Boundary-aware Pyramid Transformer for Polyp Segmentation

Jiacheng Wang[1], Yuxi Ma[1], Ruochen Mu[1] and Liansheng Wang[1]

[1]Department of Computer Science at School of Informatics, Xiamen University

Abstract

According to the World Health Organization (WHO), colorectal cancer(CRC) leads to a growing death rate in recent years, whose major cause comes from adenomatous polyps. Early polyp diagnosis can help to lower the incidence of CRC, which is achieved by colonoscopy as the gold standard. In this direction, polyp segmentation, on the other hand, is still a time-consuming and labor-intensive process. Although deep learning has made significant progress in the group of automatic polyp segmentation recently, however, these models have the following drawbacks: (i) lesions with small size are hard to detect since the pooling layers miss detailed contexts, and (ii) the boundaries are sometimes blurry and ambiguous which are extremely hard to determine. In this paper, we propose to equip the **P**yramid vision **Trans**former with **B**oundary-aware supervision, so-called **BP-Trans**, which can build multi-scale feature maps for dense prediction tasks and attentive boundary knowledge for precise boundary segmentation. We perform five-fold cross-validation on the Endoscopic computer vision challenges 2.0, in which the results on all metrics and folds consistently indicate the advantage of our method.

Keywords

Transformer, Boundary-aware Supervision, Polyp Segmentation

1. Introduction

Colorectal cancer(CRC) is the third most prevalent cause of cancer mortality worldwide, with more than 1.85 million cases and 850,000 deaths per year [1], whose major cause is owing to adenomatous polyps. The numbers can bring a more intuitive feeling: $50\%70\%$ of colon cancer comes from adenoma, and the cancer rate of adenomatous polyps is $2.9\%.4\%$ [2]. Colonoscopy is a vital medical screening technique for the illnesses of the lower digestive system. It may be used to check for intestinal polyps, bleeding, intestinal blockage, and the exclusion of lesions. With the developing agreement on artificial intelligence, people's interest has shifted to the subject of health as deep learning-based polyp segmentation is able to aid in clinician diagnosis. In this group, CNNs have made significant progress in the application of numerous imaging applications and automatic polyp segmentation is a popular topic among them. Thanks to the strong ability of infeasible and robust feature representation, FCN[3], U-Net[4], U-Net++[5], DoubleU-Net[6] and ResUNet[7] series, etc, have good results compared with traditional methods.

However, these methods have certain limitations. In general, polyp lesions have closed hue with the patients' own intestinal environment so that their appearance will be obviously different under different environmental conditions. As a result, during model training, the significant association between color and polyp segmentation tends to be overly concentrated, which is harmful to model training. Wei et al.[8] present the color exchange (CE) operation as a solution to this problem. They also propose the Probability correction method (PCS), which can improve positive sample prediction while reducing negative sample interference. Furthermore, the majority of polyp regions is rather small. When simple CNN is utilized for feature extraction, these small areas are frequently overlooked. To solve this issue, Wang et al. introduce Pyramid Vision Transformer(PVT)[9] that can yield multi-scale feature maps for dense prediction tasks by combining pyramid structure of the transformer. Dong et al[10] extend PVT with other modules and propose Poly-PVT for polyp segmentation, which effectively suppresses the noise in the features and greatly improves their expressiveness.

Despite the success of Polyp-PVT, it still lacks the ability to address the tricky situation when boundaries are blurry to recognize. To mitigate this problem, we propose a **B**oundary-aware **P**yramid **Trans**former (**BP-Trans**) for the multi-scale feature extraction along with boundary knowledge at multiple levels. BP-Trans extends PVT with a boundary-aware self-attention module, which is supervised by the boundary key-point map and refines features to yield more powerful representations for boundaries. To assess our method, we conduct five-fold cross-validation on the given dataset of Endoscopic computer vision challenges 2.0. The experimental results consistently demonstrate that our proposed framework improves the segmentation performance significantly.

