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, liuywen}@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.

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.
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 Transformer 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 input are then aggregated into 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-art 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 cosine 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 invariant 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., 2022]...
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 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 $3 \times 3$ convolutions. This experiment shows that more informative frequency band different from the high-frequency one 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.

SBA works as follows. The input encoder features $f_E$ 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 via subtraction. This sub-band is then converted back to spatial features which act as attention map to boost the target features $f_{D_1}$.

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.
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 \( f_C \) is convolved by \( 3 \times 3 \) (\( C_{3 \times 3} \)) two times and then max pooled (M) to obtain the convolution features, which are then attended by the corresponding Transformer features \( f_T \) from PVT to obtain augmented features \( f'_E \),

\[
f'_E = M(C_{3 \times 3}(C_{3 \times 3}(f_C))) \otimes f_T
\]  

(2)

Then, \( f'_E \) are non-linearly enhanced by \( 1 \times 1 \) and \( 3 \times 3 \) convolutions (\( C_{1 \times 1} \), \( C_{3 \times 3} \)), ReLU activation (R) and residual connection to obtain the final boosted convolution features \( f_E \) which are also the input features \( f_C \) of the next TAC in SBA-Net. Formally, \( f_E \) is computed as

\[
f_E = f'_E + C_{1 \times 1}(R(C_{1 \times 1}(f'_E)))
\]  

(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

\[
\mathcal{L} = \mathcal{L}_{BCE}^{w}(P, \hat{P}) + \mathcal{L}_{IoU}^{w}(P, \hat{P}) + \gamma \mathcal{L}_{IF}(P, \hat{P}),
\]  

(4)

where: \( \mathcal{L}_{BCE}^{w}, \mathcal{L}_{IoU}^{w} \) and \( \mathcal{L}_{IF} \) are the weighted BCE and IoU losses, and the IF loss, respectively; and \( \gamma \) 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 \( 352 \times 352 \) 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 (\( F_\beta \)), S-measure (\( S_\alpha \)), E-measure (\( E_\xi \)) and Mean Absolute Error (MAE).

DSC and IoU are similarity measures at the regional level, focusing on the internal object consistencies. \( F_\beta \) comprehensively considers the recall and precision with all pixels and thus is affected less by the individual pixels. \( S_\alpha \) measures the structural similarities, while \( E_\xi \) 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, \( \max E_\xi \), 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], Polypp-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 effectiveses 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 downsamplo 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>CVC-ClinicDB</th>
<th>Kvasir</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mDSC</td>
<td>mIoU</td>
</tr>
<tr>
<td>U-Net [Ronneberger et al., 2015]</td>
<td>0.879</td>
<td>0.818</td>
</tr>
<tr>
<td>PraNet [Fan et al., 2020]</td>
<td>0.899</td>
<td>0.849</td>
</tr>
<tr>
<td>TransNet [Chen et al., 2021]</td>
<td>0.828</td>
<td>0.771</td>
</tr>
<tr>
<td>Polyp-PVT [Wang et al., 2021]</td>
<td>0.911</td>
<td>0.853</td>
</tr>
<tr>
<td>FedDG [Liu et al., 2021]</td>
<td>0.855</td>
<td>0.791</td>
</tr>
<tr>
<td>MKDCNet [Tomar et al., 2022b]</td>
<td>0.672</td>
<td>0.613</td>
</tr>
<tr>
<td>TGANet [Tomar et al., 2022a]</td>
<td>0.901</td>
<td>0.858</td>
</tr>
<tr>
<td>SSFormer-L [Wang et al., 2022a]</td>
<td>0.906</td>
<td>0.855</td>
</tr>
<tr>
<td>SBA-Net</td>
<td>0.937</td>
<td>0.889</td>
</tr>
</tbody>
</table>

Table 1: Statistical comparisons of different methods for CVC-ClinicDB and Kvasir. The best results are shown in bold.
the SBA can be easily applied to more applications. TAC is an adaptive way to extract the sub-bands is expected so that the CNN features with stronger long-range information by the attention module, SBA, is adopted, which boosts the object recognition by high or middle sub-bands from encoder features. A 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


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

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 effects 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


