Improved U-Net Segmentation Model for Thyroid Nodules

Rong Hu, Hao Wang, Suzhang Zhang*, Weiping Zhang, Pan Xu

***Abstract*—To address the challenges posed by the irregular boundaries and varying sizes of thyroid nodules, an improved U-Net encoder incorporating residual convolutional blocks and multi-scale attention modules is proposed. This modification aims to improve the U-Net model's ability to accurately segment thyroid nodules with diverse features and sizes. In particular, the paper focuses on refining the skip connections within the U-Net model to better handle the complex boundary structures of thyroid nodules and their fusion with surrounding tissues. A multi-level feature fusion module is proposed to enhance the skip connections, allowing the model to capture boundary details more precisely. To mitigate the risk of overfitting introduced by these enhancements, an auxiliary supervisory branch mechanism is integrated into the decoder. Furthermore, given the class imbalance inherent in thyroid ultrasound images, a joint loss function incorporating Focal loss is employed to facilitate pixel-level predictions and ensure overall structural consistency of the target.**

Comprehensive experiments were conducted on a real-world dataset to evaluate the performance of the improved U-Net model. The results demonstrated its effectiveness in thyroid nodule segmentation. Compared with six other state-of-the-art segmentation models, the improved U-Net achieved superior performance with a Dice coefficient of 89.05%, a Jaccard index of 83.61%, pixel accuracy of 98.29%, sensitivity of 94.69%, and specificity of 98.41%, indicating that the improved model can effectively assist thyroid ultrasonographers in clinical diagnoses.

***Index Terms*—U-net, thyroid nodules, multi-scale attention module, multi-feature fusion, deep supervision.**

I. Introduction (related research)

ACCURATE segmentation of thyroid nodule contours not only provides precise localization but also facilitates the extraction of shape, size, and other features of the nodule. These features can then be used to assess the nature of the nodule, which plays a critical role in personalized and intelligent treatment [1]. Nodule segmentation methods in thyroid ultrasound images are conventionally classified into three principal methodological frameworks: (i) contour-based techniques employing edge detection paradigms, (ii) region-based strategies utilizing intensity thresholding mechanisms, and (iii) machine/deep learning-based architectures incorporating CNN.

Early investigations in thyroid nodule segmentation predominantly emphasized contour-based methodologies. Maroulis et al. [2] introduced the Variable Background Active Contour (VBAC) model for ultrasound-based thyroid nodule segmentation. Subsequently, Iakovidis et al. [3] enhanced this framework through their Genetic Algorithm-VBAC (GA-VBAC) model. Gui et al. [4] innovatively integrated level set theory with isoperimetric constraints, replacing traditional contour length regularization in the ACWE model with a compact shape prior derived from isoperimetric inequality. This modification helps address boundary ambiguities in sonographic imaging. Savelonas et al. [5] developed the Joint Echogenicity-Texture (JET) model, which accurately segments thyroid nodules by leveraging their echogenicity and texture features. Although contour-based methods have shown some utility in thyroid nodule segmentation, they often perform poorly when segmenting nodules with weak boundaries and are sensitive to the initial manual placement of contours, making automated segmentation of thyroid nodules challenging.

Poudel et al. [6] explored thyroid nodule segmentation using a graph-based method. In their approach, the thyroid ultrasound image is represented as a graph, and the segmentation result is determined by defining connection weights and performing a minimum cut. However, this method requires manual labeling to obtain prior information, which introduces significant subjectivity and limits the potential for fully automated segmentation. Alrubaidi et al. [7] devised a method for thyroid nodule segmentation leveraging the Variance Reduction Statistic (VRS). This method first identifies the region of interest (ROI) containing the nodule, then calculates the variance of each pixel within the ROI using VRS, which serves as the segmentation criterion. An appropriate threshold is then selected to separate pixels with significantly different variances, thereby segmenting the nodule region. The selection of this threshold requires the involvement of medical experts, making fully automated segmentation challenging.

The proliferation of machine learning techniques, particularly deep learning architectures, has established deep learning-based medical image segmentation as a predominant research focus. Chang et al. [8] implemented a decision tree (DT) model for thyroid nodule segmentation. This method can automatically learn rules and conditions from image features to segment nodules and background regions. However, this method is sensitive to noise and incompleteness in the data, which limits the model's generalization ability. With the rise of deep learning, many researchers have proposed deep learning-based semantic segmentation methods for thyroid nodule segmentation. Ying et al. [9] proposed

Manuscript received September 9, 2024; revised April 12, 2025.

This work is supported by the National Natural Science Foundation of China (No. 82360347) and Jiangxi Provincial Natural Science Foundation of China(No. 20224BAB216079).

Rong Hu is an Engineer of Space Star Technology Co., Ltd, Beijing 100080, China. (e-mail: 475790894@qq.com)

Hao Wang is a Postgraduate of School of Mathematics and Computer Sciences, Nanchang University, Nanchang 330031, Jiangxi, China. (e-mail: 1442529991@qq.com)

Suzhang Zhang* is a Postgraduate of School of Mathematics and Computer Sciences, Nanchang University, Nanchang 330031, Jiangxi, China. (corresponding author to provide e-mail: zhangsuzhang@email.ncu.edu.cn)

Weiping Zhang is a Lecturer of School of Mathematics and Computer Sciences, Nanchang University, Nanchang 330031, Jiangxi, China. (e-mail: zhangweiping@ncu.edu.cn)

Pan Xu is a Doctor of First Affiliated Hospital of Nanchang University, Nanchang 330006, Jiangxi, China. (e-mail: xupan 1989@126.com)

a cascade convolutional neural network-based method for thyroid nodule segmentation. This network first locates the region of interest (ROI), generates candidate nodule regions, and then manually labels these regions. Finally, CNN is used for more refined segmentation of the nodules within the ROI. However, this method requires physician involvement and has difficulty in accurately segmenting nodules with complex edge details. Ding et al. [10] introduced residual connections into the U-Net architecture, improving the model's performance in nodule segmentation. However, this network performs poorly on low-contrast thyroid ultrasound images and cannot handle multi-nodule segmentation. Chen et al. [11] incorporated an attention mechanism into the semantic segmentation model, generating attention weights for feature maps to enhance the network's sensitivity to lesion regions and improve segmentation accuracy. Wang et al. [12] integrated residual structures and multi-scale convolutions into the encoder path of U-Net and added an attention module to the long-range skip connections, preserving edge contours in the feature tensors and achieving more accurate nodule segmentation.

