Journal of Imaging Science and Technology® 66(4): 040416-1–040416-15, 2022.

An End-to-end Medical Image Segmentation Model based on Multi-scale Feature Extraction

Tingzhong Wang, Hailong Wang, Nanjie Li, Junhong Xian, Zhongqian Zhao, and Deguang Li
School of Information Technology, Luoyang Normal University, Luoyang, 471934, China
E-mail: lideguang@lynu.edu.cn

Abstract. *With the development of technology, medical image examination methods are steadily diversified, including ultrasound, pathology, endoscopy, computed tomography, computed radiography, magnetic resonance imaging, radionuclide imaging, positron emission computed tomography, angiography, etc. More than 70% of clinical diagnoses depends on medical imaging. Owning to the fact that medical image data grows faster, the problem of lack of radiodiagnosis doctors are more serious. Although the accuracy of doctors' diagnosis is high, the accuracy of judgment is questionable due to energy constraints, emotional fluctuations and other factors. To assist accurate and effective diagnosis and reduce the probability of misdiagnosis, in this paper, a novel end-to-end model with an encoder network and a decoder network (MS-UNet) is proposed. The encoder network extracts feature information from the input image, and the decoder network restores the encoded content to the input size. Among them, a multi-scale module (MSM) is built to help extract more accurate information that represents the feature maps of objective segmentation by fusing multiple different convolution information at different stages. The experimental results on two well-known data sets show that the proposed model outperforms other state-of-the-art algorithms, and the introduction of MSM can significantly improve performance.* © *2022 Society for Imaging Science and Technology.*
[DOI: 10.2352/J.ImagingSci.Technol.2022.66.4.040416]

1. INTRODUCTION

In recent years, artificial intelligence has become a research hotspot because of its outstanding intelligent learning ability, and has been successfully applied in many aspects. Medical image refers to the non-invasive method, technology and processing which is used in a part of the human body to obtain images of the internal organ and tissue. The application of artificial intelligence to medical images mainly focuses on the analysis and judgment of medical images by machines through in-depth learning. Medical imaging, as an auxiliary tool, can help doctors process image information faster and assist doctors in facilitating diagnosis and treatment by conducting qualitative and quantitative analysis, and improve efficiency when doctors look at images to find hidden lesions [1]. Through image classification, target detection, image segmentation, image retrieval and other methods, artificial intelligence completes the functions of lesion recognition and labeling, three-dimensional reconstruction, automatic target delineation and adaptive radiotherapy, which is applied in the stage of disease screening, diagnosis and treatment [2].

The data from the medical community of 1,241 imaging doctors show that more than 50% of doctors work more than 8 hours a day, 20.6% of doctors work more than 10 hours, and more than 71% of imaging doctors expect to finish their work on time and have more time to spend with families. In addition, the shortage of physicians in China is becoming more and more serious, and the imaging department is the hardest hit area. At present, the annual growth rate of medical imaging data in China is about 30%, while the annual growth rate of the number of radiologists is only 4.1%. The growth in the number of imaging physicians is far less than the growth in image data, and doctors need a long time of training and learning, which means that imaging physicians will have more and more pressure to process image data in the future, and it is difficult to bear this huge workload. The disadvantages of manual analysis are also obvious: first, a lot of mental work and long-time work are susceptible to visual fatigue and human visual errors; Second, massive image information increases the chance of misdiagnosis; Third, based on doctors' experience, lack of quantitative standards, high misdiagnosis rate, and it is very easy to have inconsistent interpretation results by different doctors.

According to statistics [2], 12 million people are misdiagnosed in the United States and 57 million in China. Additionally, according to misdiagnosis data of the Chinese Medical Association, the total misdiagnosis rate of clinical medicine in China is 27.8%, of which the average misdiagnosis rate of malignant tumors is 40%, the misdiagnosis rate of ectopic organs is 60%, and the average misdiagnosis rate of extra pulmonary tuberculosis such as liver tuberculosis and gastric tuberculosis is also more than 40%. These misdiagnoses mainly occur in Grass-roots Medical Institutions [2]. "Artificial intelligence can improve the medical effect by 50% and reduce the medical cost by up to 50%" is the conclusion made by Frost & Sullivan, a reputed consulting company in the United States [3]. There is an urgent need for an automatic segmentation method to assist doctors in rapid diagnosis, improve work efficiency and reduce doctors' work burden, which is also a problem that researchers have been trying to solve [4].

At present, the vigorous development of computer vision technology will greatly promote the realization of this goal. Especially with the rise of deep learning technology, computer vision technology is more and more popular in

Received Jan. 12, 2022; accepted for publication June 6, 2022; published online June 24, 2022. Associate Editor: Bo Wang.
1062-3701/2022/66(4)/040416/15/$25.00

medical auxiliary diagnosis. Among them, convolutional neural network (CNN) is a classical and widely used deep neural network structure that has strong image feature extraction ability, small parameter quantity and strong generalization ability. For example, classical GoogLeNet, ResNet and other networks have been widely used in the field of image processing [5, 6]. These networks are suitable for image recognition and classification, but lack pertinence for image segmentation. Fully Convolutional Networks for Semantic Segmentation (FCN) [7] proposed by Berkeley University is a milestone in image segmentation. It is the first end-to-end, point-to-point semantic segmentation model that uses full convolution to solve the prediction problem per pixel. The biggest difference between FCN network and previous networks is that it uses a cross layer with skip connection structure to integrate the shallow texture information and deep semantic information to achieve accurate segmentation tasks. It contains three core ideas of CNN. One is the full convolution network without full connection layer, which can adapt to the output of any size. Second, the deconvolution layer that increases the data size can output accurate targets. Third, the step-by-step structure combined with different deep-seated results can ensure robustness and accuracy at the same time. Because of its excellent performance, it has become the basic framework of semantic segmentation since it was proposed. Convolutional neural network based on encoder and decoder architecture is the most basic and important segmentation model in medical image segmentation. U-Net [8] is one of the first algorithm to use full convolution neural network for semantic segmentation. It is a classical end-to-end architecture, includes an encoder (contraction path) for feature extraction, a skip connection structure to realize feature stitching, and a decoder (expansion path) to restore the image size to the input image resolution to realize segmentation.

Medical images have fixed structure with simple and important semantic information. Compared with automatic driving, semantic information in medical images is relatively single, so there is minimal trivial information to filter when learning feature representation by convolutional neural network. Specially, all features of medical images are very important, so low-level features and high-level semantic features are very important, the skip connection widely used in U-shape structure carries low-level and high-level semantic feature information and is very suitable for solving the problem. In addition, there are few medical image data and it is difficult to obtain. The amount of data may be only a few hundred or even less than 100. Therefore, it is easy to over fit if using large networks, such as deeplabv3+. The advantage of large-scale network is stronger image representation ability, while the relatively simple and small number of medical images do not have so much content to express. Consequently, it was also found that in the decimal order, the segmented state-of-the-art (SOTA) model has no advantage over the lightweight U-Net. Moreover, medical images are often multimodal. For example, in the ISLES cerebral infarction competition, officials provided multimodal data such as CBF, MTT and CBV. Therefore, in medical imaging tasks, we often need to design our own network to extract different scale features.

U-Net is a classic feature pyramid architecture, including encoder and decoder network. Encoder extracts different levels of features (such as edge, shape, semantics, etc.) through multiple down sampling, and decoder integrates multi-level features for decoding to restore the encoded content to the same as the input images. Compared with FCN network that directly decodes for the high-level semantic features, U-Net increases the interaction and fusion of different levels of features, and has stronger feature representation ability. As for the characteristics of medical images, organs or lesions show the characteristics of multi-scale and large morphological differences. Therefore, U-Net can improve learning of different scale features. From the perspective of network optimization, the horizontal connection of different levels of features of encoder is similar to short connection. It strengthens the feature transmission of the whole network and can alleviate the problem of gradient disappearance; Strengthening feature reuse can also improve the learning ability of the network. Focusing on the application of U-Net and its improved network in the field of medical image processing, the improvement measures and effects of relevant networks in the process of practical application are discussed. Finally, it summarizes the content of the full text, analyzes the key and difficult points of U-Net network, and puts forward its own suggestions and prospects for the development of related work in the future.

In this paper, we propose a novel end-to-end model with an encoder network and a decoder network. In view of the importance of low-level and high-level features in medical images for diagnosis, we design a module based on multi-scale. The module is arranged in each stage of coding and decoding network in our model to assist the model to obtain accurate low-level and high-level target semantic features in each stage. The remainder of this paper is organized as follows. Section 2 reviews the related research work of medical segmentation and U-Net. Section 3 describes our proposed network. Section 4 reports the implementation process, experimental results and analysis of the model design. Section 5 concludes our work.

