Santa Clara University

Department of Computer Science and Engineering

Date: June 06, 2024

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY

Ruopu He

ENTITLED

Residual Transformer Unet for Medical Image Segmentation

BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE

OF

MASTER OF SCIENCE IN COMPUTER SCIENCE AND ENGINEERING

N. Ling

N. Ling (Jun 12, 2024 12:20 PDT)

Thesis Advisor Dr. Nam Ling

Ying Liu

Ying Liu (Jun 12, 2024 12:29 PDT)

Thesis Reader Dr. Ying Liu

Silvia Figueira

Silvia Figueira (Jun 12, 2024 12:55 PDT)

Department Chair Dr. Silvia Figueira

Residual Transformer Unet for Medical Image Segmentation

By

Ruopu He

COEN 497

Master's Thesis Research

Thesis

Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science
in Computer Science and Engineering
in the School of Engineering at
Santa Clara University, 2024

Santa Clara, California, USA
Date: June 06, 2024

Acknowledgements

I would like to express my deepest gratitude to my thesis advisor, Dr. Nam Ling, for his invaluable guidance, support, and encouragement throughout the development of this thesis. His profound knowledge and insightful feedback have been instrumental in shaping the direction and quality of my research. I am sincerely grateful for his patience and unwavering commitment to my academic growth. I also wish to extend my heartfelt thanks to Dr. Ying Liu for being my thesis reviewer, whose meticulous review and thoughtful suggestions have significantly enhanced the clarity and coherence of this thesis.

This thesis would not have been possible without the generous support and mentorship of both Dr. Nam Ling and Dr. Ying Liu. I am deeply appreciative of their dedication and the time they have invested in helping me achieve this milestone. Last but not the least, I would like to thank my family for their love and support.

Abstract

With the continuous development of deep learning theory in the field of medical images, information technology-assisted treatment methods represented by medical image segmentation technology can help doctors to quickly determine the shape and location of the lesions and improve the diagnosis efficiency of brain tumors. Based on deep learning technology, this thesis carries out related research work on MRI image segmentation. The main contents are as follows:

To begin with, acquire and prepare the brain tumor (MRI) image segmentation dataset from the official MICCAI Society website. This involves normalizing the images, cropping, and slicing, as well as scaling the data to ensure that the dataset meets the input specifications of the deep learning model. This paper presents a medical image segmentation method based on an improved Swin U-net. Initially, an atrous spatial pyramid pooling module is introduced at the end of the encoder to capture multi-scale features, allowing the model to effectively understand image at different scales and fully extract contextual information. Subsequently, in the encoder, the original blocks are replaced with residual Swin Transformer Blocks, and on the decoder side, replaced with deep residual convolution blocks. This replacement preserves the original information and alleviates the gradient vanishing issues. Lastly, an attention gate mechanism is introduced in the skip connections, enabling the model to focus more on important features within the feature map and suppress irrelevant information, thereby improving the model's segmentation accuracy.

The experimental results show that the improved segmentation model reached a validation Intersection over Union (IoU) of 89.47%, an increase of 4.36% over the Swin U-net model, demonstrating that it can effectively enhance the accuracy of image segmentation and optimize the results of the original model.

Keywords: semantic segmentation; Residual Transformer U-Net; attention mechanism gate; atrous spatial pyramid pooling; Swin Transformer; deep residual convolution

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Research Background

Brain tumor is a common malignant disease of the brain system. It is a tumor formed by the tissue proliferation and partial cells of the human brain under the influence of complex internal and external environments, including primary tumors and secondary tumors [1]. The mortality rate accounts for 2.4% of the incidence of human tumors. Brain tumors can cause very serious harm to the body, causing symptoms such as confusion and memory loss, seriously reducing the patient's quality of life. Therefore, the early diagnosis and treatment of patients with brain tumors are of great significance.

Currently, the use of Magnetic Resonance Imaging (MRI) is the best way to diagnose brain tumors. It can emit different pulse sequences to obtain different multi-channel brain tumor medical images [2], which can be used for image segmentation tasks in different lesion areas of brain tumors. However, MRI images may have problems such as random field noise.

In recent years, the application of computer vision technology in the field of medical image segmentation has developed rapidly. Since traditional machine learning requires manual annotation of many areas, the actual segmentation process is inefficient, and the result accuracy is low. Deep learning technology can effectively help doctors improve the efficiency of image processing, complete deep feature extraction of medical images, and improve the accuracy of image segmentation.

1.2 Literature Review

In recent years, with the development of computer technology, deep learning theory has attracted the interest of researchers for its learning method in processing large-scale data and its powerful prediction ability [3]. And the application of deep learning in brain tumor (MRI) images has also aroused the interest of more and more scientific researchers [4,5].

Segnet [6] is a deep image semantic segmentation model proposed by Vijay Badrinarayanan in 2015, in which each network encoding layer corresponds to a network decoding layer. The encoder of the SegNet model uses the first 13 convolutional layers of the VGG16 [7] network, and the results of the decoder are input to the SoftMax classifier to complete the classification of image pixels. Compared with other semantic segmentation models, the Segnet model inputs low-resolution feature parameters into the decoder for the first time, significantly reduces the parameters, and finally achieves pixel-level classification of images.

The Fully Convolutional Networks (FCN) segmentation model was proposed by Long Jonathan et al. [8] in 2014. The FCN network is a representative work of deep learning theory in the field of image segmentation. This model eliminates the constraints of neural networks regarding the size of the input image. It is capable of learning features from and performing segmentation on images of any size. Although the fully convolutional network achieves end-to-end pixel-level classification of images, its disadvantage is that the details of the segmented images are not perfect enough.

The symmetric U-shaped segmentation network U-Net [9] is a semantic segmentation network improved and proposed by Ronneberger et al. based on the fully convolutional network architecture. The U-Net model structure is mainly divided into three parts: down-sampling, up-sampling and skip connections. The main components of the down-sampling contraction part and the up-sampling expansion part are completely symmetrical. The skip connection fuses the low-level information (providing the basis for object classification) and the high-level information (providing the basis for accurate segmentation) in up and down sampling to improve the model.

