LACFormer: Toward accurate and efficient polyp segmentation

Quan Van Nguyen
nguyen.van.quan@sun-asterisk.com

Mai Nguyen
nguyen.mai@sun-asterisk.com

Thanh Tung Nguyen
nguyen.tung.thanh@sun-asterisk.com

Huy Quang Trinh
trinh.quang.huy@sun-asterisk.com

Linh Bao Doan
doan.bao.linh@sun-asterisk.com

Toan Van Pham
pham.van.toan@sun-asterisk.com

R&D Lab, Sun* Inc
Hanoi, Vietnam

Abstract

Polyp segmentation is an essential task in medical image analysis for early detection of colorectal cancer. Deep learning models, particularly encoder-decoder architectures, have been successful in polyp segmentation. However, these models often struggle to capture long-range dependencies and exhibit limited performance on small polyps. In this paper, we propose LACFormer, a novel hierarchical Transformer-CNN model incorporating the Laplacian pyramid for polyp segmentation. The proposed model combines the strengths of Transformers and CNNs along with Laplacian images to overcome the limitations of previous models. Specifically, the hierarchical Transformer backbone captures long-range dependencies and hierarchically processes the features to generate multi-scale representations. These representations are then fused with a novel CNN decoder, which enhances feature representations and refines the segmentation masks. Besides, many novel modules for effective polyp segmentation are also proposed. We evaluated our model on five popular benchmark datasets for polyp segmentation, including Kvasir, CVC-Clinic DB, CVC-ColonDB, CVC-T, and ETIS-Larib. Experimental results show that LACFormer outperforms state-of-the-art models, achieving a Dice similarity coefficient (DSC) of 0.927 and a mean intersection-over-union (mIoU) of 0.878 on CVC-ClinicDB, a DSC of 0.831 and mIoU of 0.753 on CVC-ColonDB and a DSC of 0.824 and mIoU of 0.753 on ETIS-Larib. Code is available at: https://github.com/sun-asterisk-research/LACFormer.

1 Introduction

Over 2 million colorectal cancer (CRC) cases were detected in 2020, and it is the third most prevalent cancer form globally. Over 1 million fatalities yearly are the second most frequent reason for cancer mortality. Asia has the most remarkable rate of colorectal cancer burden. The use of colonoscopy remains the conventional approach for detecting colorectal polyps. It enables the acquisition of adequate data concerning the polyps' dimensions, color, and placement, allowing medical professionals to identify and remove them before colorectal cancer can occur. Notwithstanding its benefits, colonoscopy has its limitations. Previous studies have demonstrated that as many as 18% polyps may avoid identification throughout the diagnostic process [20, 22]. The accuracy of diagnosis is contingent on the expertise and technical proficiency of the endoscopist conducting the examination. With the vigorous development of deep learning. In the past few years, several studies have applied image segmentation techniques to the segmentation of polyps with promising results. CNN-based models such as UNet [26] consist of an encoder on the left and a decoder on the right-hand side and leverage skip-connection to aggregate feature maps with several stages. In addition to the standard U-Net architecture, there exist several powerful variants of typical Unet, such as UNet++ [43], ResUNet [19], and ResUNet++ [18]. CNN-based models often exhibit a significant weakness in their limited capacity to learn global information.

Recently, there has been a significant advancement in the field of computer vision with the emergence of transformer-based models such as ViT [10], Swin-Transformer [24], and Segformer [39]. These models have demonstrated superior performance to the traditional convolutional neural network (CNN) models. In this paper, we propose a novel approach for polyp segmentation tasks by leveraging the Segformer encoder along with a newly designed head. As described in Figure 1 and Table 2, our success is mainly based on four important factors: Laplacian Pyramid, Atrous Sequential Aggregation Module (ASAM), Scale Guidance and Polarized Self-Attention (PSA). To give the model overall structure instruction, we propose using Laplacian Pyramid, which contains both high-frequency and low-frequency information. This helps ease the learning process of model through the guidance of Laplacian images. Noticed that feature map at the last stage lack of local information and bias toward global information, which is not balanced enough to produce a good output map. To solve this problem, we design an efficient and highly compatible with polyp segmentation task called Atrous Sequential Aggregation Module (ASAM). The Atrous Sequential Aggregation Module (ASAM) aims to enhance feature map produced by last stage for further generating better global map to guide next stage while still remaining efficient through leveraging large kernel depth-wise convolution layers. In order to create connection between stages, we use the Scale Guidance to link and guide each stage in model. Moreover, we also propose Soft Guidance and Absolute Guidance in the domain of Scale Guidance to make stages 3 and 4 more robust and harmonize the affection of these last two global maps. Lastly, we utilize Polarize Self-Attention (PSA) to emphasize important regions that need to attend by using parallelly or sequentially spatial and channel re-weighting technique.

