Sub-Band Based Attention for Robust Polyp Segmentation

Xianyong Fang, **Yuqing Shi**, **Qingqing Guo**, **Linbo Wang*** and **Zhengyi Liu**
School of Computer Science and Technology, Anhui University
fangxianyong@ahu.edu.cn, e21201044@stu.ahu.edu.cn, guoqingad@sina.com, {wanglb, liuzywen}@ahu.edu.cn

Abstract

This article proposes a novel spectral domain based solution to the challenging polyp segmentation. The main contribution is based on an interesting finding of the significant existence of the middle frequency sub-band during the CNN process. Consequently, a Sub-Band based Attention (SBA) module is proposed, which uniformly adopts either the high or middle sub-bands of the encoder features to boost the decoder features and thus concretely improve the feature discrimination. A strong encoder supplying informative sub-bands is also very important, while we highly value the local-and-global information enriched CNN features. Therefore, a Transformer Attended Convolution (TAC) module as the main encoder block is introduced. It takes the Transformer features to boost the CNN features with stronger long-range object contexts. The combination of SBA and TAC leads to a novel polyp segmentation framework, SBA-Net. It adopts TAC to effectively obtain encoded features which also input to SBA, so that efficient sub-bands based attention maps can be generated for progressively decoding the bottleneck features. Consequently, SBA-Net can achieve the robust polyp segmentation, as the experimental results demonstrate.

1 Introduction

Polyp segmentation of colonoscopy images is very important for the treatment of colorectal cancer which is one of the severe diseases of the world [Sawicki *et al.*, 2021]. However, polyps usually look like the surrounding tissues and also have various shapes and sometimes blurry boundaries. It is very difficult to recognize them, even though there have been impressive results in the literature [Tomar *et al.*, 2022a; Tomar *et al.*, 2022b; Guo *et al.*, 2022; Guo *et al.*, 2023; Li *et al.*, 2022].

One recent trend attracts us is the spectral domain feature boosting methods [Huang *et al.*, 2022; Suvorov *et al.*, 2022]. Based on the global spectrum analysis, they supplement important frequency distributions to spatial signals for high performances. Existing ideas often take the complete frequency band [Huang *et al.*, 2022; Liu *et al.*, 2021; He *et al.*, 2022], high or low frequencies [Zhou *et al.*, 2022; Liu *et al.*, 2022]), or special frequency channels [Xu *et al.*, 2020; Huang *et al.*, 2021] for further feature engineering.

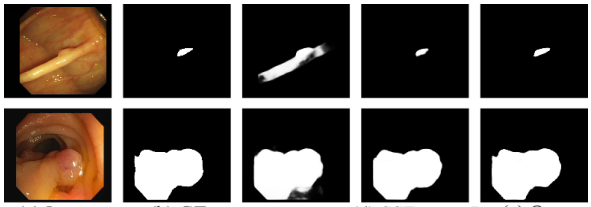

(a) Image (b) GT (c) TGANet (d) SSFormer-L (e) Ours

Figure 1: Example of polyp segmentation results from two state-of-the-arts methods, TGANet [Tomar *et al.*, 2022a] and SSFormer-L [Wang *et al.*, 2022a], and ours (SBA-Net). Note: These two polyps have quite different sizes and shapes and similar appearances to the background.

However, the polyp images call for a new spectral domain based method. They are often low contrasted with low resolutions and almost flatten textures that the polyps are insignificant from their surroundings (Figure 1). These appearances make them lack of enough high-frequency local details and significant low-frequency global polyp contexts.

Even worse, our experiment (Figure 2) show that the limited high and low frequencies will be further reduced after recursive CNN convolutions and, therefore, those frequencies may mess up the polyps and background tissues when doing final decoding. But, interestingly, the magnitude ratio of middle sub-band frequencies between the high and the low ones to the whole spectrum gets higher and higher after passing the convolution stages. As global data, these strong sub-band frequencies contain the rich global information on polyps. Therefore, this finding let us think of the middle sub-band for effective feature boosting.

Furthermore, the high frequencies of the features from the first network stage include salient object details and thus they can also be taken as a sub-band to boost features. Consequently, a unified module called the Sub-Band based Attention (SBA) module is proposed, so that the middle sub-band and the initial high one from encoding features can all be utilized to enhance the features.

*Corresponding author

Apparently, the encoder supplying features to SBA for sub-bands is very important for robust recognition [Fang *et al.*, 2021]. One of the popular ways is to combine CNN and Transformer [Shamshad *et al.*, 2022] together, as either a serial branch [Chen *et al.*, 2021; Ye *et al.*, 2022] or two parallel branches [Li *et al.*, 2022; Lin *et al.*, 2022].

We highly value the CNN based encoder because it can discover rich local and global information through layer-by-layer convolution. But Transformer has been demonstrated to be very good at extracting the long-range information or global object contexts by the self-attention mechanism. Apparently, if taking Transformer to attend the CNN features, the global object contexts can be augmented while the local details can still be kept for efficient recognition. Therefore, we argue that Transformer can act as an attention tool to enhance CNN features for better polyp recognition.

Consequently, a new SBA based polyp segmentation framework SBA-Net is proposed, where a novel Transformed attended CNN encoder is proposed to obtain strong features for the SBA based decoder. A new module called Transformer Attended Convolution (TAC) module is introduced as the main building block of the encoder to boost the CNN features with Transformer ones. The features from this encoder are input to SBAs for extracting their concrete sub-bands to attend the progressive decoding of the bottleneck features. Experiments including the comparisons with the state-of-the-arts methods on public colonoscopy image datasets show the effectiveness of the proposed method.

In summary, the contributions of our work are as follows:

- A spectral based feature augmentation module SBA which builds on the finding on the significant existence of the middle frequency sub-bands and uniformly treats the high and middle sub-bands from encoder features as attention maps to boost feature discrimination.
- A Transformer attended CNN module TAC which takes the Transformer features as attention map to boost the CNN features and, therefore, enhances the rich CNN features with stronger long-range object contexts.
- A polyp image segmentation method SBA-Net which takes TAC as the main block to supply boosted CNN features and empower SBA with concrete sub-bands for progressive decoding and final robust segmentation.

2 Related Work

2.1 Spectral Domain Based Deep Methods

Spectral domain based deep methods have been proved to be effective [Frank *et al.*, 2020; Xu *et al.*, 2020]. For example, Xu *et al.* [Xu *et al.*, 2020] showed the high accuracies from learning in the frequency domain with static channel selection. More specific methods with frequency based ideas are also proposed, either by Fourier transform [Huang *et al.*, 2022; Liu *et al.*, 2021; Xu *et al.*, 2020], discrete consine transform (DCT) [Magid *et al.*, 2021; Huang *et al.*, 2021; Frank *et al.*, 2020] or discrete wavelet transform (DWT) [Ji *et al.*, 2021; He *et al.*, 2022].