Combining U-Net architecture with Transformer has become one of the research hotspots in the past two years. Swin U-Net [10] is a U-shaped network with a Transformer as its backbone, which utilizes the Transformer to compensate for the deficiencies of U-Net in capturing long-range dependencies, thereby improving the semantic segmentation effects of multi-scale and multi-regional edematous areas. However, the medical images extracted by Swin U-Net still have issues such as blurred edges and missed segmentation targets.

1.3 Contribution of the Thesis

Addressing the issues such as gradient vanishing and the loss of spatial information in classical network segmentation models, this paper proposes an image segmentation model based on an enhanced Swin U-net architecture. The key contributions of this work are as follows:

(1) The application of digital image processing techniques such as image normalization, and image cropping and slicing to preprocess the dataset, thereby enhancing the dataset's image features.

(2) The introduction of an atrous spatial pyramid pooling at the encoder's end to extract multi-scale brain tumor image features and expand the receptive field.

(3) The implementation of residual Swin Transformer Blocks in the decoder and residual convolutional modules in the encoder to prevent model overfitting.

(4) The incorporation of attention gate mechanisms within the skip connections to bolster important features and suppress irrelevant information.

Chapter 2

Method

2.1 Architecture Overview

This paper builds upon the foundation of Swin U-net to achieve segmentation of brain tumor images, with the overall network structure illustrated in the figure 2-1. The enhanced Swin U-net network comprises three main parts: an encoder, a decoder, and skip connections. In the encoder part, the input image is initially processed through Patch Partition operation, dividing the image into equal-sized blocks, which are then altered in channel number through Linear Embedding. These blocks are fed into multiple Residual Swin Transformer Blocks and Patch Merging layers. At the end of the encoder, an Atrous Spatial Pyramid Pooling (ASPP) [11] module is introduced to extract information across different scales, enlarge the receptive field, and capture more detailed information. The Residual Swin Transformer Blocks which is motivated by the Swin Transformer [12] are tasked with feature extraction, while the Patch Merging serves as a down-sampling operation, halving the dimensions of the feature map and doubling the number of channels. The decoder part includes multiple residual convolutional modules and transposed convolutions. The incorporation of residual connections in both Res-Swin Transformer Blocks and residual convolutional modules effectively prevents model overfitting and enhances the model's generalization ability. Transposed convolutions mainly serve as up-sampling operations, doubling the dimensions of the feature maps and halving the number of channels. The final transposed convolution increases the feature map dimensions by four times, without changing the number of channels, and then passes through a convolutional layer with a 1×1 kernel to map the learned

features to the required number of output classes. In the skip connection segment, cross-layer connections are made between the residual Swin Transformer Blocks at the encoder end and the residual convolutional modules at the decoder end to compensate for any loss of information. Additionally, attention gate mechanisms [13] are integrated within the skip connections to focus on important information in the feature map and suppress the irrelevant information, thus enhancing the precision of image segmentation.

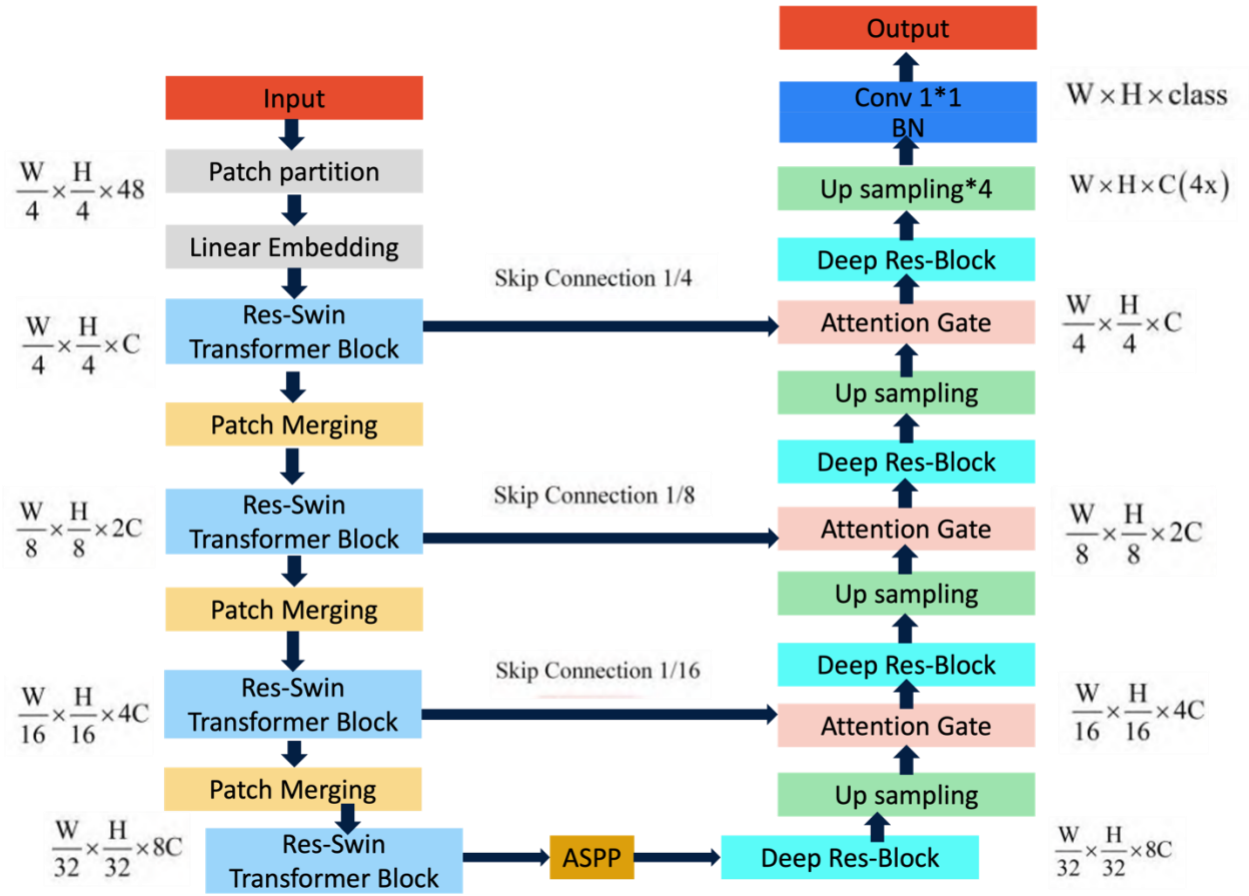


Figure 2-1 Deep Residual Transformer U-net structure.

