Tumor Segmentation Based on Deeply Supervised Multi-Scale U-Net

Lei Wang[1,2,⋆], Bo Wang[1,2,⋆], Zhenghua Xu[1,2,†]
[1]State Key Laboratory of Reliability and Intelligence of Electrical Equipment,
Hebei University of Technology, China
[2]Key Laboratory of Electromagnetic Field and Electrical Apparatus Reliability of Hebei Province,
Hebei University of Technology, China
⋆Co-first authors, contributed to this work equally
†Corresponding author, email: zhenghua.xu@hebut.edu.cn

***Abstract*—Although deep learning has achieved great success in the field of medical image processing, the existing deep learning based medical image segmentation solutions still cannot obtain satisfactory performances for abdominal small organs and lesions due to their small object size and shape-variability. In this work, a Deeply Supervised Multi-Scale U-Net (DSMS U-Net) is proposed for more accurate segmentation performances on abdominal small organs images. DSMS U-Net integrate the existing U-Net model with a restoration decoder module and some multi-scale convolution modules. Our experiment results demonstrate that the proposed DSMS U-Net approach has much better segmentation performances than the state-of-the-art baselines.**

***Index Terms*—Tumor Segmentation, Multi-Scale, Deep Supervision**

I. Introduction

With the considerable growth of computer image processing technology, computer-assisted medical images diagnosis and treatment has grown rapidly in the last few decades [1]. Particularly, the success of deep learning technology for image classification, object detection, and semantic segmentation, has also attracted increasing attention in medical image analysis. The semantic segmentation for organ and its tumor medical image are regarded as an important step in the clinic medical image analysis process. For example, segmenting a tumor target from patients images is a critical step for radiotherapy treatment, measurement of chemotherapy and surgery.

In recent years, we have witnessed significant segmentation works on medical organs image by CNNs architectures. Ronneberger et al. [2] trained the U-shaped network (U-Net) end-to-end from neural structures in electron microscopic stacks and achieved well performance for segmentation. Drozdzal et al. show that a low-capacity fully convolution network (FCN) [3] can serve as a preprocessor to obtain normalized images, which are then iteratively refined by a Fully Convolutional Residual Networks (FC-ResNets) to generate an improving segmentation prediction on CT images of liver lesions [4]. And Xue et al. [5] proposed a novel adversarial critic network to force the critic and segmentor to learn both global and local features. As a result, Segan framework leads to better performance than U-Net for its effectivity and stability.

Although U-Net has become the state-of-the-art method on the medical image segmentation task for its accuracy and efficiency, the original U-Net with the relatively simple encoder-decoder structure is inappropriate for the segmentation on abdominal small organs. This is because the small organs and their tumor in CT image, compared with the common object in natural image segmentation tasks, are much smaller against the whole background. And U-Net will fail to capture the sufficient associated features of global and local details due to the impacts of spatial resolution. In this work, to achieve more accurate segmentation in abdominal small organ images, we propose a novel image segmentation method.

To solve the above problems, we propose DSMS U-Net that integrates the existing U-Net model with a restoration decoder module and some multi-scale convolution modules for more accurate segmentation performances on abdominal small organ images. The advantages of our proposed method are shown as follows: (i) Multidimensional features extracting are supervised by adding a restoration decoder to the bottom of the model. So that the principal components of the image are better abstracted from the sparse and specific images, and more effective features are encoded for segmentation tasks. (ii) Based on the structure of U-Net, multi-scale convolutional modules are proposed to increase the receptive density during feature maps fusion. Then the model more effectively captures the global information and fine-grained details of the foreground regions when high-dimension features from the encoder network are gradually enriched prior to fusion with the corresponding semantically rich feature maps from the decoder network. Therefore, the proposed method captures the features of different scale. Our experiment reveals that the proposed method yield better performance over the state-of-the-arts.

