SCIPEDIA Revista Internacional de Métodos Numéricos para Cálculo y Diseño en Ingeniería

Revista Internacional Métodos numéricos para cálculo y diseño en ingeniería RIMNI

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH In cooperation with CIMNE

Research on thyroid nodule segmentation using an improved U-Net network

Peng Xu[1]

1 College of Optical and Electronic Technology, China Jiliang University, Hangzhou 310018, China

Abstract

To develop a precise neural network model designed for segmenting ultrasound images of thyroid nodules. The deep learning U-Net network model was utilized as the main backbone, with improvements made to the convolutional operations and the implementation of multilayer perceptron modeling at the lower levels, using the more effective BCEDice loss function. The modified network achieved enhanced segmentation precision and robust generalization capabilities, with a Dice coefficient of 0.9062, precision of 0.9153, recall of 0.9023, and an F1 score of 0.9062, indicating improvements over the U-Net and Swin-Unet to various extents. The U-Net network enhancement presented in this study outperforms the original U-Net across all performance indicators. This advancement could help physicians make more precise and efficient diagnoses, thereby minimizing medical errors.

OPEN ACCESS

Published: 06/06/2024

Accepted: 24/05/2024

Submitted: 21/05/2024

DOI: 10.23967/j.rimni.2024.05.012

Keywords:
U-Net
Image Segmentation
Thyroid Nodule Ultrasound Imaging
Deep Learning

1. Introduction

Thyroid diseases, frequently characterized by nodular lesions, are prevalent in the general population. These thyroid nodules are lumps found within the human thyroid gland [1]. According to research statistics, the likelihood of discovering thyroid nodules in asymptomatic adults can be as high as 68%. Among these nodules, 7%-15% are eventually diagnosed as thyroid cancer, the fastest-growing type of malignant tumor [2], which significantly impacts individuals' physical health.

Although thyroid nodule ultrasound imaging technology is mature, the quality of imaging cannot be guaranteed, and shortcomings such as blurred edges of thyroid nodules in images are unavoidable. Differences in the model and type of ultrasound equipment also lead to significant differences in the collected ultrasound images. Additionally, fine-needle aspiration biopsy surgery requires a large amount of medical and human resources and is somewhat invasive for patients. Therefore, this diagnostic method heavily relies on the subjective judgment of attending physicians, which can easily lead to misdiagnosis due to differences in doctors' operational experience and techniques. Unnecessary biopsy surgeries can also cause patients more suffering. Therefore, improving the accuracy of segmentation for ultrasound images of thyroid nodules in computational fields will notably enhance the precision and efficacy of clinical diagnosis and treatment.

Addressing thyroid nodule segmentation, the U-Net model, as introduced by Ronneberger et al. [3], revolutionized deep learning techniques for medical image segmentation by integrating skip connections within its encoder-decoder architecture. This advancement marked a significant milestone, heralding a new era in the field. In a parallel development, Wu Junxia et al. [4] improved the network by introducing a multi-dilation convolutional block. This enhancement enables more accurate segmentation of nodule regions, resulting in the creation of more precise binary masks for medical image segmentation. Hu Yishan et al. [5] introduced attention mechanisms for thyroid nodule segmentation, optimizing low-dimensional features of images and preserving important features through the fusion of high and low-dimensional features. Zhao Kefu et al. [6] fused different feature layers with the U-Net as the backbone network and introduced SE attention mechanisms to further improve segmentation accuracy. Chu et al. [7] introduced a thyroid nodule segmentation network utilizing U-Net architecture, substantially enhancing segmentation accuracy with limited datasets, thereby effectively aiding physicians in diagnosing thyroid nodules. Oktay et al. [8] introduced the Attention-UNet, a novel network model designed to automatically prioritize targets of diverse sizes and shapes. This approach effectively accentuates significant features while mitigating attention towards irrelevant areas. Zhou et al. [9] developed the Deeply Supervised Encoder-Decoder UNet++ network. This diminishes the semantic disparity between the feature maps of encoder and decoder subnetworks. Meanwhile, Chen et al. [10] enhanced the DeepLabv3+ model by integrating a decoder module to refine segmentation outcomes and integrating depth-wise separable convolutions into both the spatial pyramid pooling and decoder modules. Badrinarayanan et al. [11] introduced the SegNet segmentation network. It symmetrically performs downsampling and upsampling. Many models adopt multi-stage segmentation methods, further increasing computational complexity, indicating the need to improve the segmentation speed of many thyroid nodule models.

