SAM-EG: Segment Anything Model with Edge Guidance framework for efficient Polyp Segmentation

Quoc-Huy Trinh[1,2], Hai-Dang Nguyen [1,2], Bao-Tram Nguyen Ngoc[1,2]

Debesh Jha[3], Ulas Bagci[3]

Minh-Triet Tran*[1,2]

[1] University of Science, VNU-HCM Ho Chi Minh City, Vietnam

[2] Viet Nam National University, Ho Chi Minh City, Vietnam

[3] Northwestern University Chicago, Illinois, USA

Abstract

Polyp segmentation, a critical challenge in medical imaging, has prompted numerous proposed methods to enhance the quality of segmented masks. While current state-of-the-art techniques produce impressive results, these models' size and computational cost pose challenges for practical industry applications. Recently, the Segment Anything Model (SAM) has been proposed as a robust foundation model, showing promise for adaptation to medical image segmentation. Inspired by this concept, we propose SAM-EG, a framework that guides small segmentation models for polyp segmentation to address the computation cost challenge. Additionally, in this study, we introduce the Edge Guiding module, which integrates edge information into image features to assist the segmentation model in addressing boundary issues from the current segmentation model in this task. Through extensive experiments, our small models showcase their efficacy by achieving competitive results with state-of-the-art methods, offering a promising approach to developing compact models with high accuracy for polyp segmentation and in the broader field of medical imaging. The implementation of the paper can be found at https://github.com/huyquoctrinh/SAM-EG

1 Introduction

Colorectal Cancer (CRC) is one of the most perilous diseases worldwide, emerging as a prevalent affliction affecting approximately one-third of the global population. To help with efficient treatment, early diagnosis of CRC and Polyp segmentation tools are proposed to support early treatment, while this tool can help localize the polyp region.

In recent years, numerous deep-learning methods have been proposed to tackle this issue. UNet [18] initially introduced an encoder-decoder based architecture as an efficient approach for segmenting polyps in the early stages. Subsequent methods, such as ResUNet [6], ResUNet++ [9], UNet++ [32], DDANet [21], Refined-UNet [25], PEFNet [17], and M^2UNet, propose improved versions of the encoder-decoder architecture to better capture the polyp

object in endoscopic images. PraNet [7], one of the remarkable methods, employs supervision techniques to enhance the segmented polyp mask. Since then, several supervision-based methods [3, 7, 15] have been proposed, yielding positive results. State-of-the-art methods, including MetaPolyp [23], Polyp2SEG [15], MEGANet [5], and Polyp-PVT [3], demonstrate impressive results and have the potential for further enhancement through real-world data application. However, these methods encounter challenges when deployed in real-world applications due to computational costs, which can be a critical problem, due to the resource limitations of the hospital facility. For this reason, several compacted models such as ColonSegNet [11], TransResUNet [22], TransNetR [12], MMFIL-Net [16], and KDAS [24] have been proposed. Nonetheless, they also exhibit issues with boundary problems, as mentioned in MEGANet [5], which affect the learning of the model on overall polyp shape, which can affect the results when doing segmentation on small polyp objects.

Recently, the Segment Anything Model (SAM) [14] has emerged as a fundamental model for segmentation and has undergone extensive fine-tuning for medical segmentation tasks. Specifically, Med-SA [27] and SAMed [30] have showcased promising results by training SAM's image encoder with a segmentation decoder. However, SAM's computational cost and its inability to provide boundary information are suboptimal for polyp segmentation. Despite this limitation, SAM can offer rich semantic knowledge for segmentation models, as highlighted by Julka et al. [13].

To deal with previous challenges in Polyp Segmentation, we propose **SAM-EG**: **S**egment **A**nything **M**odel with **E**gde **G**uidance framework for efficient Polyp Segmentation, which is inspired by the SAM application. The primary goal of this framework is to guide the small segmentation model by using SAM semantic feature during the learning phase. By doing so, the small segmentation model can acquire semantic features from SAM, thereby enhancing its performance in the segmentation process. Moreover, we propose the Edge Guidance module (EG) to address boundary issues to capture edge information from the input image and integrate it with the learned features from segmentation and SAM embedding. As a result, the segmentation model can prioritize edge information of the polyp, enriching boundary details during the model's learning process. In summary, our contributions are in three folds:

- We propose SAM-EG, a guiding framework from the Segment Anything Model image encoder to the small model for efficient polyp segmentation.
- We introduce the Edge Guidance module (EG), to help the small model prioritize the edge information, which can leverage the boundary problem in polyp segmentation.
- We conduct extensive experiments to assess the performance and efficiency of our methods on diverse datasets, including Kvasir[10], Clinic-DB [2], Colon-DB[20], and Etis[19] dataset.

