MufiNet: Multiscale Fusion Residual Networks for Medical Image Segmentation

Chun Wang
Xi'an Jiaotong University
School of Software Engineering
Xi'an, China
wangchunlhq@foxmail.com

Zhi Wang*
Xi'an Jiaotong University
School of Software Engineering
Xi'an, China
zhiwang@xjtu.edu.cn

Wei Xi
Xi'an Jiaotong University
School of Computer Science & Technology
Xi'an, China
xiwei@xjtu.edu.cn

Zhao Yang
Xi'an Jiaotong University
School of Computer Science & Technology
Xi'an, China
zhaoyang9425@gmail.com

Gairui Bai
Xi'an Jiaotong University
School of Computer Science & Technology
Xi'an, China
grbai2018@stu.xjtu.edu.cn

Ruimeng Wang
The University of New South Wales
School of Photovoltaics & Renewable Energy
Sydney, Australia
brianrwangmsecs@gmail.com

Meichen Duan
Xi'an Jiaotong University
School of Computer Science & Technology
Xi'an, China
1917106079@qq.com

***Abstract*—U-Net has been considered as an outstanding deep learning neural network in medical image segmentation problems. The segmentation results of the U-Net based model, however, are always too conservative and smooth. MufiNet, a segmentation model using multiple U-Net chains (with multiple encoder-decoder branches), is proposed in this paper. It can fuse the receptive fields obtained from different scales. The convolution layer of 1×1 is introduced to add the residual connection to enhance the adaptability to the depth of the network. The multi-scale fusion module with residuals is combined with the U-Net chain architecture to retain more information flow paths, and the multi-scale context information is used to improve the performance and robustness of the segmented network. MufiNet model is extensively evaluated on three datasets in this paper, including two benchmark datasets (lung segmentation and skin cancer lesion segmentation) and cervical cancer dataset jointly constructed with a hospital. The experimental results show that MufiNet could yield better performance in medical image segmentation tasks than U-Net and LadderNet models.**

***Index Terms*—U-Net chain, medical image segmentation, multi-scale, fusion**

I. Introduction

Image segmentation is a fundamental problem and complex task in the field of image processing and computer vision, because the segmentation result will directly affect the performance of the subsequent processing steps [1]. The purpose is to segment areas of interest (such as tumor areas, organs) in medical images to extract relevant features, and then classify them at the pixel level to cluster similar pixels together. This could greatly improve clinical processes. For disease diagnosis, disease progression detection and treatment planning are crucial. However, it is very difficult to obtain large amounts of labeled data in the medical field.

The emergence of U-Net brings the possibility to solve this problem. U-Net [2] simply splices the feature map generated by the encoder with the upper sampling feature map of the corresponding decoder in each stage to form a trapezoid structure. By skipping the connection, each stage allows the decoder to learn the relevant features lost in the corresponding encoder pooling. U-Net achieved the best results on the EM dataset and still performed well without large enough medical image data. After that, many researchers continued to conduct in-depth research on U-Net, and also proposed many U-Net variants. Oktay et al. [3] proposed a new attention gate (AG) model for medical imaging. Alom et al. [4] combined a recursive residual convolutional neural network with U-Net for medical image segmentation. Zhuang [5] proposed LadderNet, which formed a more complex network structure by adding more U-Nets. However, during the encoding process of high-dimensional features, the original pixel context will gradually lose spatial resolution in the convolution process. In order to obtain fine segmentation results, a multi-scale context [6] is proposed. Zhao et al. [7] uses the pyramid pool module and the proposed Pyramid Scenario Analysis Network (PSPNet) to aggregate the context information based on different regions to mine global context information. However, the pyramid pool module proposed by PSPNet may lose pixel-level positioning information. In medical image segmentation, accurate segmentation of the

Corresponding author: zhiwang@xjtu.edu.cn

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

region of interest is crucial. Chen et al. [8] proposed to use different ratios of hole convolution and atrous spatial pyramid pooling (ASPP). Hole convolution expands the receptive field of convolution kernel, and ASPP used the input convolution feature layer of multiple sampling rates and effective field filters to capture the object and context information on multiple scales. But hole convolution will generate a lot of computing resources and may also cause the grid effect, which has no advantages or disadvantages for small object segmentation.

