MEDU-Net+: a novel improved U-Net based on multi-scale encoder-decoder for medical image segmentation

Zhenzhen Yang[1,2*], Xue Sun[1], Yongpeng Yang[1,3*], and Xinyi Wu[2]
[1]Key Laboratory of Ministry of Education in Broadband Wireless Communication and Sensor Network Technology, Nanjing University of Posts and Telecommunications, Nanjing, 210023, China.
[e-mail: yangzz@njupt.edu.cn]
[2]College of Science, Nanjing University of Posts and Telecommunications, Nanjing, 210023, China.
[3]School of Network and Communication, Nanjing Vocational College of Information Technology, Nanjing, 210023, China.
[e-mail:yangyp@njcit.cn]
[*]Corresponding author: Zhenzhen Yang, Yongpeng Yang

Received July 4, 2023; revised November 16, 2023; accepted July 2, 2024; published July 31, 2024

Abstract

The unique U-shaped structure of U-Net network makes it achieve good performance in image segmentation. This network is a lightweight network with a small number of parameters for small image segmentation datasets. However, when the medical image to be segmented contains a lot of detailed information, the segmentation results cannot fully meet the actual requirements. In order to achieve higher accuracy of medical image segmentation, a novel improved U-Net network architecture called multi-scale encoder-decoder U-Net+ (MEDU-Net+) is proposed in this paper. We design the GoogLeNet for achieving more information at the encoder of the proposed MEDU-Net+, and present the multi-scale feature extraction for fusing semantic information of different scales in the encoder and decoder. Meanwhile, we also introduce the layer-by-layer skip connection to connect the information of each layer, so that there is no need to encode the last layer and return the information. The proposed MEDU-Net+ divides the unknown depth network into each part of deconvolution layer to replace the direct connection of the encoder and decoder in U-Net. In addition, a new combined loss function is proposed to extract more edge information by combining the advantages of the generalized dice and the focal loss functions. Finally, we validate our proposed MEDU-Net+ and other classic medical image segmentation networks on three medical image datasets. The experimental results show that our proposed MEDU-Net+ has prominent superior performance compared with other medical image segmentation networks.

Keywords: Medical image segmentation, Deep learning, U-Net, Multi-scale feature fusion, Skip connection

[1]This work is sponsored by the National Natural Science Foundation of China (Nos.61501251, 62071242), the Open Research Fund of Key Lab of Broadband Wireless Communication and Sensor Network Technology (No.JZNY202113), the Postgraduate Research & Practice Innovation Program of Jiangsu Province (Nos.KYCX22_0955, SJCX23_0251, KYCX24_1125, SJCX24_0279), and the NUPTSF (No.NY220207).

https://doi.org/10.3837/tiis.2024.07.001 **ISSN : 1976-7277**

1. Introduction

Traditional image segmentation methods are mainly based on human intervention. These methods believe that the image can be divided to different regions according to features such as grayscale, texture and color, and there are different features between regions while the same features within regions [1]. It has been demonstrated that these traditional methods are difficult to process complex medical images efficiently and conveniently.

The image segmentation methods based on deep learning (DL) [2,3] have been attracted wide attention and many efficient, convenient and simple DL-based methods have been proposed gradually. Most DL-based image segmentation methods require a large number of images for processing, however, the number of medical images that can be used for training and testing is always limited, and this problem has become a big challenge for DL-based medical image segmentation methods [4]. These DL-based methods use a certain depth of network architecture to extract rich semantic information from images. For example, Jonathan et al. [5] proposed fully convolutional network (FCN) for image semantic segmentation. It changed the visual geometry group (VGG) mechanism [6] and some other networks, added the fully connected layers to its architecture, and introduced the convolution structure for image semantic segmentation. Lots of experiments have indicated that FCN can achieve good image semantic segmentation performance. And then, a U-shaped architecture with complementary the encoder and decoder and a feature fusion with splicing was firstly proposed in U-Net network [7]. U-net is one of the most popular FCN models and it is widely used in medical image processing [3]. Thanks to the data augmentation [8], U-Net only needs few annotated images to get impressive performance and has a reasonable training time. It is worth mentioning that except for data augmentation, transfer learning [9-11] can also generate images to effectively alleviate the negative effects of small datasets on medical image processing tasks, and improve the generalization ability and performance of the models. For example, Zhang et al. [9] placed a particular focus on the field of deep learning, including the utilization of convolutional neural networks, transfer learning, and semi-supervised learning. Lu et al. [10] utilized transfer learning to obtain the image level representation based on the backbone deep convolutional neural network. Lu et al. [11] proposed a novel abnormal brain detection algorithm based on improved AlexNet and ELM optimized by chaotic bat algorithm to obtain better generalization ability. However, when the image to be segmented contains a lot of detailed information, the segmentation results cannot fully meet the actual requirements, and it also has its inherent disadvantages. For example, U-Net only uses 3×3 convolution for encoding and 3×3 deconvolution for decoding, the obtained feature information is less and the information cannot be fully utilized. Moreover, only a small part of the extracted feature information of the encoder can be provided by employing the skip connection between the single corresponding layers, which wastes the rest of the feature information. In addition, the object is biased to the background by adopting the cross-entropy loss function, which may lead to poor segmentation results.

Aiming at some existing shortcomings of U-Net, researchers have put forward some improved U-Net network methods. For example, 3D U-Net [12] was proposed for solving the limitation of the medical image segmentation only operating on 2D images. And V-Net network [13] was proposed for volumetric medical image segmentation. To reduce the semantic loss of the pooling of U-Net network, Zhou et al. [14] used the convolution to replace the pooling operation, and proposed a nested U-Net (U-Net++) network. This network pays more attention to the skip connection and uses the integration of U-Net of different depths to

balance the unknown network depth and the resulting excessive parameters. Gu et al. [15] proposed the context encoder network (CE-Net) which combined a residual module similar to ResNet [16] with dilated convolution to obtain more comprehensive and detailed semantic information. In addition, some other improved U-Net networks [17-22] are proposed for improving image segmentation performance by changing the basic convolution block of the encoder of U-Net. For example, MultiResUNet [18] proposed a combination of a multi residual (MultiRes) module and U-Net, replacing the convolutional module in the U-Net with the MultiRes module to complete the function of collecting image feature information. MDU-Net [19] not only utilized the multi-scale feature fusion to collect more semantic information, but also improved network segmentation accuracy by establishing residual connections. MA-Unet [20] included common modules such as encoder, decoder, and skip connection, and introduced the attention mechanism in the decoder to improve the performance of the model in image segmentation tasks. DIU-Net [21] integrated the inception module and the dense connections into the U-Net architecture. It contains the analysis path and the synthesis path, and these two paths mainly consist of the inception-res, the dense-inception, the down-sample and the up-sample four kinds of blocks. MECAU-Net [22] reduced the required parameters for segmentation by utilizing multi-scale even convolutions to reduce additional computational overhead, and then utilized the convolutional attention module to extract richer information without adding additional computational complexity.

