MEGANet: Multi-Scale Edge-Guided Attention Network for Weak Boundary Polyp Segmentation

Nhat-Tan Bui[1], Dinh-Hieu Hoang[2,3], Quang-Thuc Nguyen[2,3], Minh-Triet Tran[2,3], Ngan Le[1]
[1]AICV Lab, University of Arkansas, Fayetteville, Arkansas, USA
[2]University of Science, and John von Neumann Institute, VNU-HCM
[3]Vietnam National University, Ho Chi Minh City, Vietnam

Abstract

*Efficient polyp segmentation in healthcare plays a critical role in enabling early diagnosis of colorectal cancer. However, the segmentation of polyps presents numerous challenges, including the intricate distribution of backgrounds, variations in polyp sizes and shapes, and indistinct boundaries. Defining the boundary between the foreground (i.e. polyp itself) and the background (surrounding tissue) is difficult. To mitigate these challenges, we propose **M**ulti-Scale **E**dge-**G**uided **A**ttention Network (MEGANet) tailored specifically for polyp segmentation within colonoscopy images. This network draws inspiration from the fusion of a classical edge detection technique with an attention mechanism. By combining these techniques, MEGANet effectively preserves high-frequency information, notably edges and boundaries, which tend to erode as neural networks deepen. MEGANet is designed as an end-to-end framework, encompassing three key modules: an encoder, which is responsible for capturing and abstracting the features from the input image, a decoder, which focuses on salient features, and the Edge-Guided Attention module (EGA) that employs the Laplacian Operator to accentuate polyp boundaries. Extensive experiments, both qualitative and quantitative, on five benchmark datasets, demonstrate that our MEGANet outperforms other existing SOTA methods under six evaluation metrics. Our code is available at* `https://github.com/UARK-AICV/MEGANet`.

1. Introduction

Colorectal cancer (CRC) is a major health concern due to its high prevalence, ranking as the top gastrointestinal cancer and the third most common cancer. It is second in cancer-related mortality, trailing only behind lung cancer in both genders [27]. Thus, early CRC detection is of utmost importance. Colonoscopy is the gold standard for CRC examination, yet manual detection and localization of polyps in colonoscopic images are labor-intensive, requiring skilled experts. Consequently, accurate computer-aided polyp segmentation is vital for clinicians to evaluate patients.

In the field of Deep Learning (DL), particularly within the context of computer vision where Convolutional Neural Networks (CNNs) have established dominance, encoder-decoder network architectures [6, 7, 11–13, 22, 24, 26, 32, 35, 39, 40] have demonstrated significant success in the realm of medical image segmentation. While methods such as U-Net++ [40], SFA [7], PraNet [6], MSNet [39], and SANet [35] are mainly designed for polyp segmentation, the accuracy of the segmentation heavily hinges on the amalgamation of encoded feature maps from various scales in the contracting path and the semantically enriched decoded feature maps in the expanding path. Despite notable advancements, these methodologies still grapple with the challenge of preserving high-frequency information, a critical aspect of medical imaging. Particularly, the presence of variable mucous membranes surrounding the polyps, differing in shape, color, and texture, contributes to complex and diverse polyp borders. This complexity, combined with the downscaled encoding, challenges the maintenance of border details and the improvement of segmentation during decoding, resulting in imprecise polyp boundary generation. Our insight underscores that edge information obtained through conventional image processing techniques tends to be more straightforward and precise than edges extracted by CNN-based methods, especially when training data is scarce. Hence, a promising strategy involves revisiting classical image processing-based edge extraction techniques. This approach holds the potential for addressing the issue of weak boundaries in medical imaging. In our present study, we harness the capabilities of classical edge features by introducing the Edge-Guided Attention (EGA) module. This module is designed to function across multiple scales, spanning from low-level to high-level features. Its primary objective is to compel the model to focus on edge-related information, thereby enhancing predictions at each decoder level. Importantly, the EGA module achieves this objective without succumbing to noise or encountering the semantic gap. Our EGA approach capitalizes on classical edge extraction methods to augment the accuracy of medical image segmentation. The EGA module, operating at multiple scales and targeting edge-related features, addresses the challenge of weak boundaries in a noise-resistant manner, enriching the segmentation predictions across the decoder's levels.

As a result, this paper introduces the **M**ulti-Scale **E**dge-**G**uided **A**ttention Network (MEGANet), a novel and innovative approach that integrates an EGA module into the U-Net architecture during the decoding process. The primary goal of MEGANet is to preserve crucial edge and boundary information effectively. In essence, MEGANet comprises three key modules: (i) Encoder, responsible for capturing the visual representation of the input image, akin to the encoder in the U-Net architecture. (ii) Decoder aims to extract salient features, following similar settings as in the U-Net decoder architecture. (iii) EGA module, this distinctive module leverages the Laplacian operator to preserve high-frequency information, particularly edges. The EGA module operates on both the embedding feature and multi-level predictions. This strategic combination empowers the model to accentuate intact edge details and polyp boundaries across various scales. In summary, this paper's main contributions are: (i) We explore the potential of the Laplacian operator, a parameter-free method, to enhance the segmentation of weak boundary objects like polyps by preserving high-frequency edge information. (ii) We present a novel architecture, MEGANet, that addresses the challenge of supplementing low-level boundary information using the Laplacian operator. (iii) We extensively evaluate our method on five benchmark datasets, i.e., Kvasir-SEG [14], CVC-ClinicDB [1], CVC-ColonDB [30], ETIS [28], and EndoScene [34], to demonstrate its effectiveness.

2. Related Work

The widely adopted U-Net [26] architecture, known for its effectiveness in medical image segmentation, has been applied to polyp segmentation. U-Net++ [40], an enhanced variant, addresses semantic gaps through nested skip connections. Although these concepts have broader applications beyond medical imaging, they predominantly focus on enhancing feature learning rather than medical-specific challenges.

Various approaches heavily leverage boundary information in addressing the challenge of weak boundaries in medical imaging. SFA [7] introduces an extra decoder for boundary prediction and employs a boundary-sensitive loss to exploit area-boundary relationships. PraNet [6] uses parallel partial decoders and reverse attention modules to progressively extend object regions by incorporating edge features. ACSNet [38] combines local and global context to accommodate varying polyp sizes. NB-AC [16] introduces narrow band active contour attention when considering weak boundary is a confusing case that needs more attention. [17] presents offset curve loss to give more attention to the boundary. DAM-AL [12] employs dilated attention for long-range relationships and introduces a novel loss mechanism. MSNet [39] introduces the multi-scale subtraction module for mitigating redundant and obtaining complementary information. PEFNet [23] focuses on the positional information of the polyp objects in the skip connection with the EfficientNetV2 [31] encoder. M^{2}UNet [33] integrates MetaFormer [37] and multi-scale information for enhanced context exploitation.