We proposed an effective multi-scale fusion module, called MufiNet, based on the U-Net chain. In the encoder and decoder part, three 3×3 convolutions are used to get the feature representation of different scales, and the receptive fields obtained from different scales are fused together. In addition, a 1×1 [9] convolution layer is introduced to add the residual connection to improve segmentation performance by preserving more information flow paths and fusing multi-scale receptive fields.

The contributions can be summarized as follows:

- By using 1×1 convolution to keep the size of the feature map unchanged (i.e. without loss of resolution), nonlinear characteristics are added, and the residual connection is introduced to enhance the adaptability to network depth.
- A "multi-scale fusion" strategy is proposed to fuse features at different scales to generate more complex semantic information. Combined with U-Net chain architecture, more accurate target boundary segmentation is achieved, which makes the model have better performance and robustness.
- Construct a cervical cancer image segmentation dataset. The performance of the medical image segmentation task model based on end-to-end is evaluated. It is experimentally observed that the proposed model achieves the best performance in different medical datasets including lung segmentation, skin cancer lesion segmentation, and cervical cancer segmentation.

II. Relation Work

A. Medical Image Segmentation

In recent years, the development of deep learning in the field of medical treatment has been widely concerned. Medical image segmentation [1] has also become a hot topic in recent years. Most models that achieve excellent performance in medical image segmentation tasks are improved by FCN or U-Net. In the FCN [10] architecture, the fully connected layer of the classic CNN after the convolution layer is replaced with the convolution layer. It can accept input images of any size and upsample the feature map obtained by the last convolution layer. And add skip architecture between the same depth layers, combining the local information learned from the shallow layer of the network with the more complex information learned from the deep layer. In the U-Net [2] structure, it includes a shrink path that captures contextual information and a symmetrical expansion path that allows precise positioning. Unlike FCN, the U-Net model does not directly perform the upsampling operation on the feature map obtained by the last convolution layer, but instead maps the high-dimensional feature to the low-level feature by transposing convolution. [11], [12], [13] are all improved from these networks. A classic idea is used in these networks, which is the encoder-decoder structure.