2. RELATED WORK

2.1 *Characteristics of Medical Image Segmentation*

Medical images have high complexity with essentially no linear features. Thus, it is difficult to segment the target region of interest [9]. In the face of medical image data, the segmentation target and direction should be clarified first, whether to segment organs or lesions. Segmentation of different organs also has different detailed objectives, such as segmentation of pulmonary nodules, brain tissue and brain tumor, etc. [10–12]. Due to the limitations of current technology, the unified model cannot be used to complete different task segmentation. The fusion and forcible segmentation will lead to the failure of prior

knowledge, thus reducing the accuracy of segmentation. Before segmentation of medical images, it is necessary to understand the characteristics of medical images, which can be generally divided into the following five categories [13]: First, medical images are multimodal, with many different formats, such as CT, MRI, PET, ultrasound imaging and RGB images of the retina of the fundus, etc. Different hospitals and even the same hospital often use different types of imaging equipments, resulting in a variety of imaging results. Different formats of medical images have different emphasis on patient information feedback. MRI is suitable for observing soft tissues such as liver, abdomen, brain, chest and pelvis with high resolution, but there are also artifacts. It takes a long time to acquire and the image quality is difficult to standardize. CT images can reflect bleeding in organs and tissues. The X-ray image will give a clearer view of the bone. Second, medical image segmentation is easily affected by noise and artifacts. The appearance of noise will lead to the appearance of grainy, mottled, snowflake or texture in the image, and will reduce the visibility of some features in the image. Noise in medical imaging is unavoidable because of the difference between imaging principle and equipment. In the actual segmentation, noise should be suppressed as much as possible and more details of the image should be retained to achieve effective segmentation. Artifacts are usually generated in 3D reconstruction or registration of images, which cannot be avoided and eliminated at present. Therefore, when deep learning network is used for segmentation, data imaging characteristics should also be considered for segmentation design. Third, the dynamic distribution of pixel values in medical images is very large. Natural images generally range from 0 to 255 pixels, but medical images can be in the thousands (such as CT images), and even floating-point data (such as X-ray imaging). Fourth, the effective sample size of medical image data is generally small. Medical image datasets are collected from patients and are difficult to be generated manually. From the point of medical ethics, medical images cannot be collected artificially. In contrast, in the field of natural images, sample data can be generated in large quantities. Finally, medical image annotation generally needs to be completed by medical practitioners, and it is extremely difficult to obtain the annotation data, which will involve patient privacy issues and cannot be solved by spending money. In other fields, ordinary people can annotate images, even with the help of deep learning technology, and the cost of annotation is very low. The high cost of annotation also determines that the data volume of medical image samples should not be too large.

2.2 *Preprocessing of Medical Image Segmentation*

When the segmentation algorithm based on encoding and decoding structure is used for medical image segmentation, it is necessary to preprocess the image. Due to different datasets and segmentation tasks, there are differences in data preprocessing methods, but there are also some common operations. The following sections will cover both sampling and data augmentation.

For sampling, it is necessary to fully understand the format and characteristics of corresponding medical images. Different medical devices produce different medical image benchmarks, resulting in different pixel sizes and actual sizes. Furthermore, the resolution of medical images is generally too high. In many cases, the model is not trained because of insufficient memory. The input layer cannot directly accept an unprocessed image and must be sampled. Current sampling methods can be divided into two categories. One is to use the operation of direct resizing, which has a great impact on the accuracy. The second type is patch-based training, which involves the crop operation of the original image. Under this strategy, the image is clipped and sampled into multiple patches according to the target size, which may have overlapping parts or may not be correlated. In general, random patch clipping is carried out to obtain higher difference between patches and better convergence in the training process.

For data augmentation, usually only for training sets, it is a good method to reduce the generalization error of computer vision models. To make the data more conducive to network training, there are two common operations. The first is to improve the generalization ability of the model, mainly using elastic deformation, such as translation, rotation, mirror transformation, contrast enhancement, histogram equalization and other operations to enhance the data in real time during the training process. The second is to use deep learning methods to enhance the data set. For example, the generative adversarial network (GAN) is used to generate medical image data with lesions, and the texture and spatial attributes of the image itself are used for learning transformation to enhance the data set [14].

2.3 *Application of U-Net in Medical Image Segmentation*

U-Net is the most famous network architecture in the field of medical image segmentation. In 2015, Ronneberger et al. [8] participated in the ISBI Challenge and proposed a segmentation network based on FCN. The modified and extended U-Net can adapt to a small training set and output more accurate segmentation results. As shown in Figure 1, there are still a large number of channels in the up-sampling process of U-Net, which makes the network propagate the context information to a higher resolution, and its expansion path and contraction path are symmetrical to form a U-shaped shape segment and fuse the feature maps from different stages by jumping connection.

Once the U-Net network architecture was proposed, it attracted the attention of many researchers in the field of medical image segmentation. For example, Gordienko et al. [15] used the U-Net for the lung segmentation experiment of Chest X-ray images. The results show that U-Net can implement fast and accurate medical image segmentation. When facing the task of medical image segmentation, U-Net coding-decoding network architecture composed of the extended path and contracted path has become the first choice. At the same time, driven by new technology, researchers have developed many variants based on U-Net.

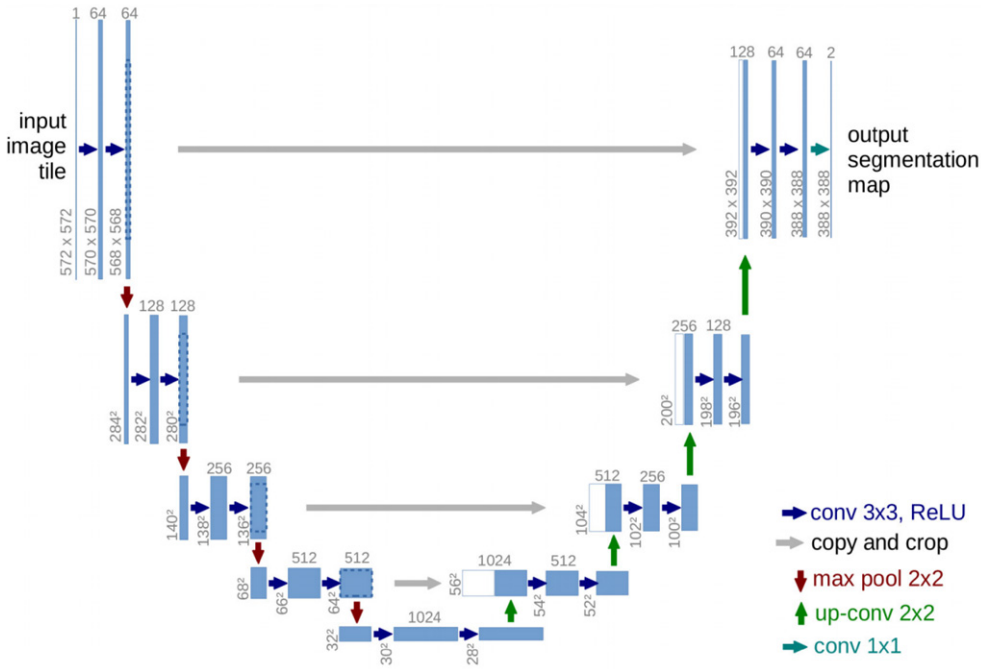
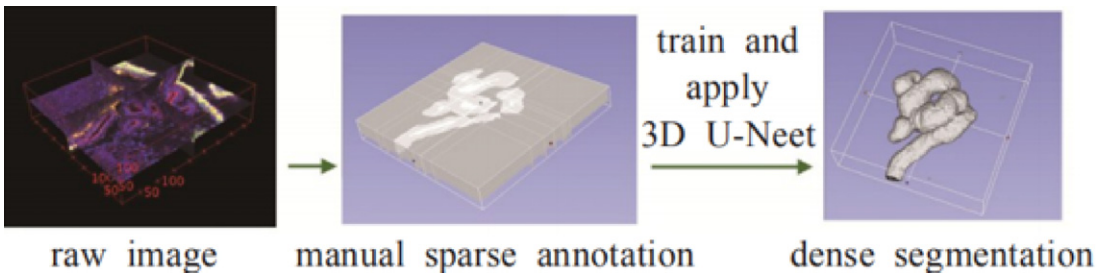


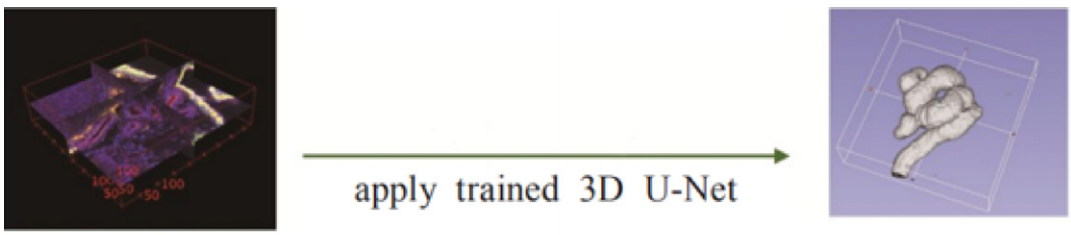
Figure 1. The overview framework of U-Net.

