FuzzyNet: A Fuzzy Attention Module for Polyp Segmentation

Krushi Patel[1], **Fenjun Li**[1], **Guanghui Wang**[2]*
[1]Department of EECS, University of Kansas, KS, USA
[2]Department of CS, Toronto Metropolitan University, Toronto, ON, Canada
krushi92@ku.edu, wangcs@ryerson.ca (* corresponding author)

Abstract

Polyp segmentation is essential for accelerating the diagnosis of colon cancer. However, it is challenging because of the diverse color, texture, and varying lighting effects of the polyps as well as the subtle difference between the polyp and its surrounding area. To further increase the performance of polyp segmentation, we propose to focus more on the problematic pixels that are harder to predict. To this end, we propose a novel attention module named Fuzzy Attention to focus more on the difficult pixels. Our attention module generates a high attention score for fuzzy pixels usually located near the boundary region. This module can be embedded in any convolution neural network-based backbone network. We embed our module with various backbone networks: Res2Net, ConvNext and Pyramid Vision Transformer and evaluate the models on five polyp segmentation datasets: Kvasir [11], CVC-300 [30], CVC-ColonDB [29], CVC-ClinicDB [2], and ETIS [28]. Our attention module with Res2Net as the backbone network outperforms the reverse attention-based PraNet by a significant amount on all datasets. In addition, our module with PVT as the backbone network achieves state-of-the-art accuracy of 0.937, 0.811, and 0.791 on the CVC-ClinicDB, CVC-ColonDB, and ETIS, respectively, outperforming the latest SA-Net, TransFuse and Polyp-PVT. The source code is available at: https://github.com/krushi1992/FuzzyNet.

1 Introduction

Polyp segmentation is an essential task to accelerate the diagnosis of colorectal cancer[19][22][28], which is considered the most prevalent cancer worldwide. If the polyp is detected earlier, the mortality rate can further be reduced. Colonoscopy is considered the effective technique for CRC screening, which detects the polyps that may cause colon cancer. Detecting polyps is a complicated process because of their similar appearance to background pixels. Sometimes, even an experienced clinician finds it very difficult to recognize, and thus leads to missing detection of polyps because of their subtle difference [14] [23]. In addition, polyps are widely varied in size, texture, and color. Therefore an accurate and automatic polyp segmentation method is required to detect the cancerous polyp in the early-stage to reduce the mortality rate [13].

Convolution neural networks have achieved tremendous performance gain on various medical image segmentation tasks, including the polyp segmentation [1][3][9][12] [25][37]. Various methods have been proposed to tackle the issue of detecting difficult boundary pixels accurately, either by using separate edge supervision [6][20] or attention modules. However, the use of edge supervision reduces the generalization capability of the model and requires extra boundary annotations, which are expensive. The attention-based methods used reverse attention [5], focusing on the background region to mine the boundary clues. However, we believe the performance can be further improved if we focus more on the difficult pixels instead of the background pixel. Therefore, in this work, we

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

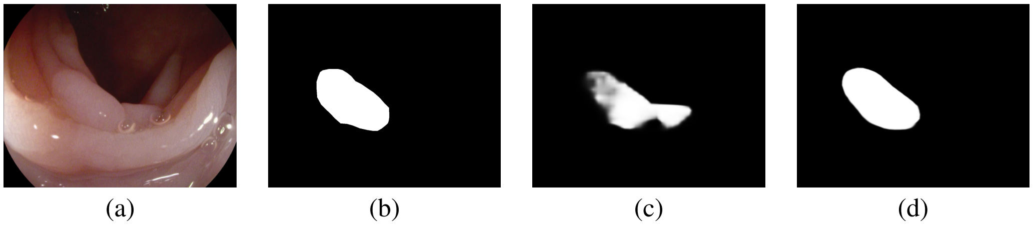

Figure 1: (a) An original image of the polyp; (b) the ground-truth mask; (c) the prediction mask generated by PraNet; and (d) the prediction mask generated by our Fuzzy-Net.

propose a novel attention module named Fuzzy attention to encourage the model to focus on the hard boundary pixels.

The pixels that are not categorized straightforwardly as foreground or background pixels are considered hard pixels. The smaller the difference between the foreground and background attention score, the higher the complexity. Our attention module uses the above observation to calculate the final attention score, which results in a high score for difficult pixels, usually lying around the boundary region, and lower weights for the easy pixels. Figure 1 shows an image of the polyp, its ground truth, and the prediction masks generated by PraNet and our FuzzyNet, respectively. It can be seen from the original image that the region around the boundary is hard to predict, resulting in an uneven edge, as shown in Figure 1-(c). However, our model predicts the mask closer to the ground truth with a smooth boundary as shown in Figure 1-(d). Like reverse attention in PraNet [5], we apply this module in parallel on the top of the last three levels of the feature map along with deep supervision.