In summary, while contour-based and region-based segmentation methods have made some progress in addressing thyroid nodule segmentation in ultrasound images, they still require manual intervention, leading to certain subjectivity and limitations. In contrast, machine learning and deep learning-based methods have achieved significant advancements in automation and segmentation accuracy. However, due to the unclear contours, varying sizes, and frequent fusion with surrounding tissues of thyroid nodules in ultrasound images, existing deep learning-based methods for thyroid nodule segmentation often have a single receptive field scale for feature extraction, which hinders the effective capture of multi-scale features. Although encoder-decoder-based semantic segmentation models improve resolution and mitigate the gradient vanishing problem through hierarchical connection between the decoder and encoder via skip connections, they still fail to fully capture multi-scale features at different levels in the decoder. This limitation restricts their ability to comprehensively capture edge features at various scales.

Based on the aforementioned research findings and challenges, U-Net is selected as the baseline model following comparative experiments. To optimize semantic feature extraction in thyroid nodule ultrasound imaging and improve segmentation performance, enhancements are made to the encoder and skip connection components of the U-Net model. Additionally, a strategy is proposed to incorporate auxiliary branches for providing extra supervision at different levels of the decoder. This approach reduces model overfitting while simultaneously guiding the network to learn feature representations more effectively at multiple levels, thereby improving its ability to segment nodules. Furthermore, a joint loss function is designed by linearly combining the Focal loss function and the Dice loss function, which is integrated with the auxiliary branches to jointly supervise model parameter updates.

II. U-Net Architecture

The U-Net architecture, as shown in Fig. 1, is a fully convolutional neural network based on an encoder-decoder structure. It consists of an encoder and a decoder connected by skip connections. In the encoder, each feature map undergoes two 3×3 convolutional operations for feature extraction, followed by a non-linear transformation using the ReLU activation function. The resulting feature map is then passed to the corresponding decoder layer via a skip connection. Additionally, max pooling is applied between each encoder layer to reduce the size of the feature map and extract higher-level features. The decoder serves to progressively reconstruct feature representations toward original spatial dimensions throughout the hierarchical architecture. This restoration process utilizes transposed convolutional operations for feature map upsampling. Unlike traditional encoder-decoder networks, the U-Net architecture implements cross-hierarchical skip connections during decoding. At each stage, encoder-derived features are concatenated with upsampled features through these connections, enabling hierarchical integration of global contextual patterns and localized structural details. Finally, a 1×1 convolutional layer coupled with a sigmoid function generates the probabilistic segmentation output.

As shown in the network architecture diagram, the structure of the network is relatively simple. In the encoder, the network's limited feature extraction capability stems from the use of a single-size convolution operation, making it difficult to capture multi-scale information of nodules. Additionally, some thyroid nodules have complex edge structures, and their boundaries often merge with surrounding tissues. Although U-Net improves resolution and mitigates the gradient vanishing problem through skip connections, the decoder fails to fully encode multi-scale semantic information at different levels. This limitation prevents the model from comprehensively capturing boundary features at various levels, thus affecting the nodule segmentation performance.

Therefore, to achieve accurate segmentation of thyroid nodules, this paper takes U-Net as the basic structure and proposes a variety of optimization strategies to address the common issues faced by this network and existing segmentation methods. These strategies aim to better tackle the difficulties and challenges in thyroid nodule segmentation.

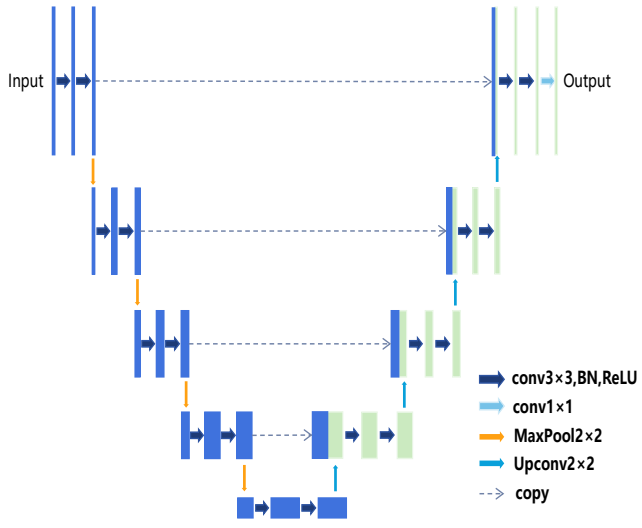

Fig. 1. U-Net architecture

III. Methods

A. U-net encoder with residual convolutional blocks and multiscale attention modules

Comprehensive and effective extraction of semantic features from thyroid nodules in ultrasound images is crucial for improving segmentation performance. Due to the irregular shapes of thyroid nodules and their often fused boundaries with surrounding tissues, an encoder that can fully capture morphological and environmental features of nodules is essential. However, the traditional U-Net encoder struggles to effectively extract the features, necessitating further optimization. To address this, a residual network is introduced to enhance the model's feature extraction capability. Additionally, given the significant variation in thyroid nodule sizes, a feature extraction network based on a single receptive field can not adequately capture multi-scale information, leading to suboptimal segmentation results. Therefore, inspired by [13], we propose a multi-scale attention module to enhance the model's feature extraction capabilities and improve segmentation accuracy for nodules of varying sizes. The encoder structure combining residual blocks and multi-scale attention is shown in Fig. 2.

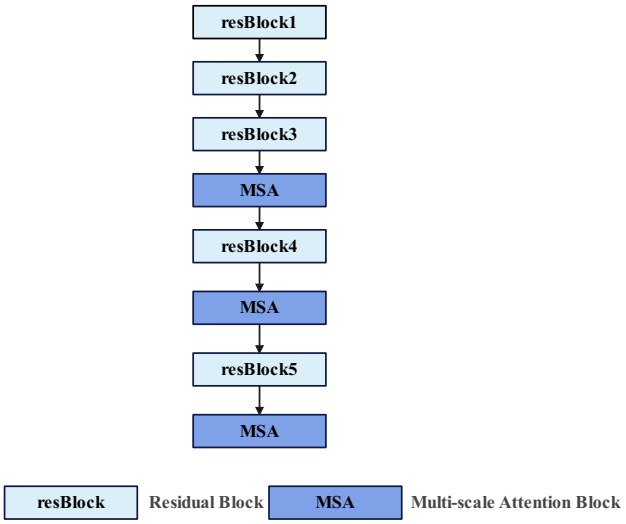

Fig. 2. Network encoder with residual blocks and multi-scale attention

1) Residual convolutional block: In the residual network structure, given an input x where the learned residual mapping is represented as F(x), such that the final output becomes F(x) + x. Therefore, the network can adaptively deactivate the residual component F(x) when it adversely affects model performance, enabling the network to continually learn additional residual information at deeper levels and improve performance. In this work, we replace the original encoding layers with residual blocks, adopting ResNet-50 as the backbone architecture. Five cascaded residual blocks are implemented to extract discriminative features from thyroid nodule ultrasound images.