Some methods [Huang *et al.*, 2022; Frank *et al.*, 2020; Ji *et al.*, 2021; Liu *et al.*, 2021; He *et al.*, 2022; Huang *et al.*, 2021] take the complete frequency band as the source for further processing. For example, FECNet [Huang *et al.*, 2022] takes a Spatial-Frequency Interaction (SFI) block for the amplitude sub-network and the phase sub-network respectively as complementary learning; Ji *et al.* [Ji *et al.*, 2021] targeted at the frequency inconsistency of super-resolution and proposed the frequency density comparator and the wavelet discriminator as guidance for consistency.

Some researchers recognized the difference between different frequencies and built their methods based on the high and low frequencies [Yin *et al.*, 2019; Magid *et al.*, 2021; Zhou *et al.*, 2022; Liu *et al.*, 2022]. For example, Magid *et al.* [Magid *et al.*, 2021] noticed that super-resolution methods often bias toward low-frequency signals and thus proposed a dynamic high-pass filtering (HPF) module to preserve high-frequency signals. An additional matrix multi-spectral channel attention (MMCA) module is also introduced to predict the attention maps in frequency domain.

More refined frequency division based methods are also proposed [Xu *et al.*, 2020; Huang *et al.*, 2021]. Xu *et al.* [Xu *et al.*, 2020] introduced a learning based frequency channel selection strategy which can achieve higher accuracy than spatial methods. Their argument is that some frequency channels are less informative than others for a specific task. Frequency Space Domain Randomization (FSDR) [Huang *et al.*, 2021] further separates frequency components into domain variant and invariant ones for more generalized models.

As far as we know, few researchers apply spectral domain based methods to medical image segmentation [Zhou *et al.*, 2022; Liu *et al.*, 2021]. Zhou *et al.* [Zhou *et al.*, 2022] introduced a random amplitude mixup (RAM) module incorporating low-level frequency from different source images to synthesize new images as data augmentation. FedDG [Liu *et al.*, 2021] uses federated learning and incorporates a continuous frequency space interpolation mechanism to transmit the distribution information across clients.

2.2 Deep Learning Based Polyp Segmentation

Early deep polyp segmentation methods [Brandao *et al.*, 2017] use fully convolutional networks (FCN), while, later on, U-Net [Ronneberger *et al.*, 2015] based methods become popular, such as PolypSeg [Zhong *et al.*, 2020] and SCR-Net [Wu *et al.*, 2021]. Multiple parallel branches are often adopted for robust features, either from the decoder [Tomar *et al.*, 2021] or intermediate stages [Yin *et al.*, 2022]. Boundaries and contours are often adopted as constraints explicitly [Du *et al.*, 2022] or implicitly [Nguyen *et al.*, 2021].

Recently, Transformer based ideas turn popular [Shamshad *et al.*, 2022]. Some take pure Transformers for feature abstraction [Wang *et al.*, 2022a; Dong *et al.*, 2021] or generation [Li *et al.*, 2021]. Many studies combine both Transformer and CNN, either serially by taking Transformer as an intermediate layer for better performance [Ye *et al.*, 2022; Chen *et al.*, 2021], or in parallel before fusing them together [Li *et al.*, 2022; Zhang *et al.*, 2021; Lin *et al.*, 2022]. For example, TransFuse [Zhang *et al.*, 2021] uses both in parallel to capture global dependencies and low-level details.

Various feature boosting methods have been proposed [Kim *et al.*, 2021; Nguyen *et al.*, 2021; Tomar *et al.*,

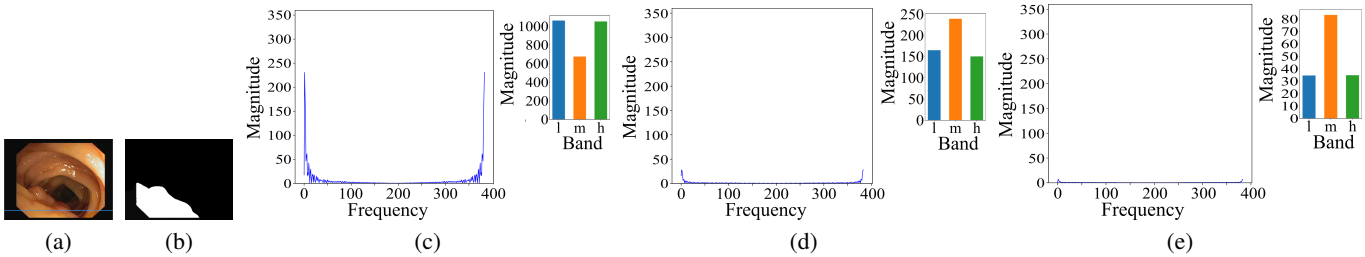


Figure 2: Demonstration of frequency spectra and informative sub-bands in the CNN stages. The row shown in blue crossing the foreground polyp of the image in (a) is extracted for this demo, where the typical double-3×3-convolution based CNN stage is applied twice. (a): Image; (b) GT; and (c), (d) and (e) are the frequency distributions and their corresponding histograms on their top right before CNN convolutions, after one CNN stage, and after two CNN stages, respectively. Note that the frequency histograms of (c), (d) or (e) are generated according to three adjacent frequency sub-bands, *i.e.*, low (l), middle (m) and high (h) sub-bands.

2022a; Tomar *et al.*, 2022b]. For example, the PAA module [Kim *et al.*, 2021] takes several axial attentions for both horizontal and vertical axes so that global dependencies and local representation can be obtained; TGANet [Tomar *et al.*, 2022a] adopts text guided attention to learn different features according to the polyp number and size.

There are also other approaches except the general encoder-decoder style, such as Generative Adversarial Networks (GAN) based methods [Ahmed and Ali, 2020] and Mask R-CNN based methods [Kang and Gwak, 2019]. Nanni *et al.* [Nanni *et al.*, 2021] leveraged on differences of various classifiers by an ensemble of CNNs with different methods.

3 Sub-Band Based Attention

3.1 Why Sub-Band?

Our common view is that the high frequencies include the local object details while the low ones depict the global object contexts. However, the high frequencies will gradually decrease to be less significant during the CNN stages; and (2) the low frequencies are not so important for polyps because those from the polyps and their surroundings are difficult to distinguish due to their similar looks.