4th International Workshop and Challenge on Computer Vision in Endoscopy (EndoCV2022) in conjunction with the 19th IEEE International Symposium on Biomedical Imaging ISBI2022, March 28th, 2022, IC Royal Bengal, Kolkata, India

jiachengw@stu.xmu.edu.cn (J. Wang);
Correspondingauthor:lswang@xmu.edu.cn (L. Wang)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3148/paper1.pdf

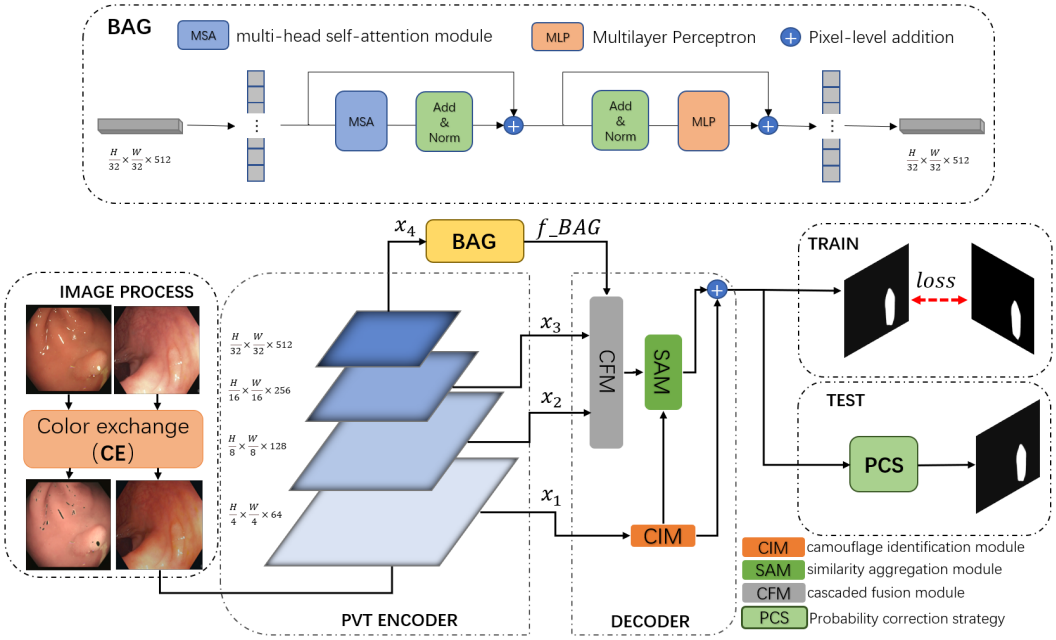

Figure 1: An overview of the boundary-aware pyramid transformer (BP-Trans) for polyp segmentation.

2. Method

2.1. Overall Architecture

As illustrated in 1, to minimize the influence of backdrop colors on model training, we first utilize the CE[8] module to preprocess the input images. Then, the PVT encoder proposed by Wang et al. [9] is used for the coarse feature extraction due to its superior ability in multi-scale representation. After the extraction, features at four different scales are obtained, in which the feature of the highest level, x_4, is sent into a Boundary-aware Attention Gate (BAG) [11] to retrieve boundary information, resulting in the feature f_{BAG}. Finally, we assemble different levels of features, send them into the prediction head, and predict the segmentation map. During inference, we employ the PCS[8] module to correct tiny polyps' excessive pixel imbalance.

2.2. Pyramid Feature Extraction

PVT [9] introduces a pyramid structure of the transformer framework to generate multi-scale feature maps for intensive prediction tasks, which contains four stages to generate feature maps at different scales. All stages share a similar architecture and the details are as follows. Assumed that the feature map with size of $\frac{H}{4} * \frac{W}{4} * C_1$, the input image of size $H * W * C$ is divided into $\frac{H*W}{4*4}$ patches. Each flatten patch is linearly projected to get the embedding and concatenated with the position embedding. After that, the resulted vector is passed through the transformer encoder of this layer and the output is reconstructed. The computation of PVT is greatly reduced by using progressively shrinking pyramids to reduce large feature maps.

We remove the decoder layer, and use the PVT encoder on top of four multi-scale feature maps (i.e., $\{x_i\}_{i=1}^{4}$) generated by different stages. Among these feature maps, x_1 is the lowest level feature, which contains a lot of information, but has a lot of noise. In comparison, x_2, x_3, x_4 provide high-level semantic cues.

2.3. Boundary-aware Knowledge Modeling

The main task of BAG[11] is to extract enough local details to handle blurred boundaries. We argue that the equipment of boundary information can also let the transformer obtain more power in addressing lesions with ambiguous boundaries. At the end of each transformer encoder layer, we add a BAG to enhance the converted features. In addition, the BAG has a key-patch map generator that uses the modified features as the input and outputs a binary patch-wise attention map, where the identity 1 means that the associated patch is at the fuzzy border, similar to a classic spatial attention gate. BAG learns the robust feature representation of fuzzy borders

in a variety of ways thanks to its architecture, which is critical to managing the segmentation of fuzzy boundary lesions.