The encoder is considered the backbone network in the segmentation task, which extracts the row fine level to coarse level features and is further processed by various small architectures modules to enhance the feature representation. Therefore, to observe the impact of various backbone architecture types, we embed our module in three different networks: Res2Net [8], ConvNext [15], and PVT [31], and compare the performance. Our result shows that the proposed module with Res2Net as a backbone network significantly outperforms PraNet with the same backbone [5] on various polyp segmentation datasets, including Kvasir[11], CVC-ClinicDB[2], CVC-ColonDB [29], CVC-300, and ETIS[28]. In addition, our attention module with the PVT as a backbone network achieves state-of-the-art accuracy on CVC-ClinicDB, CVC-ColonDB and ETIS by exceeding the performance of recently proposed SA-Net, TransFuse, and Polyp-PVT.

The main contributions of this work are summarized below:

1. We propose a novel attention module, named Fuzzy attention, to focus more on the difficult pixels which usually lie near the boundary region. It can be embedded in any backbone network in parallel after the last three feature maps.
2. We investigate the impact of various types of backbone networks: Res2net [8], ConvNext[15], and PVT[31], along with our attention module through extensive experiments.
3. Extensive experimental results show that our Fuzzy attention module outperforms the reverse attention-based model, PraNet [5], by a significant margin with the same Res2Net backbone on the polyp segmentation datasets: Kvasir, CVC-ColonDB [29], CVC-ClinicDB[2], CVC-300[30]. With PVT [31] as a backbone network, we achieve state-of-the-art accuracy on the CVC-ClinicDB, CVC-ColonDB, and ETIS datasets.

2 Related Work

Various approaches have been proposed to segment the polyp in colonoscopic images using either handcrafted features or deep features extracted by deep learning networks. All approaches can be broadly divided into two categories:

Classical computer vision approaches: Early polyp segmentation approaches use low-level handcrafted features, including texture[18] and geometric features. [17] used a simple linear iterative clustering superpixel to segment the polyp. As mentioned above, all the method has a high false detection rate because of the high similarity between polyps and the surrounding area.

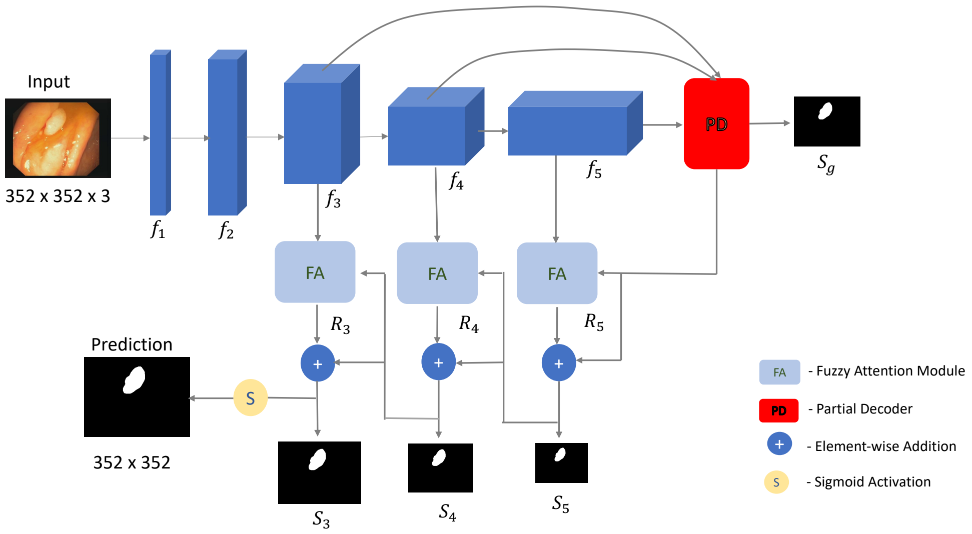

Figure 2: Overall architecture of FuzzyNet. It includes a partial decoder denoted as "PD" in red block and a series of fuzzy attention modules denoted as "FA" blocks. It generates the global map from the partial decoder and passes through the series of fuzzy attention modules which focuses on the difficult and fuzzy boundary pixels. The deep supervision is applied at the end of the output of each fuzzy attention module and partial decoder.