Experiment on the spectrum (Figure 2) further shows that the high and low frequencies all become less and less significant after several CNN stages. Here, the frequency distributions along the 250th row in blue across the foreground polyp of the 384×288 image (Figure 2a) are collected. The spectra are obtained from the typical CNN stages for the image [Ronneberger *et al.*, 2015], with each stage consisting of two 3×3 convolutions. This experiment shows that more informative frequency band different from the high-frequency or low-frequency band is expected.

Further checking these spectra show that the middle frequencies between the high and the low ones seem almost unchanged. It means their overall contributions to the whole signal band increase during the CNN stages. This inspires us to further check their collective distributions through histograms (Figure 2), where three bins according to the low (frequencies between 0 and 30), middle (frequencies between 31 and 353) and high sub-bands (frequencies between 354 and 384) are computed for each spectrum.

We can see that the middle frequency band always goes up after the CNN stages and reaches 0.546 proportionately in the histogram of Figure 2e after two CNN stages, even though it is only 0.242 proportionately in the histogram of Figure 2c. This interesting finding justifies our intuition that the middle sub-band turns more and more important for subsequent CNN stages.

Considering this middle sub-band globally depicts the scene, we believe the rich information from this increasingly-important sub-band should be adopted to boost the features for robust recognition. Note that the detail preserved high frequencies are also very useful for object discovery and can also be considered as a sub-band. Therefore, we unify the effective sub-bands from the middle and high-frequency ones and propose the sub-band base attention method, SBA, for utilizing both the high and middle frequencies.

3.2 The Structure of SBA

Figure 3 shows the structure of SBA which mainly relies on two Gaussian filters to obtain the sub-band frequencies.

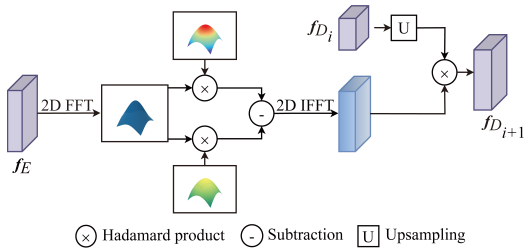


Figure 3: The structure of SBA. Two Gaussian filters are applied to FFT converted features of the encoder ones $\boldsymbol{f}_E$ to obtain the sub-band via subtraction. This sub-band is then converted back to spatial features which act as attention map to boost the target features $\boldsymbol{f}_{D_i}$.

SBA works as follows. The input encoder features $\boldsymbol{f}_E$ is first transformed by fast Fourier transform (FFT) and then processed by two Gaussian filters to obtain two frequency spectra with different high-frequency bounds. Then these spectra subtract each other to obtain the sub-band which is restored to spatial features by the inverse fast Fourier transform (IFFT). These spatial features then act as attention map

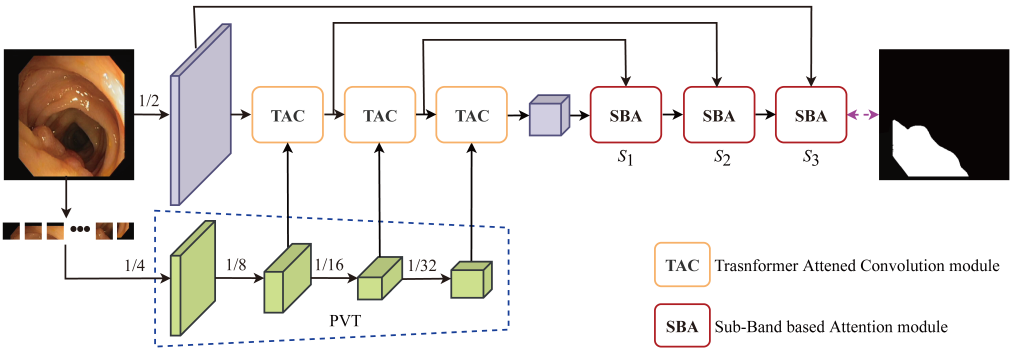


Figure 4: The pipeline of SBA-Net. It takes three TACs as encoder block to boost CNN features with PVT defined Transformer features and uses three SBAs to gradually decode bottleneck features attended by the sub-bands of the TAC boosted features. First two SBA S_1 and S_2 take the middle sub-bands of corresponding encoding features with the last SBA S_3 using the high sub-band of the coarsest encoding features. Note the numbers denote the zooming ratios of the original image size.

to augment the upsampled features of target features $\boldsymbol{f}_{D_i}$ and finally obtain the sub-band attended ones $\boldsymbol{f}_{D_{i+1}}$. The process can be formulated as:

$$\boldsymbol{f}_{D_{i+1}} = \mathrm{I}((K_u - K_d) \otimes \mathrm{F}(\boldsymbol{f}_E)) \otimes \mathrm{U}(\boldsymbol{f}_{D_i}) \qquad (1)$$

where: F, I and U represent the FFT, IFFT and upsampling operations respectively; K_u and K_d are the two Gaussian filters; and $\otimes$ denotes the Hadamard product. Experimentally, the mean values of the two Gaussian Filters are all set to the mean width and height of the image, while their standard deviations are set to 7 Hertz and 10 Hertz, respectively.

SBA generally computes the middle sub-band. For the high-frequency sub-band, the real highest frequencies are unstable according to our experiment, perhaps due to the affection of noise. Therefore, in practice, this sub-band can be obtained by SBA with only one Gaussian filter having the higher up-frequency bound, i.e., K_u. More details on the possible ways to compute the high sub-band in SBA can be found in the supplementary document.

4 SBA-Net: Application of SBA to Polyp Image Segmentation

Intuitively, SBA can be directly deployed into existing methods by replacing their decoders or individual decoding blocks with SBA. Experiments in Section 6.4 demonstrate the possibility of such a design.

However, we want a more efficient framework with SBA. Specifically, we need a powerful encoder, considering that the robust encoder features can provide SBA with concrete sub-bands for augmentation. We are interested in the locally-and-globally rich CNN features but also impressed with Transformer's strong long-range discovery ability. Consequently, a Transformer boosted CNN encoder is proposed, which mainly consists of three TAC modules using PVT [Wang *et al.*, 2022b] defined lower-stage Transformers as attention maps to progressively augment the CNN features. Combining TAC and SBA leads to the new framework SBA-Net (Figure 4) whose decoder consists of three SBA modules for progressively decoding the bottleneck features under the enhancement of the sub-bands from the encoder features.