The contributions of this work are summarized as follows:

- We propose a **L**aplacian **A**trous **C**ascaded Transformer(LACFormer) model for polyp segmentation task that is capable of effectively capturing polyps of various sizes.

- We design a novel decoder head suitable for hierarchical encoder transformer architecture with newly developed modules: Atrous Sequential Aggregation Module (ASAM), and Scale Guidance.

- Our proposed LACFormer improves the SOTA performance on CVC-ClinicDB, CVC-ColonDB, and ETIS-LaribLarib by 1%, 2% and 2.9% respectively.

2 Related Works

2.1 Polyp segmentation

Deep learning has seen significant advancement over the past decade and has seen real-world applications. It also plays an assisting role in medical diagnosis, including polyp segmentation. UNet [26] is a remarkable medical image segmentation deep learning model that uses a CNN encoder-decoder architecture. Inspired by UNet, many polyp segmentation models also employ the encoder-decoder architecture such as: [43], [18], [16]. These models focus on enhancing feature fusion at different scales to achieve better performance. PraNet [12] revolve around reversed attention mechanism to better distinguish polyp and its surrounding mucosa. Other methods [4], [40] use combine Transformer with CNN to achieve stronger representation for polyp, thus enhancing predicting performance. LAPFormer [25] builds a lightweight CNN decoder on top of a Transformer encoder with proper feature connections to achieve a light model but promising results.

2.2 Attention mechanism

Attention mechanisms in deep learning are inspired by the human perception process. The intuition behind the mechanism is that not all the features are equally informative. By selectively focusing on relevant features the performance should be beneficial. This is usually done by computing weights for features. The high values of weights indicate the features should be considered in the following computation. In contrast, low-weighted features have little contribution to the later computation.

Attention mechanisms have become popular and applied in various vision tasks including image classification [10], object detection [2], and semantic segmentation [12]. We categorize these mechanisms into four types: spatial attention, channel attention, mixed attention, and self-attention. Jaderberg et al. [17] utilize Spatial Transformer module providing spatial attention capability to increase computational efficiency. Hu et al. [15] propose SE block to perform channel attention by considering dependencies between feature channels. Spatial and channel attention can also stack together to form mixed attention [32, 37]. Self-attention is another type of attention mechanism. Its efficiency has been demonstrated in various vision tasks [10, 24, 39].

3 Proposed Method

In this section, we describe the proposed LACFormer in detail. An overview of our model is presented in Fig 1. We adopt MiT (Mix Transformer) proposed in SegFormer [39] as our encoder. We will answer two main questions: Why do we need this, and How to use it effectively? In Section 3.1, we first introduce the laplacian images into LACFormer's training pipeline. Then in Section 3.2, we present how we come up with the idea of Atrous Sequential Aggregation Module and the detail of it. Moving to Section 3.3, we propose to utilize Polarize Self-Attention for enhancing feature reason. Finally, in Section 3.4, we generalize Scale Guidance to handle features at different stages under adaptive manner.

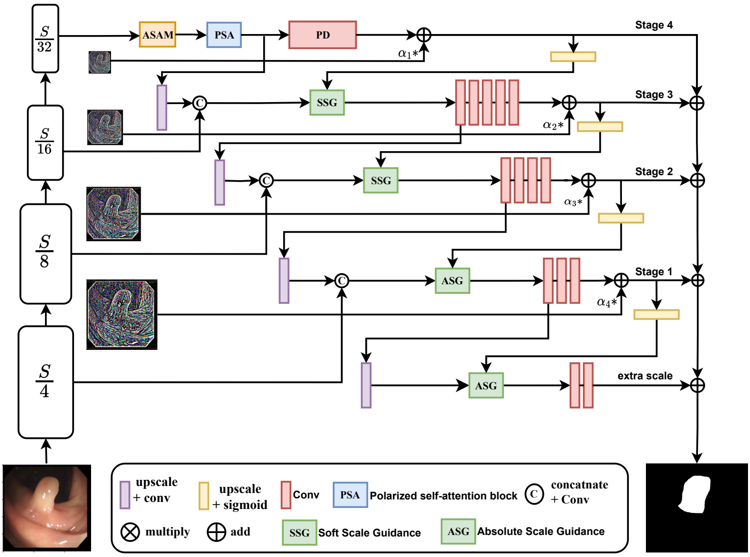


Figure 1: The architecture of proposed LACFormer. "Conv" denotes pre-activation based convolution block: BatchNorm + ReLU + Convolution. "Upscale + conv" is pre-activation based upsampling convolution block: BatchNorm + ReLU + Upsampling + Convolution. "Upscale + sigmoid" is normally an upscale operation then sigmoid.

