ARTICLE

Boundary-aware Feature and Prediction Refinement for Polyp Segmentation

Jie Qiu[a] Yuichiro Hayashi[a] Masahiro Oda[b,a] Takayuki Kitasaka[c] and Kensaku Mori[a,b,d]

[a]Graduate School of Informatics, Nagoya University, Nagoya, Japan; [b]Information and Communications, Nagoya University, Nagoya, Japan; [c]Department of Information Science, Aichi Institute of Technology, Toyota, Japan; [d]Research Center for Medical Bigdata, National Institute of Informatics, Tokyo, Japan

ARTICLE HISTORY
Compiled September 14, 2022

ABSTRACT
Polyp segmentation from colonoscopy videos is an essential task in medical image processing for detecting early cancer. However, segmenting a precise boundary is still challenging, even with powerful deep neural networks. We consider the difficulty can be caused by: 1) the ambiguity boundary and 2) some complicated shape makes polyps hard to segment. To address these problems, we propose the Boundary-aware Feature and Prediction Refinement framework (BaFPR) for polyp segmentation. Specifically, we design a segmentation decoder for representation learning with boundary prior and propose a novel consistency loss to learn clues from the polar coordinate. The decoder mainly consists of a boundary prior module (BPM) and a bi-directional fusion module (BiFM). BPM is designed to learn the boundary prior, while BiFM learns to fuse representations of BPM and multi-scale representations from an encoder. To handle these complicated shapes of polyps, we maintain an extra segmentation network that learns with polar transformations of data to provide extra clues for the main segmentation network by our proposed consistency loss. We evaluated BaFPR with 5 challenging datasets for polyp segmentation. and the results showed that our proposal consistently improves the segmentation performance of polyps. Code available at: https://github.com/MoriLabNU/BaFPR.

KEYWORDS
polyp segmentation; boundary-aware; consistency learning

1. Introduction

Polyp segmentation is an essential task in medical image processing. For example, the regions of polyps can provide valuable information such as the size to evaluate how possible these polyps are/would be cancer (Ponugoti et al. 2017), which potentially allows an automatic polyp analysis system. In recent years, with the development of powerful deep neural networks (DNNs), a large proportion of DNNs-based research has investigated the segmentation problem and achieved significant improvement (Chen et al. 2018; Fan et al. 2020; Fang et al. 2019; Jha et al. 2019; Lee et al. 2020; Wang et al. 2021; Wei et al. 2021; Zhang et al. 2021, 2018). A typical example is called

Contact Jie Qiu Email: jieqiu@mori.m.is.nagoya-u.ac.jp; Kensaku Mori Email: kensaku@is.nagoya-u.ac.jp

U-Net (Ronneberger et al. 2015), which is one of the earliest approaches using the advantages of encoder-decoder structure for medical image segmentation. The encoder part is responsible for representation embedding, and the decoder part generates the prediction mask. Based on U-Net, a series of segmentation models further improve the performance. U-Net++ (Zhou et al. 2018) develops the U-Net into a nested structure and ResUNet++ (Jha et al. 2019) takes the additional advantages of squeeze and excitation block (Hu et al. 2018), atrous spatial pyramid pooling (Chen et al. 2018), and attention block to improve performance.

For polyp segmentation, PraNet (Fan et al. 2020) aggregates a multi-level feature and uses reverse attention to build the relation among regions for polyp segmentation, while SANet (Wei et al. 2021) maintains the multi-level feature maps but uses a shallow 1×1 convolution attention to obtain a faster speed. UACANet (Kim et al. 2021) designs a progressive saliency map propagation module to refine the predictions. More recently, vision transformer has been widely applied in segmentation models to capture the long-range information in images. For example, Polyp-PVT (Dong et al. 2021) takes PVT (Wang et al. 2022a) as backbone to fuse the multi-scale representations from different encoder layers.

However, providing a fine-level segmentation of polyps is still challenging, especially for the boundary. In this work, we consider the difficulty for the following two reasons: 1) the ambiguity boundary, and 2) some complicated shape makes the polyp hard to segment. To handle these problems, we propose a polyp segmentation framework that 1) we design a segmentation decoder for representation learning with boundary prior and 2) we propose a novel consistency loss to learn the clues from polar coordinate, where the round-shaped polyps are transformed into a straight-line-like object. The decoder mainly consists of a boundary prior module (BPM) and a bi-directional fusion module (BiFM). BPM is designed to learn the boundary prior information by the distance transformation of ground-truth segmentation masks. We utilize the learned representations of BPM to provide boundary prior information to BiFM module, where BiFM fuses the boundary prior information from BPM and the multi-scale representations from the encoder. The main reason we choose a polar view to provide extra clues is that polar transformation transfers some complicated boundaries with round shapes to the easy-to-segment straight-line-like boundaries (Zhang et al. 2020; Xu et al. 2022). In order to learn the clues from the polar view, we maintain an extra segmentation network that learns with the polar transformation of the input images and the ground-truth segmentation mask. The center of polar transformation is calculated from the ground-truth segmentation mask for training and from the prediction mask of the main segmentation network for inference. The proposed loss encourages the main segmentation model to learn from the consistency between the predictions of the main segmentation network and the inverse polar transformation of the predictions of the polar segmentation network.

In this work, we aim to provide more precise semantic information such as shape for polyps for further analysis. In summary, The main contributions of this paper are the following 3-folds:

- We design a novel decoder for polyp segmentation. The decoder consists of BPM module and BiFM module. BPM learns boundary prior information of polyps, and BiFM utilizes such boundary prior for representation learning.
- We present a consistency loss to learn the prediction clues from a polar view.
- Extensive experiments have been conducted in five datasets for polyp segmentation to validate the advance of our proposal. Further ablation studies show the