This paper is organized as follows: in Section 2, we briefly review existing methods related to this research. Then, we propose our methods in Section 3. Experiments setup are in Section 4. Results of the experiment and the discussion are in Section 5. Finally, we present the conclusion in Section 6.

2 Related Work

2.1 Polyp Segmentation

Polyp segmentation is a task aimed at segmenting polyp objects from endoscopic images, thereby assisting doctors in making more accurate decisions during early diagnosis. The initial method addressing this challenge is UNet [18], which employs an Encoder-Decoder architecture for segmentation. Subsequently, various methods based on the UNet architecture, such as UNet++[32], PEFNet [17], and M^2UNet [26], have emerged, focusing on improving feature extraction from the encoder and utilizing skip connections to address the limitations of boundary delineation in UNet. Additionally, PraNet [7] introduces Supervision Learning as a novel approach to mitigate the sharpness boundary gap between a polyp and its surrounding mucosa. Building on this concept, subsequent methods, including MSNet [31], have been proposed to enhance further the drawbacks associated with redundant feature generation at different levels of supervision. HarDNet-CPS [29] has been introduced to concentrate on the lesion area specifically. Recently, Polyp-PVT [3] has been proposed to help suppress noises in the features and improve expressive capabilities, leading to significant results in polyp segmentation. While state-of-the-art methods such as MetaPolyp [23], Polyp2SEG [15], MEGANet [5], and Polyp-PVT [3] have demonstrated impressive results, the challenge remains in implementing these approaches in products due to the heavy computational demands of large models. In response to this issue, several methods have been proposed. For instance, ColonSegNet [11] introduces a lightweight architecture to mitigate the trade-off between prediction accuracy and model size. TransNetR [12] are Transformer-based architectures that incorporate outlier detection methods, enhancing the generalization of lightweight models. Additionally, MMFIL-Net [16] proposes a lightweight model with integrated modules to address issues related to varying feature sizes in lightweight models, and KDAS [24] proposes a Knowledge Distillation framework with Attention mechanism for efficient Polyp Segmentation. While these methods show promising initial results, they still exhibit a significant performance gap compared to state-of-the-art models and the lack of learning of the boundary information, which can affect the segmentation result of the model (pointed out by MEGANet [5] and KDAS [24]).

To alleviate previous work limitations, we propose **SAM-EG**, a framework that employs the Segment Anything Model to guide the segmentation model during the learning stage. Moreover, we incorporate Edge information via the edge detector, especially by using the Sobel edge detector to help the model prioritize the edge of the polyp, thus aiding it in learning the boundary information of the polyp tissue.

2.2 Segment Anything Model (SAM)

The Segment Anything model has become a popular foundation for prompting image segmentation. However, the robustness and generality of the image encoder from SAM enable its use in various segmentation tasks, particularly in medical image segmentation. Several methods, such as SAM3D [4], SAM-UNETR[1], AFTer-SAM [28], Med-SA [27], and SAMed[30] propose methods that employ the image encoder of SAM by fine-tuning it with a decoder for segmentation. These methods have shown promising results in medical segmentation and can be further improved for real-world scenarios. However, the main challenge of using the full image encoder from SAM lies in the computational cost during the inference stage, which poses difficulties with hospital hardware requirements.

In addition, our exploration reveals that the semantic knowledge from the SAM image encoder can be transferred to the segmentation model to enhance its segmentation learning by guiding the semantic features to the segmentation model. Inspired by this discovery, SAM-EG employs the SAM image encoder to guide the segmentation model in the framework for polyp segmentation. Furthermore, it is notable that the basic features from the Segment Anything Model lack boundary features, which may hinder the segmentation model's ability to segment polyps in endoscopic images. This is why we introduce the EG module, which incorporates edge features to enhance boundary information in the image features from SAM when guiding the segmentation model. This enables our framework to address this challenge effectively.