Although these above-mentioned improved U-Net networks have achieved good results, they only improve some modules of U-Net network, do not consider its overall network architecture and improve each module of the whole framework according to its own shortcomings. Besides, these networks are always more concerned with the internal information of images, but the research of medical images pays more attention to the edge part and detailed information of images, which also brings new challenges. To further improve the U-Net network performance and achieve higher segmentation accuracy, we propose a novel improved U-Net called multi-scale encoder-decoder U-Net+ (MEDU-Net+) network. This proposed network focuses on not only the changes of the convolution block at the encoder, but also the process of restoring the semantic information. We design the GoogLeNet [23,24] for achieving more information at the encoder of the proposed MEDU-Net+, and present the multi-scale feature extraction for fusing semantic information of different scales. At the same time, we introduce the single skip connection to connect the information of each layer, so that there is no need to encode the last layer and then return the information, ensure that the feature information extracted from each layer of the encoder is not wasted. In addition, we use the multi-scale technology combined with GoogLeNet to improve the decoder of our proposed MEDU-Net+. Finally, we propose a new combined loss function to extract more edge information by combining the advantages of the generalized dice and the focal loss functions, and get better performance without increasing too many parameters. The innovations of our proposed MEDU-Net+ network are given as follows:

(1)A multi-scale encoding method combined with GoogLeNet is proposed.

To solve the problem that the encoder of U-Net only uses 3×3 convolution to obtain less feature information, a multi-scale encoding method combined with GoogLeNet is proposed. The main purpose is to aggregate feature maps from different branches of kernels of different sizes, which can make the network wider and capable of learning more features. The proposed encoding method uses convolutions of different scales to extract the output of the previous layer, and then aggregates several extracted feature information through different convolution layers. In addition, GoogLeNet is used to reduce the dimensionality of each convolution operation through 1×1 convolution to reduce the computational complexity. Through the

proposed encoding method, we can get more comprehensive feature information, and achieve better segmentation performance.

(2)A layer-by-layer skip connection is proposed.

Only establishing a path for transmitting information at the corresponding layers of the encoding and decoding through skip connections can result in a series of problems. For example, due to the large number of layers, transmission span, and semantic differences of feature information, the effectiveness of information fusion on both parts can be reduced. In addition, skip connection only establishes connection at the corresponding layer, which also causes the waste of information of other layers. In order to solve the problem that the skip connection between a single corresponding layer of U-Net can only provide a part of the extracted feature information by the encoding layer during decoding, and wastes the rest of the feature information. We propose to use deconvolution to get the result after each encoding step, and add the layer-by-layer skip connection to return to the previous layer, and connect the returned information after each encoding with the final decoding, so as to preserve each information which generated by the encoding layer. In this way, all the information obtained by the encoding layer can be completely transmitted back to the decoding layer to retain all the extracted feature information so as to obtain better segmentation results.

(3)A multi-scale decoding method is proposed.

Although as much information as possible has been reserved through the proposed layer-by-layer skip connection and the multi-scale encoding combined with GoogLeNet, these information cannot be fully utilized in decoding. Inspired by the improvement of the encoder of our proposed network, we adopt the multi-scale method to improve the decoder. Referring to the idea of multi-scale encoding, we carry out similar operations on the decoder, which can restore the features more completely. The proposed multi-scale decoding is presented in 3×3 on the basis of deconvolution, parallel addition of 1×1 deconvolution and 5×5 deconvolution branches. Through 1×1, the deconvolution operation greatly enhances the nonlinear features while ensuring the invariance of feature information, helping the network achieve better segmentation performance. By making full use of multi-scale decoding and the layer-by-layer skip connection, all feature information is extracted and retained to obtain better segmentation images.

(4)A new combined loss function is proposed.

The cross-entropy loss function of U-Net is a classic loss function that has achieved good results in most image segmentation. However, when the number of current object pixels is much smaller than the number of background pixels, this loss function causes the object to be severely biased toward the background, and leads to poor segmentation results. The generalized dice loss function can solve the problem of imbalanced positive and negative samples and can improve the segmentation performance of small objects. The Focal loss function is suitable for solving the problem of sample imbalance in medical images where the background pixels are much larger than the object pixels. Based on the advantages of these two loss functions, a new combined loss function is proposed by combining the generalized dice and focal loss functions. This combined loss function solves the problem that the segmentation result is seriously biased to the background when the object pixels are less than the background pixels in medical image, and makes small object medical images get better segmentation performance.

The rest structure of the paper is organized as follows. The second section introduces some related works of this paper, and the third section presents our proposed MEDU-Net+ network. In the fourth section, our proposed network is scattered for ablation experiments and

comparative experiments on three typical datasets to demonstrate its superior performance. Finally, the fifth section gives the concluding remarks.

2. Related Works

In recent years, U-Net and its improved networks have attracted much attention in image segmentation. Since U-Net has the architecture such as skip connection and unique U-shape structure, it can get more detailed image information according to the combination of deep and shallow features of the image, and basically extract some related features from the image and obtain accurate image segmentation results.

2.1 The U-Net network

It was generally believed that the successful training of DL depends on a huge number of labeled image samples. However, the emergence of U-Net network provides a more effective method to use available image samples. U-net can accurately capture the feature information in the available images through skip connection and unique U-shaped symmetry structure. This special structure makes it possible to produce more accurate segmentation results by processing a small number of training samples.

The U-Net network [7] has a unique U-shaped symmetrical structure, which includes mainly three parts: down-sampling, skip connection and up-sampling. Firstly, this network is divided into left and right parts to analyze. The left part is the encoder, which reduces the image size and extracts some simple features through convolution and down-sampling. The right part is the decoder, which obtains some depth features through convolution and up-sampling. In the middle, the obtained feature maps in encoding and decoding are combined by concatenation, and the image is refined by combining the deep and shallow features, and the prediction and segmentation are carried out according to the obtained feature maps. It is noteworthy that the feature maps sizes of the two layers need to be cut since they are different. And the last layer is classified by the 1×1 convolution.

2.2 The GoogLeNet network

VGG and some other DL networks have achieved good training performance by increasing their network depth, but the increase of layers leads to many negative impacts, such as over fitting, gradient disappearance, and gradient explosion. GoogLeNet [23,24] was proposed to improve the training results from another perspective, which can use computing resources more effectively and extract more features under the same amount of computing.

The GoogLeNet proposed the concept of an inception module to construct a sparse and high performance computing network architecture. The main advantage of this module is that it improves the utilization rate of the computing resources by increasing the depth and width of the network, while maintaining the computational budget constant by using the inception module. The GoogLeNet introduces the inception module, which uses 1×1 convolution to upgrade and reduce dimensions, and simultaneously performs convolution and aggregation on multiple dimensions. A large number of 1×1 convolution kernels are added before 3×3 and 5×5 convolution kernels in the inception module. This structure can aggregate feature maps from different branches of kernels of different sizes, which can make the network wider and capable of learning more features. It can extract more rich features of different scales and enhance segmentation performance since the inception module perform convolution reassembly on multiple sizes. Meanwhile, it can also reduce the dimension and the

computational complexity. Based on these advantages, we introduce the GoogLeNet to the encoder of our proposed MEDU-Net+ network.

