Machine Vision and Applications (2023) 34:87
https://doi.org/10.1007/s00138-023-01442-x

ORIGINAL PAPER

A dual-path U-Net for pulmonary vessel segmentation method based on lightweight 3D attention

Rencheng Wu[1] · Yu Xin[1] · Yihong Dong[1] · Jiangbo Qian[1]

Received: 10 March 2022 / Revised: 22 April 2023 / Accepted: 3 August 2023 / Published online: 21 August 2023

Abstract
In recent years, pulmonary vessel segmentation has aroused widespread interest in medical image analysis. However, the structure of pulmonary vessels is complex and CT images have a lot of noise. Therefore, it is a difficult task to extract pulmonary vessels accurately. In terms of pulmonary vessel segmentation, medical image segmentation methods based on deep learning only utilize single volume data in CT image, but cannot fully fuse multiple volume data, resulting in the accuracy of pulmonary vessel segmentation is low. In order to fully utilize the complementary advantages of multiple volume data, we propose a DS-ResUNet to segment pulmonary vessels in multi-view. The DS-ResUNet uses feature fusion module to fuse 2 mm and 5 mm volume data and makes full use of the detailed vessel textures in 2 mm volume data and coarse vessel structures in 5 mm volume data. By this multi-view fusion module, the accuracy of pulmonary vessel segmentation can be effectively improved. In addition, in order to strengthen the spatial weight of vessels and reduce the model parameters, we design a lightweight 3D axial attention module by separable convolution. To confirm the improved performance, we design some comparison experiments with the state-of-the-art segmentation methods. As a result, our DS-ResUNet has a better performance than other state-of-the-art methods on pulmonary vessel segmentation, but also has fewer parameters.

Keywords Pulmonary vessel segmentation · Deep learning · Multi-view fusion · Separable convolution

1 Introduction

In recent years, lung tissue disease has gradually become one of the diseases with the highest morbidity and mortality. The traditional examination of lung lesions is very invasive to patients and can only reflect the state of the whole lung. It is difficult to detect local lesions and not conducive to early diagnosis and treatment. With the continuous development of computed tomography (CT), it has critical clinical value at the diagnosis of vessel diseases, pulmonary hypertension and tumor. Therefore, CT has become the main tool on lung diseases diagnosis. However, due to the complexity of lung tissues, organs, vessels and nerves, CT images reading is time-consuming and inefficient manually. Therefore, 3D reconstruction technology can be used to obtain intuitive and clear pulmonary vessels structure, which can avoid human error and has high efficiency. This technology can provide auxiliary diagnostic functions such as surgical planning, postoperative effect prediction in the diagnosis and treatment of pulmonary nodules, pulmonary embolism and lung tumors.

The 3D reconstruction of pulmonary vessels requires the first labeling of vessels, and then, the labeling results can be converted into 3D view. At present, in the process of pulmonary vessels labeling, the amount of data is large, and the pulmonary vessels labeling usually needs to be completed manually by professional doctors. Therefore, pulmonary vessels labeling is time-consuming and error-prone. Compared with manual labeling, AI-based pulmonary vessel labeling has advantages of low cost, high accuracy and efficiency. The AI technology on CT diagnosis is the trend of medical image processing and medical image segmentation.

✉ Yu Xin
xinyu@nbu.edu.cn

Rencheng Wu
710528275@qq.com

Yihong Dong
dongyihong@nbu.edu.cn

Jiangbo Qian
qianjiangbo@nbu.edu.cn

[1] Ningbo University, 818 Fenghua Road, Ningbo 315211, China

In terms of medical image segmentation, the traditional medical image segmentation techniques mainly include thresholding [1, 2], region growing [3–5], watershed [6] and so on. These methods cannot deal with the complex nonlinear relationship in medical image, but also are time-consuming and inaccurate. With the significant improvement of computing capability, deep learning has become one of the cutting-edge technologies of AI. Its ability of complex nonlinear relational data processing can fully solve the issues of inefficiency and poor accuracy of traditional medical image segmentation technology. Consequently, more and more researchers have begun to use deep learning techniques to solve the problem of pulmonary vessel segmentation.

Currently, medical image segmentation methods based on deep learning are mainly based on FCN [7] and U-Net [8]. Such methods need to conduct down-sampling and pooling of images in feature extraction. By down-sampling and pooling images would lose key information, resulting in the mis-segmentation of tiny vessels. Due to the small image area and similar morphology of lung tissues, nodules and vessels, down-sampling and pooling will seriously affect the accuracy of vessel segmentation.

There are many kinds of volume data in lung CT images, such as 2 mm volume data and 5 mm volume data. The 2 mm volume data contains a lot of detailed pulmonary vessels, while the 5 mm volume data has clear vessels texture and less image noise. In addition, the 3D positions of pulmonary vessels in lung CT data are relatively fixed; 3D spatial attention mechanism should be adopted to enhance the spatial impact of vessels. Therefore, in order to improve the accuracy and efficiency of pulmonary vessel segmentation, it is necessary to establish a fusion module of 2 mm and 5 mm volume data and simplify the model parameters of 3D spatial attention mechanism. In this study, a dual-path U-Net for pulmonary vessel segmentation based on lightweight 3D attention is proposed. The main contributions are summarized as follows:

1. We propose a 3D fusion module based on 2 mm and 5 mm volume data which is designed to solve the problem of tiny vessel losing and mis-segmentation of vessel-like regions caused by down-sampling and pooling.
2. The optimized network structure introduces a lightweight 3D spatial attention module which is designed to decrease the massive parameters and improve the efficiency of the 3D attention model.
3. Extensive experimental results demonstrate that DS-ResUNet consistently improves the performance of previous state-of-the-art approaches.

2 Related work

Image segmentation is one of the key technologies in medical image processing. Image segmentation technology has experienced the stage from manual participation to semi-automatic image segmentation. The classical methods such as thresholding [1, 2], region growing [3–5] and watershed [6] are proposed. Such methods are difficult to meet the requirements of high efficiency and accuracy in the medical field due to the poor ability of nonlinear data processing. With the development of deep learning technology, image processing technology based on deep learning can significantly improve the performance of image segmentation and has better effects on complex data processing. In addition, compared with the traditional AI algorithm, the deep learning model has a complex model structure, which can extract more potential features and greatly improve the efficiency and accuracy of image segmentation. Therefore, image segmentation based on deep learning is the trend of medical image processing in future.