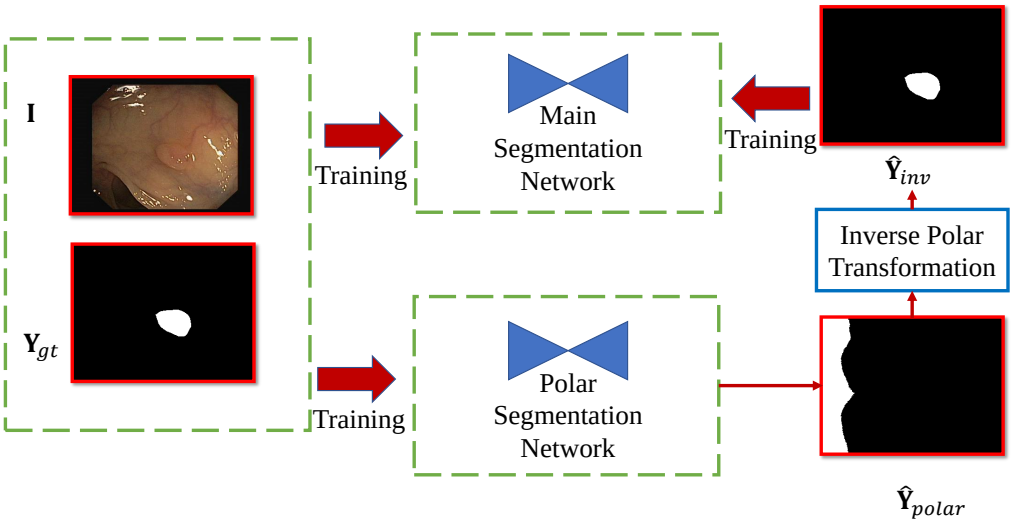

Figure 1.: Overview of our proposed BaFPR. We maintain a main segmentation network and a polar segmentation network for training. $\mathbf{I}$: input image. $\mathbf{Y}_{gt}$: corresponding ground-truth for $\mathbf{I}$. $\hat{\mathbf{Y}}_{polar}$: prediction from polar segmentation network. $\hat{\mathbf{Y}}_{inv}$: inverse polar transformed $\hat{\mathbf{Y}}_{polar}$.

effectiveness of our decoder design as well as the proposed loss.

2. Methodology

Figure 1 shows the overall framework of our proposal. We name it boundary-aware feature and prediction refinement (BaFPR). We maintain one main segmentation network and one polar segmentation network for training. The details of the main segmentation network are given in Fig. 2, where a boundary prior module (BPM), and a bi-directional fusion module (BiFM) are proposed for representation learning. The detailed structure of BiFM is given in Fig. 3. Besides learning from Cartesian view, Fig. 1 and Fig. 2 show our main segmentation network learns from a polar view of polar segmentation network at the same time. And a detailed training scheme of polar segmentation networks is given in Fig. 4.

In Fig. 2, an encoder outputs multi-scale feature maps at the beginning stage. Most of the multi-scale backbones can be used as the encoder, such as Res2Net (Gao et al. 2019), and PVT (Wang et al. 2022b). In this section, we will follow an important previous work (Dong et al. 2021) to use PVT as our encoder. By considering the input image to be $\mathbf{I} \in \mathbf{R}^{H\times W\times 3}$, where H and W are the height and width of the image, we denote the output feature maps of the PVT encoder from the largest scale as $\mathbf{X}_1 \in \mathbf{R}^{\frac{H}{4}\times\frac{W}{4}\times C}, \mathbf{X}_2 \in \mathbf{R}^{\frac{H}{8}\times\frac{W}{8}\times C}, \mathbf{X}_3 \in \mathbf{R}^{\frac{H}{16}\times\frac{W}{16}\times C}, \mathbf{X}_4 \in \mathbf{R}^{\frac{H}{32}\times\frac{W}{32}\times C}$ sequentially, and C is the embedded channel number which set to be 32 in our experiments. Further experiments with different architecture can be found in the experiment section.

For the decoder, we design a boundary prior module (BPM) for the feature map $\mathbf{X}_1$ as shown in Fig. 2 (b). The BPM learns from the distance transformation (denoted

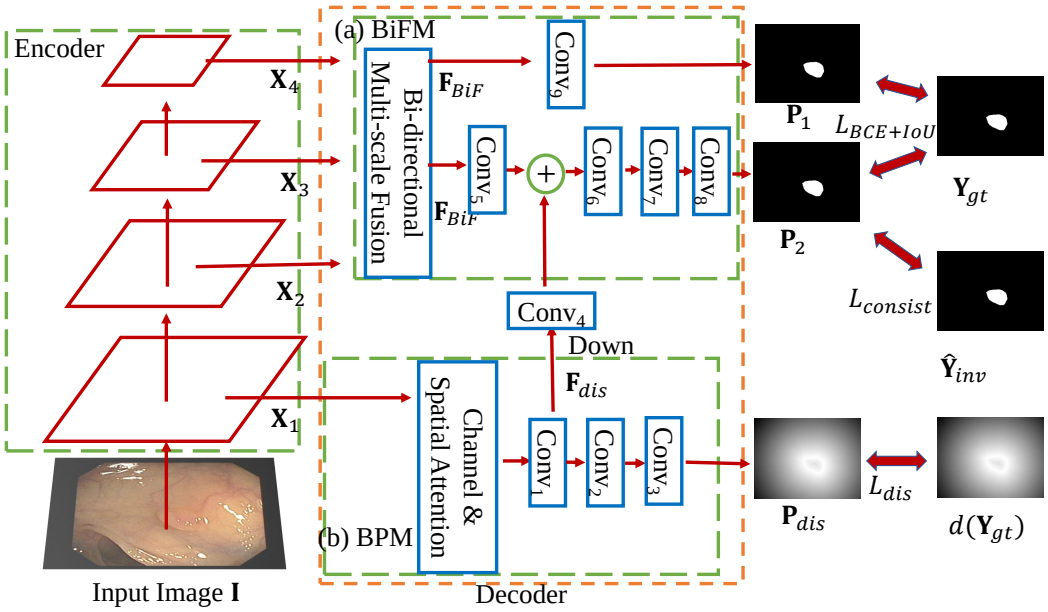

Figure 2.: The overall framework of the main segmentation network. Block (a): Bi-directional Fusion Module (BiFM). Block (b): Boundary Prior Module (BPM).

as $d(\cdot)$) of the binary ground-truth mask $\mathbf{Y}_{gt}$ during training by a mean square error loss L_{dis}, and passes the middle representation maps $\mathbf{F}_{dis}$ to the bi-directional fusion module (BiFM, Fig. 2 (a)) as the boundary prior. The BiFM also benefits from the multi-scale feature maps by a bi-directional multi-scale fusion block (Fig. 3). The multi-scale feature maps are passed to a bi-directional multi-scale fusion component (Fig. 2 (a), left), thus fused from the top scale to the button scale and versa vice. We denote the fused feature maps as $\mathbf{F}_{BiF}$, which are passed to a convolution prediction block to form the prediction $\mathbf{P}_1$ and are combined with the boundary prior $\mathbf{F}_{dis}$ to form the prediction $\mathbf{P}_2$. $L_{BCE+IoU}$ is a combination of IoU loss and binary cross-entropy to optimize the predictions $\mathbf{P}_1$ and $\mathbf{P}_2$.