In contrast, our approach explicitly addresses the weak boundary issue by incorporating high-frequency edge information obtained through typical image processing techniques. A fundamental limitation of prior methods is their inability to accurately reconstruct input image edges, as CNN-based features are not optimized for this purpose. We opt for the Laplacian operator to extract and retain high-frequency features, particularly edge details, in polyp images. Laplacian, a second-order derivative operator, yields more meaningful edge structures than hand-crafted first-order derivative methods like Sobel [29] or Prewitt [25] without adding computational complexity. Laplacian has been successfully applied in many image processing problems, such as style transfer [2, 19], image super-resolution [15], image synthesis [18], image deraining [8], image-to-image translation [20], etc.

3. Proposed MEGANet

Our proposed MEGANet architecture comprises three main modules: an encoder, a decoder, and an EGA (Edge-Guided Attention) module, as depicted in Figure 1. The encoder, located in the contracting path, captures context and high-level features from the input polyp image. It encompasses five convolutional blocks and results in encoding feature f_i^e at the i^{th} layer. On the other hand, the decoder, situated in the expanding path, leverages the high-level features acquired by the encoder to generate decoding maps f_i^d at the i^{th} layer that matches the original resolution of the input image. To showcase the effectiveness of our MEGANet as well as to conduct a fair comparison with existing work, we evaluate its performance using two distinct backbone networks: ResNet-34 [10] and Res2Net-50 [9].

In the expanding path, pooling and strided convolution layers are employed to downsample feature maps, reducing the volume of information for processing. While downsampling layers offer significant advantages for constructing deep architectures, it's noteworthy that conventional CNNs often suffer from the loss of information as downsampling layers accumulate at deeper levels. Recognizing the im-

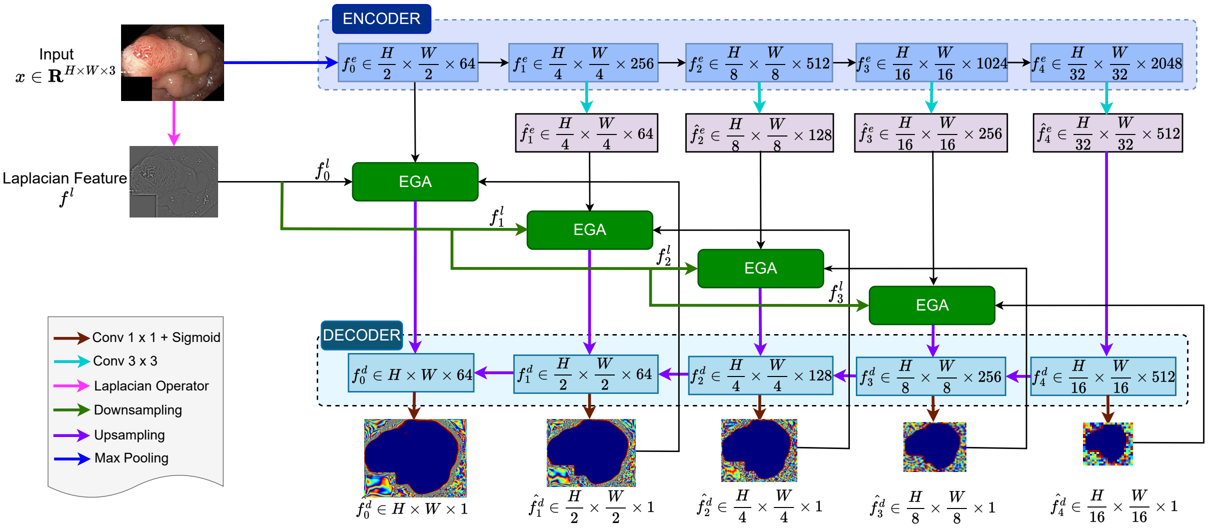
ENCODER
Input
$x \in \mathbf{R}^{H\times W\times 3}$
$f_0^e \in \frac{H}{2}\times\frac{W}{2}\times 64$
$f_1^e \in \frac{H}{4}\times\frac{W}{4}\times 256$
$f_2^e \in \frac{H}{8}\times\frac{W}{8}\times 512$
$f_3^e \in \frac{H}{16}\times\frac{W}{16}\times 1024$
$f_4^e \in \frac{H}{32}\times\frac{W}{32}\times 2048$
$\hat{f}_1^e \in \frac{H}{4}\times\frac{W}{4}\times 64$
$\hat{f}_2^e \in \frac{H}{8}\times\frac{W}{8}\times 128$
$\hat{f}_3^e \in \frac{H}{16}\times\frac{W}{16}\times 256$
$\hat{f}_4^e \in \frac{H}{32}\times\frac{W}{32}\times 512$
Laplacian Feature
f^l
f_0^l
EGA
f_1^l
EGA
f_2^l
EGA
f_3^l
EGA
DECODER
$f_0^d \in H\times W\times 64$
$f_1^d \in \frac{H}{2}\times\frac{W}{2}\times 64$
$f_2^d \in \frac{H}{4}\times\frac{W}{4}\times 128$
$f_3^d \in \frac{H}{8}\times\frac{W}{8}\times 256$
$f_4^d \in \frac{H}{16}\times\frac{W}{16}\times 512$
$\hat{f}_0^d \in H\times W\times 1$
$\hat{f}_1^d \in \frac{H}{2}\times\frac{W}{2}\times 1$
$\hat{f}_2^d \in \frac{H}{4}\times\frac{W}{4}\times 1$
$\hat{f}_3^d \in \frac{H}{8}\times\frac{W}{8}\times 1$
$\hat{f}_4^d \in \frac{H}{16}\times\frac{W}{16}\times 1$
Conv 1 x 1 + Sigmoid
Conv 3 x 3
Laplacian Operator
Downsampling
Upsampling
Max Pooling

Figure 1. Overall architecture of proposed MEGANet(Res2Net-50), including three modules, i.e., *encoder* and *decoder*, which utilizes a U-Net to extract visual representations, and the novel *edge attention* module, denoted as the Edge-Guided Attention (EGA) module, designed to retain high-frequency details effectively. In this context, H, W represents the input height and width.

portance of preserving such critical information in medical segmentation, we introduce a novel Edge-Guided Attention module (EGA) that operates between the two aforementioned paths at every resolution level. In MEGANet, the output of the contracting path serves as an input to the EGA module, and subsequently, the output of the EGA module feeds into the expanding path. This establishes a coherent linkage between the contracting and expanding paths through the intermediary EGA modules. The subsequent subsection will delve into a comprehensive explanation of the EGA module's functioning and attributes. We detail the EGA module in the following section.