Note that the first stage of the encoder is a convolution layer because it is already rich with the high-frequency object details. This layer is obtained directly from $1/2$ downsampled input images by two 3×3 convolutions and max pooling. Also note that the Transformer and CNN features input to each TAC are the same sizes as the correspondingly double downsampled PVT features (Figure 4).

Now let's discuss the new Transformer attended CNN encoder, especially its core block TAC.

4.1 Transformer Attended CNN Encoder

Transformer and CNN are typically combined in serial [Ye *et al.*, 2022; Chen *et al.*, 2021] (Figure 5(b)) or parallel [Li *et al.*, 2022; Zhang *et al.*, 2021; Lin *et al.*, 2022] (Figure 5(c)). For the serial encoder, CNN and Transformer blocks are subsequently applied to obtain the encoded features, while, for the parallel encoder, both CNN and Transformer are independently applied before fuse their features. Both types of methods demonstrate impressive results by the fused features.

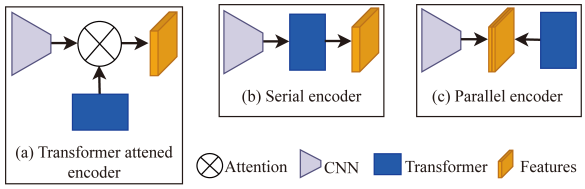


Figure 5: Different types of CNN and Transformer combined encoders.

However, CNN features contain rich global and local information which are all important for the global contexts and local details of polyps, while Transformer features focus more on the long-range information. Therefore, CNN can be taken as the main informative contributor and we opt to consider Transformer as supplementary of the long-range distribution to CNN. Consequently, the Transformer features can be taken as attention maps to enhance the CNN features with more global object contexts while maximally keeping the existing information of CNN features (Figure 5(a)). Then, a Transformer attended CNN encoder can be obtained.

As shown in Figure 4, the core part to fulfill this decoder is TAC (Figure 6) which uses the Transformer features as the attention maps to augment the convolution features. Here, the input features $\boldsymbol{f}_C$ is convolved by 3×3 ($\mathrm{C}_{3\times3}$) two times and then max pooled (M) to obtain the convolution features, which are then attended by the corresponding Transformer features $\boldsymbol{f}_T$ from PVT to obtain augmented features $\boldsymbol{f}'_E$,

$$\boldsymbol{f}'_E = \mathrm{M}(\mathrm{C}_{3\times3}(\mathrm{C}_{3\times3}(\boldsymbol{f}_C))) \otimes \boldsymbol{f}_T \tag{2}$$

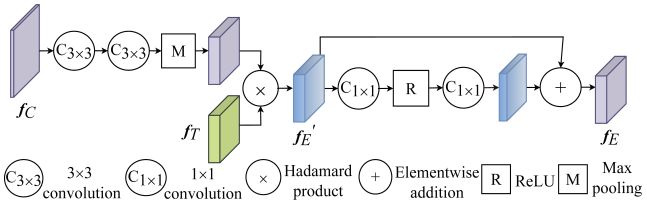


Figure 6: The structure of TAC. The input convoluted features $\boldsymbol{f}_C$ are first convoluted again, then augmented by Transformer features $\boldsymbol{f}_T$ and further non-linearly processed as the output features $\boldsymbol{f}_E$.

Then, $\boldsymbol{f}'_E$ are non-linearly enhanced by 1×1 convolutions ($\mathrm{C}_{1\times1}$), ReLU activation (R) and residual connection to obtain the final boosted convolution features $\boldsymbol{f}_E$ which are also the input features $\boldsymbol{f}_C$ of the next TAC in SBA-Net. Formally, $\boldsymbol{f}_E$ is computed as

$$\boldsymbol{f}_E = \boldsymbol{f}'_E + \mathrm{C}_{1\times1}(\mathrm{R}(\mathrm{C}_{1\times1}(\boldsymbol{f}'_E))) \tag{3}$$

With TAC, SBA-Net can obtain strong encoder features for SBA to do further boosting and discriminative recognition.

5 Losses

Three losses are considered: The binary cross entropy loss (BCE) [Murphy, 2012], Intersection over Union loss (IoU) [Rahman and Wang, 2016] and inverse-transformation loss (IF) [Borse *et al.*, 2021]. BCE is most widely used based on pixel-level constraints and IoU optimizes the global structure rather than individual pixels. We take their weighted forms [Dong *et al.*, 2021] to estimate. IF estimates the boundary loss with an inverse-transformation network.

Assuming the prediction and ground truth are P and $\hat{P}$ respectively, the complete loss function can be formulated as

$$\mathcal{L} = \mathcal{L}^w_{BCE}(P, \hat{P}) + \mathcal{L}^w_{IoU}(P, \hat{P}) + \gamma\mathcal{L}_{IF}(P, \hat{P}), \tag{4}$$

where: $\mathcal{L}^w_{BCE}$, $\mathcal{L}^w_{IoU}$ and $\mathcal{L}_{IF}$ are the weighted BCE and IoU losses, and the IF loss, respectively; and γ is the weight which is set to 0.5 experimentally.

6 Experiments

This section reports some experimental results. For the complete codes and more experimental settings, results and ablation studies, please check the supplementary document.

SBA-Net is implemented in PyTorch with the CUDA library, a GeForce RTX 3090 Ti GPU and an Intel Core i7-12700KF Processor. Adam optimizer is adopted with the learning rate 5e-5. Batch size is set to 8 with the epoch 80.

Eight methods with open codes are taken for performance comparison, including U-Net [Ronneberger *et al.*, 2015], PraNet [Fan *et al.*, 2020], TransUNet [Chen *et al.*, 2021], Polyp-PVT [Dong *et al.*, 2021], FedDG [Liu *et al.*, 2021], MKDCNet [Tomar *et al.*, 2022b], TGANet [Tomar *et al.*, 2022a] and SSFormer [Wang *et al.*, 2022a]. Especially, FedDG is the spectral augmentation method. All the experimental results are taken directly from their provided ones, except for FedDG, TransUNet, MKDCNet and TGANet whose original experimental datasets are different from our polyp datasets and, therefore, re-trained by us for polyp segmentation according to their directions.

6.1 Datasets and Data Augmentation

Five colonoscopy image datasets are adopted, including ETIS [Silva *et al.*, 2014], CVC-ClinicDB [Bernal *et al.*, 2015], CVC-ColonDB [Tajbakhsh *et al.*, 2015], CVC-300 [Vázquez *et al.*, 2017] and Kvasir [Jha *et al.*, 2020].

