

PAPER • OPEN ACCESS

Semantic Segmentation of Surgical Instruments based on Enhanced Multi-scale Receptive Field

To cite this article: Yu Dong *et al* 2021 *J. Phys.: Conf. Ser.* **2003** 012006

View the [article online](#) for updates and enhancements.



240th ECS Meeting

Digital Meeting, Oct 10-14, 2021

We are going fully digital!

Attendees register for free!

REGISTER NOW



Semantic Segmentation of Surgical Instruments based on Enhanced Multi-scale Receptive Field

Yu Dong¹, Hongbo Wang^{1,2,3,4,5}, Jingjing Luo¹, Zhiping Lai¹, Fuhao Wang¹ and Jiawei Wang¹

¹Academy for Engineering and Technology, Fudan University, Shanghai 200433, China

²Parallel Robot and Mechatronic System Laboratory of Hebei Province, Yanshan University, Qinhuangdao 066004, China

³Intelligent Robot Engineering Research Center of Ministry of Education, Shanghai 200433, China

⁴Shanghai intelligent robot engineering technology research center, Shanghai 200433, China

⁵Corresponding Author: Hongbo Wang. E-mail: Wanghongbo@fudan.edu.cn; 19210860071@fudan.edu.cn.

Abstract. With the rapid development of robot assisted surgery, the segmentation of surgical instruments becomes more and more important. However, compared with the natural scene segmentation, surgical instrument segmentation is more difficult. To solve this problem, we improve a high and low resolution fusion module, which aims to extract detail information and context information from the fusion feature map of high and low resolution. Then, in the last layer of the encoder, we propose the Enhanced Multi-scale Receptive Field module to generate more available receptive fields. Our method is validated on 2017 MICCAI EndoVis Robotic Instrument Segmentation Challenge dataset, and the result is better than the other methods. The extended experiment is carried out on the dataset of our surgical soft robot which has a content implementation.

1. Introduction

In robot assisted surgery, surgical video plays a very important role. Doctors need to get the position of the current surgical instruments in the operation video to accurately control the execution path of the surgical instruments in the human body [1]. However it is very difficult to segment the surgical instruments, as it is in a dynamic environment, and there will be biological tissue [2]. For the surgical soft robot, on account of its own characteristics, it is more vulnerable to the deformation caused by the collision and extrusion of biological tissues in the body.

The segmentation of surgical instruments is mainly divided into three tasks: binary segmentation, types segmentation and parts segmentation. Since the first two are used to segment the whole surgical instrument, and cannot distinguish one part of the surgical instrument [3], the parts segmentation is more practical for the use of automatic robots in surgery. And combined with our actual situation, we segment three parts of the soft robot, so we choose the third segmentation task.

Our contributions include:



Content from this work may be used under the terms of the [Creative Commons Attribution 3.0 licence](https://creativecommons.org/licenses/by/3.0/). Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

- In order to obtain local information and context information, a high and low resolution fusion module is proposed;
- For the purpose of getting more available receptive fields, we propose an enhanced multi-scale receptive field;
- we produced a dataset of our surgical soft robots for verifying the effect of our algorithm in a real scene.

2. Related Work

2.1. Multi-resolution fusion

In medical image segmentation, since the emergence of U-Net [4], many algorithms were improved based on U-Net [5][8]. What they had in common is that they used a symmetric skip connection between encoder and decoder to supplement the information lost in encoder. Although these methods had achieved good results, they weren't enough to recover the missing information generally. Recently, there were also some multi-resolution fusion methods, such as SwiftNet [9], FasterNet [10] which either used add or concat to fuse two feature maps with different resolutions, and then used a single convolution for further processing. Despite this kind of method obviously reduced the complexity, it couldn't fully mine the useful information after experimental verification.

2.2. Multi-scale feature aggregation

For semantic segmentation, especially for surgical instrument segmentation, the target objects will present different sizes, so we need to pay attention to the context information of different sub-regions in the image for obtaining higher segmentation accuracy. PSPNet [6] used pyramid pool to create different receptive fields, and then fused these receptive fields to obtain information from different sub-regions. DeepLabV3+ [7] and DenseASPP [12] consisted of parallel atrous convolution layers with different rates which could attend to multi-scale context information. These models had achieved top results in the field of semantic segmentation, which could get the context information of different size objects by using the pyramid architecture to generate multi-scale receptive field. Generally speaking, how to obtain multi-scale receptive fields is very important for semantic segmentation.

