M3FPOLYPSEGNET: SEGMENTATION NETWORK WITH MULTI-FREQUENCY FEATURE FUSION FOR POLYP LOCALIZATION IN COLONOSCOPY IMAGES

Ju-Hyeon Nam, Seo-Hyeong Park, Nur Suriza Syazwany, Yerim Jung, Yu-Han Im and Sang-Chul Lee

Department of Computer Science and Engineering, Inha University, Incheon, Republic of Korea

ABSTRACT

Polyp segmentation is crucial for preventing colorectal cancer a common type of cancer. Deep learning has been used to segment polyps automatically, which reduces the risk of misdiagnosis. Localizing small polyps in colonoscopy images is challenging because of its complex characteristics, such as color, occlusion, and various shapes of polyps. To address this challenge, a novel frequency-based fully convolutional neural network, *Multi-Frequency Feature Fusion Polyp Segmentation Network (M3FPolypSegNet)* was proposed to decompose the input image into low/high/full-frequency components to use the characteristics of each component. We used three independent multi-frequency encoders to map multiple input images into a high-dimensional feature space. In the *Frequency-ASPP Scalable Attention Module (F-ASPP SAM)*, ASPP was applied between each frequency component to preserve scale information. Subsequently, scalable attention was applied to emphasize polyp regions in a high-dimensional feature space. Finally, we designed three multi-task learning (i.e., region, edge, and distance) in four decoder blocks to learn the structural characteristics of the region. The proposed model outperformed various segmentation models with performance gains of 6.92% and 7.52% on average for all metrics on CVC-ClinicDB and BKAI-IGH-NeoPolyp, respectively.

Index Terms— Deep learning, Fully convolutional neural network, Polyp segmentation, Frequency domain

1. INTRODUCTION

Colorectal Cancer is one of the most common types of cancer worldwide [1]. Because colorectal cancer typically starts as polyps and progresses to cancer, medical experts recommend regular colonoscopies for the early detection of polyps. However, manual polyp detection is not highly accurate, because of dependence on the ability of doctors and the limitations of colonoscopy equipment, which results in a decreasing survival rate.

This work was supported in part by the National Research Foundation of Korea (NRF) under Grant NRF-2021R1A2C2010893 and in part by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2022-00155915, Artificial Intelligence Convergence Innovation Human Resources Development (Inha University).

With the advancement of deep learning [2, 3], automatic polyp segmentation has been developed rapidly to reduce misdiagnosis resulting from overworked doctors and obsolete equipment. U-Net [4] has been widely adopted in polyp segmentation tasks because of its remarkable performance in biomedical image segmentation. In U-Net++ [5], the ensemble nested U-Net of various depths and deep supervision are used. ResNet++ [6] is focused on attention mechanisms and multi-scale feature extraction. In PraNet [7], reverse attention is used to clarify the relationship between areas and boundary cues to mitigate misaligned prediction.

The localization of small polyps in polyp segmentation is challenging because of complex structures such as, colors, occlusion, and various shapes of polyps and affects model performance. Frequency-based methods exhibit considerable potential for image segmentation. In FRCU-Net [8], the Laplacian pyramid and Frequency Re-Calibration module that implement frequency attention to the basic U-Net architecture was applied. By contrast, in FDA [9], discrete Fourier transform (DFT) is applied to each image, replacing the low-frequency component of the target image with the source image and the source image with the target style is reconstructed through inverse DFT.

In this paper, we propose a novel frequency-based fully convolutional neural network (FCNN), Multi-Frequency Feature Fusion Polyp Segmentation Network (M3FPolypSegNet). M3F PolypSegNet extracts feature maps by decomposing the input image into low/full/high-frequency to learn unique characteristics. The Frequency-ASPP Scalable Attention Module (F-ASPP SAM) combines the modified ASPP and attention modules to preserve scale information and emphasize polyp regions in a high-dimensional feature space. Finally, a multi-task (i.e., region, edge, and distance) loss is designed for learning the structural characteristics of the polyp region during training. The proposed model outperformed various segmentation models with performance gains of 6.92% and 7.52% on average for all metrics on CVC-ClinicDB and BKAI-IGH-NeoPolyp, respectively. The contributions of the study are as follows:

- We propose a novel polyp segmentation model (M3F PolypSegNet) based on a multi-frequency encoder and a single-decoder architecture that utilizes unique char-

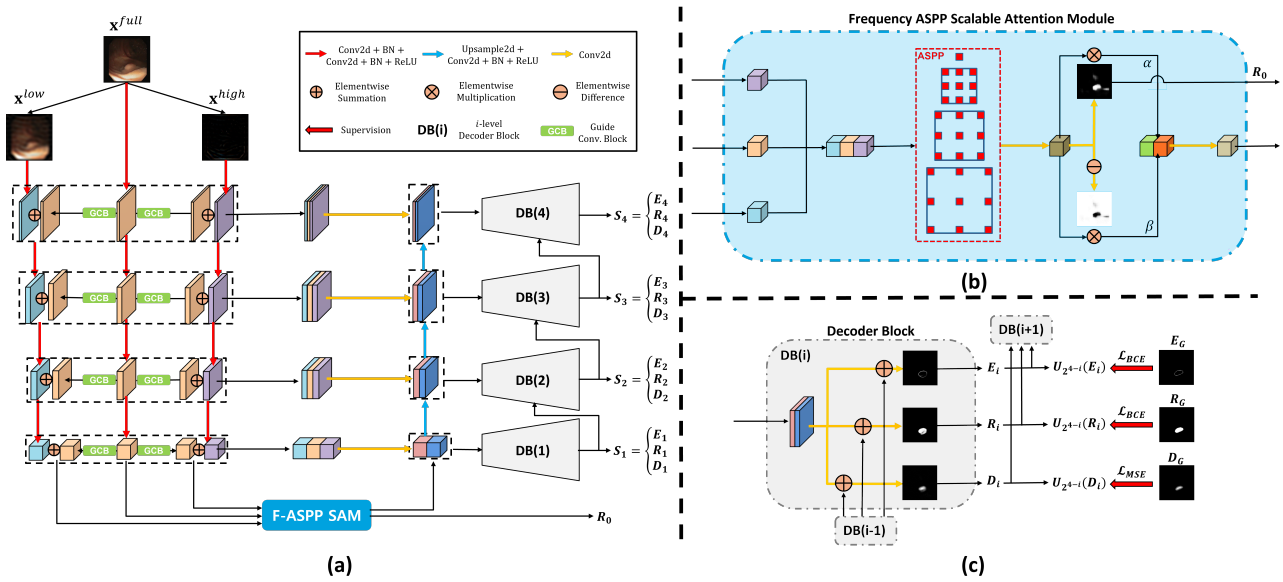

Fig. 1. The proposed M3FPolypSegNet architecture. (a) Overall block diagram of our network, (b) Overview of F-ASPP SAM, and (c) Training procedure of i-level decoder block **DB(i)** with multi-task learning and $S_i = \{R_i, E_i, D_i\}$ is multiple output from each decoder block **DB(i)** for $i = 1, 2, 3, 4$.

acteristics for each frequency component.

- F-ASPP SAM introduces trainable parameters between the foreground/background attention of frequency and scale to prevent information loss during the gradual upsampling of the decoder block. Furthermore, the vanishing gradient problem was prevented by performing multi-task deep supervision training in each decoder block.
- We experimentally achieved state-of-the-art performance in various evaluation metrics when comparing various polyp image segmentation models on two datasets (CVC-ClinicDB and BKAI-IGH-NeoPolyp).

2. METHOD

M3FPolypSegNet is a novel polyp segmentation architecture with multiple encoders and a single decoder-based FCNN for end-to-end training. Our model consists of three primary components, namely multi-frequency encoder, frequency-ASPP scalable attention module, and single decoder with multi-task learning. Figure 1 (a) display the overall architecture of M3FPolypSegNet.

2.1. Multi-Frequency Encoder

We were motivated by [10], in which the low/high-frequency components exhibit distinct cues that contain the approximate location and details from the low/high-frequency polyp image, respectively. Therefore, our motivation is that if latent feature maps of images are extracted and combined with various frequencies, a multi-modality perspective can be used. For this purpose, we applied DFT to transform the input image in the frequency domain.