3.1 Laplacian pyramid

Several empirical findings in research have shown that CNN models recognize objects through two types of biases: texture bias and shape bias [14]. Typically, CNN models tend to have a stronger texture bias, which is completely opposite to human behavior. Nowadays, the direction of deep learning development is to mimic human behavior and shift toward shape bias [14]. Although applying self-attention to visual tasks has partly addressed this issue, the features extracted from the encoder of vision transformer are still quite coarse. Representing image in the frequency domain helps to easily perceive and extract useful properties rather than on its raw pixels, since compressed representations in the frequency domain contain rich patterns for image understanding tasks, thus improving generalization performance [3, 30, 38]. To achieve this, we apply a Laplacian pyramid, which represents the image at different frequency levels. High-frequency levels represent texture information, while low-frequency levels represent shape information, see fig 2. Applying the Laplacian pyramid in a reasonable manner can provide a soft guide for the network, and also adding global shape to the feature map of each stage. This gives an overall picture of the input image for the model, helping to identify the areas that require attention.

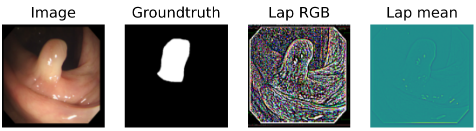


Figure 2: Laplacian image

Building a Laplacian pyramid is implemented by gradually downsampling the input to a small size and then upsampling it according to the previously downsampled levels. This allows each level of feature to capture the structural representation of the image at a different scale. Formally, let $d_i(\cdot)$ be a downsampling operator, $u_i(\cdot)$ be an upsampling operator, and i be the exponential scale factor of 2. The Laplacian pyramid is obtained as follows:

$$L_i(X) = d_i(X) - u_i(d_{i+1}(X)),\ i \in \{0, 1, ..., K\}$$

where K is the number of levels in the pyramid.

Although Laplacian Pyramid is necessary, Laplacian images often contain a significant amount of noise from high-frequency texture bias. Therefore, Laplacian images will be used with a small degree of soft-guide for the network through concatenation operator with laplacian RGB image and addition with coefficient α with the laplacian mean-channels image for the output mask at the end. Since how Laplacian images will affect the output mask is unknown, the model will decide this on its own through coefficient α rather than straightforward addition. Therefore the alpha coefficient is a learnable parameter, initialized through the following process:

```
alpha = nn.Parameter(
        torch.ones(1, dtype=torch.float32),
        requires_grad=True)
```

3.2 Atrous Sequential Aggregation Module

Inspired by [5] and [41], we want a module that can enhance features from the last scale via creating multi-scale context information which enlarges receptive field of model and captures special insight of data. Despite being relatively lightweight and useful when it can extract contextual information quite well, Pyramid Pooling Module (PPM) [41] has the disadvantage that the pooling operation of PPM can cause loss of information. Pixel-wise regression task cannot tolerate the loss of resolution caused by such large pooling operations, as the complex non-linearities associated with object edges and body parts are challenging to represent accurately in low-resolution features [28, 34].

The limitation of the Atrous Spatial Pyramid Pooling (ASPP) module [5] is rooted in the use of dilated convolution, which entails a non-continuous kernel. Although this operation can somewhat enlarge the receptive field, the non-continuous nature of the kernel seems to significantly impact the effectiveness of dense prediction tasks. The aim of the ASPP module is to capture long-range dependencies, but its efficacy is only apparent for large objects. For small objects, this approach presents a weakness with the rates of $\{6, 12, 18, 24\}$, which is further exacerbated in the polyp dataset, where small objects comprise a substantial proportion [13]. As proposed in [8], despite the convolutions in the last stage already involving

