2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Surgical instrument segmentation based on multi-scale and multi-level feature network

Yiming Wang[1], Zhongxi Qiu[2], Yan Hu[2], Hao Chen[1], Fangfu Ye[3], Jiang Liu[2]

Abstract— **Surgical instrument segmentation is critical for the field of computer-aided surgery system. Most of deep-learning based algorithms only use either multi-scale information or multi-level information, which may lead to ambiguity of semantic information. In this paper, we propose a new neural network, which extracts both multi-scale and multi-level features based on the backbone of U-net. Specifically, the cascaded and double convolutional feature pyramid is input into the U-net. Then we propose a DFP (short for Dilation Feature-Pyramid) module for decoder which extracts multi-scale and multi-level information. The proposed algorithm is evaluated on two publicly available datasets, and extensive experiments prove that the five evaluation metrics by our algorithm are superior than other comparing methods.**

I. Introduction

With the continuous development of science and technology, surgical robots and computer-aided surgery systems have gradually become important clinical tools. Segmentation of surgical instruments is an important task in the field of computer-aided surgery (CAS) system. The goal of image semantic segmentation is to give each pixel a category label, which belongs to the underlying image perception problem and is used as an intermediate task for instrument tracking, pose estimation and surgical phase estimation. It is critical to improve the surgeon's environmental awareness during the operation, thus high-accuracy surgical instrument segmentation is the fundamentals for the CAS system.

Deep-learning-based methods have proved their effectivity in natural and medical image segmentation fields. Fully Convolution Network (FCN) [1] usually addresses the semantic segmentation task and achieves superior results among some segmentation benchmarks. But it downsamples input images by stride convolutions and/or spatial pooling layers, resulting in a final feature map with low resolution. Wu et.al. improved the FCN for its high computational complexity as Rethinking Dilated Convolution in the Backbone for Semantic Segmentation (FastFCN) [2]. U-net [3] is another widely applied algorithm in medical image segmentation, which upsamples for 4 times and uses skip connection in the same stage. U-net ensures that more low-level feature maps are fused. Based on U-net [3], M-net [4] is proposed to adopt multi-level semantic information and eliminate the need of any post-processing step to become an end-to-end structure.

Most of present algorithms are only based on a single type of information, such as M-net and U-net [3] base on multi-scale information, and FCN [1] series base on multi-level information, which may lead to ambiguity of semantic information. In this paper, the method of combining multi-level and multi-scale is adopted. we also adopt the U-net as the backbone for the surgical instrument segmentation. Since the featurized image pyramids of the U-net [3] and its improvements, are used as their inputs, which increase the time considerably. For the symmetry structure characteristic of U-net [3], it does not consider the multi-scale information, which is helpful for segmentation. Thus, in this paper, we propose to adopt multi-level features as the input to reduce the computational complexity. Then a dilated convolution is adopted in the network to extract the multi-scale information for segmentation.

Therefore the contributions of the paper are concluded as: 1) We propose a new deep-learning based algorithm for surgical instrument segmentation, which adopts the spatial multi-scale and multi-level information. 2) Based on the backbone of U-net [3], we propose a DFP module, short for Dilation Feature-Pyramid module for decoder to extract multi-scale and multi-level features. The feature-pyramid is adopted as the input of our proposed algorithm. 3) We prove the effectivity of the proposed algorithm on two public datasets, including an Endoscopic vision dataset and a cataract surgical dataset.

II. Proposed Method

For the surgery instrument segmentation, we propose a new deep-learning based algorithm, as shown in the Fig. 1. U-net [3] is adopted as the primary structure. To extract more features for the network, multi-level features are considered as the input in encoder. Then we propose a DFP module, short for Dilation Feature-Pyramid module for decoder structure based on depthwise-seperable convolution [5], [6] to capture multi-scale feature and multi-level feature. The details of the proposed framework are illustrated in the following.

A. Encoder

To reduce the semantic loss caused by downsampling, [7] introduces multi-scaled image as inputs to provide semantic

Yiming Wang and Zhongxi Qiu are the co-first authors.

[1] Eye Hospital and School of Ophthalmology & Optometry, School of Biomedical Engineering, Wenzhou Medical University

[2] Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, 518055, China

[3] Wenzhou Institute, University of Chinese Academy of Sciences, Wenzhou 325001, China

*This work was supported in part by The Science and Technology Innovation Committee of Shenzhen City (JCYJ20200109140820699 and 20200925174052004),Guangdong Basic and Applied Basic Research Foundation(2021A1515012195)

Corresponding author: Yan Hu (huy3@sustech.edu.cn), and Jiang Liu (liuj@sustech.edu.cn).