3 Our Method

In this section, we introduce the overall SAM-EG framework, as illustrated in Figure 1. The framework consists of three main parts: the SAM image Encoder model (Sam Encoder), the Segmentation model, and the Edge Guiding Module. Within this framework, the SAM Encoder serves as a teacher model, transferring knowledge to the segmentation model, which acts as the student model. Additionally, we introduce the Edge Guiding module (EG) to address the boundary problem mentioned in MEGANet [5].

The input images X with shape $B \times 352 \times 352 \times 3$ (with B is the batch size) are processed through the SAM Encoder f_{sam} and the Segmentation Model f_{seg}. Notably, f_{sam} remains frozen during the training phase, while the f_{seg} is unfrozen. Subsequently, the global features from the SAM Encoder z_{sam} and the Segmentation model z_{seg} undergo guidance through our Edge Guiding Module f_{eg} to form two global features z_{eg}^{sam} and z_{eg}^{seg} before the knowledge transfer process begins. The f_{eg} modules are learned during the training phase.

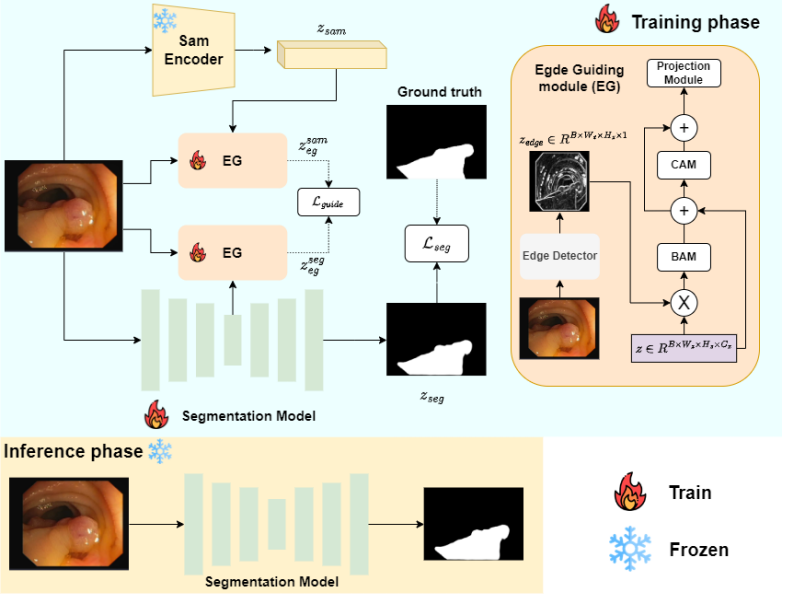

Figure 1: General SAM-EG framework

3.1 SAM Encoder

In the SAM Encoder branch, we utilize the image encoder from the Segment Anything Model (SAM), which is a pretrained using the MAE Vision Transformer [8]. This image encoder takes the input image X and generates global features $z_{sam} \in \mathbb{R}^{B\times 16\times 16\times 256}$. Leveraging this encoder allows us to benefit from a well-trained model on a large dataset, thereby producing rich-information semantic features for guiding the segmentation model.

3.2 Segmentation Model

In the Segmentation Model (f_{seg} branch), we adopt the Polyp-PVT based architecture [3] with the PVTV2-B0 backbone, which is the smallest version. Initially, the backbone encoder encodes the image X into four scale features $f_i^h \in \mathbb{R}^{\frac{W}{2^{i+1}}\times\frac{H}{2^{i+1}}\times C_i}$, where $C_i \in \{64, 128, 320, 512\}$, and $i \in \{1, 2, 3, 4\}$, with W and H being the dimensions of the input image. Afterward, the features in the last scale of the encoder are utilized for guiding with the SAM encoder, while the remaining features are forwarded to the proposed modules in the Polyp-PVT [3] architecture to obtain segmentation results. In addition, in the inference stage, only the small segmentation model is used to segment the polyp object.