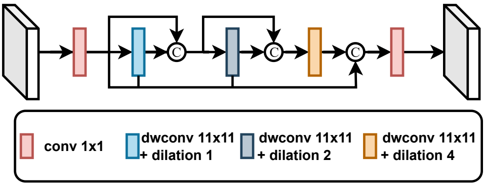

C
C
C
conv 1x1
dwconv 11x11 + dilation 1
dwconv 11x11 + dilation 2
dwconv 11x11 + dilation 4

Figure 3: Atrous Sequential Aggregation Module

a very large receptive field, increasing the kernel sizes further still results in performance improvements. Meanwhile, [35] pointed out a limitation of dilated convolution framework, so-called the "gridding" problem: Zeros padded between pixels in a convolutional kernel create a receptive field that only covers areas with checkerboard patterns, resulting in the loss of some neighboring information. The problem gets worse when increasing dilation rates, particularly in higher layers with larger receptive fields, as the sparse convolutional kernel fails to cover local information due to non-zero values being too far apart.

Synthesize ideas from [8] and [35], we propose a lightweight and efficient Atrous Sequential Aggregation Module that is highly compatible with polyp segmentation task denoted as ASAM, as illustrated in fig 3. The intuition of this module is through using large-kernel depth-wise convolution layers to extract larger and denser features and aggregate supplementary information so that adapts to polyps of diverse sizes and extraordinary shapes. Our proposed module consists of three sequential large-kernel depth-wise convolution blocks with low atrous rates $r = \{1,2,4\}$ and all depth-wise convolution use kernel 11×11, which is equal to the size of feature of the last stage. For efficiency, a conv1×1 is utilized to reduce number of channels of original feature map from $\{B,C,H,W\}$ to $\{B,\frac{C}{2},H,W\}$, then forward sequentially to three conv 11×11 blocks with different atrous rate in order of $\{1,2,4\}$. To maintain the dimension of original feature map and aggregate information extracted from atrous depth-wise convolution, the original feature map is concatenated in parallel with output of three branches above and fed into a conv 1×1.

3.3 Polarized self-attention

In order to highlight feature maps containing crucial information while suppressing redundant details after Scale Guidance, we further feed the feature to polarize self-attention module (PSA) [23]. PSA focuses on polarize filtering which aims to preserve high-resolution information in both channel and spatial attention computations while also collapsing the dimensionality of inputs along their orthogonal direction and employing non-linear functions to accurately match the output distribution of the typical granularity regression task. Basically, PSA has two variations of mechanism: sequential and parallel. We have done several experiments with both mechanisms and observed better performance with parallel PSA. In parallel framework, feature map **X** is fed simultaneously through two branches shortly called: channel branch and spatial branch

Formally, the parallel PSA mechanism is instantiated as below:

- **Channel branch** $A^{ch}(\mathbf{X}) \in R^{C\times 1\times 1}$:

$$A^{ch}(\mathbf{X}) = F_{SG}\left[\mathbf{W}_{z|\theta_1}\left(\left(\sigma_1\left(\mathbf{W}_v(\mathbf{X})\right) \times F_{SM}\left(\sigma_2\left(\mathbf{W}_q(\mathbf{X})\right)\right)\right)\right]$$

where $\mathbf{W}_z, \mathbf{W}_v, \mathbf{W}_q$ are convolution 1×1, σ_1 and σ_2 are reshape operators, $F_{SM}(\cdot)$ is Softmax operator, $F_{SG}(\cdot)$ is the Sigmoid operator and "×" is the matrix dot-product operation. The output of channel only branch is $\mathbf{Z}^{ch} = \mathbf{A}^{ch}(\mathbf{X}) \odot^{ch} \mathbf{X} \in R^{C\times H\times W}$

- **Spatial branch** $A^{sp}(\mathbf{X}) \in R^{1\times H\times W}$:

$$A^{sp}(\mathbf{X}) = F_{SG}\left[\sigma_3\left(F_{SM}\left(\sigma_1\left(F_{GP}\left(\mathbf{W}_q(\mathbf{X})\right)\right)\right) \times \sigma_2\left(\mathbf{W}_v(\mathbf{X})\right)\right)\right]$$

where $\mathbf{W}_v, \mathbf{W}_q$ are convolution 1×1, $\sigma_1, \sigma_2, \sigma_3$ are reshape operators. The output of spatial only branch is $\mathbf{Z}^{sp} = \mathbf{A}^{sp}(\mathbf{X}) \odot^{sp} \mathbf{X} \in R^{C\times H\times W}$