The idea of dense connection comes from DenseNet [16]. Before DenseNet appeared, the evolution of convolutional neural networks was generally carried out by deepening or widening the number of layers. DenseNet proposed a new structure through reusing features, which not only slowed down the phenomenon of gradient disappearance but also had fewer parameters of the model. The U-Net++ network architecture was proposed by Zhou et al. [17] in 2018. The innovation lies in adding dense connections to the U-Net, introducing the idea of in-depth supervision [18], and integrating u-net structures of different sizes into a network through redesigned jump connection paths. On the original U-Net, U-Net++ adds more hop connection paths and up sampling convolution blocks to bridge the semantic gap between encoder and decoder. On the one hand, the depth supervision used in the middle-hidden layers can solve the gradient disappearance problem during U-Net++ training, on the other hand, it allows the network to prune in the test stage and reduce the inference time of the model. Huang et al. proposed U-Net3+ [19] in 2020. For the deficiency that U-Net++ cannot extract enough information from multi-scale, U-Net3+ uses full-scale jump connection and depth supervision to improve this problem. The full-scale jump connection combines the high-level semantics and low-level semantics from different scale feature maps, while the deep supervision learns the hierarchical representation from the multi-scale aggregated feature map. In addition, there is proposed a hybrid loss function and designed a classification guidance module in U-Net3+ to enhance the organ boundary and reduce the over-segmentation of non-organ images, so as to obtain more accurate segmentation results. The primary advantage of U-Net++ and U-Net3+ architectures is to improve accuracy, which is brought by its integration of different levels of characteristics. The second is the flexible network structure combined with deep supervision so that the deep neural network with huge.

In 2018, Alom et al. [20] proposed the R2U-Net network architecture, which integrates the structure of U-Net, ResNet, and RCNN [21]. The R2U-Net network obtained good experiments on multiple medical image segmentation tasks such as blood vessels, lungs, and retinas. The RCNN combines convolutional neural networks and recurrent neural networks, allowing the neural network to memorize serialized input information. It not only implicitly increases the original CNN depth, but also enhances the model's ability to capture the long-term dependence of features. The Recurrent Residual Convolution Unit (RRCU) in R2U-Net integrates the design ideas of encoding-decoding, residual connection, and recurrent convolution by replacing each convolution unit in the original U-Net network architecture with RRCU. This allows the R2U-Net network to extract better features and achieve better performance with the same model parameters. BCDU-Net, proposed by Azad et al. [22] in 2019, is another strategy for applying recurrent neural networks in U-Net networks. LSTM [23] is a special RNN, mainly to solve the problem of gradient disappearance and gradient explosion in the training process. ConvLSTM [24] is a combination of CNN and LSTM, and BCDU-Net adds a bidirectional ConvLSTM to the skip connection path to merge the corresponding feature maps of the encoding and decoding stages in a non-linear manner to produce more refined segmentation results. The author also added dense connections in the last encoding stage of U-Net to generate more diverse features to enhance the expressive ability of the model. Although the recurrent neural network can better capture the semantic information



(a)



(b)

Figure 2. Application scenario of 3D U-Net. (a) Semi-automatic segmentation. (b) Fully automatic segmentation.

in the sequence data, it is difficult to expand in parallel due to the inherent characteristics of its computing state. Therefore, RNN-based medical imaging algorithms are more suitable for segmenting lesions or organs that span multiple slices, and may be less competitive than other algorithms for segmentation tasks of a single slice. The 3D U-Net [25] network architecture is a simple extension of the original U-Net network architecture, proposed by the U-Net research team in 2016 and applied to 3D image segmentation. Because only two-dimensional slices can be displayed on the computer screen, it is difficult to mark the segmentation labels directly on the three-dimensional level. At the same time, adjacent two-dimensional slices often contain similar picture information. Based on the above two facts, the author proposes a 3D U-Net network architecture that only requires sparsely labeled two-dimensional images for training. 3D U-Net realizes the direct segmentation of 3D medical images by replacing U-Net's original 2D convolution, pooling, and up sampling operations with corresponding 3D operations, and adding the Batch normalization layer [26]. There are two ways to apply 3D U-Net network architecture for medical image segmentation. Figure 2(a) is the sparse annotation of 3D medical images, and only part of the 2D slices are annotated. 3D U-Net is trained to output dense segmentation results of 3D medical images. Fig. 2(b) shows that the three-dimensional medical image that needs to be segmented must have a similar representative training set, and then the trained 3D U-Net calculates and outputs the segmentation map on the unlabeled three-dimensional medical image.

In 2016, Milletari et al. [27] proposed the V-Net network architecture, which is another 3D implementation of the original U-Net network architecture. Compared with 3D U-Net, the biggest innovation of V-Net is that it has absorbed the idea of ResNet and added a residual learning mechanism in each stage of the network expansion path and contraction path. At the same time, inspired by the research of Spring Enberg et al. [28], V-Net uses a $2 \times 2 \times 2$ convolution kernel with a stride of 2 to replace the pooling operation in the expansion path to reduce the resolution of the feature map. In order to solve the problem of imbalance between segmentation target and background in medical images, V-Net also designed a new Dice objective function. V-Net and 3D U-Net are both end-to-end deep convolutional neural networks built directly for 3D medical imaging. The purpose is to use 3D convolution to encode from 3D, so as to well segment certain pathologies in the absence of obvious 2D Characterizing. The 3D segmentation algorithm has advantages in using the three-dimensional tissue connectivity of medical images, but it has more parameters than the 2D segmentation algorithm, and the training and reasoning process requires higher computing power of the equipment. A comprehensive summary of the main functions of the above algorithm framework is shown in Table I.

In addition, some improvements are based on the overall structure and are tabulated in Table II. On the basis of the basic codec structure, by repeatedly using the U-shaped network structure, step by step extraction of coarse and fine features for training, speed up the network convergence, and improve segmentation accuracy, for instance, Chen [29] used a bridging method to connect two U-Net structures and proposed the Bridged U-Net model. Compared with the U-Net network, this model has a higher average Dice coefficient and a more stable network performance. Some scholars have added a network structure to meet specific needs and achieve better segmentation effects. For example,

Table I. Representative algorithms in the field of medical image segmentation.

Network architecture	Advantages	Disadvantages	Main function
FCN, U-Net	The traditional classification network is transformed into a segmentation network	The upper sampling results are rough, Insensitive to image details	The size of the output result image is consistent with that of the input image
U-Net++, U-Net3+	The segmentation accuracy is improved and the amount of parameters is reduced	The training data is redundant and occupies more video memory	It is used to bridge the semantic gap between the feature map of encoder and decoder and improve the gradient flow
MultiResUNet, ResUNet++	The phenomenon of gradient explosion and gradient disappearance is slowed down	Increased network training time	Superimpose the input and output of stage to improve the training effect
R2U-Net, BCDU-Net	Better capture the information in sequence data	Parallelization is difficult to expand and is suitable for segmenting targets across multiple slices	The network can extract better features and obtain better performance with the same model parameters
Attention U-Net, ANU-Net	It flexibly captures the relationship between global features and local features, and increases the interpretability of the model	It may destroy the deep characteristic information of the network and affect the learning ability of the model	Suppress feature activation in irrelevant regions to improve the sensitivity and accuracy of the model, and achieve high segmentation accuracy without additional positioning modules
3DU-Net, V-Net	The direct segmentation of 3D medical images is realized	Too many parameters and slow segmentation speed	Convert 2D operation to 3D operation
nnU-Net, MPUNet	It can dynamically adapt to different medical image data sets	Training time is too long	Better learn the characteristics of the data itself and strengthen the robustness of the algorithm
NAS-UNet, V-NAS	The parameters and segmentation performance of the model are balanced	It is only a simple arrangement and combination with poor interpretability	Let the network automatically discover more efficient lightweight architecture
Y-Net	It can constrain the model to alleviate the over fitting problem	It is easy to be dominated by other tasks, which reduces the performance of the model	Use related tasks to improve the generalization effect of the main task model
HyperDenseNet	Maximum use of existing data	The multi-modal fusion effect and segmentation accuracy still need to be improved	Overcome the limitations of single imaging technology and use different modal data to complement each other
Cascaded 3D U-Net	It has more advantages in detecting small targets	It increases the additional computing cost and has poor real-time performance	After generating candidate regions, refine the segmentation results of the model
MS-Dual-Guided Net, CE-Net	Maximize information extraction	Artificial traces are obvious and cannot be generalized to different segmentation tasks	Capture global dependencies, restore spatial information, and maximize the use of existing features

Fu et al. [30] added a multi-scale input layer to the U-Net network structure. The side-output layer (side-output layer) constitutes the M-Net model architecture, and this network can well solve the problem of segmentation of the optic disc and the optic cup in a single-stage multi-label system.