2) Multi-scale attention module: By replacing the encoding layers of the original U-Net with residual convolutional blocks, the feature extraction capability of the network's encoder is enhanced. However, due to the considerable variation in the sizes of nodules in thyroid ultrasound images, the U-Net segmentation network's single receptive field scale is insufficient for effectively capturing multi-scale features. Furthermore, when aggregating multi-scale information, methods such as simple concatenation or element-wise summation are often used, which overlook the correlation and differing importance of features from various receptive fields. To optimize the utilization of multi-scale features, we introduce an attention-driven multi-scale module that dynamically selects discriminative features across varying receptive fields, thus enhancing segmentation performance for nodules with diverse dimensions. The proposed architecture is depicted in Fig. 3.

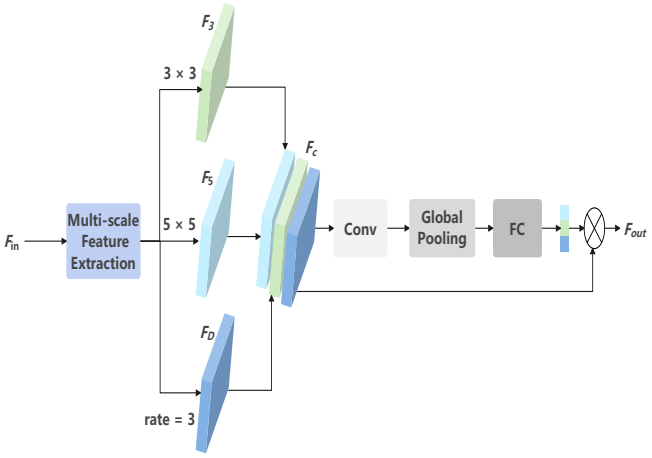

Fig. 3. Multi-scale attention module

In Fig. 3, the input image is first processed through three parallel convolutional layers to generate feature maps with three distinct receptive fields. These layers employ a 3×3 convolution, a 5×5 convolution, and a dilated convolution with a dilation rate of 3, respectively. The multi-scale feature extraction process is shown in (1), (2) and (3):

$$F^3 = conv3(F_{in}) \tag{1}$$

$$F^5 = conv5(F_{in}) \tag{2}$$

$$F^D = convD(F_{in}) \tag{3}$$

Where F_{in} denotes the input feature map, F_3 and F_5 denote the feature maps captured by the 3×3 and 5×5 convolutions, respectively. F_D denotes the feature maps captured by the dilated convolution layer. Convolution kernels of different sizes provide distinct receptive fields, allowing focus on different spatial ranges.

Concatenation of multi-scale feature maps. The feature maps from the three different receptive fields are concatenated along the channel dimension to obtain a multi-scale feature map, as shown in (4):

$$F_c = concat(F^3, F^5, F^D) \tag{4}$$

Feature aggregation. The concatenated multi-scale feature map undergoes feature aggregation to capture long-range dependencies. The multi-scale feature map is initially processed through a 1×1 convolutional layer for channel

compression and feature integration. Subsequently, the resultant feature map undergoes parallel global max-pooling and average-pooling, to model spatial long-range dependencies. This process is shown in (5) and (6):

$$F_{avg} = Avg(conv(F_c)) \tag{5}$$

$$F_{max} = Max(conv(F_c)) \tag{6}$$

Where Avg(·) denotes global average pooling and Max(·) denotes global max pooling.

Attention weight generation. The sum of F_{avg} and F_{max} is fed into a fully connected module consisting of two cascaded fully connected layers. After passing through an activation function, the multi-scale attention weights are generated. The weight generation process is shown in (7):

$$weights = \sigma(conv(\delta(conv(F_{avg} + F_{max})))) \tag{7}$$

Where $\sigma(\cdot)$ denotes the sigmoid loss function and $\delta(\cdot)$ denotes the ReLU loss function.

Finally, the generated attention weights are multiplied by the multi-scale feature vector to obtain a feature vector F_{out} with multi-scale information weights, addressing multi-scale attention challenges and enhancing the model's ability to adapt to features at varying scales. This improvement ultimately enhances segmentation performance for nodules of different sizes.

In this paper, the multi-scale attention module is incorporated only into the final three layers of the encoder to improve the model's ability to segment nodules of varying sizes while reducing the number of model parameters.

B. Multi-level feature fusion module for optimizing U-Net skip connections

Enhancing the U-Net encoder allows the model to capture nodule features more effectively, improving segmentation performance for nodules of various sizes. However, this optimized model exhibits poor performance in segmenting the edges of thyroid nodules in scenarios where the edge structures are complex and boundaries merge with adjacent tissues. Two primary factors may contribute to this limitation. The conventional U-Net architecture progressively downsamples feature maps in its encoding pathway to extract features. However, this hierarchical resolution degradation inevitably compromises critical boundary preservation during global semantic feature handling. Second, the model lacks integration and effective utilization of multi-level features, which hinders its ability to capture long-range dependencies and accurately delineate the edges of complex nodules, thereby affecting overall segmentation accuracy.

Based on the above analysis, to address this problem, first, in the feature extraction stage of the encoder, high-resolution feature maps containing rich image details can be upsampled to capture sufficient semantic information, thus enhancing the accuracy of the final segmentation results. Second, multi-level feature fusion is applied to the outputs of each encoder layer to extract the complex semantic features of nodules with intricate structures, further improving segmentation performance. To this end, we propose a multi-level feature fusion module to augment U-Net's skip connections, as depicted in Fig. 4.

Upsampling with feature maps from different levels. The outputs of each encoder layer, denoted as F_1, F_2, F_3, and F_4, are fed into the multi-level feature fusion module. Each level of features is upsampled using transposed convolution to match the size of the highest-resolution feature map in the network. Simultaneously, to reduce model complexity, the number of channels is reduced during upsampling. Notably, the feature map F1 from the first layer undergoes only channel compression without upsampling. After this process, high-resolution feature maps $F_1^{'}$, $F_2^{'}$, $F_3^{'}$, and $F_4^{'}$ of identical size are obtained.