Currently, the widely used deep learning neural network in

Correspondence: Peng Xu (xupeng336@163.com).

medical imaging is the U-Net network. However, U-Net still faces limitations in thyroid nodule segmentation, such as ineffective utilization of pixel-space information and long training times. The primary contribution of this paper is the research and development of an optimal network structure designed to accurately segment nodules in the thyroid region.

2. Relevant Work

2.1 U-Net Network

In 2015, Ronneberger and colleagues presented the U-Net architecture, which ingeniously integrated skip connections. This innovation marked a significant milestone in medical image segmentation through deep learning methodologies.

As illustrated in Figure 1, the U-Net network is distinguished by its architecture, which consists of three key components: an encoder, a decoder, and a bottleneck layer. The encoder is responsible for feature extraction and learning from the target object through four stages of convolutional and pooling operations, progressively decreasing the size of the feature maps. During the decoding process, the feature maps are upsampled to restore them to the original image size. Concurrently, the innovative skip connection algorithm integrates shallow and deep feature information. This architecture allows U-Net to effectively learn from small-scale datasets in medical imaging.

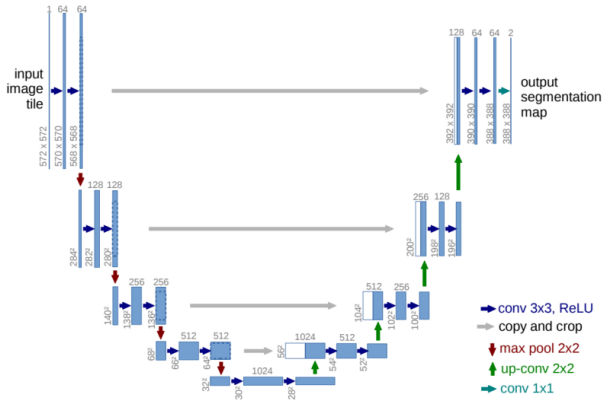


Figure 1 U-Net Network Architecture Diagram

2.2 Swin-Unet Network

Within the domain of medical image segmentation, the demands for precision are exceedingly high. While CNN segmentation algorithms have achieved significant advancements in recent years, they still fall short of the stringent criteria required for medical applications. To address this gap, the Swin-Unet network was introduced, merging the capabilities of U-Net with the Swin Transformer. Figure 2 illustrates the comprehensive structure of this network.

Encoder Part: The Swin-Unet network significantly modifies the convolutional pooling operations of the original U-Net network, replacing them with multiple basic unit blocks from the Swin Transformer network. Each unit block in the network is capable of computing self-attention through local and global perception layers, allowing it to capture image features at various scales.

Decoder Component: Comparable to the U-Net architecture, Swin-Unet incorporates skip connections during upsampling, reinstating the reduced feature maps to the original image dimensions. However, the key distinction lies in replacing conventional convolutional operations with Swin Transformer blocks for feature learning.

