Cascade Transformer Encoded Boundary-Aware Multibranch Fusion Networks for Real-Time and Accurate Colonoscopic Lesion Segmentation

Ao Wang[1(✉)], Ming Wu[1(✉)], Hao Qi[1], Wenkang Fan[1], Hong Shi[3(✉)], Jianhua Chen[3], Sunkui Ke[4], Yinran Chen[1], and Xiongbiao Luo[1,2(✉)]

[1] Department of Computer Science and Technology, Xiamen University, Xiamen, China
awang.xmu@gmail.com, xiongbiao.luo@gmail.com, wuming@stu.xmu.edu.cn
[2] National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, China
[3] Fujian Cancer Hospital, Fujian Medical University Cancer Hospital, Fuzhou, China
endoshihong@hotmail.com
[4] Zhongshan Hospital, Xiamen University, Xiamen 361004, China

Abstract. Automatic segmentation of colonoscopic intestinal lesions is essential for early diagnosis and treatment of colorectal cancers. Current deep learning-driven methods still get trapped in inaccurate colonoscopic lesion segmentation due to diverse sizes and irregular shapes of different types of polyps and adenomas, noise and artifacts, and illumination variations in colonoscopic video images. This work proposes a new deep learning model called cascade transformer encoded boundary-aware multibranch fusion networks for white-light and narrow-band colorectal lesion segmentation. Specifically, this architecture employs cascade transformers as its encoder to retain both global and local feature representation. It further introduces a boundary-aware multibranch fusion mechanism as a decoder that can enhance blurred lesion edges and extract salient features, and simultaneously suppress image noise and artifacts and illumination changes. Such a newly designed encoder-decoder architecture can preserve lesion appearance feature details while aggregating the semantic global cues at several different feature levels. Additionally, a hybrid spatial-frequency loss function is explored to adaptively concentrate on the loss of important frequency components due to the inherent bias of neural networks. We evaluated our method not only on an in-house database with four types of colorectal lesions with different pathological features, but also on four public databases, with the experimental results showing that our method outperforms state-of-the-art network models. In particular, it can improve the average dice similarity coefficient and intersection over union from (84.3%, 78.4%) to (87.0%, 80.5%).

A. Wang and M. Wu—Shows the equally contributed authors.
X. Luo and H. Shi are the corresponding authors.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14228, pp. 718–727, 2023.
https://doi.org/10.1007/978-3-031-43996-4_69

1 Introduction

Colorectal cancer (CRC) is the third most commonly diagnosed cancer but ranks second in terms of mortality worldwide [11]. Intestinal lesions, particularly polyps and adenomas, are usually developed to CRC in many years. Therefore, diagnosis and treatment of colorectal polyps and adenomas at their early stages are essential to reduce morbidity and mortality of CRC. Interventional colonoscopy is routinely performed by surgeons to visually examine colorectal lesions. However these lesions in colonoscopic images are easily omitted and wrongly classified due to limited knowledge and experiences of surgeons. Automatic and accurate segmentation is a promising way to improve colorectal examination.

Many researchers employ U-shaped network [7,13,18] for colonoscopic polyp segmentation. ResUNet++ [7] combines residual blocks and atrous spatial pyramid pooling and Zhao et al. [18] designed a subtraction unit to generate the difference features at multiple levels and constructed a training-free network to supervise polyp-aware features. Unlike a family of U-Net driven segmentation methods, numerous papers have been worked on boundary constraints to segment colorectal polyps. Fan et al. [2] introduced PraNet with reverse attention to establish the relationship between boundary cues from global feature maps generated by a parallel partial decoder. Both polyp boundary-aware segmentation methods work well but still introduce much false positive. Based on PraNet [2] and HardNet [1], Huang et al. [6] removed the attention mechanism and replaced Res2Net50 by HardNet to build HardNet-MSEG that can achieve faster segmentation. In addition, Kim et al. [9] modified PraNet to construct UACANet with parallel axial attention and uncertainty augmented context attention to compute uncertain boundary regions. Although PraNet and UACANet aim to extract ambiguous boundary regions from both saliency and reverse saliency features, they simply set the saliency score to 0.5 that cannot sufficiently detect complete boundaries to separate foreground and background regions. More recently, Shen et al. [10] introduced task-relevant feature replenishment networks for cross-center polyp segmentation, while Tian et al. [12] combined transformers and multiple instance learning to detect polyps in a weakly supervised way.