Deep learning based approaches: There have been various deep learning-based approaches proposed for the polyp segmentation task. It started with the study [1], which first employs a convolutional network for the polyp segmentation and outperforms the traditional methods. The U-shaped encoder and decoder architectures: U-Net [25], U-Net++ [37], ResUNet++ [12], ACS-Net [35] and Enhanced U-Net [22] started dominating the segmentation field because of their tremendous performance gain.

To alleviate the issues of complex boundary regions, SFA [6] and PSI [20] include an extra edge supervision branch. However, it requires extra boundary annotation and has an overfitting problem. PraNet [5] introduces the reverse attention mechanism to mine the boundary region gradually by focusing more on background pixels. In contrast, to reverse attention, our module focuses more on the complex pixels usually lying around the boundary region. ACS-Net [35] also introduces the attention mechanism to focus more on the hard pixels, however, it employs a predefined pixel score to classify the pixel as hard or easy.

Other attention-based models, SA-Net [32] and Enhanced U-Net [22] use different attention mechanisms to give more attention to the foreground region and adaptively select the features. TranFuse [36] and Polyp-PVT [4] use the latest vision transformer for the polyp segmentation task and achieve an excellent result. The attention mechanism has also been successfully applied in many other applications [7][16][26][27]. In one of our experiments, we also embedded our attention module in Pyramid Vision Transformer and established state-of-the-art accuracy on various datasets [21].

3 Method

The overview of our proposed network is shown in Figure 2. We follow the architecture used in the PraNet [5] and replace the reverse attention module in PraNet with our proposed Fuzzy attention module. Specifically, our model takes the RGB image as an input and passes it through the backbone network, followed by the partial decoder, which employs multi-resolution feature maps to generate the initial global semantic map. This global map is then passed through a series of fuzzy attention modules, which gradually mine the boundary cues. We apply deep supervision after each attention module and the initial global map. The map generated by the last attention layer is considered the final prediction map. A detailed explanation of each element of the architecture is elaborated below.

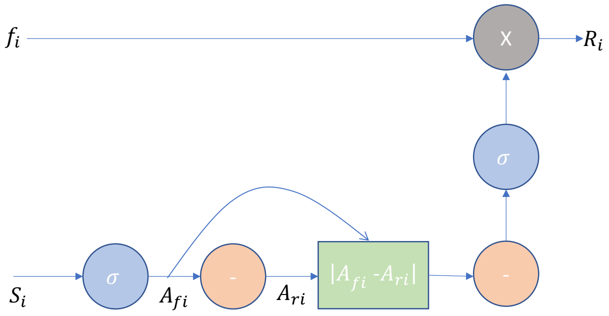

Figure 3: The block diagram of the Fuzzy attention module. It takes the S_i as input and passes it through the Sigmoid and the reverse block to calculate the forward and reverse attention map. The absolute difference between these attention maps is calculated followed by the reverse attention and sigmoid activation. $'-'$ represents the $1 - X$ operation, where X is the input.

3.1 Backbone Network

In the segmentation task, the encoder is considered as the backbone network, which generates the essential row multi-resolution features from fine level to abstract $\{f_i, i = 1, ...5\}$. It is considered the heart of the segmentation model because the model performance heavily relies on the features generated by it. Therefore, to observe the impact of various types of backbone networks (either convolution-based or transformer-based), we use three different networks in our experiments: Res2Net [8], ConvNext [15], and PVT [31].

3.2 Partial Decoder

As mentioned in the previous section, the encoder generates five levels of multi-resolution feature maps $f_i, i = 1, ...5$. These feature maps are further divided into two types: low level $\{f_i, i = 1, 2\}$ and high level $\{f_i, i = 3, 4, 5\}$. As observed in [34], low-level features increase the computation cost by a large amount and have less contribution towards increasing the performance. Therefore, we employ the parallel partial decoder proposed in [34], which aggregates only the high-level features to generate the global initial semantic map, and it is further refined in attention modules.

