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

EMS-Net: Enhanced Multi-Scale Network for Polyp Segmentation

Miao Wang, Xingwei An*, Yuhao Li, Ning Li, Wei Hang and Gang Liu*

Abstract— **In recent years, polyp segmentation plays an important role in the diagnosis and treatment of colorectal cancer. Accurate segmentation of polyps is very challenging due to different sizes, shapes, and unclear boundaries. Making full use of multi-scale contextual information to segment polyps may bring better results. In this paper, we propose an enhanced multi-scale network for accurate polyp segmentation. It is composed of a multi-scale connected baseline (U-Net+++), a multi-scale backbone (Res2Net), three Receptive Field Block (RFB) modules, and four Local Context Attention (LCA) modules. Specifically, the baseline's multi-scale skip connections can aggregate features in both low-level and high-level layers. We have evaluated our model on three publicly available and challenging datasets (EndoScene, CVC-ClinicDB, Kvasir-SEG). Compared with other methods, our model achieves SOTA performance. It is noteworthy that our model is the only network that has achieved over 0.900 mean Dice on EndoScene and CVC-ClinicDB.**

I. INTRODUCTION

Colorectal cancer (CRC) is one of the most common malignant tumors in the world. It has the third-highest mortality rate among all cancers for many years. Studies have shown that 95% of colorectal cancers are caused by colorectal adenomatous polyps. Therefore, early detection of polyps becomes particularly important. Colonoscopy is considered to be the best diagnostic tool for early examination and removal of polyps. However, early colonoscopy annotation requires an endoscopist to perform it manually, which tests the doctor's ability and endurance during the operation. Therefore, automatic and accurate polyp segmentation has important clinical significance.

In recent years, with the rapid development of deep learning technology in a variety of computer vision tasks, polyp segmentation based on deep learning has also benefited. The Fully Convolutional Networks [5] is the pioneer of the auto image segmentation task, which replaces the full connection layer of the neural network with the convolution layer. Later, Ronneberger et al. introduced U-Net [6] for the task of medical image segmentation. It has a symmetrical U-shaped encoder-decoder architecture. The skip connection between encoder and decoder is responsible for enhancing the fusion of shallow and deep features to improve the segmentation performance. At present, U-Net [6] has become the most popular baseline in medical image segmentation. Inspired by the success of U-Net [6], U-Net++ [7] and ResUNet++ [8] expanded the original U-Net's architecture respectively. For example, U-Net++ [7] designed nested and dense skip connections to achieve the fusion of different features. And ResUNet++ [8] utilized residual learning mechanisms, attention modules, etc. They achieved better performance in the task of medical image segmentation. SFANet [9] used one encoder and two decoders to predict the region and boundary of polyps respectively, and it proposed a loss function to improve the region segmentation and boundary detection of polyps. In addition, PraNet [10] introduced at MICCAI 2020 used a parallel partial decoder (PPD) to combine features to obtain rich contextual information and leveraged reverse attention (RA) module to further mine the boundary cues, which has outperformed most cutting-edge models by a large margin. However, we believe that most methods do not make full use of multi-scale information and better attention mechanisms to deal with the size and shape of polyps and the boundaries of polyps.

Polyp segmentation needs a powerful multi-scale learning strategy and attention mechanism due to the challenging characteristics of polyps, such as size, shape, unclear boundaries, etc. Thus, we assume that the simultaneous processing of polyp regions and boundaries is the core of accurate polyp segmentation.

In this paper, we propose a new network called Enhanced Multi-Scale Network (EMS-Net) for accurate polyp segmentation. Firstly, we use a series of multi-scale methods to enhance the extraction of multi-scale features. Then we utilize the Local Context Attention (LCA) module [4] to increase attention to polyp boundaries. We have conducted three different experiments to verify our proposed model. Please refer to our experiments (Part.III) for more details.

In summary, the main contributions of our paper are as follows:

- We propose a novel multi-scale learning network EMS-Net, which is a medical image segmentation neural network for accurate polyp segmentation tasks. It uses Res2Net [2] as the backbone and U-Net+++ [1] as the baseline. So it is designed with an encoder-decoder architecture. In addition, our model takes advantage of the Receptive Field Block (RFB) module and Local Context Attention (LCA) module. Since we have introduced an enhanced multi-scale learning strategy, our model has a stronger ability to

This work was supported in part by National Natural Science Foundation of China (grant 81925020 and 81630051) and Tianjin Science and Technology Project of China (grant 20JCZDJC00620).