- Parallel mechanism is the composition of the above two branches:

$$PSA_p(\mathbf{X}) = \mathbf{Z}^{ch} + \mathbf{Z}^{sp}$$

3.4 Scale Guidance

MiT uses the hierarchical Transformer encoder architecture to obtain four stages f_i where $i \in \{1,2,3,4\}$ corresponds to four different level features at $\{1/4, 1/8, 1/16, 1/32\}$ of the original image resolution. At each stage, the partial decoder (PD) or stack of convolutional layers receives high-level feature information from PSA module and produces the global prediction segmentation map M_i where $i \in \{1,2,3,4\}$ respectively, which serves as Scale Guidance for parallel branch decoder in order to further refine the multi-scale segmentation prediction map. Particularly, global map from higher stage will give contextual guidance for lower stage with two approaches: one gives early instruction to feature before PSA module through a multiplicative operator, and another is to be fused with salient map produced by stack of convolutional layers to generate global prediction map.

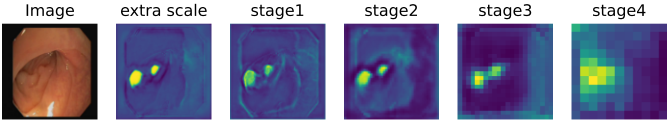
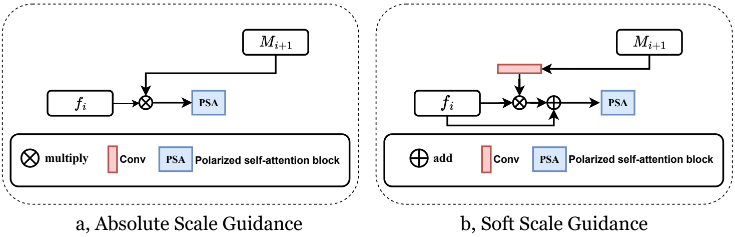


Figure 4: Global maps generated by each stage

With the early instructions, there are totally five scales as described in Fig 1, so there will be 4 steps: 4 - 3, 3 - 2, 2 - 1, 1 - Extra Scale(0). It is worth noting that guidances of 4 - 3 and 3 - 2 are different from those of 2 - 1, and 1 - Extra Scale(0). Not smooth like global map generated by stages 1 and 2, maps generated by stages 3 and 4 are very sparse and not really accurate because of tiny size and lack of non-linearity (See Fig. 4). Therefore, global map from stage 4 cannot be used to directly guide stage 3 or stage 3 to directly guide stage 2. Instead, we propose Soft Guidance to smoothen this process with skip connection and convolutional layer, and Absolute Guidance for the instruction of stages 2 and 1; this helps stabilize learning process and reduce bad behavior of global map from stages 4 and 3:


M_{i+1}
f_i
PSA
multiply
Conv
PSA
Polarized self-attention block
M_{i+1}
f_i
PSA
add
Conv
PSA
Polarized self-attention block

a, Absolute Scale Guidance

b, Soft Scale Guidance

Figure 5: Absolute Scale Guidance and Soft Scale Guidance

- **Absolute Scale Guidance** associates with PSA at stage i:

$$G_A^i(\mathbf{X}) = \mathrm{PSA}(\mathrm{Sigmoid}(M_{i+1}) \times f_i),\ i \in \{0,1\}$$

- **Soft Scale Guidance** associates with PSA at stage i :

$$G_S^i(\mathbf{X}) = f_i + \mathrm{PSA}(\mathrm{W}(\mathrm{Sigmoid}(M_{i+1}) \times f_i)),\ i \in \{2,3\}$$

where f_i is feature map at stage i, M_{i+1} is global map generated by stage $i+1$, W is convolution 1×1

4 Experiments

Dataset and Evaluation Metrics: We conduct experiments on five polyp segmentation datasets: Kvasir [19], CVC-ClinicDB [1], CVC-ColonDB [29], CVC-T [31] and ETIS-Larib Polyp DB [27]. We follow the experimental scheme mentioned in PraNet [12], and UACANet [21] which randomly extract 1450 images both from Kvasir and CVC-ClinicDB to construct a training dataset. We used the same training dataset as in PraNet and UACANet. Then we perform evaluation on the rest of Kvasir and CVC-ClinicDB. We also evaluate on CVC-ColonDB, CVC-T, and ETIS-Larib which relatively contain 380 images, 60 images and 196 images to show our model's generalization ability on unseen datasets. For performance measuring, we use mean Dice and mean IoU score as evaluation metrics for our experiments.