Unfortunately, limited field of view and illumination variations usually result in insufficient boundary contrast between intestinal lesions and their surrounding tissues. On the other hand, various polyps and adenomas with different pathological features have similar visual characteristics to intestinal folds. To address these issues mentioned above, we explore a new deep learning architecture called cascade transformer encoded boundary-aware multibranch fusion (CTBMF) networks with cascade transformers and multibranch fusion for polyp and adenoma segmentation in colonoscopic white-light and narrow-band video images. Several technical highlights of this work are summarized as follows. First, we construct cascade transformers that can extract global semantic and subtle boundary features at different resolutions and establish weighted links between global semantic cues and local spatial ones for intermediate reasoning, providing long-range dependencies and a global receptive field for pixel-level segmentation. Next, a hybrid spatial-frequency loss function is defined to compensate for loss

features in the spatial domain but available in the frequency domain. Additionally, we built a new colonoscopic lesion image database and will make it publicly available, while this work also conducts a thorough evaluation and comparison on our new database and four publicly available ones (Fig. 2).

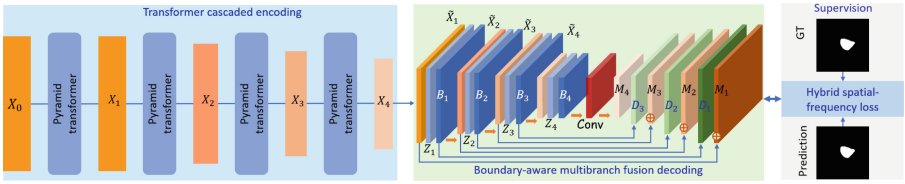


Fig. 1. CTBMF consists of cascade transformers, boundary-aware multibranch fusion, and hybrid spatial-frequency loss.

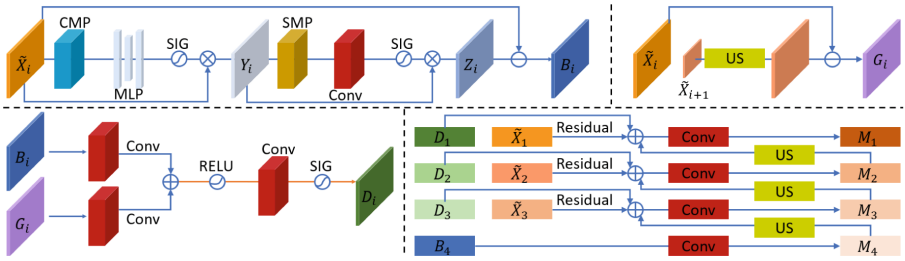


Fig. 2. The boundary-aware multibranch fusion decoder employs the boundary-aware attention module to compute B_i, G_i, and D_i and introduces residual multibranch fusion to calculate M_i.

2 Approaches

This section details our CTBMF networks that can refine inaccurate lesion location, rough or blurred boundaries, and unclear textures. Figure 1 illustrates the encoder-decoder architecture of CTMBF with three main modules.