*Corresponding author: anxingwei@tju.edu.cn; liugang60@aliyun.com

Miao Wang, Xingwei An, Yuhao Li and Ning Li are with Academy of Medical Engineering and Translational Medicine, Tianjin University, Tianjin, 300072 China.

Xingwei An is also with Tianjin Center of Brain Science, Tianjin, 300072 China.

Wei Hang and Gang Liu are with Department of Otorhinolaryngology Head and Neck Surgery, Huanhu Hospital of Tianjin University, Tianjin, 300350 China.

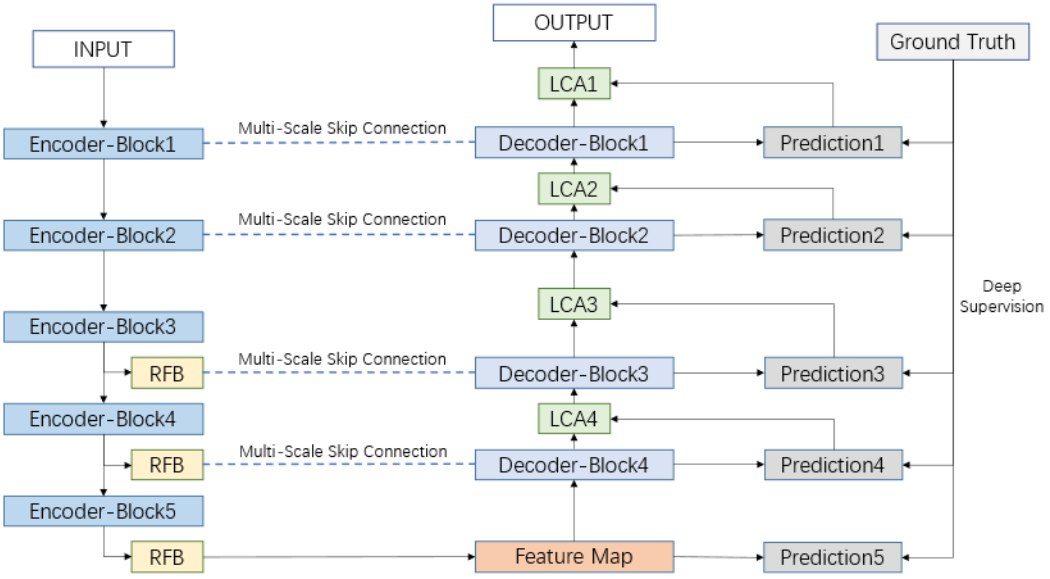

Fig. 1. Block diagram of the proposed EMS-Net architecture.

extract and learn polyp features to better cope with the challenges in polyp segmentation.

- Our experiments demonstrate that the proposed EMS-Net achieves state-of-the-art results on three widely used public and challenging datasets.

II. METHOD

A. Architecture

Fig. 1 shows the architecture of our EMS-Net, which is consisted of an encoder backbone and a decoder. We use Res2Net [2] as our encoder to extract fine-grained features efficiently, which contains five encoder blocks. Differently, the decoder branch only has four decoder blocks and a series of Local Context Attention (LCA) modules for accurate polyp boundary segmentation. Then the deepest feature map (the orange block) plays a role in multi-scale skip connections. Each decoder block is composed of five sub-blocks from EMS-Net and generates a prediction map with a different resolution. We adopt deep supervision for these prediction maps. Inspired by U-Net+++ [1], we use the RFB module to enhance its full-scale skip connections to obtain richer multi-scale features. Each component will be elaborated as follows.

B. Multi-Scale Skip Connection

We use the U-Net+++ [1] as the baseline for its advantage of multi-scale skip connections. Fig. 2 illustrates the construction of the multi-scale feature map of Decoder-Block4. Encoder-Block1, Encoder-Block2, and Encoder-Block3 apply different max pooling and convolution operations to obtain a feature map generated by Encoder-Block4 with the same size and channels. The Encoder-Block5 utilizes bilinear interpolation and convolution operation to get the corresponding feature map. Then, we concatenate these five feature maps together. After concatenation, we perform a 3×3 convolution operation, each of which is followed by batch normalization and then by a ReLU activation function. Eventually, we get the multi-scale feature map of the Decoder-Block4. This sufficient context information can improve the performance of segmentation.