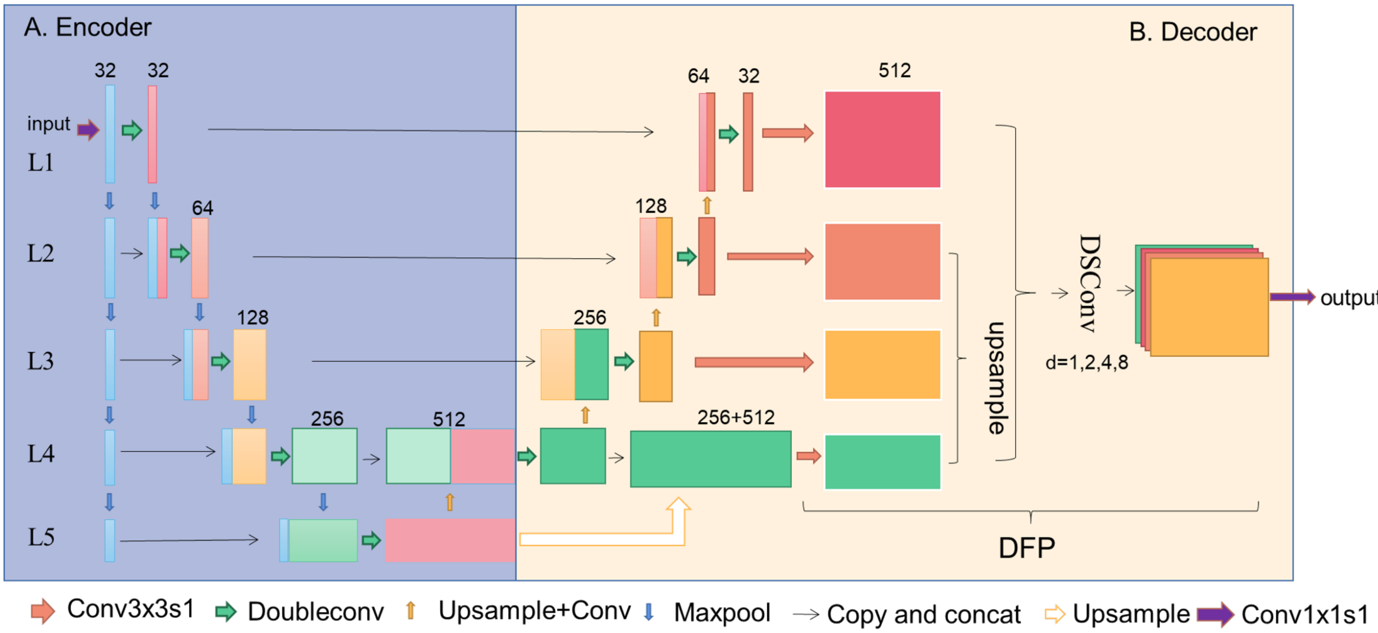

Fig. 1. The flowchart of our proposed algorithm.

context gain at each level. But this way increases the computation complexity greatly. Inspired by the [8], we build a feature-pyramid based encoder with lower computation complexity. As shown in Fig. 1 leftmost two columns, we first use conv1x1 with stride= 1 to produce 32-dimension feature. The feature pyramid is constructed by 5 layers. In order to reduce the semantic context loss caused by downsampling, every layer has a cascade of two maxpoolings, which include maxpooling the last layer feature and the double convolutional feature. The output of double convolutional feature is also the decoder input of the same level.