In Fig. 4, we give a training scheme of the polar segmentation network. We propose a consistency loss $L_{consist}$ to learn the clues from $\hat{\mathbf{Y}}_{inv}$, the inverse polar transformation of the prediction masks of the polar segmentation network. The polar segmentation network utilizes the same network structure but without $L_{consist}$ or L_{dis}.

The organization of this section will be as follows: we will introduce the main components of our decoder BiFM and BPM in Sec. 2.1, and in Sec. 2.2 we will talk about the details of the consistency learning between the main segmentation network and the polar segmentation network. In Sec. 2.3, the overall optimization function will be introduced.

2.1. Decoder Design

The main decoder components consist of 1) a boundary prior module (BPM) to learn the boundary prior information and 2) a bi-directional fusion module (BiFM) to fuse the multi-scale feature maps of the encoder and the boundary prior information from BPM.

2.1.1. Boundary Prior Module

Boundary provide essential information for segmentation. However, boundaries in ground-truth masks are not always reliable in polyp segmentation. Therefore, direct optimization on boundary distance may lead to degradation of performance. Instead of directly optimizing boundary distance in previous works (Kervadec et al. 2019; Wang et al. 2020), we leverage an additional boundary prior module to learn a relaxation of boundary information called boundary prior for later segmentation usage. The details of the boundary prior module are shown in the button part of Fig. 2. It takes the largest feature map $\mathbf{X}_1$ of the encoder as an input and outputs its middle representation $\mathbf{F}_{dis}$ as the boundary prior and $\mathbf{P}_{dis}$ as distance map prediction.

Channel and spatial attention is commonly used to extract crucial information for noisy feature maps (Dong et al. 2021; Chen et al. 2017; Roy et al. 2018). In this work, we adopt the same channel and spatial attention design as Polyp-PVT (Dong et al. 2021) to extract the crucial information about the boundary. Therefore, the feature map $\mathbf{X}_1$ is fed into the channel and spatial attention block denoted as $\text{attn}_\text{S}(\text{attn}_\text{C}(\mathbf{X}_1))$, where attn_S and attn_C are the function for spatial and channel attention, respectively. We indicate $\text{Conv}_i(\cdot)$ as a convolution block with kernel size 3×3 and padding size 1 followed by batch normalization and a relu activation function, thus $\mathbf{F}_{dis}$ and $\mathbf{P}_{dis}$ can be written as

$$\mathbf{F}_{dis} = \text{Down}(\text{Conv}_1(\text{attn}_\text{S}(\text{attn}_\text{C}(\mathbf{X}_1)))) \tag{1}$$

$$\mathbf{P}_{dis} = \text{Sigmoid}(\text{Conv}_3(\text{Conv}_2(\text{Conv}_1(\text{attn}_\text{S}(\text{attn}_\text{C}(\mathbf{X}_1)))))), \tag{2}$$

where $\text{Down}(\cdot)$ is a downsampling function for the feature map and $\text{Sigmoid}(\cdot)$ indicates a sigmoid function. The output $\mathbf{P}_{dis}$ is optimized with the distance transformation of the ground-truth mask for training by L_{dis}, and $\mathbf{F}_{dis}$ is treated as boundary prior for the feature fusion in BiFM, which will be introduced in Sec. 2.1.2.

2.1.2. Bi-directional Fusion Module

The BiFM module (Fig. 2 (a)) consists of a bi-directional multi-scale fusion block and a convolution fusion block as main components. The design of the bi-directional multi-scale fusion block is based on the feature fusion block of (Dong et al. 2021), and we extend it to a bi-direction manner to capture the coarse-to-fine and fine-to-coarse boundary information. After obtaining the fused feature maps, a convolution fusion block takes the $\mathbf{F}_{dis}$ from BPM as guidance for further feature fusion. The whole module takes feature maps $\mathbf{X}_2, \mathbf{X}_3$ and $\mathbf{X}_4$ as inputs and outputs the prediction maps $\mathbf{P}_1$ and $\mathbf{P}_2$.

The details of bi-directional multi-scale fusion block are given in Fig. 3. $\mathbf{F}_{Up}$ and $\mathbf{F}_{Down}$ are the coarse-to-fine and fine-to-coarse feature maps. Coarse-to-fine feature map help to refine the details of the boundary, while the fine-to-coarse feature map will preserve a reliable boundary. Thus, we then fuse $\mathbf{F}_{Up}$ and $\mathbf{F}_{Down}$ as

$$\mathbf{F}_{BiF} = \text{Conv}(\text{Concat}(\mathbf{F}_{Up}, \mathbf{F}_{Down})), \tag{3}$$

where $\text{Concat}(\cdot)$ concatenates the feature maps along with the channel dimension and

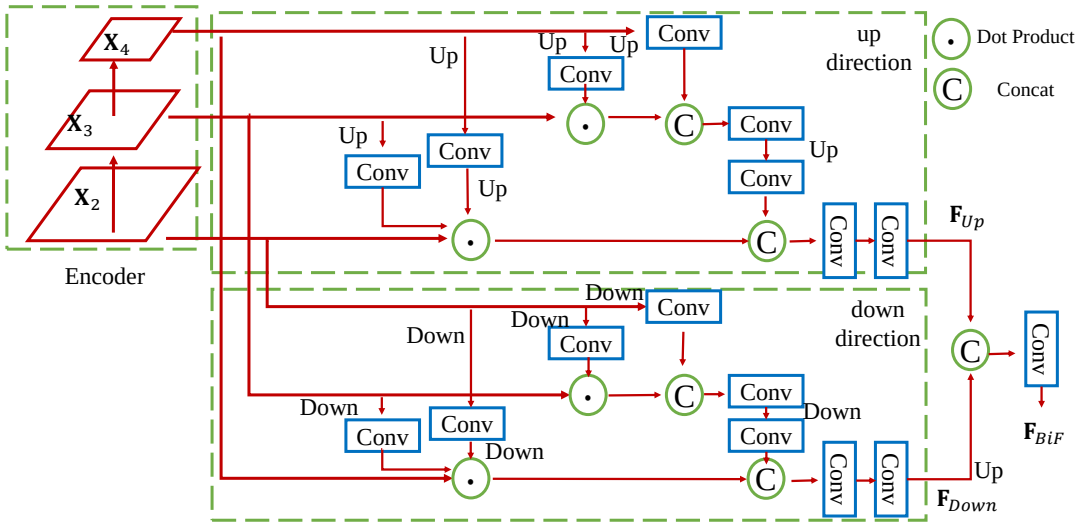

Figure 3.: The framework of the bi-directional multi-scale fusion block. Down and up indicate down-sampling and up-sampling, respectively. The upper block follows a up-sampling scheme and the button follows a down-sampling scheme.

