LS-PolypSeg: A Parameter-Efficient LoRA-Enhanced SAM3 Framework for Polyp Segmentation

1st Sanjana Jhansi Ganji
Department of Artificial Intelligence and Machine Learning
Chaitanya Bharathi Institute of Technology (CBIT)
Hyderabad, India
sanjanajhansi99@gmail.com

2nd Panigrahi Srikanth
Department of Artificial Intelligence and Machine Learning
Chaitanya Bharathi Institute of Technology (CBIT)
Hyderabad, India
srikanth.panigrahi@gmail.com

3rd Kaushal Sambanna
Department of Artificial Intelligence and Machine Learning
Chaitanya Bharathi Institute of Technology (CBIT)
Hyderabad, India
kaushal.sambanna@gmail.com

4th Routhu Srinivasa Rao
CureBay
Acharya Vihar
Bhubaneswar, Odisha, India
routh.srinivas@gmail.com

***Abstract*—Accurate polyp segmentation is essential for early detection of colorectal cancer, where precise delineation of lesion boundaries directly impacts clinical decision-making. Despite significant progress, existing convolutional methods often struggle to capture global contextual information, while transformer-based and foundation models introduce high computational complexity and require extensive fine-tuning. In this work, we propose LS-PolypSeg, a parameter-efficient framework that leverages a pretrained SAM3 vision encoder with Low-Rank Adaptation (LoRA) for domain-specific learning. By selectively adapting key transformer layers while keeping most of the encoder frozen, the proposed approach preserves rich pretrained representations while significantly reducing training overhead. To complement global feature extraction, a lightweight UNet-style decoder performs multi-scale feature fusion, enabling accurate recovery of fine-grained spatial details. Extensive experiments on three benchmark datasets, namely Kvasir-SEG, CVC-ClinicDB, and BKAI-IGH, demonstrate that LS-PolypSeg achieves competitive and state-of-the-art performance across multiple evaluation metrics. These results highlight the effectiveness of combining foundation model representations with efficient adaptation and hierarchical decoding for robust, scalable polyp segmentation.**

***Index Terms*—Polyp segmentation, Medical image segmentation, Vision Transformers, SAM3, Low-Rank Adaptation (LoRA), Parameter-efficient learning, Multi-scale feature fusion**

I. Introduction

Colorectal cancer is a leading cause of cancer-related mortality worldwide, where early detection and removal of polyps are critical for improving survival rates. Colonoscopy is the standard diagnostic procedure; however, manual inspection is time-consuming, operator-dependent, and prone to missing subtle lesions. Consequently, automated polyp segmentation has become an essential component of computer-aided diagnosis (CAD) systems for precise lesion delineation.

Recent deep learning approaches have significantly advanced polyp segmentation. Convolutional neural networks (CNNs), such as U-Net and its variants, effectively capture local spatial features, while attention-based and multi-scale methods improve robustness under challenging conditions. More recently, transformer-based and hybrid architectures have been introduced to model global context. In parallel, foundation models such as the Segment Anything Model (SAM) and its medical variants (e.g., MedSAM and MedSAM3) have demonstrated strong generalization. However, CNN-based models often lack global context, transformer-based methods incur high computational cost, and foundation models typically require extensive fine-tuning or prompt engineering, limiting practical deployment.

To address these challenges, we propose **LS-PolypSeg**, a hybrid framework that combines a pretrained SAM3 encoder with efficient task-specific adaptation. We employ Low-Rank Adaptation (LoRA) to selectively adapt key transformer layers while preserving most pretrained parameters. In addition, a lightweight UNet-style decoder fuses multi-scale features and recovers fine-grained spatial details for accurate segmentation.

The main contributions of this work are summarized as follows:

- We propose LS-PolypSeg, a hybrid framework integrating a pretrained SAM3 encoder with LoRA-based adaptation and a multi-scale UNet decoder for polyp segmentation.
- We introduce a parameter-efficient adaptation strategy that updates low-rank components of transformer layers, enabling effective domain adaptation while preserving pretrained knowledge.
- We conduct extensive experiments on three benchmark datasets (Kvasir-SEG, CVC-ClinicDB, and BKAI-IGH),

where the proposed method demonstrates competitive and state-of-the-art performance compared to existing approaches.