The contribution of this work is provided as follows: Firstly, we present several problems of segmentation tasks on abdominal small organ images. Secondly, to meet the challenges of segmentation, which are the individual specificity and low proportion target region, we propose a novel model with deeply supervised multi-scale U-net. Thirdly, our method experiment is conducted on two CT datasets. And we only present the results achieved on the pancreas dataset due to the limited page.

II. Related Work

Some researchers point out improved CNNs-based methods, whose structure has multi-layers on medical segmentation tasks. FCN and U-Net are both improved versions of CNNs, which start the research domain named semantic segmentation end to end. FCN changed its last layer with the convolution layer. Additionally, up-sampling and skip connection [6] are applied to segment target by classification at the pixel level. Apparently, for lack of relationships among global pixels information, FCN obtains a fuzzy and smooth result, which is not refined enough. U-Net contains a compressed path to capture semantic features and an extension path to locate accurately. This is the trade-off between higher resolution and more abstract features. Compared with FCN, U-Net presents a fine segmentation ability via several combinations between low level-features and deep features. However, there are unsolved problems around the appropriate depth of U-Net for different datasets and space for improvement on the features extracting of local details and global information.

To deal with the problem of model depth for U-Net, Zhou et al. [7] filled the framework on every layer and cut off the redundant network on the process of training. Then U-Net++ has more abundant and stable representation ability but a large number of parameters. To learn more principal features and discard useless information or noise, Mao et al. [8] built a convolutional auto-encoders with symmetric skip connections. And the architectural decisions [9] were designed based on the Hebbian principle and the multi-scale processing, to optimize the quality of features extraction. The classication and detection tasks were achieved via utilizing the one-side convolution [10] to adjust the channels of feature maps in the network. Another pooling mode, like spatial pyramid pooling, was used to eliminate the impact of interest regions scale and achieve the deformation robust [11]. He et al. proved that SPP-net is also significant in classification and detection by generating a fixed-length representation. And Zhao et al. [12] exploited the global context information through the pyramid pooling module together with the proposed pyramid scene parsing network (PSPNet). After that, based on the PSPNet, Lin et al. [13] pointed out the Feature Pyramid Network (FPN), whose top-down architecture with lateral connections was developed for building high-level semantic feature maps. Then, Men et al. [14] applied PSPNet into rectal tumors CT images for segmentation task. They applied the proposed CNN with atrous convolution and spatial pyramid pooling (SPP) module to extract high-resolution features, meanwhile maintaining large receptive fields for tumors of different sizes. In spite of that, PSPNet extracts the spatial location information with not sufficient abilities enough, because the pyramid pooling module focuses more on the existence of image features rather than location. And we think that the rich location details are significant for accurate segmentation of the medical image.

III. Methods

The proposed framework mainly contains the body of U-Net, restoration decoder and several multi-scale convolutional modules. Our proposed framework is illustrated in Fig. 1.

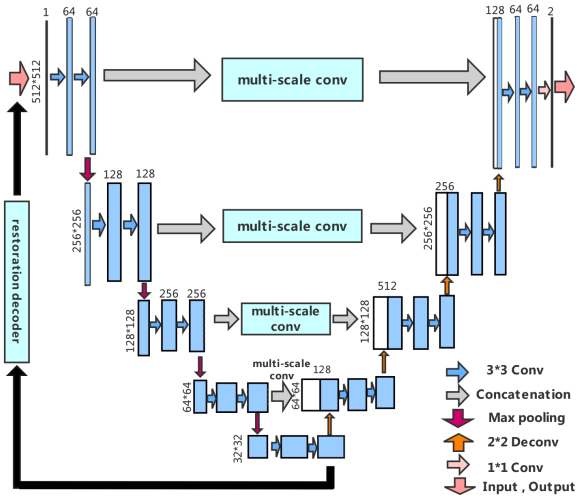


Fig. 1. Deeply Supervised Multi-Scale U-Net architecture.