B. Multi-Scale Context

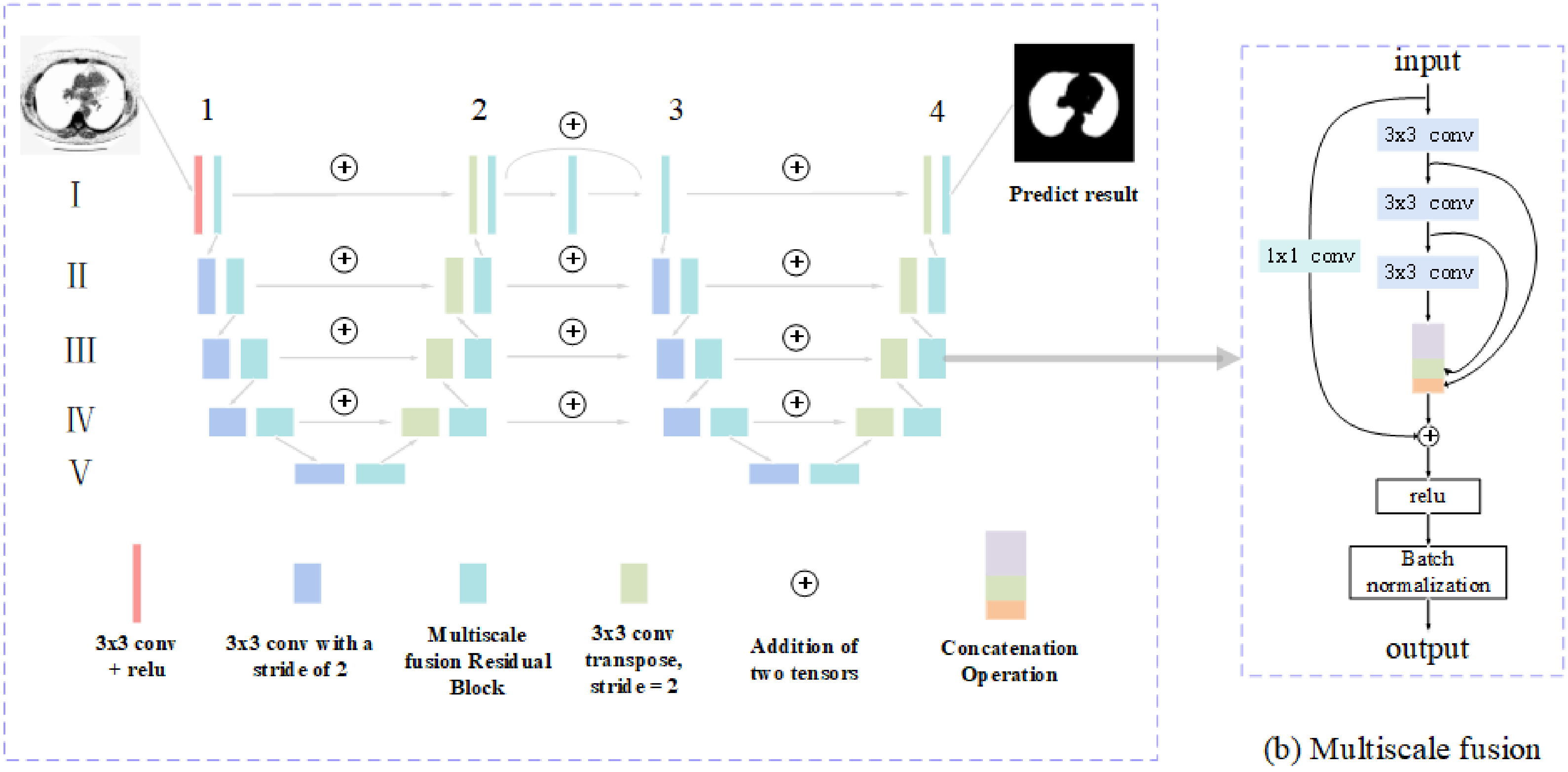
The smaller receptive field can only see smaller objects. Due to there are big differences between the shapes and sizes of various organs and tumor areas in medical images, a larger receptive field is needed to see larger objects. The simplest way to enhance the receptive field is downsampling, and upsampling simply restore the results of the downsampling to the original size. and it is impossible to completely recover the lost information, so using only a simple convolution operation cannot solve this problem. He et al. [14] proposed spatial pyramid pooling, which divides the features of the convolutional layer into different sizes, extracts features of fixed dimensions from each size and then fuses the features extracted from each block together. EncNet [15] introduces the Context Encoding Module to capture the global context information and highlight the category information associated with the scene. Using multi-scale context features, these methods also achieve better performance on various benchmarks. As far as we know, these parallel multi-scale processing operations have not been used in the U-Net chain structure.

Different from previous work, in the task of medical image segmentation, we combine the receptive fields obtained from different scales and add residual connections to combine low-level semantic information with high-level semantic information. And then transfer it to the network structure composed of multiple pairs of encoder-decoders to capture more complex features. Besides, the performance of proposed method is verified on multiple widely different datasets.

III. Method

A. FCNs

MufiNet consists of multiple pairs of encoder-decoder branches and has additional skip connection between each corresponding branch pair can be regarded as a collection of multiple FCNs. Its unique structure greatly increases the number of effective paths, and retains more information flow paths, thereby capturing more and more complex feature information and generating more accurate segmentation results. Due to computational constraints, our experiment takes two pairs of encoder-decoders as examples, as shown in Fig. 1(a). It can be seen that a variety of information flow paths are provided in the U-Net chain, and the number of effective paths has increased from 5 to 108, for example: (1) $I1 \rightarrow II1 \rightarrow II2 \rightarrow II3 \rightarrow II4 \rightarrow I4$; (2) $I1 \rightarrow I2 \rightarrow I3 \rightarrow I4$; (3) $I1 \rightarrow I2 \rightarrow I3 \rightarrow II3 \rightarrow II4 \rightarrow I4$; (4) $I1 \rightarrow I2 \rightarrow I3 \rightarrow II3 \rightarrow III3 \rightarrow III4 \rightarrow II4 \rightarrow I4$. However, due to the complex structure of many pairs of encoders and decoders, the model hierarchy is getting deeper and deeper, which leads to the loss of low-level information of the image,



(a) Network Architecture

(b) Multiscale fusion Residual Block

Fig. 1. The overall architecture of the proposed image segmentation network. We replace the sequence of two convolutional layers with shared weights in the LadererNet architecture with the proposed Multiscale fusion Residual Block.

which has an impact on the boundary information. In addition, for different medical image segmentation tasks, the region of interest that needs to be segmented is often irregular and have large differences in size, so the model should be more robust. To solve these problems, we propose to fuse receptive fields of different scales and add residual connections.