In terms of traditional medical image segmentation methods, the main methods include thresholding [1, 2], region growing [3–5], watershed [6] and so on. The thresholding method is a widely used and efficient segmentation method. The commonly used threshold segmentation methods include minimum error method, maximum category difference method and maximum entropy method [2]. The thresholding method is sensitive to noise, and its segmentation effect is poor when the gray difference is small or segmentation targets are overlapping. Therefore, thresholding-based methods are not suitable for pulmonary vessel segmentation. The region growing method clusters the target regions according to prior parameters. This method requires manual participation in the selection of seed points, and the effect is not well when the vessels and other organs are highly similar. Accordingly, the practicability of this method is poor. The watershed method takes the similarity of adjacent pixels as the measurement of target space clustering. The watershed method is sensitive to noise and easy to segment vessel-like regions. In summary, the traditional medical image segmentation methods mostly require manual participation, the segmentation accuracy is poor and the efficiency is low.

In terms of medical image segmentation based on machine learning, the main methods include support vector machine (SVM) [9, 10], Clustering method [11, 12] and AdaBoost method [13, 14]. SVM is a binary classification model, which aims to find the hyperplane of sample segmentation. The accuracy of the SVM-based methods is low. Clustering method is represented by K-means clustering method and

fuzzy clustering method. K-means method [12] is a classical partition-based clustering method, which selects k points as cluster centers initially. Then, optimize the clusters formed by the k cluster centers iteratively, until the best clustering results are obtained. Fuzzy clustering [11] mainly obtains the similarity of each sample point to all clustering centers by optimizing the objective function, so as to realize the automatic classification of pixels. AdaBoost is an iterative method, whose core idea is to train different classifiers for the same training set. The trained classifiers are weak classifiers. AdaBoost combines these weak classifiers into a stronger final classifier. The methods above need to manually design feature models for the segmentation target. The designed feature models have many prior parameters with idealized conditions. Practically, these feature models will lead to a serious decline on image segmentation when the images exist large differences.

With the development of AI technology, deep learning has replaced traditional machine learning methods and become the cutting-edge technology of AI. Currently, the image segmentation based on deep learning has surpassed the traditional segmentation methods in segmentation speed and accuracy and is widely used in the field of medical image segmentation. In terms of deep learning algorithm, Yan LeCun proposed LeNet [15], which has high accuracy and is widely used in handwritten character recognition and classification, laying the foundation of convolutional neural network (CNN). Based on CNN, there are many excellent models such as AlexNet [16], VggNet [17] and ResNet [18]. In 2018, Eleftherios et al. [19] proposed an image segmentation method using 3D convolutional neural network, which can directly segment 3D tomography data by a $3 \times 3 \times 3$ convolution kernel. 3D convolutional neural network can effectively enhance the relationship between volume data. Compared with 2D convolutional neural network, it can greatly improve the accuracy of 3D segmentation. Although convolutional neural network has made significant breakthroughs in the field of medical image segmentation, it still cannot achieve pixel-level segmentation. In this regard, Jonathan Long et al. [7] proposed fully convolutional networks (FCN). FCN converts the fully connected layer at the end of CNN into a convolutional layer, which not only allows the network have elastic input size, but also realizes pixel-level classification, greatly improving segmentation accuracy. Based on FCN, many researchers have proposed excellent fully convolutional neural networks such as U-Net [8], SegNet [20], DeepLab [21] and PSPNet [22], which have shown their effectiveness in the field of image segmentation. In 2017, Yu et al. [23] proposed a novel 3D FCN based on dense connections, which uses 3D fully convolutional architecture to predict volume data, and can automatically segment heart and vessel structures from 3D images. In 2018, Alom M Z et al. [24] proposed a U-Net-based recurrent neural network model, which gains the advantages of U-Net, residual network and RCNN, greatly improves the performance of segmentation. This method achieves good results in retinal vessel segmentation, skin cancer segmentation and lung segmentation tasks. Wang et al. [25] proposed DEU-Net that can use both spatial and semantic context information. The model uses the attention mechanism for feature reuse, and achieves good results on the retinal vessel segmentation task.

In conclusion, the U-Net-based segmentation methods have a good effect in medical image segmentation. Therefore, we proposed a dual-path U-Net for pulmonary vessel segmentation based on lightweight 3D attention (DS-ResUNet), which can solve the problems of low segmentation accuracy caused by the mis-segmentation of tiny vessels and massive parameters.

3 Method

3.1 DS-ResUNet

In terms of image segmentation based on deep learning, full convolutional neural network is mainly used as the basic methods of image feature extracting, among which U-Net is a representative framework. U-Net uses full convolutional neural network and skip connection layer to combine high-level semantic features with low-level detailed features of image, which can greatly improve the accuracy of segmentation results. U-Net includes encoder, decoder and skip connection, whose structure is symmetrical and u-shaped as a whole. The encoder uses the multi-scale transformation method to extract the high-level semantic features of the image; the decoder maps high-level semantic features to original size; the skip connection is a bridge for multi-scale fusion of encoder and decoder.

The depth of the neural network has a great influence on the segmentation results. When the depth of the neural network is shallow, the extracted high-level semantic features of the image is less, and the image segmentation accuracy is low. When the depth is deep, the phenomenon of gradient vanishing or explosion is easy to occur, which would negatively impact the convergence on model training.

To solve this problem, residual structure is adopted into U-Net, namely ResUNet. ResUNet has two advantages: (1) Increasing the nonlinear structure of the network can deepen the network structure, to avoid the problem of overfitting and non-convergence. (2) U-Net realizes the fusion of high-level semantic features and low-level detailed features by skip connection, which can improve the receptive field from multiple dimensions. Therefore, the structural framework of ResUNet is more suitable for vessel segmentation in lung CT images.

There are 2 mm and 5 mm volume data in CT images. The 2 mm volume data has a clear fine structure, which can accurately identify lung lesions. The 5 mm volume data is sub-thin layer CT scan data, with less noise and clear structure, which can accurately identify pulmonary vessel. Therefore, the accuracy of vessel features extraction can be improved by fully utilizing the complementary advantages of the two kinds of volume data. Based on ResUNet, a dual-path U-Net for pulmonary vessel segmentation based on lightweight 3D attention (DS-ResUNet) which can fuse 2 mm and 5 mm volume data is designed.