Table 1: Statistics of each experimental dataset

Dataset	Average Resolution	Train Samples	Test Samples
Kvasir	618 x 539	838	100
CVC-ClinicDB	384 x 288	612	62
CVC-ColonDB	574 x 500	0	380
CVC-T	574 x 500	0	60
ETIS-LaribPolypDB	1225 x 966	0	196

Implementation details: Our implementation is based on PyTorch and MMSegmentation [6] toolbox. Training is performed with $2\times$ NVIDIA RTX 3090 GPU and 64GB RAM. We used AdamW optimizer with initial learning rate of 0.0001. We resize images to 352^2 for training and testing. For data augmentations, we employ flip, slight color jittering, and cutout [7]. Our loss function is a combination of Binary Cross Entropy and Dice Loss. Our model is trained 5 times for 50 epochs with batch size of 16. Reported results are averaged over 5 runs.

4.1 Ablation Study

For ablation study, we use MiT-B4 backbone and train model for 50 epochs average over 5 runs. All results are reported under Table 2

ASAM. ASAM helps model explore polyp information in multi-view through the prism of large kernels with different dilation rates thus produce more accurate masks, as shown in Fig 6

Scale Guidance. Flexible usage of Soft Guidance and Absolute Guidance helps ease the learning process and filter out harmful features in the last two scales. Their effect of them can be seen in Fig 7

Table 2: Ablation study on each component

Methods	GFLOPs	Params (M)	Kvasir		ClinicDB		ColonDB		CVC-T		ETIS	
			mDice	mIou	mDice	mIou	mDice	mIou	mDice	mIou	mDice	mIou
w/o Laplacian Pyramid	19.87	65.1	0.923	0.872	0.925	0.855	0.808	0.730	0.889	0.820	0.813	0.740
w/o ASAM	19.81	64.64	0.921	0.869	0.920	0.861	0.819	0.750	0.888	0.814	0.809	0.727
w/o Scale Guidance	19.3	63.62	0.924	0.868	0.922	0.866	0.817	0.744	**0.901**	0.824	0.803	0.719
w/o PSA	19.09	63.32	0.920	0.872	0.915	0.861	0.810	0.731	0.884	0.817	0.811	0.740
LACFormer Head	19.87	65.1	**0.927**	**0.878**	**0.932**	**0.885**	**0.831**	**0.753**	0.892	**0.825**	**0.824**	**0.753**

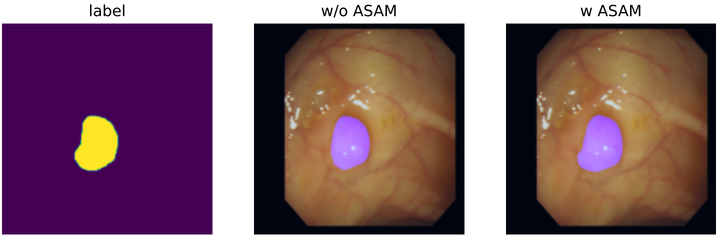


Figure 6: Impact of ASAM in LACFormer

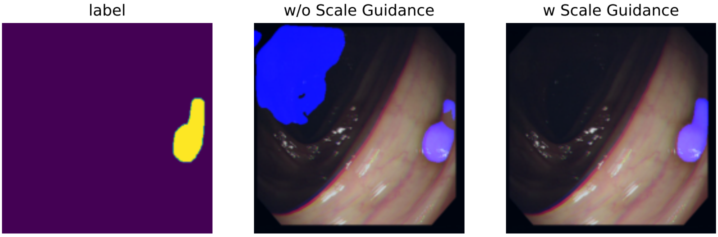


Figure 7: Impact of Scale Guidance in LACFormer

Table 3: Evaluation on different variations of LACFormer

Methods	Backbones	GFLOPs	Params (M)	Kvasir		ClinicDB		ColonDB		CVC-T		ETIS	
				mDice	mIou	mDice	mIou	mDice	mIou	mDice	mIou	mDice	mIou
LACFormer-XS	MiT-B1	6.17	17.41	0.911	0.859	0.915	0.865	0.792	0.707	0.863	0.789	0.785	0.708
LACFormer-S	MiT-B2	9.15	28.46	0.920	0.869	0.920	0.872	0.815	0.732	0.887	0.815	0.774	0.695
LACFormer-M	MiT-B3	14.59	48.33	0.921	0.870	0.924	0.865	0.819	0.733	0.887	0.815	0.804	0.724
LACFormer-L	MiT-B4	19.87	65.1	**0.927**	**0.878**	**0.932**	**0.885**	**0.831**	**0.753**	**0.892**	**0.825**	**0.824**	**0.753**

4.2 Comparison with State-of-the-Art

We compare our results with existing approaches on 5 benchmark datasets. Table 4 shows the results of SOTA methods.