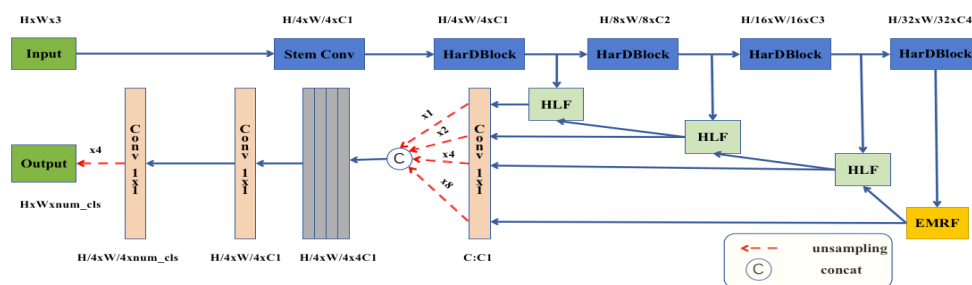


Figure 1. Architecture of the proposed Network

3. Methods

In this section, we firstly introduce our proposed network architecture, including the encoder and decoder. Then we introduce the High-Low resolution Fusion (HLF) which to obtain context information and local information, and the Enhanced Multi-scale Receptive Field (EMRF) that to further increase available receptive fields at a more granular level.

3.1. Architecture of the proposed Network

Our proposed network is based on encoder-decoder architecture, as shown in figure 1. Inspired by FASSD-Net[11], we choose HardNet [13], which is a state-of-the-art network based on DenseNet [14], as encoder of our network. In the decoder, we directly use bilinear interpolation to upsample the

feature maps generated by EMRF module and HLF module of each layer to the quarter of the original image. Then these feature maps are concatenated, and two 1×1 convolutions are used to reduce channels to the number of categories followed by an upsampling layer. What's different between our decoder and FASSD-Net is that we directly upsample after feature fusion from each layer which can avoid the tedious operation of extracting features. And because our proposed HLF module and EMRF module can extract both local information and global information well, so our network architecture has a better effect.

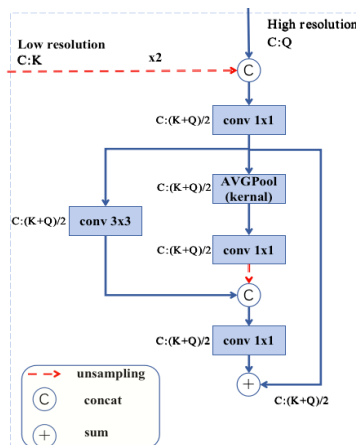


Figure 2. HLF: the detail of High-Low resolution Fusion.

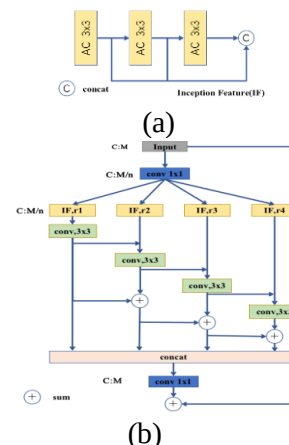


Figure 3. EMRF: the detail of Enhanced Multi-scale Receptive Field, (a)(b) are the Inception Feature (IF) and the architecture of EMRF.

3.2. High-Low resolution Module

we design the High-Low resolution Fusion (HLF) module to acquire detail information and context information from the fusion feature map of high and low resolution feature maps base on FASSD-Net. As shown in figure 2, the low resolution feature map from the previous layer is upsampled by bilinear interpolation to get the same resolution as that from the encoder. Then they are concatenated to get the feature map with $(K+Q)$ channels. To reduce the amount of computation, a 1×1 convolution is applied to reduce the number of channels to $(K+Q)/2$. Next there are two parallel branches. One branch obtains local details by a 3×3 convolution while another branch uses average pooling with different strides to get context information and then restored to its original feature by upsampling. Then the two branches are concatenated followed by a 1×1 convolution to fuse feature. Finally, a residual connection is used to avoid vanishing gradient. We reduce the amount of parameters by using average pooling instead of atrous convolution compared to FASSD-Net. Although average pooling will lose some information, the other branch uses 3×3 convolution which can supplement the lost details.