2.3 The multi-scale feature fusion network

Convolutional neural network (CNN) [25,26] has been widely used in image segmentation. It abstractly extracts image features through a layer-by-layer convolution. The image features extracted by deep convolution have a large receptive field and strong semantic information expression ability, but the resolution of the feature map is low and almost does not contain spatial information. Shallow image features have a stronger ability to express spatial information and some details due to their relative small receptive field, but contain less abstract information. In the process of image segmentation, both deep and shallow features are very important. In order to make full use of different scales of information, we use the multi-scale feature fusion network to add all these different scale features together for enhancing the expression ability of image features and improve the segmentation performance of the network.

The multi-scale feature network which can improve image segmentation accuracy includes the multi-scale input, the multi-scale feature prediction, and the multi-scale feature fusion. The multi-scale input is also known as spatial pyramid pooling. It operates on the input images to get the images resolution of different sizes and input them together. The multi-scale feature prediction is mainly to predict and output different features, and then the output results are fused to get a final output. This method is widely used in object detection such as the single shot multibox detector (SSD) network [27]. The multi-scale feature fusion network is widely used in image segmentation and object detection. It mainly includes the parallel multi-branch and the serial layer skip connection structures, and their structures are given as **Fig. 1**. Both of these two structures extract the features under different receptive fields. The parallel multi-branch structure can obtain the characteristics of different receptive fields at the same level, and then transfer them to the next layer after fusion, which can balance the amount of calculation and model capabilities more flexibly. This structure uses different sizes of convolution kernels to extract the features of receptive fields with different sizes and concentrate different results from different channels. The serial structure merges the features of different abstract levels, which is indispensable for boundary-sensitive image segmentation. U-Net [7] is a typical structure of serial multi-scale feature. In order to achieve feature combination, they need to be connected by layer skip connections. Notably, our proposed MEDU-Net+ network adopts the parallel multi-branch structure for the multi-scale feature fusion.

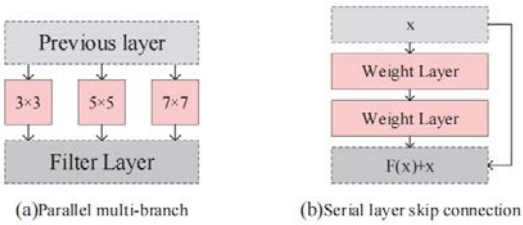

Fig. 1. The multi-scale feature fusion.
(a) is the parallel multi-branch structure; (b) is the serial layer skip connection structure

2.4 Skip connection

In general, the significant depth of the network can be useful for the image segmentation and get better performance. However, the deep network is difficult to train and always brings

gradient disappearance and network degradation. The recent researches have shown that these problems can be addressed with the skip connection during the processing of training. And the successful application of the skip connection in U-Net proves that it can recover the lost spatial information effectively. However, U-Net only establishes connections between the corresponding layers of the encoder and decoder, and always ignores the relationship between other layers, which leads to loss of the shallow information. It is worth mentioning that we design a new skip connection structure in the proposed MEDU-Net+, which uses a layer-by-layer returning method to obtain enough information for image segmentation.

3. The proposed MEDU-Net+ network

We introduce the motivations, the multi-scale encoder, the layer-by-layer skip connection, the multi-scale decoder, the overall architecture, and the loss function in the following subsection.

3.1 Motivations

If the accuracy of medical image segmentation cannot meet the actual requirements, it will lead to worse user experience in the clinical environment. Therefore, medical image segmentation needs higher accuracy compared with other image segmentation tasks. Since it is difficult to obtain appropriate medical image, the number of medical image datasets is very limited. The feature information of the image is very important, which requires us to extract and use all the feature information as much as possible in the segmentation process. Although U-Net has obtained some good results, its segmentation effect cannot completely meet the needs of medical image with a lot of detailed information. For example, only a single scale convolution kernel is used in the encoder, that is, the convolution operation of each layer from top to bottom can only obtain fixed information within the image, and the extracted image features are limited. And then, when a skip connection is created between encoding and decoding to transfer information, the connection between single corresponding layers aggravates waste of the image feature information. In addition, the decoding of the single deconvolution operation makes the successful transfer of image feature information cannot be used reasonably. Therefore, it is difficult to make full use of the image feature information due to the incomplete information extracted by the U-Net network encoder, the waste in the process of transferring information by skip connection, and the poor information recovery method of the decoder. Therefore, we propose to improve the encoder, the skip connection and the decoder of U-Net. Finally, the cross-entropy loss of U-Net has achieved good performance in most image semantic segmentation. However, when the number of object pixels only accounts for a small part of the whole image pixels, this loss function will lead to serious bias of the foreground to the background. Hence, a new loss function is proposed to solve the problem that the segmentation result is seriously biased to the background when the object pixels are less than the background pixels in medical image.

3.2 Multi-scale encoder

We usually want to segment medical images of cells, blood vessels and tumors with different shapes and sizes for segmentation task, and cannot only pay attention to the feature information of a certain fixed scale in the process of collecting information. To get a better segmentation image, the medical image segmentation network should analyze different scales objects as much as possible. Although the traditional multi-scale encoding can collect the changes of different scales, it also brings more parameters. Inspired by the GoogLeNet, a novel

multi-scale encoder is proposed which integrates the inception module into the U-Net encoder. It uses the inception structure of GoogLeNet network to add the 1×1 convolution to the multi-scale encoder in this paper. By convoluting multi-feature information with 1×1, not only more image features can be collected, but also dimension reduction can be reduced which alleviates the huge amount of parameters.

Our proposed MEDU-Net+ network introduces the GoogLeNet convolution to the basic convolution block of the encoder, and uses the 1×1 convolution in each branch. The improved convolution block of the multi-scale encoder is given as **Fig. 2**. It consists of input block, three convolution with different kernels, the combination of pooling operation and convolution, concatenation and output block.

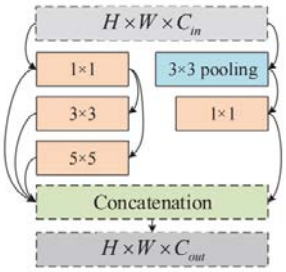

Fig. 2. The convolution block of the multi-scale encoder. It consists of input block, three convolution with different kernels, the combination of pooling operation and convolution, concatenation and output block.

The multi-scale encoder with the inception structure of GoogLeNet network can get the effect of the information interaction between channels by reducing the dimensionality. In addition, the proposed MD-UNet+ network adds convolution kernels of different sizes to each branch which can expand the receptive field and extract richer semantic information.

3.3 Layer-by-layer skip connection

The skip connection can improve the image segmentation performance by using the horizontal connection of fused image features. The U-shape structure in U-Net is directly used in encoder and decoder, and a channel for information transmission is added between them by using the skip connection.

However, only establishing a path for both encoder and decoder can only transfer the information of the corresponding layer, and the differences between the information reduce the effect of information fusion since the transmission span is too large. To reduce the differences in the connection process and further improve the segmentation results, we design a new skip connection, which establishes a more suitable path for transmitting information between encoding and decoding through the basic steps of convolution, deconvolution, and connection. This layer-by-layer skip connection can be seen as a cascade of multiple U-shaped networks, our proposed MEDU-Net+ achieves the goal of collecting and transferring semantic information of all encoding layers by using this transfer form. The process of up reverse decoding is added after each encoding, and the features with similar semantics and small span are fused together to get more image feature information. The improved connection is shown in **Fig. 3**. The red part is our developed connections, the black part is the connections of U-Net, and these dashed lines indicate skip connections. Different from U-Net++ [14], each part of the middle connection is the transposed convolution of the next adjacent layer. As shown in **Fig. 3**, we add decoding after each step of encoding, fuse feature maps with similar

semantics and a small span, and deal with the segmentation details.