3.1. Edge-Guided Attention Module (EGA)

The primary objective of the EGA module is to robustly preserve edge information across multiple scales, effectively addressing the issue of weak boundaries in polyp segmentation. Additionally, the EGA module plays a pivotal role in bridging the semantic gap between the low-level features extracted by the encoder and the high-level features produced by the decoder prior to their fusion.

At the i-th layer, the EGA module takes three inputs: the embedding feature $\hat{f}_i^e$ from the encoder, the high-frequency feature f_i^l obtained through classical image processing methods, i.e., edge detector, and the higher-level predicted feature $\hat{f}_{i+1}^d$ generated by the decoder. The EGA module processes these inputs and generates an output feature map denoted as f_i^d. The detailed process by which the inputs are processed and the specific operations performed by the EGA module are elaborated below. Refer to Figure 2 for a visual representation of the EGA module.

3.1.1 EGA Input

The EGA module takes three inputs: encoded visual feature f^e from the contracting path, decoded predicted feature $\hat{f}^d$ from a higher layer in the expanding path, and high-frequency feature f^l from the Laplacian operator.

Encoded visual feature. Each resultant encoding feature map f_i^e from the contracting path then undergoes a 3×3 convolutional operation to reduce the number of channels, producing a distinctive encoded visual feature denoted as $\hat{f}_i^e \in \mathbb{R}^{H_i\times W_i\times N_i}$ at the i^{th} layer. However, we use f_0^e for the first EGA module as its number of channels is already small. For convenience, we still write $\hat{f}_0^e$ instead of f_0^e.

Decoded predicted feature. The second input of the EGA at the i^{th} layer is the decoded predicted feature from a higher layer, specifically at $(i+1)^{th}$ layer, denoted as $\hat{f}_{i+1}^d \in \mathbb{R}^{H_i\times W_i\times 1}$. This decoded predicted feature, $\hat{f}_{i+1}^d$, is derived from the decoding feature f_{i+1}^d, which is the output of the EGA at the $(i+1)^{th}$ layer.

High-frequency feature. To validate the effectiveness of EGA, we opt for the Laplacian pyramid method, an efficient technique for preserving high-frequency details, namely, image edge information. It is important to note that the Laplacian operator is a second-order derivative operator primarily employed for edge detection. Nevertheless, due to its susceptibility to noise, practical application entails an initial smoothing of the original image using a Gaussian filter. This modified process is known as the Laplacian of Gaussian (LoG). To enhance computational efficiency, the LoG method is itself approximated using the Difference of Gaussian (DoG) operator. This operator essentially functions as a highpass filter, proficiently retaining the most salient

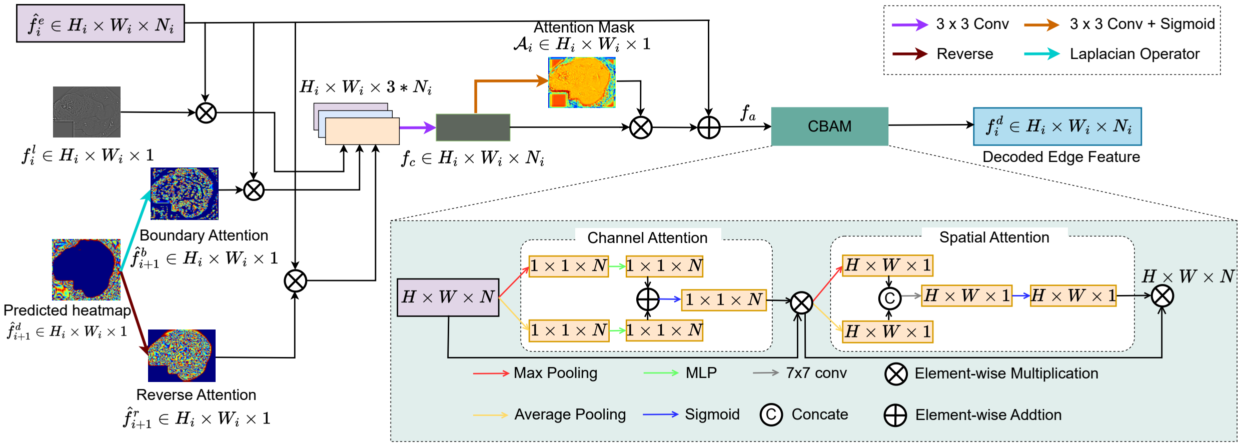

Figure 2. The architecture of our EGA block, which takes embedding feature $\hat{f}_i^e \in H_i \times W_i \times N_i$, edge information by Laplacian feature $f_i^l \in H_i \times W_i \times 1$ and higher-level predicted feature $\hat{f}_{i+1}^d \in H_i \times W_i \times 1$ as its input. H_i, W_i, N_i denote height, width, and the number of channels of the input at the layer i^{th}.

high-frequency attributes within the image. The Laplacian pyramid can be considered a sequence of cascading approximations of the DoG operator. In essence, the Laplacian pyramid encapsulates crucial low-level details across various scales.

$$\begin{aligned} I_k &= I, \text{if } k = 0. \\ I_k &= d(g(I_{k-1})), \text{if } k \geq 1. \end{aligned} \tag{1}$$

where I is the input image, g is the convolution operator with Gaussian filter, and d denotes the 2 × downsampling operation, respectively. Each level L_k of the Laplacian pyramid is attained from the Gaussian pyramid by subtracting from the current level I_k the upsampled version (u) of the smaller one I_{k+1}.

$$L_k = I_k - u(I_{k+1}). \tag{2}$$

The Laplacian operator captures second-order variations within the input image, rendering it a parameter-free and harmonious approach to extracting high-frequency details, including edges, contours, and more. These high-frequency features play a pivotal role in discerning the unique characteristics of the polyp, particularly in distinguishing it from the adjacent mucous membrane. As a result, the high-frequency feature at i^{th} level $f_i^l \in \mathbb{R}^{H_i \times W_i \times 1}$ corresponding to the level-1 image in the Laplacian pyramid is supplied to the i^{th} EGA module. In general, we denote $f^l = L_1(I)$.

This particular pyramid level is selected because, as resolution diminishes, the finer edge details undergo significant degradation owing to the repeated application of the Gaussian filter. Kindly refer to the supplementary material provided for a visual representation of the level-1 Laplacian pyramid. It's important to note that our selection of the Laplacian operator and its associated techniques serves as a proof of concept for our approach, albeit without an exhaustive practical evaluation. The decision to employ the Laplacian pyramid as the mechanism for extracting high-frequency information is grounded in its simplicity, minimal computational overhead, and the quality of information it retains. It's worth mentioning that the model's performance could potentially be enhanced through an exhaustive search for an appropriate edge detection technique, as the choice of such a technique can significantly impact the overall results.