B. Encoder-Decoder

In the encoder part, we use the convolution operation with a step size of 2 from small receiving field features to large receiving field features, so that the feature map is halved and the number of channels is doubled. In the decoder part, we use a deconvolution operation with a step size of 2 from large receiving field features to small receiving field features, so that the feature map is doubled and the number of channels is halved. And adding skip connections between corresponding layers to fuse features at different levels of abstraction.

C. "Multi-Scale Fusion" Strategy

In the LadderNet model, after each convolutional layer and the transposed convolutional layer, two 3×3 convolutional layers sharing the same weight are used. The simplest way to extract features from different receptive fields is a parallel multi-branch network, including the basic Inception module of the Inception network [17], hole convolution [8], and directly using different size pooling operations. It can be seen from these models that the parallel structure can extract the features of different scale receptive fields at the same level, and then transfer them to the next level after fusion, which can balance the calculation amount and performance of the model more flexible. The simplest way to extend U-Net is to combine 3×3, 5×5 and 7×7 parallel convolution operations, which is helpful to improve the performance and robustness of the network. However, MufiNet is based on the U-Net chain, and the additional convolutional layer makes the calculation amount and memory requirements of the network increase exponentially. Inspired by the DeepLabv3+ [18] network, two 3×3 convolution operations are similar to a 5×5 convolution operation, and three 3×3 convolution operations are similar to a 7×7 convolution operation. Therefore, a series of 3×3 convolution sequences can be used instead of 5×5 and 7×7 convolution operations, as shown in Fig. 1(b). The output of the second and third 3×3 convolution operations can be considered to be equivalent to the 5×5 and 7×7 convolution operations. The feature maps extracted from the receptive fields of different scales are then subjected to concatenation operation. In order to achieve the same number of input and output channels, we set the number of filters of the three consecutive convolution layers to (0.2,0.3,0.5)×$Input_{channel}$, where $Input_{channel}$ is the number of input channels, because such parameter setting makes MufiNet has better convergence

TABLE I
ARCHITECTURE AND NUMBER OF NETWORK PARAMETERS.

Model	Number of parameter(million)
U-Net 32→64→128→25→512	0.78
MufiNet(3x3, 5x5, 7x7) 20→40→80→160→320	42.89
MufiNet(improved 3 3x3) 20→40→80→160→320	**0.73**

in both the training and testing phases.

D. Residual Connection

Instead of adding the original input directly to the output of the receptive field fused with different scales, we convolute the original input with a 1×1 [9] convolution so that each pixel can be linearly combined on different channels to achieve cross-channel information integration, which can extract features from the spatial dimension and the channel dimension respectively. The non-linear activation is added to the learning representation of the previous layer, which improves the expression ability of the network, preserves the low-level information of the image, makes the final segmentation result more precise, and realizes the accurate target boundary location.

IV. THE EXPERIMENTS

A. Implementation Details

MufiNet adopts the complex structure of U-Net chain and evaluates it with 20→40→80→160→320 architecture. In the traditional five-layer U-Net model, in order to make the number of parameters comparable to proposed model, the number of filters is set to 32→64→128→256→512. The required network parameters are 0.73M and 0.78m respectively. The network parameters required for the unimproved parallel 3×3, 5×5, and 7×7 models are 42.89M. The improvement in this paper greatly reduces the network parameters and speeds up the model training. In addition, semantic segmentation is performed using ADAM optimization techniques and cross-entropy loss. And 70% of the samples were used for training and 30% for testing. During the training process, 10% of the training samples are randomly selected as validation data, and the remaining 90% of the data is used to train the model. During the training, we set the total learning epoch to 100. Due to there will be no performance improvement in later learning.