Conv$(\cdot)$ indicates a convolution block. $\mathbf{F}_{BiF}$ is further passed to another convolution layer to form $\mathbf{P}_1$. Furthermore, the convolution fusion block takes the fused representation $\mathbf{F}_{BiF}$ and the boundary prior $\mathbf{F}_{dis}$ as inputs to generate prediction $\mathbf{P}_2$ as

$$\mathbf{P}_2 = \text{Sigmoid}(\text{Conv}_8(\text{Conv}_7(\text{Conv}_6(\text{Conv}_4(\mathbf{F}_{dis}) + \text{Conv}_5(\mathbf{F}_{BiF}))))). \quad (4)$$

2.2. Consistency Learning from Polar View

Some polyps have complicated boundaries with approximately round shapes. And such shapes make them hard to be segmented precisely by a DNN-based model. Polar transformation can reorganize the coordinate from Cartesian view to polar view according to a given center, and makes the round boundary straight-line-like, which leads to easier segmentation and analysis in some cases. This is why we choose a polar view to provide extra clues to our main segmentation network. On the other hand, many previous works have applied polar transformation in segmentation tasks and achieved significant improvements (Xu et al. 2022; Zhang et al. 2020; Benčević et al. 2021). Most of these works adopt a centroid estimation network to estimate centroids of mass for polar transformation, and a polar segmentation network to learn to segment on transformed images. By contrast, we use a main segmentation network to do segmentation in Cartesian coordinate and calculate centroids based on ground-truth masks for training and predictions of main segmentation network for inference. Polar segmentation network and polar transformation mainly provide clues from polar view to our main segmentation network by our proposed consistency loss. And we ensemble prediction results from main segmentation network and polar segmentation network to refine prediction masks.

Figure 4 shows how we apply polar transformation to inputs to train our polar segmentation network and inverse polar transformation to provide clues. We first calculate

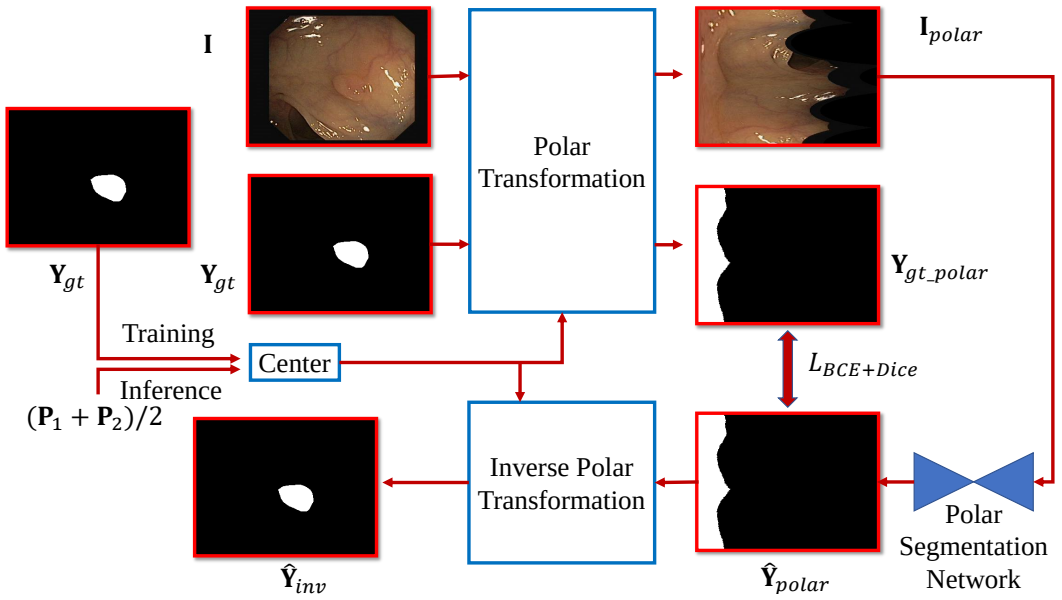

Figure 4.: Polar transformation and polar segmentation network.

the polar origin (center of mass) c_X and c_Y from the ground-truth segmentation mask $\mathbf{Y}_{gt}$ for training and from $\frac{\mathbf{P}_1+\mathbf{P}_2}{2}$ for inference. Given the polar origin, the position (ρ, ϕ) in the polar coordinate can be written as [1]

$$\rho = \frac{H}{2\pi}\frac{180}{\pi}\tan^{-1}\left(\frac{y - c_Y}{x - c_X}\right) \tag{5}$$

$$\phi = \frac{W}{\sqrt{(W/2)^2 + (H/2)^2}}\sqrt{(x - c_X)^2 + (y - c_Y)^2} \tag{6}$$

where x, y indicate the position in Cartesian coordinate and H and W are the height and width. And by applying this transformation, we can transfer $\mathbf{I}$ to $\mathbf{I}_{polar}$. Therefore, we can use $\mathbf{I}_{polar}$ and $\mathbf{Y}_{gt\_polar}$ to train the polar segmentation network and leverage the inverse polar transformed prediction $\hat{\mathbf{Y}}_{inv}$ of the prediction $\hat{\mathbf{Y}}_{polar}$ to provide consistency clues to the main segmentation network.

2.3. Overall Optimization

The objective function of our main segmentation network consists of 4 terms, $L_{BCE\_IoU}(\mathbf{P}_1, \mathbf{Y}_{gt})$, $L_{BCE\_IoU}(\mathbf{P}_2, \mathbf{Y}_{gt})$, $L_{consist}(\mathbf{P}_2, \hat{\mathbf{Y}}_{inv})$ and $L_{dis}(\mathbf{P}_{dis}, d(\mathbf{Y}_{gt}))$. Within these loss functions, L_{BCE_IoU} combines binary cross-entropy and weighted intersection over union. Both $L_{consist}$ and L_{dis} are mean square error (MSE) loss but used for different purposes. The distance transformation $d(\mathbf{Y}_{gt})$ can be written as

[1] We refer to (Benčević et al. 2021) and *cv.linearPolar* in OpenCV documentation for the definitions of Eq. 5 and Eq. 6. We use atan2 implementation for $\tan^{-1}$ to limit the range.

$$d(\mathbf{Y}_{gt}) = \left(1 - \frac{dist(\mathbf{Y}_{gt})}{\max dist(\mathbf{Y}_{gt})}\right), \tag{7}$$

where

$$dist(\mathbf{Y}_{gt}) = \begin{cases} \min_{y \in \partial \mathbf{Y}_{gt}} \text{euc}(x, y), & x \in \mathbf{Y}_{gt} \ , \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

where $\partial \mathbf{Y}_{gt}$ indicates the boundary, x and y are two pixels in $\mathbf{Y}_{gt}$ and $\text{euc}(\cdot)$ is a function to calculate the Euclidean distance between two pixel positions. By applying the distance transformation, the values next to the boundary in the segmentation mask will be close to 1, and the values will decrease gradually to 0 as they are far from the boundary.