2.4. Prediction with Multi-level Fusion

To fuse features of different levels, we employ three sub-modules: cascaded fusion module(CFM), camouflage identification module(CIM), and similarity aggregation module(SAM) [10]. CFM is used to extract the semantic and geographical information of polyps in advanced features from x_2, x_3, and x_4, while CIM is utilized to collect the information about polyps camouflaged in x_1. SAM processing extends the pixel features of the polyp region with improved semantic location information on the whole polyp region, successfully integrating the cross-layer features. We utilize the Probability correction technique (PCS)[8] module to cope with little polyps with very uneven foreground and background pixels during the test phase. The primary idea of PCS is to explicitly adjust the forecast probability using logarithmic weighting. This module can significantly increase the accuracy of the final forecast.

3. Experiments

3.1. Datasets and Evaluation Metrics

Table 1
Detailed statistics of each fold in the training data.

fold	Train		Test	
	sequence	sample	sequence	sample
0	36	2571	10	719
1	37	2827	9	463
2	37	2684	9	606
3	37	2465	9	825
4	37	2613	9	677

We employ the Endocv2022 dataset [12][13][14] to conduct the experiments, which includes 46 video sequences with a total of 3390 images. Five-fold cross-validation is adopted here for fair and thorough comparison, statistics of each fold have been shown in Table 1. Dice score is used as the final evaluation metric, which mainly focuses on the internal consistency of segmented objects.

$$\text{Dice score}(y_{true}, y_{pred}) \triangleq \frac{2\mathrm{y_{true}}\mathrm{y_{pred}}.\mathrm{sum}\,() + 1e^{-15}}{\mathrm{y_{true}}.\mathrm{sum}\,() + \mathrm{y_{pred}}.\mathrm{sum}\,() + 1e^{-15}} \quad (1)$$

Here, we report the mean value of Dice score. y_{true} stands for label and y_{pred} stands for prediction. A smoothing factor is used here to avoid the training collapse when meeting empty labels.

3.2. Implementation Details

We utilize the PyTorch framework to create our BP-Trans and a NVIDIA GeForce RTX 3080 Ti to speed up the computations. We use a multi-scale technique in the training phase since each polyp has a unique form and size. The following are the other details. First, because of the variations in picture size from one sequence to the next. we resize the picture altered by CE[8] to a consistent size of 352*352. The picture is then flipped horizontally and vertically with 0.5 probabilities, rotated randomly, then subjected to GaussianBlur with 0.1 probabilities. With an initial learning rate of $1e-4$, we utilize the AdamW optimizer to update network parameters. The batch size was set to 8 for a total of 120 epochs. To oversee model training, we employ a mix of IoU and binary cross entropy with logic as the loss function.

3.3. Compared Models

We tried many models, like PraNet [15], SANet [8], TransFuse [16], HarDNet-MSEG [17].The experimental results are shown in Table 2. Among these, by backward noticing, PraNet [15] first predicts the rough areas and then implicitly models the borders. As a result, when compared to certain traditional models, it delivers a significant improvement in performance. SANet[8] recommends CE to decouple the image's color and content, and shallow attention to decrease data noise for tiny polyps that are difficult to separate, Reduce the interference of irrelevant factors to the model. TransFuse [16] combines Transformers and CNNs in a parallel style, where both global dependency and low-level spatial details can be efficiently captured in a much shallower manner, also achieving good results. In contrast to the above model, HarDNet-MSEG [17] uses a simple encoder-decoder architecture without any attention modules. The backbone of HarDNet-MSEG [17] is a low-storage traffic CNN paired with a decoder that offers outstanding accuracy and fast inference times. Experimental results prove that it is 1.3 times faster than PraNet and more than 2 times faster than other models.

We employ the same encoder-decoder structure as HarDNet-MSEG [17], but instead of using low-storage traffic CNN as the backbone, we use PVT[9], which is a versatile backbone for the dense prediction that we combine with BAG[11] to increase the focus on boundary information. As can be seen in Table 2, the BP-Trans model is superior to the current methods, demonstrating that it has a better learning ability.