Let $\mathbf{x}^{full} \in \mathbb{R}^{H\times W\times 3}$ be the input image where H and W are the height and width of the input image, respectively. First, we transform the input image into the frequency domain using DFT as follows:

$$\hat{\mathbf{x}}^{full} = \mathcal{F}\{\mathbf{x}^{full}\} \in \mathbb{C}^{H\times W\times 3} \tag{1}$$

where $\mathcal{F}\{\cdot\}$ denotes the DFT with a shift operator that multiplies $(-1)^{u+v}$ in the frequency domain to translate the DC component to the center of the image where u and v are the frequency index. The low/high-frequency components based on the proportion of the total power spectrum of the image are defined. Let T and P be the total and partial power spectrum of the image, respectively. For a power spectrum ratio $0 \leq r \leq 1$, we define the low-frequency component satisfying $P/T \leq r$. After finding the maximum frequency index (u_{max}, v_{max}) satisfying the condition, we define the low-frequency pass mask M_{low} as follows:

$$M_{low}(u,v) = \begin{cases} 1 \text{ if } u^2 + v^2 \leq R^2 \\ 0 \text{ if Otherwise} \end{cases} \tag{2}$$

where $R^2 = u_{max}^2 + v_{max}^2$. We decompose $\hat{\mathbf{x}}$ into the low/high-frequency components by two binary masks, namely M_{low} and $1 - M_{low}$ to create a new input image as follows:

$$\begin{cases} \mathbf{x}^{low} = \mathcal{F}^{-1}\{\mathbf{x}^{full} \otimes M_{low}\} \\ \mathbf{x}^{high} = \mathcal{F}^{-1}\{\mathbf{x}^{full} \otimes (1 - M_{low})\} \end{cases} \tag{3}$$

where $\mathcal{F}^{-1}\{\cdot\}$ and $\otimes$ represent the inverse DFT with a shift operator that multiplies $(-1)^{x+y}$ and element-wise multiplication, respectively. We design an encoder-decoder architecture similar to U-Net [4] in this paper. The encoder consists of four blocks with a convolutional layer, batch normalization, and ReLU activation function. First, a high-dimensional

feature map was extracted from a full-frequency image for each $i = 1, 2, 3, 4$ as follows:

$$\mathbf{x}_i^{full} = e_i^{full}\left(\mathbf{x}_{i-1}^{full}\right) \tag{4}$$

where $\mathbf{x}_0^{full} = \mathbf{x}^{full}$, and $e_i(\cdot)$ is i-th encoder block. The multi-frequency encoder has higher representation power than a single encoder by training multiple encoders from images in various frequency components. However, $\mathbf{x}^{low}$ and $\mathbf{x}^{high}$ exhibit information loss, because specific frequency ranges are completely removed. To solve this problem, we add a residual connection, Guided Convolution Block (GCB), from a full-frequency encoder into low/high-frequency encoders while extracting high-dimensional feature maps as follows:

$$\mathbf{x}_i^{z} = e_i^{z}\left(\mathbf{x}_{i-1}^{z}\right) + GCB\left(\mathbf{x}_i^{full}\right) \tag{5}$$

where $z = low, high$ and $GCB(\cdot)$ consists of 1×1 convolution and ReLU activation function. Because the three inputs have distinct characteristics, we use three independent encoders that do not share parameters. Through this method, each encoder extracts feature maps corresponding to each frequency component while minimizing information loss.

2.2. Frequency-ASPP Scalable Attention Module

In the F-ASPP SAM, heterogeneous feature maps are fused to enhance polyp regions. The architecture of F-ASPP SAM is displayed in Figure 1 (b). First, the three feature maps are concatenated as $\mathbf{X} = \left[\mathbf{x}_4^{low}, \mathbf{x}_4^{full}, \mathbf{x}_4^{high}\right] \in \mathbb{R}^{H/16, W/16, 3C_4}$. Convolutional layers with various atrous rates are then used for efficient multi-frequency and scale fusion. In this paper, after modifying the original ASPP architecture, the four branches use convolution operations with various atrous rates to extract feature maps and then sum them.