3. PROPOSED MODEL

3.1 *The Overall Architecture of Proposed Model*

The U-Net [8] network model used in this paper is a classic encoder and decoder architecture that is proposed in 2015. It is famous for its special U-shaped symmetrical structure and has achieved good results in medical image segmentation. U-Net has also become a milestone algorithm in medical image segmentation with its excellent performance. Of course, U-Net also has its own limitations, this is also the reason why many improved U-Net algorithms will emerge later. The limitations of U-Net network mainly lie in: (1) Due to the shallow depth of U-Net and the limit of feature extraction, the network cannot extract some small lesion targets. (2) The high complexity of different parts of medical images brings difficulties to U-Net. To capture the geometric transformation of the lesion area, a modulated deformable convolution block is integrated in the proposed model. In the process of network training, the deformable convolution kernel can adjust its shape according to the proportion and shape of the lesion object, so that the complex skin lesion structure can also be well detected. As shown in Fig. 1, the left side is the encoder part, the right side is the decoder part; the blue arrow in the middle indicates the skip connection that can splice the features of different scales in encoder part and their corresponding scales in decoder part to enhance the sampling amount of convolution operation. The modulation deformable convolution is added to the convolution operations of the second and third

Table II. Summary of corresponding improvement measures for different purposes.

Purpose	Improvement measure	Associated network
A. Avoiding overfitting	Multiple types of dense connections; Multiple technologies of data enhancement;	MDU-Net;
B. Reducing the number of parameters	Inception module; The global pooling layer; Deeply supervision; Dense skip connection; Full-scaled skip connection; Skip connection using addition	MultiResUNet; BSU-Net; GP-Unet; DENSE-Inception U-net; Unet++; Unet3+; LadderNet; Bridged U-net;
C. Focusing on effective features and suppressing irrelevant features	Attention module; SE module;	Attention U-Net; AnatomyNet; RA-Unet; ANU-Net;
D. Enhancing feature fusion	Residual module; Dense module; Dense skip connection; Full-scaled skip connection; Feature pyramid; Bidirectional feature network;	Vnet; MDU-Net; FD-Unet; Bridged U-net; Dense Multi-path U-Net; Unet++; Unet3+; RA-Unet; DPSN; U-Det;
E. Speeding up convergence	Residual module; Dense skip connection; Full-scaled skip connection; Bridged U-Net; Data normalization;	Vnet; GP-Unet; Bridged U-net; Unet++; nnUNet;
F. Enlarging receptive field	Deformable convolution; Dilated convolution;	DU-Net (Deformable U-Net); DMFNet; BSU-Net; 3D-HDC-Unet;
G. Avoiding gradient vanishing or gradient explosion	Residuals module; Attention residuals module;	BSU-Net; RA-Unet; DENSE-Inception U-net;

layers to form the modulation deformable convolution block. Each modulation deformable convolution block is composed of a convolution offset layer, a convolution layer, a batch normalization layer and an excitation layer (ReLU in this paper). In each layer, there are two convolutions, batch normalization, and down sampling path. In each encoder step, the number of feature channels is doubled until the end of the fourth down sampling, and then the high-dimensional feature extraction vector representations are obtained through multiple multi-step learning process. The obtained feature extraction vectors are sent to our multi-scale fusion extraction module. In the decoder part, we aim to recover the high-level semantic features extracted by the encoder part by continuously up sampling different scale feature vectors and changing the number of channels of feature maps in each convolution operation. We use skip connection to obtain more detailed information from different stages in the encoder part, so as to make up for the information loss due to continuous pooling and striding revolutionary operations. In the decoding process, two operations are usually used: up sampling and deconvolution. The up-sampling operation increases the image size with linear interpolation, and the deconvolution is to fill the remaining position with 0, and then the large feature map undergoes a convolution operation, which uses convolution to enlarge the image to some fixed size at that time.

3.2 *Multi-scale Feature Fusion*

The perception network extracts the information of different scales of the image through multiple convolution kernels, and finally fuses it to obtain a better feature representation of the image. At the same time, the dilated convolution can make convolution operations obtain a larger receptive field under the same feature map, so as to obtain dense and effective feature representations. Under the same feature map, the larger receptive field can improve the effect of a model when recognizing small object recognition and segmentation in the tasks of target detection and semantic segmentation. Atrus spatial pyramid polling (ASPP) was originally proposed in deeplabv2 [31] as an idea of multi-scale information fusion. Larger receptive fields can help a model extract and generate more abstract and accurate features for large targets, while smaller receptive fields are fit for small targets. By combining dilated convolution with different expansion rates, the features of targets with various sizes can be extracted. Deeplabv3 [32] further discusses dilated convolution, which is a powerful tool in the semantic segmentation task: it can adjust the filter field of view and control the feature response resolution calculated by convolution neural network. In order to solve the problem of target segmentation in multi-scale, we designed dilated convolution cascade or dilated convolution parallel architecture with different sampling rates. In addition, we emphasize the ASPP (atrous spatial pyramid pooling) module. The BN layer is used in the ASPP. The Batch normalization (BN) layer is applied in the ASPP. With the increase of sampling rate, the effective weight in the filter decreases (the effective weight decreases and it is difficult to capture the original distance information, which requires reasonable control when setting different sampling rates), Using the global average pooling of the last feature map of the model (in order to overcome the problem that reduces the effective weight at a long distance), the module can obtain convolution features at multiple scales to further improve the performance of medical image segmentation. At the same time, we shared the implementation details and training methods. Compared with the previous version, the proposed deeplabv3 has significantly improved the effect

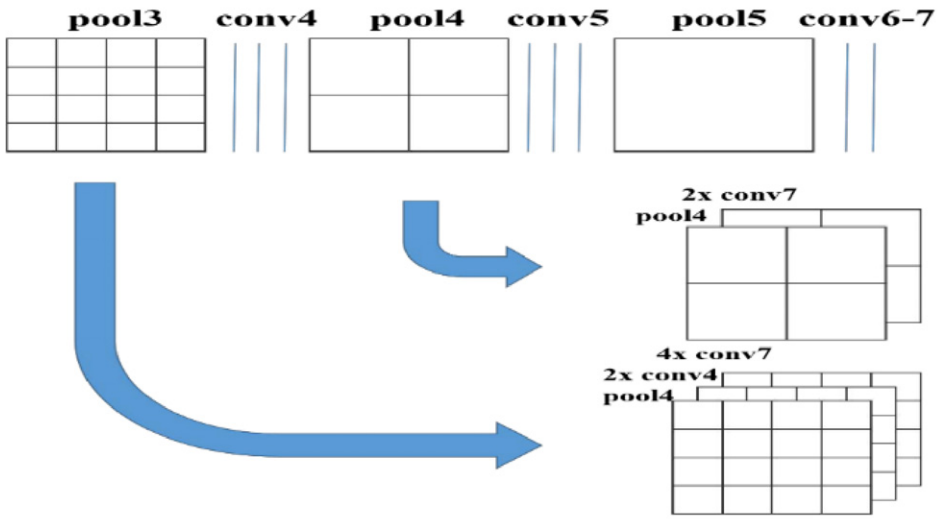


Figure 3. Multichannel schematic in up sampling of FCN [35].

and achieved advanced performance on Pascal VOC 2012. The DRNFnet network proposed by Chen et al. [33] adopts expansion convolution, and forms expansion fibers through three three-dimensional expansion convolution branches with different expansion rates to establish multi-scale feature representation, so that the network can actively learn the most valuable information in different sensory fields, so as to improve the segmentation accuracy. One of the challenges of medical image segmentation is the huge change of target size; For example, intermediate or advanced tumors may be much larger than early tumors. According to this type of problem, we adopt the operation of multi branch set. It mainly relies on multiple effective fields of view to detect objects of different sizes. The size of the perception field in convolution operation determines the amount of information we can use. Aggregation operations of different sizes produce perceptual fields of different sizes. At present, most networks use end-to-end multi-channel combination to solve the multi-scale problem. For example, the FCN structure in Figure 3 uses three scales in the subsequent upper sampling part to alleviate the shortage of detail segmentation. By connecting different depth feature maps, U-Net enables the network to combine feature maps of different sizes. MCNN [34] directly uses different scales of data size in the input to deal with the detailed features. These methods of fusing multi-detail features are mainly independent channels, which reduces and weakens the connection between features.

The multi-scale convolution kernel module proposed in this paper, more specifically, there are multiple columns of parallel serial convolutions with different kernel sizes, two convolution kernels of the same size are connected in series in each column, and the output results are fused to the next layer of the network by concatenating operation. Multi-scale feature extraction on a specific feature map not only ensures the diversity of features, but also makes the relationship between features closer and hierarchical.