Aggregating multi-level feature maps. These feature maps are then concatenated along the channel dimension, resulting in a preliminary aggregated multi-level feature map $F_S^{'}$, as calculated in (8):

$$F_S^{'} = concat(F_1^{'}, F_2^{'}, F_3^{'}, F_4^{'}) \tag{8}$$

Where concat(·) denotes the concatenation operation along the channel dimension.

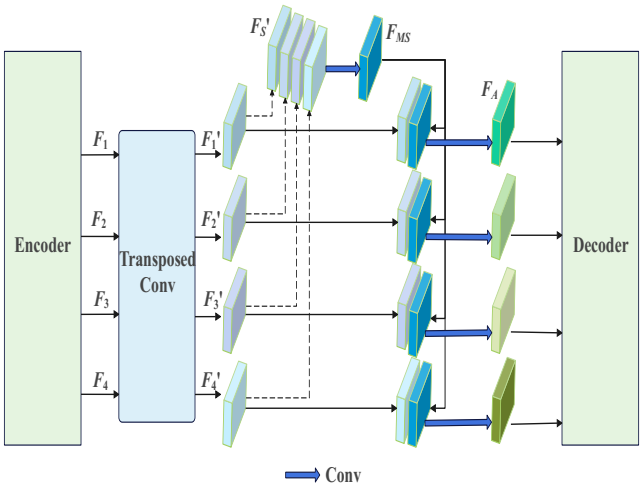

Fig. 4. Multi-level feature fusion module

Following feature concatenation, the multi-level feature map $F_S^{'}$ undergoes channel reduction through a convolutional layer to generate the fused multi-scale representation F_{MS}. This process integrates cross-scale contextual information. Through this design, F_{MS} simultaneously captures both low-level structural patterns and high-level semantic concepts acquired through distinct network stages. The generation process of F_{MS} is shown in (9):

$$F_{MS} = conv(F_S^{'}) \tag{9}$$

Upsampled feature maps of each layer fusing global and local features. To further integrate the global features obtained from the encoder with the local feature of each layer, F_{MS} is concatenated with $F_1^{'}$, $F_2^{'}$, $F_3^{'}$, and $F_4^{'}$ along the channel dimension, respectively. A convolution operation is then applied to aggregate the multi-level feature information, resulting in the fused feature map F_A, as calculated in (10):

$$F_{Ai} = conv(concat(F_i^{'}, F_{MS})) \tag{10}$$

Where F_{Ai} denotes the multi-level feature map for the ith layer and $F_i^{'}$ denotes the upsampled feature map for the ith layer.

Finally, F_A is downsampled to match the scale of the corresponding feature map in the decoder and then passed to the respective decoder layer to participate in the decoding process.

C. Deep supervision as auxiliary branches to prevent overfitting

To enhance the model's ability to segment nodules, the U-Net was optimized. However, this also increased the number of network parameters, potentially leading to degraded convergence performance and overfitting. To address this issue, a deep supervision mechanism is introduced. Deep supervision is a training strategy that introduces additional supervisory signals by adding extra supervision tasks to intermediate layers of a deep neural network [14]. In the task of segmenting thyroid nodules in ultrasound images, utilizing information from various levels of the decoder to supervise the network during training enables the segmentation network to better learn feature representations at different levels, further improving its segmentation capability. Additionally, deep supervision branches can prevent the network from encountering gradient vanishing and slow convergence during training. As illustrated in Fig. 5, deep supervision branches are added to the decoder section of the segmentation network to enhance nodule segmentation performance.

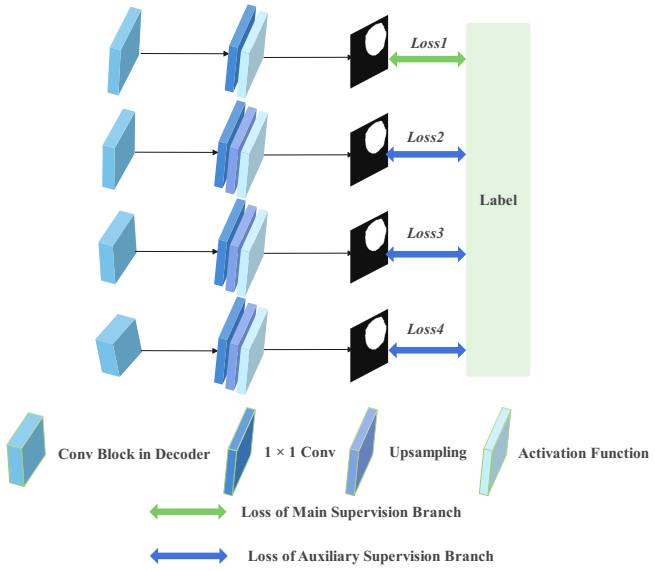

Fig. 5. Design of deep supervision branch

In Fig. 5, auxiliary branches are appended to each layer of the decoder to more fully utilize the feature representations at various levels. The decoder's output feature map first undergoes dimensionality adjustment through a 1×1 convolution, serving dual purposes of channel reduction and parameter optimization while extracting more representative feature information. Subsequently, the compressed feature map is upsampled to match the size of the ground truth label map. Next, the full-size feature map is mapped to a range between 0 and 1 via a sigmoid function, resulting in a pixel-wise binary classification prediction. Finally, the pixel-wise classification errors between these auxiliary branch predictions and the ground truth labels are calculated, and these errors are backpropagated along with the loss of the main branch output layer.

D. Joint loss function incorporating Focal loss

Loss functions quantify the discrepancy between a model's predictions and the ground truth labels, directly influencing training efficacy and model performance. In semantic segmentation tasks, a joint loss function combining cross-entropy loss and Dice loss is often employed. While cross-entropy loss primarily focuses on individual pixels, neglecting the overall structural information of the target, Dice loss places more emphasis on the overall structure. Therefore, combining the cross-entropy loss and the Dice loss allows consideration of both pixel-level accuracy and the consistency of the target structure. However, in ultrasound images of thyroid nodules, nodules typically occupy only a small portion of the image, while normal tissues dominate, resulting in a severe class imbalance problem. Cross-entropy loss assumes that all class samples are of equal importance, and thus, for imbalanced datasets like ultrasound images of thyroid nodules, using the cross-entropy loss can lead the model to focus more on the majority class (normal tissue) and ignore the minority class (nodule regions). Additionally, in nodule segmentation tasks, easily classifiable samples dominate, meaning that gradients are primarily determined by these samples. Meanwhile, pixels along nodule boundaries, which are more challenging to distinguish from surrounding tissues, require greater attention.