II. Related Work

Polyp segmentation plays a crucial role in computer-aided diagnosis for colorectal cancer, where accurate delineation of lesion boundaries is essential. Early deep learning approaches based on convolutional neural networks, such as U-Net, DoubleU-Net [1], and TransResU-Net [2], demonstrated strong capability in capturing local spatial features and multi-scale context. Subsequent models, including GMSRF-Net [3], MNet-SAt [4], and PAM-UNet [5], improved performance through attention mechanisms and multi-scale feature fusion, enhancing robustness under challenging imaging conditions. More recently, hybrid architectures combining convolutional networks and transformers have been proposed to model local and global dependencies [6]. In this direction, Fuse-Polyp [7] introduced a dual-encoder framework that integrates transformer-based semantic modeling with convolutional feature extraction for precise boundary delineation, along with an effective cross-modal fusion strategy. In parallel, foundation models such as the Segment Anything Model (SAM) and its medical variants, including MedSAM2 [8] and MedSAM3 [9], have demonstrated strong generalization across diverse medical imaging tasks. However, these models often require extensive fine-tuning, large annotated datasets, or prompt-based inputs, limiting efficiency and clinical applicability. Additionally, semi-supervised and multimodal approaches, such as depth-guided mutual learning frameworks [10] and vision-language models [11], reduce annotation dependency but introduce additional complexity or require auxiliary modalities that may not always be available.

To address these challenges, parameter-efficient adaptation techniques have emerged as a promising direction, particularly Low-Rank Adaptation (LoRA) [12], which enables efficient fine-tuning of large transformer-based models by updating only a small subset of parameters. Recent SAM-based approaches employing lightweight adaptation strategies, such as PP-SAM [13] and SAM-Mamba [14], have demonstrated improved robustness and generalization with reduced computational overhead. Nevertheless, existing methods either modify the backbone architecture or still involve considerable training complexity. Motivated by these limitations, we propose **LS-PolypSeg**, a parameter-efficient framework that adapts the SAM3 vision encoder using LoRA and a lightweight UNet-style decoder for segmentation. By freezing the pretrained encoder and selectively injecting low-rank adapters into key transformer layers, the proposed method significantly reduces trainable parameters while preserving rich multi-scale feature representations. This design bridges the gap between foundation-model generalization and the efficiency required for clinical deployment, enabling accurate and scalable polyp segmentation without additional modalities or extensive fine-tuning.

III. Methodology

A. Overview

We propose **LS-PolypSeg**, shown in Fig. 1, a parameter-efficient segmentation framework that leverages a pretrained SAM3 vision encoder augmented with Low-Rank Adaptation (LoRA) and a lightweight UNet-style decoder. The overall architecture consists of two stages: (i) a *parameter-efficient feature encoding stage* using a frozen SAM3 encoder with LoRA adapters, and (ii) a *multi-scale decoding stage* that progressively fuses hierarchical features to produce the final segmentation mask.

Given an input image $\mathbf{X} \in \mathbb{R}^{H\times W\times 3}$, the goal is to predict a segmentation map $\mathbf{Y} \in \mathbb{R}^{H\times W}$.

B. SAM3 Encoder with LoRA Adaptation

The input image is first resized to a fixed resolution of 1008×1008 to match the pretrained SAM3 configuration:

$$\mathbf{X}_{enc} = \mathcal{R}(\mathbf{X}), \tag{1}$$

where $\mathcal{R}(\cdot)$ denotes bilinear interpolation.

The resized image is then passed through the SAM3 vision encoder, which consists of a Vision Transformer (ViT) backbone with positional encoding and a feature pyramid network (FPN) neck:

$$\{\mathbf{F}_0, \mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3\} = \mathcal{E}_{SAM3}(\mathbf{X}_{enc}), \tag{2}$$

where $\mathbf{F}_i \in \mathbb{R}^{H_i\times W_i\times C}$ represent multi-scale feature maps with $C = 256$ channels and progressively decreasing spatial resolutions.

To achieve parameter efficiency, we freeze the pretrained SAM3 weights and introduce Low-Rank Adaptation (LoRA) modules within the transformer layers. Specifically, for a given linear transformation $\mathbf{W} \in \mathbb{R}^{d\times k}$, LoRA decomposes the update as:

$$\mathbf{W}' = \mathbf{W} + \Delta\mathbf{W}, \quad \Delta\mathbf{W} = \mathbf{BA}, \tag{3}$$

where $\mathbf{A} \in \mathbb{R}^{r\times k}$ and $\mathbf{B} \in \mathbb{R}^{d\times r}$ with $r \ll \min(d, k)$.