3.3 Fuzzy Attention Module

The human's natural tendency is to roughly locate the object and then gradually mine the complex region by concentrating more on that area. We apply a similar approach for medical image segmentation to clearly distinguish the background area and foreground objects (polyp or skin lesion) by focusing more on the hard pixels using the Fuzzy attention module. We apply this module parallelly on the high-level feature maps $\{f_i, i = 3, 4, 5\}$ which produce the resultant feature map $\{R_i, i = 3, 4, 5\}$, where R_3 is used to generate the final prediction map. The block diagram of the fuzzy attention module is shown in Figure 3

Specifically, the resultant feature maps are calculated as:

$$R_i = f_i.A_i \tag{1}$$

The attention maps A_i have high scores for difficult pixels and low scores for easy pixels. It is mathematically formulated as:

$$A_i = \sigma(1- \mid A_{fi} - A_{ri} \mid) \tag{2}$$

where. A_{fi} represents the forward attention map, with a high score for the foreground object and a low score for the background area, and A_{ri} indicates the reverse attention map, which has a high score for the background pixels and low score for the foreground pixels. The attention maps can be mathematically formulated as:

$$A_{fi} = \sigma(Up(S_{i+1})) \tag{3}$$

$$A_{ri} = 1 - \sigma(Up(S_{i+1})) \tag{4}$$

where Up indicates the upsampling operation, $\sigma(.)$ represents the sigmoid activation, and S_i is the global map from the previous layer. Pixel's difficulty can be associated with the absolute difference between the forward and reverse attention score; the lower the absolute difference, the higher the difficulty. To focus more on the complex pixels, we further subtract the absolute difference from 1 followed by sigmoid activation as shown in equation 2.

3.4 Loss Function

We use the combination of weighted IoU loss L_{IOU}^{w} and weighted cross-entropy loss L_{BCE}^{w} as our main loss function [24][33]. We apply deep supervision after each resultant map generated by the attention module along with the initial global map. The total loss can be formulated as:

$$L_{total} = L(G, S_g^{up}) + \sum_{i=3}^{5} L(G, S_i^{up}) \tag{5}$$

where S_g is the global map and S_3, S_4, and S_5 are the output maps generated by the attention module.

4 Experiments

4.1 Datasets

We conducted experiments on five publicly available polyp segmentation datasets: ETIS, CVC-ClinicDB, CVC-ColonDB, CVC-300, and KVasir. ETIS is an old dataset with 196 polyp images and its ground truth mask. CVC-ClinicDB and CVC-300 comprise 612 and 300 images from 29 and 13 colonoscopy video sequences, respectively. CVC-ColonDB is a small-scale dataset containing 380 images from 15 short colonoscopy sequences. Kvasir dataset is relatively new, with 1000 polyp images. We compare our FuzzyNet with state-of-the-art models: PraNet, Enhanced U-Net, ACSNet, MSEG[10], SA-Net, TransFuse, and Polyp-PVT, along with the previous approaches U-Net, U-Net++, and ResU-Net++.

4.2 Evaluation Metrics

We utilize the Dice coefficient and Intersection over Union (IOU) as our evaluation metrics which are defined below:

Dice coefficient: It is defined as:

$$DSC(A, B) = \frac{2 \times (A \cap B)}{A + B} \tag{6}$$

where A denotes the predicted set of pixels and B is the ground truth of the image.

Intersection over union (IoU): It is another standard metric to evaluate the performance of the segmentation task. It is defined as:

$$IoU(A, B) = \frac{A \cap B}{A \cup B} \tag{7}$$

where A denotes the predicted set of pixels and B is the corresponding ground truth of the set of pixels.

4.3 Implementation Details

In our experiments, we follow the same training settings used in PraNet for the Res2Net backbone and Polyp-PVT for ConvNext and PVT backbone. All the models are trained on a V100 GPU, with batch-size 16 and Adam optimizer with an initial learning rate of 0.0001. We employ multi-scale training for all the backbone networks instead of data augmentation techniques by following the PraNet and PVT. We employ the backbone networks Res2Net, ConvNext, and Pyramid Vision Transformer, initialize the weights with pretrained weights trained on ImageNet-1K and train them from scratch.

model	CVC-ClinicDB		Kvasir	
	mDice	mIoU	mDice	mIoU
U-Net	0.823	0.755	0.818	0.746
U-Net++	0.794	0.729	0.821	0.743
SFA	0.700	0.607	0.723	0.611
ACSNet	0.882	826	0.898	0.838
PraNet	0.899	0.849	0.898	0.840
EU-Net	0.902	0.846	0.908	0.854
SA-Net	0.916	0.859	0.904	0.847
TransFuse	0.918	0.868	0.918	0.868
Polyp-PVT	0.937	0.889	0.917	0.864
Fuzzy-Net(Res2Net)	0.919	0.867	0.889	0.830
Fuzzy-Net(ConvNext)	0.922	0.863	0.907	0.848
Fuzzy-Net(PVT)	**0.937**	**0.889**	0.913	0.864

Table 1: Results on CVC-ClinicDB and Kvasir, which represents the learning capability of the model. It shows that our model outperforms the other models by a significant margin on the CVC-ClinicDB dataset and achieves a comparable result on the Kvasir dataset. The reported result is the average of three experiments.