B. DFP for Decoder

The main structure of decoder is illustrated as Fig.1 (B) based on the U-net [3]. We propose a DFP module to extract multi-scale features of multi-level features in the decoder part. The first three-level (from L1 to L3) features connect to the decoder module directly, and the other two-level (L4 and L5) features are merged into one level as input to the DFP module of the decoder part. As shown in the Fig.1 (B), the last three-level features (from L2 to L4) are upsampled to the same size of the first-level L1 feature, 512 dimensions. The first layer is concatenated with other three upsampled layers to construct as 4 layers. They connect to DSConv block composed by four depth-wise separable convolutions [3] with different dilation rates, which extract different scale features from the input of decoder. As depth-wise separable convolution with dilation is formulated as:

$$F(x) = x \rightarrow SC_{dw}M \tag{1}$$

where C_{dw} is depth-wise separable convolution, extracting multi-scale features from the input. S and M are split and merge operations, respectively. DSConv block is expressed as:

$$\begin{aligned} F(x) = x \rightarrow & \left\{ \begin{matrix} SC_{dw}M \\ SC_{dw}M \\ SC_{dw}M \\ SC_{dw}M \end{matrix} \right\} \rightarrow concat \\ = & \left\{ \begin{matrix} x \rightarrow SC_{dw}M \\ x \rightarrow SC_{dw}M \\ x \rightarrow SC_{dw}M \\ x \rightarrow SC_{dw}M \end{matrix} \right\} \rightarrow concat \\ = & \left\{ \begin{matrix} x_0 \rightarrow C_{dw}M \\ x_1 \rightarrow C_{dw}M \\ x_2 \rightarrow C_{dw}M \\ x_3 \rightarrow C_{dw}M \end{matrix} \right\} \rightarrow concat \end{aligned} \tag{2}$$

where $concat$ delegates as concatenate operation. In our experiments, the depth-wise is valued as $d = 1, 2, 4, 8$, and four-scales features are extracted for surgery instrument segmentation. For the propose algorithm, cross-entropy is adopted as loss function during training, defined as:

$$L_{CE} = -\frac{1}{n}\sum_{i=0}^{n} y_i \log P_i \tag{3}$$

where n is the number of classes, y is the label with one-hot format, p represents the probability of class i.

III. Experiments

A. Dataset

In this paper, we use two publicly available datasets to evaluate the effectivity of our proposed algorithm.

- Rigid Instrument dataset: It is from MICCAI 2015 endoscopic vision challenge-instrument segmentation and tracking sub-challenge [9]. This dataset consists of two sub-datasets, robotic and non-robotic. The training data for the non-robotic subdataset is formed by 4 laparoscopic colorectal surgeries with a total of 160 images, and the test data is formed by 140 images. The size of each image is 480×640. All the input images and labels are adjusted to 160×160 and normalized as preprocessing.

• Cataract surgical dataset: The instrument segmentation dataset is released as a part of the CATARACTS: Challenge on automatic tool annotation for cataract surgery [10]. The dataset consists of 25 different surgical fragments and each fragment is composed of about 200 frames. The size of each image is 540×960. In the experiment, the 3267 images from 18 fragments are used for training, 816 images from other 4 fragments are for verification and the rest 575 images from rest 3 fragments are for testing. All the input images are resized to 160×160 and normalized as preprocessing.

B. Evaluation Metrics

In the experiments, we list the following five metrics from common semantic segmentation to evaluate the proposed algorithm. In this paper, we consider the foreground and background as two classes. Let i be the foreground class and j be the background class. To be more convincing, the evaluation metrics are counted for the entire test set and the average parameters are listed in the tables. The evaluation metrics are defined as:

• Precision:

$$Precision = \frac{\sum_{k=0}^{z} n_{ii}^{k}}{\sum_{k=0}^{z} n_{ii}^{k} + \sum_{k=0}^{z} n_{ji}^{k}}$$

• Recall:

$$Recall = \frac{\sum_{k=0}^{z} n_{ii}^{k}}{\sum_{k=0}^{z} n_{ii}^{k} + \sum_{k=0}^{z} n_{ij}^{k}}$$

• Accuracy:

$$Accuracy = \frac{\sum_{k=0}^{z} n_{ii}^{k} + \sum_{k=0}^{z} n_{jj}^{k}}{z}$$

• F1-Score:

$$F1 - Score = \frac{2 \times \sum_{k=0}^{z} n_{ii}^{k}}{2 \times \sum_{k=0}^{z} n_{ii}^{k} + \sum_{k=0}^{z} n_{ji}^{k} + \sum_{k=0}^{z} n_{ij}^{k}}$$

• IoU:

$$IoU = \frac{\sum_{k=0}^{z} n_{ii}^{k}}{\sum_{k=0}^{z} n_{ij}^{k} + \sum_{k=0}^{z} n_{ji}^{k} + \sum_{k=0}^{z} n_{ii}^{k}}$$

where n is the number of pixels, and z is the sum number of the image.