However, this method cannot capture polyp regions, and the results tend to be scattered. Therefore, we simultaneously applied foreground/background attention and concatenated them to preserve the information of polyp regions during progressive decoding. We introduced two trainable parameters (α and β) to determine two attention ratios as follows:

$$O = C_{3\times3}([(\alpha g(f(\mathbf{X})))\,\mathbf{X}, (\beta(1 - g(f(\mathbf{X}))))\,\mathbf{X}]) \tag{6}$$

where $f(\cdot)$ and $g(\cdot)$ are four-branch ASPP module which we modified from original module [11] and 1×1 convolution block, respectively. The result of the concatenation goes through a 3×3 convolutional layer and continues with the decoding.

2.3. Training and Inference Process

By applying deep supervision to each block of the decoder, we obtained four additional outputs R_0, S_1, S_2, S_3 and the final output S_4. We denote that $S_i = \{R_i, E_i, D_i\}$ is multiple outputs from each decoder block, **DB(i)**, that performs multi-task learning for each $i = 1, 2, 3, 4$. At this stage, each task prediction is upsampled to the same size as the ground truth to calculate the loss function. The edge ground truth, E_G, is obtained by applying the anisotropic Sobel edge detection filter from R_G. The distance map ground truth, D_G, is obtained from R_G by applying a distance transform and normalizing the distances from pixels in the region to edges between 0 and 1. The loss function for the ith decoder block, **DB(i)**, is computed as follows:

$$\begin{aligned}\mathcal{L}^i = \mathcal{L}_{BCE}(U_{2^{4-i}}(R_i), R_G) + \mathcal{L}_{BCE}(U_{2^{4-i}}(E_i), E_G) \\ + \mathcal{L}_{MSE}((U_{2^{4-i}}(D_i), D_G)\end{aligned} \tag{7}$$

where $U_{2^{4-i}}(\cdot)$ is bi-linear interpolation with a 2^{4-i} scale factor. Finally, the total loss function in M3FPolypSegNet is $\mathcal{L}_{total} = \mathcal{L}_{BCE}(R_0, R_G) + \sum_{i=1}^{4}\mathcal{L}^i$. Additionally, the final prediction map R_{final} can be obtained by applying the sigmoid function to $R_4 \in S_4$.

3. EXPERIMENTAL RESULTS

3.1. Experimental Settings and Implementation Details

We implemented M3FPolypSegNet in Pytorch 1.11 and Python 3.8, and used two datasets (CVC-ClincDB [12] and BKAI-IGH-NeoPolyp [13]) for training and evaluation. All input images were resized at the same resolution of 256×256. We compared the proposed M3FPolypSegNet with ten existing segmentation networks (FCN8s [2], DeepLabv3+ [11], SegNet [14], U-Net [4], U-Net++ [15], CENet [16], ResU-Net [17], ResU-Net++ [6], PraNet [7]). We optimized parameters using the Adam optimizer in an end-to-end approach. The initial learning rate started from 10^{-4} and decreased to 10^{-6} by using the cosine annealing learning rate scheduler, and the training settings were set to a batch size of 16 and epochs of 200. During the training phase, a random horizontal flipping with a probability of 50% and a random non-extended rotation between -5° and 5° were applied. We use five representative segmentation metrics (pixel accuracy, precision, recall, F1-Score, and IoU) for comparison. We fixed $r = 0.5$ to equally set the importance between low/high-frequency[1].

3.2. Results Analysis

Table 1 summarizes experiment results. As presented in Table 1, M3FPolypSegNet outperformed on all metrics but exhibited a much higher recall on the CVC-ClnicDB dataset. In particular, IoU is improved by approximately 2.5%, and

[1]The code is available in our M3FPolypSegNet.git

Table 1 Experiment results on the CVC-ClinicDB and BKAI-IGH-NeoPolyps datasets. **Bold** and *italic* denote best and second-best performance, respectively.