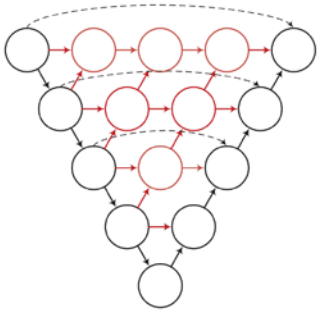

Fig. 3. The improved connection. The red part is our developed connections, the black part is the connections of U-Net, and these dashed lines indicate skip connections.

3.4 Multi-scale decoder

As we all know, multi-scale feature fusion is commonly used in the encoding. It can collect image features of different sizes of receptive fields through the multi-scale method. Because of this advantage, we introduce the multi-scale feature fusion to the decoding of our proposed MEDU-Net+ for getting better image segmentation performance.

In recent years, the decoder has been a hot research topic to restore the feature map to the network input through a series of operations. It consists of three parts: up-sampling convolution layer, concatenation feature obtained from skip connection and 3×3 deconvolution layer. We use multi-scale feature fusion to decoder instead of 3×3 deconvolution layer since it cannot obtain good image segmentation performance. The multi-scale feature fusion of decoding adopts different sizes convolution kernels to segment the object region as accurately as possible. The improved deconvolution block of the multi-scale decoder is shown in **Fig. 4**. It consists of input block, four deconvolution with different kernels and output block. The 1×1 and 5×5 convolution kernel branches are added to the deconvolution block of the multi-scale decoding in our proposed MEDU-Net+ network. Through the 1×1 deconvolution, the nonlinear characteristics are greatly enhanced while keeping the scale of the feature map unchanged, and better performance can be obtained.

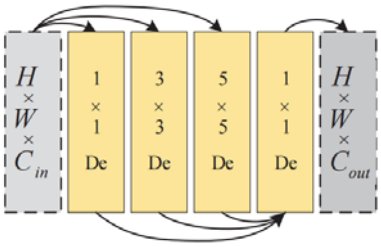

Fig. 4. The improved deconvolution block of the multi-scale decoder. It consists of input block, four deconvolution with different kernels and output block.

3.5 The overall architecture of our proposed MEDU-Net+

In recent years, the U-Net-based medical image segmentation has attracted much attention,

and many improved U-Net networks have been proposed gradually, such as V-Net [13], U-Net++ [14], 3D U-Net [12], MultiResUNet [18], CE-Net [15] and U2-Net [28]. Some improved methods mainly improve the basic convolution block in the encoding, add residual connections between convolution blocks, or deepen and widen the network. However, these methods often bring the parameters increase explosively.

It is different from these previous improved U-Net networks which only improve the encoder and/or the connection, our proposed MEDU-Net+ network integrates the information between encoder and decoder, and also improves the encoder, decoder and skip connection for getting better performance. The overall architecture of our proposed MEDU-Net+ is shown in **Fig. 5**. The proposed MEDU-Net+ network still uses the U-shaped structure. Compared with U-Net, the proposed network performance have been greatly improved, and the number of parameters has a certain amount of increase since the U-shape structure is completely filled with the deconvolution transfer term.

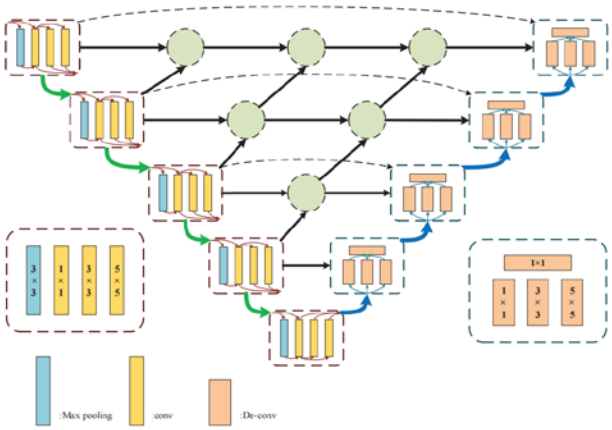

Fig. 5. The overall architecture of our proposed MEDU-Net+. It contains the multi-scale encoding combined with GoogLeNet, the new layer-by-layer skip connection and the multi-scale feature fusion of the decoder.

The proposed MEDU-Net+ contains the multi-scale encoding combined with GoogLeNet, the new layer-by-layer skip connection and the multi-scale feature fusion of the decoder. The function of the encoding block is to extract image feature information through a series of operations such as convolution and pooling. This block contains four submodules, each submodule containing four branches: 1×1 convolution, the concatenated 1×1 convolution and 3×3 convolution, the concatenated 1×1 convolution and 5×5 convolution, and the concatenated 3×3 maximum pooling and 1×1 convolution. The output is the result of concatenating these four branches. After each submodule, a down-sampling layer is implemented through the maximum pooling, which sequentially collects deeper semantic information. The connection block introduces a new layer-by-layer return skip connection, which establishes a more suitable path for transmitting information between encoding and decoding through the basic steps of convolution, deconvolution, and connection. The layer-by-layer skip connection can be seen as a cascade of multiple U-shaped networks, using this transmission form to collect and transmit semantic information of all coding layers. The corresponding decoding block consists of four layers, each layer containing a submodule which composes of up-sampling and deconvolution. Different scales convolution of 1×1, 3×3, and 5×5 are selected for deconvolution the information. The results of each branch are then aggregated to obtain the

output of each deconvolution block at the decoding. During decoding, the resolution is restored by up-sampling until it is consistent with the resolution of the input image. The detailed contributions of our proposed MEDU-Net+ are given as follows:

(1)The encoding convolution block of the MEDU-Net+ is replaced by the inception structure of the above-mentioned GoogLeNet convolution block. This inception structure includes 4 branches: the 1×1 convolution, the concatenated 1×1 and 3×3 convolutions, the concatenated 1 × 1 and 5 × 5 convolutions, and the concatenated 3×3 max pooling and 1×1 convolution. And the output of the encoding convolution block is the result of the concatenating of the four branches.

(2)We introduce a new skip connection in our proposed MEDU-Net+ network. The layer-by-layer skip connection is the addition of a transfer item based on deconvolution, which can be seen as a cascade of multiple U-shaped networks. And the process of this new skip connection is convolution, deconvolution and connection. At last, we aggregate all the collected semantic information.

(3)We introduce the multi-scale feature fusion to the decoding part for getting better image segmentation performance. The receptive fields in the aggregation process are equally important, and the extracted features can be recovered more comprehensively by using large and different scale receptive fields, so we choose convolution kernels of different sizes. The decoding part deconvolutes the convolution kernel sizes of 1 × 1, 1 × 1, 3 × 3 and 5 × 5, and then aggregates the results of each branch to obtain the output of the basic deconvolution block.

Generally, our proposed network has only one main structure, without adding redundant branch structure. This novel network architecture has considerable flexibility, and it is convenient to add modules to further improve the performance.