2.2 Atrous Spatial Pyramid Pooling

Image segmentation networks are mainly based on an encoder-decoder structure, which employs down-sampling to expand the receptive field and up-sampling to restore the original image size. However, this network structure can easily result in the loss of crucial semantic feature information,

and often fail to adequately consider contextual information, leading to significant accuracy degradation. The ASPP (Atrous Spatial Pyramid Pooling) module was first introduced by Chen et al. [11]. It utilizes parallel atrous convolutions with varying dilation rates to capture features at different scales of the image, obtaining varied receptive field features and fusing them. Thereby, it can fully consider contextual information, and improve the network's ability to extract detail features.

The ASPP module mainly has five branches: the first branch is a 1×1 convolution, the second, third, and fourth branches are 3×3 atrous convolutions with dilation rates of 4, 8, and 12, respectively, and the fifth branch is global average pooling. Atrous convolution allows for the expansion of the feature map's receptive field without increasing the number of model parameters. The image is then restored to its original size via bilinear interpolation. The overall structure of the ASPP module is depicted in Figure 2-2. In this paper, by integrating the ASPP module at the end of the residual Swin-transformer encoder, it is possible to fully extract multi-scale information and enlarge the receptive field, which also helps the decoder to recover detail information.

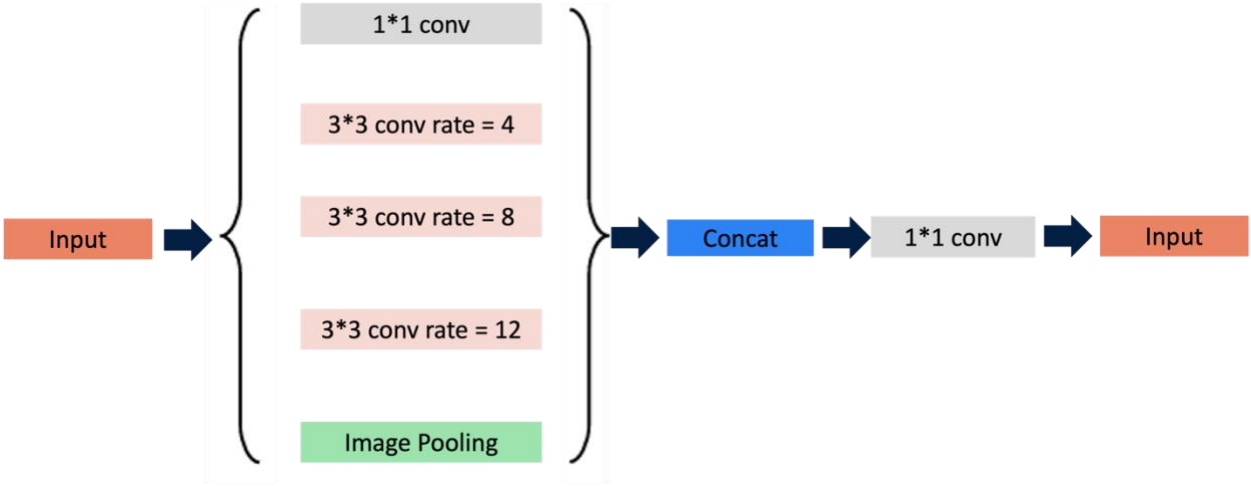


Figure 2-2 Atrous spatial pyramid pooling module.

2.3 Residual Swin Transformer Module

With the continuous increase in the number of layers in deep learning network models, issues such as overfitting and network degradation become more prevalent, further limiting enhancements in model accuracy. To address this issue, He et al. [14] introduced the Res-Net model, which incorporates a directed shortcut connection that directly connects the input across layers to the output. This approach effectively mitigates the gradient vanishing problem without increasing the parameter count, while also capturing more rich semantic features, thereby improving the accuracy of image recognition.

Inspired by Res-Net, this paper incorporates the residual concept into the Swin Transformer Block by using a shortcut connection to link the beginning and end of a Swin Transformer Block. This is achieved through identity mapping to prevent the phenomenon of gradient disappearance in the network model. In this work, modules paired with W-MSA (Window Multi-Head Self Attention) and SW-MSA (Shifted Windows Multi-Head Self-Attention) are considered as a single Swin Transformer Block. The structure of the residual Swin Transformer Block is depicted in Figure 2-3. Experiments have shown that replacing the Swin Transformer Blocks at the encoder end with residual Swin Transformer Blocks effectively improves the model's accuracy in segmenting brain tumor images.