3.3 Edge Guiding module (EG)

As pointed out in MEGANet [5], the main challenge of the polyp segmentation task is the boundary problem. Additionally, the guiding through SAM just helps the segmentation model learn the basis feature, which is the pattern from the polyp, not the general polyp shape. This can cause the segmentation model to lack focus on the boundary information, leading to the segmentation's bad result. To deal with this challenge, the Edge Guiding module (EG) f_{eg} is to help the segmentation learn and prioritize the boundary information of the image.

This module takes two inputs: the input images and the input feature $z \in \mathbb{R}^{B\times W_z\times H_z\times C_z}$, where W_z, H_z, and C_z represent the width, height, and channel of the input feature, respectively. The input images are initially converted to grayscale images and passed through the Edge Detector to generate the edge features. These edge features are resized to $z_{edge} \in \mathbb{R}^{B\times W_z\times H_z\times 1}$ (the effect of the different Edge Detectors is depicted in the Section 5.3). To fuse the edge feature with the image feature, we first resize the edge feature to match the shape of the image feature z, and then perform element-wise multiplication between the edge feature and the image feature z to produce the fused feature z_{fused}. Subsequently, the Boundary Attention Module (BAM) and Channel Attention Module (CAM) are applied to form the z_{bam} and z_{cam} features, which focus on vital parts of the image feature and edge feature, thus improve the prioritize the boundary information, particularly the polyp region. Moreover, the skip connection is applied between each module to mitigate the vanishing gradient. Finally, the Projection Module is utilized to transform these features into embeddings for the guiding process.

The three modules, BAM, CAM, and Projection Module, are described as follows.

Boundary Attention Module (BAM): From our experiments, we observed that the edge feature may contain noise, adversely affecting the learning process. This is the motivation behind our introducing the Boundary Attention Module (BAM) to help the model focus on the vital region of the edge feature, as depicted in Equation 1.

$$z_{bam} = z_{fused} \otimes \sigma(Conv(z_{fused})) \tag{1}$$

The fused edge feature z_{fused} is initially extracted through a convolution operation with sigmoid activation $\sigma(.)$, forming the attention mask. Subsequently, the multiplication operation between z_{fused} and the attention mask is applied to create z_{bam}. This approach enables our model to focus on the vital region of the edge feature of the polyp, instead of learning from the noisy background.

Channel Attention Module (CAM): To highlight the edges and important regions to help the model learn the general shape of the polyp from the fused results, we incorporate the Channel Attention Module (CAM). Within this block, the input from the previous stage, z_{bam}, aggregates spatial information from a feature map through both average-pooling and max-pooling operations, generating two distinct spatial context descriptors. These descriptors are subsequently fed into a shared network to produce our channel attention map, highlighting the importance. The resulting channel attention maps are then multiplied with z_{bam} to generate the output feature, z_{cam}, that focuses on the important features of the polyp, as illustrated in Equation 2.

$$z_{cam} = z_{bam} \otimes \sigma(W_1(W_0(AvgPool(z_{bam}))) + W_1(W_0(MaxPool(z_{bam})))) \tag{2}$$

where $W_0 \in \mathbb{R}^{(\frac{C}{r}) \times C}$ and $W_1 \in \mathbb{R}^{C \times (\frac{C}{r})}$, with C and r are channel and the reduction ratio, denote the weights of two MLP layers, σ denotes the sigmoid activation function. Following the first fully connected layer, the ReLU activation is also applied after multiplying W_0.

By applying this module, the model can prioritize the crucial parts from the fusion of the image feature map and the edge information, which can benefit it in learning the polyp general feature and the boundary information from the input image.

Projection Module: The primary objective of the Projection Module is to convert the final feature map, after passing through CAM, into an embedding for guiding between SAM and the Segmentation model. Given the input z_{cam}, the output of this module z_{eg} represents the fused feature with edge information. In this stage, z_{cam} undergoes several operations, including global average pooling, linear projection, batch normalization, and the non-linear ReLU activation function. These operations generate an implicit embedding representing the polyp features with shape $B \times d$ (d value denotes the embedding dimension, which is 256 in our implementation).