Therefore, the final objective function for the main segmentation network can be written as

$$\begin{aligned} L = & L_{BCE\_IoU}(\mathbf{P}_1, \mathbf{Y}_{gt}) + L_{BCE\_IoU}(\mathbf{P}_2, \mathbf{Y}_{gt}) + \\ & \lambda L_{consist}(\mathbf{P}_2, \hat{\mathbf{Y}}_{inv}) + L_{dis}(\mathbf{P}_{\mathbf{dis}}, d(\mathbf{Y}_{gt})) \end{aligned} \tag{9}$$

where λ is a hyper-parameter to balance the contribution of clues from polar view (set to 2 for all the experiments). Besides, for the polar segmentation network, it learns from $L_{BCE\_IoU}(\hat{\mathbf{Y}}_{polar}, \mathbf{Y}_{gt\_polar})$ as shown in Fig. 4.

3. Experiment

3.1. Experimental Settings and Datasets

Following previous works in polyp segmentation from colonoscopy videos, we validated our proposal on 5 datasets, Kvasir-SEG (Jha et al. 2020), CVC-ClinicDB (Bernal et al. 2015), CVC-ColonDB (Bernal et al. 2012), EndoScene (Vázquez et al. 2017), and ETIS (Silva et al. 2014). Kvasir-SEG consists of 1000 images for gastrointestinal polyp segmentation with various resolutions, while CVC-ClinicDB comprises 612 images from 31 colonoscopy sequence. And CVC-ColonDB, EndoScene and ETIS contain 380, 60 and 196 images, respectively. For a fair comparison, we adopted the same training and test split as PraNet (Fan et al. 2020), SANet (Wei et al. 2021) and Polyp-PVT (Dong et al. 2021). Specifically, 900 and 550 images from Kvasir-SEG and CVC-ClinicDB are used as training sets, while the other 100 and 62 images are kept for testing. In order to examine the generalization ability, the remaining 3 datasets are kept for testing.

We resized images to $(352, 352)$ pixels for training and testing. We performed only multi-scale training with scale ratio of $(0.75, 1, 1.25)$ for data augmentation, which is consistent with the previous works (Dong et al. 2021; Wei et al. 2021; Fan et al. 2020). For polar transformation, we set the size of polar image to be the same size as in Cartesian view (H, W for height and width in our case). We used AdamW (Loshchilov and Hutter 2018) as an optimizer for 40 epochs training with an initial training rate at 10^{-4}, and a batch size of 16. We reported the results for models trained with 40 epochs in all experiments unless specified, rather than reporting results from the best

model. We implemented the whole framework with Pytorch, and all the experiments were conducted with a single NVIDIA Tesla V100 GPU. It took about 2.5 hours for training, and single inference cost about 14.75 ms for main segmentation network.

Mean IoU, mean Dice, weighted F-measure, and S-measure are adopted as evaluation metrics. Mean IoU (mIoU) and mean Dice (mDice) are widely-used metrics considering the intersection and union between the prediction and ground truth. Weighted F-measure (Margolin et al. 2014) F_β^w utilizes a Gaussian dependency weighting matrix to weight the error maps by considering the importance of each pixel. S-measure (Fan et al. 2017) S_α further combines structure similarity (Wang et al. 2004) to its metric. We applied our training setting (training epochs, data augmentation, etc.) to Polyp-PVT[2] (Dong et al. 2021) and conducted all experiments for 4 times if not specified. Results for U-Net, UNet++, PraNet, and SANet are cited from PraNet (Fan et al. 2020) and Polyp-PVT (Dong et al. 2021).

Table 1.: Quantitative evaluation on different datasets.

(a) Quantitative evaluation of CVC-ColonDB. Best result in bold.

method	mDice	mIoU	F_β^w	S_α
U-Net	0.512	0.444	0.498	0.712
UNet++	0.483	0.410	0.467	0.691
PraNet	0.712	0.640	0.699	0.820
SANet	0.753	0.670	0.726	0.837
Polyp-PVT	0.783	0.704	0.775	0.852
BaFPR(w/o ens.)	0.791	0.709	0.779	0.854
BaFPR	**0.797**	**0.720**	**0.787**	**0.859**

(b) Quantitative evaluation of EndoScene. Best result in bold.

method	mDice	mIoU	F_β^w	S_α
U-Net	0.710	0.627	0.684	0.843
UNet++	0.707	0.624	0.687	0.839
PraNet	0.871	0.797	0.843	0.925
SANet	0.888	0.815	0.859	0.928
Polyp-PVT	0.893	0.824	0.882	0.930
BaFPR(w/o ens.)	0.896	0.828	**0.885**	0.931
BaFPR	**0.897**	**0.832**	**0.885**	**0.932**

(c) Quantitative evaluation of ClinicDB. Best result in bold.

method	mDice	mIoU	F_β^w	S_α
U-Net	0.823	0.755	0.811	0.889
UNet++	0.794	0.729	0.785	0.873
PraNet	0.899	0.849	0.896	0.936
SANet	0.916	0.859	0.909	0.939
Polyp-PVT	0.924	0.875	0.923	0.941
BaFPR(w/o ens.)	0.926	0.877	0.924	0.942
BaFPR	**0.930**	**0.883**	**0.928**	**0.944**

(d) Quantitative evaluation of ETIS. Best result in bold.

method	mDice	mIoU	F_β^w	S_α
U-Net	0.398	0.335	0.366	0.684
UNet++	0.401	0.344	0.390	0.683
PraNet	0.628	0.567	0.600	0.794
SANet	0.750	0.654	0.685	0.849
Polyp-PVT	0.779	0.695	0.754	0.866
BaFPR(w/o ens.)	0.778	0.698	0.754	0.868
BaFPR	**0.793**	**0.720**	**0.774**	**0.879**

(e) Quantitative evaluation of Kvasir. Best result in bold.

method	mDice	mIoU	F_β^w	S_α
U-Net	0.818	0.746	0.794	0.852
UNet++	0.821	0.743	0.808	0.862
PraNet	0.898	0.840	0.885	0.915
SANet	0.904	0.847	0.892	0.915
Polyp-PVT	0.911	0.860	**0.905**	0.922
BaFPR(w/o ens.)	**0.915**	0.861	0.904	**0.923**
BaFPR	0.914	**0.862**	0.904	0.922

3.2. Result Analysis

Tables 1(a), 1(b), 1(c), 1(d), and 1(e) showed the quantitative evaluation for CVC-ColonDB, Endoscene, CVC-ClinicDB, ETIS, and Kvasir datasets, respectively. For fair

[2]For reference, we run the released code by the original authors of Polyp-PVT with their settings. The mDice for the model trained with 100 epochs are 0.881, 0.935, 0.804, 0.781, and 0.918 for Endoscene, CVC-ClinicDB, CVC-ColonDB, ETIS, and Kvasir, respectively.

comparisons, we listed our results with and without prediction ensemble. We referred BaFPR (w/o ens.) as training with all the proposed components but inferring with merely the main segmentation network, and BaFPR to inferring with ensemble by average predictions of main segmentation network and polar segmentation network.