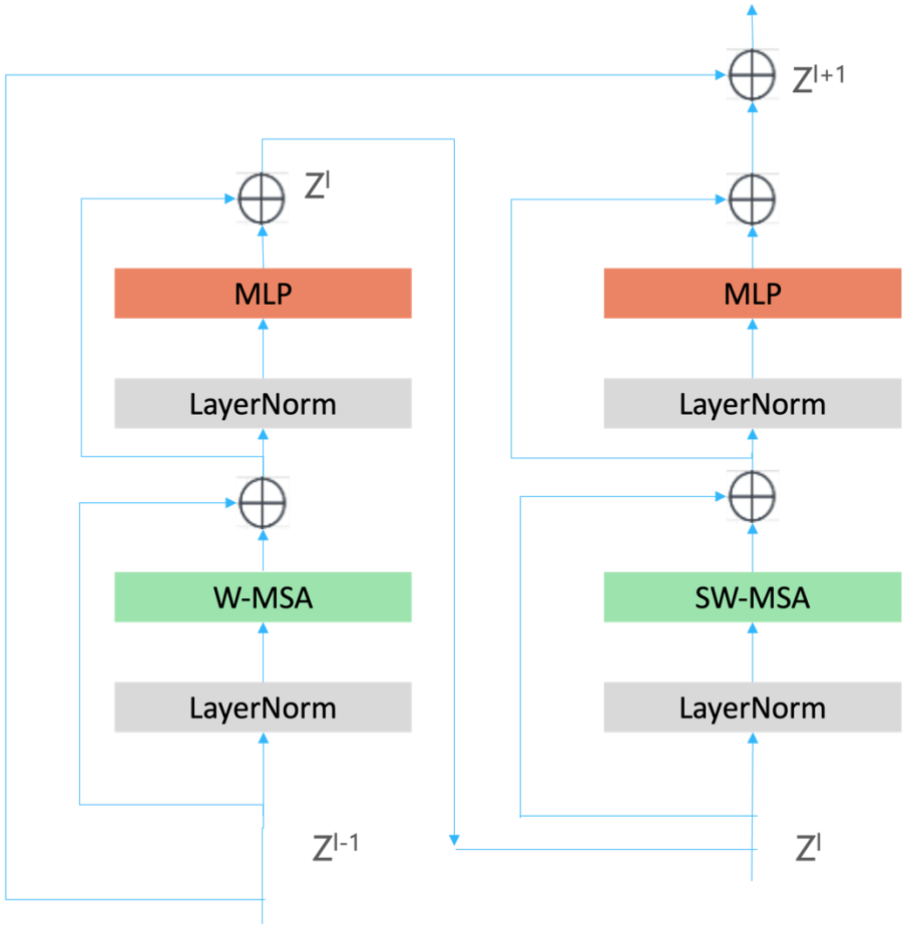


Figure 2-3 Residual Swin Transformer Block.

2.4 Residual Convolutional Modules

This paper incorporates the concept of residual blocks from deep residual networks [14], utilizing improved convolutional modules with Batch Normalization (BN) and Identity Mapping to replace the standard convolutional blocks in the U-Net network. Originally, each convolutional block in U-Net consisted of two 3×3 convolution layers and two ReLU activation functions. In the improved version, each convolutional block first normalizes the output from the upper layer (BN), then extracts image features through a ReLU activation function followed by a 3×3 convolution layer. This process is then repeated once more, and finally, the input and output ends are connected via direct mapping. The residual convolutional module is illustrated in the figure 2-4.

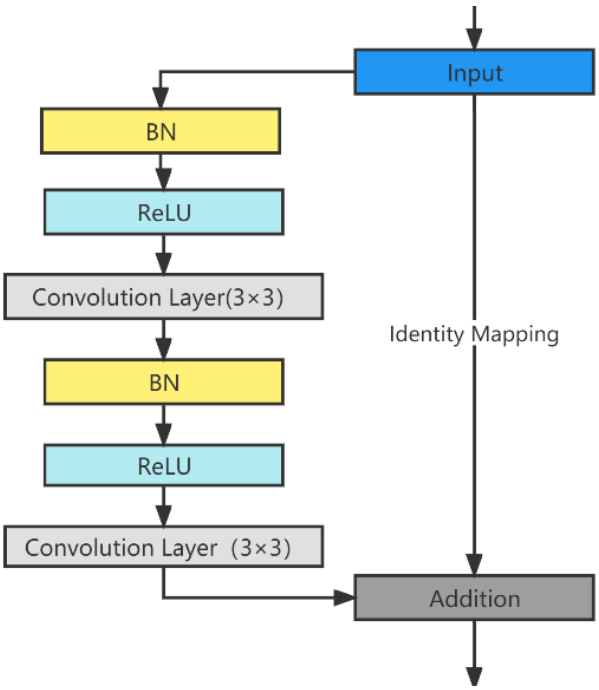


Figure 2-4 Residual Convolutional Block.

2.5 Attention Gate Mechanism

Skip connections serve to mitigate information loss by concatenating low-level features from the encoder with high-level features from the decoder. However, low-level features often contain a significant amount of redundant information and lack specific semantic details, which can impact the accuracy of brain tumor image segmentation. To address this, our study introduces attention gate mechanism [13] into the skip connections.

By combining skip connect with attention gate mechanism, the model automatically learns the shape and size of the target, emphasizes salient features, and suppresses the feature response of irrelevant areas. This is accomplished by a probability-based soft attention to improving model sensitivity and accuracy with minimal computational overhead. The detailed structure of attention gate is shown in Figure 2-5.

Initially, feature maps *Xu* with dimensions $H \times W \times C$, upscaled, and *Xs*, extracted by Residual Swin Transformer block with dimensions $H \times W \times C$, are processed in parallel. Both undergo a 3×3

convolution and batch normalization (BN) operations, resulting in feature maps of dimensions $H\times W\times(C/2)$, denoted as Xu' and Xs'. Then, corresponding elements of Xu' and Xs' are added together, followed by a ReLU operation. This is succeeded by a 1×1 convolution with an output channel count of 1, and then applying BN and a sigmoid activation function to obtain attention coefficient weights α with dimensions $H\times W\times 1$. Finally, Xs is multiplied by α to produce the attention feature map Xr with dimensions $H\times W\times C$.

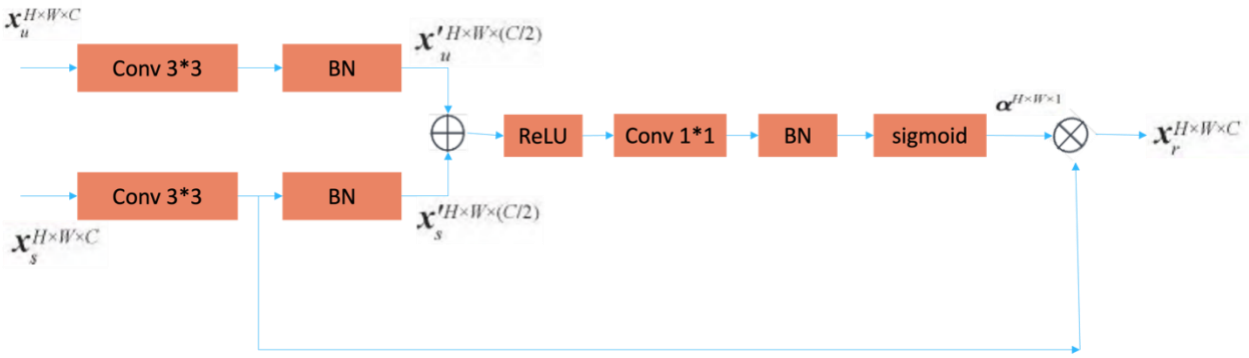


Figure 2-5 Attention Gate.