3.1.2 EGA Procedure

At the i^{th} layer, the EGA module integrates multiple components, including the encoded visual feature $\hat{f}_i^e \in \mathbb{R}^{H_i \times W_i \times N_i}$, the predicted feature map at a higher layer $(i+1)^{th}$, represented as $\hat{f}_{i+1}^d \in \mathbb{R}^{H_i \times W_i \times 1}$ and the high-frequency feature $f_i^l \in \mathbb{R}^{H_i \times W_i \times 1}$. Generally, a Laplacian pyramid consists of high-frequency components at multiple scales. However, the high-frequency component at the i^{th} level of the Laplacian pyramid is obtained by performing Gaussian filtering, followed by a 2 × downsampling of level $(i-1)^{th}$. This process can reduce the magnitude of the high-frequency information at the i^{th} level. To preserve the high-frequency information f_i^l at every level, we propose to derive f_i^l from the base level f_0^l, which retains the high-frequency information most effectively. The formulation for calculating f_i^l is presented in Equation 3.

$$\begin{aligned} f_0^l &= f^l, \text{ where } f^l = L_1(I). \\ f_i^l &= (d(f^l))^i, \text{ if } i \geq 1. \end{aligned} \tag{3}$$

where d is 2 × downsamling and $(d(f^l))^i$ is i times 2 × downsamling, i.e., $d(d(...d(f^l)))$.

Drawing upon insights from [3], we apply a decomposition method to the higher-level predicted map $\hat{f}_{i+1}^d$, generating two distinct attention maps: i) a reverse attention map $\hat{f}_{i+1}^r$, calculated as $\hat{f}_{i+1}^r = 1 - \hat{f}_{i+1}^d$ to re-evaluate and refine the imprecise prediction map from the higher layer, and ii) a boundary attention map $\hat{f}_{i+1}^b$, derived by applying the

Laplacian operator, i.e., $\hat{f}_{i+1}^b = L_0(\hat{f}_{i+1}^d)$. Subsequently, we execute element-wise multiplication between the three attention maps, namely f_i^l, $\hat{f}_{i+1}^b$, and $\hat{f}_{i+1}^r$, with the current encoder features $\hat{f}_i^e$. This process culminates in the creation of a combined feature f_c, characterized as follows:

$$f_i^c = \text{Conv}([(f_i^l \otimes \hat{f}_i^e), (\hat{f}_{i+1}^b \otimes \hat{f}_i^e), (\hat{f}_{i+1}^r \otimes \hat{f}_i^e)]) \quad (4)$$

where [.] denotes concatenation. Recognizing that edge information could potentially encompass noise and superfluous details unhelpful for polyp segmentation; we introduce an attention mask denoted as $\mathcal{A}_i$ at level i. This mask serves the purpose of guiding the model's attention toward vital regions while simultaneously suppressing background noise and redundant information. The attention feature map at the i^{th} layer, f_i^a, is defined as follows:

$$f_i^a = \hat{f}_i^e + (f_i^c \otimes \mathcal{A}_i), \text{where } \mathcal{A}_i = \sigma(\text{Conv}(f_i^c)) \quad (5)$$

In this context, the symbol σ signifies the sigmoid function. As depicted in Figure 3, the attention masks $\mathcal{A}_i$ exhibit markedly elevated values precisely at pixels located along the edges of the polyp. In simpler terms, the fusion of deep features and Laplacian features empowers the model to prioritize the polyp's edge accurately. Building upon insights from [36], we subject the attention feature f_i^a to a `CBAM` (Convolutional Block Attention Module) for recalibration purposes. This step facilitates the capture of feature correlations between the boundary and the background region. The `CBAM` comprises two consecutive blocks: channel attention, concentrating on the channel dimension, and spatial attention, centering on the spatial dimension. In the channel attention, the attention feature map f_i^a is refined by convolution with kernels $1 \times 1 \times N_i$. In the spatial attention, spatial kernels of $H_i \times W_i \times 1$ are used. The configuration of this module is visually depicted in Figure 2. This figure uses H, W, N for a general case. Consequently, this process yields a refined decoding feature f_i^d, i.e., $f_i^d = \texttt{CBAM}(f_i^a)$.

3.2. Objective function

We employ a combination of the binary cross-entropy loss ($\mathcal{L}_{BCE}$) and the dice loss ($\mathcal{L}_{Dice}$) as our network's objective functions for training, taking into account their established efficacy as demonstrated in [38]. Consequently, the objective function for our MEGANet can be formally defined as follows:

$$\mathcal{L}_{EGA} = \sum_{i=1}^{D} \mathcal{L}_{BCE}(\hat{f}_i^d, f_i) + \mathcal{L}_{Dice}(\hat{f}_i^d, f_i) \quad (6)$$

where D is the number of decoder layers. We maintain the standard setting of 5 as inherited from U-Net. $\hat{f}_i^d$ is predicted feature map at the i^{th} decoding layer and f_i is groundtruth polyp segmentation at scale i^{th}.

4. Experiments

4.1. Datasets and Evaluation Metrics

Datasets. We evaluate our proposed MEGANet on five standard benchmark datasets: Kvasir-SEG [14], CVC-ClinicDB [1], CVC-ColonDB [30], ETIS [28] and EndoScene [34]. Note that the EndoScene [34] composes 912 images of two subsets, CVC-ClinicDB and CVC-300.

To conduct a fair comparison, we follow the same experiment setup in [6], which selects 1,450 images from Kvasir (900 images), and CVC-ClinicDB (550 images) for the training set while 798 images from Kvasir (100 images), CVC-ClinicDB (62 images), CVC-ColonDB (380 images), ETIS (196 images), and CVC-300 (60 images) for testing. This setting is challenging since the evaluation procedure is conducted across the different datasets with a wide range of resolutions (720 x 576 up to 1,920 x 1,072 in Kvasir, 384 x 288 in CVC-ClinicDB) and varied image-acquiring processes, which introduce high variance across these datasets in the size and shape of the polyps.

Evaluation Metrics. To conduct a comprehensive evaluation and comparisons with other methodologies, we follow existing SOTA approaches, employing five different metrics, i.e., mean Dice (mDice), mean IoU (mIoU), the weighted F-measure (F_β^w) [21], the structure measure (S_α) [4], the enhanced-alignment measure (E_ϕ^{max}) [5] and mean absolute error (MAE). These metrics serve the dual purpose of assessing the performance of our method in relation to ground truth labels, i.e., between prediction $\hat{f}^d$ and ground truth f, as well as facilitating comparative analysis with other existing techniques. The effectiveness of those metrics in polyp segmentation is discussed in [6, 39].