4.4 Learning Ability

Setting: We evaluate the learning ability of our model on the dataset ClinicDB and Kvasir-Seg. Clinic-DB consists of 612 images extracted from 31 colonoscopy videos, whereas Kvasir-Seg consists of a total of 1000 polyp images. We follow the same setting as PraNet and Polyp-PVT, which include 900 and 548 images from ClinicDB and Kvasir-Seg datasets as the train set, and the remaining 64 and 100 images are used as the test set.

Results: Table 1 shows the result on Kvasir and CVC-ClinicDb datasets. Our FuzzyNet model with Res2Net as a backbone network achieves 2% higher mean dice than PraNet on CVC-ClinicDB and achieves comparable results on the Kvasir dataset which demonstrates the better learning ability of our model. Our model with ConvNext as a backbone network outperforms the ACSNet, PraNet, EU-Net, and SA-Net on the Clinic-DB dataset by 3.4%, 2.3%, 2%, and 0.6%, respectively, in terms of mean-dice. In addition, it also achieves 0.9%, 0.9%, and 0.03% higher mean dice than the ACSNet, PraNet, and SANet, respectively, on the Kvasir dataset. With Pyramid Vision Transformer as a backbone network, we achieve the state-of-the-art accuracy 0.937 on the CVC-ClinicDB dataset and comparable mean dice on the Kvasir dataset as Polyp-PVT.

4.5 Generalization Ability

Setting: To evaluate the generalization ability of the model, we use three unseen datasets: ETIS, ColonDB, and CVC-300. The ETIS, ColonDB, and CVC-300 datasets consist of a total of 190, 380, and 60 images, respectively. The images of these datasets belong to different medical centers, which means that the training and testing sets are different and the model has not seen the test images before during training.

Results: The result is shown in the table 2. It can be seen from the result that our model has a better generalization performance compared to state-of-the-art models. On ColonDB and ETIS, we achieved the highest performance by outperforming the Polyp-PVT. On CVC-300, our model achieves comparable performance to Polyp-PVT. In addition, our model with Res2Net as a backbone outperforms PraNet by 3.1%, 2.7%, and 10% on CVC-300, CVC-ColonDB, and ETIS respectively in terms of mean dice, which demonstrates that Fuzzy Attention module has an outstanding generalization ability compared with the Reverse Attention module in PraNet. With PVT as a backbone network, our model outperforms the latest SANet and TransFuse by 5.8% and 6.7% on CVC-ColonDB, 4.2% and 5.5% on ETIS respectively. It also achieves 1.1% higher mean dice than SA-Net and comparable mean dice as TransFuse on CVC-300.

model	CVC-ColonDB		ETIS		CVC-300	
	mDice	mIoU	mDice	mIoU	mDice	mIoU
U-Net	0.512	0.444	0.398	0.335	0.710	0.627
U-Net++	0.483	0.410	0.401	0.344	0.707	0.624
SFA	0.469	0.347	0.297	0.217	0.467	0.329
ACSNet	0.716	0.649	0.578	0.509	0.863	0.787
PraNet	0.712	0.640	0.628	0.567	0.871	0.797
EU-Net	0.756	0.681	0.687	0.609	0.837	0.765
SA-Net	0.753	0.670	0.750	0.654	0.888	0.815
TransFuse	0.773	0.696	0.733	0.659	**0.902**	**0.833**
Polyp-PVT	0.808	0.727	0.787	0.706	0.900	0.833
Fuzzy-Net(Res2Net)	0.739	0.662	0.731	0.658	0.894	0.825
Fuzzy-Net(ConvNext)	0.784	0.696	0.740	0.648	0.877	0.795
Fuzzy-Net(PVT)	**0.811**	**0.728**	**0.791**	**0.702**	0.891	0.818

Table 2: Results on CVC-ColonDB, ETIS, and CVC-300 demonstrate the generalization capability of the model. It shows that our model outperforms the other models by a significant margin on CVC-ColonDB and ETIS and achieves a comparable result on the CVC-300 dataset. The reported result is the average of three experiments.

model	Attention	mDice		mIoU		GFlops	Parameters
		Seen	Unseen	Seen	Unseen		
Pra-Net	Reverse	0.8985	0.737	0.8445	0.668	**13.11**	**32.55M**
Fuzzy-Net	Fuzzy	**0.904**	**0.788**	**0.845**	**0.715**	**13.11**	**32.55M**

Table 3: Comparison of the GFlops and the number of parameters of reverse attention-based Pra-Net and fuzzy attention-based FuzzyNet along with average mean dice and mean IoU calculated by taking the average of mean dice and mean IoU of all datasets included in type seen and unseen, respectively.