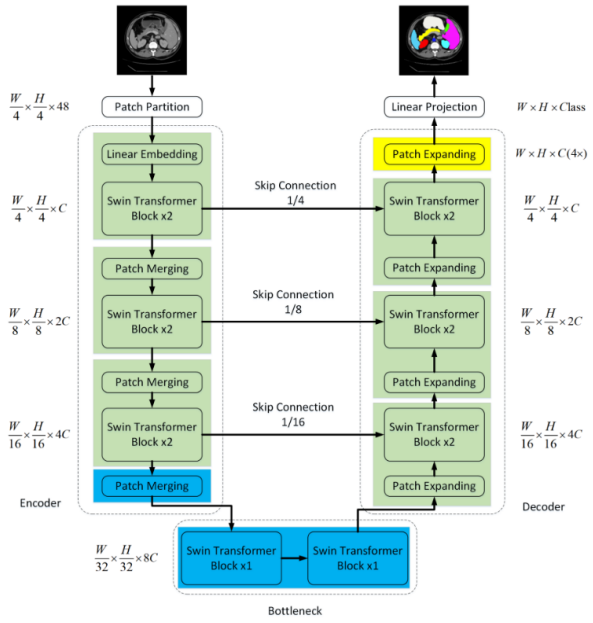


Figure 2 Swin-UNet Network Architecture Diagram

3. Improved U-Net Network

The enhanced U-Net network presented in this paper intelligently integrates the traditional U-Net with elements from Swin-Unet. It utilizes the U-Net structure as the backbone, replacing its convolutional units with Swin Transformer Blocks derived from Swin-Unet. This integration allows for the effective combination of global and local contextual information, thus enhancing segmentation accuracy and robust generalization. Additionally, multi-layer perceptrons are employed at lower levels to model complex features, which significantly reduces both computational complexity and the number of parameters, while still maintaining high segmentation accuracy. The comprehensive framework of this model is showed in Figure 3.

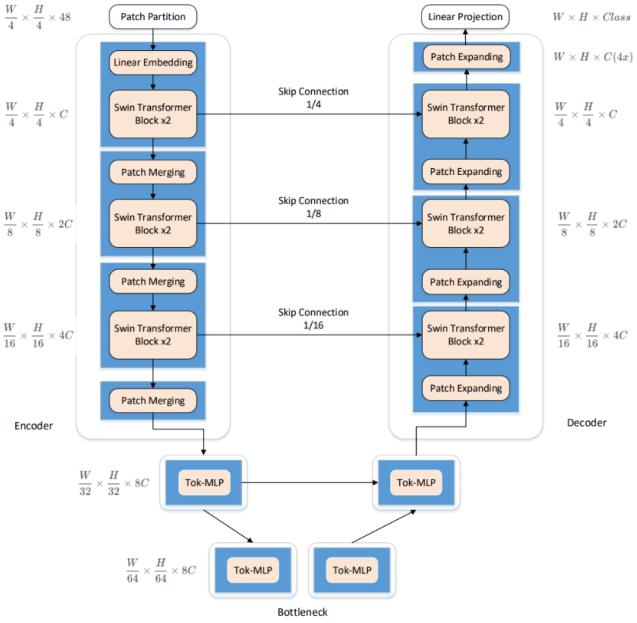


Figure 3. Schematic of the Improved U-Net Network Structure

3.1 Improved Network Composition

3.1.1 Encoder

In the encoder, traditional convolutional operations, typical in the U-Net network, are replaced with Swin Transformer Blocks from the Swin-Unet architecture to enhance feature learning. Following each operation, a patch merging layer is utilized for downsampling, reducing the size and channel count of the feature maps. Downsampling takes place at a factor of 2 during each operation, wherein elements are selected at fixed positional intervals along both row and column directions before being concatenated.

Distinguishing itself from traditional modules, the concept of a movable window is introduced into the improved network units. As shown in Figure 4, the structure diagram of a basic unit block is presented.