2.1 Transformer Cascade Encoding

This work employs a pyramid transformer [15] to build a transformer cascaded encoder. Let X_0 and X_i be input patches and the feature map at stage i, respectively. Overlapping patch embedding (OPE) separates an image into fixed-size patches and linearly embeds them into tokenized images while making adjacent windows overlap by half of a patch. Either key K_i or value V_i is the input

sequence of linear spatial reduction (LSR) that implements layer normalization (LN) and average pooling (AP) to reduce the input dimension:

$$\mathrm{LSR}(K_i) = \mathrm{AP}(\mathrm{Reshape}(\mathrm{LN}(K_i \oplus \Omega(K_i), R_i)W_{K_i}) \tag{1}$$

where $\Omega(\cdot)$ denotes the output parameters of position embedding, $\oplus$ is the element-wise addition, W_{K_i} indicates the parameters that reduces the dimension of K_i or V_i, and R_i is the reduction ratio of the attention layers at stage i. As the output of LSR is fed into multihead attention, we can obtain attention feature map A_i^j from head j ($j = 1, 2, \cdots, N$, N is the head number of the attention layer) at stage i:

$$A_i^j = \mathrm{Attention}(QW_{Q_i}^j, \mathrm{LSR}(K_i)W_{K_i}^j, \mathrm{LSR}(V_i)W_{V_i}^j) \tag{2}$$

where Attention$(\cdot)$ is calculated as the original transformer [14]. Subsequently, the output $\mathrm{LSRA}(Q_i, K_i, V_i)$ of LSRA is

$$\mathrm{LSRA}(Q_i, K_i, V_i) = (A_i^1 \odot \cdots A_i^j \odot \cdots A_i^N)W_{A_i} \tag{3}$$

where $\odot$ is the concatenation and W_{A_i} is the linear projection parameters. Then, $\mathrm{LSRA}(Q_i, K_i, V_i)$ is fed into convolutional feed-forward (CFF):

$$\mathrm{CFF}(Q_i, K_i, V_i) = \mathrm{FC}(\mathrm{GELU}(\mathrm{DC}(\mathrm{FC}(\mathrm{LN}(\Psi))))) \tag{4}$$

$$\Psi = ((Q_i, K_i, V_i) \oplus \Omega(Q_i, K_i, V_i)) \oplus \mathrm{LSRA}(Q_i, K_i, V_i) \tag{5}$$

where DC is a 3×3 depth-wise convolution [5] with padding size of between the fully-connected (FC) layer and the Gaussian error linear unit (GELU) [4] in the feed-forward networks. Eventually, the output feature map X_i of the pyramid transformer at stage i can be represented by

$$X_i = \mathrm{Reshape}(\Psi \oplus \mathrm{CFF}(Q_i, K_i, V_i)) \tag{6}$$

2.2 Boundary-Aware Multibranch Fusion Decoding

Boundary-Aware Attention Module. Current methods [2,9] detect ambiguous boundaries from both saliency and reverse-saliency maps by predefining a saliency score of 0.5. Unfortunately, a predefined score cannot distinguish foreground and background of different colonoscopic lesions [3]. Based on [17], this work explores an effective boundary-aware attention mechanism to adaptively extract boundary regions.

Given the feature map X_i with semantic cues and rough appearance details, we perform convolution (Conv) on it and obtain $\tilde{X}_i = \mathrm{Conv}(X_i)$, which is further augmented by channel and spatial attentions. The channel attention performs channel maxpooling (CMP), multilayer perceptron (MLP), and sigmoid (SIG) to obtain the intermediate feature map Y_i:

$$Y_i = \tilde{X}_i \otimes \mathrm{SIG}(\mathrm{MLP}(\mathrm{CMP}(\tilde{X}_i))) \tag{7}$$

where $\otimes$ indicates the elementwise product. Subsequently, the detail enhanced feature map Z_i of the channel-spatial attention is

$$Z_i = Y_i \otimes \mathrm{SIG}(\mathrm{Conv}(\mathrm{SMP}(Y_i))) \tag{8}$$

where SMP indicates spatial maxpooling. We subtract the feature map $\tilde{X}_i$ from the enhanced map Z_i to obtain the augmented boundary attention map B_i, and also establish the correlation between the neighbor layers X_{i+1} and X_i to generate multilevel boundary map G_i:

$$B_i = Z_i \ominus \tilde{X}_i, i = 1, 2, 3, 4 \qquad G_i = \tilde{X}_i \ominus \mathrm{US}(\tilde{X}_{i+1}), i = 1, 2, 3 \tag{9}$$

where $\ominus$ and US indicate subtraction and upsampling.