The only preprocessing for the input images is to adjust their size to 256×256 pixels to fit the GPU memory for calculation, then divide the pixel value by 255. Set the image as a grayscale image with pixel values in [0 ... 1]. Again, no application-specific post-processing is performed. Finally, at the end of the network, a softmax function is used for pixel-level classification to calculate the probability of the target category, and its output range is also in [0 ... 1]. Therefore, we set the threshold to 0.5 to get the final segmentation result.

B. Database Summary

a) Lung Segmentation: The Lung Nodule Analysis (LUNA) competition held at the Kaggle Data Science Bowl in 2018 aims to find lung lesions in 2D and 3D CT images. The dataset contains 267 2D samples. The initial resolution of each sample is 512×512, which is scaled to 256×256 due to computational limitations.

b) Skin Cancer Lesion Segmentation: This dataset is from the 2017 kaggle competition dataset on skin lesion segmentation. The dataset contains a total of 2000 samples. Each sample has a variety of resolutions, and due to computational limitations, we scale it to 256×256.

c) Cervical cancer Segmentation: Finally, we performed experiments on our own data set, which was provided by the hospital with the original data, including the colposcopy images and test reports of 6974 patients. According to the inspection report and image quality, the images that meet the inspection criteria were selected, and a total of 2363 images were obtained.The initial resolution of each sample is 5184×3456, which is scaled to 256×256 due to computational limitations.

The ground truth for lung segmentation and skin lesion segmentation datasets are provided by the official website, and the ground truth for cervical segmentation dataset is marked by the professional doctor. However, the regions of interest segmented by different doctors also differ to some extent. In addition, the borders of the lesion are very hard to be defined even for specialist, especially true in the case of skin lesions.

C. Quantitative Analysis Approaches

In order to analyze the experimental results quantitatively, we used several indicators to evaluate the performance of MufiNet, including accuracy (AC), sensitivity (SE), specificity (SP) and Jaccard Similarity (JS). To do this, we first calculated True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). The different indicators are calculated as follows:

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

$$SE = \frac{TP}{TP + FN} \tag{2}$$

$$SP = \frac{TN}{TN + FP} \tag{3}$$

$$JS = \frac{V_{gt} \cap V_{pred}}{V_{gt} \cup V_{pred}} \tag{4}$$

Among them, V_{gt} represents the segmentation result of ground truth, and V_{pred} represents the predicted segmentation result.

In addition, area under curve (AUC) is common evaluation methods for medical image segmentation. To further evaluate the performance of different neural networks, we also compare the indicator.

TABLE II

Experimental results of the proposed method for lung segmentation and comparison with the U-Net, LadderNet and ResModelW/OConv.ResModelW/OConv represents the improved residual model without 1×1 convolution in this paper.

Method	AC	SE	SP	JS	AUC
U-Net	0.9774	0.9734	0.9827	0.9774	0.9691
LadderNet	0.9853	0.9781	0.9874	0.9853	0.9759
ResModelW/OConv	0.9869	0.9802	0.9879	0.9869	0.9764
MufiNet	**0.9878**	**0.9839**	**0.9889**	**0.9874**	**0.9793**

D. Results

a) Lung Segmentation: In the past fifty years, many countries have reported a significant increase in the incidence and mortality of lung cancer. Accurate segmentation and localization of lung areas of interest to doctors in CT images has also become critical. In this implementation, set the batch size to 32, and set the learning rate as 0.001, 0.0001 on the 0th and 20th epoch respectively.

Table II shows the quantitative results of different methods in the lung segmentation task. In terms of accuracy, specificity, sensitivity, JS and AUC, we will compare the proposed model with U-Net, LadderNet and the improved residual model without 1×1 convolution in this paper (ResModelW/OConv). Compared with U-Net and LadderNet, ResModelW/OConv achieves better performance in each indicator. It can be proved that the proposed multi-scale fusion strategy is effective. The proposed model provided the highest AUC and reached 0.9793. The accuracy of the proposed method reaches 0.9878, which is 1.04%, 0.25%, and 0.09% higher than U-Net, LadderNet, and ResModelW/OConv, respectively. In addition, we calculated JS during the test phase and reached 0.9878. The proposed model also produced higher SE (0.9839) and SP (0.9889). This proves that adding a 1×1 convolution residual model can enhance the adaptability of the model to the depth of the network. The experimental results show the superiority of the proposed network.