To better segment nodule regions, the Focal loss function is incorporated into the aforementioned joint loss function, aiming to address the class imbalance and hard example mining problems in segmentation while preserving the advantages of Dice loss.

Focal loss was originally introduced to address class imbalance in object detection tasks, with the goal of enhancing detection accuracy. Its mathematical expression is shown in (11):

$$L_{FL} = \begin{cases} -\alpha(1-P)^{\gamma}\log(P), & y=1 \\ -(1-\alpha)P^{\gamma}\log(1-P), & y=0 \end{cases} \tag{11}$$

where P represents the model's prediction, y represents the ground truth label (1 for nodule regions and 0 for other tissues), α is a balancing factor between 0 and 1 that regulates the influence of positive and negative samples, and γ is a modulating factor that adjusts the weights of easy and hard examples. When γ is greater than 0, Focal Loss assigns higher weights to hard examples, allowing the model to focus more on these samples. As verified by the experiments in this paper, the model achieves better segmentation performance when α is set to 0.35 and γ is set to 2.0. The joint loss function designed in this paper is shown in (12):

$$L = L_{FL} + L_{Dice} \tag{12}$$

With the addition of auxiliary deep supervision branches, the losses generated by these branches also contribute to the model's parameter updates. Therefore, based on the joint loss function, the final loss function is further designed, as shown in (13):

$$L_{final} = L + \eta \sum_{d \in D} L_d \tag{13}$$

Where L denotes the loss of the final layer of the decoder, L_d denotes the loss at decoder layer level d, D denotes the set of layer level indexes, and η is the depth supervision coefficient that gradually decays during training.

IV. Experimental findings and discussion

A. Experimental setup and parameterization

All experiments were executed on an Ubuntu 20.04 system utilizing an Intel Xeon Skylake processor (8 vCPUs with IBRS) paired with an NVIDIA Tesla T4 GPU (16GB VRAM). The implementation leveraged Python 3.8 with PyTorch 2.0.0 for architectural development and experimental procedures. CUDA 11.8 was utilized to leverage the GPU's parallel computing capabilities for model training. The segmentation framework employed standardized input images of 448×448 pixels, trained across 200 epochs with a batch size of 4. Network optimization utilized the Adam algorithm following an adaptive learning schedule: initialized at 0.005, reduced to 0.0005 at epoch 50 (×0.1 scale), and further decreased to 0.00005 at epoch 150 (×0.01 scale). The loss function employed was the joint loss function from Section III-D, where α is set to 0.35 and γ is set to 2.0.

B. Experimental dataset and performance evaluation metrics

The experimental dataset was sourced from real-world clinical ultrasound examinations conducted at the Department of Ultrasound in a hospital in Nanchang. It comprised 375 patients' ultrasound images of thyroid nodules. All experiments utilized a preprocessed ultrasound image dataset, which was partitioned into three distinct subsets: a training set containing 8,168 images (80% of total data), a validation set of 681 images (10%), and a test set comprising 681 images (remaining 10%). Segmentation performance was validated on the validation set every 5 epochs during training. To ensure the stability and reliability, each experiment was repeated 4 times, with identical parameter settings and datasets but different random seeds and the mean value of the results was taken as the final evaluation outcome.

Five evaluation metrics were employed for assessment: Dice similarity coefficient, pixel accuracy, sensitivity, specificity, and Jaccard index. Their definitions are as follows:

$$Dice = \frac{2 \times |P \cap G|}{|P| + |G|} \tag{14}$$

$$PA = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

$$SE = \frac{TP}{TP + FN} \tag{16}$$

$$SP = \frac{TN}{TN + FP} \tag{17}$$

$$Jaccard = \frac{|P \cap G|}{|P \cup G|} \tag{18}$$

Where P represents the model's predicted result, while G represents the ground truth. Among these, TP denotes true positive, TN denotes true negative, FP denotes false positive, and FN denotes false negative.

TABLE I
Performance of baseline networks

Model	Jaccard	PA	Dice	SE	SP
FCN	64.63	91.16	72.50	79.79	92.69
U-Net	**69.10**	**95.79**	**77.42**	86.56	96.25
DeeplabV3+	67.24	94.56	74.89	84.24	**97.16**
SegNet	67.71	95.12	76.39	**87.31**	96.04

C. Baseline network selection and performance comparison

To select a suitable baseline semantic segmentation network, four representative models-FCN, U-Net, DeepLabv3+, and SegNet—were trained and tested on the thyroid nodule ultrasound image dataset. Each model was tested under identical settings and dataset conditions across four experiments but with different random seeds, as shown in Table I. The experimental results presented are the averages of the four experiments, and all data in the table are percentages. Bolded values indicate the highest performance for each metric.

As shown in the table, U-Net demonstrated the highest segmentation performance among all models, achieving the highest Jaccard index (69.10%) and Dice coefficient (77.42%), both directly reflecting nodule segmentation accuracy. Additionally, U-Net attained a pixel accuracy of 95.79%, which is also higher than other models. However, the results indicate that U-Net's nodule segmentation performance remains suboptimal, likely due to the encoder's limited feature extraction capability, which hampers comprehensive nodule feature capture. Furthermore, U-Net's lower specificity and sensitivity scores suggest a degree of mis-segmentation, indicating areas where improvement is needed.

D. Comparative study of encoder improvements

To address the insufficient feature extraction capability of the original U-Net, we introduced residual networks to replace the convolutional blocks in the encoder. Simultaneously, a multi-scale attention (MSA) module was developed to improve segmentation performance across varying nodule dimensions. To validate the proposed enhancements, four comparative experiments were implemented using thyroid nodule ultrasound image dataset, employing the U-Net architecture as the baseline. Quantitative results are presented in Table II, with Backbone denoting the base network and MSA indicating the proposed multi-scale attention module. All values in the table are percentages, with bold font indicating the highest value for each metric.