Residual Multibranch Fusion Module. To highlight salient regions and suppress task-independent feature responses (e.g., blurring), we linearly aggregate B_i and G_i to generate discriminative boundary attention map D_i:

$$D_i = \mathrm{SIG}(\mathrm{Conv}(\mathrm{RELU}(\mathrm{Conv}(B_i) \oplus \mathrm{Conv}(G_i)))), \tag{10}$$

where $i = 1, 2, 3$ and RELU is the rectified linear unit function.

We obtain the fused feature representation map M_i $(i = 1, 2, 3, 4)$ from the elementwise addition or summation of M_{i+1}, D_i, and the residual feature $\tilde{X}_i$ by

$$M_i = \mathrm{Conv}(\tilde{X}_i \oplus D_i \oplus \mathrm{US}(M_{i+1})), i = 1, 2, 3 \qquad M_4 = \mathrm{Conv}(B_4) \tag{11}$$

Eventually, the output M_1 of the boundary-aware multibranch fusion decoder is represented by the following equation:

$$M_1 = \mathrm{Conv}(\tilde{X}_1 \oplus D_1 \oplus \mathrm{US}(M_2)) \tag{12}$$

which precisely combines global semantic features with boundary or appearance details of colorectal lesions.

2.3 Hybrid Spatial-Frequency Loss

This work proposes a hybrid spatial-frequency loss function $\mathcal{H}_L$ to train our network architecture for colorectal polyp and adenoma segmentation:

$$\mathcal{H}_L = \mathcal{S}_L + \mathcal{F}_L \tag{13}$$

where $\mathcal{S}_L$ and $\mathcal{F}_L$ are a spatial-domain loss and a frequency-domain loss to calculate the total difference between prediction P and ground truth G, respectively. The spatial-domain loss $\mathcal{S}_L$ consists of a weighted intersection over union loss and a weighted binary cross entropy loss [16].

The frequency-domain loss $\mathcal{F}_L$ can be computed by [8]

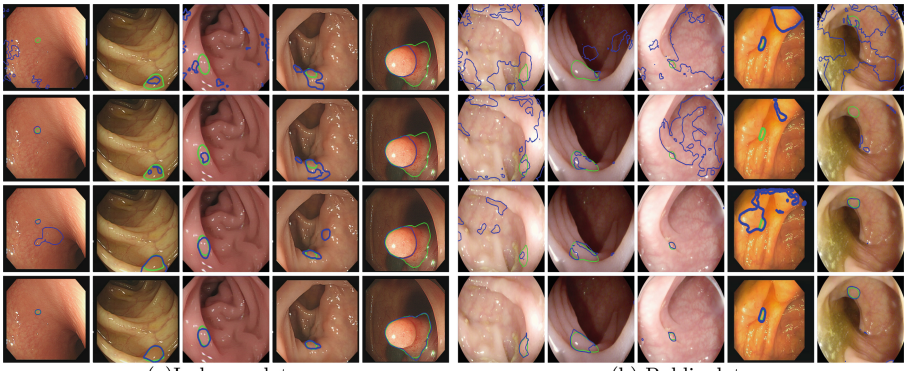
$$\mathcal{F}_L = \lambda \frac{1}{WH} \sum_{u=0}^{W-1} \sum_{v=0}^{H-1} \gamma(u, v) |\mathcal{G}(u, v) - \mathcal{P}(u, v)|^2 \tag{14}$$

where $W \times H$ is the image size, λ is the coefficient of $\mathcal{F}_L$, $\mathcal{G}(u, v)$ and $\mathcal{P}(u, v)$ are a frequency representation of ground truth G and prediction P using 2-D discrete Fourier transform. $\gamma(u, v)$ is a spectrum weight matrix that is dynamically determined by a non-uniform distribution on the current loss of each frequency.