Table 1(a) showed results for ColonDB dataset. BaFPR brought about 2% improvement over Polyp-PVT of mDice, mIoU, and F_β^w. Even without prediction ensemble, BaFPR(w/o ens.) consistently surpassed Polyp-PVT and other methods. Table 1(b) showed results for Endoscene dataset. Though both BaFPR (w/o ens.) and BaFPR were ahead of other comparisons, the advance seems less convincing. We carefully investigated the Endoscene dataset, and we found some frames in the Endoscene dataset were relatively easy to predict, while others were difficult due to motion blur and occlusion. For Table 1(c), less than 1% improvements of BaFPR compared to Polyp-PVT were observed. Many non-round-shaped polyps exist in ClinicDB dataset, which we assumed would limit the performance of our polar transformation. And for other round-shaped polyps, either SANet, Polyp-PVT, or our proposal could achieve satisfactory predictions. Table 1(d) presented results of ETIS dataset. BaFPR exceeded Polyp-PVT by about 2% and 3% of mDice and mIoU, and surpassed 6% performance than SANet, PraNet. Table 1(e) listed results for the Kvasir dataset. Although either BaFPR (w/o ens.) or BaFPR outperformed other methods, we observed an interesting finding that using the prediction ensemble harmed the final performance scores. We carefully inspected Kvasir dataset, and found it common that a single frame in Kvasir contains multiple polyps. The existence of many polyps made the center of polar transformation hard to be calculated, which could lead to bad performance scores.

We further examined the variance of mDice of predictions for 4 runs as shown in Table 2. We noticed our proposal worked more stable in Endoscene and CVC-colonDB datasets. For ETIS, neither Polyp-PVT nor our proposal could achieve a stable performance score. In CVC-ClinicDB and Kvasir, our proposal had larger variances than Polyp-PVT, which matched the our statement of improvement above.

Table 3.: Ablation study on proposed components. Results are given in mDice.

BiFM	BPM	Polar	EndoScene	CVC-ClinicDB	CVC-ColonDB	ETIS	Kvasir
			0.866	0.902	0.774	0.726	0.905
✓			0.887	0.924	0.795	0.778	0.914
✓	✓		0.893	0.925	0.790	0.780	0.910
✓	✓	✓	0.897	0.930	0.797	0.793	0.914

Table 4.: Quantitative results with Res2Net backbone. Results shown in mDice.

Method	EndoScene	CVC-ClinicDB	CVC-ColonDB	ETIS	Kvasir
UACA	0.873	0.897	0.681	0.604	0.883
BaFPR-Res2Net (w/o ens.)	0.869	0.918	0.656	0.568	0.892
BaFPR-Res2Net	0.887	0.925	0.709	0.618	0.898

Table 5.: Quantitative results with different fusion blocks (w/o BPM or Polar consistency). Results given in mDice.

Fusion Block	Endoscene	CVC-ClinicDB	CVC-ColonDB	ETIS	Kvasir
Single-directional Fusion	0.894	0.911	0.788	0.767	0.911
Bi-directional Fusion	0.887	0.924	0.795	0.778	0.914

Table 2.: Stability analysis on different datasets.

(a) Stability analysis on Endoscene and CVC-ClinicDB datasets. Results shown in (mean ± std).

method	Endoscene		CVC-ClinicDB	
	mDice	mIoU	mDice	mIoU
Polyp-PVT	0.893±0.010	0.824±0.010	0.924±0.005	0.875±0.004
BaFPR	0.897±0.003	0.832±0.004	0.930±0.006	0.883±0.008

(b) Stability analysis on ETIS and CVC-ColonDB datasets. Results shown in (mean ± std).

method	ETIS		CVC-ColonDB	
	mDice	mIoU	mDice	mIoU
Polyp-PVT	0.779±0.008	0.695±0.006	0.783±0.006	0.704±0.006
BaFPR	0.793±0.009	0.720±0.010	0.797±0.002	0.720±0.002

(c) Stability analysis on Kvasir dataset. Results shown in (mean ± std).

method	Kvasir	
	mDice	mIoU
Polyp-PVT	0.911±0.002	0.860±0.003
BaFPR	0.914±0.004	0.862±0.003

3.3. Ablation Studies

We further conducted ablation studies on each component, and the results were shown in Table 3. The first row in Table 3 represent PVT (Wang et al. 2022b) with a simple convolution layer as its decoder. Table 3 indicated BiFM introduced a large improvement. BPM advanced the performance in EndoScene, ETIS, and CVC-ClinicDB dataset but dropped the performance in CVC-ColonDB and Kvasir dataset. Consistency learning from polar view and prediction ensemble advanced the mDice from 0.893, 0.925, 0.790, 0.780, and 0.910 to 0.897, 0.930, 0.797, 0.793, and 0.914 in EndoScene,CVC-ClinicDB, CVC-colonDb, ETIS, and Kvasir, respectively.

To validate the backbone generalization ability of our proposal, we replaced the PVT encoder with Res2Net (Gao et al. 2019) and compared our proposal with UACANet[3] (Kim et al. 2021) (built on the top of Res2Net encoder). Our proposal consistently surpassed UACA [4] in all the datasets. Especially for CVC-ClinicDB dataset, we observed an improvement of 3% over UACAnet.

For the fusion block, we conducted experiments with only single-directional multi-scale fusion block (with paths to $\mathbf{F}_{up}$ only in Fig. 3) or bi-directional multi-scale fusion block (Fig. 3) and the results were shown in Table 5. These results showed the advance of bi-directional multi-scale fusion.

3.4. Discussion on Polar Transformation Centers

In previous sections, we used predicted segmentation mask from main segmentation

[3]we aligned the training settings to 100 epochs training for better converge of Res2Net backbone.