4.2. Implementation Details

We implement MEGANet using Pytorch and an NVIDIA RTX 3090. We train our network with a batch size of 16 and a general training strategy as the ACSNet [38], stochastic gradient descent (SGD) optimizer with a momentum of 0.9 and a weight decay of 1e-5. The learning rate scheduler is defined as $lr = init\ lr \times (1 - \frac{epoch}{nEpoch})^{power}$, where $init\ lr$ = 1e-3, $power$ = 0.9, $nEpoch$ = 200. We resize the input images to 352×352 for both the training and inference stages and then resize them back to the original size for calculating evaluation metrics. For data augmentation, we employ random flipping on both horizontal and vertical, random rotation, and a multi-scale training strategy $\{0.75, 1, 1.25\}$.

4.3. Performance Comparisons

Corresponding to two backbone networks, ResNet-34 [10] and Res2Net-50 [9], we qualitatively and quantitatively compare our MEGANet with eight SOTA methods, including U-Net [26], U-Net++ [40], SFA [7], PraNet [6],

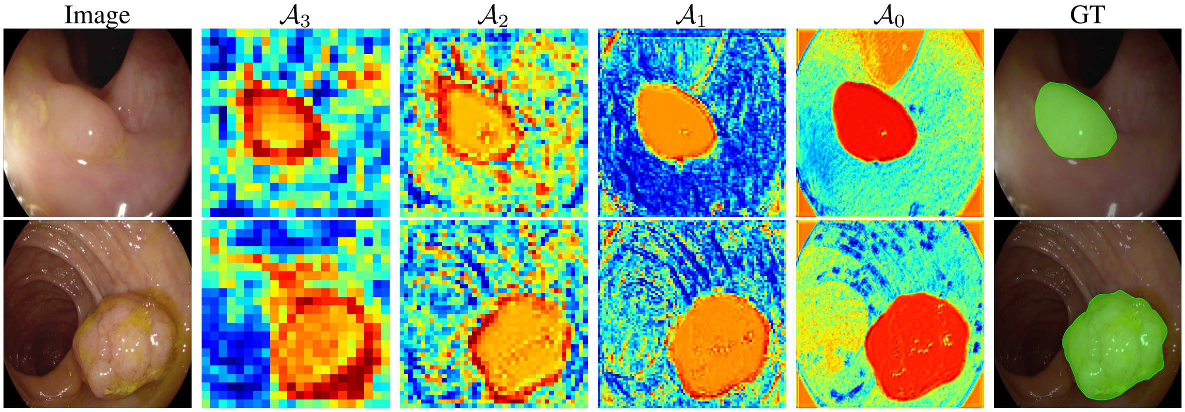

Figure 3. Heatmap visualization of attention mask $\mathcal{A}_i$ in EGA module at different i^{th} layer.

Table 1. Performance comparison between our MEGANet and other existing SOTA methods on Kvasir, CVC-300 (EndoScene), ColonDB, and ETIS datasets. The highest and second highest scores are shown in **bold** and underline, respectively. All metrics are in (%).

	Methods	Kvasir-SEG (seen)						CVC-300 (EndoScene) (unseen)					
		mDice ↑	mIoU ↑	F^w_β ↑	S_α ↑	E^{max}_ϕ ↑	MAE ↓	mDice ↑	mIoU ↑	F^w_β ↑	S_α ↑	E^{max}_ϕ ↑	MAE ↓
SOTA methods	U-Net [26]	81.8	74.6	79.4	85.8	89.3	5.5	71.0	62.7	68.4	84.3	87.6	2.2
	U-Net++ [40]	82.1	74.3	80.8	86.2	91.0	4.8	70.7	62.4	68.7	83.9	89.8	1.8
	SFA [7]	72.3	61.1	67.0	78.2	84.9	7.5	46.7	32.9	34.1	64.0	81.7	6.5
	PraNet [6]	89.8	84.0	88.5	91.5	94.8	3.0	87.1	79.7	84.3	92.5	**97.2**	1.0
	SANet [35]	90.4	84.7	89.2	91.5	95.3	2.8	88.8	81.5	85.9	$\underline{92.8}$	**97.2**	$\underline{0.8}$
	MSNet [39]	90.7	$\underline{86.2}$	89.3	**92.2**	94.4	2.8	86.9	80.7	84.9	92.5	94.3	1.0
	PEFNet [23]	89.2	83.3	–	–	–	2.9	87.1	79.7	–	–	–	1.0
	M^{2}UNet [33]	90.7	85.5	–	–	–	**2.5**	$\underline{89.0}$	$\underline{81.9}$	–	–	–	**0.7**
MEGANet(ResNet-34)		$\underline{91.1}$	85.9	$\underline{90.4}$	91.6	$\underline{95.4}$	$\underline{2.6}$	88.7	81.8	$\underline{86.3}$	92.4	95.9	0.9
MEGANet(Res2Net-50)		**91.3**	**86.3**	**90.7**	$\underline{91.8}$	**95.9**	**2.5**	**89.9**	**83.4**	**88.2**	**93.5**	$\underline{96.9}$	**0.7**

		ColonDB (unseen)						ETIS (unseen)					
		mDice ↑	mIoU ↑	F^w_β ↑	S_α ↑	E^{max}_ϕ ↑	MAE ↓	mDice ↑	mIoU ↑	F^w_β ↑	S_α ↑	E^{max}_ϕ ↑	MAE ↓
SOTA methods	U-Net [26]	51.2	44.4	49.8	71.2	77.6	6.1	39.8	33.5	36.6	68.4	74.0	3.6
	U-Net++ [40]	48.3	41.0	46.7	69.1	76.0	6.4	40.1	34.4	39.0	68.3	77.6	3.5
	SFA [7]	46.9	34.7	37.9	63.4	76.5	9.4	29.7	21.7	23.1	55.7	63.3	10.9
	PraNet [6]	70.9	64.0	69.6	81.9	86.9	4.5	62.8	56.7	60.0	79.4	84.1	3.1
	SANet [35]	75.3	67.0	72.6	83.7	87.8	4.3	$\underline{75.0}$	65.4	68.5	$\underline{84.9}$	$\underline{89.7}$	**1.5**
	MSNet [39]	75.5	67.8	73.7	83.6	88.3	4.1	71.9	66.4	67.8	84.0	83.0	2.0
	PEFNet [23]	71.0	63.8	–	–	–	**3.6**	63.6	57.2	–	–	–	$\underline{1.9}$
	M^{2}UNet [33]	76.7	68.4	–	–	–	**3.6**	67.0	59.5	–	–	–	2.4
MEGANet(ResNet-34)		$\underline{78.1}$	$\underline{70.6}$	$\underline{76.6}$	$\underline{84.5}$	**89.9**	$\underline{3.8}$	**78.9**	**70.9**	**75.3**	**86.6**	**91.5**	**1.5**
MEGANet(Res2Net-50)		**79.3**	**71.4**	**77.9**	**85.4**	$\underline{89.5}$	4.0	73.9	$\underline{66.5}$	$\underline{70.2}$	83.6	85.8	3.7

SANet [35], MSNet [39], PEFNet [23] and M^{2}UNet [33]. The result of PEFNet is adapted from the M^{2}UNet, while the others are reported based on the original papers.