Method	Parameters	CVC-ClinicDB [12]					BKAI-IGH-NeoPolyps [13]				
		Acc	F1-Score	Recall	Precision	IoU	Acc	F1-Score	Recall	Precision	IoU
FCN8s [2]	18.64M	0.9723	0.8015	0.8175	0.8299	0.7285	0.9659	0.7234	0.8399	0.7092	0.6726
DeepLabV3+ [11]	59.34M	0.9791	0.8373	0.8306	0.8736	0.7844	0.9806	0.8822	0.9066	0.8881	0.8422
SegNet [14]	16.50M	0.9631	0.5787	0.6862	0.5707	0.5484	0.9501	0.6440	0.7238	0.6374	0.6115
U-Net [4]	34.53M	0.9792	0.8585	0.8635	0.8962	0.7985	0.9842	0.9052	0.9217	0.9138	0.8696
U-Net++ [15]	36.63M	0.9827	0.8817	*0.8897*	0.9157	0.8257	0.9861	0.9178	**0.9515**	0.9194	0.8878
CENet [16]	29.00M	*0.9845*	*0.8854*	0.8699	**0.9249**	*0.8370*	0.9869	0.9133	0.9285	0.9161	0.8790
ResU-Net [17]	10.81M	0.9522	0.7812	0.7563	0.8720	0.7022	0.9791	0.8860	0.9222	0.8878	0.8362
ResU-Net++ [6]	14.48M	0.9684	0.7643	0.7907	0.7948	0.7026	0.9476	0.7846	0.8887	0.7611	0.7152
PraNet [7]	32.55M	0.9797	0.8748	0.8654	*0.9244*	0.8248	**0.9918**	*0.9370*	*0.9461*	*0.9395*	*0.9081*
M3FPolySegNet (Ours)	22.39M	**0.9883**	**0.8937**	**0.9015**	0.9062	**0.8507**	*0.9891*	**0.9399**	0.9441	**0.9450**	**0.9147**

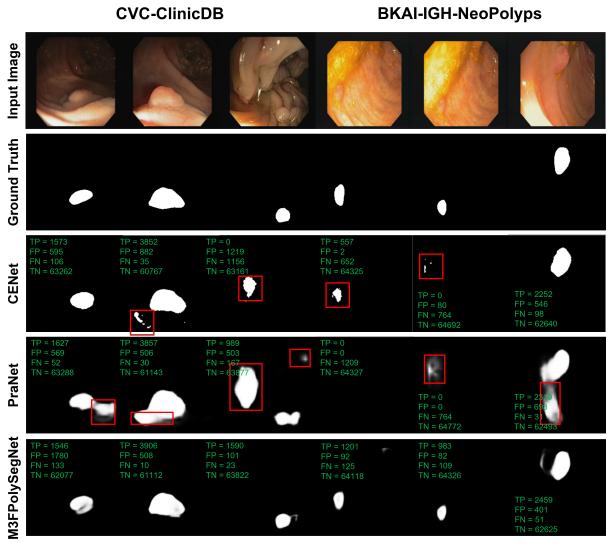

Fig. 2. Qualitative results of PraNet [7], CENet [16] and M3FPolypSegNet for each input image. Red boxes indicate misdiagnosis of the input images.

1.3% compared with PraNet and CENet, respectively. We introduce two trainable attention ratios (α and β) between foreground and background in F-ASPP SAM. Therefore, M3FPolypSegNet's prediction mask can observe higher details of polyp edges compared to PraNet [7], which only uses existing reverse attention. Furthermore, M3FPolySegNet obtained higher performance than PraNet even if approximately 10M lower number of parameters. In Figure 2, we observe the qualitative results of each method. In colonoscopy image, a single lighting source makes difficulties that edges of polyps are ambiguous. Our model leverages the high-frequency components to increase invariance and perform specialized segmentation on colonoscopy images. High-frequency components tend to reveal detailed information, such as the edges of polyps, and by utilizing this information, the accuracy and performance of polyp segmentation can be improved.

3.3. Ablation Study

In this section, we measure the performance of each component of M3FPolypSegNet; Frequency(FD), GCB, Multi-Task Learning (MTL) and Frequency-ASPP Scalable Attention Module (F-ASPP SAM), separately on two datasets to gain a deeper understanding of our model. The results are summarized in the Table 2.