Chapter 3

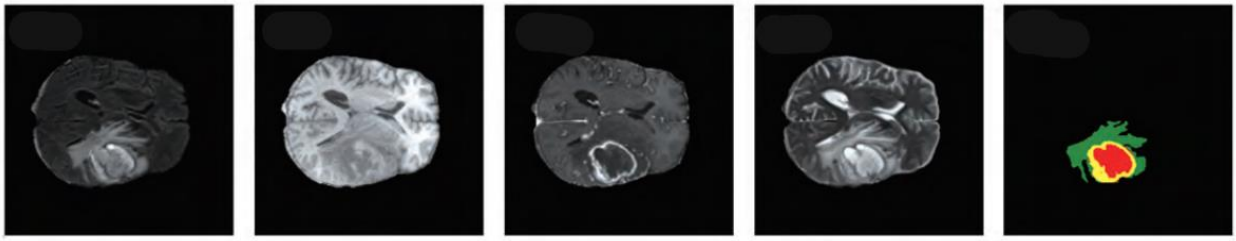
Experiments

3.1 Brain Tumor (MRI) Introduction

3.1.1 Dataset Background

This paper obtained the BraTs2019 dataset through the official website of the American MICCAI Society, which comprises 259 cases of High-Grade Gliomas (HGG) medical images and 76 cases of Low-Grade Gliomas (LGG) medical images. Each case includes four modalities: Flair, T1, T2, and T1ce, with each modality having an image size of 155×240×240.

The BraTs dataset is primarily composed of brain tumor (MRI) images and its four modalities are: T1-weighted imaging(T1), T2-weighted imaging(T2), T1-weighted contrast-enhanced (T1ce), and Fluid-Attenuated Inversion Recovery (Flair). T1 sequences can depict various cross-sectional anatomical details; T2 sequences can determine the location and size of lesions; T1ce sequences involve administering a contrast agent before MRI to differentiate enhanced tumors from necrosis; Flair suppresses the bright signal of cerebrospinal fluid, clearly displaying the entire tumor, but it cannot distinguish necrotic components. The brain tumor imaging results are shown in Figure 3-1, with images from left to right corresponding to Flair, T1, T2, and T1ce, along with the ground truth (GT).



(a) Flair (b) T1 (c) T2 (d) T1ce (e) GT

Figure 3-1 Brain tumor (MRI) image.

3.1.2 Dataset Label

In the BraTs dataset, the segmentation of brain tumor (MRI) images is standardized as follows: the background brain tissue is labeled as 0, necrotic tumor (NET) as 1, peritumoral edema (ED) as 2, and the enhancing tumor region (ET) as 4 [15]. The segmentation task in this paper requires the accurate differentiation of three distinct cancerous regions within the cases: the whole tumor region (WT), the tumor core (TC), and the enhancing tumor region (ET). The whole tumor (WT) encompasses necrotic (NET), edematous (ED), and enhancing tumor (ET) regions, which means WT includes labels 1, 2, and 4. The tumor core (TC) consists of necrotic (NET) and enhancing tumor (ET) regions, indicated by labels 1 and 4, while the ET (enhancing tumor region) is denoted by label 4. The specific segmentation tasks will be as indicated in Table 3-1, and the brain tumor (MRI) images will be illustrated as shown in Figure 3-2.

Table 3-1 Brain tumor segmentation tasks.

Serial	Task Region	Region	Label Data
1	Whole Tumor (WT)	Whole Tumor	1，2，4
2	Tumor Core (TC)	Excluding Edema	1，4
3	Enhancing Tumor (ET)	Severe Tumor	4

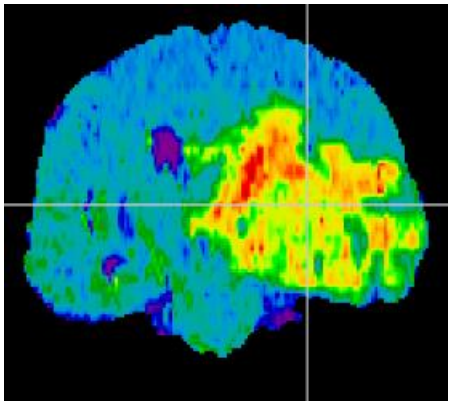


Figure 3-2 Brain tumor (MRI) image.

(The red area represents necrosis, the yellow area indicates the enhancing tumor, and the green area signifies edema.)

3.2 Image Preprocessing

3.2.1 Dataset Acquisition

The training dataset for BraTs2019 was accessed through the official website of the MICCAI Society. However, since the dataset does not include a test set, and considering that a small number of training samples might lead to model overfitting during network training, it was planned to divide the BraTs2019 dataset into a training set, test set, and validation set in a 3:1:1 ratio.

3.2.2 Data Reading

Using the SimpleITK toolkit, which is dedicated to medical image processing in Python, MR (Magnetic Resonance) medical images in nii.gz format were converted into arrays in the form of W×H×Z three-dimensional .npy files.

3.2.3 Multimodal Image Standardization

As each sequence in the BraTs dataset represents a different modality, the image contrast varies across sequences. Therefore, the z-score method is applied to standardize all modal images, setting the mean of the data within each modality to 0 and the standard deviation to 1, resulting in a normal distribution for the entire modality data.

During the standardization of modality data, the concept of percentiles from statistics was employed, using the 99th and 1st percentiles as the boundaries to identify and correct outliers, improving the accuracy of the model algorithm training.

3.2.4 Image Cropping

The MRI images in the BraTs dataset are all sized 240×240, but a large proportion of the image is occupied by black (non-informative) space, while the brain tumor is relatively small, affecting the balance of the dataset. To improve the model's performance in segmenting the target regions of images and reduce background noise, the original modality images were cropped to 224×224.

3.2.5 Slicing and Integration

Given that medical images are three-dimensional, but the plan is to construct a two-dimensional neural network, the NumPy library was used to slice the three-dimensional .npy files into two-dimensional data. Additionally, since each case data is multimodal, slices from the same case of different modalities were combined into multi-channel data. Since the size of each modality slice is 224×224, and there are four modalities per case, the images saved in .npy format are 224×224×4 in size. For the label images' slicing, they are directly saved in a 224×224 .npy data format.

3.3 Experiment Environment

The experiments in this article were performed on a system running the Windows 10 operating system with DirectX 12 support. The computer is equipped with 16GB of RAM and an Nvidia RTX 4090 GPU. Python 3.8 was utilized as the development language, with PyTorch 1.8.1 serving as the development framework.