The same training and testing data as PraNet [Fan *et al.*, 2020] are adopted for fair comparison. The training set contains 1450 images selected from Kvasir [Jha *et al.*, 2020] and CVC-ClinicDB [Bernal *et al.*, 2015] with all the left images taken as testing image. The sizes of training images are set to 352×352 and normalized with means $\{0.485, 0.456, 0.406\}$ and variances $\{0.229, 0.224, 0.225\}$. Images are augmented with three scales $\{0.75, 1, 1.25\}$ and color exchange.

6.2 Evaluation Metrics

Several popular metrics used by previous studies [Fan *et al.*, 2020; Dong *et al.*, 2021] are adopted, including the Dice Similarity Coefficient (DSC), Intersection over Union (IoU), Weighted F-measure (F^w_β), S-measure (S_α), E-measure (E_ξ) and Mean Absolute Error (MAE).

DSC and IoU are similarity measures at the regional level, focusing on the internal object consistencies. F^w_β comprehensively considers the recall and precision with all pixels and thus is affected less by the individual pixels. S_α measures the structural similarities, while E_ξ evaluates the segmentation results at both image and pixel levels. MAE is a pixel-by-pixel comparison index, denoting the average absolute error between the predicted value and the true value.

This paper additionally takes mDSC and mIoU to separately represent the means of DSCs and IoUs of all test images, and the max value of E-measure, $maxE_\xi$, to represent the segmentation similarity. All measures tell better models with bigger values, except MAE: The lower MAE, the better.

6.3 Qualitative Results

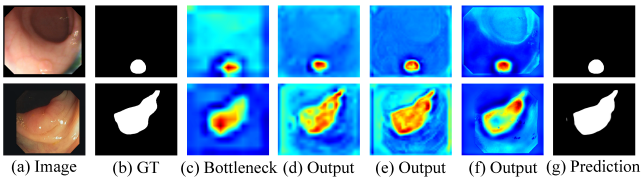


Figure 7: The effectiveness of SBA. Bottleneck features gradually boosted by different SBAs in SBA-Net (Figure 4).

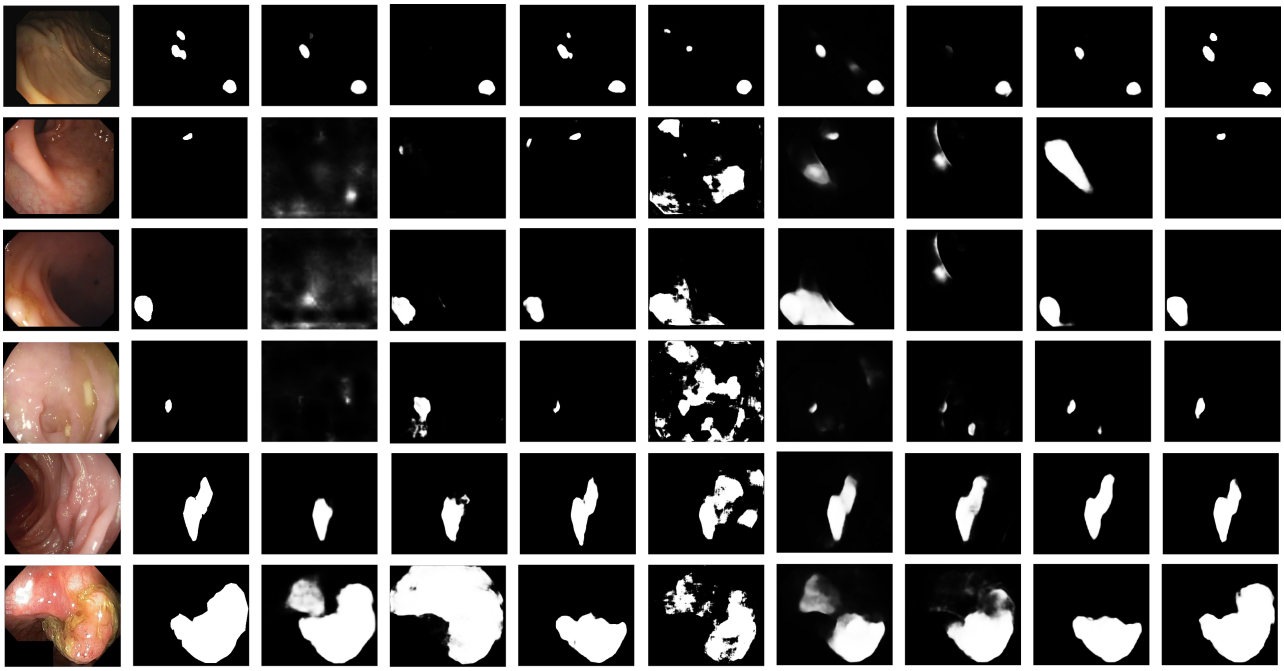


Figure 8: Qualitative comparison for the challenging images with very blurry polyp boundaries and low contrasts among PraNet [Fan *et al.*, 2020], TransUNet [Chen *et al.*, 2021], Polyp-PVT [Dong *et al.*, 2021], FedDG [Liu *et al.*, 2021], MKDCNet [Tomar *et al.*, 2022b], TGANet [Tomar *et al.*, 2022a], SSFormer-L [Wang *et al.*, 2022a] and our method SBA-Net.

Figure 7 shows the feature boosting performances of SBAs in SBA-Net. The features are gradually augmented with the polyps more and more salient after passing different decoding stages, thanks to the discriminative booster SBA and its robust feature supplier TAC.

Figure 8 shows experimental results with challenging polyp images of very blurry polyp boundaries and low contrasts, which demonstrate the efficiency of SBA-Net.

The statistical comparisons on learning ability are undertaken first on CVC-ClinicDB and Kvasir (Table 1). SBA-Net is generally the best among all methods, with eight times being the best and the left four times being the second best.

The generalization abilities are experimented with unseen datasets, CVC-ColonDB, ETIS and CVC-300 (Table 2). Again, SBA-Net is generally the best.

DSCs under different thresholds on four datasets are also collected (Figure 9), which shows that SBA-Net consistently outperforms other models.