Table 2 Ablation study on the CVC-ClinicDB and BKAI-IGH-NeoPolyps (BKAI) datasets. **Bold** and *italic* denote best and second-best performance, respectively.

FD	GCB	MTL	F-ASPP SAM	CVC-ClincDB	BKAI
✗	✗	✗	✗	0.7985	0.8696
✓	✓	✗	✗	0.7909	0.8684
✓	✗	✓	✗	0.8103	0.8771
✓	✓	✓	✗	*0.8282*	*0.8820*
✓	✓	✓	✓	**0.8507**	**0.9147**

First, when FD and GCB are applied to U-Net, performance degradation occurs on CVC-ClinicDB and BKAI. However, after MTL is applied for each decoder, the performance is improved by 2.97% and 1.24% compared with U-Net. Table 2 reveals that keeping MTL and removing GCB results in performance degradation of 3.49% and 4.04%, respectively. This result indicates that GCB supplement information loss due to FD and gain performance improvement. Finally, when F-ASPP SA is added, the performance of both datasets is improved by a large margin. It can be seen that CVC-ClinicDB and BKAI-IGH-NeoPolyp improve from 0.8282 and 0.882 to 0.8507 and 0.9147, respectively.

4. CONCLUSION

We propose M3FPolypSegNet, a polyp segmentation model based on frequency-domain automated colonoscopy images. Experiment results revealed that M3FPolypSegNet exhibits higher learning and evaluation capabilities than existing polyp segmentation models (CVC-ClinicDB: $> 6\%$ & BKAI-IGH-NeoPolyp: $> 7\%$). In particular, the power spectrum-based frequency decomposition technique and multi-frequency-based feature fusion method enable high-performance improvements by preventing spatial information loss during training. Furthermore, we demonstrated that the proposed model can be applied to various datasets because it does not require any initialization techniques or post-processing techniques. We conducted additional research in areas such as various topics related to biomedical images (brain tumor segmentation, liver segmentation, etc.), rather than restricting M3FPolypSegNet to polyp segmentation task.

5. REFERENCES

[1] Jorge Bernal, Nima Tajkbaksh, Francisco Javier Sanchez, Bogdan J Matuszewski, Hao Chen, Lequan Yu, Quentin Angermann, Olivier Romain, Bjørn Rustad, Ilangko Balasingham, et al., "Comparative validation of polyp detection methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge," *IEEE transactions on medical imaging*, vol. 36, no. 6, pp. 1231–1249, 2017.

[2] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[3] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[5] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE transactions on medical imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.

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

[7] 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 *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23*. Springer, 2020, pp. 263–273.

[8] Reza Azad, Afshin Bozorgpour, Maryam Asadi-Aghbolaghi, Dorit Merhof, and Sergio Escalera, "Deep frequency re-calibration u-net for medical image segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3274–3283.

[9] Yanchao Yang and Stefano Soatto, "Fda: Fourier domain adaptation for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4085–4095.

[10] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing, "High-frequency component helps explain the generalization of convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8684–8694.

[11] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[12] 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*, vol. 43, pp. 99–111, 2015.

[13] Phan Ngoc Lan, Nguyen Sy An, Dao Viet Hang, Dao Van Long, Tran Quang Trung, Nguyen Thi Thuy, and Dinh Viet Sang, "Neounet: Towards accurate colon polyp segmentation and neoplasm detection," in *Advances in Visual Computing: 16th International Symposium, ISVC 2021, Virtual Event, October 4-6, 2021, Proceedings, Part II*. Springer, 2021, pp. 15–28.

[14] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[15] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*. Springer, 2018, pp. 3–11.

[16] Zaiwang Gu, Jun Cheng, Huazhu Fu, Kang Zhou, Huaying Hao, Yitian Zhao, Tianyang Zhang, Shenghua Gao, and Jiang Liu, "Ce-net: Context encoder network for 2d medical image segmentation," *IEEE transactions on medical imaging*, vol. 38, no. 10, pp. 2281–2292, 2019.

[17] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang, "Road extraction by deep residual u-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, 2018.