The internal structure of presented multi-scale module (MSM) is shown in Figure 4. The kernel sizes of the three column convolutions are respectively 1×1, 3×3 and 5×5. A variety of convolution kernel sizes are used to extract image features of different scales. In order to obtain more effective feature representations from the same input feature graph, we use different convolution kernels. 1×1 convolution kernel adds nonlinear feature learning and representation without changing the size. 3×3 and 5×5 can further increase the receptive field of convolution operation on the input characteristic map.

3.3 *The Internal Structure of Proposed Model*

The proposed model is also an encoder and decoder structure and the overview architecture of proposed model shown in Figure 5. Combined with the medical segmentation task, it mainly presents the following characteristics: (1) the number of channels used by our model in the last stage of encoder part is only 128, which is far less than the number of 1024 channels used by the original U-Net model in the last stage of encoder part; (2) There are only 3×3 convolution operations in the original U-Net model, while 1×1, 3×3 and 5×5 convolution operations are used in the proposed model; (3) Moreover, in order to obtain and make use of the advantages of multi-scale, the multi-scale fusion module is adopted in this model.

In each layer of U-Net network, there are usually only two ordinary 3×3 convolutions that cannot effectively extract features of different scales and are easy to lead to some local small features instead of learning feature representation from are not easy to be used to learn in a large range. Inspired by the paper of MCNN [36–38], head size counting detection is carried out for dense people with different distances and proximity. In this paper, we use multiscale convolution blocks with different kernel sizes to replace two series 3×3 convolutions in each block of standard U-Net model. At the same time, the symmetrical U-shaped structure is used to cut

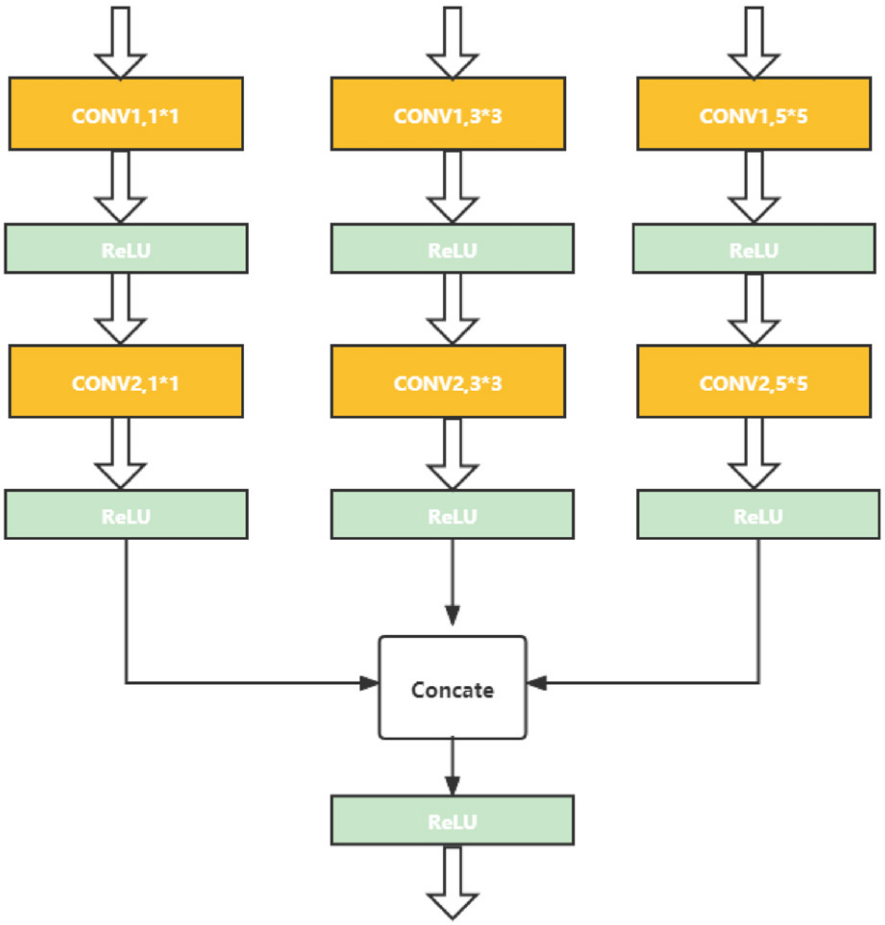


Figure 4. The internal structure of multi-scale module (MSM).

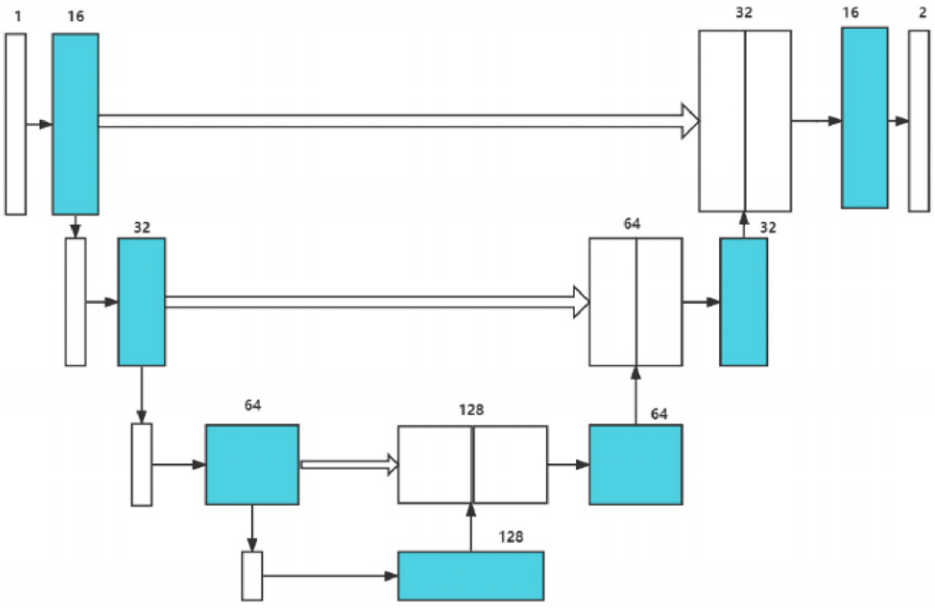
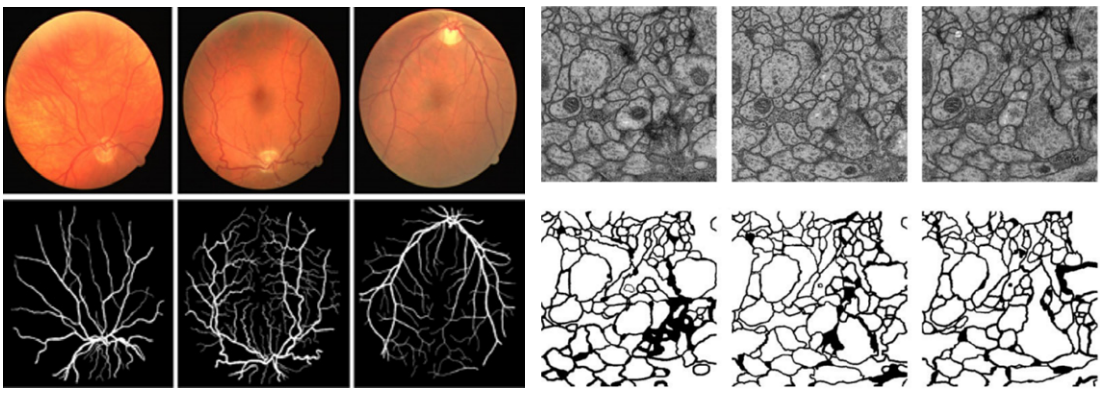


Figure 5. The overview architecture of proposed model (MS-UNet). The blue module indicates the multi-scale module that consists of three parallel and serial convolution operations with different kernel sizes.



(a) (b)

Figure 6. Dataset sampling.

Table III. Segmentation algorithm encoder of MS-UNet.

MSM	Kernel	Stride	Padding	Filter-number
	–	–	–	–
Conv1_1,1_2	1×1	1	–	16
Conv2_1,2_2	3×3	1	1	16
Conv3-1,3-2	5×5	1	2	16
Maxpooling	2×2	2	–	16
Conv4_1,4_2	1×1	1	–	32
Conv5_1,5_2	3×3	1	1	32
Conv6_1,6_2	5×5	1	2	32
Maxpooling	2×2	2	–	32
Conv7_1,7_2	1×1	1	–	64
Conv8_1,8_2	3×3	1	1	64
Conv9_1,9_2	5×5	1	2	64
Maxpooling	2×2	2	–	64
Conv10_1,10_2	1×1	1	–	128
Conv11_1,11_2	3×3	1	1	128
Conv12_1,12_2	5×5	1	2	128

and splice the features of the encoder and decoder of the same layer with down sampling and up sampling, and transpose convolution is used for up sampling.