Each unit block within the network is configured to encompass layer normalization (LN), a multi-head self-attention mechanism (MSA), a residual connection, and two multilayer perceptrons (MLP). The block incorporates two types of attention mechanisms: a window-based multi-head self-attention mechanism (W-MSA) and a shifted window-based multi-head self-attention mechanism (SW-MSA) [12]. This design facilitates the employment of continuous unit blocks that utilize a movable window concept, enhancing the flexibility and effectiveness of the attention mechanisms in capturing varying spatial features.

$$Z^{l} = W - MSA\left(LN\left(z^{l-1}\right)\right) + z^{l-1} \quad (1)$$

$$z^{l} = MLP\left(LN\left(^{l}\right)\right) + ^{l} \quad (2)$$

$$Z^{l+1} = SW - MSA\left(LN\left(z^{l}\right)\right) + z^{l} \quad (3)$$

$$z^{l+1} = MLP\left(LN\left(^{l+1}\right)\right) + ^{l+1} \quad (4)$$

In equations (1) and (2), ^{l} and z^{l} represent the outputs of the (SW-MSA) module and the MLP module of the first block, respectively. The computational approach for self-attention is as follows:

$$Attention\left(Q, K, V\right) = SoftMax\left(\frac{QK^{T}}{\sqrt{d}} + B\right)V \quad (5)$$

In equation (5), $Q,K,V \in R^{M^2 \times d}$ represent the query, key, and value, respectively. M^2 and d denote the number of patches in the window and the dimension of the query or key, respectively. The value is derived from the bias matrix $\in R^{(2M-1)\times(2M+1)}$.

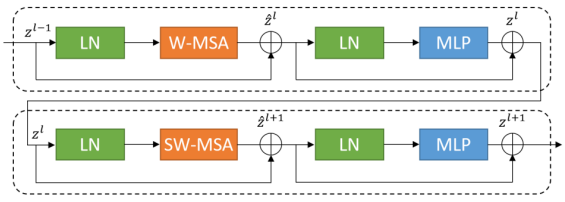


Figure 4. Schema of Swin Transfomer Block

3.1.2 Shifted MLP

During the shifted MLP stage, before tokenization, the first operation performed is to shift the axes of the convolutional feature channels, aiding the multilayer perceptrons in focusing solely on the positional features of the convolutional features. To introduce more locality into the originally entirely global model, a window-based attention mechanism is employed at this stage, enabling the model to better integrate both global and local feature information. As illustrated in Figure 5, the shifted MLP schematic depicts features moving across width and height within two blocks, dividing features into different partitions and shifting their positions along the specified axes.

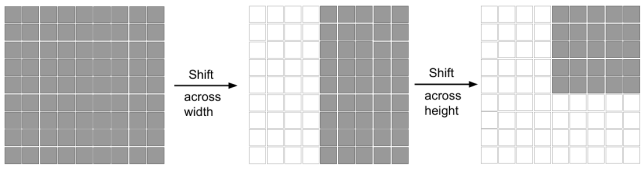


Figure 5. Diagram of the Shifted MLP (Multilayer Perceptron)

3.1.3 Tokenized MLP Stage

In the tokenized MLP stage, features undergo an initial transformation and projection onto tokens, where the channel count is adapted to align with the number of tokens. This step ensures proper correspondence between feature dimensions and the token structure. Subsequently, the tokens are forwarded to the shifted MLP for cross-width movement. The entire process employs depth-wise separable convolution (DWConv) for the following reasons:

1. Depth-wise separable convolution is advantageous for encoding positional information of features. Experimental results indicate that convolutional layers in MLPs are sufficient for encoding positional information and outperform standard positional encoding in practical performance.

2. DWConv has fewer parameters, In the tokenized MLP stage, features are initially transformed and projected onto tokens, with the channel count adjusted to match the number of tokens.