4.6 Effectiveness of Fuzzy Attention

The effectiveness of the Fuzzy attention module can be verified by comparing the result of our model with the Res2Net backbone with PraNet. For a fair comparison, apart from the type of attention module, we utilize a similar backbone, hyperparameters, augmentation, and regularization technique as used in PraNet.

It can be seen from the table our model outperforms PraNet on the CVC-ClinicDB dataset by 2% and achieves a comparable result on the Kvasir dataset. It improves the result on unseen datasets by a significant margin of 2.7%, 10%, and 1.9% on CVC-ColonDB, ETIS, and CVC-300 respectively. The overall result of all datasets along with GFlops and the total number of parameters are listed in Table 3. We can observe a significant performance gain of 0.9% and 5% on the average of seen and unseen datasets respectively. The outstanding results on the unseen dataset prove the generalization capability of the proposed fuzzy attention, which is significantly higher than reverse attention.

Furthermore, fuzzy attention does not add an extra number of parameters and the computation cost. With the same amount of parameters and GFlops: 32M and 13.11 GMac respectively, we achieve a significant performance gain compared to reverse attention-based PraNet.

In addition, we visualize the segmentation mask generated by our model and PraNet. It can be seen from the result shown in Figure 4 that our model correctly classifies the fuzzy pixels near the boundary compared to PraNet model. For most of the images, PraNet seems to be misclassified background pixels as foreground pixels because of the high focus on the background pixels in reverse attention. In contrast to PraNet, the resultant mask generated by our model has well-defined boundaries, and the results are closer to the ground truth. In addition, our model's performance is consistent irrespective of the lighting and reflection condition in the image. Furthermore, we also visualize the segmentation mask generated by Polyp-PVT and SA-Net. It can be seen from the figure that our resultant mask has fewer false positive pixels and a smooth boundary than all other methods.