In order to ensure that the convolution output features of three columns with different sizes in MS modules that are fused by add operations. The feature map is padded after each convolution operation in the multi-scale convolution core module to maintain the input size. The padding size and other parameters are shown in Table III.

The effectiveness of the U-shaped segmentation network model has been verified in various segmentation tasks, especially in the application of medical images. Many subsequent studies have improved or added more effective components based on this structure. In order to fully illustrate the effectiveness and reliability of multi-scale feature extraction, another coding strategy is used in this paper.

4. EXPERIMENTS

4.1 *Dataset*

The DRIVE database [39] is derived from the diabetes retinopathy screening project in Holland. It is used to create a comparative study of blood vessel segmentation in retinal images. A total of 400 people of different ages from 25 to 90 participated. Among them, the signs of retinopathy in each picture are marked in varying degrees. Of the 40 pictures selected, 20 were used for training and 20 for testing. For training images, a single manual segmentation of the vascular system can be performed. For test cases, two manual segments can be used; one is used as the gold standard, and the other can be used to compare computer-generated segmentation with that of independent human observers. In addition, a mask image may be used for each retinal image to indicate a region of interest. All human observers who manually segment the vascular system are guided and trained by experienced ophthalmologists. They were asked to label all pixels, and at least 70% were convinced that they were blood vessels. The results shall be submitted as a zip file containing a binary png file containing the vascular predictions of the corresponding files in the test set. The images in the dataset are shown in the following Figure 6, the first row is original images, and the second row lists corresponding ground truth.

The International Symposium on biomedical imaging (ISBI) challenge training data set is composed of continuous slice images that collected from transmission electron microscopy of the ventral nerve cord (VNC) of the first instar larvae of Drosophila melanogaster, and the resolution size of each image is 512 × 512.

4.2 *Data Enhancement*

This part of research uses the standard data enhancement technology to increase the performance of proposed model by image rotation (Figure 7), scaling (Figure 8), flipping (Figure 9)and combined data enhancement (Figure 10). In

Figure 7. Image rotation.

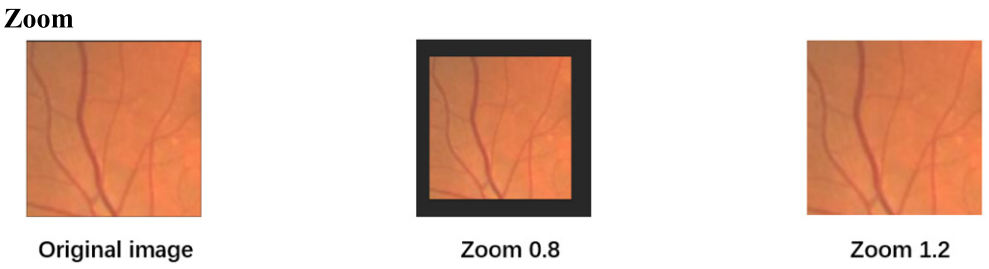


Figure 8. Image zoom.

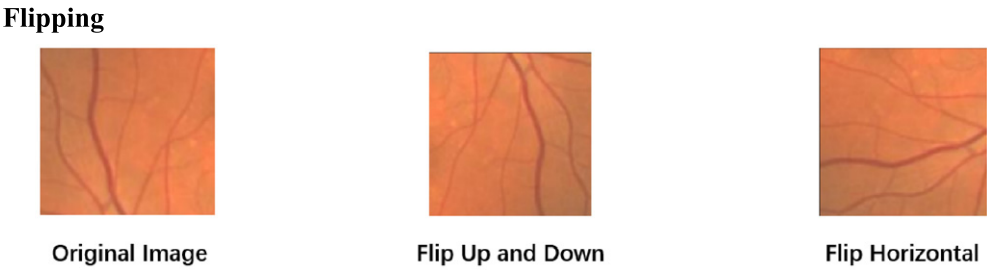


Figure 9. Image flipping.

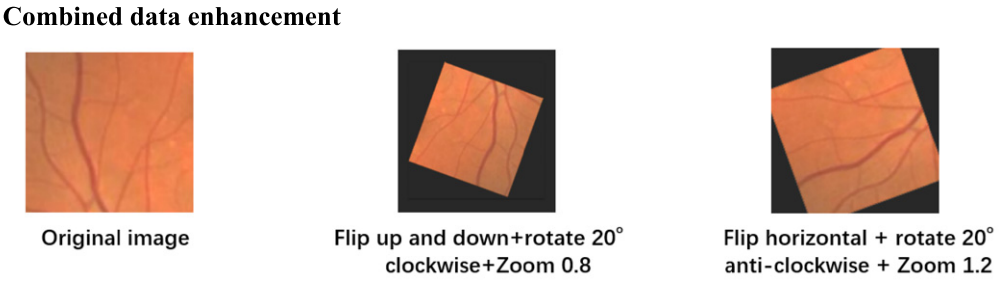


Figure 10. The combinations of different enhancement operations.

addition, in order to speed up the training of the network, the mean and pixel values of the channels in the image data are normalized. To help understand the diversity of the train data and the robustness of proposed model, we briefly introduce the data enhancement methods: rotation, scaling, flipping and different combinations of them.

We hope that the enhanced data set is a sub distribution of the original overall sample distribution, and at least their intersection contains the original data set. In this regard, we may slightly expand the sample distribution space of the original data set, so that the network captures some additional information. We call the hypothesis that the enhanced data set contains the original data set as the enhanced assurance hypothesis, that is, the network trained under the enhanced assurance hypothesis will improve the generalization ability. For the drive data set, this paper directly uses the method of random clipping to expand the data, and randomly cuts each picture in the training set to 5,000 150 $\times$ 150 local blocks, 3,000 for training and 2,000 for verification. For example, Figure 11 shows the details after enhancing the original drive dataset.

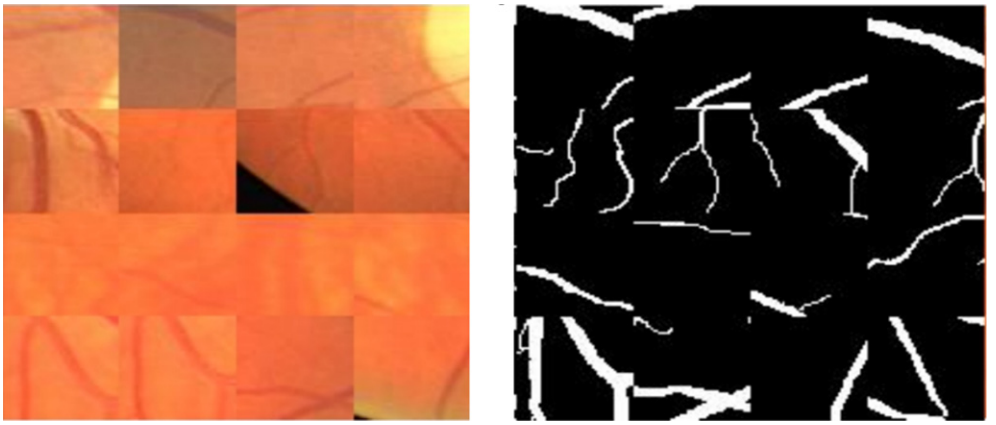


Figure 11. Detail display from data enhancement.

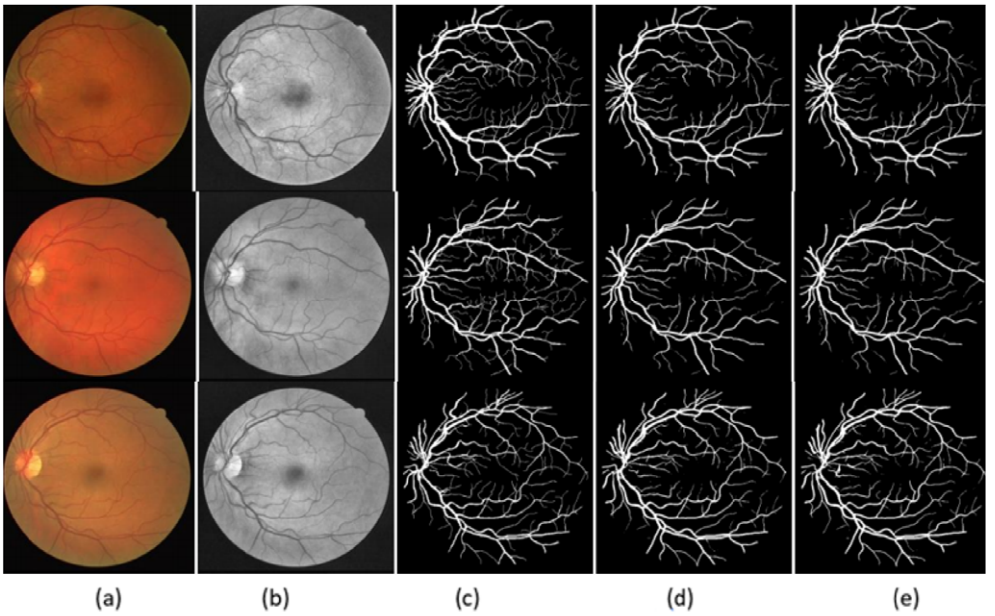


Figure 12. (a) Original images. (b) Pre-processing. (c) Label. (d) U-Net. (e) Proposed model.

