HCMUS at MediaEval2021: PointRend with Attention Fusion Refinement for Polyps Segmentation

E-Ro Nguyen[1,3], Hai-Dang Nguyen[1,3], Minh-Triet Tran[1,2,3]
[1]University of Science, VNU-HCM, [2]John von Neumann Institute, VNU-HCM
[3]Vietnam National University, Ho Chi Minh city, Vietnam
{nero,nhdang}@selab.hcmus.edu.vn
tmtriet@fit.hcmus.edu.vn

ABSTRACT

The Medico task in MediaEval 2021 explores the challenge of building accurate and high-performance algorithms to detect all types of polyps in endoscopic images. This paper introduces our approach for the automatic segmentation of polyp images. We employ a ResNeXt as an encoder backbone with a UNet decoder. Further, the addition of PointRend and Attention Fusion Refinement on the network improves our segmentation performance. The experimental results show the efficiency of the proposed method, which achieves a Jaccard index of 0.7572, an accuracy of 0.9634, and a dice score of 0.8326.

1 INTRODUCTION

Medico: Transparency in Medical Image Segmentation 2021[6] task aims to develop automatic segmentation systems for segmenting polyps in images taken from endoscopies that are transparent and explainable, and reduce the chance that diagnosticians overlook a polyp during a colonoscopy. A modified version of the segmentation part of HyperKvasir [2] is given with more than 1000 training polyp images with their corresponding masks labeled by medical experts and 200 testing polyp images to challenge the participants for the robust, transparent, and efficient algorithms for polyp segmentation.

In recent years, the task of automatic polyp segmentation using deep learning-based [1, 3, 4] methods has gained a lot of achievements. Especially, the appearance of attention strategies [3] effectively improves polyp detection and segmentation performance. However, it still has some challenges, including the varieties of polyp's appearance (size, texture, and color). The boundary between a polyp and its neighbor regions is usually blurred and hard to be segmented.

In this paper, we propose an accurate and real-time framework **P**oint**R**end with **A**ttention **F**usion Refinement (PRAFNet) for the polyp segmentation. Fig. 1 shows the overview of our proposed framework. PRAFNet utilizes the Attention Fusion Refinement to decode an effective high-level semantic segmentation, and the PointRend [8] module to generate high-quality polyp segmentation from the colonoscopy images. The following section will introduce our approach and elaborate details about our network.

2 APPROACH

2.1 Attention Fusion Refinement

Current popular medical image segmentation networks usually rely on a U-Net architecture (e.g., U-Net [9], U-Net++ [13], ResUNet [7], etc). These models are essentially encoder-decoder frameworks, which aggregate all multi-level features extracted with a simple decoder, which does not effectively leverage these features. Woo et al. introduce a Convolutional Block Attention Module (CBAM) [11], which applies attention-based feature refinement with two distinctive modules, channel and spatial, to learn what and where to emphasize or suppress and refines intermediate features effectively.

We propose an Attention Fusion Refinement(AFR) module to better aggregate high-level features and focus on important regions, combining high-level features with upsampled features by CBAM as a core module. More specifically, for an input image, five levels of features $\{f_i, i = 1, .., 5\}$ can be extracted from a ResNeXt [5, 12] backbone network. We introduce a new decoder component, AFR, to aggregate the high-level features with upsampled features. As shown in Fig. 1, An AFR module inputs a high-level feature f_i with the previous upsampled feature d_{i+1} and we obtain the upsampled feature d_i.

2.2 PointRend

The U-Net [9, 13] model gives decent accuracy. However, it still has some drawbacks like predicting classes with very near distinguishable features, not being able to predict precise boundaries, etc. We have used the PointRend [8] module to address these drawbacks.

PointRend constructs point-wise features at selected points by concatenating two features, fine-grained to render fine segmentation details and coarse prediction features to gain more contextual and semantic information. We use the features f_2 as our fine-grained features and select top $K = 3136$ uncertain points in each subdivision step. In general, the uncertain points are located near the boundary of classes, so it can help refine the polyp's boundary effectively. As shown in Fig. 1, we use two subdivision steps of PointRend to obtain the final segmentation, which is the same size as the input image. We plot the uncertain points used in PointRend as blue dots in the coarse predictions m_2, m_1.