3.4. Ablation Study

We explore the effectiveness of each component in detail. **Effectiveness of PVT:** As shown in 2, compared to

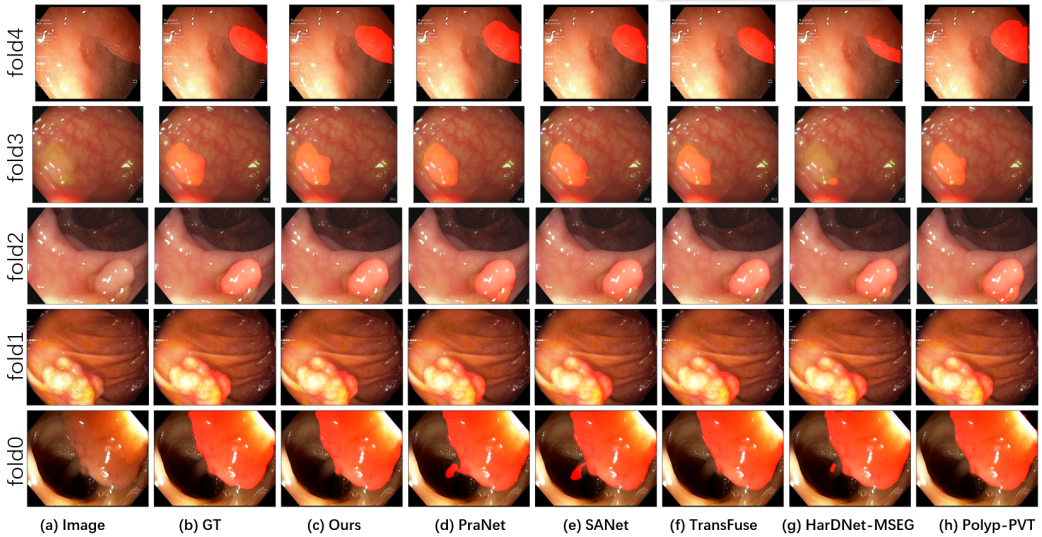

Figure 2: Visualized comparison of (a) Image; (b) Ground-truth (GT); (c) Result of our model; (d-h) Results of compared methods.

Table 2
Quantitative results of compared methods.

model	Dice	Dice_std
Deeplabv3+[18]	0.3019	0.455
Deeplabv3+ + pointrend[19]	0.4884	0.4347
PSPNet[20]	0.2933	0.455
SANet[8]	0.634	0.41
PraNet[15]	0.665	0.39
TransFuse[16]	0.6755	0.3876
HarDNet-MSEG[17]	0.7247	0.3685
Polyp-PVT[10]	0.7321	0.3662
BP-Trans**(ours)**	**0.7429**	**0.352**

other models, Polyp-PVT[10] can achieve the best performance on the training set compared to the other models. HarDNet-MSEG[17] follows closely behind. Their combination with several temporal or aggregation modules are discussed in Table 3. The experimental results prove that PVT[9] outperforms HarDNet-MSEG[17] in all aspects.

Table 3
Ablation study about the key components.

model	Dice	Dice_std
HarDNet-MSEG[17]	0.7247	0.3685
HarDNet-MSEG+CONVLSTM	0.6143	0.382
PVT[9]	0.7321	0.3662
PVT+CONVLSTM	0.7146	0.369
BP-Trans	**0.7429**	**0.352**

Effectiveness of BAG: since the Polyp-PVT[10] model is relatively complex, it is difficult to surpass HarDNet-MSEG[17] in terms of speed, so we try to improve the accuracy of the model. Our tries include the combination of PVT[9] and convlstm[21], in order to make use of the temporal information to further improves the model effect. However, this attempt has not improved the model's performance. Finally, we decided to start from the boundary information and use BAG[11] to obtain more information. It turns out that the model has been improved to a certain extent.

4. Conclusion

This paper introduces a boundary-aware pyramid transformer for polyp segmentation, which leverages the abundant knowledge and multi-scale information of boundaries to boost segmentation performance. We conduct five-fold cross-validation to assess the performance and the results consistently support the wonderful advantage of our method. Furthermore, our method has achieved first place in the first and second rounds during the official evaluation. In the future, temporal performance and generalization ability will be improved and we hope that our findings may inspire new approaches to the problem of polyp segmentation.