The structure of DS-ResUNet is shown in Fig. 1a. DS-ResUNet consists of three modules: encoder, decoder and skip connection. The encoder is composed of four blocks, and the high-level semantic features of image are extracted between the blocks through the down-sampling operation. In the process of down-sampling, the encoder fuses 2 mm and 5 mm volume data at multi-scales in multiple layers, which can fully guarantee the information integrity of high-level semantic features. In Fig. 1a, the decoder and the encoder are symmetrical. Each decoder layer has a corresponding encoder layer in the network. And hence each decoder layer has the same number of size and channels as their encoder inputs. The final decoder output is fed to a multi-class softmax classifier to produce class probabilities for each pixel independently.

Each block in the encoder is composed by a residual unit and a feature fusion module. The residual unit is shown in Fig. 1c, and the feature map of the input block is convolution, and the results are superimposed with the original feature. The residual unit can effectively extract detailed vessel features from 2 mm volume data. The feature fusion module is shown in Fig. 1d. In this module, the enhanced features of 2 mm volume data would be extracted by residual unit. And then we combine the features of 5 mm data with the enhanced 2 mm data, to realize feature fusion.

The skip connection in Fig. 1a is the bridge for multi-scale fusion of encoder and decoder. We consider that multi-scale fusion could become less effective if there is a large semantic or resolution gap between encoder and decoder. To address it, we adopt a 3D axial attention module as shown in Fig. 1b to strengthen the weight of vessel position and suppress irrelevant information. In addition, in order to improve the convergence speed of model training and avoid the gradient vanishing, we adopt PReLU as activation function after each convolution.

3.2 3D axial attention

For the U-Net, skip connection is a bridge for multi-scale fusion of encoder and decoder, but the low-level features contain a lot of noise, which affects the feature representation. Therefore, the spatial attention mechanism is adopted to strengthen the weight of vessel position when the low-level features are connected by skip connection, which can effectively reduce the negative influence of noise on segmentation.

The classical 3D attention mechanism obtains the weight matrix by a $k \times k \times k$ convolution layer. The weight matrix can strengthen original feature spatially; however, the $k \times k \times k$ convolution kernel will increase parameters sharply. To solve this problem, we propose a 3D axial attention, which uses separable convolution instead of classical convolution to calculate 3D spatial attention. It can perform convolution in three directions, respectively, which can greatly reduce the parameters and maintain segmentation accuracy. The structure of 3D axial attention is shown in Fig. 1b.

When separable convolution is used to calculate spatial attention of 3D CT image, the axis order of spatial attention should be considered from x-axis, y-axis and z-axis, respectively. If the calculation order of separable convolution is $x \rightarrow y \rightarrow z$, the convolution order is $1 \times k \times 1$, $k \times 1 \times 1$, $1 \times 1 \times k$. Therefore, when calculating spatial attention of 3D CT image by separable convolution, there are six arrangements according to the calculating order of x-axis, y-axis and z-axis. To avoid the effect of calculation order, the sum of the output results of six arrangements can be used as spatial attention. SA is calculated as follows:

$$\begin{aligned} C_1 &= conv_3(conv_2(conv_1(X))) \\ C_2 &= conv_2(conv_3(conv_1(X))) \\ C_3 &= conv_3(conv_1(conv_2(X))) \\ C_4 &= conv_1(conv_3(conv_2(X))) \\ C_5 &= conv_1(conv_2(conv_3(X))) \\ C_6 &= conv_2(conv_1(conv_3(X))) \end{aligned} \tag{1}$$

$$SA = \sigma(C_1 + C_2 + C_3 + C_4 + C_5 + C_6) \tag{2}$$

$$\widetilde{X} = X \cdot SA \tag{3}$$

where σ is Sigmoid activation function. $conv_1, conv_2, conv_3$ denote the 3D convolution with kernel $1 \times k \times 1$, $k \times 1 \times 1$, $1 \times 1 \times k$.

4 Segmentation strategy design

4.1 Semantic-enhanced module

CNN-based methods usually adopt down-sampling to increase receptive field and reduce calculation cost. However, the down-sampling method, such as pooling and strided convolution, will cause image information loss. At present, more and more CNN-based methods utilize atrous convolution as down-sampling method. Compared with pooling and strided convolution, the atrous convolution has a larger receptive field, which can effectively reduce the image information