2.3 Training strategy

We apply the Bootstrapped Cross Entropy loss to prevent the models from overfitting on simple pixels and force them to focus on more challenging cases. With the Bootstrapped Cross Entropy, we calculate the loss for the top K percent pixels with the largest losses at each step in the training process. We would also add a "warm-up"

MediaEval'21, December 13-15 2021, Online

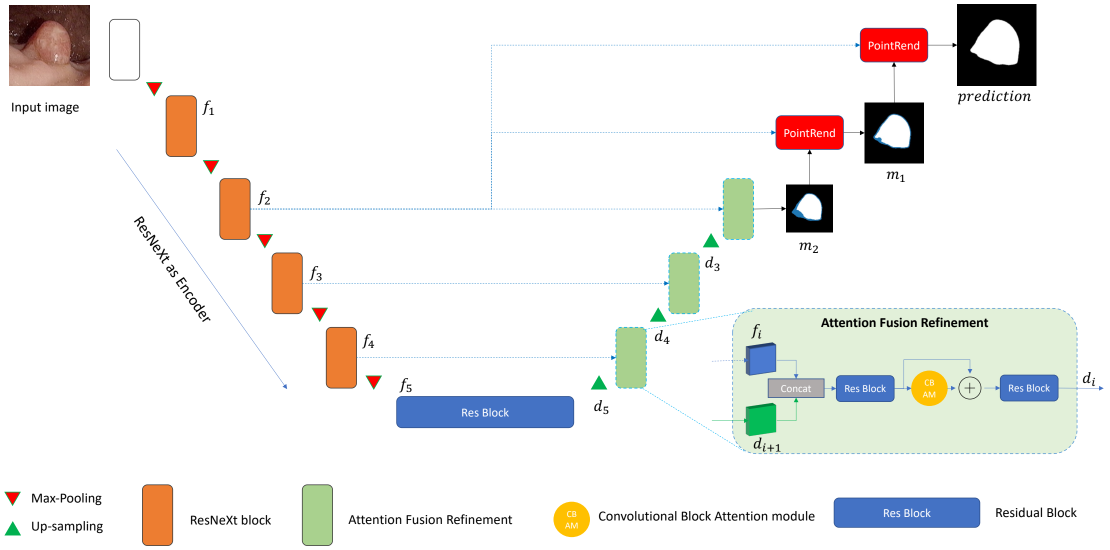

Figure 1: Overview of our proposed method PRAFNet, which consists of three attention fusion refinement (AFR) modules with two adaptive subdividion steps of PointRend module. Please refer to section 2 for more details.

Method	Acc	Jaccard	Dice	F1	R	P
2	0.9580	0.7252	0.8059	0.8059	0.7942	0.8871
3	0.9595	0.7283	0.8093	0.8093	0.7941	0.8831
4	0.9608	0.7441	0.8188	0.8188	0.8110	0.8741
5	0.9613	0.7497	0.8290	0.8290	**0.8352**	0.8639
6	**0.9634**	**0.7572**	**0.8326**	**0.8326**	0.8153	**0.8956**

Table 1: Medico polyp segmentation task 1's result. Acc denotes the accuracy, R and P denote the recall and precision, respectively.

Method	FPS	Accuracy	Jaccard	Dice	F1
1	**76.38**	0.9580	0.7210	0.8054	0.8054
4	47.86	**0.9608**	**0.7441**	**0.8188**	**0.8188**

Table 2: Medico polyp segmentation task 2's result.

period to the loss with $K = 100$ such that the network can learn to adapt to the easy regions first. Then transit to the harder areas by gradually decaying K to 15 in a polynomial manner.

3 RESULTS AND ANALYSIS

We performed experiments on six different settings for two tasks: Method 1 uses the UNet with ResNeXt50 [12] backbone as a baseline model. Method 2 extends Method 1 with the PointRend. Method 3 extends Method 2 with the Attention Fusion Refinement. Method 4 uses ResNeXt101 as a backbone with the same settings as Method 3. Method 5 uses EfficientNetB6 [10] as as backbone with the same setting as Method 3. Method 6 ensembles the results of Method 3, Method 4, and Method 5 together.

For task 1, we submit five runs from Method 2 to Method 6. For task 2, we submit two runs. In the first run, we use Method 4. And the second run is Method 1 for the lightweight architecture.