3.3. Enhanced Multi-scale Receptive Field Module

Our proposed EMRF module draws inspiration from CFPNet [15] which merges three convolution kernels into one channel, then apply skip connection to concatenate features that are obtained from each convolution block to create multi-scale feature maps. We use asymmetric convolutions (AC) instead of convolution as shown in figure 3(a). Then, we use a pyramid structure and atrous convolutions in a layer to get multi-scale receptive fields with rates are r_1 , r_2 , r_3 and r_4 as shown in figure 3(b). However, if the obtained feature maps are concatenated directly which may lead to insufficient information. Based on [19], we further enhance the receptive fields of different scales by using a 3×3 convolution on one branch which is produced by a different rate. Specifically, we use hierarchical residual connection to send the output feature of the current branch to the next 3×3 convolution along with the input feature of the next branch. This process is repeated some times until all branch's input features are transformed to output features. Additionally, considering the

checkerboard effect of pyramid atrous convolution, we apply the fusion method of HFF [16] to fuse the output features. Although the additional 3x3 convolutions increases the number of parameters, the final output feature maps can represent more scale features.

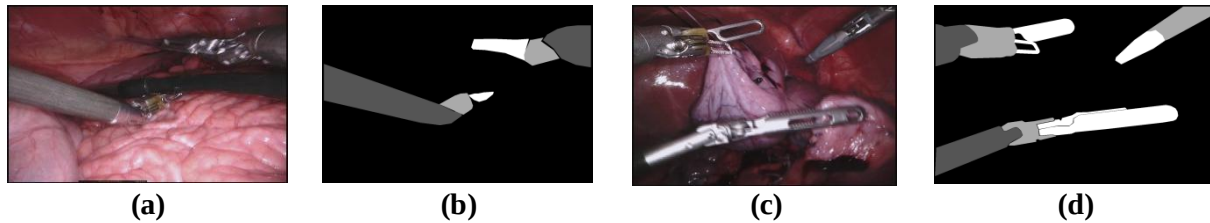


Figure 4. Examples of EndoVis 2017 dataset, (a)(c) are original images, (b)(d) are groundtruths of parts of surgical instrument.

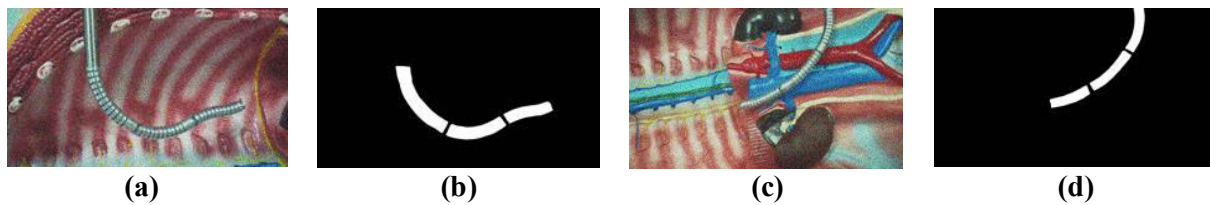


Figure 5. Examples of our dataset, (a)(c) are original images. (b)(d) are groundtruths of soft robot.

4. Experiments

4.1. Datasets and experimental configuration

The first dataset is EndoVis 2017 [17]. This dataset is obtained by da Vinci Xi robot in different surgical scenes which has eight videos and each containing 225 frames. The size of the original image is 1920x1080 as shown in figure 4. There are seven types of surgical instruments and each type of them is divided into three parts, namely axis, wrist, and buckle.