3 Experiments

Our clinical in-house colonoscopic videos were acquired from various colonoscopic procedures under a protocol approved by the research ethics committee of the university. These white-light and narrow-band colonoscopic images contain four types of colorectal lesions with different pathological features classified by surgeons: (1) 268 cases of hyperplastic polyp, (2) 815 cases of inflammatory polyp, (3) 1363 cases of tubular adenoma, and (4) 143 cases of tubulovillous adenoma. Additionally, four public datasets including Kvasir, ETIS-LaribPolypDB, CVC-ColonDB, and CVC-ClinicDB were also used to evaluate our network model.

We implemented CTBMF on PyTorch and trained it with a single NVIDIA RTX3090 to accelerate the calculations for 100 epochs at mini-batch size 16. Factors λ (Eq. (14)) were set to 0.1. We employ the stochastic gradient descent



(a)In-house data (b) Public data

Fig. 3. Visual comparison of the segmentation results of using the four different methods tested on those in-house and public datasets. *Green* and *blue* show ground truth and prediction. (Color figure online)

Table 1. Results and computational time of using five databases(our in-house and four public databases)

Average	DSC	IoU	F_β	In-house	Public	Average
PraNet [2]	0.813	0.751	0.795	28.7 FPS	30.5 FPS	29.6 FPS
HardNet [6]	0.830	0.764	0.812	39.2 FPS	39.9 FPS	39.5 FPS
UACANet [9]	0.843	0.784	0.824	16.5 FPS	16.7 FPS	16.6 FPS
CTBMF (Ours)	**0.870**	**0.805**	**0.846**	**33.1 FPS**	**33.6 FPS**	**33.4 FPS**

algorithm to optimize the overall parameters with an original learning rate of 0.0001 for cascade transformer encoding and 0.05 for other parts and use warm-up and linear decay strategies to adjust it. The momentum and weight decay were set as 0.9 and 0.0005. Further, we resized input images to 352×352 for training and testing and the training time was nearly 1.5 h to achieve the convergence. We employ three metrics to evaluate the segmentation: Dice similarity coefficient (DSC), intersection over union (IoU), and weighted F-measure (F_β).

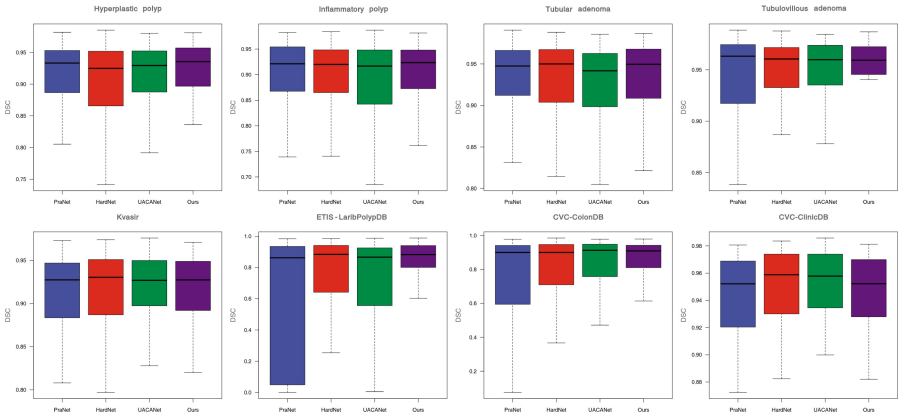


Fig. 4. DSC boxplots of using the four methods evaluated on our in-house and publicly available databases

Table 2. Public data segmented results of our ablation study

Modules	DSC	IoU	F_β
D_1	0.681	0.593	0.634
D_2	0.822	0.753	0.798
D_3	0.820	0.748	0.793
Residual	0.828	0.757	0.802
$\mathcal{F}_L$	0.834	0.764	0.804

4 Results and Discussion

Figure 3 visually compares the segmentation results of the four methods tested on our in-house and public databases. Our method can accurately segment polyps in white-light and narrow-band colonoscopic images under various scenarios, and CTBMF can successfully extract small, textureless and weak boundary and colorectal lesions. The segmented boundaries of our method are sharper and clear than others especially in textureless lesions that resemble intestinal lining.