Table 1 and 2 shows our results on task 1 and task 2, respectively. Method 2 is slightly better than method 1 in all metrics, which shows that PointRend helps improve the results. In method 3, we use AFR, and the results also improve compared to method 2. With a stronger backbone (ResNeXt101 instead of ResNeXt50) in method 4, the results are improved with a Jaccard index of 0.7441. Method 5 with an EfficientNetB6 backbone is better than method 4 in several metrics except for precision. In method 6, we ensemble our methods 3, 4, 5 to achieve our best result in this task with the Jaccard index of 0.7572.

In task 2, although our method 1 is 1.5 faster than method 2, method 2 has higher accuracy with a real-time efficiency (**48 FPS**)

4 CONCLUSION

This paper presents a fast and accurate method for automatic polyps segmentation. The proposed methods use an encoder-decoder architecture. ResNeXt is used as an encoder backbone with the UNet decoder. Further, PointRend and Attention Fusion Refinement are applied to improve the segmentation result. PointRend helps refine the uncertainty points, especially with the boundary regions. The Attention Fusion Refinement enhances the fusion between high-level features and upsampled features in the decoder. In the future, we plan to apply better architecture such as ResUnet++ or PraNet for our work and further improve the results.

ACKNOWLEDGMENTS

This work was funded by Gia Lam Urban Development and Investment Company Limited, Vingroup and supported by Vingroup Innovation Foundation (VINIF) under project code VINIF.2019.DA19.

REFERENCES

[1] Mojtaba Akbari, Majid Mohrekesh, Ebrahim Nasr Esfahani, S.M.Reza Soroushmehr, Nader Karimi, Shadrokh Samavi, and Kayvan Najarian. 2018. Polyp Segmentation in Colonoscopy Images Using Fully Convolutional Network. *Conference proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference* 2018, 69–72. https://doi.org/10.1109/EMBC.2018.8512197

[2] Hanna Borgli, Vajira Thambawita, Pia H Smedsrud, Steven Hicks, Debesh Jha, Sigrun L Eskeland, Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, Dag Johansen, Carsten Griwodz, Håkon K Stensland, Enrique Garcia-Ceja, Peter T Schmidt, Hugo L Hammer, Michael A Riegler, Pål Halvorsen, and Thomas de Lange. 2020. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data* 7, 1 (2020), 283. https://doi.org/10.1038/s41597-020-00622-y

[3] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. 2020. PraNet: Parallel Reverse Attention Network for Polyp Segmentation. (2020). arXiv:eess.IV/2006.11392

[4] Yuqi Fang, Cheng Chen, Yixuan Yuan, and Kai-yu Tong. 2019. *Selective Feature Aggregation Network with Area-Boundary Constraints for Polyp Segmentation.* 302–310. https://doi.org/10.1007/978-3-030-32239-7_34

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. (2015). arXiv:cs.CV/1512.03385

[6] Steven Hicks, Debesh Jha, Vajira Thambawita, Hugo Hammer, Thomas de Lange, Sravanthi Parasa, Michael Riegler, and Pål Halvorsen. 2021. Medico Multimedia Task at MediaEval 2021: Transparency in Medical Image Segmentation. In *Proceedings of MediaEval 2021 CEUR Workshop.*

[7] Debesh Jha, Pia H. Smedsrud, Michael A. Riegler, Dag Johansen, Thomas De Lange, Pål Halvorsen, and Håvard D. Johansen. 2019. ResUNet++: An Advanced Architecture for Medical Image Segmentation. In *2019 IEEE International Symposium on Multimedia (ISM).* 225–2255. https://doi.org/10.1109/ISM46123.2019.00049

[8] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. 2020. PointRend: Image Segmentation as Rendering. (2020). arXiv:cs.CV/1912.08193

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. (2015). arXiv:cs.CV/1505.04597

[10] Mingxing Tan and Quoc V. Le. 2020. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. (2020). arXiv:cs.LG/1905.11946

[11] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. CBAM: Convolutional Block Attention Module. (2018). arXiv:cs.CV/1807.06521

[12] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated Residual Transformations for Deep Neural Networks. (2017). arXiv:cs.CV/1611.05431

[13] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. 2018. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. (2018). arXiv:cs.CV/1807.10165