6.4 Ablation Study

The effectivenesses of the SBA, TAC and proposed full SBA-Net are evaluated on CVC-ColonDB, ETIS and CVC-300 (Table 3). Six different configurations are included for this evaluation:

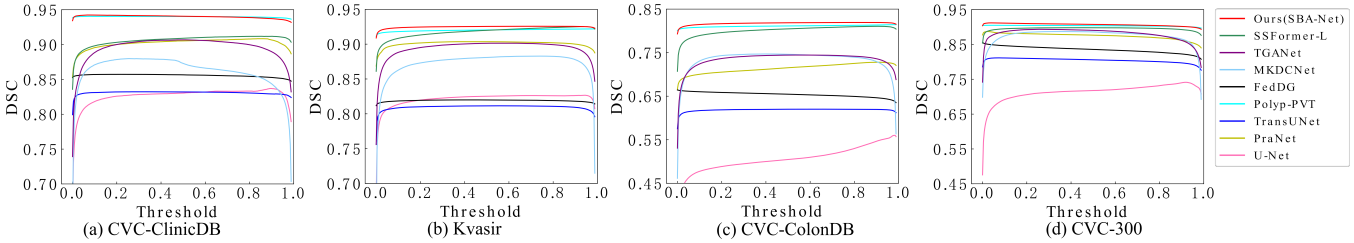
- CNN: It takes the typical CNN stage as the encoder with max pooling to downsample features and directly decodes the bottleneck features.
- Transformer: It takes the Transformer, *i. e.*, PVT, as the encoder and directly decodes as CNN.
- CNN+SBA: It replaces the decoder of CNN with the SBA based decoder in SBA-Net..
- Transformer+SBA: It replaces the decoder of TRANS with the SBA based decoder in SBA-Net.
- CNN+Transformer+TAC: It takes TAC as the main block of the encoder, *i. e.*, the same encoder as SBA-Net and directly decodes as CNN.

Method	CVC-ClinicDB						Kvasir					
	mDSC	mIoU	F_β^w	S_α	$maxE_\xi$	MAE	mDSC	mIoU	F_β^w	S_α	$maxE_\xi$	MAE
U-Net [Ronneberger *et al.*, 2015]	0.879	0.818	0.876	0.917	0.964	0.017	0.811	0.726	0.780	0.848	0.897	0.051
PraNet [Fan *et al.*, 2020]	0.899	0.849	0.896	0.936	0.979	0.009	0.898	0.84	0.885	0.915	0.948	0.030
TransUNet [Chen *et al.*, 2021]	0.828	0.771	0.822	0.887	0.941	0.023	0.807	0.725	0.756	0.880	0.892	0.053
Polyp-PVT [Dong *et al.*, 2021]	0.937	0.889	**0.936**	0.949	0.989	0.006	0.917	0.864	0.911	0.925	0.962	0.023
FedDG [Liu *et al.*, 2021]	0.853	0.791	0.844	0.892	0.928	0.026	0.817	0.731	0.792	0.849	0.899	0.055
MKDCNet [Tomar *et al.*, 2022b]	0.872	0.813	0.861	0.911	0.951	0.018	0.877	0.817	0.869	0.898	0.932	0.039
TGANet [Tomar *et al.*, 2022a]	0.901	0.850	0.899	0.932	0.968	0.010	0.897	0.840	0.887	0.910	0.956	0.027
SSFormer-L [Wang *et al.*, 2022a]	0.906	0.855	0.913	0.929	0.970	0.008	0.917	0.864	**0.916**	0.922	**0.964**	**0.022**
SBA-Net	**0.937**	**0.889**	0.934	**0.951**	**0.990**	**0.006**	**0.922**	**0.871**	0.913	**0.928**	0.963	0.023

Table 1: Statistical comparisons of different methods for CVC-ClinicDB and Kvasir. The best results are shown in bold.

Method	CVC-ColonDB						ETIS						CVC-300					
	mDSC	mIoU	F_β^w	S_α	$maxE_\xi$	MAE	mDSC	mIoU	F_β^w	S_α	$maxE_\xi$	MAE	mDSC	mIoU	F_β^w	S_α	$maxE_\xi$	MAE
U-Net [Ronneberger *et al.*, 2015]	0.584	0.493	0.559	0.740	0.807	0.052	0.395	0.320	0.357	0.662	0.714	0.034	0.743	0.648	0.708	0.840	0.902	0.015
PraNet [Fan *et al.*, 2020]	0.712	0.640	0.699	0.820	0.872	0.043	0.628	0.567	0.600	0.794	0.841	0.031	0.871	0.797	0.843	0.925	0.972	0.010
TransUNet [Chen *et al.*, 2021]	0.617	0.529	0.583	0.760	0.828	0.052	0.416	0.348	0.373	0.659	0.688	0.073	0.801	0.704	0.756	0.880	0.931	0.016
Polyp-PVT [Dong *et al.*, 2021]	0.808	0.727	0.795	0.865	0.919	0.031	0.787	0.706	0.750	0.871	0.910	**0.013**	0.900	0.833	0.884	0.935	0.981	0.007
FedDG [Liu *et al.*, 2021]	0.652	0.550	0.609	0.755	0.813	0.069	0.407	0.338	0.358	0.618	0.645	0.114	0.832	0.737	0.787	0.892	0.953	0.014
MKDCNet [Tomar *et al.*, 2022b]	0.741	0.658	0.721	0.829	0.870	0.038	0.667	0.582	0.615	0.797	0.831	0.030	0.880	0.810	0.865	0.928	0.971	0.008
TGANet [Tomar *et al.*, 2022a]	0.739	0.661	0.724	0.829	0.868	0.035	0.663	0.586	0.633	0.800	0.865	0.017	0.885	0.819	0.869	0.931	0.959	0.007
SSFormer-L [Wang *et al.*, 2022a]	0.802	0.721	0.798	0.860	0.909	0.031	**0.796**	**0.720**	**0.771**	0.873	0.912	0.014	0.895	0.827	0.881	0.933	0.976	0.007
SBA-Net	**0.815**	**0.737**	**0.799**	**0.871**	**0.919**	**0.029**	0.790	0.712	0.748	**0.878**	**0.912**	0.016	**0.903**	**0.840**	**0.887**	**0.941**	**0.981**	**0.007**

Table 2: Statistical comparisons of different methods for CVC-ColonDB, ETIS and CVC-300. The best results are shown in bold.



(a) CVC-ClinicDB (b) Kvasir (c) CVC-ColonDB (d) CVC-300

Figure 9: Comparison of the DSC curves under different thresholds for different datasets.