In our project, the soft surgical robot can not accurately perceive the shape during the complex surgical environment, so we need to obtain the curvature and Azimuth angle of the soft robot to further control it. For this purpose, we segment the soft robot and then reconstruct its main curve, so the second dataset of the surgical soft robot is created. We use a binocular camera to collect the changing shape of the soft robot in the abdominal cavity of a mannequin. The dataset has 1440 pictures in total which consists of eight videos, and the image size is 1280x720. As shown in figure 5, the soft robot we use is composed of three segments, and the curvature of each segment is same.

Our training model uses Python 1.8.1. Using Adam optimizer, the learning rate is 0.0001. Because we can't find the testing set's labels of the first dataset, and consider the actual situation of our second dataset, we divide all the data into two parts, one is the training set, the other is the testing set. And to test the generalization ability of the model in limited time and hardware resources, we use uniform distribution to generate two numbers randomly, which are used as the serial number of the testing set, and then the rest as the training set. The partition results are 2, 4, 5, 6, 7, 8 as the training set and the remaining two for testing. All loss functions adopt the cross-entropy loss function.

Table 1. The results compared with other methods

Method	Miou (%)	Dice (%)
LinkNet34	71.75	82.67
BiseNet	74.38	84.53
DeepLabV3+	76.54	86.0
FASSD-Net	76.16	85.82
Ours	76.80	86.32

4.2. Results and Comparison

We use Miou and Dice as the measurement criteria to evaluate our model. Firstly, we choose LinkNet34[5] that is classical in medical images, then BiseNet [18], DeeplabV3+[7] and FASSD-Net [11] are chosen, which have prominent semantic segmentation effect in other areas of network. To compare with other works fairly, we train the networks in the same experimental configuration. As can be seen from the result table 1, the accuracy of Linknet34, a classic network for medical image segmentation, can reach 71.75%. While compared with other methods, such as DeepLabV3+ which add the module of multi-scale receptive field in the network, its miou accuracy is up to 76.54%. It can also be seen that our method which uses the EMRF module has a significant improvement, and the accuracy is improved from 76.16% to 76.80%.

Table 2. Ablation Study of our methods

Method	Miou (%)
Baseline	76.16
+simple decoder	75.57
+simple decoder+EMRF	76.59
+simple decoder+EMRF+HLF	76.80

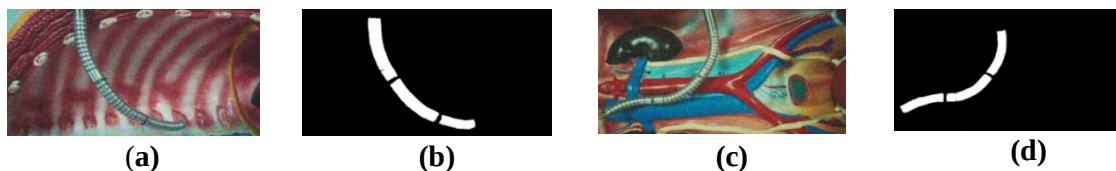


Figure 6. Visualization result examples of our dataset

4.3. Ablation Study

To evaluate the effectiveness of our improved module: HLF and EMRF, we compare the baseline we use, FASSD-Net, the result as shown in table 2. “+ simple decoder” indicates using our simple decoder. As you can see from the results, we start with a simple decoder straight away, compared to the baseline, our results decrease from 76.16% to 75.57%. But when we use EMRF module, the miou increases from 75.57% to 76.59%. At the same time, with the addition of HLF, the accuracy is also improved a little. Generally our improved modules can improve the accuracy of surgical instrument segmentation.

4.4. Generalization Experiment

We employ our model on the dataset of our surgical soft robot to verify the ability of our model in the actual scene. We use the same training as above. The accuracy of the test result reaches 91%, and some examples of segmentation results are shown in figure 6. And we can meet the requirement of the soft robot main curve reconstruction when we carry out the next operation of this result.

5. Conclusion

In this paper, we construct EMRF module to get more available receptive fields, so that we could better deal with objects of different sizes in surgical scene. We have also improved a module for multi-resolution fusion to obtain local information and context information. Through the improvement of the above two modules, we use a simpler decoder to restore the image size, and get a high score. In addition, we applied our model to the dataset of our surgical soft robot which could meet the requirements of the subsequent shape detection of soft robot. In future work, we will collect data in the body to verify the effect of our model, and speed it up for using in soft robotic laparoscopic surgery.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFB1311700, in part by National Natural Science Foundation of China under Grant U1713219, and in part by Fudan-ZiMaoYiHao Medical Device Joint Experimental Center Project under Grant SGH2310020/007. (Corresponding author: Hongbo Wang.)