3.4 Objective function

The overall objective function for this framework (depicted in Equation 5) is the sum of $\mathcal{L}_{guide}$ from the knowledge transfer of SAM embedding (z_{eg}^{sam}) with segmentation embedding (z_{eg}^{seg}), and the $\mathcal{L}_{seg}$ between the segmentation prediction z_{seg} and the ground truth y. We employ the L2 Loss for the guide loss, as demonstrated in Equation 3. Meanwhile, the segmentation loss ($\mathcal{L}_{seg}$) is the combination of binary cross-entropy loss ($\mathcal{L}_{BCE}$) and the dice loss ($\mathcal{L}_{dice}$), as depicted in Equation 4.

$$\mathcal{L}_{guide} = \frac{1}{N}||z_{eg}^{seg} - z_{eg}^{sam}||_2^2 \tag{3}$$

$$\mathcal{L}_{seg} = \sum_{i}^{D} \mathcal{L}_{BCE}(z_{seg}^{i}, y^{i}) + \mathcal{L}_{dice}(z_{seg}^{i}, y^{i}) \tag{4}$$

$$\mathcal{L}_{total} = \mathcal{L}_{guide} + \mathcal{L}_{seg} \tag{5}$$

where D denotes the number of decoded layers, and N is the number of samples. In the implementation of Polyp-PVT [3], D equals 2.

By minimizing the guide loss, the segmentation model can learn the semantic feature from the SAM model. Additionally, the integration of the Edge Guiding module can help the transfer information contain and prioritize the boundary information, which benefits the segmentation model.

4 Experiment

4.1 Datasets

To conduct the fair comparison, the experiment's dataset follows the merged dataset from the PraNet [7] experiment for the training stage which includes 900 samples from Kvasir-SEG [10] and 550 samples from CVC-ClinicDB [7]. The remaining images of Kvasir-SEG [10] and CVC-ClinicDB [7] with two unseen datasets such as ColonDB [20], and ETIS [19] are used for benchmarking our method.

4.2 Implementation Detail

In our implementation, we conducted experiments based on the Polyp-PVT baseline [3]. We used the PyTorch framework and a Tesla H100 80GB for training. The images were resized to 352×352, and a batch size of 16 was set. The AdamW optimizer was employed with a learning rate of $1e-4$, and weight decay value is $1e-4$. The best weights were obtained after 100 epochs, with a total training time of approximately 2 hours. During testing, the images were resized to 352×352.

4.3 Evaluation Metrics

Two commonly used evaluation metrics, mean Dice (mDice) and mean Intersection over Union (mIoU), are employed for assessment. Higher values for both mDice and mIoU indicate better performance. In the context of polyp segmentation, the mDice metric holds particular significance as it is considered the most important metric for determining a model's effectiveness. Moreover, we also compare the number of parameters and FLOPs values to illustrate the efficiency and light-weight of our method.

4.4 Performance Comparisons

To assess its effectiveness, we compare it with state-of-the-art methods and real-time approaches.

Comparison with State-of-the-art To assess the effectiveness of our model, we compare our PVTV0 distilled model (approximately 3.7 million parameters) with several methods with a higher parameter count. These include UNet [18], UNet++ [32], PraNet [7], MSNet [31], Polyp-PVT [3],PEFNet [17], M^2UNet [26], and HarDNet-CPS [29]. In this comparison, we also provide information about the number of parameters to evaluate the impact of model size on overall performance. As the datasets used in PEFNet [17] differ, we retrain this method in our dataset setting, which follows PraNet[7] for a fair and consistent comparison. Moreover, in the Polyp-PVT setting, we reproduce the training and testing experiments with the backbone PVTV2-B0 for a fair comparison with our method.

Comparison with real-time methods: To evaluate the performance of our small model in terms of both prediction accuracy and computational efficiency, we conducted comparisons with several methods, namely ColonSegNet[11], TransNetR[12], MMFIL-Net[16], and KDAS [24]. Except for MMFIL-Net, we reproduced the training processes for the remaining models using the same training dataset and testing dataset as our models. The comparison weights were carefully selected to ensure a fair and meaningful comparison.