Quantitative Evaluation. Table 1 presents a quantitative performance comparison between our MEGANet and other SOTA methods on the Kvasir-SEG, CVC-300 (a subset of EndoScene), ColonDB, and ETIS datasets. We provide the corresponding performance results for both MEGANet backbones, namely ResNet-34 and Res2Net-50. Like other SOTA methods, our models were trained on the Kvasir-SEG and CVC-ClinicDB training sets. The performance metrics reported on the Kvasir-SEG dataset are classified as seen, while those reported on CVC-300, ColonDB, and ETIS datasets are considered unseen. From the insights presented in Table 1, it is evident that MEGANet (Res2Net-50) excels in numerous metrics, demonstrating superior performance. Especially, our MEGANet with the backbone ResNet-34 leads the other methods by remarkable gaps on most metrics when tested on the ETIS dataset. On investigating this, we observe that the polyps in the ETIS dataset are smaller than the others. Thanks to the bounded capacity of the ResNet-34 [10] backbone, MEGANet with this backbone is prone to avoid overfitting, achieving preferable results on this dataset.

In addition to the performance evaluation, we also factor in network efficiency for a comprehensive comparison, as outlined in Table 2. This evaluation is conducted specifically on the ClinicDB dataset. Consider the ex-

Table 2. Performance and network efficiency comparison between our MEGANet with other existing SOTA methods on ClinicDB dataset. The highest and second highest scores are shown in **bold** and $\underline{\text{underline}}$, respectively. All metrics are in (%).

	Methods	ClinicDB (seen)						Backbone	Params(M)
		mDice ↑	mIoU ↑	F^w_β ↑	S_α ↑	E^{max}_ϕ ↑	MAE ↓		
SOTA methods	U-Net [26]	82.3	75.5	81.1	88.9	95.4	1.9	–	7.76
	U-Net++ [40]	79.4	72.9	78.5	87.3	93.1	2.2	–	9.0
	SFA [7]	70.0	60.7	64.7	79.3	88.5	4.2	–	–
	PraNet [6]	89.9	84.9	89.6	93.6	97.9	0.9	Res2Net-50	32.55
	SANet [35]	91.6	85.9	90.9	93.9	97.6	1.2	Res2Net-50	23.89
	MSNet [39]	92.1	87.9	91.4	$\underline{94.1}$	97.2	$\underline{0.8}$	Res2Net-50	29.74
	PEFNet [23]	86.6	81.4	–	–	–	1.0	EfficientNetV2-S	27.98
	M^{2}UNet [33]	90.1	85.3	–	–	–	$\underline{0.8}$	MetaFormer	28.77
MEGANet(ResNet-34)		$\underline{93.0}$	$\underline{88.5}$	$\underline{93.1}$	**95.0**	$\underline{98.0}$	$\underline{0.8}$	ResNet-34	29.27
MEGANet(Res2Net-50)		**93.8**	**89.4**	**94.0**	**95.0**	**98.6**	**0.6**	Res2Net-50	44.19

Table 3. Impact of each component of the EGA module on model performance. The highest scores are shown in **bold**. All metrics in (%).

Exp.	EGA				Kvasir-SEG (seen)		ClinicDB (seen)		ETIS (unseen)		CVC-300 (unseen)	
	$\hat{f}^r$	$\hat{f}^b$	f^l	CBAM	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑
#1	✗	✗	✗	✗	90.0	84.7	91.7	86.8	71.6	63.7	87.1	80.0
#2	✗	✓	✓	✓	90.7	85.3	92.1	87.5	76.9	68.7	86.3	79.4
#3	✓	✗	✓	✓	90.5	85.3	92.8	88.3	78.3	69.9	88.2	81.5
#4	✓	✓	✗	✓	**91.5**	**86.6**	92.0	87.5	75.5	67.3	86.9	80.3
#5	✓	✓	✓	✗	91.0	85.7	92.1	87.8	76.5	68.6	88.4	81.4
#6	✓	✓	✓	✓	91.1	85.9	**93.0**	**88.5**	**78.9**	**70.9**	**88.7**	**81.8**

ample of M^{2}UNet [33]: our MEGANet (ResNet-34) possesses a comparable number of network parameters, yet it outperforms M^{2}UNet with a significant 2.9% improvement in mDice and a 3.2% enhancement in mIOU. Furthermore, when contrasted with all existing SOTA methods, our MEGANet (Res2Net-50) obtains the best performance across all metrics, even with a relatively small number of network parameters.

Qualitative Evaluation. Visual comparisons of each polyp segmentation challenge are illustrated in Figure 4. In particular, (a1 and a2) highlight the complex background issue, (b1 and b2) exemplify the variability in polyp sizes and configurations, while (c1 and c2) demonstrate the challenge of dealing with indistinct boundaries. Notably, our approach reflects the capability to address various sizes and shapes within each challenge. Particularly, the results in the case of (b2 and c1) underscore the exceptional performance of our method in maintaining an impressively low false positive rate, accurately refraining from misclassifying healthy regions as tumors.

4.4. Ablation Study

As previously mentioned, the EGA module is composed of three key components: the encoded visual feature $\hat{f}^e$, the decoded predicted feature $\hat{f}^d$, which is further decomposed into reverse attention $\hat{f}^r$, boundary attention $\hat{f}^b$, and the high-frequency feature f^l. Additionally, the EGA incorporates the CBAM module. To assess the effectiveness of each input component ($\hat{f}^r$, $\hat{f}^b$, f^l) and the CBAM module within the EGA, we systematically remove each of these components as well as CBAM, conducting an ablation study on both seen datasets (Kvasir and ClinicDB) and unseen datasets (ETIS and CVC-300). The results are presented in Table 3.

Based on the empirical findings, each component within the EGA framework distinctly contributes to enhancing predictive performance. Comparing #5 to #6, we can observe the impact of the CBAM component. Focusing on #1 and #5, we can discern the effect of the combined use of $\hat{f}^r$, $\hat{f}^b$, and f^l. The comparison between #2 and #4 highlights the influence of the combined utilization of $\hat{f}^r$ and $\hat{f}^b$, which is also indicative of the influence of $\hat{f}^d$. Experiments #2 and #3 specifically isolate the effects of $\hat{f}^r$ and $\hat{f}^b$, respectively. Notably, the comparison between #4 and #6 underscores the pivotal role played by the high-frequency feature f^l. It's essential to note that the Kvasir dataset's mucous membrane (background) presents a highly intricate composition, causing the high-frequency feature f^l of the input image to contain noise. Consequently, the version without f^l attains the highest score within the Kvasir dataset.