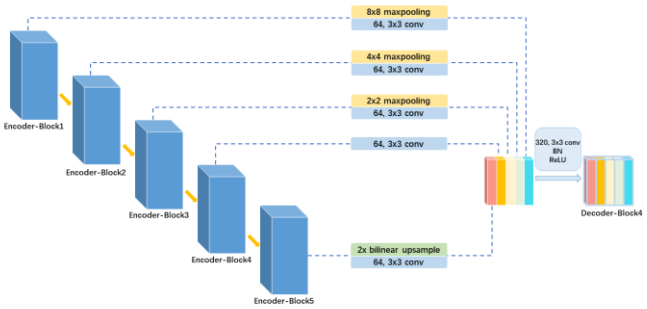

Fig. 2. Overview of the Multi-Scale Skip Connection.

C. RFB Module

The Receptive Field Block module [3] can generate different receptive fields to capture multi-scale information and enhance the deep features learned from the backbone. It consists of a multi-branch with different kernel size convolution and dilated convolution layers. Finally, it utilizes a 1×1 convolution to combine these features and produce the corresponding output. We apply the RFB module to the deep skip connections and after Encoder-Block5. On the one hand, it strengthens the transmitted deep features between the decoder and the encoder. On the other hand, it emphasizes the deepest features to better locate polyps.

D. LCA Module

The attention mechanism has been widely employed in semantic segmentation tasks. The attention mechanism determines which part of the neural network requires more

attention and enhances the quality of the features to promote the results. Therefore, we utilize the Local Context Attention (LCA) module [4] in the decoder branch. LCA module contains two inputs, the feature map generated by the decoder and its corresponding prediction map. LCA can better refine the multi-scale feature maps generated by the decoder to enhance the attention to the polyp boundary.

E. Loss Function

In binary cross-entropy (BCE) loss, it assigns equal weights to all pixels. However, in colonoscopy images, the number of pixels of some polyps is far less than the number of background pixels. It will make the model biased to the background and lead to poor results. Different from the standard BCE loss, the weighted BCE loss pays more attention to foreground pixels rather than background pixels.

In this paper, we utilize the combination of a weighted binary cross-entropy (BCE) loss and a Dice loss as the loss function. Therefore, our loss function is defined as follows:

$$\mathcal{L}_{\text{total}} = \lambda \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{Dice}}$$

where λ is set to 0.5.

III. EXPERIMENTS AND RESULTS

A. Dataset

Our three polyp segmentation datasets are the same as [10], namely EndoScene [12], CVC-ClinicDB [13], and Kvasir-SEG [14]. The three datasets have a total of 1762 images. The training set consists of 1540 images without data augmentation, including 900 images in Kvasir-SEG and 550 images in CVC-ClinicDB. And the testing set has the untrained remaining data from the three datasets mentioned above.

B. Implementation Details and Evaluation metrics

Our model is implemented in the PyTorch framework, which is trained and tested on an NVIDIA Tesla P100 GPU with 16GB memory. We train our model for 100 epochs using Adam optimizer, with a learning rate is of 1e-4 and batch size 8. All the images are resized to 352×352 and employ a multi-scale training strategy {0.75, 1, 1.25} rather than data augmentation.

In the task of polyp segmentation, most cutting-edge models[8,10] use two metrics (mean dice and mean IoU) for quantitative evaluation, so we also use these two metrics to evaluate our model. Moreover, [10] presents a publicly available and comprehensive benchmark for existing SOTA models. We also use the other four metrics (wFm, Sm, MAE, maxEm) mentioned in this benchmark for comparison so that we can better demonstrate the advantages of the performance of our model.

C. Results on the EndoScene Dataset

We compare our EMS-Net with UNet [6], UNet++ [7], SFANet [9], and PraNet [10] on the test set of EndoScene. As shown in Table 1, our method achieves the best performance over five metrics and outperforms other SOTA methods by large margins. Notably, our method is the only model that has achieved 0.900 mean Dice, a 3.33% improvement over the second-best algorithm. And we can also see that EMS-Net outperforms the PraNet [10] by 4.64% in mIoU. Especially, the dice coefficient and mIoU scores are important metrics for semantic segmentation tasks.