3.6 Loss Function

The loss function [29] is used to measure the difference between the real value and the predicted value of a network. As a classical loss function, the cross-entropy loss of U-Net has achieved good performance in most image semantic segmentation. However, when the number of object pixels only accounts for a small part of the whole image pixels, this loss function will lead to serious bias of the foreground to the background, which makes poor segmentation results. The dice loss function can solve the problem of imbalanced positive and negative samples, but it is very disadvantageous to the prediction of small objects. Once some pixels of small objects are mispredicted, the dice coefficient may fluctuate greatly, resulting in great gradient change and unstable training. The generalized dice loss function is to integrate multiple types of dice loss functions, which improves the small objects segmentation performance. This loss function formula is given as follows:

$$L_{gd} = 1 - 2\frac{\sum_{i=1}^{2} w_i \sum_{n=1}^{N} y_{in}\hat{y}_{in}}{\sum_{i=1}^{2} w_i \sum_{n=1}^{N} \left(y_{in} + \hat{y}_{in}\right)} \tag{1}$$

Where y_{in} is the ground truth of category i at the n-th pixel, N is the total number of pixels, $\hat{y}_{in}$ is the probability predicted value of y_{in}, and $w_i = \frac{1}{(\sum_{n=1}^{N} y_{in})^2}$ is the weight of each category. The generalized dice loss function can solve the problem of imbalanced positive and

negative samples and can improve the segmentation performance of small objects.

The focal loss function is suitable to solve the sample imbalance problem when the background pixels of medical image are far more than the object pixels. This loss function formula is given as follows:

$$L_f = -\frac{1}{N}\sum_{n=1}^{N}\left(\alpha\left(1-\hat{y}_n\right)^{\gamma} y_n \log\left(\hat{y}_n\right)+\left(1-\alpha\right)\hat{y}_n^{\gamma}\left(1-y_n\right)\log\left(1-\hat{y}_n\right)\right) \tag{2}$$

Where y_n is the ground truth of the *n*-th pixel, N is the total number of pixels, $\hat{y}_n$ is the probability predicted value of y_n, α and γ are parameters. We set α = 0.75 and γ = 2 in our experiment, since the network segmentation performance is the best under these parameter settings.

A new combined loss function is proposed by combining the generalized dice and focal loss functions, which has both advantages of them. This novel combined loss function formula can be expressed as follows:

$$L = \lambda L_{gd} + \left(1-\lambda\right) L_f \tag{3}$$

Where λ ($0 \le \lambda \le 1$) is the trade-off parameter, and we set λ = 0.1 in our experiment, since the performance is the best in this case. The novel combined loss function solves the problem that the segmentation result is seriously biased to the background when the object pixels are less than the background pixels in medical image, and makes the small objects of medical image get better segmentation performance.

4. Experimental results and analysis

4.1 Datasets and evaluation indexes

We choose the DRIVE [30], the ISBI2012 [31], and the CHAOS [32] datasets to train the network model to show the superiority of our proposed MEDU-Net+ network. The DRIVE is a retinal dataset with many small blood vessels, the most important thing for this dataset is how to extract as much detailed information as possible in the image. The ISBI2012 is a very common cell segmentation dataset in medical image analysis, which is challenged by ISBI. Its goal of segmentation is to extract cell edges which is a two-class classification problem. The CHAOS dataset includes MRI images and the ground truth of four organs including the spleen, liver, left kidney and right kidney [33]. Its goal of segmentation is to successfully distinguish the tumor area from other parts. In addition, these three datasets all have a small number of samples and a large amount of detailed information such as fundus blood vessels. It is noteworthy that we adjust the size of each image to a uniform value in the training process.

In order to quantitatively evaluate the performance of our proposed MEDU-Net+ network more clearly and intuitively, we choose the pixel accuracy (PA), the intersection over union (IoU), and the mean intersection over union (MIoU) as evaluation indicators in simulation experiment. The values of these three indicators are all between 0 and 1, and the larger the values, the better the segmentation performance.

4.2 Implementation Details

It is noteworthy that we adjust the size of each image to a uniform value in the training process. And the environment of our experiments is a server with Linux operating system. The CPU of server is intel Xeon E5-2695 which is with 220GB RAM while the GPU is NVIDIA

Tesla M40 (12 GB).

We adjust the size of each image to a uniform value during the training process. However, there are few images that can be trained in the medical image dataset, so we enhance these images in the process of image processing, flip the image 90 degrees horizontally and vertically, and try to use more complex elastic transformation to process these image, so as to increase the number of images through image enhancement. In order to avoid over fitting in the training process and reasonably evaluate the performance of our proposed network, we choose to use 10-fold cross-validation to optimize the whole network. We split the training and test sets, and then randomly combine them to generate ten different combinations of training and verification sets. In addition, we set the proportion of the verification set to 0.1, that is, 10% of the combined dataset is extracted as the test set.

In the training process, we firstly try to start with a larger bitch size, but the selected size is too large, resulting in very unstable training results. By gradually reducing the value of bitch size, we finally set the bitch size as 1. Although the training time becomes longer, the performance has also been greatly improved. We select Adam optimizer with the fuzzy factor eps = 1e − 8 and the initial learning rate is 0.01 which reduces by 0.1 dynamically.

According to [29], the network performance is the best when $\gamma = 2$, so we first refer to this value, and then increase or decrease by 0.5 with the center of 2 to observe the changes of the network performance. And then further increase or decrease the value in the unit of 0.1, and record the experimental results in turn. Finally, we found that the network performance is the best when $\gamma = 2$ through the experiment, so the value of γ is determined as 2. The existence of α can balance the attention to positive and negative samples in the training process. In order to adjust the proportion of background, we set the parameter $\alpha = 0.75$. As for the combined loss function, we firstly set the value of λ as 0.5, record the experimental results, and then set the value of λ as 0.4 and 0.6 respectively, and record the experimental results. The experimental results at $\lambda = 0.5$ are better than those at $\lambda = 0.6$ and worse than those at $\lambda = 0.4$, so we further set the range of 0.1 to 0.4. The best experimental results are obtained by similar methods when $\lambda = 0.1$, so we set the parameter $\lambda = 0.1$.

4.3 Ablation Study

To verify the effectiveness of our proposed multi-scale encoder with GoogLeNet, layer-by-layer skip connection, multi-scale decoder, and the loss function, qualitative and quantitative comparative experiments of different networks are carried out on the DRIVE datasets under the same experimental conditions. And we make the following four changes to compare with U-Net in this ablation experiment. The ablation experimental results are shown in **Table 1**.

Table 1. The ablation experimental results

Network	PA	IoU	MIoU
U-Net	0.9209	0.748	0.740
GU-Net	0.9330	0.759	0.767
SSCU-Net	0.9498	0.766	0.775
MU-Net	0.9447	0.763	0.773
CLU-Net	0.9354	0.757	0.756

First of all, we use the inception convolution to replace the 3 ×3 convolution in U-Net to verify the performance of multi-scale encoder with GoogLeNet, and we call this method as the GU-Net in this paper. According to the experiment results of **Table 1**, we can see that after the convolution of the encoding is replaced with the inception structure, the PA, IOU, and

MIoU three quantitative indicators all have a certain improvement. In the improved encoding, not only 3×3 convolution is used to extract information, 1×1 and 5×5 convolutions are also added to the inception structure. The improved encoding increases the receiving field and is conducive to collecting more information, hence improves the network performance.