We also conducted an ablation experiment to assess the methodology for computing high-frequency features f^l_i computation, as defined in Equation 3 by comparing it with high-frequency features f^l_i obtained from Laplacian pyramid (Equation 1, 2). In other words, we compare the performance of MEGANet when using Equation 3 and the following equation

$$f^l_i = L_i(I) = I_i - u(I_{i+1}) = I_i - u[d(g(I_i)] \qquad (7)$$

Table 4 presents the results of two scenarios: using our proposed Equation 3 and obtaining the features from the Lapla-

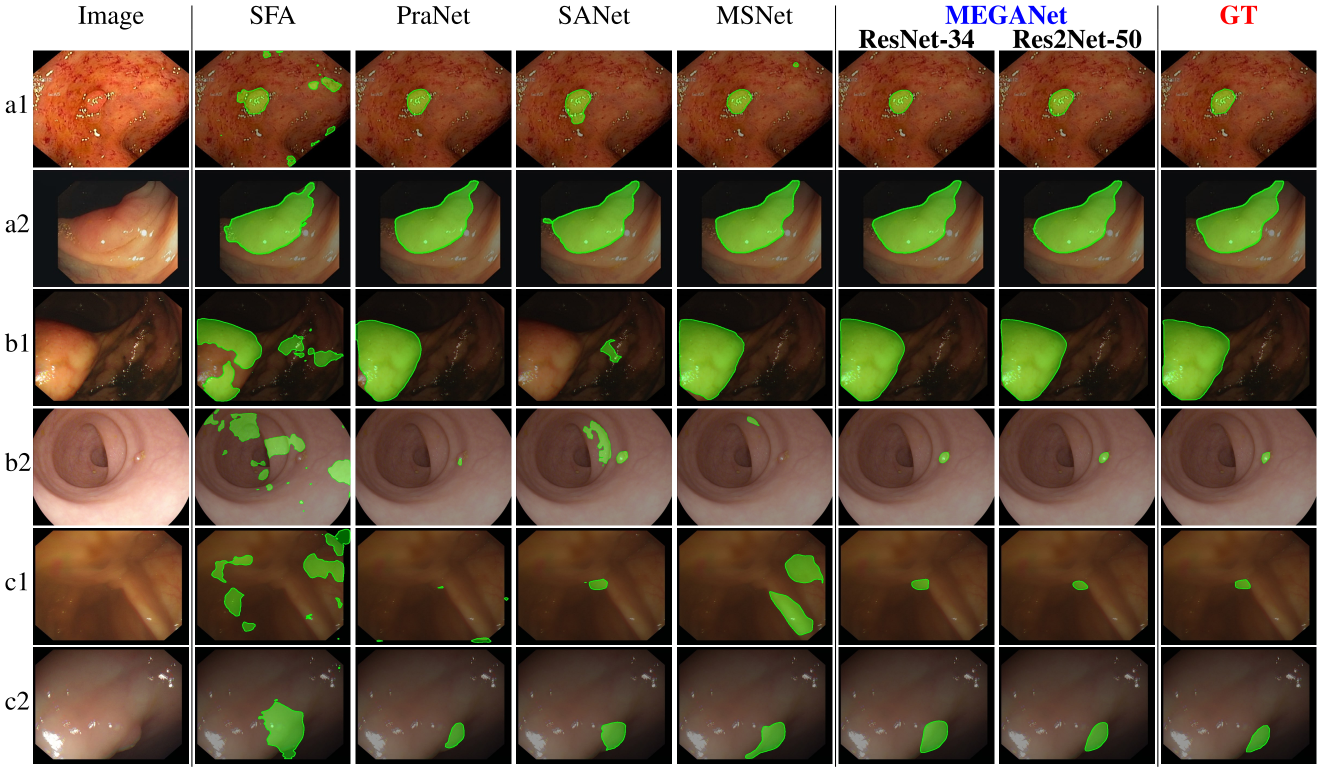

Figure 4. Qualitative comparison between our MEGANet and existing SOTA methods. The a1 image comes from the Kvasir dataset, while the a2 is from ClinicDB. The b2 image is derived from the ETIS dataset, and the rest are from the ColonDB dataset.

Table 4. Ablation study to assess the methodology for computing high-frequency features f_i^l with two formulations: our high-frequency feature (Equation 3) and Laplacian pyramid features (Equation 7). The highest scores are shown in **bold**. All metrics in (%).

f_i^l	Kvasir-SEG (seen)		ClinicDB (seen)		ETIS (unseen)		CVC-300 (unseen)	
	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑
Equation 3	**91.1**	**85.9**	**93.0**	**88.5**	**78.9**	**70.9**	**88.7**	**81.8**
Equation 7	89.9	85.8	92.6	88.1	78.0	70.2	88.0	81.4

cian pyramid as described in Equation 7.

5. Conclusion

This paper introduces a novel approach called Multi-Scale Edge-Guided Attention Network (MEGANet) for polyp segmentation. The key innovation is the integration of the Edge-Guided Attention (EGA) module, designed to retain crucial high-frequency details (such as edges) to enhance the detection of weak boundary polyp objects. Our EGA module at the i^{th} level amalgamates information from the i^{th} layer encoder, the i^{th} layer's high-frequency component, and the $(i+1)^{th}$ layer decoder. To maintain the integrity of high-frequency information, we propose deriving the high-frequency component from the base level rather than applying Gaussian filtering at each layer. Experimental results underscore the effectiveness of our MEGANet in polyp segmentation. The assessment is based on a range of metrics, including localization measures (mDice, mIoU), accuracy (F^w_β), and structural assessments (S_α, E^{max}_ϕ, MAE), all of which demonstrate the advantages of our proposed MEGANet.

Acknowledgement. Nhat-Tan Bui and Ngan Le are sponsored by the National Science Foundation (NSF) under Award No OIA-1946391 RII Track-1, NSF 1920920 RII Track 2 FEC, NSF 2223793 EFRI BRAID, NSF 2119691 AI SUSTEIN, NSF 2236302, NIH 1R01CA277739-01. Dinh-Hieu Hoang and Quang-Thuc Nguyen are funded by Vingroup Joint Stock Company and supported by the Domestic Master/ PhD Scholarship Programme of Vingroup Innovation Foundation (VINIF), Vingroup Big Data Institute (VINBIGDATA), code VINIF.2022.ThS.JVN.04 and VINIF.2022.ThS.JVN.10, respectively. Minh-Triet Tran is sponsored by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number DS2020-42-01.