[4]Note that the results for UACA here are lower than the results reported in the original paper, probably due to fewer training epochs (240) and data augmentations (random flipping, rotation, dilation and erosion were used). UACANet is sensitive to hyper-parameters as discussed in https://github.com/plemeri/UACANet/issues/2

network to compute the center of mass, which is slightly different with the center of mass of ground-truth due to prediction error. Considering that, we conducted further analysis of segmentation performance in test set for our trained polar segmentation networks by calculating origin by 1) center of mass of ground-truth mask $\mathbf{C}$; 2) center of mass of ground-truth mask with additional Gaussian noise $\mathcal{N}(\mathbf{C}, \alpha\mathbf{I})$; 3) center of mask of prediction mask. Results are shown in Table 6.

Table 6.: Mean Dice of pre-trained polar segmentation networks with different centers. GT center: calculate mass center from ground-truth mask; PD center: calculate mass center from prediction mask of main segmentation network. Note that GT centers are not available in practice for inference.

	EndoScene	CVC-ClinicDb	CVC-ColonDB	ETIS	Kvasir
GT Center	0.928	0.941	0.874	0.899	0.915
$\alpha = 1$	0.922	0.935	0.870	0.892	0.912
$\alpha = 3$	0.914	0.930	0.863	0.881	0.910
$\alpha = 5$	0.908	0.919	0.858	0.869	0.905
$\alpha = 7$	0.899	0.912	0.853	0.861	0.902
PD Center	0.893	0.913	0.800	0.785	0.872

From this table, it is not difficult to tell the center used in polar transformation has large impact on segmentation result. Basically, polar segmentation network improves as being closer to GT center.

3.5. Qualitative Results

Figure 5 showed the visualization of predictions. The first and the third rows showed our proposal could give more precise segmentation compared to UACANet, and introduce fewer false positives compared to Polyp-PVT. The second row showed that our proposal has better boundary segmentation than Poly-PVT. Figure 6 showed some examples that our proposal failed to segment. The first row is for an image in Kvasir containing multiple polyps. In such a case, BaFPR introduced more noise than without ensemble prediction. The second and the third rows were from Endoscene. We observed BPM helped to segment a rough region for polyps than with only BiFM. BaFPR (w/o ens.) further learned the clues from the polar view and reduced the false positives.

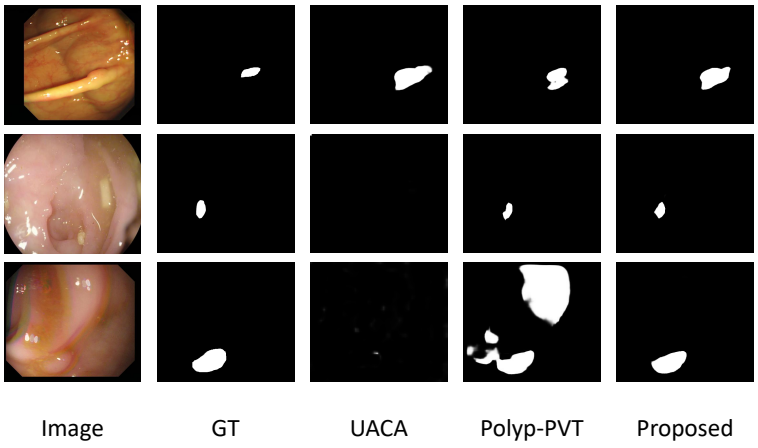

Figure 5.: Visualization of segmentation results. First, second, and third rows are from Endoscene, ETIS, and CVC-ColonDB datasets.

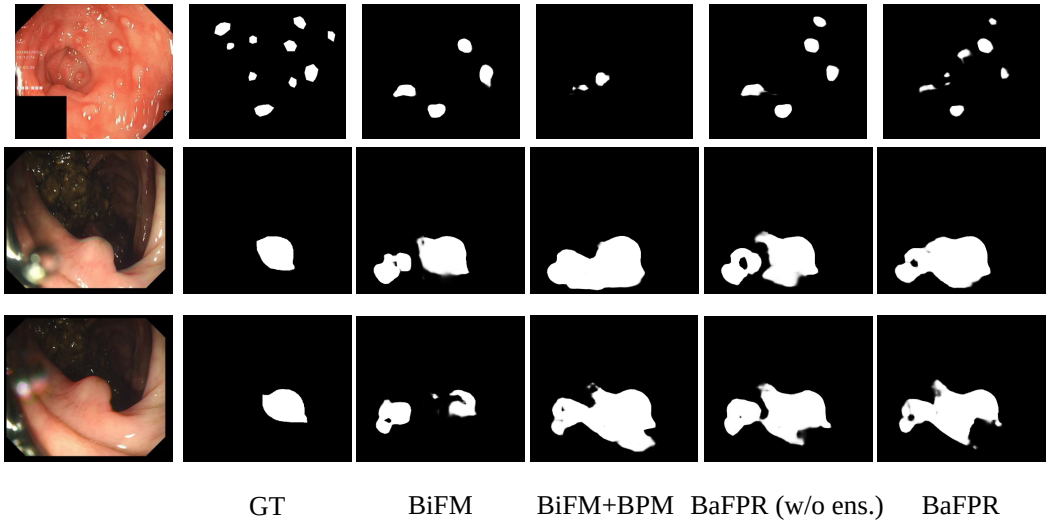

Figure 6.: Visualization of failure cases. Row 1 is from Kvasir dataset, and row 2, 3 are from CVC-ColonDB dataset.

4. Conclusion

In this paper, we design a decoder on the top of PVT by considering the boundary prior and multi-level bi-direction feature fusion. Besides, we propose a consistency loss to learn the clues from a polar view, where the polar view handles better on the round polyps. We conducted experiments in 5 polyp segmentation datasets, and the results showed the advance of our proposal. Further ablation studies showed the effectiveness of each component of our proposal.

However, we found the proposed boundary prior module improved the performance in Endoscene, CVC-ClinicDB, and ETIS datasets but decreased the performance in CVC-ColonDb and Kvasir datasets. Better fusion strategies for boundary prior and feature maps should be investigated. And the use of prediction ensemble increases the inference time. To handle this, more efficient backbone such as mobileNet (Howard et al. 2019), and knowledge distillation (Gou et al. 2021) to better learn the clues from other views should be explored. Besides, the predictions for successive frames are varying, which suggests we require a temporal smoothing strategy for our framework.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

Parts of this research were supported by the JST CREST Grant Number JPMJCR20D5, Japan and the MEXT/JSPS KAKENHI Grant Numbers 17H00867, 21K19898.