To prevent overfitting due to overtraining, this study adopted the Early Stopping strategy to control whether learning should be halted. Training would continue only if the results were better than the best recorded; otherwise, it would terminate. The training was scheduled for 250 epochs with an early_stop parameter of 20 and a batch size of 18. At the beginning of training, the learning rate (Lr) was set at 0.0003 with a momentum of 0.9, decreasing by 0.0001 each epoch until training

concluded. The Adam optimizer was used to adjust the learning rate and update the network weights.

3.4 Loss Function

The loss function is utilized to quantify the discrepancy between the predicted values and the ground truth; a smaller loss function value signifies better model robustness. Dice Loss was proposed to address the issue of imbalance between positive and negative samples in semantic segmentation tasks. It originates from the Dice Similarity Coefficient, a metric used to evaluate the similarity between samples. The calculation is given by

$$S(X,Y) = 1 - \frac{2|X \cap Y|}{|X| + |Y|} \tag{3-1}$$

where X and Y represent the set of points contained in the actual and predicted contour regions, respectively.

Dice Loss is a region-based loss function, meaning that the loss and gradient value for the current pixel is related to the prediction of that pixel as well as the true results (ground truth) of other pixels. A Dice coefficient value closer to 0 indicates higher similarity between predictions and ground truth, thus a higher model accuracy; conversely, a value closer to 1 indicates lower similarity and, therefore, lower model accuracy.

3.5 Evaluation Metrics

To assess the segmentation accuracy of the model, this paper employs three evaluation metrics: Dice Similarity Coefficient (DSC), Sensitivity, and Positive Predictive Value (PPV) [16]. The Dice Similarity Coefficient measures the degree of closeness between the model's segmentation predictions and the annotated results. A Dice value closer to 1 indicates a smaller disparity between prediction and annotation, thus more accurate predictions. The Positive Predictive Value (PPV) is the ratio of the predicted tumor regions that are also annotated as tumor to all regions predicted as

tumor by the model and a higher PPV value suggests a lower rate of false positives. Sensitivity reflects the ratio of correctly predicted tumor regions to all annotated tumor regions, representing the true positive rate of the network's segmentation and a lower value indicates a higher rate of false negatives. The formulas are as follows:

$$Dice = \frac{2TP}{FP+2TP+FN} \tag{3-2}$$

$$PPV = \frac{TP}{FP+TP} \tag{3-3}$$

$$Sensitivity = \frac{TP}{TP+FN} \tag{3-4}$$

Where TP (True Positive) is the number of pixels predicted as tumor that are also annotated as tumor, FP (False Positive) is the number of pixels predicted as tumor but annotated as non-tumor, TN (True Negative) is the number of pixels predicted as non-tumor that are also annotated as non-tumor, and FN (False Negative) is the number of pixels predicted as non-tumor but annotated as tumor.

Chapter 4

Analysis of Experiment Result

4.1 Comparison Experiment

To objectively assess the segmentation performance of the proposed method, deep Residual Transformer U-Net, were compared with those of prominent CNN segmentation models such as U-Net, Attention U-Net, Dense U-Net, Nested U-Net, DeepRes U-Net, and Transformer segmentation models like Trans U-Net and Swin U-Net under the same experimental conditions and dataset. As shown in Table 4-1.

Table 4-1 Results of model comparison experiments.

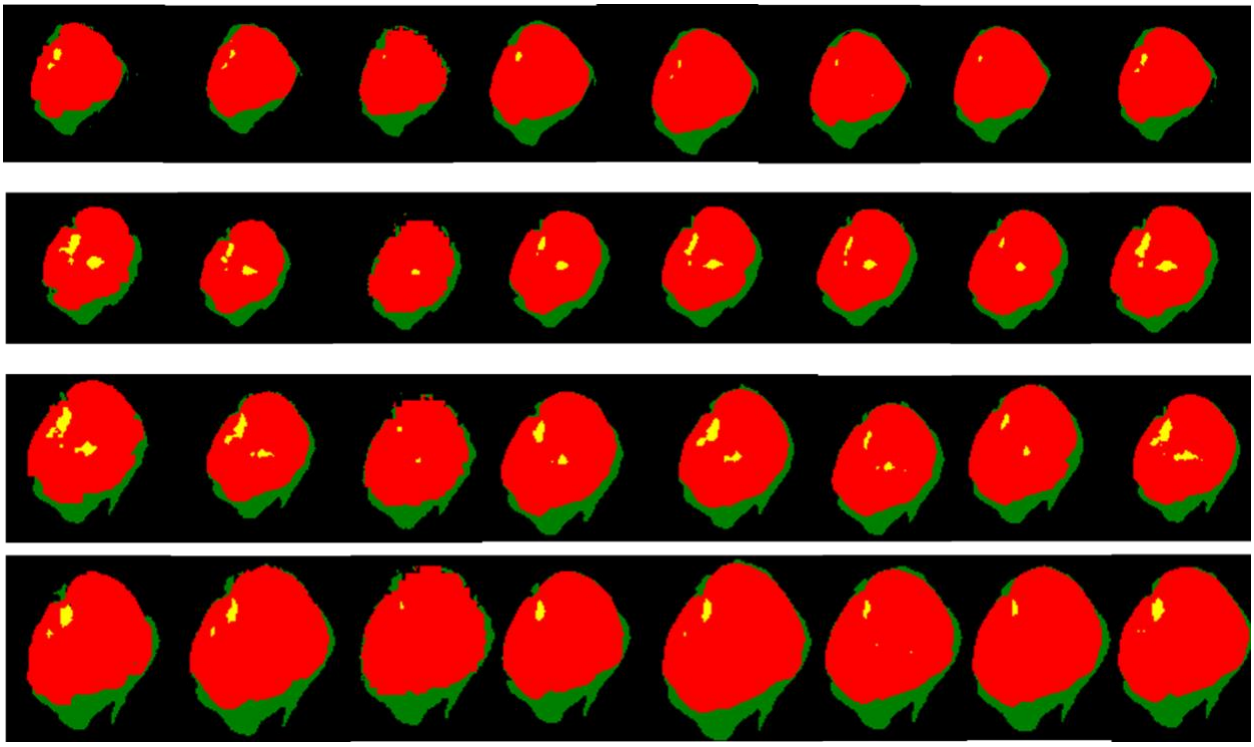
Method	Avg Sensitivity	Avg PPV	Avg Dice
FCN8s	0.8715	0.8809	0.8587
U-net	0.8931	0.8895	0.8742
Attention U-Net	0.8795	0.9026	0.8743
Dense U-Net [17]	0.8743	0.8882	0.8621
Nested U-Net [18]	0.8926	0.9044	0.8833
DeepRes U-Net [19]	0.9059	0.9160	0.8981
Swin U-Net	0.8581	0.8529	0.8316
Trans U-Net [20]	0.9001	0.9166	0.8943
Proposed Method	**0.9178**	**0.9254**	**0.9092**

(The best value under each metric is bolded.)