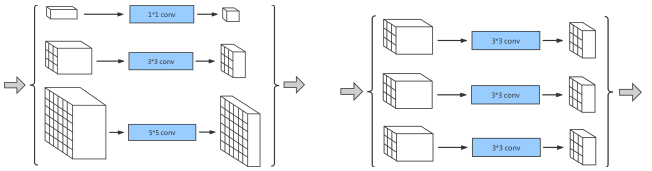
A. Network architecture

In this work, our segmentation framework is built based on U-Net, and consisted of a restoration decoder and a chain of multi-scale convolutional modules. An restoration decoder is introduced to restore input images from the bottom feature maps. Then the abilities of feature presentation of encoder branch are strengthened by minimizing loss function between the input images and restoration images. Moreover, multi-scale convolutional modules are adopted instead of pooling in our proposed network. Since pooling, to some extent, tend to discard useful details that are essential for segmentation. The multi-scale convolutional modules capture the features on regions of different sizes, and the modules are added before concatenation on every level. Based on the above architecture, Decoder Restoration Supervised U-Net (DRS U-Net), Decoder Restoration Supervised U-Net with 3×3 Kernel Convolution Module (KCMDRS U-Net), and DSMSC U-Netare is conducted respectively in this work.

B. Restoration decoder module

There are decoder models that contain up-sampling and convolution layers [15] for semantic segmentation in recent years. Besides the up-sampling branch on the right of U-Net, another independent restoration decoder branch with deconvolution is built to restore input images and calculate loss function for training.

In order to improve the extracting abilities of the left branch of U-Net, we adopt the deep supervision, that it is conducted by a restoration decoder, to the bottom of U-Net. The feature maps are decoded to the same size as input with deconvolution operation. There are two layers of 3×3 kernel convolution and one 2×2 kernel deconvolution on every level in the restoration decoder module. Then reconstruction loss is calculated between the input images and restoration images. Then, the proposed network is trained to obtain more efficient and rapidly convergent performance.



(a) Different size kernel convolutional module (b) Same size kernel convolutional module

Fig. 2. Multi-Scale Convolutional Module

C. Multi-scale convolution module

The repeated down-sampling in U-Net reduces the spatial resolution of feature maps and decline the fine connection among global-local information. In order to strengthen the capture fusion of global and local context without changing the image resolution, we place a multi-scale convolutional module before the concatenation on every level. Those multi-scale convolutional modules are composed of one-side convolution, 3×3 kernel and 5×5 kernel convolution, which apparently can efficiently overcome the problem of variable target sizes. The different size kernel convolutional module architecture is shown in Fig. 2(a). Compared with multi-scale max-pooling, the convolution operation preserves more pixels location information for accurate segmentation for its less information discard. In order to analyze whether the reason of work is the multi-scale design of convolution or wider network, we contrast the multi-scale convolutional module with parallel 3×3 kernel convolutional module, whose architecture is illustrated in Fig. 2(b). Then we can analyze the feature capture abilities of multi-scale convolution for segmentation and discriminate the representation of multi-scale convolution from that of the single size of kernel module.

D. Objective Optimization

Generally, there are two loss functions, which are termed as $Loss1$ and $Loss$ 2 and illustrated in Fig. 3. $Loss$ 1 is calculated with the prediction and manual annotation, while $Loss$ 2 present the distance between input images and restoration images.

We introduce the Binary of Cross-Entropy (BCE) as the loss function of $Loss$ 1 and $Loss$ 2, which is described as:

$$L(Y, \hat{Y}) = -\frac{1}{N}\sum_{i=1}^{N}(Y_i log\hat{Y}_i + (1 - Y_i)log(1 - \hat{Y}_i)), \quad (1)$$

where Y_i and $\hat{Y}_i$ denotes the ground truth and prediction output $F(X_i; \Theta)$ of i_{th} image respectively, and N indexes the batch size. The final loss optimization is obtained through weighted summing $Loss$ 1 and $Loss$ 2 by a ratio of 10 to 1.