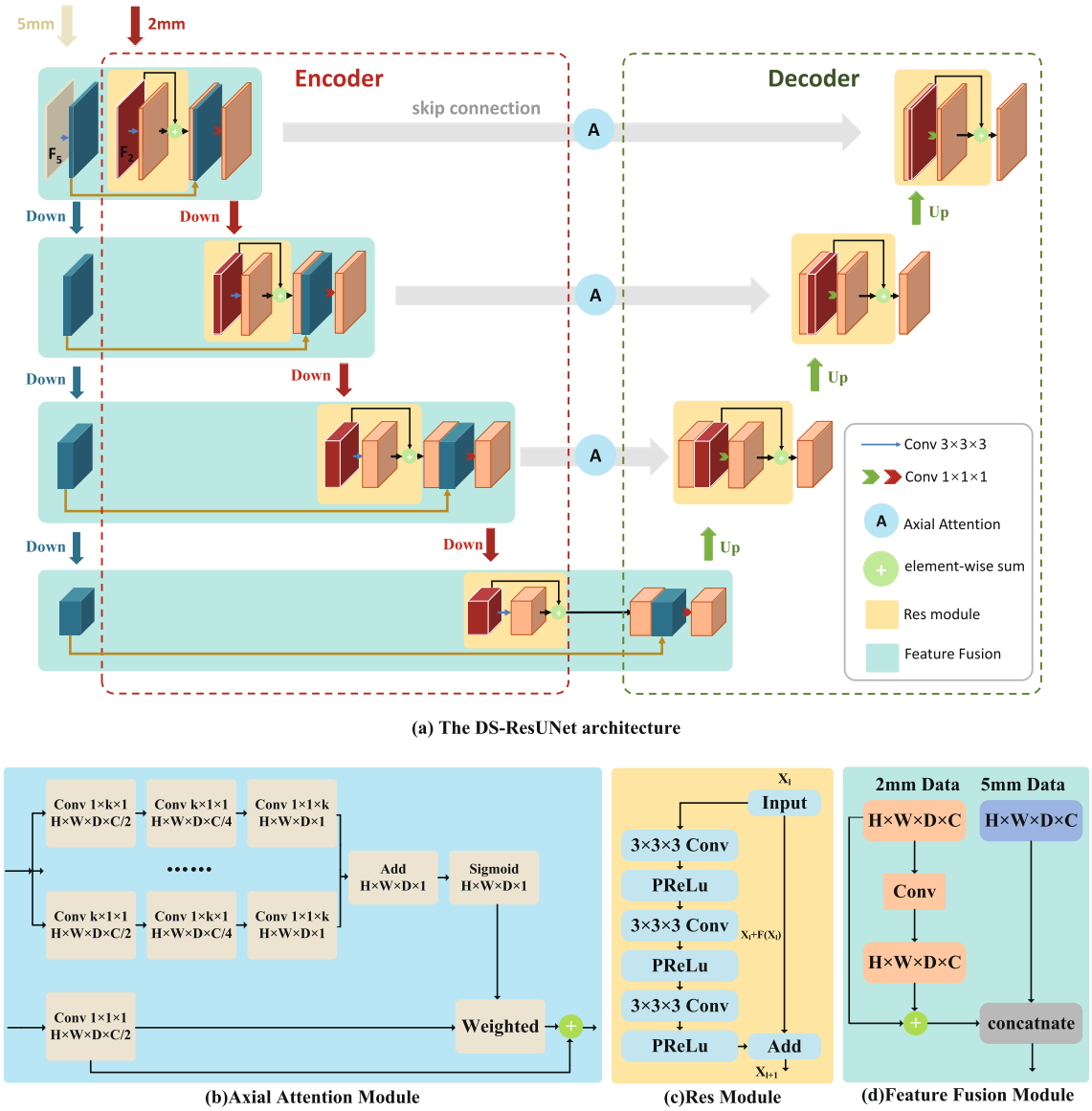


Fig. 1 Illustration of the architecture of the proposed DS-ResUNet

loss. Therefore, an SE-AC module is used based on atrous convolution, as shown in Fig. 2, which first employ a convolutional layer with $r \times r \times r$ kernels to extract local features. In SE-AC module, dilation rate dr is the controller of atrous convolution, which can extend the receptive field of atrous convolution. By adjusting the dilation rate dr, the global features of image can be extracted. The SE-AC can greatly preserve the spatial features of images without increasing model parameters.

4.2 ASPP with full sampling

To improve the efficiency and effectiveness, the ASPP-FS is proposed, as shown in Fig. 3, to extract multi-scale contexts and enhance vessel texture features. The ASPP-FS is composed of four parallel atrous convolution branches. After channel reduction by $1 \times 1 \times 1$ convolutional layer in each branch, we first employ a convolutional layer with $r_i \times r_i \times r_i$ kernels to extract local features. The following atrous convolutional layer with dilation rate dr_i is able to extract global features in a full sampling manner. Then, we fuse the enhanced features of four branches through a concatenation operation. The channel fusion can be realized by a $1 \times 1 \times 1$ convolutional layer.

4.3 Joint pyramid upsampling

The atrous convolutions play an important role in maintaining the spatial features, which has a superior performance compared to most encoder–decoder-based methods. How-

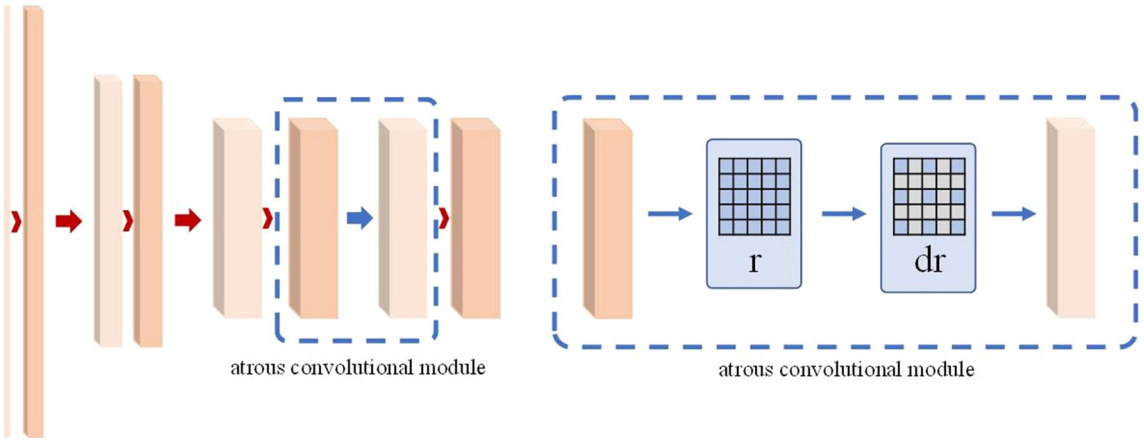


Fig. 2 The architecture of SE-AC module

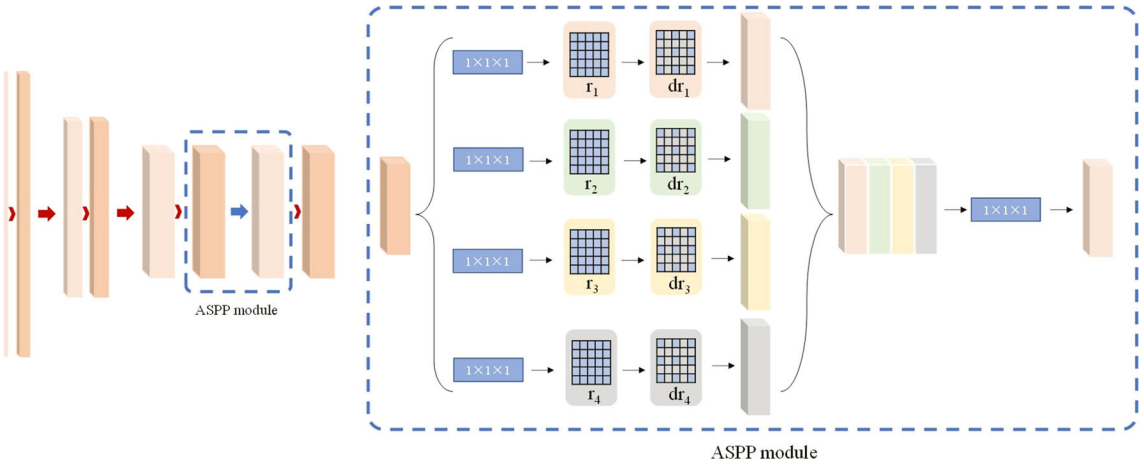


Fig. 3 The architecture of ASPP-FS module

ever, the introduced atrous convolutional layer bring heavy computation complexity and memory usage, which limits its application in many real-time scenes. Aiming at the aforementioned issue caused by atrous convolutional layer, we use a novel joint upsampling module to replace the time and memory consuming atrous convolutions, namely joint pyramid upsampling (JPU) [26]. The JPU takes the last three features as the inputs of ASPP, to segment vessels in multi-scale.