The Attention U-Net showed a significant improvement in average PPV among the CNN architecture comparison segmentation models, with an increase of 1.31% over U-Net. This indicates that the integration of AG (Attention Gates) into the U-Net base significantly enhances the model's performance. Therefore, the incorporation of AG based on Transformer for brain tumor image segmentation was considered in the design process of this study.

The proposed deep Residual Transformer U-Net model achieved the best results in three evaluation metrics of average sensitivity, average PPV, and average Dice. The model's average sensitivity of 0.9178 showed improvements of 2.47%, 3.83%, 4.35%, 2.52%, and 1.19% over the CNN segmentation models U-Net, Attention U-Net, Dense U-Net, Nested U-Net, and DeepRes U-Net respectively, and 5.97% and 1.77% over the Transformer segmentation models Swin U-Net and Trans U-Net respectively. The average PPV reached 0.9254, showing increases of 3.59%, 2.28%, 3.72%, 2.1%, 0.94%, 7.25%, and 0.88% compared to the seven segmentation models, reflecting the improved model's high similarity between segmentation results and true values. The average Dice coefficient reached 0.9092, indicating increases of 3.5%, 3.49%, 4.71%, 2.59%, 1.11%, 7.76%, and 1.49% compared to the seven models mentioned, demonstrating the model's accurate identification of tumor parts.

For a more intuitive comparison of different models on the segmentation results, the study visualized the results of the experiments. The visualizations, as shown in Figure 4-1, reveal that the model proposed in this paper is closer to the true labels, capable of accurately segmenting finer brain tumor targets, and significantly improves the instances of mis-segmentation and missed segmentation, resulting in the best segmentation performance.



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 4-1 Visualization of segmentation results of different methods.

((a) Ground Truth; (b) proposed method; (c) Swin U-Net; (d) Trans U-Net; (e) Nested U-Net; (f) Dense U-Net; (g) Attention U-Net; (h) DeepRes U-Net.)

In summary, the proposed model shows a high similarity between segmentation results and true values, better identifying brain tumor images and outperforming the other eight models in segmentation performance, effectively enhancing the segmentation accuracy of brain tumor images.

4.2 Ablation Study

4.2.1 The influence of proposed module on model performance

To validate the effectiveness of the introduced modules, this paper uses the Swin U-Net as the baseline model, sequentially stacking the Residual Swin Transformer Block, the Residual Convolutional Module, ASPP, and the Attention Gate mechanism onto the baseline for ablation

experiments. These experiments evaluate the impact of each module on the model's segmentation accuracy through performance metrics. The results of the ablation experiments are shown in Table 4-2. "Res-Swin" represents the Residual Swin Transformer Block, "CNN-decoder" indicates the Residual Convolutional Module, "ASPP" stands for Atrous Spatial Pyramid Pooling, and "AG" refers to the Attention Gates.

Table 4-2 Comparison of evaluation result of ablation experiment.

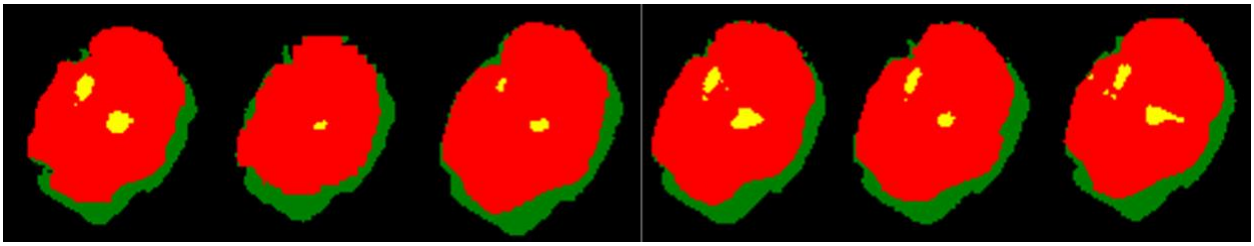
Serial	Baseline	Res-Swin	CNN-decoder	ASPP	AG	Avg Sensitivity	Avg PPV	Avg Dice
1	√					0.8581	0.8529	0.8316
2	√	√				0.9117	0.8983	0.8921
3	√	√	√			0.9087	0.9158	0.8998
4	√	√	√	√		0.9092	0.9125	0.8910
5	√	√	√	√	√	**0.9178**	**0.9254**	**0.9092**

(The best value under each metric is bolded.)

From Table 4-2, it can be observed that the introduction of the Residual Swin Transformer Block improved the model's average sensitivity by 5.36% compared to the baseline model. Further additions of the Residual Convolutional Module increased the average PPV by 1.75%, and subsequent additions of the ASPP module and AG mechanism improved the average sensitivity, average PPV, and average Dice by 0.91%, 0.96%, and 0.94%, respectively. This further confirms that the four modules proposed in this study can effectively enhance the precision of brain tumor image segmentation.

For a more intuitive comparison of the impact of each module on the segmentation results, the study visualized the results of the ablation experiments. The visualizations, as shown in Figure 4-2, reveal that the original Swin U-net model had noticeable issues with under-segmentation and mis-segmentation. However, with the gradual integration of the Res-Swin, Residual Convolutional Module, ASPP, and AG, the target became more accurate and clearer. The segmentation effect of the model became closer to the real labels, and the phenomena of mis-segmentation and under-

segmentation were reduced. This conclusively demonstrates that the model proposed in this paper can effectively improve the results of brain tumor image segmentation.



(a) (b) (c) (d) (e) (f)

Figure 4-2 Visualization of segmentation results in ablation experiments.

((a) Ground Truth; (b) Swin U-Net; (c) + Res-Swin; (d) + Res-Swin + CNN-decoder; (e) + Res-Swin + CNN-decoder + ASPP; (f) + Res-Swin + CNN-decoder + ASPP + AG.)