4.3 *Implementation Details*

The detailed information of the experimental platform are window10 and Ubuntu 16 04, 3.6 GHz Intel Core i7-8700k, NVIDIA GEFORCE GTX 1080ti 11G, and RAM 32GB. The programming language of the experiment is Python 3 6. The implementation is based on the PyToch framework (0.4.0). The optimizer selects the random gradient descent method (SGD), the momentum is 0.9, the learning rate is set to 0.00001, batch size is set 5, and iterates are 100 times.

4.4 *Metrics*

Retinal blood vessel segmentation is to divide the pixels in the color retinal image into blood vessels and background. In order to judge whether the retinal vessel segmentation method is effective, this paper adopts five commonly used indexes to measure the performance of the algorithm: accuracy (ACC), sensitivity (SN), specificity (SP), F1 (F1 score) and Jaccard index. Their calculation formulas are respectively shown from formula 1 to formula 5, where TP (true positive) represents the correctly classified vascular pixels, TN (true negative) represents the correctly classified background pixels, and FP (false positive) represents the number of background pixels divided into vascular pixels; FN (false negative) indicates the number of blood vessel pixels divided into background pixels.

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{1}$$

$$\text{Sensitivity (Sn)} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{2}$$

$$\text{F1-score (F1)} = \frac{2\times\text{TP}}{2\times\text{TP}+\text{FP}+\text{FN}} \tag{3}$$

$$\text{Specificity (Sp)} = \frac{\text{TN}}{\text{TN}+\text{FP}} \tag{4}$$

$$\text{Jaccard Index} = \frac{\text{TP}}{\text{TP}+\text{FP}+\text{FN}} \tag{5}$$

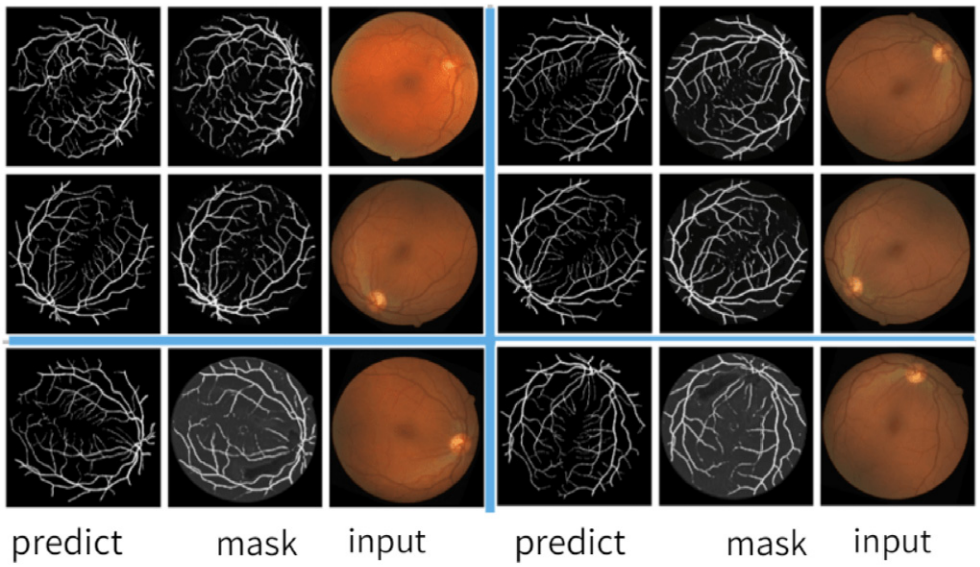


Figure 13. Comparison of mask segmentation on drive dataset.

4.5 *Results and Analysis*

We test and compare our model with baseline (U-Net) and other state-of-the-art models on the drive dataset, the retinal vessel segmentation results are shown in Figure 12. The original images are in the first column, the labels are in the second column, the results of original U-Net are shown in the third column, and the results of proposed model are displayed in the last column. It can be seen from the segmentation results that the results of proposed model better than the original U-Net model, the details are as follows: (1) compared with the original image, the preprocessing has clearly filtered some noise. (2) The detection results of this model are clearer than those of U-Net, such as the prediction results of this model in the lower half of the blood vessel in columns 4 and 5 of the first row. (3) The results of model detection show more details than that of U-Net, especially in the peripheral parts of blood vessels. The results predicted by this model are extended, such as columns 4 and 5 in the second row. (4) The prediction result of the model is more comprehensive than that of U-Net. The fifth column of the third row predicts the contents that are not predicted by U-Net and consistent with the label. In addition, to further prove the effectiveness of proposed model, we still test it and compare the results with corresponding masks that can ensure that the model selects the best region of interest of the target, they are shown in Figure 13.

The experimental results are shown in Table IV. On the drive test set, the proposed model achieves high performance. The ACC value is 0.9693, the Sn value is 0.7456, the SP value is 0.9891, the F1 value is 0.8264 and the Jaccard value is 0.7621. Compared with other algorithms, most of the values from the proposed model are better than other algorithms. Additionally, to further illustrate the effectiveness of proposed model, we test our model and other models on the International Symposium on biomedical imaging (ISBI) challenge dataset (Figure 14), the detailed values are listed in Table V. On the ISBI test set, the ACC value is 0.9356, Sn value is 0.9523, SP value is 0.9321, F1 value is 0.9652 and Jaccard value is 0.9132.

Table IV. Performance metrics for different algorithms in DRIVE test database.

Model	Acc	Sn/ Recall	F1	Sp	Jaccard
U-Net	0.9632	0.7911	0.8186	0.9861	0.7434
U-Net+Ags	0.9652	0.7966	0.8189	0.9863	0.7523
MultiResUNet	0.9534	0.7432	0.7812	0.9798	0.7322
Liskowski [33]	0.9645	0.7864	0.8064	0.9641	0.7552
Proposed method	0.9693	0.7456	0.8264	0.9891	0.7621

Table V. Comparison of set metrics of different improved algorithms on ISBI dataset.

Model	Acc	Sn/ Recall	F1	Sp	Jaccard
U-Net	0.9112	0.8432	0.9159	0.9032	0.9057
U-Net+Ags	0.9136	0.9056	0.9235	0.9289	0.9094
MultiResUNet	0.9322	0.9189	0.9532	0.9148	0.9103
Liskowski [33]	0.9295	0.9645	0.9423	0.9156	0.9084
Proposed method	**0.9356**	0.9523	**0.9652**	**0.9321**	**0.9132**

The possible reasons why our model can achieve better performance than other models are as follows: (1) different from other models which only use one fixed convolution kernel size, e.g. 3×3 convolution kernel size. In our model, 1×1, 3×3 and 5×5 convolution kernel sizes are used to help proposed model expand its perceptional ranges when learning feature representations of the targets (Eyeball and blood vessels in our task). (2) We use multi-scale fusion in our model to ensure rich and effective feature information

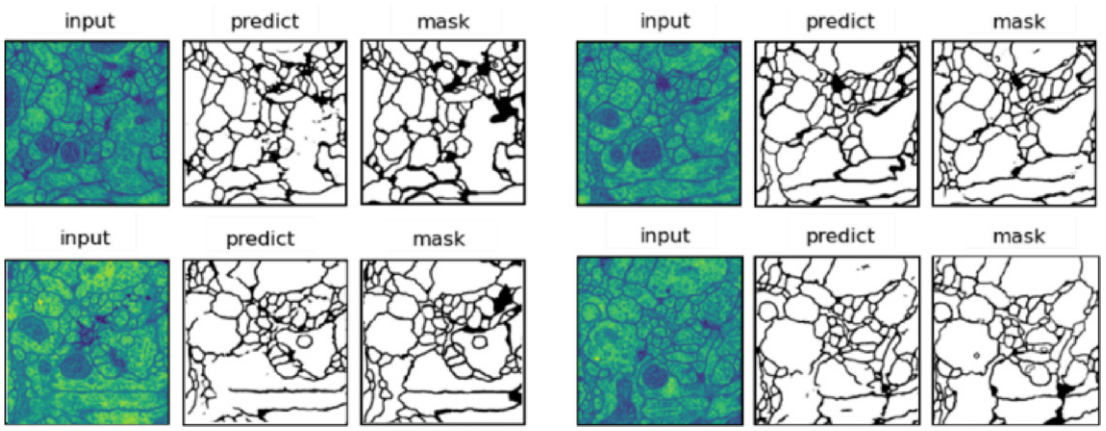


Figure 14. Comparison of mask segmentation on ISBI dataset.

can be learned for each step. Subsequently, the concatenation operation further make our model obtain enough and effective samples when our model begins to learn feature information.