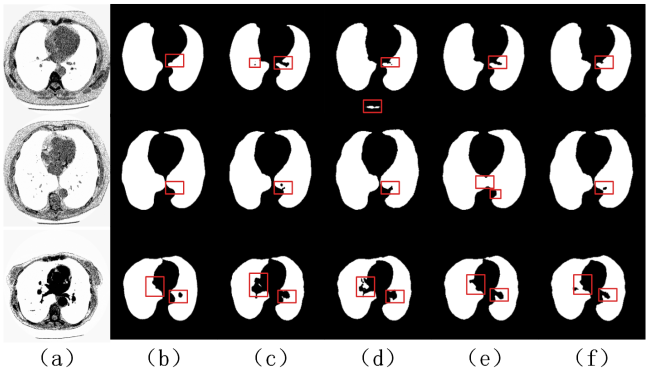


Fig. 2. Comparison of lung segmentation results. From left to right: (a) Input image. (b) Ground Truth. (c) U-Net segmentation results. (d) LadderNet segmentation results. (e) ResModelW/OConv segmentation results. (f) MufiNet segmentation results.

TABLE III

Experimental results of the proposed method for skin cancer lesion segmentation and comparison with U-Net and LadderNet methods.

Method	AC	SE	SP	JS	AUC
U-Net	0.9314	**0.9479**	0.9263	0.9314	0.9371
LadderNet	0.9422	0.8792	0.9532	0.9422	0.9428
ResModelW/OConv	0.9502	0.8733	0.9644	0.9502	0.9487
MufiNet	**0.9510**	0.8708	**0.9650**	**0.9510**	**0.9492**

Some sample output from the test phase is shown in Fig. 2. U-Net, LadderNet, ResModelW/OConv and the proposed model are almost perfect in the task of lung segmentation. The proposed model shows state-of-the-art performance for sensitive areas. Observing the first line of Fig. 2, we can see that LadderNet may segment some discrete areas incorrectly. ResModelW/OConv and the proposed model avoid the discrete regions caused by the wrong classification of some regions and concentrate the target regions together. From the second and third lines, we can see that in order to avoid the holes in the lung, U-Net and LadderNet are more conservative for the performance of boundary segmentation. ResModelW/OConv and the proposed model greatly avoid this kind of interference. However, in the second line of Fig. 2, the ResModelW/OConv model incorrectly links the left and right lungs, and the proposed model successfully achieves finer segmentation details.

b) Skin Cancer Lesion Segmentation: The skin lesion dataset contains melanoma images and non-melanoma images. For non-melanoma images, the segmentation area is less obvious, and other interference items may be included in the melanoma images. So in this implementation, we set the initial number of channels to 30, and learn more features by increasing the number of filters. Set the batch size to 16, and set the learning rate as 0.01, 0.001 on the 0th and 20th epoch, respectively.

Table III shows the quantitative results of different methods in the skin cancer segmentation task. In terms of accuracy, specificity, sensitivity, JS and AUC, we compare the proposed model with U-Net, LadderNet and ResModelW/OConv. MufiNet produces the highest accuracy, SP, JS, and AUC for this task, and also produces a high SE. It is easy to produce higher SE or SP if only one type of prediction, while other indicators are based on the prediction of two types of prediction to evaluate the whole model. ResModelW/OConv and the proposed model generate the next highest and highest accuracy, SP, JS, and AUC for this task, respectively, and a higher SE. It is easy to produce higher SE or SP if only one type of prediction, while other indicators are based on the prediction of two types of prediction to evaluate the whole model. ResModelW/OConv and the proposed model have the second highest and highest accuracy, JS and AUC. Therefore, compared with U-Net and LadderNet, ResModelW/OConv and the proposed model achieve better performance in the task of segmentation of complex medical images with large differences in regions of interest and sensitive boundaries.