4.2.2 The influence of the number of AG on model performance

To further investigate the impact of the number of Attention Gates (AG) in skip connections on model performance, experiments were conducted with varying quantities of AG within the skip connections of the model structure. AGs were added sequentially at the resolutions of 1/16, 1/8, and 1/4 scale in the skip connections. The experimental results are shown in Table 4-3. Where AG = 0 indicates no AG incorporation in skip connections, and AG = 3 corresponds to the deep Residual Transformer U-Net model proposed in this paper. As can be observed from Table 4-3, the segmentation accuracy of the model improves with an increasing number of skip connections. Therefore, to enhance the final segmentation accuracy, this paper sets the number of AGs to 3, resulting in the optimal model performance.

Table 4-3 The influence of the number of AG connections on the model performance.

The number of AG	Avg Sensitivity	Avg PPV	Avg Dice
AG = 0	0.8880	0.9041	0.8801
AG = 1	0.8938	0.8973	0.8800
AG = 2	0.8999	0.8964	0.8838
AG = 3	**0.9178**	**0.9254**	**0.9092**

(The best value under each metric is bolded.)

Chapter 5

Conclusion

This paper combines the residual Swin Transformer, residual convolutional modules, AG, and ASPP to propose an improved Swin U-Net based image segmentation model, the deep Residual Transformer U-Net, which achieves precise segmentation of brain tumor MRI images. Building upon the original Swin U-Net architecture, the model integrates an ASPP module to merge features of brain tumor images at different scales and increase the receptive field. It employs residual Swin Transformer blocks and residual convolutional modules for the encoder and decoder layers, respectively, preserving original feature information and effectively preventing overfitting. Attention Gate mechanisms are introduced into the skip connections to enhance important features and suppress irrelevant information.

Experimental results show that the proposed model outperforms the other eight CNN segmentation models and Transformer image segmentation models in brain tumor image segmentation tasks. The study also demonstrates that incorporating Attention Gates into the hybrid Transformer and CNN segmentation models can effectively improve the segmentation accuracy of brain tumor images.

Although the proposed method has achieved significant improvements, there still exists an issue with insufficient detail in the segmentation and the model's complexity is relatively high. Future work will aim to further optimize the deep Residual Transformer U-Net network to reduce its complexity. Additionally, we will utilize traditional signal processing algorithms to improve our

model [21], classify different medical images for higher segmentation fidelity [22], apply our model to new areas [23] and design lightweight segmentation model to reduce training cost for green AI [24].

References

[1] Rebecca L. Siegel MPH, Kimberly D. Miller MPH, et al. Cancer statistics [J]. CA: A Cancer Journal for Clinicians, 2015, 65(1):5-29.

[2] Rupal Snehkunj, Ashish N. Jani, Nalin N. Jani. Brain MRI/CT Images Feature Extraction to Enhance Abnormalities Quantification[J]. Indian Journal of Science and Technology, 2018, 11(1): 1-12.

[3] Ma X, Yu H, Wang Y, et al. Large-scale transportation network congestion evolution prediction using deep learning theory[J]. PloS one, 2015, 10(3): e0119044.

[4] Akkus Z, Galimzianova A, Hoogi A, et al. Deep learning for brain MRI segmentation: state of the art and future directions[J]. Journal of Digital Imaging, 2017, 30(4):449-459.

[5] Pereira S, Pinto A, Alves V, et al. Brain tumor segmentation using convolutional neural networks in MRI images[J]. IEEE Transactions on Medical Imaging, 2016, 35(5):1240-1251.

[6] Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-2495.

[7] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[8] Long J, Shelhamer E, Darrell T. Fully Convolutional Networks for Semantic Segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015,39(4): 640-651.

[9] Ronneberger O, Fischer P, Brox T. U-net: Convolutional Networks for Biomedical Image segmentation[J]. Springer, Cham, 2015, 9351: 234-241.

[10] Cao H, Wang Y, Chen J, et al. Swin-Unet: Unet-like pure transformer for medical image segmentation [EB/OL]. (2021-05-12) [2023-06-06]. https://arxiv.org/abs/2105.05537.

[11] Chen L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation [EB/OL]. (2017-06-17) [2023-06-06].https://arxiv.org/abs/17 06.05587.

[12]. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” CoRR, vol. abs/2103.14030, 2021.

[13] Oktay O, Schlemper J, Le Folgoc L, Lee M, Heinrich M, Misawa K, Mori K, McDonagh S, Hammerla N Y, Kainz B, Glocker B and Rueckert D. 2018. Attention U-Net: learning where to look for the pancreas [EB/OL], [2022-11-23]. https://arxiv.org/pdf/1804.03999.pdf

[14] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770-778.

[15] Menze B, Jakab A, Bauer S, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS) [J]. IEEE Transactions on Medical Imaging, 2015, 30(10):1993-2024.

[16] Wu Y, Lim J, Yang M H. Online object tracking: A benchmark[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, 2013: 2411-2418.

[17] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-denseunet: Hybrid densely connected unet for liver and tumor segmentation from ct volumes," IEEE Transactions on Medical Imaging, vol. 37, no. 12, pp. 2663–2674, 2018.

[18] Z. Zhou, M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation." Springer Verlag, 2018, pp. 3–11.

[19] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual unet," IEEE Geoscience and Remote Sensing Letters, vol. 15, no. 5, pp. 749–753, 2018.

[20] 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.

[21] Y. Pei, Y. Liu, and N. Ling, "Deep learning for block-level compressive video sensing," 2020 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE, 2020, pp. 1-5.

[22] Y. Pei, Y. Liu, N. Ling, L. Liu, and Y. Ren, "Class-specific neural network for video compressed sensing,"2021 IEEE International Symposium on Circuits and Systems (ISCAS), 2021, pp. 1–5.

[23] Y. Pei, Y. Liu and N. Ling, "MobileViT-GAN: A Generative Model for Low Bitrate Image Coding," 2023 IEEE International Conference on Visual Communications and Image Processing (VCIP), Jeju, Korea, Republic of, 2023, pp. 1-5.

[24] Y. Pei, Y. Liu, N. Ling, Y. Ren and L. Liu, "An End-to-End Deep Generative Network for Low Bitrate Image Coding," 2023 IEEE International Symposium on Circuits and Systems (ISCAS), 2023, pp. 1-5.