As shown in Table II, both the residual convolutional blocks and the multi-scale attention module improved the model's ability to segment nodules. Moreover, the improvement was more significant when both modules were introduced to modify the encoder. Comparing the results of the first and second experiments, ResNet50 as the encoder achieved superior segmentation across all metrics relative to the original model. Likewise, the comparison between the first and third experiments shows that the proposed multi-scale attention module also effectively improved the segmentation performance of the model. Observing the second and fourth experiments reveals that combining the multi-scale attention module with ResNet50 provided further improvement across all metrics. These findings indicate that

TABLE II
PERFORMANCE OF IMPROVED ENCODERS

	Backbone	MSA	Jaccard	PA	Dice	SE	SP
1	×	×	69.10	95.79	77.42	86.56	77.42
2	Resnet50	×	75.54	97.49	82.78	88.98	97.35
3	×	✓	73.36	96.51	80.85	87.39	96.89
4	Resnet50	✓	**78.02**	**97.86**	**85.26**	**91.57**	**97.76**

TABLE III
PERFORMANCE OF MULTI-SCALE ATTENTION MODULES WITH DIVERSE RECEPTIVE FIELDS

	parameters	Jaccard	PA	Dice	SE	SP
1	1×1;3×3;3×3,r=2	77.13	96.25	84.42	90.16	96.25
2	3×3;5×5;3×3,r=3	**78.02**	**97.86**	**85.26**	**91.57**	**97.76**
3	5×5;7×7;3×3,r=3	77.64	96.83	84.75	90.47	96.88

using ResNet50 as the network encoder and using deeper convolutional blocks can enhance the segmentation model's feature extraction capability, improve nodule feature capture, and enable more accurate nodule region segmentation.

Furthermore, by introducing a multi-scale convolutional block and adaptively adjusting weights among feature maps with different receptive fields, the obtained multi-scale feature maps enable the network to utilize multi-scale features more effectively, enhancing segmentation accuracy for nodules of varying sizes and improving overall performance on the nodule dataset.

In this set of experiments, we explored the impact of convolutional operations with different receptive field sizes on the multi-scale attention module. To this end, three different combinations of convolutional kernels with varied receptive field sizes were evaluated. The first group consisted of smaller-sized convolutions, including 1×1 and 3×3 standard convolutions, and a 3×3 dilated convolution with a dilation rate of 2. The second group consisted of 3×3 and 5×5 standard convolutions, and a 3×3 dilated convolution with a dilation rate of 3. The third group used larger-sized convolutions, including 5×5 and 7×7 standard convolutions, and a 3×3 dilated convolution with a dilation rate of 3. Table III presents the comparative results. As shown in the table, both smaller and larger receptive fields are detrimental to nodule segmentation, indicating that the multi-branch convolutional kernel size combination selected in this paper is reasonable.

E. Comparative study of multi-level feature fusion module and deep supervision branch

To further enhance the model's ability to segment nodule boundaries, a feature fusion module was proposed based on previous improvements. Simultaneously, a deep supervision auxiliary branch was introduced to mitigate overfitting and further improve the model's segmentation performance. Through comparative experiments, the impact of multi-level feature fusion and deep supervision on the model's segmentation performance was explored. The baseline model in this experiment employed the improved encoder. Table IV presents the experimental results, where FFM indicates whether the multi-feature fusion module was added, Supervise indicates whether the deep supervision branch was added, and all data in the table are percentages. Bolded values represents the highest performance for each metric.

TABLE IV
PERFORMANCE WITH/WITHOUT MULTI-LEVEL FEATURE FUSION MODULE AND WITH/WITHOUT DEEP SUPERVISION BRANCH

	FFM	Supervise	Jaccard	PA	Dice	SE	SP
1	×	×	78.02	97.86	85.26	91.57	97.76
2	✓	×	81.09	98.06	87.49	93.02	98.18
3	×	✓	79.95	97.52	86.61	93.13	97.70
4	✓	✓	**83.61**	**98.29**	**89.05**	**94.69**	**98.41**

TABLE V
PERFORMANCE OF DIFFERENT LOSS FUNCTION COMBINATIONS

	Loss function	Jaccard	PA	Dice	SE	SP
1	CE	81.73	98.06	87.26	92.79	**98.60**
2	DSC	82.91	97.48	88.78	91.13	97.87
3	CE+DSC	82.74	98.12	88.61	93.56	98.34
4	FL	82.51	**98.37**	88.43	94.05	98.23
5	FL+DSC	**83.61**	98.29	**89.05**	**94.69**	98.41

Comparing the results of the first and second experimental groups reveals that adding only the feature fusion module improves all metrics, achieving higher pixel accuracy, sensitivity, and specificity. This indicates that the feature fusion module allows the segmentation model to more comprehensively capture the edge details of thyroid nodules in ultrasound images, resulting in more accurate segmentation. Similarly, the comparison between the first and third groups shows that introducing only the deep supervision branch also enhances performance. Notably, the model incorporating both the feature fusion module and the deep supervision branch achieves the highest performance across all metrics. In conclusion, the combined addition of the feature fusion module and deep supervision branch significantly enhances the model's segmentation capability.

F. Comparative study of different loss functions

To evaluate the effectiveness of the proposed loss function, a comparative experiment was conducted. The final improved model was used in the experiment. Quantitative comparisons are presented in Table V, with CE, DSC, and FL denoting cross-entropy loss, Dice loss, and Focal loss, respectively. All table data are presented as percentages, with the highest values in bold.

Comparing the results of the third group with those of the first and second groups, it can be observed that combining cross-entropy and Dice losses enhances the model's segmentation performance across several metrics compared to using either loss alone. Similarly, observing the second, fourth, and fifth groups shows that the combined loss function achieves superior segmentation results compared to using Dice loss or Focal loss independently.

The analysis above indicates that the combined loss function leverages the strengths of different loss functions to more effectively supervise model training. Comparing the results of the first and fourth groups, it can be found that the model

TABLE VI
PERFORMANCE OF THE DESIGNED LOSS FUNCTION WITH DIFFERENT PARAMETER SETTINGS

	α	γ	Jaccard	PA	Dice	SE	SP
1	0.25	1.5	83.22	98.15	88.56	94.07	98.26
2	0.35	1.5	83.37	98.06	88.79	94.22	98.37
3	0.45	1.5	83.09	97.72	88.61	93.92	98.30
4	0.25	2.0	83.42	98.11	88.93	94.45	**98.56**
5	0.35	2.0	**83.61**	**98.29**	**89.05**	**94.69**	98.41
6	0.45	2.0	83.23	97.88	88.76	94.27	98.51

using the Focal loss outperformed the model using the cross-entropy loss in all metrics except for specificity. Additionally, comparing the third and fifth groups, it can be found that combining Focal loss with Dice loss yielded improvements in all metrics over the combination of cross-entropy and Dice losses. These results demonstrate that training the model jointly with Focal and Dice losses enhances segmentation performance.