6. References

- [1] Allan, M., Kondo, S., Bodenstedt, S., Leger, S., Kadkhodamohammadi, R., Luengo, I., ... & Speidel, S. (2020). 2018 robotic scene segmentation challenge. arXiv preprint arXiv:2001.11190.
- [2] Lin, S., Qin, F., Li, Y., Bly, R. A., Moe, K. S., & Hannaford, B. (2020). LC-GAN: Image-to-image Translation Based on Generative Adversarial Network for Endoscopic Images. arXiv preprint arXiv:2003.04949.
- [3] Shvets, A. A., Rakhlin, A., Kalinin, A. A., & Iglovikov, V. I. (2018, December). Automatic instrument segmentation in robot-assisted surgery using deep learning. In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 624-628). IEEE.
- [4] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham
- [5] Shvets, A. A., Rakhlin, A., Kalinin, A. A., & Iglovikov, V. I. (2018, December). Automatic instrument segmentation in robot-assisted surgery using deep learning. In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 624-628). IEEE.
- [6] Zhao, H. , Shi, J. , Qi, X. , Wang, X. , & Jia, J. . (2016). Pyramid scene parsing network. IEEE Computer Society.
- [7] Chen, L. C. , Zhu, Y. , Papandreou, G. , Schroff, F. , & Adam, H. . (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. Springer, Cham.
- [8] Isensee, F. , & Maier-Hein, K. H. . (2020). Or-unet: an optimized robust residual u-net for instrument segmentation in endoscopic images. arXiv e-prints.
- [9] Ori, M. , Kreo, I. , Bevandi, P. , & Egvi, S. . (2019). In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images.
- [10] Chen, W. , Gong, X. , Liu, X. , Zhang, Q. , Li, Y. , & Wang, Z. . (2019). Fasterseg: searching for faster real-time semantic segmentation.
- [11] Rosas-Arias, L. , Benitez-Garcia, G. , Portillo-Portillo, J. , Sanchez-Perez, G. , & Yanai, K. . (2021). Fast and Accurate Real-Time Semantic Segmentation with Dilated Asymmetric Convolutions. 2020 25th International Conference on Pattern Recognition (ICPR).
- [12] Yang, M. , Yu, K. , Chi, Z. , Li, Z. , & Yang, K. . (2018). DenseASPP for Semantic Segmentation in Street Scenes. CVPR.
- [13] Chao, P. , Kao, C. Y. , Ruan, Y. , Huang, C. H. , & Lin, Y. L. . (2020). HardNet: A Low Memory Traffic Network. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE.
- [14] Huang, G. , Liu, Z. , Laurens, V. , & Weinberger, K. Q. . (2016). Densely Connected Convolutional Networks. IEEE Computer Society. IEEE Computer Society.
- [15] Lou, A. , & Loew, M. . (2021). Cfpnet: channel-wise feature pyramid for real-time semantic segmentation.
- [16] Mehta, S. , Rastegari, M. , Caspi, A. , Shapiro, L. , & Hajishirzi, H. . (2018). Espnet: efficient spatial pyramid of dilated convolutions for semantic segmentation. Springer, Cham.
- [17] Allan, M. , Shvets, A. , Kurmann, T. , Zhang, Z. , & Azizian, M. . (2019). 2017 robotic instrument segmentation challenge.
- [18] Yu, C. , Gao, C. , Wang, J. , Yu, G. , Shen, C. , & Sang, N. . (2020). Bisenet v2: bilateral network with guided aggregation for real-time semantic segmentation.

- [19] Gao, S. , Cheng, M. M. , Zhao, K. , Zhang, X. Y. , & Torr, P. . (2019). Res2net: a new multi-scale backbone architecture. IEEE Transactions on Pattern Analysis and Machine Intelligence, PP(99), 1-1.