As shown in Fig. 4, JPU can be divided into three stages. After the channel fusion by the $1 \times 1 \times 1$ convolutional layer, the generated features are up-sampled and concatenated, and then four atrous convolutions with different dilation rates are employed in parallel to deeply extract features. Then, we fuse the four branches through a concatenation operation. Finally, another regular convolution block is employed, to segment vessels.

5 Experiments

5.1 Datasets and pre-processing

It takes professionals a lot of time and energy to make high-quality lung CT images for image collection and labeling. Therefore, there are few publicly available datasets of lung CT images, and such datasets have low contrast and lots of noise, resulting in poor pulmonary vessel segmentation effect.

In this regard, to evaluate the proposed method, we apply our method on the original CT images of 110 patients from a hospital in Zhejiang Province. The images were generated by scanning the lungs from the apex to the base of the lung by a high-speed spiral CT machine, and the data were scanned every 2 mm and 5 mm, respectively. The image format was DICOM. All the data used in this study are manually labeled one by one under the guidance of a medical expert, and there are two categories (artery and vein).

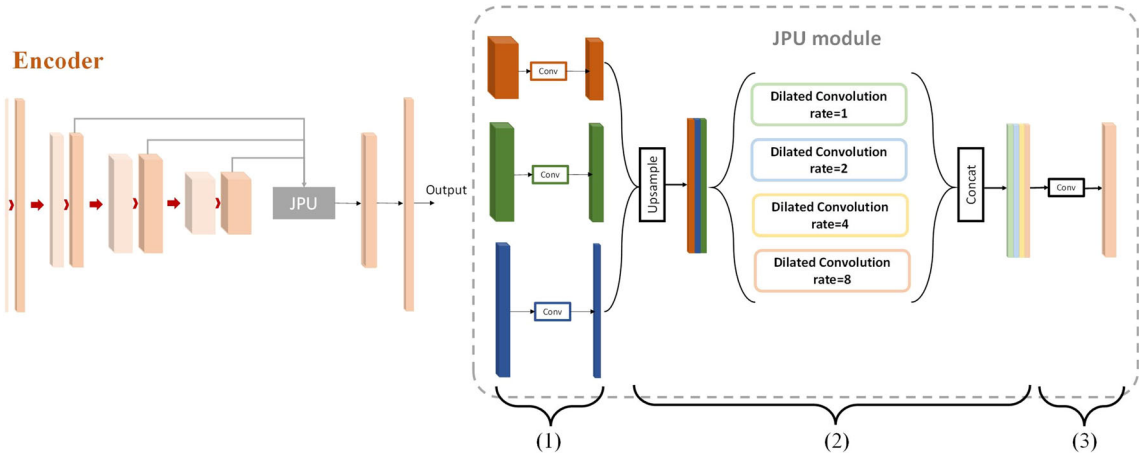


Fig. 4 The architecture of JPU

Different from natural images, the voxel in a CT volume is the Hounsfield unit (HU) value with a range from −1000 (air at standard pressure and temperature) to more than +3000 (dense bone). To remove irrelevant information, we truncate the HU values of all volumes to range [−950, +350], and then normalized them linearly to [0, 1]. In addition, the linear gray transformation is used to enhance the contrast between vessels and other background regions. The CT images after gray transformation can enhance the vessel texture, and the background can be well suppressed. This method lays a good foundation for pulmonary vessel segmentation. Finally, the experimental data is preprocessed, where 70% is used as the training set and 30% as the validation set.

5.2 Implementation details

The PyTorch framework is employed to implement our network on a NVIDIA Tesla V100 GPU. We use the U-Net as our basic framework in the following experiments. During training phase, the framework is trained by an Adam optimizer for 200 iterations, with an initial learning rate 0.001 decayed by 0.1 every 40 iterations. For data augmentation, random cropping (crop_size 32) is employed as the network input. During the testing phase, we use a sliding window strategy to obtain the final results.

In this paper, we use Dice similarity coefficient (DSC) [27] to evaluate the segmentation performance of DS-ResUNet. DSC is a set similarity metrics, which is usually used to calculate the similarity of two samples. The value ranges from 0 to 1, and the best value of segmentation results is 1, and the worst value is 0. The definition of DSC is as follows:

$$\mathrm{DSC} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{4}$$

where TP, FP and FN represent the number of true positives, false positives and false negatives, respectively.

5.3 Results on various modules

(1) In order to solve the problem of detailed vessel texture loss in down-sampling, we utilize the SE-AC module as the last down-sampling layer of DS-ResUNet. The feature size can be kept by an atrous convolution in SE-AC, and the dilation rate of the atrous convolution is 4.

(2) In this experiment, in order to enhance the perception ability of the model on multiple scales, the last layer of down-sampling is replaced by ASPP-FS module. In Fig. 3, there are four traditional convolutions r_{1-4} and four atrous convolutions dr_{1-4} to extract multi-scale features on parallel. The four convolutions r_{1-4} have different kernel sizes, and four atrous convolutions dr_{1-4} have different dilation rates. By ASPP-FS, image feature can be extracted on four scales, whose channels are down to 32. Finally, the extracted multi-scale features are fused. In this experiment, the kernel sizes in the ASPP-FS module are (1, 3, 5, 7) and the dilation rates are set as (1, 2, 4, 8).

(3) In order to reduce computing cost and memory usage in model training, we use JPU module as upsampling module on DS-ResUNet. The JPU module can extract multi-scale features in multiple layers using parallel atrous convolution. The dilation rates of the four atrous convolutions in JPU are (1, 2, 4, 8). Finally, the multi-scale features are fused and mapped to the size of ground truth.

The three modules mentioned above can enhance the semantic classification ability of DS-ResUNet. We carry out DS-ResUNet with the three modules and evaluate the pulmonary vessels segmentation by DSC. The vessels segmentation results are shown in Table 1. In Table 1, the DS-ResUNets with the three modules have greater DSC than normal DS-ResUNet. The three modules can significantly improve the vessel segmentation performance on veins (difficult samples).