In addition, to prove the effectiveness of the improved skip connection, we apply the proposed layer-by-layer skip connection to the connection between U-Net encoder and decoder, and we call this method as the single skip connection U-Net (SSCU-Net) in this paper. Experiments on the DRIVE dataset in **Table 1** show that the values of PA, IoU and MIoU are increased to 0.9498, 0.766 and 0.775, respectively. The performance of the proposed layer-by-layer skip connection network are improved since the previous skip connection does not completely transmit the depth information to the decoder, so we return the extracted information in time through the layer-by-layer method. The decoder can provide more available image feature information for the decoding process.

And then, we improve the decoding of U-Net to the multi-scale decoding to compare it with U-Net on the DRIVE dataset, and we call this method as the multiscale U-Net (MU-Net) in this paper. Experiments on **Table 1** show that the values of PA, IoU and MIoU are increased to 0.9447, 0.763 and 0.773, respectively. The U-Net network has been significantly improved by the multi-scale decoding. The reason for the improvement performance is that the multi-scale decoding can make full use of the collected feature information, and achieve better image segmentation results. Finally, we replace the cross-entropy with the combined loss function, and we call this method as the combined loss U-Net (CLU-Net). Experiments on **Table 1** show that the values of PA, IoU, and MIoU are correspondingly improved to 0.9354, 0.757, and 0.756, respectively. The new combined loss function improves the problem that the object pixel is smaller than the background pixel, so that the small object medical image can obtain better segmentation performance.

4.4 Comparisons with other networks

To verify the effectiveness of our proposed MEDU-Net+, we compared it with other image segmentation networks including U-Net, V-Net, MA-Unet, MDU-Net, MultiResUNet and CE-Net on these three datasets. We give the visual effects of different network for the image segmentation of the three different datasets in **Figs. 6-8**, respectively.

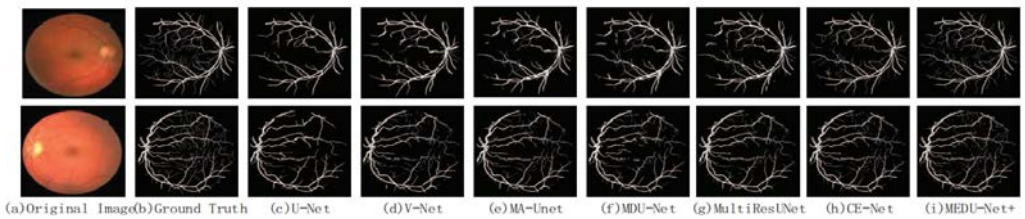

Fig. 6. The visual segmentation results of DRIVE.

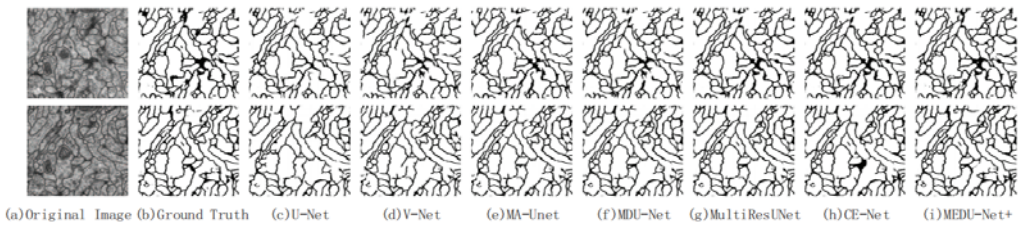

Fig. 7. The visual segmentation results of ISBI2012.

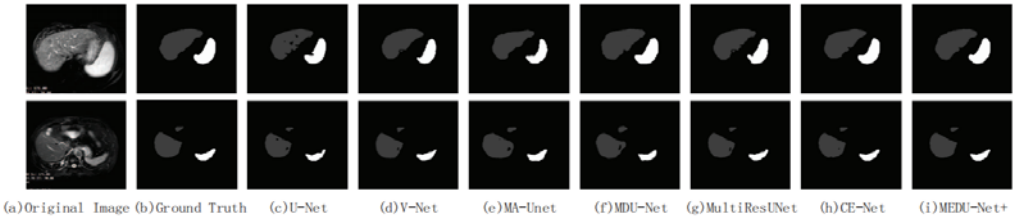

Fig. 8. The visual segmentation results of CHAOS.

The DRIVE dataset contains the most detailed information, so in the segmentation process, the extract information ability of the segmentation network is particularly important. As shown in **Fig. 6**, the image segmented by the U-Net network can only segment basic contours and shapes, and there are almost no small blood vessels in the result image. Compared with MA-Unet network with multi-scale structure in the encoder, the segmentation effect of V-Net is not obvious. The image segmented by MA-Unet network is better than V-net network, and some medium-sized details can be segmented. MDU-Net network adopts multi-scale densely connection, which can extract more detailed information, and the segmented image is also finer, but the edge processing is still lacking; MultiResUnet network adds residual connection on the basis of multi-scale, and the detail processing is also better. CE-Net network adopts multi-scale and attention mechanism, which has better performance. It can not only completely segment the main information, but also extract more complete detailed information. Through our proposed MEDU-Net+ segmentation, the extracted image information is the most complete. Compared with the ground truth, it is not difficult to find that the segmented images of our proposed MEDU-Net+ misses some very detailed parts, and the edge information is almost not lost.

The ISBI2012 dataset is a representative segmented dataset in this experiment, the image of this dataset does not contain too much detailed information. As shown in **Fig. 7** that the visual segmentation results of U-Net, V-Net, MA-Unet, MDU-Net, MultiResUnet, and CE-Net networks become better in turn. The segmented results of our proposed network have the most complete outline, and are the most similar to their ground truths.

The edge information of CHAOS is better than the DRIVE dataset, but there is interference information in the image, so it is easy to cause misjudgment when the feature information is insufficient in decoding. This situation is very obvious in U-Net and V-Net networks. As shown in **Fig. 8**, the results segmented by U-Net and V-Net networks are misjudged, and the background information is misjudged as the object information. Although the performance of MA-Unet, MDU-Net and MultiResUnet have been improved, there are also some errors. CE-Net network does not have a judgment error, but the contours of the segmented images are incomplete. The proposed MEDU-Net+ network not only correctly segments the object, but also the contours of the segmented images are more in line with their ground truths.

In conclusion, we can see from **Figs. 6-8** that our proposed MEDU-Net+ can get better segmentation results than the other compared networks such as U-Net, V-Net, MA-Unet, MDU-Net, MultiResUnet, and CE-Net on the DRIVE, ISBI2012 and CHAOS datasets.

In order to further clearly demonstrate the performance of our proposed MEDUNet+ network, we use PA, IoU and MIoU as quantitative indicators to conduct experiments on the DRIVE, ISBI2012 and CHAOS datasets, and compare it with other medical image segmentation networks such as U-Net, V-Net, MA-Unet, MDU-Net, MultiResUnet, CE-Net, DIU-Net, and MECAU-Net. The experimental results are shown in **Table 2**.