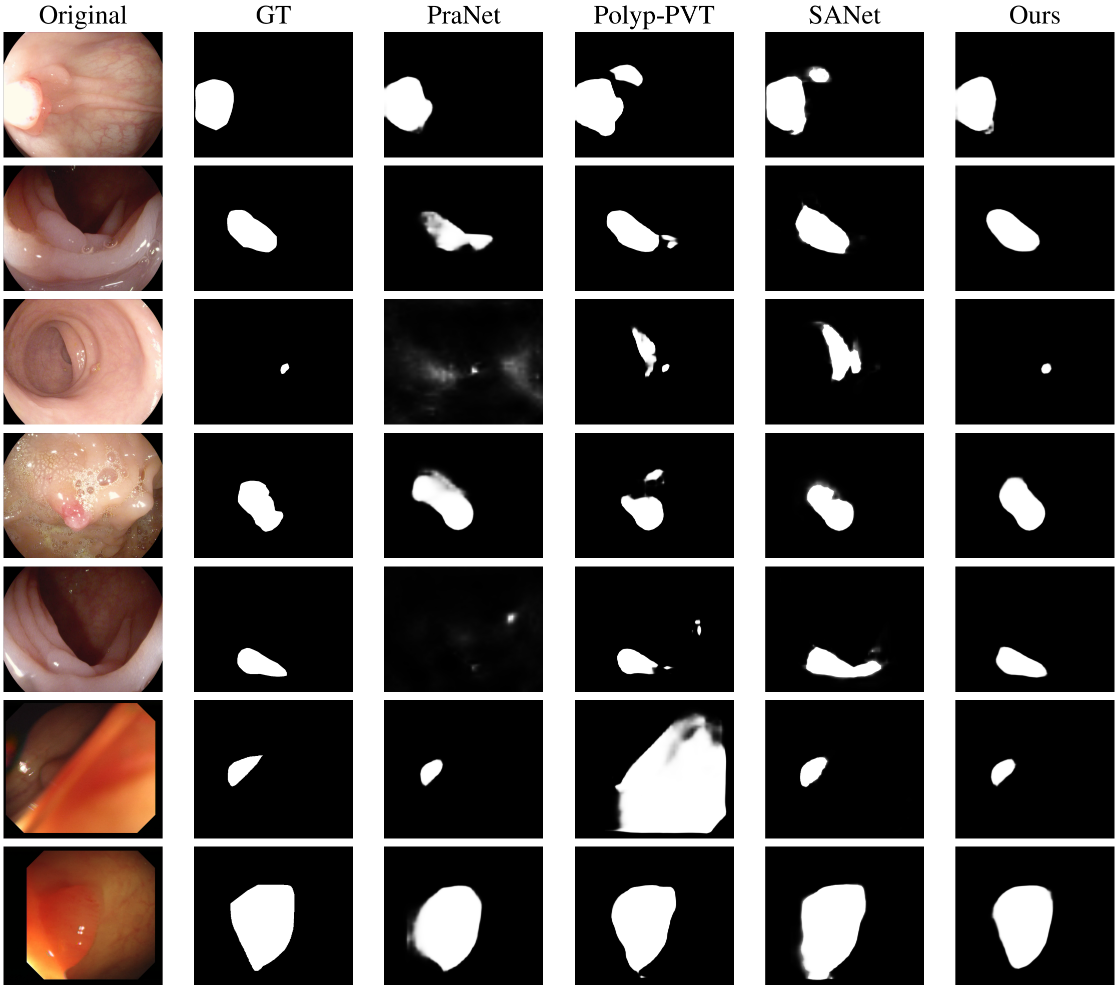

Figure 4: A comparison of the segmentation maps generated by our model, PraNet, Polyp-PVT and SA-Net along with the original image and ground-truth mask. Our model has more sharp boundaries than PraNet and the maps are closer to the ground-truth mask compared to other methods.The images are taken from ETIS and CVC-ColonDB dataset.

5 Conclusion

This paper has presented a novel attention mechanism to encourage the network to focus more on the fuzzy region, which usually lies around the boundary. We embed our attention module with various backbone networks: Res2Net, ConvNext, and Pyramid Vision Transformer (PVT) for polyp segmentation. Our result shows that the fuzzy attention module significantly outperforms PraNet, which employs the reverse attention mechanism on all polyp segmentation datasets. With the PVT as the backbone network, our model achieves state-of-the-art accuracy on the CVC-ClinicDB, CVC-ColonDB, and ETIS dataset for polyp segmentation.

Acknowledgement

This work was partly supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) under grant nos. RGPIN-2021-04244 and ALLRP 576612-22, and the National Institutes of Health (NIH) under grant no. 1R03CA253212-01.

References

[1] Mojtaba Akbari, Majid Mohrekesh, Ebrahim Nasr-Esfahani, SM Reza Soroushmehr, Nader Karimi, Shadrokh Samavi, and Kayvan 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)*, pages 69–72. IEEE, 2018.

[2] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015.

[3] Patrick Brandao, Evangelos Mazomenos, Gastone Ciuti, Renato Caliò, Federico Bianchi, Arianna Menciassi, Paolo Dario, Anastasios Koulaouzidis, Alberto Arezzo, and Danail Stoyanov. Fully convolutional neural networks for polyp segmentation in colonoscopy. In *Medical Imaging 2017: Computer-Aided Diagnosis*, volume 10134, pages 101–107. SPIE, 2017.

[4] Bo Dong, Wenhai Wang, Deng-Ping Fan, Jinpeng Li, Huazhu Fu, and Ling Shao. Polyp-pvt: Polyp segmentation with pyramid vision transformers. *arXiv preprint arXiv:2108.06932*, 2021.

[5] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 263–273. Springer, 2020.

[6] Yuqi Fang, Cheng Chen, Yixuan Yuan, and Kai-yu Tong. Selective feature aggregation network with area-boundary constraints for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 302–310. Springer, 2019.

[7] Kamala Gajurel, Cuncong Zhong, and Guanghui Wang. A fine-grained visual attention approach for fingerspelling recognition in the wild. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3266–3271. IEEE, 2021.

[8] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. *IEEE transactions on pattern analysis and machine intelligence*, 43(2):652–662, 2019.

[9] Lei He, Jiwen Lu, Guanghui Wang, Shiyu Song, and Jie Zhou. Sosd-net: Joint semantic object segmentation and depth estimation from monocular images. *Neurocomputing*, 440:251–263, 2021.

[10] Chien-Hsiang Huang, Hung-Yu Wu, and Youn-Long Lin. 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.

[11] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International Conference on Multimedia Modeling*, pages 451–462. Springer, 2020.