Table 1 Comparison of three modules on vessels segmentation by DSC (%)

Modules	Artery_DSC	Vein_DSC
DS-ResUNet(SE-AC)	0.6860 ± 0.0107	0.7124 ± 0.0193
DS-ResUNet(ASPP-FS)	0.6896 ± 0.0258	0.6639 ± 0.0321
DS-ResUNet(JPU)	0.6944 ± 0.0163	0.6544 ± 0.0217
DS-ResUNet	0.7161 ± 0.0282	0.6531 ± 0.0411

Table 2 Performance of the network with different image preprocessing strategies

Settings	Artery_DSC	Vein_DSC
DS-ResUNet w/o enhanced	0.6897 ± 0.0133	0.6239 ± 0.0324
DS-ResUNet w/ enhanced	0.7161 ± 0.0282	0.6531 ± 0.0411

5.4 Ablation studies

To verify the rationality of the DS-ResUNet, we conduct extensive ablation experiments with different settings. There are mainly the following aspects, image enhancement, attention mechanism and training strategy.

5.4.1 Image enhancement strategy

For pulmonary vessel segmentation, it is a challenging problem to pre-process CT images to reduce the interference of noise. In the original CT images, most of them contain noise, which has a great impact on the analysis and processing of CT images. Therefore, it is necessary to select appropriate image enhancement and denoising schemes to eliminate or reduce the noise in the images, to reduce its adverse effects in medical image processing. In this paper, we use an image enhancement scheme including image thresholding, normalization, linear gray-level transformation and Gaussian filter to enhance the original lung CT data.

As shown in Table 2, without the use of image enhancement and denoising, the feature representation ability of DS-ResUNet would be disturbed by the noise in CT images, which will seriously affect the segmentation performance of the model. On the contrary, in the stage of image preprocessing, we use image enhancement and denoising to effectively reduce the influence of noise on model training, alleviate the problem of the training difficulty and improve the accuracy of pulmonary vessel segmentation to a certain extent.

5.4.2 Comparisons of attention module

In this section, we verify the effectiveness of the axial attention module. In DS-ResUNet, if the skip-layer connection is directly used to fuse the encoder features to recover the lost spatial detail information, a large amount of noise would be introduced, which will affect the feature representation ability of the network. Therefore, we propose an axial attention module to strengthen the weight of vessel position, which can effectively reduce the interference of noise. As shown in Table 3, it can be seen from the experimental results that DS-ResUNet with the axial attention module can achieve better results. On the contrary, if the attention module is not used, the feature representation ability of the network would be reduced, thus reducing the accuracy of pulmonary vessel segmentation.

In addition, the classic 3D attention modules usually use $k \times k \times k$ convolution kernel to calculate the weight matrix. However, DS-ResUNet is a 3D model. If $k \times k \times k$ convolution kernel is used to calculate the weight matrix in each layer of network, the parameters of the network will increase sharply, thus increasing the training difficulty of the network. In this regard, in axial attention module, we use separable convolution instead of classic convolution to calculate the weight matrix. The module not only fully considers the characteristics of 3D data when calculating the weight matrix, that is, calculating the weight matrix from three directions, respectively, but also effectively reduces the amount of feature calculation. As shown in Table 3, we extend the classic attention modules (SE module and CBAM module) to 3D version and compare them with the axial attention module proposed in this paper. As shown in Table 3, it can be seen from the experimental results that the axial attention module reduces the training difficulty by effectively reducing the amount of feature calculation of DS-ResUNet, so that the segmentation performance of DS-ResUNet using this module is better than that of DS-ResUNet using the other two classic attention modules.

5.4.3 Comparisons of training strategy

The image size after encoder processing is down to one-eighth, so a lot of detailed vessel textures are lost in the down-sampling process. Therefore, it is very difficult to recover the features through the decoder. In this regard, we use deep supervision mechanism to adjust the parameters of each layer, to keep the detailed vessel texture in down-sampling. The deep supervision mechanism makes the parameters in each layer to be trained sufficiently, by adding additional loss into each layer. The deep supervision mechanism can also avoid gradient vanishing in model training. In the upsampling process of DS-ResUNet, the output of each layer is directly mapped to the size of ground truth, and supervise the outputs by ground truth. We adopt the training strategy above to deeply supervise the model training for DS-ResUNet. For the training criterion, we use a linear combination of deep supervision loss in each layer as total

Table 3 Evaluation of the proposed network with different attention blocks

Settings	Artery_DSC	Vein_DSC
DS-ResUNet w/o att	0.6823 ± 0.0072	0.6193 ± 0.0145
+SE	0.7076 ± 0.0236	0.6397 ± 0.0104
+CBAM	0.7094 ± 0.0187	0.6435 ± 0.0350
+Axial attention	0.7161 ± 0.0282	0.6531 ± 0.0411

Table 4 Comparison of different training strategies to segment vessels in terms of DSC (%)

Methods	Artery_DSC	Vein_DSC
Ours with single loss	0.6698 ± 0.0421	0.5920 ± 0.0377
Ours with deep supervision	0.7161 ± 0.0282	0.6531 ± 0.0411

Table 5 Quantitative comparison between our proposed method and other methods

Methods	Artery_DSC	Vein_DSC
3D U-Net	0.6998 ± 0.0122	0.6389 ± 0.0051
3D Attention U-Net	0.6972 ± 0.0254	0.6509 ± 0.0157
3D ResUNet	0.7044 ± 0.0038	0.6364 ± 0.0344
3D SegNet	0.6530 ± 0.0311	0.5698 ± 0.0432
3D DeepLab	0.7033 ± 0.0110	0.6436 ± 0.0066
Ours with attention gates	0.7102 ± 0.0163	0.6707 ± 0.0330
Ours	0.7161 ± 0.0282	0.6531 ± 0.0411

loss. The coefficient of loss combination is trainable. In this strategy, the initial coefficient of deep supervision loss is 0.4.

In this section, the proposed DS-ResUNet model using deep supervision training strategy is compared with that using classical training strategy, and the results are shown in Table 4. It can be seen from the experimental results that the network using deep supervision can precisely extract vessel features, and perform better than that using single loss on artery and vein segmentation.

5.5 Comparison with state-of-the-art methods

In this experiment, we compare our method with five widely used methods about medical image segmentation on the same dataset. The comparison is shown in Table 5. We can see that our method has a better accuracy than the five state-of-the-art methods in terms of DSC. The DS-ResUNet with attention gates works well on vein segmentation, but it has massive parameters and performs worse on artery segmentation. Compared with U-Net, ResUNet improves the segmentation accuracy by a large margin, indicating that residual learning can bring performance gain. We also visualize segmentation results by some classical methods in Fig. 5, which further show the superiority of our proposed method.