Table 2. The experimental results of different networks

Network	DRIVE			ISBI2012			CHAOS		
	PA	IoU	MIoU	PA	IoU	MIoU	PA	IoU	MIoU
U-Net	0.9209	0.748	0.740	0.9339	0.747	0.738	0.9172	0.743	0.741
V-Net	0.9283	0.766	0.755	0.9482	0.761	0.764	0.9230	0.758	0.747
MA-Unet	0.9312	0.767	0.762	0.9467	0.761	0.759	0.9307	0.760	0.753
MDU-Net	0.9330	0.766	0.767	0.9484	0.765	0.762	0.9346	0.763	0.762
MultiResUNet	0.9392	0.767	0.763	0.9501	0.764	0.763	0.9378	0.778	0.776
CE-Net	0.9393	0.764	0.778	0.9478	0.762	0.769	0.9308	0.773	0.768
DIU-Net	0.9395	0.768	0.774	0.9485	0.769	0.770	0.9398	0.776	0.773
MECAU-Net	0.9491	0.777	0.786	0.9488	0.779	0.780	0.9489	0.780	0.779
MEDU-Net+	**0.9587**	**0.783**	**0.790**	**0.9543**	**0.786**	**0.789**	**0.9548**	**0.782**	**0.780**

As shown in **Table 2**, the performance of V-Net, MA-Unet, MDU-Net, MultiResUnet, CE-Net, DIU-Net and MECAU-Net are better than that of U-Net, and our proposed MEDU-Net+ achieves the best performance on the DRIVE, ISBI2012 and CHAOS datasets. Specifically, on the DRIVE dataset, the PA, IoU and MIoU values of U-Net are 0.9209, 0.748 and 0.740 respectively, and the PA, IoU and MIoU values of our proposed MEDU-Net+ are 0.9587, 0.783 and 0.790 respectively, which are 4.105%, 4.679% and 6.757% relatively higher than that of U-Net respectively. On the ISBI2012 dataset, U-Net has also achieved good performance, and the performance of our proposed MEDU-Net+ are still the best. The values of PA, IoU and MIoU are 0.9543, 0.786 and 0.789 respectively, which are 2.184%, 5.221% and 6.911% relatively higher than that of U-Net respectively. On the CHAOS dataset, the performance of our proposed MEDU-Net+ are also the best. The values of PA, IoU and MIoU are 0.9548, 0.7832 and 0.780 respectively, which are 4.099%, 5.249% and 5.263% relatively higher than that of U-Net respectively. In conclusion, compared with the other medical image segmentation networks, our proposed MEDU-Net+ achieves the best performance. The first reason of the best results of our proposed MEDU-Net+ is that the GoogLeNet is designed in the encoding to obtain more semantic information; In addition, the new layer-by-layer skip connection is designed between encoding and decoding to connect the information of each layer, preserving all feature information as much as possible to reduce the semantic gap between the left and right parts of feature information. Besides, the deconvolution block at the decoding with multi-scale processing helps better recover the segmented image. Finally, the new combined loss function makes the proposed network segment small object medical images and extract more edge feature information.

In order to demonstrate the complexity of our proposed MEDUNet+ network, we use parameters size (MB) as the quantitative indicator to conduct experiments on the DRIVE datasets, and compare it with other medical image segmentation networks such as U-Net, V-Net, MA-Unet, MDU-Net, MultiResUnet, CE-Net, DIU-Net, and MECAU-Net. The experimental results are shown in **Table 3**.

Table 3. The experimental parameters size of different networks

Network	U-Net	V-Net	MA-Unet	MDU-Net	Multi-ResUNet	CE-Net	DIU-Net	MECAU-Net	MEDU-Net+
Parameter	**5.43**	8.93	10.57	12.33	12.45	15.24	12.38	5.89	12.25

It can be seen from **Table 3** that parameter size of the proposed MEDU Net+ is smaller than that of MDU-Net, MultiResUnet, CE-Net, and DIU-Net, and larger than that of U-Net, V-Net, MA-Unet, and MECAU-Net. Parameter size is an effective representation of network complexity. Experimental results have verified that the complexity of the proposed MEDU Net+ is lower than that of MDU-Net, MultiResUnet, CE-Net, and DIU-Net, and higher than U-Net, V-Net, MA-Unet, and MECAU-Net, but its performance is better than other comparative networks.

5. Conclusion

In this paper, we propose a novel improved U-Net called MEDU-Net+ for medical image. We introduce GoogLeNet to obtain more information at the encoder, and present multi-scale feature extraction to fuse different scales of semantic information at the encoder and decoder. At the same time, we also introduce a layer-by-layer skip connection to connect the information of each layer, so there is no need to encode to the last layer and then return the information. In addition, we divide the unknown depth network into each part of the deconvolution layer, replacing the original U-Net network directly connected to the encoder and decoder. And then, combining the advantages of the generalized dice and the focal loss function, we propose a new combined loss function to extract more edge information. Finally, we test our proposed MEDU-Net+ network and other classical medical image segmentation networks on three classical datasets. The experimental results show that our proposed MEDU-Net+ has a significant improvement compared with other medical image segmentation networks. We also acknowledge limitations in the data, the model, and the experimental setup of our proposed network to show deep understanding of our research and provide avenues for future work. For the data, we select three representative medical image datasets for segmentation experiments, but their proportion in medical images is still very small. In future work, we will consider conducting experiments on more medical image datasets to improve the universality of segmentation networks in the field of medical images. For the model, the proposed medical image segmentation network in this paper mainly focuses on the segmentation of two-dimensional medical images. For three-dimensional medical image segmentation, it can be sliced into two-dimensional images for processing one by one, but the relationships between images and the characteristics of different medical images are not considered. In further work, the relationships between images and the characteristics of different medical images will be considered to improve the segmentation accuracy of the network for three-dimensional medical images. For the experimental setup, the experimental parameters in this paper are manually selected through a large number of experiments. In future work, optimization algorithms for parameter selection will be studied and implemented to obtain the optimal parameters.

References

[1] Nirkin Y, Wolf L, Hassner T, “Hyperseg: Patch-Wise Hypernetwork for Real-Time Semantic Segmentation,” in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.4061-4070, 2021. Article(CrossRefLink)

[2] Shijie Hao, Yuan Zhou, Yanrong Guo, “A Brief Survey on Semantic Segmentation with Deep Learning,” *Neurocomputing*, vol.406, pp.302-321, 2020. Article(CrossRefLink)