Table 4: Comparison with other approaches on 5 benchmark datasets

Methods	Kvasir		ClinicDB		ColonDB		CVC-T		ETIS	
	mDice	mIou	mDice	mIou	mDice	mIou	mDice	mIou	mDice	mIou
PraNet [12]	0.898	0.840	0.899	0.849	0.709	0.640	0.871	0.797	0.628	0.567
Polyp-PVT [9]	0.917	0.864	0.937	0.889	0.808	0.727	0.900	0.833	0.787	0.706
SANet [36]	0.904	0.847	0.916	0.859	0.753	0.670	0.888	0.815	0.750	0.654
MSNet [42]	0.907	0.862	0.921	0.879	0.755	0.678	0.869	0.807	0.719	0.664
TransFuse-L* [40]	0.920	0.870	**0.942**	**0.897**	0.781	0.706	0.894	0.826	0.737	0.663
SSFormer-L [33]	0.917	0.864	0.906	0.855	0.802	0.721	0.895	0.827	0.796	0.720
ColonFormer-L [11]	0.924	0.876	0.932	0.884	0.811	0.733	**0.906**	**0.842**	0.801	0.722
LACFormer-L (Ours)	**0.927**	**0.878**	0.932	0.885	**0.831**	**0.753**	0.892	0.825	**0.824**	**0.753**

Kvasir and CVC-ClinicDB are in-domain datasets characterized by large polyp sizes. Furthermore, the polyp's colors are considered different from the backgrounds, so the performance on these two datasets is accordingly high. The same scenario also appears with CVC-T. Meanwhile, we categorize CVC-ColonDB and ETIS-Larib Polyp DB as the out-of-domain datasets with smaller polyp size and high similarity between polyps and background regions. The detection of precise polyp edges from the colon skin poses a considerable challenge in such datasets. Therefore, we adopt a novel approach by incorporating the ASAM and PSA module as a microscope to assist the model in detecting small polyps. This is the reason why our model outperforms other methods on CVC-ColonDB and ETIS-Larib Polyp DB by a large margin of relatively 2.5% and 4.8% Dice score. However, it is important to notice that our model may not outperform other approaches on CVC-ClinicDB and CVC-T. This limitation can be attributed to different factors, including differences in training strategies, data augmentations, and model architectures. We use the traditional strategy with only one output map while ColonFormer [11] uses the deep supervision training scheme for better performance on in-domain datasets. Furthermore, Transfuse-L [40] leverage two strong pretrained models ResNetV2-50 and ViT-B, and a few attention modules, which give it strength to outperform on in-domain dataset.

5 Conclusion And Future Works

In this work, we propose a novel deep neural network architecture called LACFormer for colon polyp segmentation. The proposed approach holds great potential in applications of laplacian image for medical image analysis. Together with Laplacian Pyramid, Atrous Sequential Aggregation Module and polarize self-attention also play an important role in searching and refining potential polyp regions. The experimental results on the five public datasets demonstrate the significant performance of our model compared to state-of-the-art methods.

In the future, firstly, we intend to keep improving the network and creating more effective models. Secondly, we will consider complex empirical environments such as video polyp datasets to design better models for real-world applications.

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. *Computerized medical imaging and graphics*, 43:99–111, 2015.

[2] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[3] Peter J. Burt and Edward H. Adelson. The laplacian pyramid as a compact image code. *IEEE Trans. Commun.*, 31(4):532–540, 1983. doi: 10.1109/TCOM.1983.1095851. URL https://doi.org/10.1109/TCOM.1983.1095851.

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

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[6] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020.

[7] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[8] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11963–11975, 2022.

[9] Bo Dong, Wenhai Wang, Deng-Ping Fan, Jinpeng Li, Huazhu Fu, and Ling Shao. Polyp-pvt: Polyp segmentation with pyramid vision transformers. *arXiv preprint arXiv:2108.06932*, 2021.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[11] Nguyen Thanh Duc, Nguyen Thi Oanh, Nguyen Thi Thuy, Tran Minh Triet, and Dinh Viet Sang. Colonformer: An efficient transformer based method for colon polyp segmentation. *arXiv preprint arXiv:2205.08473*, 2022.

