion," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*. Springer, 2019, pp. 759–767.

[67] C. Chen, X. Liu, M. Ding, J. Zheng, and J. Li, "3d dilated multi-fiber network for real-time brain tumor segmentation in mri," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22*. Springer, 2019, pp. 184–192.

[68] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference,*

Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24. Springer, 2021, pp. 171–180.

[69] W. Sun and R. Wang, "Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with dsm," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 3, pp. 474–478, 2018.

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

[71] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-net and its variants for medical image segmentation: A review of theory and applications," *Ieee Access*, vol. 9, pp. 82 031–82 057, 2021.

[72] G. Du, X. Cao, J. Liang, X. Chen, and Y. Zhan, "Medical image segmentation based on u-net: A review," *Journal of Imaging Science and Technology*, 2020.

[73] C. Chen, C. Biffi, G. Tarroni, S. Petersen, W. Bai, and D. Rueckert, "Learning shape priors for robust cardiac mr segmentation from multi-view images," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*. Springer, 2019, pp. 523–531.

[74] X. Fang, B. Du, S. Xu, B. J. Wood, and P. Yan, "Unified multi-scale feature abstraction for medical image segmentation," in *Medical Imaging 2020: Image Processing*, vol. 11313. SPIE, 2020, pp. 282–288.

[75] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*. Springer, 2018, pp. 3–11.

[76] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE transactions on medical imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.

[77] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. De Lange, P. Halvorsen, and H. D. Johansen, "Resunet++: An advanced architecture for medical image segmentation," in *2019 IEEE International Symposium on Multimedia (ISM)*. IEEE, 2019, pp. 225–2255.

[78] X. Yan, H. Tang, S. Sun, H. Ma, D. Kong, and X. Xie, "After-unet: Axial fusion transformer unet for medical image segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 3971–3981.

[79] K.-L. Tseng, Y.-L. Lin, W. Hsu, and C.-Y. Huang, "Joint sequence learning and cross-modality convolution for 3d biomedical segmentation," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 6393–6400.

[80] J. Zhang, Y. Xie, Y. Wang, and Y. Xia, "Inter-slice context residual learning for 3d medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 2, pp. 661–672, 2020.

[81] J. Sun, Y. Peng, Y. Guo, and D. Li, "Segmentation of the multimodal brain tumor image used the multi-pathway architecture method based on 3d fcn," *Neurocomputing*, vol. 423, pp. 34–45, 2021.

[82] V. Mnih, N. Heess, A. Graves *et al.*, "Recurrent models of visual attention," *Advances in neural information processing systems*, vol. 27, 2014.

[83] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021.

[84] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, and S.-M. Hu, "Attention mechanisms in computer vision: A survey," *Computational visual media*, vol. 8, no. 3, pp. 331–368, 2022.

[85] L. R. Medsker and L. Jain, "Recurrent neural networks," *Design and Applications*, vol. 5, no. 64-67, p. 2, 2001.

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

[87] A. Sinha and J. Dolz, "Multi-scale self-guided attention for medical image segmentation," *IEEE journal of biomedical and health informatics*, vol. 25, no. 1, pp. 121–130, 2020.

[88] T. Henry, A. Carré, M. Lerousseau, T. Estienne, C. Robert, N. Paragios, and E. Deutsch, "Brain tumor segmentation with self-ensembled, deeply-supervised 3d u-net neural networks: a brats 2020 challenge solution," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers, Part I 6*. Springer, 2021, pp. 327–339.

[89] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical image analysis*, vol. 53, pp. 197–207, 2019.

[90] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

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

[92] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in *International MICCAI Brainlesion Workshop*. Springer, 2021, pp. 272–284.

[93] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*. Springer, 2021, pp. 171–180.

[94] X. Yan, H. Tang, S. Sun, H. Ma, D. Kong, and X. Xie, "After-unet: Axial fusion transformer unet for medical image segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 3971–3981.

[95] J. Zhou, Z. Wu, Z. Jiang, K. Huang, K. Guo, and S. Zhao, "Background selection schema on deep learning-based classification of dermatological disease," *Computers in Biology and Medicine*, vol. 149, p. 105966, 2022.

[96] D. Gut, Z. Tabor, M. Szymkowski, M. Rozynek, I. Kucybała, and W. Wojciechowski, "Benchmarking of deep architectures for segmentation of medical images," *IEEE Transactions on Medical Imaging*, vol. 41, no. 11, pp. 3231–3241, 2022.

[97] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[98] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[99] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1314–1324.

[100] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[101] 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.

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

[103] H. Yang, W. Huang, K. Qi, C. Li, X. Liu, M. Wang, H. Zheng, and S. Wang, "Clci-net: Cross-level fusion and context inference networks for lesion segmentation of chronic stroke," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22*. Springer, 2019, pp. 266–274.

[104] W. Yu, Z. Huang, J. Zhang, and H. Shan, "San-net: Learning generalization to unseen sites for stroke lesion segmentation with self-adaptive normalization," *Computers in Biology and Medicine*, vol. 156, p. 106717, 2023.

[105] N. Tomita, S. Jiang, M. E. Maeder, and S. Hassanpour, "Automatic post-stroke lesion segmentation on mr images using 3d residual convolutional neural network," *NeuroImage: clinical*, vol. 27, p. 102276, 2020.

[106] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 11–19.

[107] T. Lei, R. Sun, X. Du, H. Fu, C. Zhang, and A. K. Nandi, "Sgu-net: Shape-guided ultralight network for abdominal image segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 3, pp. 1431–1442, 2023.

[108] Y. Weng, T. Zhou, Y. Li, and X. Qiu, "Nas-unet: Neural architecture search for medical image segmentation," *IEEE access*, vol. 7, pp. 44 247–44 257, 2019.

[109] X. Liu, L. Song, S. Liu, and Y. Zhang, "A review of deep-learning-based medical image segmentation methods," *Sustainability*, vol. 13, no. 3, p. 1224, 2021.

[110] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai *et al.*, "Recent advances in convolutional neural networks," *Pattern recognition*, vol. 77, pp. 354–377, 2018.

[111] Q. Ji, J. Wang, C. Ding, Y. Wang, W. Zhou, Z. Liu, and C. Yang, "Dmagnet: Dual-path multi-scale attention guided network for medical image segmentation," *IET Image Processing*, 2023.

[112] J. Shi, C. Guo, and J. Wu, "A hybrid robust-learning architecture for medical image segmentation with noisy labels," *Future Internet*, vol. 14, no. 2, p. 41, 2022.

[113] B. Landman, Z. Xu, J. Igelsias, M. Styner, T. Langerak, and A. Klein, "Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge," in *MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, 2015.

[114] T. Henry, A. Carré, M. Lerousseau, T. Estienne, C. Robert, N. Paragios, and E. Deutsch, "Brain tumor segmentation with self-ensembled, deeply-supervised 3d u-net neural networks: a brats 2020 challenge solution," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers, Part I 6*. Springer, 2021, pp. 327–339.