Next, the settings of parameters α and γ in the designed loss function were explored. Six parameter combinations were selected, and experiments were conducted with the final improved model. The results are shown in Table VI, with all values in percentages, and the highest values in bold.

Through the comparison of experimental results with different parameter settings in the above table, it can be found that when α is 0.35 and γ is 2.0, the model outperforms other parameter combinations in terms of Jaccard index, pixel accuracy, Dice similarity coefficient, and sensitivity. The results indicate that combining the Focal loss function with the Dice loss function under this parameter configuration for model training enhances the model's focus on nodule regions, thereby improving the accuracy of segmentation outcomes.

G. Comparative study of different segmentation networks

To validate the superiority of the proposed segmentation method, this section compares the nodule segmentation model constructed in this paper with other state-of-the-art segmentation models, including SA-Unet (2021) [15], Unet3+ (2020) [16], Focus U-net (2021) [17], MA-Net (2022) [18], TA-Net (2021) [19], CE-Net (2019) [20] and Mask2Former (2022) [21]. All network models are based on the "encoder-decoder" structure. SA-UNet introduces a spatial attention mechanism to enhance the network's focus on specific spatial locations. Unet3+ modifies the U-Net framework by introducing full-scale skip connections, connecting each decoder stage to multiple encoder stages to improve feature transfer and reconstruction. Focus U-net introduces short-range skip connections and deep supervision to increase feature diversity and provide additional paths for propagating losses, better updating parameters. MA-Net eliminates semantic ambiguity in skip connections through attention gate integration, while incorporating a multi-scale prediction fusion mechanism to effectively leverage global contextual information across spatial resolutions. TA-Net designs a multi-scale dilated convolution module to strengthen feature extraction capability, integrating channel and spatial attention mechanisms to improve focus on the target region. CE-Net proposes the dense atrous convolution block and residual multi-kernel pooling block to capture higher-level abstract features and retain more spatial information. This paper analyzes the segmentation performance of each method using five metrics: Jaccard, PA, Dice, SE, and SP. The experimental results are shown in Table VII, where the data is the average of four experiments, expressed as a percentage, with the bolded data indicating the highest value for each metric.

TABLE VII
PERFORMANCE OF THIS PAPER'S METHOD AND OTHER SEGMENTATION METHODS

	Model	Jaccard	PA	Dice	SE	SP
1	baseline	69.10	95.79	77.42	86.56	96.25
2	SA-UNet	74.23	96.01	81.19	87.80	96.51
3	U-Net3+	76.50	97.23	83.34	89.42	97.54
4	Focus U-net	79.38	98.08	86.31	93.17	98.27
5	MA-Net	81.26	97.95	87.30	91.42	98.22
6	TA-Net	80.58	97.63	86.97	92.83	97.83
7	CE-Net	11.21	94.58	37.10	46.31	97.53
7	Mask2Former	75.05	88.53	85.75	88.53	**98.77**
8	method in this paper	**83.61**	**98.29**	**89.05**	**94.69**	98.41

The quantitative comparisonsin the table demonstrate that our method surpasses comparative models across nearly all evaluation metrics. Notably, CE-Net exhibited substantially inferior Dice and Jaccard indices relative to other approaches; in contrast, our framework achieved statistically significant improvements, demonstrating a positive correlation between prediction accuracy and regional overlap consistency. In addition, our method outperforms other models in sensitivity while slightly underperforming Mask2Former in specificity, indicating that the proposed segmentation method based on MSA and FFM can effectively predict foreground and background while reducing erroneous segmentation. In terms of the PA metric, our method achieved 98.29%, indicating that the model is more accurate in pixel-level prediction of the overall image.

Fig. 6 shows the qualitative comparisons, illustrating segmentation results of our proposed method and other advanced models across four thyroid nodules with varied sizes and edge structures. The top row displays the original ultrasound images, the second row shows the corresponding ground truth masks, and the subsequent rows present segmentation outputs from the comparison models and our method.

By comparing the segmentation results of the proposed method with those of other models, it can be observed that the segmentation masks generated by our method are closest to the standard masks. As shown in the figure, whether facing larger or smaller nodules, our method can accurately identify the nodule region, while other methods show notable mis-segmentation. Additionally, when facing nodules with complex boundary structures (such as the second nodule), our method still significantly outperforms other models, providing more precise segmentation of edge details.

Comprehensive analysis of both the quantitative results (see Table VII) and the qualitative segmentation results (see Fig. 6) demonstrates that the proposed thyroid nodule segmentation method, incorporating multi-scale attention and feature fusion, surpasses the comparison methods in overall

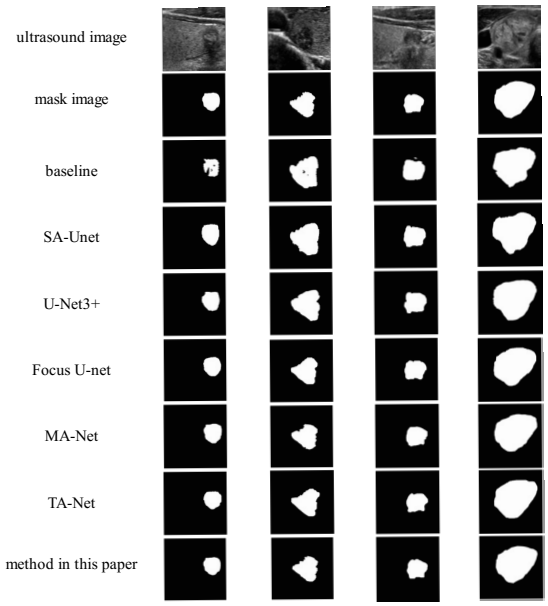

Fig. 6. Segmentation results of different models

performance.