C. Implementation Details

The proposed algorithm is implemented with pytorch framework, which runs on a workstation equipped with NVIDIA TITAN V GPU. For the training parameters, we set SGD optimization with the batch size 5 with a pair of images as input, and the lr is set as 0.01.

TABLE I
THE RESULTS OF EVALUATION METRICS FOR ABLATION STUDY.

FP Input	DFP	Precision	Recall	Accuracy	F1-Score	IoU
x	x	0.843	0.729	0.962	0.782	0.642
✓	x	0.857	0.783	0.960	0.818	0.693
x	✓	0.869	0.842	0.966	0.855	0.747
✓	✓	**0.922**	**0.878**	**0.976**	**0.899**	**0.817**

D. Ablation Study

There are two improvements of our proposed algorithm, the feature-pyramid as input of encoder and the DFP module for decoder. Thus for the ablation study, we prove the two improvements step by step. The ablation study is based on the Rigid Instrument dataset [9], and results are show at Table I. In the table, the two improvements of FP Input and DFP stand as the feature-pyramid as input and dilation feature pyramid module respectively. The ✓means including the improvement, and x means not containing the improvement. The first line in the Table I is the results of our backbone U-net. The higher evaluation metrics in the second and third lines express that both the two improvements are helpful to improve the segmentation accuracy. The last line including both two improvements is the proposed algorithm in this paper, which produces the superior evaluation metrics, proving its effectivity for instrument segmentation.

Some segmentation experimental examples of ablation study shown at Fig. 2, which further proves the effectivity of our improvements. The details inside red square frame in the figure further emphasize that our algorithm improves the segmentation results.

E. Comparison Experiments

To prove the effectivity of our proposed algorithm, it is compared with other 6 related algorithms, including FCN-8s [1], FCN-16s [1], FCN-32s [1], M-net [7], Pyramid Scene Parsing Network (PSPnet) [11] and DeepLabv3 [12]. For DeepLabv3 architecture, we use Resnet101 to be the backbone. The parameters as illustrated their papers are applied in the experiments. Moreover, our proposed algorithm adopts 5-layer-pyramid input, thus we improve the M-net with 5-layer input named as M-net 5l.

TABLE II
THE RESULTS OF EVALUATION METRICS ON THE CATARACT SURGICAL DATASET.

	Precision	Recall	Accuracy	F1-Score	IoU
FCN-8s	0.731	0.783	0.984	0.756	0.608
FCN-16s	0.682	0.643	0.979	0.662	0.490
FCN-32s	0.667	0.623	0.978	0.644	0.478
M-Net	0.740	**0.824**	0.985	0.780	0.608
M-Net 5l	0.783	0.705	0.984	0.742	0.590
PSPnet	0.578	0.449	0.972	0.506	0.330
DeepLabv3	0.470	0.754	0.965	0.579	0.408
AttU-net	0.770	0.690	0.983	0.728	0.572
Ours	**0.814**	0.775	**0.987**	**0.794**	**0.659**

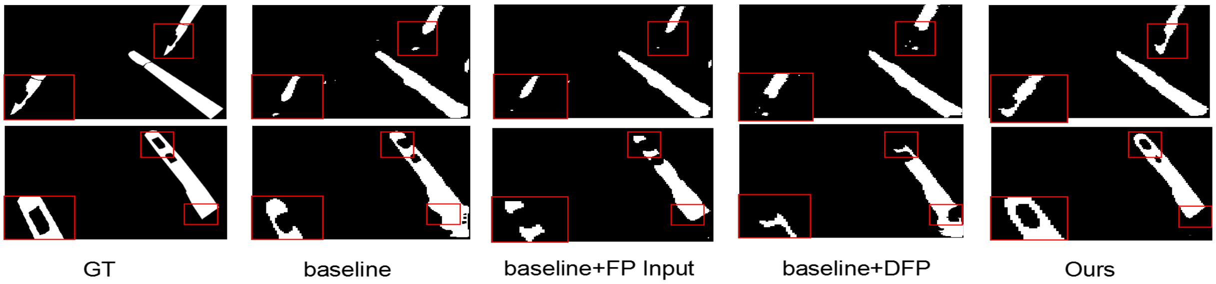

Fig. 2. From left to right: the segmentation groundtruth, segmentation results by the baseline, the baseline plus one improvement and the last our proposed algorithm with two improvements. We give priority to enlarging the details when dealing with more than one difference in the same picture.