Figure 4 shows the DSC-boxplots to evaluate the quality of segmented polyps and adenomas, which still demonstrate that our method works much better than the others. Figure 5 displays the enhanced feature maps using the boundary-aware attention module. Evidently, small and weak-boundary or textureless lesions can be enhanced with good boundary feature representation.

Table 1 summarizes the quantitative results in accordance with the three metrics and computational time of four methods. Evidently, CTBMF generally works better than the compared methods on the in-house database with four types of colorectal lesions. Furthermore, we also summarizes the average three metrics computed from all the five databases (the in-house dataset and four public datasets). Our method attains much higher average DSC and IoU of (0.870, 0.805) than the others on the five databases.

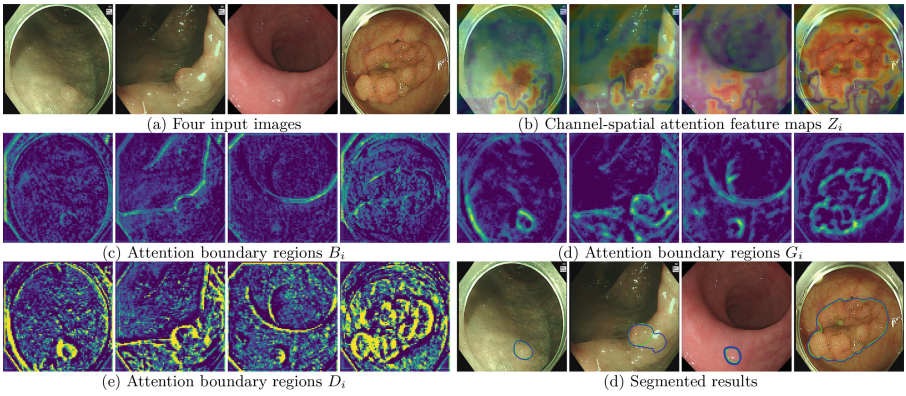


Fig. 5. Effectiveness of the boundary-aware multibranch fusion decoder generated various boundary-aware feature maps

We performed an ablation study to evaluate the effectiveness of each module used in CTBMF. The baseline is the standard version of cascade pyramid transformers. Modules D_1, D_2, D_3, residual connections, and frequency loss $\mathcal{F}_L$ are gradually added into the baseline, evaluating the effectiveness of each module and comparing the variants with each other. We tested these modules on the four public databases. Table 2 shows all the ablation study results. Each module can improve the segmentation performance. Particularly, the boundary-aware attention module critically improves the average DSC, IoU, and F_β.

Our method generally works better than the other three methods. Several reasons are behind this. First, the cascade-transformer encoder can extract local and global semantic features of colorectal lesions with different pathological characteristics due to its pyramid representation and linear spatial reduction attention. While the pyramid operation extracts multiscale local features, the attention mechanism builds global semantic cues. Both pyramid and attention strategies facilitate the representation of small and textureless intestinal lesions

in encoding, enabling to characterize the difference between intestinal folds (linings) and subtle-texture polyps or small adenomas. Next, the boundary-aware attention mechanism drives the multibranch fusion, enhancing the representation of intestinal lesions in weak boundary and nonuniform lighting. Such a mechanism first extracts the channel-spatial attention feature map, from which subtracts the current pyramid transformer's feature map to enhance the boundary information. Also, the multibranch fusion generates multilevel boundary maps by subtracting the next pyramid transformer's upsampling output from the current pyramid transformer's output, further improving the boundary contrast. Additionally, the hybrid spatial-frequency loss was also contributed to the improvement of colorectal lesion segmentation. The frequency-domain information can compensate loss feature information in the spatial domain, leading to a better supervision in training.