The computational process of the tokenized MLP stage module involves:

$$X_{shift} = Shift_{W}(X); T_{W} = \&Tokenize\left(X_{shift}\right), \quad (6)$$

$$Y = f\left(DWConv\left(\left(MLP\left(T_{W}\right)\right)\right)\right) \quad (7)$$

$$Y_{shift} = Shift_{H}(Y); T_{H} = \&Tokenize\left(Y_{shift}\right), \quad (8)$$

$$Y = f\left(LN\left(T + MLP\left(GELU\left(T_{H}\right)\right)\right)\right), (9) \quad (9)$$

3.1.4 Decoder

The decoder has a symmetric structure to the encoder, both composed of Swin Transformer block unit modules. The key distinction involves the use of patch expansion operations in the decoder, which essentially serve as the inverse of patch merging operations. This reversal process is critical for reconstructing the image from compressed features to its original dimensionality during the decoding phase. It performs upsampling operations on the features extracted by the decoder and reassembles the feature maps into higher-resolution ones.

3.1.5 Skip Connection

Similar to most U-shaped network structures, skip connection operations fuse the feature information of downsampling and upsampling, effectively reducing information loss during downsampling to achieve better thyroid nodule segmentation.

4. Experiment and Results

4.1 Experimental Dataset

The thyroid nodule ultrasound scan images used in this study are sourced from the publicly available TN3K (thyroid nodule 3 thousand) dataset (Gong et al., 2021), which includes 3493 ultrasound images with pixel labels. The dataset comprises high-quality nodule mask annotations from various devices and views.

4.2 Experimental Design

The experimental setup includes the following parameters: image dimensions are uniformly adjusted to 256×256×1, with an initial learning rate set at 0.0005, and a batch size of 8. The Adam optimizer is employed for model optimization. The models undergo training across over a span of 100 epochs.

The BCE loss function treats each pixel as an independent binary classification problem, calculating the loss for each pixel. It offers good and stable focus on individual pixels. While the Dice loss function demonstrates excellent experimental performance for small target objects, maintaining stable training results is challenging. Based on the above studies, this paper aims to balance stability and accuracy in thyroid nodule segmentation training. Therefore, the BCEDiceLoss is utilized during training, combining the advantages of both loss functions to achieve better experimental results.

The BCE (Binary Cross-Entropy) loss function is defined as follows:

$$L_{\mathrm{BCE}}(X, Y, \hat{Y})^{\leftrightarrow} = \frac{1}{P}\sum_{\mathrm{ij}} -(R_{ij}\log(P_{ij}) + (1 - R_{ij})\log(1 - P_{ij})) \quad (10)$$

In equation (10), X represents the initial thyroid nodule image, Y the true labels, and $\hat{Y}$ the corresponding predicted labels.

The Dice loss function is defined as:

$$L_{\mathrm{Dice}}(X, Y, \hat{Y}) = 1 - \frac{\sum_{ij} P_{ij}R_{ij} + \varepsilon}{\sum_{ij}(P_{ij} + R_{ij}) - \sum_{ij} P_{ij}R_{ij} + \varepsilon} \quad (11)$$

In equation (11), a represents a smoothing coefficient, which prevents situations like zero denominators. The computation of BCEDiceLoss is as follows:

$$L_{BCEDice} = \alpha L_{BCE} + (1 - \alpha)L_{Dice} \quad (12)$$

5. Results and Discussion

5.1 Evaluation Metrics

The evaluation metrics employed include the Dice coefficient, precision (P), recall (R), and F1-score. The Dice coefficient measures the similarity between sets. Values for the BCE loss function fall within the range of 0 to 1, with higher values denoting improved segmentation performance achieved by the model. Precision, recall, and F1-score provide additional insights into the accuracy and reliability of the segmentation.

The formula for calculating the Dice coefficient is as follows:

$$Dice = \frac{2\,|A \cap B|}{|A| + |B|} \quad (13)$$

In equation (13), A and B respectively represent the actual and predicted areas of the image.

The formulas for calculating precision (P), recall (R), and the F1-score are as follows:

$$P = \frac{T_P}{T_P + F_P} \quad (14)$$

$$R = \frac{T_P}{T_P + F_N} \quad (15)$$

$$F_1 = 2 \times \frac{P \times R}{P + R} \quad (16)$$

5.2 Results and Analysis

5.2.1 Comparing the Training Effects of Different Loss Functions

The initial set of experiments aims to assess the training effects of the enhanced network model proposed in this paper, employing BCE, Dice, and BCEDice loss functions.