TABLE III
THE RESULTS OF EVALUATION METRICS ON THE RIGID INSTRUMENT TEST SET.

	Precision	Recall	Accuracy	F1-Score	IoU
FCN-8s	0.759	0.820	0.949	0.788	0.51
FCN-16s	0.756	**0.900**	0.955	0.822	0.698
FCN-32s	0.836	0.627	0.942	0.716	0.558
M-Net	0.670	0.824	0.938	0.772	0.670
M-net 5l	0.918	0.828	0.971	0.871	0.771
PSPnet	0.770	0.500	0.925	0.606	0.435
DeepLabv3	0.690	0.641	0.925	0.665	0.498
AttU-net	0.778	0.851	0.955	0.813	0.686
Ours	**0.922**	0.878	**0.976**	**0.899**	**0.817**

The evaluation metrics of our algorithm and other comparison methods based on the two datasets are listed in the Table II and III, respectively. In the table, although Recall by our algorithm is a little higher than that by FCN-8s or FCN-16s, most of the evaluation metrics by our algorithm are better than those by all other algorithms. For visual evaluation, sample images from the two datasets are shown in the Fig. 3. The the red square frame of figures by FCN-8s and FCN-16s express that they cannot segment the instruments details correctly, but the segment results by our algorithm are better.

IV. CONCLUSION

In this paper, we proposed a new neural network algorithm for surgical instrument segmentation, which extracted both the multi-scale and multi-level features. It adopted the feature pyramid instead of image pyramid to reduce the computational complexity. Then the proposed DFP (short for Dilation Feature-Pyramid) module extracted multi-scale and multi-level features for segmentation. The five evaluation metrics of ablation study expressed the effectivity of the two improvements. We also compared the proposed algorithm with other algorithms based on two datasets. Both the evaluation metrics and the segmentation samples proved its superiority.

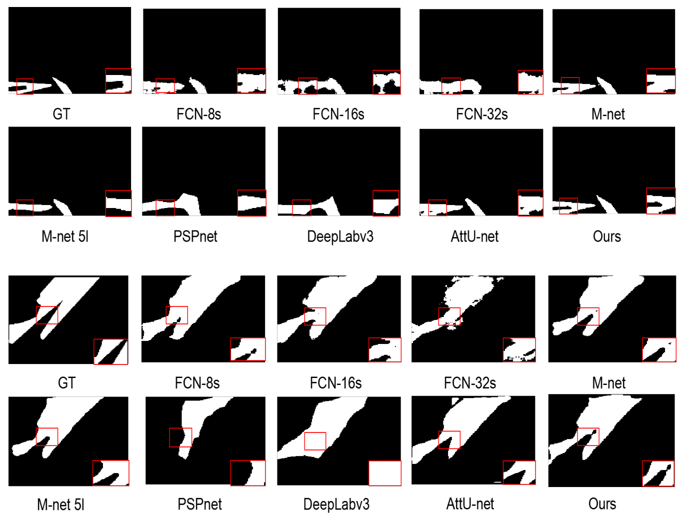

Fig. 3. Segmentation samples. The first two lines are some results of the caratact surgical dataset, and the last two lines are results of the rigid instrument test dataset.

REFERENCES

[1] Evan Shelhamer, Jonathan Long, and Trevor Darrell. *Fully Convolutional Networks for Semantic Segmentation.* IEEE Computer Society, 2017.

[2] Huikai Wu, Junge Zhang, Kaiqi Huang, Kongming Liang, and Yu Yizhou. Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation. In *arXiv preprint arXiv:1903.11816*, 2019.

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[4] Raghav Mehta and Jayanthi Sivaswamy. M-net: A convolutional neural network for deep brain structure segmentation. In *IEEE International Symposium on Biomedical Imaging*, pages 437–440, 2017.

[5] Francois Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[6] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[7] Huazhu Fu, Jun Cheng, Yanwu Xu, Damon Wing Kee Wong, Jiang Liu, and Xiaochun Cao. Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. *IEEE transactions on medical imaging*, 37(7):1597–1605, 2018.

[8] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[9] Miccai. `http://endovissub-instrument.grand-challenge.org/`.

[10] Miccai. `https://cataracts.grand-challenge.org/`.

[11] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[12] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017.