V. Conclusion

Replacing the original network's encoding blocks with residual networks incorporating multi-scale attention modules effectively improves the encoder's feature extraction capability. By adaptively weighting feature maps with different receptive field sizes, the network can flexibly select appropriate receptive field sizes to adapt to nodules of various scales, thereby enhancing its nodule segmentation capability. The multi-level feature fusion module effectively integrates global and local information from the network's encoder, and transmits fused multi-scale global feature maps to the corresponding decoder levels. This effectively addresses the problem of difficult-to-obtain semantic information caused by the complex structure of thyroid nodules and their intermingling with surrounding tissues, enabling better capture of spatial information at nodule boundaries and more accurate segmentation of nodule regions. The strategy of using deep supervision branches at multiple decoder levels mitigates gradient vanishing and slow convergence during training, guiding the network to better learn feature representations at different levels, and further improving its nodule segmentation capability. The combined loss function incorporating the Focal loss function can better supervise model training, enhance the model's focus on nodule regions, and help to solve the class imbalance problem. The optimized U-net model achieves excellent results on real datasets in terms of Jaccard index, Dice coefficient, pixel accuracy (PA), sensitivity (SE), and specificity (SP) when compared with state-of-the-art methods. Compared with current mainstream semantic segmentation models, the optimized U-net model can more accurately segment the contours of thyroid nodules and accurately locate the position of nodules.

Future work includes: (1) Validation on multiple datasets to evaluate the model's performance under different data sources and features, thereby enhancing its generalization ability and stability. (2) Focusing on lightweight model design to reduce complexity and parameter count, improving the model's applicability and efficiency in resource-constrained environments. (3) Research on semi-supervised segmentation methods to leverage large amounts of unlabeled data alongside limited labeled data, addressing the data annotation challenges in thyroid ultrasound image segmentation.

References

[1] V. Kumar, J. Webb, A. Gregory, D. D. Meixner, J. M. Knudsen, M. Callstrom, M. Fatemi, and A. Alizad, "Automated segmentation of thyroid nodule, gland, and cystic components from ultrasound images using deep learning," *IEEE Access*, vol. 8, pp. 63 482–63 496, 2020.

[2] D. E. Maroulis, M. A. Savelonas, D. K. Iakovidis, S. A. Karkanis, and N. Dimitropoulos, "Variable background active contour model for computer-aided delineation of nodules in thyroid ultrasound images," *IEEE Transactions on Information Technology in Biomedicine*, vol. 11, no. 5, pp. 537–543, 2007.

[3] D. K. Iakovidis, M. A. Savelonas, S. A. Karkanis, and D. E. Maroulis, "A genetically optimized level set approach to segmentation of thyroid ultrasound images," *Applied Intelligence*, vol. 27, pp. 193–203, 2007.

[4] L. Gui, C. Li, and X. Yang, "Medical image segmentation based on level set and isoperimetric constraint," *Physica Medica*, vol. 42, pp. 162–173, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1120179717304489

[5] M. A. Savelonas, D. K. Iakovidis, I. Legakis, and D. Maroulis, "Active contours guided by echogenicity and texture for delineation of thyroid nodules in ultrasound images," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 4, pp. 519–527, 2009.

[6] P. Poudel, A. Illanes, D. Sheet, and M. Friebe, "Evaluation of commonly used algorithms for thyroid ultrasound images segmentation and improvement using machine learning approaches," *Journal of Healthcare Engineering*, vol. 2018, no. 1, p. 8087624, 2018. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1155/2018/8087624

[7] W. M. H. Alrubaidi, B. Peng, Y. Yang, and Q. Chen, "An interactive segmentation algorithm for thyroid nodules in ultrasound images," in *Intelligent Computing Methodologies*, D.-S. Huang, K. Han, and A. Hussain, Eds. Cham: Springer International Publishing, 2016, pp. 107–115.

[8] C.-Y. Chang, H.-C. Huang, and S.-J. Chen, "Thyroid nodule segmentation and component analysis in ultrasound images," *Biomedical Engineering Applications Basis and Communications*, vol. 22, 04 2010.

[9] X. Ying, Z. Yu, R. Yu, X. Li, M. Yu, M. Zhao, and K. Liu, "Thyroid nodule segmentation in ultrasound images based on cascaded convolutional neural network," in *Neural Information Processing*, L. Cheng, A. C. S. Leung, and S. Ozawa, Eds. Cham: Springer International Publishing, 2018, pp. 373–384.

[10] J. Ding, Z. Huang, M. Shi, and C. Ning, "Automatic thyroid ultrasound image segmentation based on u-shaped network," in *2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 2019, pp. 1–5.

[11] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3640–3649, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:206594196

[12] B. WANG, M. LI, and X. LIU, "Ultrasound image segmentation method of thyroid nodules based on the improved u-net network," *Journal of Electronics and Information Technology*, vol. 44, pp. 514–522, 2022.

[13] S. Zhou, D. Nie, E. Adeli, Q. Wei, X. Ren, X. Liu, E. Zhu, J. Yin, Q. Wang, and D. Shen, "Semantic instance segmentation with discriminative deep supervision for medical images," *Medical Image Analysis*, vol. 82, p. 102626, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841522002547

[14] X. Hu, T. Qiu, Y. Liao, J. Wang, and L. Lin, "A dau-net-convlstm model for daytime sea fog segmentation," *IAENG International Journal of Computer Science*, vol. 51, pp. 1035–1041, 2024.

[15] C. Guo, M. Szemenyei, Y. Yi, W. Wang, B. Chen, and C. Fan, "Sa-unet: Spatial attention u-net for retinal vessel segmentation," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 1236–1242.

[16] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "Unet 3+: A full-scale connected unet for medical image segmentation," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 1055–1059.

[17] M. Yeung, E. Sala, C.-B. Schönlieb, and L. Rundo, "Focus u-net: A novel dual attention-gated cnn for polyp segmentation during colonoscopy," *Computers in Biology and Medicine*, vol. 137, p. 104815, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482521006090

[18] Y. Cai and Y. Wang, "MA-Unet: an improved version of Unet based on multi-scale and attention mechanism for medical image segmentation," in *Third International Conference on Electronics and Communication; Network and Computer Technology (ECNCT 2021)*, M. K. Mohiddin, S. Chen, and S. F. EL-Zoghdy, Eds., vol. 12167, International Society for Optics and Photonics. SPIE, 2022, p. 121670X. [Online]. Available: https://doi.org/10.1117/12.2628519

[19] S. Pang, A. Du, M. A. Orgun, Y. Wang, and Z. Yu, "Tumor attention networks: Better feature selection, better tumor segmentation," *Neural Networks*, vol. 140, pp. 203–222, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608021000861

[20] 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.

[21] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 1280–1289.