As shown in Fig. 5d, the 3D fusion module designed in this study can utilize 2 mm and 5 mm volume data in CT images to segment vessels. The 2 volume data can provide more detailed texture information than single volume data, so DS-ResUNet can maintain a higher segmentation accurate even if an image slice contains less vessel texture.

We visualize the performance of basic U-Net, attention U-Net, SegNet, ResUNet and our proposed method in Fig. 6.

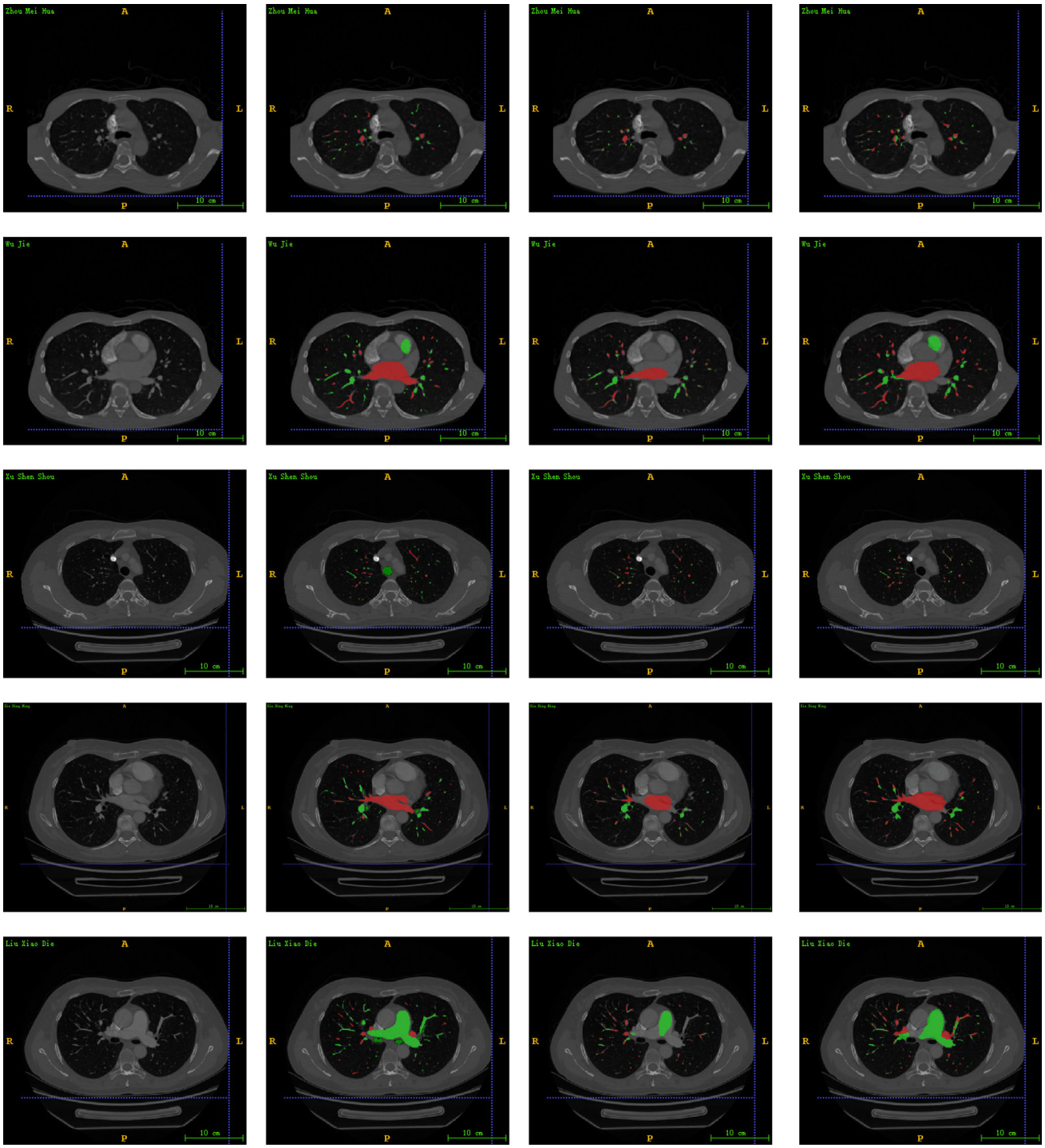
The effectiveness of DS-ResUNet is also confirmed by the improved performance evaluated by sensitivity as shown in Fig. 6. It is worth noting that our proposed 3D fusion module and 3D axial attention module contributes more performance gain for pulmonary vessel segmentation. It is consistent with our assumption that DS-ResUNet could pay more attention to tiny vessels and thus can deal with tiny vessels segmentation much better.

6 Conclusion

In this paper, we propose a dual-path U-Net for pulmonary vessel segmentation based on lightweight 3D attention. This method uses feature fusion module to fuse 2 mm and 5 mm volume data, which greatly alleviates the problem of detailed vessel texture loss caused by down-sampling and improves the segmentation accuracy of pulmonary vessels (especially tiny vessels). On this basis, we design a lightweight 3D axial attention module, which effectively reduce the massive parameters in classical 3D attention mechanism.

In this paper, we use a pulmonary vessel dataset finely labeled by medical experts for experiment. In our experiments, we train DS-ResUNet with deep supervision mechanism, so as to precisely extract vessel features. In addition, three semantic enhancement modules are further used to enhance the feature representation ability of our method and effectively improve the accuracy of vein (difficult samples) segmentation. Then, we compare the segmentation performance with the state-of-the-art methods on the same dataset. DS-ResUNet achieves remarkable advantages over other methods, evidenced by the highest DSC 71.61% for pulmonary vessel segmentation.

Our method has advantage in the segmentation of tiny vessels, and there is still a lot of room for improvement in segmentation of vessel-like regions. In designing model, we realize the fusion of 2 volume data, and further research should be undertaken to design the strategies of fusing multi-view volume data. Through the optimization and adjustment of our method, lung CT auxiliary diagnosis applications such as lung tumor and lung nodule segmentation can be realized.