[3] Siddique N, Paheding S, Elkin C, Devabhaktuni V, "U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications," *IEEE Access*, vol.9, pp.82031-82057, 2021. Article(CrossRefLink)
[4] Liu X, Yang L, Chen J, Yu S, Li K, "Region-to-boundary deep learning model with multi-scale feature fusion for medical image segmentation," *Biomedical Signal Processing and Control*, vol.71, pp.1-10, 2022. Article(CrossRefLink)
[5] Long J, Shelhamer E, Darrell T, "Fully Convolutional Networks for Semantic segmentation," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp.3431-3440, 2015. Article(CrossRefLink)
[6] Mei Y, Jin H, Yu B, Wu E, Yang K, "Visual geometry group-UNet: Deep learning ultrasonic image reconstruction for curved parts," *The Journal of the Acoustical Society of America*, vol.149, no.5, pp.2997-3009, 2021. Article(CrossRefLink)
[7] Ronneberger O, Fischer P, Brox T, "U-net: Convolutional Networks for Biomedical Image Segmentation," in *Proc. of Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp.234-241, 2015. Article(CrossRefLink)
[8] Yang Z, Shao J, Yang Y, "An Improved CycleGAN for Data Augmentation in Person Re-Identification," *Big Data Research*, vol.34, pp.1-10, 2023. Article(CrossRefLink)
[9] Zhang Y, Deng L, Zhu H et al., "Deep learning in food category recognition," *Information Fusion*, vol.98, 2023. Article(CrossRefLink)
[10] Lu S, Zhu Z, Gorriz J M et al., "NAGNN: Classification of COVID-19 based on neighboring aware representation from deep graph neural network," *International Journal of Intelligent Systems*, vol.37, no.2, pp.1572-1598, 2021. Article(CrossRefLink)
[11] Lu S, Wang S H, Zhang Y, "Detection of abnormal brain in MRI via improved AlexNet and ELM optimized by chaotic bat algorithm," *Neural Computing and Applications*, vol.33, pp.10799-10811, 2021. Article(CrossRefLink)
[12] Çiçek Ö, Abdulkadir A, Lienkamp S S et al., "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation," in *Proc. of Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, vol.9901, pp.424-432, 2016. Article(CrossRefLink)
[13] Milletari F, Navab N, Ahmadi S-A, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," in *Proc. of 2016 Fourth International Conference on 3D Vision (3DV)*, pp.565-571, 2016. Article(CrossRefLink)
[14] Zhou Z, Siddiquee M M R, Tajbakhsh N, Liang J, "Unet++: A Nested U-Net Architecture for Medical Image Segmentation," in *Proc. of Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, vol.11045, pp.3-11, 2018. Article(CrossRefLink)
[15] Gu Z, Cheng J, Fu H et al., "CE-Net: Context Encoder Network for 2D Medical Image Segmentation," *IEEE Transactions on Medical Imaging*, vol.38, no.10, pp.2281-2292, 2019. Article(CrossRefLink)
[16] He K, Zhang X, Ren S, Sun J, "Deep Residual Learning for Image Recognition," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.770-778, 2016. Article(CrossRefLink)
[17] Yang Z, Xu P, Yang Y, Bao B-K, "A Densely Connected Network Based on U-Net for Medical Image Segmentation," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol.17, no.3, pp.1-14, 2021. Article(CrossRefLink)
[18] Ibtehaz N, Rahman M S, "MultiResUNet : Rethinking the U-Net architecture for multimodal biomedical image segmentation," *Neural Networks*, vol.121, pp.74-87, 2020. Article(CrossRefLink)
[19] Zhang J, Zhang Y, Jin Y, Xu J, Xu X, "MDU-Net: multi-scale densely connected U-Net for biomedical image segmentation," *Health Information Science and Systems*, vol.11, 2023. Article(CrossRefLink)

[20] Cai Y, Wang Y, "MA-Unet: an improved version of Unet based on multi-scale and attention mechanism for medical image segmentation," in *Proc. of Third International Conference on Electronics and Communication; Network and Computer Technology (ECNCT 2021)*, vol.12167, 2022. Article(CrossRefLink)
[21] Zhang Z, Wu C, Coleman S, Kerr D, "DENSE-INception U-net for medical image segmentation," *Computer Methods and Programs in Biomedicine*, vol.192, 2020. Article(CrossRefLink)
[22] Yang Z, Sun X, Shao J, et al., "Medical image segmentation based on multiscale even convolution attention U-Net," *Journal of Signal Processing*, vol.38, no.9, pp.1912-1921, 2022.
[23] Szegedy C, Liu W, Jia Y et al., "Going Deeper With Convolutions," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.1-9, 2015. Article(CrossRefLink)
[24] Cheng X-R, Cui B-J, Hou S-Z, "Fault Line Selection of Distribution Network Based on Modified CEEMDAN and GoogLeNet Neural Network," *IEEE Sensors Journal*, vol.22, no.13, pp.13346-13364, 2022. Article(CrossRefLink)
[25] Niyas S, Pawan S J, Kumar M A, Rajan J, "Medical image segmentation with 3D convolutional neural networks: A survey," *Neurocomputing*, vol.493, pp.397-413, 2022. Article(CrossRefLink)
[26] Mishra S, Zhang Y, Chen D Z, Hu X S, "Data-Driven Deep Supervision for Medical Image Segmentation," *IEEE Transactions on Medical Imaging*, vol.41, no.6, pp.1560-1574, 2022. Article(CrossRefLink)
[27] Zhang S, Wen L, Bian X et al., "Single-Shot Refinement Neural Network for Object Detection," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.4203-4212, 2018. Article(CrossRefLink)
[28] Qin X, Zhang Z, Huang C et al., "U2-Net: Going deeper with nested U-structure for salient object detection," *Pattern Recognition*, vol.106, 2020. Article(CrossRefLink)
[29] Jadon S, "A survey of loss functions for semantic segmentation," in *Proc. of 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pp.1-7, 2020. Article(CrossRefLink)
[30] Wu Y, Xia Y, Song Y et al., "Multiscale Network Followed Network Model for Retinal Vessel Segmentation," in *Proc. of Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, vol.11071, pp.119-126, 2018. Article(CrossRefLink)
[31] Ghosh S, Das N, Das I et al., "Understanding Deep Learning Techniques for Image Segmentation," *ACM Computing Surveys*, vol.52, no.4, pp.1-35, 2019. Article(CrossRefLink)
[32] Hafiz A M, Bhat G M, "A survey on instance segmentation: state of the art," *International Journal of Multimedia Information Retrieva*l, vol.9, pp.171-189, 2020. Article(CrossRefLink)
[33] Liu Z, Tong L, Chen L et al., "Deep learning based brain tumor segmentation: a survey," *Complex & Intelligent Systems*, vol.9, pp.1001-1026, 2023. Article(CrossRefLink)

Zhenzhen Yang received her M.S. and Ph.D. degrees from Nanjing University of Posts and Telecommunications in 2011 and 2014, respectively. She was a Lecturer with the Nanjing University of Posts and Telecommunications in 2014, where she was promoted to an associate professor in 2018. Her research interests include computer vision and pattern recognition. Since 2015, she is/was the project manager of several national projects such as National Natural Science Foundation of China, the China Postdoctoral Science Foundation. She is the author of more than 40 journal and conference papers.

Xue Sun received her M.S. degree in 2022 from Nanjing University of Posts and Telecommunications. Her main research interests include medical image segmentation and classification.

Yongpeng Yang received his M.S. degree in 2011 from Nanjing University of Posts and Telecommunications and he is studying for his Ph.D. degree in Nanjing University of Posts and Telecommunications from 2021 until now. Now he is a lecturer in Nanjing Vocational College of Information Technology. His main research interests include computer vision and pattern recognition.

Xinyi Wu received her B.E degree in 2022 from Jiangsu University of Technology in China and she is studying for her M.S degree at Nanjing University of Posts and Telecommunications from 2022 until now. Her main research interests include computer vision and pattern recognition.