[12] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Dag Johansen, Thomas De Lange, Pål Halvorsen, and Håvard D Johansen. Resunet++: An advanced architecture for medical image segmentation. In *2019 IEEE International Symposium on Multimedia (ISM)*, pages 225–2255. IEEE, 2019.

[13] Xiao Jia, Xiaohan Xing, Yixuan Yuan, Lei Xing, and Max Q-H Meng. Wireless capsule endoscopy: A new tool for cancer screening in the colon with deep-learning-based polyp recognition. *Proceedings of the IEEE*, 108(1):178–197, 2019.

[14] Kaidong Li, Mohammad I Fathan, Krushi Patel, Tianxiao Zhang, Cuncong Zhong, Ajay Bansal, Amit Rastogi, Jean S Wang, and Guanghui Wang. Colonoscopy polyp detection and classification: Dataset creation and comparative evaluations. *Plos one*, 16(8):e0255809, 2021.

[15] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

[16] Wenchi Ma, Tianxiao Zhang, and Guanghui Wang. Miti-detr: Object detection based on transformers with mitigatory self-attention convergence. *arXiv preprint arXiv:2112.13310*, 2021.

[17] Omid Haji Maghsoudi. Superpixel based segmentation and classification of polyps in wireless capsule endoscopy. In *2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pages 1–4. IEEE, 2017.

[18] Alexander V Mamonov, Isabel N Figueiredo, Pedro N Figueiredo, and Yen-Hsi Richard Tsai. Automated polyp detection in colon capsule endoscopy. *IEEE transactions on medical imaging*, 33(7):1488–1502, 2014.

[19] Prashant Mathur, Krishnan Sathishkumar, Meesha Chaturvedi, Priyanka Das, Kondalli Lakshminarayana Sudarshan, Stephen Santhappan, Vinodh Nallasamy, Anish John, Sandeep Narasimhan, Francis Selvaraj Roselind, et al. Cancer statistics, 2020: report from national cancer registry programme, india. *JCO Global oncology*, 6:1063–1075, 2020.

[20] Balamurali Murugesan, Kaushik Sarveswaran, Sharath M Shankaranarayana, Keerthi Ram, Jayaraj Joseph, and Mohanasankar Sivaprakasam. Psi-net: Shape and boundary aware joint multi-task deep network for medical image segmentation. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 7223–7226. IEEE, 2019.

[21] Krushi Patel, Andres M Bur, Fengjun Li, and Guanghui Wang. Aggregating global features into local vision transformer. *arXiv preprint arXiv:2201.12903*, 2022.

[22] Krushi Patel, Andrés M Bur, and Guanghui Wang. Enhanced u-net: A feature enhancement network for polyp segmentation. In *2021 18th Conference on Robots and Vision (CRV)*, pages 181–188. IEEE, 2021.

[23] Krushi Patel, Kaidong Li, Ke Tao, Quan Wang, Ajay Bansal, Amit Rastogi, and Guanghui Wang. A comparative study on polyp classification using convolutional neural networks. *PloS one*, 15(7):e0236452, 2020.

[24] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7479–7489, 2019.

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[26] Usman Sajid, Michael Chow, Jin Zhang, Taejoon Kim, and Guanghui Wang. Parallel scale-wise attention network for effective scene text recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

[27] Usman Sajid and Guanghui Wang. Towards more effective prm-based crowd counting via a multi-resolution fusion and attention network. *Neurocomputing*, 474:13–24, 2022.

[28] Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International journal of computer assisted radiology and surgery*, 9(2):283–293, 2014.

[29] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*, 35(2):630–644, 2015.

[30] David Vázquez, Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Antonio M López, Adriana Romero, Michal Drozdzal, and Aaron Courville. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of healthcare engineering*, 2017, 2017.

[31] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021.

[32] Jun Wei, Yiwen Hu, Ruimao Zhang, Zhen Li, S Kevin Zhou, and Shuguang Cui. Shallow attention network for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 699–708. Springer, 2021.

[33] Jun Wei, Shuhui Wang, and Qingming Huang. F^3net: fusion, feedback and focus for salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12321–12328, 2020.

[34] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3907–3916, 2019.

[35] Ruifei Zhang, Guanbin Li, Zhen Li, Shuguang Cui, Dahong Qian, and Yizhou Yu. Adaptive context selection for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 253–262. Springer, 2020.

[36] Yundong Zhang, Huiye Liu, and Qiang Hu. Transfuse: Fusing transformers and cnns for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 14–24. Springer, 2021.

[37] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming 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*, pages 3–11. Springer, 2018.