Table 1 Training Results with Various Loss Functions

Loss Function	Dice	P	R	F_1
BCE	0.9016	0.9142	0.8959	0.9016
Dice	0.9031	0.9074	0.9044	0.9031
BCEDice	0.9062	0.9153	0.9023	0.9062

As shown in Table 1, based on the data analysis provided in the table above, it is apparent that for the enhanced neural network proposed in this paper, when trained using the BCE loss function, the Dice score is 0.9016, which represents the lowest score among the evaluated functions. However, when the model is trained using the Dice loss function, a slight improvement is observed compared to utilizing the BCE function, with the Dice score increasing to 0.9031.

When combining the BCE loss function with the Dice loss function for training thyroid nodule ultrasound images, experimental results show that compared to using each loss function individually, the combination of both yields the highest Dice coefficient, precision, and score, reaching 0.9062, 0.9153, and 0.9062, respectively. The results suggest that the enhanced U-Net network proposed in this paper demonstrates superior performance in balancing stability and segmentation effectiveness when utilizing the BCEDice loss function.

5.2.2 Comparison of Segmentation Performance of Different Networks

To further assess the effectiveness of the enhanced model presented in the study, the same dataset is used to train three distinct network models: U-Net, Swin-Unet, and the enhanced model. Subsequently, the segmentation accuracy is evaluated.

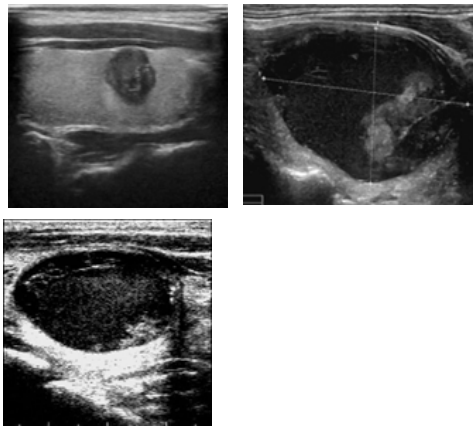
Table 2 Training Outcomes Across Different Algorithms

Network	Dice	P	R	F_1
Swin-Unet	0.7322	0.7401	0.7524	0.7322
U-Net	0.8971	0.8913	0.9106	0.8971
Improved U-Net	0.9062	0.9153	0.9023	0.9062

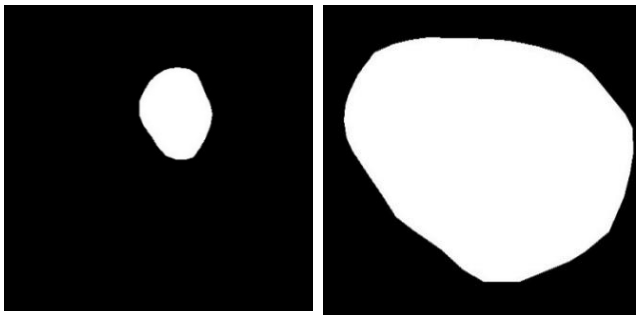
Based on Table 2, the Dice score of the Swin-Unet network reaches 0.7322, which is the lowest among the evaluated networks. The U-Net network achieves a Dice score of 0.8971,

presenting a notable improvement of 22.52% compared to Swin-Unet. After optimization, the enhanced neural network proposed in this paper achieves the highest Dice value of 0.9062 and the highest accuracy of 0.9153. Compared to Swin-Unet, it demonstrates a substantial enhancement of 23.76% and 23.67%, respectively. Furthermore, compared to U-Net, it also showcases improvements of 1.01% and 2.69%, respectively. In summary, the enhanced network proposed in this paper exhibits the most superior segmentation performance.