Method	CVC-ColonDB						ETIS						CVC-300					
	mDSC	mIoU	F_β^w	S_α	$maxE_\xi$	MAE	mDSC	mIoU	F_β^w	S_α	$maxE_\xi$	MAE	mDSC	mIoU	F_β^w	S_α	$maxE_\xi$	MAE
CNN	0.576	0.481	0.551	0.735	0.811	0.053	0.485	0.380	0.425	0.700	0.755	0.032	0.745	0.653	0.715	0.842	0.892	0.015
CNN+SBA	0.639	0.549	0.615	0.769	0.843	0.048	0.499	0.409	0.447	0.719	0.754	0.030	0.841	0.760	0.818	0.895	0.956	0.010
Transformer	0.793	0.706	0.771	0.860	0.908	0.030	0.779	0.695	0.737	0.873	0.912	0.018	0.890	0.820	0.868	0.933	0.977	0.008
Transformer+SBA	0.814	0.736	0.799	0.871	**0.920**	0.029	0.777	0.697	0.733	0.872	0.910	0.018	0.896	0.829	0.877	0.935	0.978	0.008
CNN+Transformer+TAC	0.799	0.711	0.777	0.862	0.914	0.031	0.775	0.687	0.731	0.869	0.907	0.017	0.871	0.798	0.845	0.921	0.965	0.011
CNN+Transformer+TAC+SBA	**0.815**	**0.737**	**0.799**	**0.871**	0.919	**0.029**	**0.790**	**0.712**	**0.748**	**0.878**	**0.912**	**0.016**	**0.903**	**0.840**	**0.887**	**0.941**	**0.981**	**0.007**

Table 3: Statistical comparisons of SBA-Net under different configurations for CVC-ColonDB, ETIS and CVC-300. The best results are shown in bold.

- CNN+Transformer+TAC+SBA: The full SBA-Net which includes the TAC based encoder and the SBA based decoder.

SBA consistently improves CNN and also achieves better performances than Transformer in almost all cases. The same observation can be found for TAC by CNN+Transformer+TAC, thanks to the rich CNN features and the boosting efforts of TAC. The performances of TAC alone are not significant when comparing to SBA empowered results, especially those from Transformer+SBA, showing SBA is very important for better performance. The best performances generally belong to the full SBA-Net as expected.

7 Conclusions

This paper proposes a spectral domain based polyp segmentation framework, SBA-Net. A unified sub-band based feature attention module, SBA, is adopted, which boosts the object recognition by high or middle sub-bands from encoder features. A strong Transformer attended CNN encoder is also incorporated, which takes TAC as the main block to enhance the CNN features with stronger long-range information by Transformer and supply robust features to SBA. Experimental results show the efficacy of the proposed method.

SBA-Net adopts fixed Gaussian filters for the sub-bands. An adaptive way to extract the sub-bands is expected so that the SBA can be easily applied to more applications. TAC is simple but effective. However, a more effective fusion operation other than the Hadamard product perhaps can improve the performances further. These are the future directions for more robust polyp segmentation.

Acknowledgements

This work is supported by Natural Science Foundation of Anhui Province (2108085MF210) and Key Natural Science Fund of Department of Education of Anhui Province (KJ2021A0042).

References

[Ahmed and Ali, 2020] Awadelrahman Ahmed and MA Ali. Generative adversarial networks for automatic polyp segmentation. In *Multimedia Evaluation Workshop*, 2020.

[Bernal *et al.*, 2015] 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.

[Borse *et al.*, 2021] Shubhankar Borse, Ying Wang, Yizhe Zhang, and Fatih Porikli. InverseForm: A loss function for structured boundary-aware segmentation. In *CVPR*, pages 5901–5911, 2021.

[Brandao *et al.*, 2017] 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, 2017.

[Chen *et al.*, 2021] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[Dong *et al.*, 2021] 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.

[Du *et al.*, 2022] Xiuquan Du, Xuebin Xu, and Kunpeng Ma. ICGNet: Integration context-based reverse-contour guidance network for polyp segmentation. In *IJCAI*, pages 877–883, 2022.

[Fan *et al.*, 2020] 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 *MICCAI*, pages 263–273, 2020.

[Fang *et al.*, 2021] Xianyong Fang, Xiaohao He, Linbo Wang, and Jianbing Shen. Robust shadow detection by exploring effective shadow contexts. In *ACM Multimedia*, pages 2927–2935, 2021.

[Frank *et al.*, 2020] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *ICML*, pages 3247–3258, 2020.

[Guo *et al.*, 2022] Qingqing Guo, Xianyong Fang, Linbo Wang, and Enming Zhang. Polyp segmentation of colonoscopy images by exploring the uncertain areas. *IEEE Access*, 10:52971–52981, 2022.

[Guo *et al.*, 2023] Qingqing Guo, Xianyong Fang, Kaibing Wang, Yuqing Shi, Linbo Wang, Enming Zhang, and Zhengyi Liu. Parallel matters: Efficient polyp segmentation with parallel structured feature augmentation modules. *IET Image Processing*, 2023.

[He *et al.*, 2022] Yuhong He, Tao Zeng, Ye Xiong, Jialu Li, and Haoran Wei. Deep leaning based frequency-aware single image deraining by extracting knowledge from rain and background. *Machine Learning and Knowledge Extraction*, 4(3):738–752, 2022.

[Huang *et al.*, 2021] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. FSDR: Frequency space domain randomization for domain generalization. In *CVPR*, pages 6891–6902, 2021.

[Huang *et al.*, 2022] Jie Huang, Yajing Liu, Feng Zhao, Keyu Yan, Jinghao Zhang, Yukun Huang, Man Zhou, and Zhiwei Xiong. Deep Fourier-based exposure correction network with spatial-frequency interaction. In *ECCV*, pages 163–180, 2022.

[Jha *et al.*, 2020] 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 *MMM*, pages 451–462, 2020.

[Ji *et al.*, 2021] Xiaozhong Ji, Guangpin Tao, Yun Cao, Ying Tai, Tong Lu, Chengjie Wang, Jilin Li, and Feiyue Huang. Frequency consistent adaptation for real world super resolution. In *AAAI*, volume 35, pages 1664–1672, 2021.

[Kang and Gwak, 2019] Jaeyong Kang and Jeonghwan Gwak. Ensemble of instance segmentation models for polyp segmentation in colonoscopy images. *IEEE Access*, 7:26440–26447, 2019.

[Kim *et al.*, 2021] Taehun Kim, Hyemin Lee, and Daijin Kim. UACANet: Uncertainty augmented context attention for polyp segmentation. In *ACM Multimedia*, pages 2167–2175, 2021.

[Li *et al.*, 2021] Shaohua Li, Xiuchao Sui, Xiangde Luo, Xinxing Xu, Yong Liu, and Rick Goh. Medical image segmentation using squeeze-and-expansion Transformers. In *IJCAI*, pages 807–815, 2021.