5 Result

5.1 Qualitative Comparison

Comparison with State-of-the-art: As shown in Table 1, SAM-EG enhances the performance of the small model across all benchmark datasets despite having the lowest number of parameters. Particularly, in the ColonDB dataset, our method surpasses by $+1.9$ and $+1.1$ in mDice and mIoU, respectively. In the ETIS dataset, our method outperforms by $+3.8$ and $+1.7$ in mDice and mIoU, respectively. These findings highlight the generalization ability of our small model in both familiar and unfamiliar domains despite its substantially lower number of parameters than existing methods. Additionally, these results highlight the ability of the model to localize the small polyp objects, which is a limitation of previous works. This observation underscores the promising results of our contribution to creating a small model, demonstrating its accuracy in predictions comparable to that of larger models.

Method	Params(M)	ClinicDB		ColonDB		Kvasir		ETIS	
		mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
UNet (2015) [18]	7.6	0.824	0.767	0.519	0.449	0.821	0.756	0.406	0.343
UNet++ (2018) [32]	9.0	0.794	0.729	0.483	0.410	0.820	0.743	0.401	0.344
PraNet (2019) [7]	32.6	0.899	0.849	0.712	0.640	0.898	0.840	0.628	0.567
MSNet (2021) [31]	29.7	0.921	0.879	0.755	0.678	0.907	**0.862**	0.719	0.664
Polyp-PVT (2022) [3]	3.7	0.924	0.867	0.756	0.668	0.894	0.835	0.747	0.675
PEFNet (2023) [17]	28.0	0.866	0.814	0.710	0.638	0.892	0.833	0.636	0.572
M^2UNet (2023) [26]	28.7	0.901	0.853	0.767	0.684	0.907	0.855	0.670	0.595
HarDNet-CPS (2023) [29]	−−	0.917	**0.887**	0.729	0.658	0.911	0.856	0.69	0.619
SAM-EG	**3.7**	**0.931**	0.879	**0.774**	**0.689**	**0.915**	**0.862**	**0.757**	**0.681**

Table 1: Qualitative results of **SAM-EG** on various datasets

Comparison with real-time methods: As depicted in In Table 2, SAM-EG outperforms previous real-time models across four different datasets and in all metrics. Despite having similar numbers of parameters and FLOPs values as KDAS, our method gradually achieves better performance on all four datasets. Particularly notable in the ColonDB dataset, our method surpasses the second-best method by $+1.5$ and $+1.0$ in mDice and mIoU metrics, respectively. These results indicate that the strength of the SAM image encoder can benefit the learning of the segmentation model, enabling it to address the challenges of polyp segmentation and achieve competitive results.

Method	Params(M)	FLOPs (G)	ClinicDB		ColonDB		Kvasir		ETIS	
			mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
ColonSegNEt (2021) []	5.0	6.22	0.28	0.21	0.12	0.10	0.52	0.39	0.16	0.12
TransNetR (2023) []	27.3	10.09	0.87	0.82	0.68	0.61	0.87	0.80	0.60	0.53
MMFIL-Net (2023) []	6.7	4.32	0.890	0.838	0.744	0.659	0.909	0.858	0.743	0.670
KDAS (2024) []	3.7	2.01	0.925	0.872	0.759	0.679	0.913	0.848	0.755	0.677
SAM-EG	**3.7**	**2.12**	**0.931**	**0.879**	**0.774**	**0.689**	**0.915**	**0.862**	**0.757**	**0.681**

Table 2: Comparison of model from **SAM-EG** with real-time model

5.2 Qualitative visualization

Figure 2 illustrates a comparison of segmentation between SAM-EG and other methods, both state-of-the-art and real-time. The visualization demonstrates that our method effectively identifies difficult and tiny polyps from endoscopic images. This indicates that by integrating guidance from SAM and learning from edge information, our model can capture boundary information, thus achieving promising results.