Observing the first and second rows of Fig. 3, we can

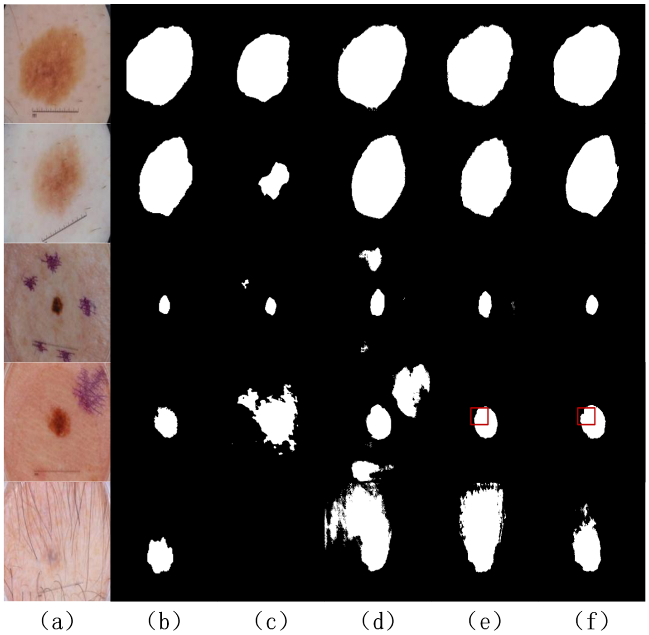


Fig. 3. Comparison of skin cancer segmentation results. From left to right: (a) Input image. (b) Ground Truth. (c) U-Net segmentation results. (d) LadderNet segmentation results. (e) ResModelW/OConv segmentation results. (f) MufiNet segmentation results.

see that the U-Net is still overly conservative and cannot be identified for areas with less cancerous changes. LadderNet, ResModelW/OConv and the proposed network perform better. But the boundaries drawn by LadderNet are too smooth. Observe that in the third and fourth lines, for images with more than one spot, U-Net and LadderNet will incorrectly segment some interference terms, and cannot obtain continuous segmented regions, but instead obtain a set of several segmented regions. The proposed network is almost perfect. It is worth mentioning that from the input image of the fifth line, the segmentation results of U-Net, LadderNet and ResModelW/OConv are not ideal, because the difference between foreground and background is too small even the human eye can hardly distinguish the region of interest. U-Net is even worse, it can not make any prediction, and can not segment the pathological area completely. The proposed model is not perfect, but its performance is far better than other models. It can be seen that for more challenging medical images, the proposed model performs better. It further demonstrates the reliability and robustness of the proposed model.

c) Cervical cancer Segmentation: Cervical cancer segmentation is very important for analyzing diseases related to cervical cancer. For example, in colposcopy, the main focus is the cervical region, so cervical cancer segmentation and cervical cancer pattern classification can be applied to identify other problems. In this implementation, set the batch size to 32 and set the learning rate as 0.001, 0.0001 on the 0th and 20th epoch, respectively.

TABLE IV
EXPERIMENTAL RESULTS OF THE PROPOSED METHOD FOR CERVICAL CANCER SEGMENTATION AND COMPARISON WITH U-NET AND LADERERNET METHODS.

Method	AC	SE	SP	JS	AUC
U-Net	0.9496	0.9203	0.9630	0.9496	0.9407
LadderNet	0.9532	0.9222	0.9725	0.9532	0.9434
ResModelW/OConv	0.9551	0.9267	0.9729	0.9551	0.9498
MufiNet	**0.9572**	**0.9295**	**0.9746**	**0.9572**	**0.9521**

Table IV shows the quantitative results of this experiment compared with U-Net, LadderNet and ResModelW/OConv in the cervical cancer segmentation task. The proposed model achieves the highest in five indicators: accuracy, specificity, sensitivity, JS and AUC. The experimental results show that the indicators of the proposed model are better than other models.