References

[1] 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. *CMIG*, pages 99–111, 2015. 2, 5

[2] Nhat-Tan Bui, Hai-Dang Nguyen, Trung-Nam Bui-Huynh, Ngoc-Thao Nguyen, and Xuan-Nam Cao. Efficient loss functions for GAN-based style transfer. In *ICMV*, 2023. 2

[3] Shuhan Chen, Xiuli Tan, Ben Wang, and Xuelong Hu. Reverse Attention for Salient Object Detection. In *ECCV*, 2018. 4

[4] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-Measure: A New Way to Evaluate Foreground Maps. *ICCV*, 2017. 5

[5] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment Measure for Binary Foreground Map Evaluation. In *IJCAI*, 2018. 5

[6] 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*, 2020. 1, 2, 5, 6, 7

[7] Yuqi Fang, Cheng Chen, Yixuan Yuan, and Kai-yu Tong. Selective Feature Aggregation Network with Area-Boundary Constraints for Polyp Segmentation. In *MICCAI*, 2019. 1, 2, 5, 6, 7

[8] Xueyang Fu, Borong Liang, Yue Huang, Xinghao Ding, and John Paisley. Lightweight Pyramid Networks for Image Deraining. *IEEE Transactions on Neural Networks and Learning Systems*, 31, 2020. 2

[9] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2Net: A New Multi-Scale Backbone Architecture. *TPAMI*, 2021. 2, 5

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016. 2, 5, 6

[11] Ngoc-Vuong Ho, Tan Nguyen, Gia-Han Diep, Ngan Le, and Binh-Son Hua. Point-Unet: A Context-aware Point-based Neural Network for Volumetric Segmentation. In *MICCAI*, 2021. 1

[12] Dinh-Hieu Hoang, Gia-Han Diep, Minh-Triet Tran, and Ngan T. H Le. DAM-AL: Dilated Attention Mechanism with Attention Loss for 3D Infant Brain Image Segmentation. In *SAC*, 2022. 1, 2

[13] Fabian Isensee, Philipp Kickingereder, Wolfgang Wick, Martin Bendszus, and Klaus H. Maier-Hein. No New-Net. *arXiv preprint arXiv:1809.10483*, 2019. 1

[14] 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*, 2020. 2, 5

[15] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution. In *CVPR*, 2017. 2

[16] Ngan Le, Toan Bui, Viet-Khoa Vo-Ho, Kashu Yamazaki, and Khoa Luu. Narrow band active contour attention model for medical segmentation. *Diagnostics*, 11(8):1393, 2021. 2

[17] Ngan Le, Trung Le, Kashu Yamazaki, Toan Bui, Khoa Luu, and Marios Savides. Offset curves loss for imbalanced problem in medical segmentation. In *ICPR*, 2021. 2

[18] Joo Ho Lee, Inchang Choi, and Min H. Kim. Laplacian patch-based image synthesis. In *CVPR*, 2016. 2

[19] Shaohua Li, Xinxing Xu, Liqiang Nie, and Tat-Seng Chua. Laplacian-Steered Neural Style Transfer. In *ACMMM*, 2017. 2

[20] Jie Liang, Hui Zeng, and Lei Zhang. High-Resolution Photorealistic Image Translation in Real-Time: A Laplacian Pyramid Translation Network. In *CVPR*, 2021. 2

[21] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to Evaluate Foreground Maps. In *CVPR*, 2014. 5

[22] Tan Nguyen, Binh-Son Hua, and Ngan Le. 3D-UCaps: 3D Capsules Unet for Volumetric Image Segmentation. In *MICCAI*, 2021. 1

[23] Trong-Hieu Nguyen-Mau, Quoc-Huy Trinh, Nhat-Tan Bui, Phuoc-Thao Vo Thi, Minh-Van Nguyen, Xuan-Nam Cao, Minh-Triet Tran, and Hai-Dang Nguyen. PEFNet: Positional Embedding Feature for Polyp Segmentation. In *MultiMedia Modeling*, 2023. 2, 6, 7

[24] Trong-Hieu Nguyen-Mau, Quoc-Huy Trinh, Nhat-Tan Bui, Minh-Triet Tran, and Hai-Dang Nguyen. Multi Kernel Positional Embedding ConvNeXt for Polyp Segmentation. In *RIVF*, 2022. 1

[25] Judith MS Prewitt. Object enhancement and extraction. *Picture processing and Psychopictorics*, 1970. 2

[26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI*, 2015. 1, 2, 5, 6, 7

[27] Semra Salimoglu, Gizem Kilinc, and Bulent Calik. *Anatomy of the Colon, Rectum, and Anus*, pages 1–22. 2021. 1

[28] Juan S. Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Towards embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. *IJCARS*, pages 283–293, 2014. 2, 5

[29] Irwin Sobel and Gary Feldman. A 3x3 isotropic gradient operator for image processing. In *A Talk at The Stanford Artificial Project*, 1968. 2

[30] Nima Tajbakhsh, Suryakanth R. Gurudu, and Jianming Liang. Automated Polyp Detection in Colonoscopy Videos Using Shape and Context Information. *TMI*, pages 630–644, 2016. 2, 5

[31] Mingxing Tan and Quoc V. Le. EfficientNetV2: Smaller Models and Faster Training. *arXiv preprint arXiv:2104.00298*, 2021. 2

[32] Minh Tran, Loi Ly, Binh-Son Hua, and Ngan Le. SS-3DCAPSNET: Self-Supervised 3d Capsule Networks for Medical Segmentation on Less Labeled Data. In *ISBI*, 2022. 1

[33] Quoc-Huy Trinh, Nhat-Tan Bui, Trong-Hieu Nguyen Mau, Minh-Van Nguyen, Hai-Minh Phan, Minh-Triet Tran, and Hai-Dang Nguyen. M^{2}UNet: MetaFormer Multi-scale Upsampling Network for Polyp Segmentation. *arXiv preprint arXiv:2306.08600*, 2023. 2, 6, 7

[34] David Vázquez, Jorge Bernal, Francisco Javier Sánchez, Glòria Fernández-Esparrach, Antonio M. López, Adriana Romero, Michal Drozdzal, and Aaron C. Courville. A Benchmark for Endoluminal Scene Segmentation of Colonoscopy Images. *Journal of Healthcare Engineering*, 2017. 2, 5

[35] Jun Wei, Yiwen Hu, Ruimao Zhang, Zhen Li, S. Kevin Zhou, and Shuguang Cui. Shallow Attention Network for Polyp Segmentation. In *MICCAI*, 2021. 1, 6, 7

[36] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional Block Attention Module. In *ECCV*, 2018. 5

[37] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *CVPR*, 2022. 2

[38] Ruifei Zhang, Guanbin Li, Zhuguo Li, Shuguang Cui, Dahong Qian, and Yizhou Yu. Adaptive Context Selection for Polyp Segmentation. In *MICCAI*, 2020. 2, 5

[39] Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Automatic Polyp Segmentation via Multi-Scale Subtraction Network. In *MICCAI*, 2021. 1, 2, 5, 6, 7

[40] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. UNet++: Redesigning skip connections to exploit multiscale features in image segmentation. *TMI*, pages 1856–1867, 2020. 1, 2, 5, 6, 7