[Li *et al.*, 2022] Weisheng Li, Yinghui Zhao, Feiyan Li, and Linhong Wang. MIA-Net: Multi-information aggregation network combining Transformers and convolutional feature learning for polyp segmentation. *Knowledge-Based Systems*, page 108824, 2022.

[Lin *et al.*, 2022] Ailiang Lin, Jiayu Xu, Jinxing Li, and Guangming Lu. ConTrans: Improving Transformer with convolutional attention for medical image segmentation. In *MICCAI*, pages 297–307, 2022.

[Liu *et al.*, 2021] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. FedDG: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *CVPR*, pages 1013–1023, 2021.

[Liu *et al.*, 2022] Yong Liu, Ran Yu, Jiahao Wang, Xinyuan Zhao, Yitong Wang, Yansong Tang, and Yujiu Yang. Global spectral filter memory network for video object segmentation. In *ECCV*, pages 648–665, 2022.

[Magid *et al.*, 2021] Salma Abdel Magid, Yulun Zhang, Donglai Wei, Won-Dong Jang, Zudi Lin, Yun Fu, and Hanspeter Pfister. Dynamic high-pass filtering and multi-spectral attention for image super-resolution. In *ICCV*, pages 4288–4297, 2021.

[Murphy, 2012] Kevin P Murphy. *Machine learning: A probabilistic perspective*. MIT press, 2012.

[Nanni *et al.*, 2021] Loris Nanni, Daniela Cuza, Alessandra Lumini, Andrea Loreggia, and Sheryl Brahnam. Deep ensembles in bioimage segmentation. *arXiv preprint arXiv:2112.12955*, 2021.

[Nguyen *et al.*, 2021] Tan-Cong Nguyen, Tien-Phat Nguyen, Gia-Han Diep, Anh-Huy Tran-Dinh, Tam V Nguyen, and Minh-Triet Tran. CCBANet: Cascading context and balancing attention for polyp segmentation. In *MICCAI*, pages 633–643, 2021.

[Rahman and Wang, 2016] Md Atiqur Rahman and Yang Wang. Optimizing Intersection-Over-Union in deep neural networks for image segmentation. In *ISVC*, pages 234–244, 2016.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015.

[Sawicki *et al.*, 2021] Tomasz Sawicki, Monika Ruszkowska, Anna Danielewicz, Ewa Niedźwiedzka, Tomasz Arłukowicz, and Katarzyna E Przybyłowicz. A review of colorectal cancer in terms of epidemiology, risk factors, development, symptoms and diagnosis. *Cancers*, 13(9):2025, 2021.

[Shamshad *et al.*, 2022] Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu. Transformers in medical imaging: A survey. *Intelligent Medicine*, 2022.

[Silva *et al.*, 2014] 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.

[Suvorov *et al.*, 2022] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with Fourier convolutions. In *WACV*, pages 2149–2159, 2022.

[Tajbakhsh *et al.*, 2015] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE T-MI*, 35(2):630–644, 2015.

[Tomar *et al.*, 2021] Nikhil Kumar Tomar, Debesh Jha, Sharib Ali, Håvard D Johansen, Dag Johansen, Michael A Riegler, and Pål Halvorsen. DDANet: Dual decoder attention network for automatic polyp segmentation. In *ICPR*, pages 307–314, 2021.

[Tomar *et al.*, 2022a] Nikhil Kumar Tomar, Debesh Jha, Ulas Bagci, and Sharib Ali. TGANet: Text-guided attention for improved polyp segmentation. In *MICCAI*, 2022.

[Tomar *et al.*, 2022b] Nikhil Kumar Tomar, Abhishek Srivastava, Ulas Bagci, and Debesh Jha. Automatic polyp segmentation with multiple kernel dilated convolution network. In *IEEE International Symposium on Computer-Based Medical Systems*, pages 317–322, 2022.

[Vázquez *et al.*, 2017] D Vázquez, J Bernal, FJ Sánchez, G Fernández-Esparrach, AM López, A Romero, M Drozdzal, and A Courville. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of Healthcare Engineering*, 2017:4037190–4037190, 2017.

[Wang *et al.*, 2022a] Jinfeng Wang, Qiming Huang, Feilong Tang, Jia Meng, Jionglong Su, and Sifan Song. Stepwise feature fusion: Local guides global. In *MICCAI*, pages 110–120, 2022.

[Wang *et al.*, 2022b] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. PVT v2: Improved baselines with pyramid vision Transformer. *Computational Visual Media*, pages 1–10, 2022.

[Wu *et al.*, 2021] Huisi Wu, Jiafu Zhong, Wei Wang, Zhenkun Wen, and Jing Qin. Precise yet efficient semantic calibration and refinement in ConvNets for real-time polyp segmentation from colonoscopy videos. In *AAAI*, number 4, pages 2916–2924, 2021.

[Xu *et al.*, 2020] Kai Xu, Minghai Qin, Fei Sun, Yuhao Wang, Yen-Kuang Chen, and Fengbo Ren. Learning in the frequency domain. In *CVPR*, pages 1740–1749, 2020.

[Ye *et al.*, 2022] Tao Ye, Jun Zhang, Yunwang Li, Xi Zhang, Zongyang Zhao, and Zezhong Li. CT-Net: An efficient network for low-altitude object detection based on convolution and Transformer. *IEEE TIM*, 71:1–12, 2022.

[Yin *et al.*, 2019] Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A Fourier perspective on model robustness in computer vision. *NIPS*, 32, 2019.

[Yin *et al.*, 2022] Zijin Yin, Kongming Liang, Zhanyu Ma, and Jun Guo. Duplex contextual relation network for polyp segmentation. In *ISBI*, pages 1–5, 2022.

[Zhang *et al.*, 2021] Yundong Zhang, Huiye Liu, and Qiang Hu. TransFuse: Fusing Transformers and CNNs for medical image segmentation. In *MICCAI*, pages 14–24, 2021.

[Zhong *et al.*, 2020] Jiafu Zhong, Wei Wang, Huisi Wu, Zhenkun Wen, and Jing Qin. PolypSeg: An efficient context-aware network for polyp segmentation from colonoscopy videos. In *MICCAI*, pages 285–294, 2020.

[Zhou *et al.*, 2022] Ziqi Zhou, Lei Qi, and Yinghuan Shi. Generalizable medical image segmentation via random amplitude mixup and domain-specific image restoration. In *ECCV*, pages 420–436. Springer, 2022.