TABLE I. QUANTITATIVE RESULTS ON THE ENDOSCENE DATASET, COMPARING WITH OTHER STATE-OF-THE-ART METHODS.

Method	**mDice**	**mIoU**	**wfm**	**Sm**	**maxEm**	**MAE**
U-Net	0.710	0.627	0.684	0.843	0.876	0.022
U-Net++	0.707	0.624	0.687	0.839	0.898	0.018
SFA	0.467	0.329	0.341	0.640	0.817	0.065
PraNet	0.871	0.797	0.843	0.925	0.972	0.010
EMS-Net	0.900	0.834	0.885	0.943	0.969	0.006

TABLE II. QUANTITATIVE RESULTS ON THE CVC-CLINICDB DATASET, COMPARING WITH OTHER STATE-OF-THE-ART METHODS.

Method	**mDice**	**mIoU**	**wfm**	**Sm**	**maxEm**	**MAE**
U-Net	0.823	0.755	0.811	0.889	0.954	0.019
U-Net++	0.794	0.729	0.785	0.873	0.931	0.022
ResUNet	0.779	n/a	n/a	n/a	n/a	n/a
ResUNet++	0.796	0.796	n/a	n/a	n/a	n/a
SFA	0.700	0.607	0.647	0.793	0.885	0.042
PraNet	0.899	0.849	0.896	0.936	0.979	0.009
EMS-Net	0.923	0.874	0.923	0.949	0.974	0.008

TABLE III. QUANTITATIVE RESULTS ON THE KVASIR-SEG DATASET, COMPARING WITH OTHER STATE-OF-THE-ART METHODS.

Method	**mDice**	**mIoU**	**wfm**	**Sm**	**maxEm**	**MAE**
U-Net	0.818	0.746	0.794	0.858	0.893	0.055
U-Net++	0.821	0.743	0.808	0.862	0.910	0.048
ResUNet	0.791	n/a	n/a	n/a	n/a	n/a
ResUNet++	0.813	0.793	n/a	n/a	n/a	n/a
SFA	0.723	0.611	0.670	0.782	0.849	0.075
PraNet	0.898	0.840	0.885	0.915	0.948	0.030
EMS-Net	0.897	0.842	0.889	0.915	0.943	0.026

D. Results on the CVC-ClinicDB Dataset

On CVC-ClinicDB dataset, we compare our EMS-Net with UNet [6], UNet++ [7], ResUNet [11], ResUNet++ [8], SFANet [9] and PraNet [10]. The evaluation results of all the models are listed in Table 2. It shows that our method achieves the best performance over five metrics. Similarly, our method is the only model that has achieved 0.900 mean Dice, with Dice of 0.923. Then EMS-Net achieves a mIoU of 0.874 which is 2.94% higher than PraNet [10]. Fig. 3 shows the qualitative results on CVC-ClinicDB test set. A careful visual analysis of the result shows that EMS-Net produces better segmentation results as compared to the PraNet. Our model outputs a better boundary and the prediction is more accurate.

E. Results on the Kvasir-SEG Dataset

Table 3 presents the quantitative results on Kvasir-SEG dataset. We have compared our method with UNet [6], UNet++ [7], ResUNet [11], ResUNet++ [8], SFANet [9] and PraNet [10]. It shows that EMS-Net achieves the best results on four metrics. Both we and PraNet [10] have achieved SOAT performance on this dataset and outperformed other baseline methods by large margins. Some qualitative results on Kvasir-SEG test set are shown in Fig. 4. From the figure, we can see that our model can precisely segment the polyp tissues in some challenging cases, such as small polyps and homogeneous regions.

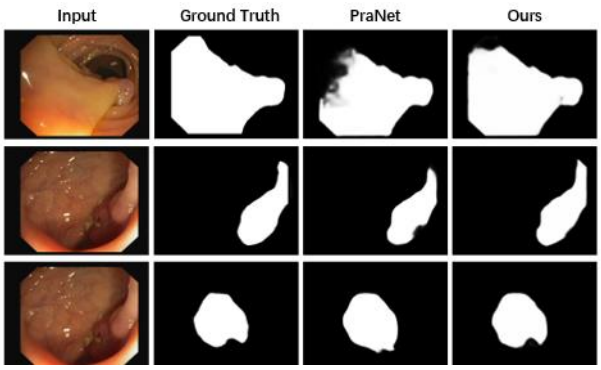

Fig. 3. Comparison of qualitative results between PraNet and EMS-Net on challenging images from CVC-ClinicDB.