IV. Experiments

We have conducted experimental studies on a kidney and a pancreas CT image datasets. However, due to the limited page, we only present the results achieved on the pancreas dataset. Two existing models, FCN and U-Net, and three proposed models, DRS U-Net, KCMDRS U-Net, and DSMSC U-Net, are conducted by using Pytorch for qualitative and quantitative evaluations. The training, validation, and testing of all models are performed by using an NVIDIA GeForce GTX 1080Ti GPU.

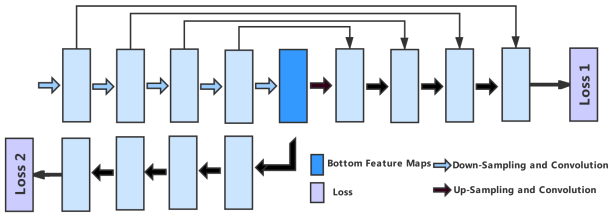


Fig. 3. Illustration of the segmentation loss and reconstruction loss.

A. Data and pre-processing

The pancreas CT images, from 283 patients (Decathlon-10) with pancreas tumor, are divided into three subsets for training (70%), validation (10%), and testing (20%). Because of the remaining target region values on red and blue channels, before training the networks, the mask of targets (pancreas and its tumor) is generated based on manual annotation. That means that there are only the target organ and its tumor with fixed values, while other regions are allocated value with zero. Thus the mask of the manual annotation is applied as our label rather than original annotation image.

B. Implementation details

FCN and U-Net have become the state-of-the-art methods on image segmentation for their high accuracy and efficiency. Therefore, we choose U-Net and FCN as the baseline models. Moreover, our proposed DSMSC U-Net and other four networks are conducted respectively in our experiment. DRS U-Net is designed to compare with U-Net for the analysis on deep supervision, and KCMDRS U-Net is implemented to compare with DRS U-Net and our DSMSC U-Net.

The bottom feature maps in Fig. 3 is up-sampled to 16 fold image, which is as same size as the original input image. There are five layers whose every step include 3×3 kernel convolution, one-side convolution, and deconvolution operation. Considering the computation and capacity, the multi-scale convolutional module is built based on the parallel construction of one-side convolution, 3×3 kernel convolution, and 5×5 kernel convolution. The feature maps obtained the same size via appropriate padding. After that, feature maps are concatenated in the channel dimension. Then one-side convolution operation is utilized to adjust the number of channels matching the extension path. And the improvement of the presentation benefits from the one-side convolution operation for nonlinearity mapping. Moreover, in order to as minimizing the variance of each observation as possible to get more stable convergence, we adopt a trick of gradient accumulation on the platform Pytorch for the limited memory.

C. Results and Discussion

As summarized in Fig. 4 and Table IV-C, we respectively present the qualitative and quantitative evaluation.

Two pancreas slices prediction on two rows are shown for five models in this paper. As shown in Fig. 4, U-Net and FCN have weakly performance so that the tumor regions can not be segmented, and its tumor is misregarded as part of the

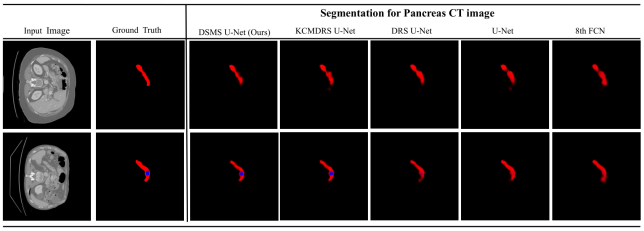


Fig. 4. The qualitative evaluation.

Table 1 The quantitative evaluation in terms of Dice, PPV, and Sensitivity

Model	Pancreas				
	FCN	U-Net	DRS U-Net	KCMDRS U-Net	DSMSC U-Net
Dice	0.8165	0.8052	0.8183	**0.8336**	0.8283
PPV	0.9073	0.8806	0.8874	0.8962	**0.8970**
Sensitivity	0.8516	0.8642	0.8725	0.8899	**0.9013**