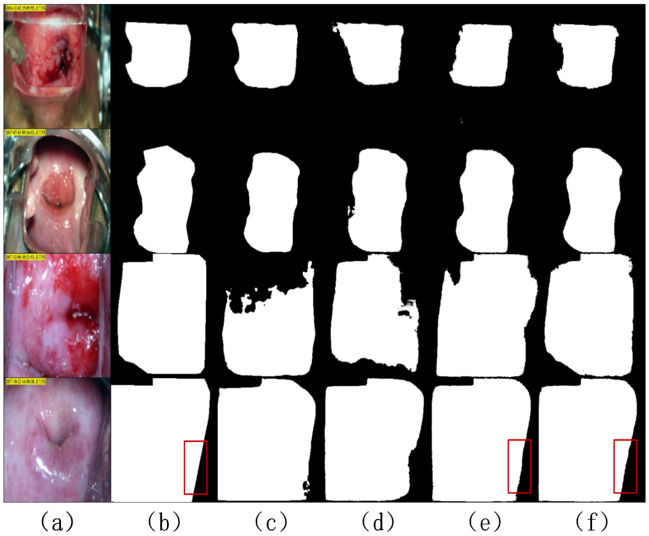


Fig. 4. Comparison of cervical cancer segmentation results. From left to right: (a) Input image. (b) Ground Truth. (c) U-Net segmentation results. (d) LadderNet segmentation results. (e) ResModelW/OConv segmentation results. (f) MufiNet segmentation results.

From the first two lines of Fig. 4, we can clearly observe that U-Net, LaddeNet, ResModelW/OConv and the proposed model can effectively segment the cervical region. But U-Net is more conservative and the segmentation result is too smooth. LadderNet will appear too smooth or over-segmentation. ResModelW/OConv performed unsatisfactorily in the first line of Fig. 4 due to the influence of interferences, and the segmentation results of the network we proposed are almost consistent with ground truth. From the third and fourth lines, it can be observed that when U-Net, LadderNet and ResModelW/OConv are large for the target area of interest, the phenomenon of insufficient or excessive segmentation will occur. However, the proposed network performs better in both cases and achieves more accurate target positioning. The proposed network can reach the highest AUC (0.9521), which is 1.14% , 0.87% and 0.23% higher than U-Net, LadderNet and ResModelW/OConv respectively. This also proves that the proposed network achieves higher performance and robustness in end-to-end image segmentation tasks.

V. Conclusion

In this paper, we propose MufiNet for semantic segmentation by fusing the features learned from different scale receptive fields and adding residual connection. It is helpful to extract complex features, which are essential for the boundary delineation of medical images, and are indispensable for edge-sensitive image segmentation tasks. We evaluated the performance of MufiNet on three different medical image segmentation tasks, including lung segmentation, skin cancer lesion segmentation, and cervical cancer segmentation. Compared with U-Net and LadderNet, MufiNet shows the best performance on all three datasets. It segments finer image boundaries and has stronger robustness. MufiNet can also be used in other semantic segmentation tasks.

Acknowledgment

This work was supported by National Key R&D Program of China 2018AAA0100500, NSFC Grant No.61832008, 61751211,61772413, 61802299.

References

[1] G. J. S. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. V. Der Laak, B. Van Ginneken, and C. I. Sanchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," pp. 234–241, 2015.

[3] O. Oktay, J. Schlemper, L. L. Folgoc, M. C. H. Lee, M. P. Heinrich, K. Misawa, K. Mori, S. Mcdonagh, N. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv: Computer Vision and Pattern Recognition*, 2018.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[4] 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: Computer Vision and Pattern Recognition*, 2018.

[5] J. Zhuang, "Laddernet: Multi-path networks based on u-net for medical image segmentation," *arXiv: Computer Vision and Pattern Recognition*, 2018.

[6] C. Galleguillos and S. Belongie, "Context based object categorization: A critical survey," *Computer Vision and Image Understanding*, vol. 114, no. 6, pp. 712–722, 2010.

[7] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," pp. 6230–6239, 2017.

[8] L. 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[9] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv: Neural and Evolutionary Computing*, 2013.

[10] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," pp. 3431–3440, 2015.

[11] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[12] S. Jegou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," pp. 1175–1183, 2017.

[13] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," pp. 5168–5177, 2017.

[15] H. Zhang, K. J. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context encoding for semantic segmentation," pp. 7151–7160, 2018.

[16] J. Zhuang, J. Yang, L. Gu, and N. C. Dvornek, "Shelfnet for fast semantic segmentation," *arXiv: Computer Vision and Pattern Recognition*, 2018.

[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," pp. 1–9, 2015.

[18] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," pp. 833–851, 2018.