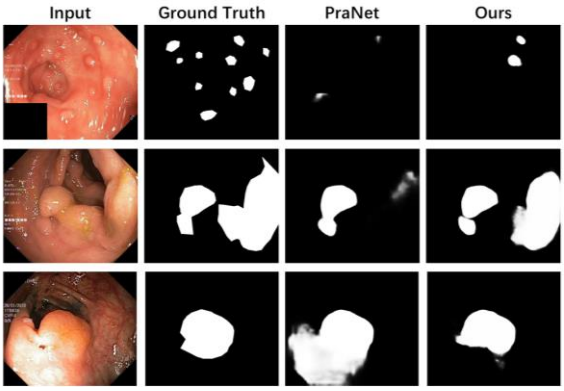

Fig. 4. Comparison of qualitative results between PraNet and EMS-Net on challenging images from Kvasir-SEG.

IV. CONCLUSION

In this paper, we suppose that more powerful and more effective multi-scale learning is essential to improve the auto segmentation of polyps. Based on this inspiration, we propose a novel encoder-decoder network named EMS-Net with a reinforced multi-scale skip connection to better capture fine-grained details and coarse-grained semantics. The RFB module was used for obtaining multi-scale receptive fields and enhancing deep skip connections. The LCA module enhanced the attention of polyp boundary. Experimental results on three public and challenging datasets demonstrate that our model achieves SOTA performance. Notedly, EMS-Net is the only model that has achieved over 0.900 mean Dice both on EndoScene and CVC-ClinicDB. In the future, we plan to optimize our model for greater performance. In this way, it can be used in more medical image segmentation tasks to contribute to the development of deep learning in the field of medicine.

REFERENCES

[1] H. Huang et al., "UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation," in ICASSP 2020 - 2020 IEEE International Con ference on Acoustics, Speech and Signal Processing (ICASSP), Barcel ona, Spain, May 2020, pp. 1055–1059.

[2] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A New Multi-scale Backbone Architecture," IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 2, pp. 652–662.

[3] S. Liu, D. Huang, and Y. Wang, "Receptive Field Block Net for Accurate and Fast Object Detection.," in Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XI, 2018, pp. 404–419.

[4] R. Zhang, G. Li, Z. Li, S. Cui, D. Qian, and Y. Yu, "Adaptive Context Selection for Polyp Segmentation," in Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, vol. 12266, A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, Eds. Cham: Springer International Publishing, 2020, pp. 253–262.

[5] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2015, pp. 3431–3440.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Cham, 2015, pp. 234–241.

[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 - 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings, 2018, pp. 3–11.

[8] D. Jha et al., "ResUNet++: An Advanced Architecture for Medical Image Segmentation.," in IEEE International Symposium on Multimedia, ISM 2019, San Diego, CA, USA, December 9-11, 2019, 2019, pp. 225–230.

[9] Y. Fang, C. Chen, Y. Yuan, and R. K.-Y. Tong, "Selective Feature Aggregation Network with Area-Boundary Constraints for Polyp Segmentation.," in Medical Image Computing and Computer Assisted Intervention - MICCAI 2019 - 22nd International Conference, Shenzhen, China, October 13-17, 2019, Proceedings, Part I, 2019, pp. 302–310.

[10] D.-P. Fan et al., "PraNet: Parallel Reverse Attention Network for Polyp Segmentation.," in Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part VI, 2020, pp. 263–273.

[11] Z. Zhang, Q. Liu, and Y. Wang, "Road Extraction by Deep Residual U-Net.," IEEE Geosci. Remote. Sens. Lett., vol. 15, no. 5, pp. 749–753, 2018.

[12] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," Computerized Medical Imaging and Graphics, vol. 43, pp. 99–111, Jul. 2015.

[13] D. Vázquez et al., "A Benchmark for Endoluminal Scene Segmentation of Colonoscopy Images," Journal of Healthcare Engineering, vol. 2017, pp. 1–9, 2017.

[14] D. Jha et al., "Kvasir-SEG: A Segmented Polyp Dataset," in MultiMedia Modeling, Jan. 2020, pp. 451–462.