pancreas. Then we add the 3×3 kernel convolution operation before concatenation on every level, for the reason of filter classification by the convolution operation. So that KCM U-Net presents a better prediction on region recognition. And DRS U-Net gains the segmentation performance in forms of much more adequacy on region recognization, that is the organ and its tumor correctly predicted with larger area. As we can see that MSCM U-Net has more accurate edge segmentation than KCM U-Net. Then the multi-scale convolutional module and restoration decoder are combined to generate our proposed method. All models' performance is evaluated by Dice Similarity Coefficient (DSC) [16], Positive Predictive Value (PPV), and Sensitivity. As shown in Table IV-C, the further convolution operation capture and filtrate the objects in feature maps to have more decline background distractions and more accurate interested regions. Moreover, DSMSC U-Net improve all the three indexes with that DSC is 82.83%, PPV is 89.70%, and Sensitivity is 90.13%. Our network can well segment the pancreas and its tumor with various size and shape, particularly for the very small tumor and special shape pancreas. Meanwhile, DSMSC U-Net can even well segment some organs and its tumor for the sharp boundary in very low contrast against the background.

V. Conclusion

To achieve more accurate segmentation, we proposed DSMSC U-Net. Fig. 4 demonstrates that there is an obvious mistake on pancreas tumor segmentation when adopting FCN and U-Net. From the perspective of segmentation details, KCMDRS U-Net and our DSMSC U-Net can obtain more accurate results at the boundary of target regions. Particularly, the multi-scale convolutional module is important for tumor segmentation because the tumors of different patients are of different sizes. In order to analyze whether the reason of network work is the multi-scale design of convolution operation or wider network, we contrast the multi-scale convolutional module with three parallel 3×3 kernel convolution operation. And we notice that the the multi-scale design of convolutional module has a more apparent effect on tumor segmentation. In spite of there are shape-variability and weak tissue contrast in pancreas images, our proposed DSMSC U-Net gets the obvious improvement when comparing with the baselines. Our experiment results demonstrate that the proposed approach has better performance than the state-of-the-art baselines.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under grant 61906063, in part by the Natural Science Foundation of Tianjin, China, under grant 19JCQNJC00400, and in part by the Yuanguang Scholar Fund of Hebei University of Technology, China.

References

[1] Y. Zhou, L. Xie, E. K. Fishman, and A. L. Yuille, "Deep supervision for pancreatic cyst segmentation in abdominal ct scans," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2017, pp. 222–230.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.

[3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[4] M. Drozdzal, G. Chartrand, E. Vorontsov, M. Shakeri, L. Di Jorio, A. Tang, A. Romero, Y. Bengio, C. Pal, and S. Kadoury, "Learning normalized inputs for iterative estimation in medical image segmentation," *Medical image analysis*, vol. 44, pp. 1–13, 2018.

[5] Y. Xue, T. Xu, H. Zhang, L. R. Long, and X. Huang, "Segan: Adversarial network with multi-scale l 1 loss for medical image segmentation," *Neuroinformatics*, vol. 16, no. 3-4, pp. 383–392, 2018.

[6] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation," in *Deep Learning and Data Labeling for Medical Applications*, 2016, pp. 179–187.

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

[8] X.-J. Mao, C. Shen, and Y.-B. Yang, "Image restoration using convolutional auto-encoders with symmetric skip connections," *arXiv preprint arXiv:1606.08921*, 2016.

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[10] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[11] 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.

[12] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2881–2890.

[13] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125.

[14] K. Men, P. Boimel, J. Janopaul-Naylor, H. Zhong, M. Huang, H. Geng, C. Cheng, Y. Fan, J. P. Plastaras, E. Ben-Josef *et al.*, "Cascaded atrous convolution and spatial pyramid pooling for more accurate tumor target segmentation for rectal cancer radiotherapy," *Physics in Medicine & Biology*, vol. 63, no. 18, p. 185016, 2018.

[15] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1520–1528.

[16] W. R. Crum, O. Camara, and D. L. Hill, "Generalized overlap measures for evaluation and validation in medical image analysis," *IEEE transactions on medical imaging*, vol. 25, no. 11, pp. 1451–1461, 2006.