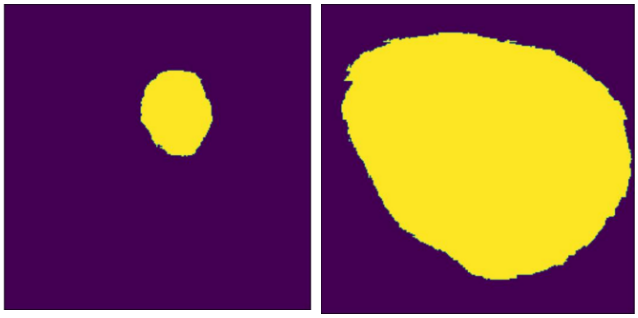
Figure 6 illustrates the segmentation outcomes attained with different neural networks using the same dataset. The enhanced U-Net network proposed in this study is evaluated alongside expert gold standards, Swin-Unet, U-Net, and other well-known network models. The segmentation results from the Swin-Unet network show jagged edges and less smooth nodule edge segmentation, leading to suboptimal outcomes. In the case of U-Net, there are evident under-segmentations with significant discrepancies in the segmented area of some nodules, resulting in inaccurate segmentation results. However, the use of the improved U-Net network introduced in this research produces smoother edges of the segmented thyroid nodules, and the edge contours more closely align with those of the expert gold standard. Moreover, the errors in shape and segmented area are smaller compared to those seen with U-Net and Swin-Unet. The findings suggest that the improved U-Net network provides superior performance in thyroid nodule segmentation.



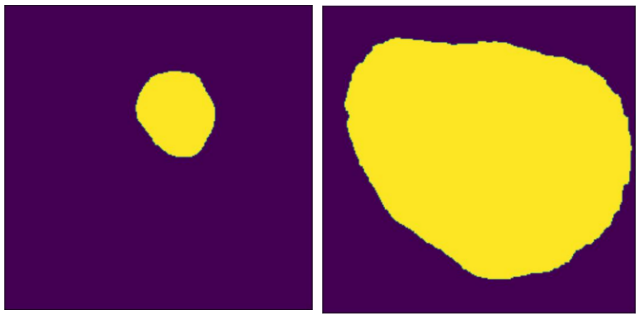
(a) Original Image



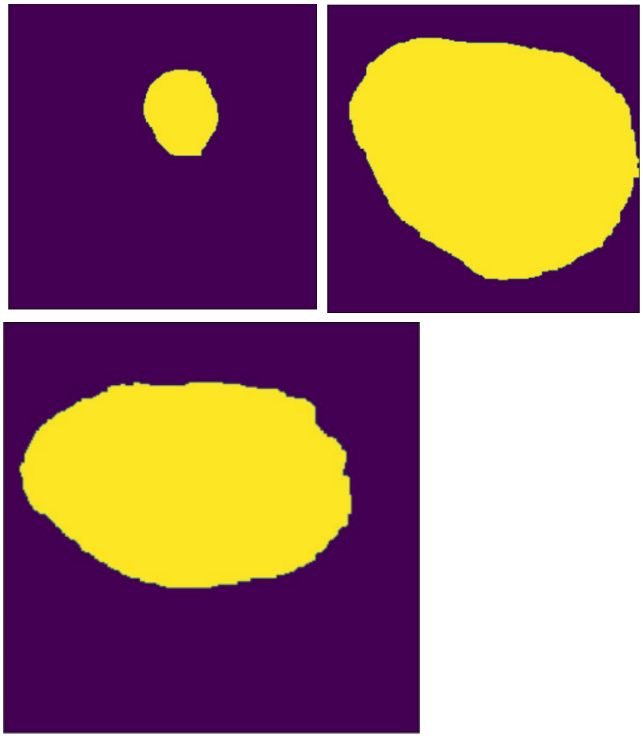
(b) Expert Gold Standard



(c) Swin-Unet Segmentation Outcome



(d) U-Net Segmentation Outcome



(e) Segmentation Results after the Improvements in this Study

Figure 6. Ultrasound Image Segmentation Results of Thyroid Nodules Using Different Networks