(a) Image (b) Ground Truth (c) ResUNet (d) Ours

Fig. 5 Segmentation results on different methods

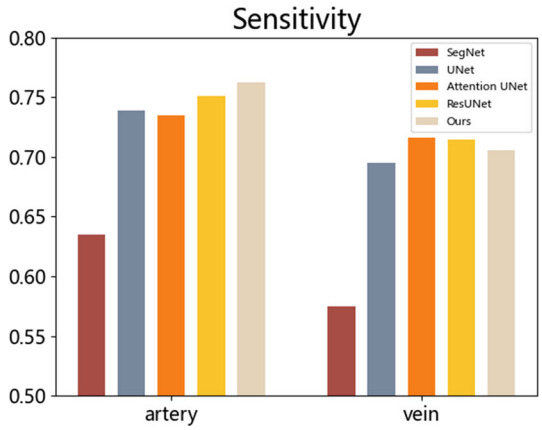


Fig. 6 Average sensitivity ratios of different methods

Acknowledgements We acknowledge the support of the Natural Science Foundation of Zhejiang Province (Grant No. LY22F020001) and the 3315 Plan Foundation of Ningbo (Grant No. 2019B-18-G).

References

1. Sookpotharom and Supot. The segmentation of dermoscopy images by k-means thresholding
2. Osuna-Enciso, V., Cuevas, E., Sossa, H.: A comparison of nature inspired algorithms for multi-threshold image segmentation. Expert Syst. Appl. **40**(4), 1213–1219 (2013)
3. Rouhi, R., Jafari, M., Kasaei, S., Keshavarzian, P.: Benign and malignant breast tumors classification based on region growing and CNN segmentation. Expert Syst. Appl. Int. J. **42**(3), 990–1002 (2015)
4. Bertolotto, M., Vo, A.-V., Truong-Hong, L., Laefer, D.F.: Octree-based region growing for point cloud segmentation. ISPRS J. Photogramm. Remote Sens. **104**(Jun), 88–100 (2015)
5. Zhao, Y.Q., Wang, X.H., Wang, X.F., Shih, F.Y.: Retinal vessels segmentation based on level set and region growing. Pattern Recognit. **47**(7), 2437–2446 (2014)
6. Shen T., Wang, Y.: Medical image segmentation based on improved watershed algorithm. In: 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), (2018)
7. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 640–651 (2015)
8. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional Networks for Biomedical Image Segmentation. Springer, Cham (2015)
9. Lizarazo, I.: SVM-based segmentation and classification of remotely sensed data. Int. J. Remote Sens. **29**(23–24), 7277–7283 (2008)
10. Park, S., Lee, H.S., Kim, J.: Seed growing for interactive image segmentation using SVM classification with geodesic distance. Electron. Lett. **53**(1), 22–24 (2016)
11. Alush, Amir: Goldberger, Jacob: Hierarchical image segmentation using correlation clustering. IEEE Trans. Neural Netw. Learn. Syst. **27**(6), 1358 (2016)
12. Burney, S.A., Tariq, H.: K-means cluster analysis for image segmentation. Int. J. Comput. Appl. **96**(4), 1–8 (2014)
13. He, B., Huang, C., Sharp, G., Zhou, S., Hu, Q., Fang, C., Fan, Y., Jia, F.: Fast automatic 3D liver segmentation based on a three-level AdaBoost-guided active shape model. Med. Phys. **43**(5), 2421–2434 (2016)
14. Lupascu, C.A., Tegolo, D., Trucco, E.: FABC: retinal vessel segmentation using AdaBoost. IEEE Trans. Inf. Technol. Biomed. A Publ. IEEE Eng. Med. Biol. Soc, **14**(5), 1267–74 (2010)
15. Abdelhamid, O., Deng, L., Yu D., Jiang, H., Penn, G., Bouvrie, J., Lawrence, S., Giles, C.L., Tsoi, A.C., Back, A.D.: Gradient-based learning applied to document recognition (2013)
16. Everingham, M., Eslami, S., Van Gool, L., Williams, C., Winn, J., Zisserman, A.: The pascal visual object classes challenge: a retrospective. Int. J. Comput. Vis. **111**(1), 98–136 (2015)
17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *arXiv*, (2014)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE, (2016)
19. Trivizakis, E., Manikis, G.C., Nikiforaki, K., Drevelegas, K., Constantinides, M., Drevelegas, A., Marias, K.: Extending 2D convolutional neural networks to 3D for advancing deep learning cancer classification with application to MRI liver tumor differentiation. IEEE J. Biomed. Health Inf. **23**(3), 923–930 (2018)
20. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2481–2495 (2017)
21. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2018)
22. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: IEEE Computer Society, (2016)
23. Yu, L., Cheng, J.Z., Qi, D., Xin, Y., Heng, P.A.: Automatic 3d cardiovascular MR segmentation with densely-connected volumetric convnets (2017)
24. Alom, M.Z., Hasan, M., Yakopcic, C., Taha, TM, Asari, V.K.: Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation (2018)
25. Wang, B., Qiu, S., He, H.: Dual Encoding U-Net for Retinal Vessel Segmentation. In: Medical Image Computing and Computer Assisted Intervention—MICCAI 2019, 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I (2019)
26. Wu, H., Zhang, J., Huang, K., Liang, K., Yu, Y.: Fastfcn: rethinking dilated convolution in the backbone for semantic segmentation (2019)
27. Yisong, H.E., Jiang, J., Hang, Y.U., Yuchuan, F.U., Radiotherapy, D.O., Hospital, W.C., University, S.: Comparison of dice similarity coefficient and Hausdorff distance in image segmentation. Chin. J. Med. Phys. **36**(11), 1307–1311 (2019)

Rencheng Wu is a graduate student of Faculty of Electrical Engineering and Computer Science at Ningbo University. His research interests include mainly deep learning, computer vision, and medical image segmentation.

Yu Xin received the Ph.D. degree in computer science and technology from Harbin Engineering University, China. He is currently an associate professor with the Faculty of Electrical Engineering and Computer Science, Ningbo University, China. His current research interests include multiple classifier and prediction systems, processing and modeling of uncertainty in predictive modeling, recommendation systems, diagnostic analysis, and decision support systems.

Yihong Dong received the Ph.D. degree in computer science from Zhejiang University, China, in 2007. He is currently a professor with the Faculty of Electrical Engineering and Computer Science, Ningbo University, China. His research interests include big data, data mining, and artificial intelligence.

Jiangbo Qian received the Ph.D. degree in computer science from Southeast University, China, in 2006. He was a visiting scholar with the Department of Computer and Information Science, the University of Michigan-Dearborn, USA. He is currently a professor with the Faculty of Electrical Engineering and Computer Science, Ningbo University, China. His research interests include database management, streaming data processing, deep learning, computer vision, and hardware/software co-design.