References

Benčević M, Galić I, Habijan M, Babin D. 2021. Training on polar image transformations improves biomedical image segmentation. IEEE Access. 9:133365–133375.

Bernal J, Sánchez FJ, Fernández-Esparrach G, Gil D, Rodríguez C, Vilariño F. 2015. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized Medical Imaging and Graphics. 43:99–111.

Bernal J, Sánchez J, Vilarino F. 2012. Towards automatic polyp detection with a polyp appearance model. Pattern Recognition. 45(9):3166–3182.

Chen L, Zhang H, Xiao J, Nie L, Shao J, Liu W, Chua TS. 2017. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. p. 5659–5667.

Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). p. 801–818.

Dong B, Wang W, Fan DP, Li J, Fu H, Shao L. 2021. Polyp-pvt: Polyp segmentation with pyramid vision transformers. arXiv preprint arXiv:210806932.

Fan DP, Cheng MM, Liu Y, Li T, Borji A. 2017. Structure-measure: A new way to evaluate foreground maps. In: Proceedings of the IEEE international conference on computer vision. p. 4548–4557.

Fan DP, Ji GP, Zhou T, Chen G, Fu H, Shen J, Shao L. 2020. PraNet: Parallel reverse attention network for polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, LNCS 12266. Springer. p. 263–273.

Fang Y, Chen C, Yuan Y, Tong Ky. 2019. Selective feature aggregation network with area-boundary constraints for polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, LNCS 11764. Springer. p. 302–310.

Gao SH, Cheng MM, Zhao K, Zhang XY, Yang MH, Torr P. 2019. Res2net: A new multi-scale backbone architecture. IEEE transactions on pattern analysis and machine intelligence. 43(2):652–662.

Gou J, Yu B, Maybank SJ, Tao D. 2021. Knowledge distillation: A survey. International Journal of Computer Vision. 129(6):1789–1819.
Howard A, Sandler M, Chu G, Chen LC, Chen B, Tan M, Wang W, Zhu Y, Pang R, Vasudevan V, et al. 2019. Searching for mobilenetv3. In: Proceedings of the IEEE/CVF international conference on computer vision. p. 1314–1324.
Hu J, Shen L, Sun G. 2018. Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. p. 7132–7141.
Jha D, Smedsrud PH, Riegler MA, Halvorsen P, Lange Td, Johansen D, Johansen HD. 2020. Kvasir-seg: A segmented polyp dataset. In: International Conference on Multimedia Modeling. Springer. p. 451–462.
Jha D, Smedsrud PH, Riegler MA, Johansen D, De Lange T, Halvorsen P, Johansen HD. 2019. ResUNet++: An advanced architecture for medical image segmentation. In: 2019 IEEE International Symposium on Multimedia (ISM). IEEE. p. 225–2255.
Kervadec H, Bouchtiba J, Desrosiers C, Granger E, Dolz J, Ayed IB. 2019. Boundary loss for highly unbalanced segmentation. In: International conference on medical imaging with deep learning. PMLR. p. 285–296.
Kim T, Lee H, Kim D. 2021. Uacanet: Uncertainty augmented context attention for polyp segmentation. In: Proceedings of the 29th ACM International Conference on Multimedia. p. 2167–2175.
Lee HJ, Kim JU, Lee S, Kim HG, Ro YM. 2020. Structure boundary preserving segmentation for medical image with ambiguous boundary. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. p. 4817–4826.
Loshchilov I, Hutter F. 2018. Decoupled weight decay regularization. In: International Conference on Learning Representations.
Margolin R, Zelnik-Manor L, Tal A. 2014. How to evaluate foreground maps? In: Proceedings of the IEEE conference on computer vision and pattern recognition. p. 248–255.
Ponugoti PL, Cummings OW, Rex DK. 2017. Risk of cancer in small and diminutive colorectal polyps. Digestive and Liver Disease. 49(1):34–37.
Ronneberger O, Fischer P, Brox T. 2015. U-Net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, LNCS 9351. Springer. p. 234–241.
Roy AG, Navab N, Wachinger C. 2018. Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks. In: International conference on medical image computing and computer-assisted intervention. Springer. p. 421–429.
Silva J, Histace A, Romain O, Dray X, Granado B. 2014. 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.
Vázquez D, Bernal J, Sánchez FJ, Fernández-Esparrach G, López AM, Romero A, Drozdzal M, Courville A. 2017. A benchmark for endoluminal scene segmentation of colonoscopy images. Journal of Healthcare Engineering. 2017.
Wang J, Wei L, Wang L, Zhou Q, Zhu L, Qin J. 2021. Boundary-aware transformers for skin lesion segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, LNCS 12901. Springer. p. 206–216.
Wang W, Xie E, Li X, Fan DP, Song K, Liang D, Lu T, Luo P, Shao L. 2022a. Pvt v2: Improved baselines with pyramid vision transformer. Computational Visual Media. 8(3):415–424.
Wang W, Xie E, Li X, Fan DP, Song K, Liang D, Lu T, Luo P, Shao L. 2022b. Pvtv2: Improved baselines with pyramid vision transformer. Computational Visual Media. 8(3):1–10.
Wang Y, Wei X, Liu F, Chen J, Zhou Y, Shen W, Fishman EK, Yuille AL. 2020. Deep distance transform for tubular structure segmentation in ct scans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. p. 3833–3842.
Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. 2004. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing. 13(4):600–612.
Wei J, Hu Y, Zhang R, Li Z, Zhou SK, Cui S. 2021. Shallow attention network for polyp segmentation. In: International Conference on Medical Image Computing and Computer-

Assisted Intervention, LNCS 12901. Springer. p. 699–708.
Xu X, Sanford T, Turkbey B, Xu S, Wood BJ, Yan P. 2022. Polar transform network for prostate ultrasound segmentation with uncertainty estimation. Medical Image Analysis. 78:102418.
Zhang H, Zhu H, Ling X. 2020. Polar coordinate sampling-based segmentation of overlapping cervical cells using attention u-net and random walk. Neurocomputing. 383:212–223.
Zhang Y, Liu H, Hu Q. 2021. Transfuse: Fusing transformers and cnns for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, LNCS 12901. Springer. p. 14–24.
Zhang Z, Liu Q, Wang Y. 2018. Road extraction by deep residual U-Net. IEEE Geoscience and Remote Sensing Letters. 15(5):749–753.
Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J. 2018. UNet++: A nested U-Net architecture for medical image segmentation. In: Deep learning in medical image analysis and multimodal learning for clinical decision support, lncs 11045. Springer; p. 3–11.