This paper presents a method that improves upon the original U-Net network by replacing the standard convolutional blocks in the U-Net architecture with Swin Transformer blocks. This modification introduces local-to-global self-attention mechanisms in the encoder, significantly improving the model's ability to generalize robustly. Additionally, a tokenized multilayer perceptron module is integrated to effectively model features using multilayer perceptrons. Following downsampling in the encoder, features are efficiently tokenized and projected. Through the adoption of a parameter-efficient design, the model attains an optimal equilibrium between segmentation accuracy and computational efficiency.

6. Conclusion

This study provides an enhanced neural network derived from U-Net for the segmentation of thyroid nodule ultrasound images. The following enhancements are incorporated into the U-Net network:

1. Integration of unit modules from Swin Transformer into the model encoder for feature learning, enabling a self-attention mechanism from local to global.

2. Employing multilayer perceptrons (MLPs) at the lower levels for feature modeling, while considering the impact of model dimensions on parameter count and computational complexity in the overall design. This design choice reduces parameter count and improves segmentation speed and accuracy.

3. Combining the advantages of both BCE and Dice loss functions by using the BCEDice loss function, which balances stability and segmentation accuracy, further enhancing the model's performance.

Experimental findings suggest that the enhanced network model proposed in this study attains superior segmentation accuracy metrics. Specifically, it achieves a Dice coefficient of 0.9062, a precision rate of 0.9153, an average recall of 0.9023, and an average F1 score of 0.9.

References

[1] Guth S, Theune U, Aberle J, Galach A, Bamberger CM. Very high prevalence of thyroid nodules detected by high frequency (13 MHz) ultrasound examination. Eur J Clin Invest 2009; 39: 699–706.

[2] Haugen BR, Alexander EK, Bible KC, et al. 2015 American Thyroid Association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: the American Thyroid Association Guidelines Task Force on thyroid nodules and differentiated thyroid cancer. Thyroid 2016; 26: 01–133.

[3] Ronneberger O, Fischer P, Brox T. U-net: Convolutional Networks for Biomedical Image Segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 2015: 234-241.

[4] Ma Xiaoxuan, Sun Boyang, Liu Weifeng, et al. AMSeg: A Novel Adversarial Architecture Based Multi-Scale Fusion Framework for Thyroid Nodule Segmentation. IEEE Access, 2023, 11: 72911-72924.

[5] Hu Yishan, Qin Pinle, Zeng Jianchao, et al. Ultrasound thyroid segmentation network based on feature fusion and dynamic multi-scale dilated convolution. Journal of Computer Applications, 2021, 41(3): 891-897.

[6] Sun J, Li C, Lu Z, et al. TNSNet: Thyroid nodule segmentation in ultrasound imaging using soft shape supervision. Computer Methods and Programs in Biomedicine, 2022, 215, 106600.

[7] Chu C, Zheng J, Zhou Y. Ultrasonic thyroid nodule detection method based on U-Net network. Computer Methods and Programs in Biomedicine, 2021, 199: 105906-105912.

[8] Oktay O, Schlemper J, Folgoc L L, et al. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999, 2018.

[9] Zhou Z, Rahman Siddiquee M M, Tajbakhsh N, et al. Unet++: A nested u-net architecture for medical image segmentation. Deep learning in medical image analysis and multimodal learning for clinical decision support. Springer, Cham, 2018: 3-11.

[10] Chen L C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 2018: 801-818.

[11] Badrinarayanan V, Kendall A, Cipolla R. Segnet: A Deep

Convolutional Encoder-Decoder Architecture for Image Segmentation. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(12): 2481-2495.

[12] Dan Y, Jin W, Wang Z, et al. Optimization of U-shaped pure transformer medical image segmentation network. PEERJ Computer Science, 2023, 9, 1515.