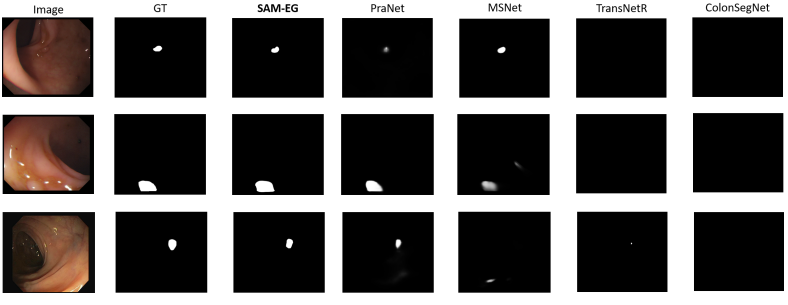

Figure 2: Visualization comparison between SAM-EG with several methods

5.3 Ablation Studies

The effect of SAM Guiding and EG Module: To assess the effectiveness of the SAM Guiding and EG Module, we conducted experiments in three scenarios. The first scenario involved training Polyp-PVT with backbone PVTV2-B0 without using SAM for guiding or the Edge Guiding Module. SAM was used for guidance in the second scenario, but the EG module was not included. Features from the SAM and segmentation models were projected into two embeddings for guidance via the $\mathcal{L}_{guide}$. The final scenario represents our full SAM-EG approach. Results are summarized in Table 3. We observe a gradual improvement from the results by incorporating SAM Guiding and the Edge Guiding module, highlighting the importance of edge information and the benefits of guiding segmentation with semantic features from the foundational model.

Method	ColonDB		Kvasir	
	mDice	mIoU	mDice	mIoU
Baseline	0.756	0.668	0.894	0.835
Sam Guiding + w/o EG	0.766	0.673	0.899	0.841
SAM-EG	**0.774**	**0.689**	**0.915**	**0.862**

Table 3: Effect of SAM guiding and the EG module

Effect of Edge Detector: To assess the effectiveness of the Edge Detector within the entire pipeline, we conducted experiments to compare results using three different detectors: Canny, Laplacian, and Sobel. The outcomes are summarized in Table 4, where the Sobel detector yielded the best performance. This suggests that the Sobel Edge detector is well-suited for the task of polyp segmentation. The reason behind this result is due to the lower sensitivity to noise of the Sobel operator, which allows its edge information to emphasize the edges of polyp tissue more effectively.

Method	ColonDB		Kvasir	
	mDice	mIoU	mDice	mIoU
Canny detector	0.772	0.687	0.898	0.842
Laplacian detector	0.772	0.689	0.901	0.853
Sobel detector	**0.774**	**0.689**	**0.915**	**0.862**

Table 4: Comparison for the effect of the Edge Detector to the segmentation results

6 Conclusion

In conclusion, we introduce SAM-EG, a framework that guides a small model using the Segment Anything Model (SAM) along with edge information to address the computational costs and boundary challenges in polyp segmentation for real-world applications. By the benefit of guiding semantic features from SAM, the segmentation model can improve the segmentation result while incorporating edge information, our segmentation model learns boundary details, enabling it to address challenges posed by difficult polyps such as tiny tissues or hard-to-identify areas. Our method achieves competitive results through extensive experiments compared to state-of-the-art and real-time models despite employing a small number of parameters and FLOPs. These results demonstrate the potential of our approach for real-world applications.

In the future, we aim to explore more efficient mechanisms and encourage researchers to delve into this topic, which could greatly benefit the integration of polyp segmentation applications into clinical environments.

Acknowledgment: Minh-Triet Tran is supported by Viet Nam National University Ho Chi Minh City (VNU-HCM) under grant number DS2020-42-01. Debesh Jha and Ulas Bagci are supported by NIH funding: R01-CA246704, R01-CA240639, U01-DK127384-02S1, and U01-CA268808.

References

[1] Jesus Alejandro Alzate-Grisales, Alejandro Mora-Rubio, Francisco García-García, Reinel Tabares-Soto, and Maria De La Iglesia-Vayá. Sam-unetr: Clinically significant prostate cancer segmentation using transfer learning from large model. *IEEE Access*, 11:118217–118228, 2023.

[2] 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.

[3] Dong Bo, Wang Wenhai, Fan Deng-Ping, Li Jinpeng, Fu Huazhu, and Shao Ling. Polyp-PVT: Polyp Segmentation with PyramidVision Transformers. *CAAI AIR*, 2023.

[4] Nhat-Tan Bui, Dinh-Hieu Hoang, Minh-Triet Tran, and Ngan Le. Sam3d: Segment anything model in volumetric medical images. *arXiv preprint arXiv:2309.03493*, 2023.

[5] Nhat-Tan Bui, Dinh-Hieu Hoang, Quang-Thuc Nguyen, Minh-Triet Tran, and Ngan Le. Meganet: Multi-scale edge-guided attention network for weak boundary polyp segmentation. In *WACV*, pages 7985–7994, 2024.