[12] 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 *International conference on medical image computing and computer-assisted intervention*, pages 263–273. Springer, 2020.

[13] Kerr Fitzgerald and Bogdan Matuszewski. Fcb-swinv2 transformer for polyp segmentation. *arXiv preprint arXiv:2302.01027*, 2023.

[14] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

[15] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *CoRR*, abs/1709.01507, 2017. URL http://arxiv.org/abs/1709.01507.

[16] Chien-Hsiang Huang, Hung-Yu Wu, and Youn-Long Lin. Hardnet-mseg: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps. *arXiv preprint arXiv:2101.07172*, 2021.

[17] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2017–2025, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/33ceb07bf4eeb3da587e268d663aba1a-Abstract.html.

[18] 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)*, pages 225–2255. IEEE, 2019.

[19] 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 *International Conference on Multimedia Modeling*, pages 451–462. Springer, 2020.

[20] Nam Hee Kim, Yoon Suk Jung, Woo Shin Jeong, Hyo-Joon Yang, Soo-Kyung Park, Kyuyong Choi, and Dong Il Park. Miss rate of colorectal neoplastic polyps and risk factors for missed polyps in consecutive colonoscopies. *Intestinal research*, 15(3):411, 2017.

[21] Taehun Kim, Hyemin Lee, and Daijin Kim. Uacanet: Uncertainty augmented context attention for polyp segmentation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2167–2175, 2021.

[22] Jeonghun Lee, Sung Won Park, You Sun Kim, Kyung Jin Lee, Hyun Sung, Pil Hun Song, Won Jae Yoon, and Jeong Seop Moon. Risk factors of missed colorectal lesions after colonoscopy. *Medicine*, 96(27), 2017.

[23] Huajun Liu, Fuqiang Liu, Xinyi Fan, and Dong Huang. Polarized self-attention: Towards high-quality pixel-wise regression. *CoRR*, abs/2107.00782, 2021. URL https://arxiv.org/abs/2107.00782.

[24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[25] Mai Nguyen, Tung Thanh Bui, Quan Van Nguyen, Thanh Tung Nguyen, and Toan Van Pham. LAPFormer: A Light and Accurate Polyp Segmentation Transformer. *arXiv e-prints*, art. arXiv:2210.04393, 2022. doi: 10.48550/arXiv.2210.04393.

[26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[27] Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International journal of computer assisted radiology and surgery*, 9(2):283–293, 2014.

[28] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019.

[29] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*, 35(2):630–644, 2015.

[30] Róbert Torfason, Fabian Mentzer, Eiríkur Ágústsson, Michael Tschannen, Radu Timofte, and Luc Van Gool. Towards image understanding from deep compression without decoding. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HkXWCMbRW.

[31] David Vázquez, Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Antonio M López, Adriana Romero, Michal Drozdzal, and Aaron Courville. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of healthcare engineering*, 2017, 2017.

[32] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6450–6458. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.683. URL https://doi.org/10.1109/CVPR.2017.683.

[33] Jinfeng Wang, Qiming Huang, Feilong Tang, Jia Meng, Jionglong Su, and Sifan Song. Stepwise feature fusion: Local guides global. *arXiv preprint arXiv:2203.03635*, 2022.

[34] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.

[35] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. Understanding convolution for semantic segmentation. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 1451–1460. Ieee, 2018.

[36] Jun Wei, Yiwen Hu, Ruimao Zhang, Zhen Li, S Kevin Zhou, and Shuguang Cui. Shallow attention network for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 699–708. Springer, 2021.

[37] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: convolutional block attention module. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, volume 11211 of *Lecture Notes in Computer Science*, pages 3–19. Springer, 2018. doi: 10.1007/978-3-030-01234-2_1. URL https://doi.org/10.1007/978-3-030-01234-2_1.

[38] Chao-Yuan Wu, Manzil Zaheer, Hexiang Hu, R. Manmatha, Alexander J. Smola, and Philipp Krähenbühl. Compressed video action recognition. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6026–6035. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00631. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Wu_Compressed_Video_Action_CVPR_2018_paper.html.

[39] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.

[40] Yundong Zhang, Huiye Liu, and Qiang Hu. Transfuse: Fusing transformers and cnns for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 14–24. Springer, 2021.

[41] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[42] Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Automatic polyp segmentation via multiscale subtraction network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 120–130. Springer, 2021.

[43] 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*, pages 3–11. Springer, 2018.