5. CONCLUSION

Medical segmentation is very important to help doctors diagnose. In recent years, the technology based on deep learning has been widely used in medicine. Among them, U-net model has important applications in medicine and is often used as the backbone model for learning and segmenting medical images. In this paper, aiming at the disadvantages of large U-net model with too many parameters and small receptive field, combined with the characteristics of eyeball and blood vessel segmentation in medicine, a multi-scale network model is proposed. Different convolution kernels can help the model obtain more visual fields to extract the target features, and multi-scale can make the model learn effective target feature information from multiple angles at the same stage as much as possible. We conduct qualitative and quantitative analysis on drive and ISBI data sets respectively, and the experimental results show the effectiveness of the model. Our future work will attempt solve the shortage of medical image data samples and further improve the accuracy of medical image segmentation from the perspective of small sample learning and antagonistic generation neural network (GAN).

ACKNOWLEDGMENT

This work was supported in part by the Science and Technology Key Project of Henan province under Grant no. 212102210520.

REFERENCES

[1] W. Mengli, "Clinical application of medical imaging technology in medical imaging diagnosis," Imaging Res. Med. Appl. **26**, 146–147 (2017).

[2] *AI+ Healthcare: The Application and Future of Intelligent Healthcare*, edited by M. Dong, W. Yu, X. Yan and F. Lin (Machinery Industry Press, Beijing, 2018).

[3] J. Zhao, P. Gao, Y. Zhu, and C. Law, "Application of artificial intelligence will improve the effectiveness of primary health care services in China," *Chinese Primary Care Medicine* (2017), pp. 6–10.

[4] Y. Lecun, Y. Bengio, and G. E. Hinton, "Deep learning," Nature **521**, 436–444 (2015).

[5] F. Y. Zhou, L. P. Jin, and J. Dong, "Review of convolutional neural network," Chin. J. Comput. **40**, 1229–1251 (2017).

[6] Y. B. Feng, D. W. Qiu, H. Cao, J. Zhang, Z. Xin, and J. Liu, "Research on coronavirus disease 2019 (COVID-19) detection method based on depthwise separable DenseNet in chest X-ray images," J. Biomed. Eng. **37**, 557–565 (2020).

[7] B. Shuai, T. Liu, and G. Wang, Improving fully convolution network for semantic segmentation. arXiv:611.08986, 2016.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Lecture Notes in Computer Science (Springer, Cham, 2015), Vol. 9351.

[9] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: Achievements and challenges," J. Digit. Imaging **32**, 582–596 (2019).

[10] S. Wang, M. Zhou, Z. Liu, Z. Liu, D. Gu, Y. Zang, D. Dong, O. Gevaert, and J. Tian, "Central focused convolutional neural networks: Developing a data-driven model for lung nodule segmentation," Med. Image Anal. **40**, 172–183 (2017).

[11] V. Cherukuri, P. Ssenyonga, B.C. Warf, A.V. Kulkarni, V. Monga, and S.J. Schiff, "Learning based segmentation of ct brain images: Application to postoperative hydrocephalic scans," IEEE Trans. Biomed. Eng. **65**, 1871–1884 (2018).

[12] P. Liskowski and K. Krawiec, "Segmenting retinal blood vessels with deep neural networks," IEEE Trans. Med. Imaging **35**, 2369–2380 (2016).

[13] S. Minaee, Y.Y. Boykov, F. Porikli, A.J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," IEEE Trans. Pattern Anal. Mach. Intell. (2021).

[14] A. Zhao, G. Balakrishnan, F. Durand, J.V. Guttag, and A.V. Dalca, "Data augmentation using learned transformations for one-shot medical image segmentation," *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Piscataway, NJ, 2019), pp. 8543–8553.

[15] Y. Gordienko, P. Gang, J. Hui, W. Zeng, Y. Kochura, O. Alienin, O. Rokovyi, and S. Stirenko, "Deep learning with lung segmentation and bone shadow exclusion techniques for chest X-ray analysis of lung cancer," *Int'l. Conf. on Computer Science, Engineering and Education Applications* (Springer, Cham, 2018), pp. 638–647.

[16] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and K. Keutzer, Densenet: implementing efficient convnet descriptor pyramids. arXiv:1404.1869, 2014.

[17] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: a nested u-net architecture for medical image segmentation," *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (Springer, Cham, 2018), pp. 3–11.

[18] P. Zhang, X. N. Xu, and H. W. Wang, "Computer-aided lung cancer diagnosis approaches based on deep learning," J. Comput.-Aided De. Comput. Graph. **30**, 90–99 (2018).

[19] 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," *IEEE Int'l. Conf. on Acoustics, Speech and Signal Processing* (IEEE, Piscataway, NJ, 2020), pp. 1055–1059.
[20] M. Z. Alom, M. Hasan, C. Yakopcic, T.M. Taha, and V.K. Asari, Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. arXiv:1802.06955, (2018).
[21] M. Liang and X. Hu, "Recurrent convolutional neural network for object recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2015), pp. 3367–3375.
[22] R. Azad, M. Asadi-Aghbolaghi, M. Fathy, and S. Escalera, "Bi-directional convLSTM U-net with densley connected convolutions," *Proc. IEEE Int'l. Conf. on Computer Vision Workshops* (IEEE, Piscataway, NJ, 2019).
[23] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: a search space odyssey," IEEE Trans. Neural **28**, 2222–2232.
[24] H. Song, W. Wang, S. Zhao, J. Shen, and K.M. Lam, "Pyramid Dilated Deeper ConvLSTM for Video Salient Object Detection," *Proc. European Conference on Computer Vision* (Springer, Cham, 2018), pp. 715–731.
[25] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: learning dense volumetric segmentation from sparse annotation," *Int'l. Conf. on Medical Image Computing and Computer-Assisted Intervention* (Springer, Cham, 2016), pp. 424–432.
[26] S. Ioffe and C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167, 2015.
[27] F. Milletari, N. Navab, and S. A. Ahmadi, "V-net: fully convolutional neural networks for volumetric medical image segmentation," *2016 Fourth Int'l. Conf. on 3D Vision (3DV)* (IEEE, Piscataway, NJ, 2016), pp. 565–571.
[28] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, Striving for simplicity: the all convolutional net. arXiv:1412.6806, (2014).
[29] W. L. Chen, Y. Zhang, J. J. He, Y. Qiao, Y. Chen, H. Shi, E.X. Wu, and X. Tang, "Prostate segmentation using 2D Bridged U-net," *2019 Int'l. Joint Conf. on Neural Networks (IJCNN)* (IEEE, Piscataway, NJ, 2019), pp. 1–7.
[30] H. Z. Fu, J. Cheng, Y. W. Xu, D.W.K. Wong, J. Liu, and X. Cao, "Joint optic disc and cup segmentation based on multi-label deep network and polar transformation," IEEE Trans. Med. Imaging **37**, 1597–1605 (2018).
[31] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A.L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," IEEE Trans. Pattern Anal. Mach. Intell. **40**, 834–848 (2018).
[32] H. Wu, J. Zhang, K. Huang, K. Liang, and Y. Yu, "Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation." arXiv preprint arXiv:1903.11816 (2019).
[33] C. Chen, X. P. Liu, M. Ding, J. Zheng, and J. Li, "3D dilated multi-fiber network for real-time brain tumor segmentation in MRI," *International Conference on Medical Image Computing and Computer-Assisted Intervention MICCAI 2019*, Lecture Notes in Computer Science (Springer, Cham, 2019), Vol 11766.
[34] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," *Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2016), pp. 589–597.
[35] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2015), pp. 3431–3440.
[36] Y. Lecun and L. Bottou, "Gradient-based learning applied to document recognition," Proc. IEEE **86**, 2278–2324 (1998).
[37] J. Zhang, C. Li, S. Kosov, M. Grzegorzek, K. Shirahama, T. Jiang, C. Sun, Z. Li, and H. Li, "LCU-Net: A novel low-cost U-Net for environmental microorganism image segmentation," Pattern Recognit. **115**, 107885 (2021).
[38] C. Sun, C. Li, J. Zhang, M.M. Rahaman, S. Ai, H. Chen, F. Kulwa, Y. Li, X. Li, and T. Jiang, "Gastric histopathology image segmentation using a hierarchical conditional random field," Biocybernetics Biomed. Eng. **40**, 1535–1555 (2020).
[39] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation." arXiv preprint arXiv:1706.05587 (2017).