References

[1] L. H. Biller, D. Schrag, Diagnosis and treatment of metastatic colorectal cancer: a review, Jama 325

(2021) 669–685.

[2] B. C. MORSON, Genesis of colorectal cancer, Clinics in gastroenterology 5 (1976) 505–525.

[3] M. Akbari, M. Mohrekesh, E. Nasr-Esfahani, S. R. Soroushmehr, N. Karimi, S. Samavi, K. Najarian, Polyp segmentation in colonoscopy images using fully convolutional network, in: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2018, pp. 69–72.

[4] A. Mohammed, S. Yildirim, I. Farup, M. Pedersen, Ø. Hovde, Y-net: A deep convolutional neural network for polyp detection, arXiv preprint arXiv:1806.01907 (2018).

[5] N. B. Le Duy Huynh, A u-net++ with pre-trained efficientnet backbone for segmentation of diseases and artifacts in endoscopy images and videos, in: CEUR Workshop Proceedings, volume 2595, 2020, pp. 13–17.

[6] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, H. D. Johansen, Doubleu-net: A deep convolutional neural network for medical image segmentation, in: 2020 IEEE 33rd International symposium on computer-based medical systems (CBMS), IEEE, 2020, pp. 558–564.

[7] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. De Lange, P. Halvorsen, H. D. Johansen, Resunet++: An advanced architecture for medical image segmentation, in: 2019 IEEE International Symposium on Multimedia (ISM), IEEE, 2019, pp. 225–2255.

[8] J. Wei, Y. Hu, R. Zhang, Z. Li, S. K. Zhou, S. Cui, Shallow attention network for polyp segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 699–708.

[9] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, L. Shao, Pyramid vision transformer: A versatile backbone for dense prediction without convolutions, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 568–578.

[10] B. Dong, W. Wang, D.-P. Fan, J. Li, H. Fu, L. Shao, Polyp-pvt: Polyp segmentation with pyramid vision transformers, arXiv preprint arXiv:2108.06932 (2021).

[11] J. Wang, L. Wei, L. Wang, Q. Zhou, L. Zhu, J. Qin, Boundary-aware transformers for skin lesion segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 206–216.

[12] S. Ali, M. Dmitrieva, N. Ghatwary, S. Bano, G. Polat, A. Temizel, A. Krenzer, A. Hekalo, Y. B. Guo, B. Matuszewski, et al., Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy, Medical image analysis 70 (2021) 102002. doi:10.1016/j.media.2021.102002.

[13] S. Ali, D. Jha, N. Ghatwary, S. Realdon, R. Cannizzaro, O. E. Salem, D. Lamarque, C. Daul, K. V. Anonsen, M. A. Riegler, et al., Polypgen: A multi-center polyp detection and segmentation dataset for generalisability assessment, arXiv preprint arXiv:2106.04463 (2021). doi:10.48550/arXiv.2106.04463.

[14] S. Ali, N. Ghatwary, D. Jha, E. Isik-Polat, G. Polat, C. Yang, W. Li, A. Galdran, M.-Á. G. Ballester, V. Thambawita, et al., Assessing generalisability of deep learning-based polyp detection and segmentation methods through a computer vision challenge, arXiv preprint arXiv:2202.12031 (2022). doi:10.48550/arXiv.2202.12031.

[15] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, L. Shao, Pranet: Parallel reverse attention network for polyp segmentation, in: International conference on medical image computing and computer-assisted intervention, Springer, 2020, pp. 263–273.

[16] Y. Zhang, H. Liu, Q. Hu, Transfuse: Fusing transformers and cnns for medical image segmentation, in: MICCAI, 2021.

[17] C.-H. Huang, H.-Y. Wu, Y.-L. S. Lin, Hardnet-mseg: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps, ArXiv abs/2101.07172 (2021).

[18] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, arXiv:1802.02611 [cs] (2018). arXiv:1802.02611.

[19] A. Kirillov, Y. Wu, K. He, R. Girshick, Pointrend: Image segmentation as rendering, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9796–9805. doi:10.1109/CVPR42600.2020.00982.

[20] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid Scene Parsing Network, arXiv:1612.01105 [cs] (2017). arXiv:1612.01105.

[21] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, W.-c. Woo, Convolutional lstm network: A machine learning approach for precipitation nowcasting, Advances in neural information processing systems 28 (2015).