[6] Foivos I Diakogiannis, François Waldner, Peter Caccetta, and Chen Wu. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2020.

[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 *MICCAI*, 2020.

[8] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[9] 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 *ISM*, 2019.

[10] 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 *Multimedia Modeling*, 2020.

[11] Debesh Jha, Sharib Ali, Nikhil Kumar Tomar, Håvard D Johansen, Dag Johansen, Jens Rittscher, Michael A Riegler, and Pål Halvorsen. Real-time polyp detection, localization and segmentation in colonoscopy using deep learning. *Ieee Access*, 9:40496–40510, 2021.

[12] Debesh Jha, Nikhil Kumar Tomar, Vanshali Sharma, and Ulas Bagci. Transnetr: Transformer-based residual network for polyp segmentation with multi-center out-of-distribution testing. *arXiv preprint arXiv:2303.07428*, 2023.

[13] Sahib Julka and Michael Granitzer. Knowledge distillation with segment anything (sam) model for planetary geological mapping. In *International Conference on Machine Learning, Optimization, and Data Science*, pages 68–77. Springer, 2023.

[14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[15] Vittorino Mandujano-Cornejo and Javier A. Montoya-Zegarra. Polyp2seg: Improved polyp segmentation with vision transformer. In *MICCAI*, 2022.

[16] Usman Muhammad, Zhangjin Huang, Naijie Gu, et al. Mmfil-net: Multi-level and multi-source feature interactive lightweight network for polyp segmentation. *Displays*, page 102600, 2023.

[17] 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.

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI*, 2015.

[19] 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.

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

[21] Nikhil Kumar Tomar, Debesh Jha, Sharib Ali, Håvard D. Johansen, Dag Johansen, Michael A. Riegler, and Pål Halvorsen. Ddanet: Dual decoder attention network for automatic polyp segmentation. In *Pattern Recognition. ICPR International Workshops and Challenges*, 2021.

[22] Nikhil Kumar Tomar, Annie Shergill, Brandon Rieders, Ulas Bagci, and Debesh Jha. Transresu-net: Transformer based resu-net for real-time colonoscopy polyp segmentation. *arXiv preprint arXiv:2206.08985*, 2022.

[23] Q. Trinh. Meta-polyp: A baseline for efficient polyp segmentation. In *2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 742–747. IEEE Computer Society, 2023.

[24] Quoc-Huy Trinh. Kdas3: Knowledge distillation via attention supervision, and symmetrical structure guiding for polyp segmentation. *arXiv preprint arXiv:2312.08555*, 2023.

[25] Quoc-Huy Trinh, Minh-Van Nguyen, Thiet-Gia Huynh, and Minh-Triet Tran. HCMUS-Juniors 2020 at Medico Task in MediaEval 2020: Refined Deep Neural Network and U-Net for Polyps Segmentation. In *MediaEval*, 2020.

[26] Quoc-Huy Trinh, Nhat-Tan Bui, Trong-Hieu Nguyen-Mau, Minh-Van Nguyen, Hai-Minh Phan, Minh-Triet Tran, and Hai-Dang Nguyen. M2unet: Metaformer multi-scale upsampling network for polyp segmentation. In *2023 31st European Signal Processing Conference (EUSIPCO)*, pages 1115–1119. IEEE, 2023.

[27] Junde Wu, Rao Fu, Huihui Fang, Yuanpei Liu, Zhaowei Wang, Yanwu Xu, Yueming Jin, and Tal Arbel. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023.

[28] Xiangyi Yan, Shanlin Sun, Kun Han, Thanh-Tung Le, Haoyu Ma, Chenyu You, and Xiaohui Xie. After-sam: Adapting sam with axial fusion transformer for medical imaging segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7975–7984, 2024.

[29] Tong Yu and Qingxiang Wu. Hardnet-cps: Colorectal polyp segmentation based on harmonic densely united network. *Biomedical Signal Processing and Control*, 2023.

[30] Kaidong Zhang and Dong Liu. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*, 2023.

[31] Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Automatic Polyp Segmentation via Multi-scale Subtraction Network. In *MICCAI*, 2021.

[32] 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*, 2018.