5 Conclusion

This work proposes a new deep learning model of cascade pyramid transformer encoded boundary-aware multibranch fusion networks to automatically segment different colorectal lesions of polyps and adenomas in colonoscopic imaging. While such an architecture employs simple and convolution-free cascade transformers as an encoder to effectively and accurately extract global semantic features, it introduces a boundary-aware attention multibranch fusion module as a decoder to preserve local and global features and enhance structural and boundary information of polyps and adenomas, as well as it uses a hybrid spatial-frequency loss function for training. The thorough experimental results show that our method outperforms the current segmentation models without any preprocessing. In particular, our method attains much higher accuracy on colonoscopic images with small, illumination changes, weak-boundary, textureless, and motion blurring lesions, improving the average dice similarity coefficient and intersection over union from (89.5%, 84.1%) to (90.3%, 84.4%) on our in-house database, from (78.9%, 72.6%) to (83.4%, 76.5%) on the four public databases, and from (84.3%, 78.4%) to (87.0%, 80.5%) on the five databases.

Acknowledgements. This work was supported partly by the National Natural Science Foundation of China under Grants 61971367 and 82272133, the Natural Science Foundation of Fujian Province of China under Grant 2020J01004, and the Fujian Provincial Technology Innovation Joint Funds under Grant 2019Y9091.

References

1. Chao, P., Kao, C.Y., Ruan, Y.S., Huang, C.H., Lin, Y.L.: Hardnet: a low memory traffic network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3552–3561 (2019)
2. Fan, D.-F., et al.: PraNet: parallel reverse attention network for polyp segmentation. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12266, pp. 263–273. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59725-2_26

3. Guo, X., Yang, C., Liu, Y., Yuan, Y.: Learn to threshold: thresholdnet with confidence-guided manifold mixup for polyp segmentation. IEEE Trans. Med. Imaging **40**(4), 1134–1146 (2020)
4. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
5. Howard, A., et al.: Efficient convolutional neural networks for mobile vision. arXiv preprint arXiv:1704.04861 (2017)
6. Huang, C.H., Wu, H.Y., Lin, Y.L.: HarDNet-MSEG: a simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps. arXiv preprint arXiv:2101.07172 (2021)
7. Jha, D., et al.: ResUNet++: an advanced architecture for medical image segmentation. In: 2019 IEEE International Symposium on Multimedia (ISM), pp. 225–2255. IEEE (2019)
8. Jiang, L., Dai, B., Wu, W., Loy, C.C.: Focal frequency loss for image reconstruction and synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 13919–13929 (2021)
9. Kim, T., Lee, H., Kim, D.: UACANet: uncertainty augmented context attention for polyp segmentation. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 2167–2175 (2021)
10. Shen, Y., Lu, Y., Jia, X., et al.: UACANet: uncertainty augmented context attention for polyp segmentation. In: Medical Image Computing and Computer Assisted Intervention (MICCAI), pp. 599–608 (2022)
11. Sung, H., et al.: Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. **71**(3), 209–249 (2021)
12. Tian, Y., Pang, G., Liu, F., et al.: Contrastive transformer-based multiple instance learning for weakly supervised polyp frame detection. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. MICCAI 2022. LNCS, vol. 13433, pp. 88–98. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_9
13. Tomar, N.K., et al.: DDANet: dual decoder attention network for automatic polyp segmentation. In: Del Bimbo, A., et al. (eds.) ICPR 2021. LNCS, vol. 12668, pp. 307–314. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-68793-9_23
14. Vaswani, A., et al.: Attention is all you need. Adv. Neural Inf. Process. Syst. **30** (2017)
15. Wang, W., Xie, E., Li, X., et al.: PVT v2: improved baselines with pyramid vision transformer. Comput. Vis. Media **8**, 415–424 (2022)
16. Wei, J., Wang, S., Huang, Q.: F^3net: fusion, feedback and focus for salient object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 12321–12328 (2020)
17. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: convolutional block attention module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 3–19. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_1
18. Zhao, X., Zhang, L., Lu, H.: Automatic polyp segmentation via multi-scale subtraction network. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 120–130. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_12