The adapted transformation becomes:

$$\mathbf{y} = \mathbf{W}\mathbf{x} + \alpha \cdot \mathbf{BA}\mathbf{x}, \tag{4}$$

where α is a scaling factor. In our implementation, LoRA is applied to the attention and feed-forward layers, including *qkv*, *projection*, *fc1*, and *fc2* modules.

This design allows efficient adaptation of the large-scale encoder while keeping the majority of parameters frozen.

C. Multi-Scale UNet Decoder

The multi-scale features extracted from the encoder are fused using a UNet-style decoder. The feature maps are ordered from low to high resolution:

$$\mathbf{F}_3 \rightarrow \mathbf{F}_2 \rightarrow \mathbf{F}_1 \rightarrow \mathbf{F}_0.$$

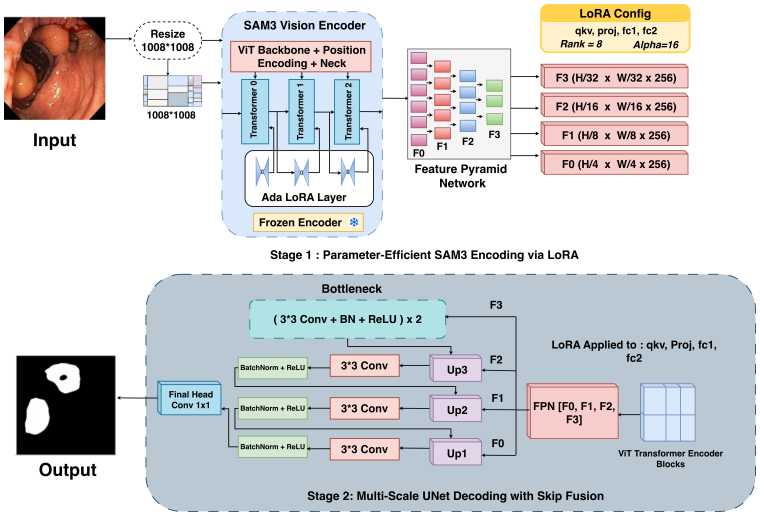

Fig. 1. Figure 1: LS-PolypSeg architecture: SAM3 with LoRA extracts multi-scale features, which are fused via a UNet decoder to produce accurate polyp segmentation.

The decoding process begins with a bottleneck transformation applied to the lowest-resolution feature map:

$$\mathbf{Z}_3 = \mathcal{B}(\mathbf{F}_3), \tag{5}$$

where $\mathcal{B}(\cdot)$ denotes a convolutional block consisting of two consecutive 3×3 convolutions with Batch Normalization and ReLU activation.

At each decoding stage, the feature map is upsampled and fused with the corresponding skip connection:

$$\mathbf{Z}_2 = \mathcal{U}_3(\mathbf{Z}_3, \mathbf{F}_2), \tag{6}$$

$$\mathbf{Z}_1 = \mathcal{U}_2(\mathbf{Z}_2, \mathbf{F}_1), \tag{7}$$

$$\mathbf{Z}_0 = \mathcal{U}_1(\mathbf{Z}_1, \mathbf{F}_0), \tag{8}$$

where each $\mathcal{U}_i(\cdot)$ denotes an upsampling block that performs bilinear interpolation followed by channel-wise concatenation and convolution:

$$\mathcal{U}(x, s) = \mathcal{C}(\text{Concat}(\text{Up}(x), s)), \tag{9}$$

with $\mathcal{C}(\cdot)$ representing the convolutional block.

D. Segmentation Head

The final feature map is passed through a segmentation head:

$$\mathbf{Y}_{logits} = \mathcal{H}(\mathbf{Z}_0), \tag{10}$$

where $\mathcal{H}(\cdot)$ consists of a 3×3 convolution followed by Batch Normalization and ReLU, and a final 1×1 convolution to project features to the desired number of classes.

The output logits are then resized back to the original image resolution:

$$\mathbf{Y} = \mathcal{R}^{-1}(\mathbf{Y}_{logits}). \tag{11}$$

Overall, LS-PolypSeg aims to provide a unified framework that integrates foundation model representations with efficient adaptation and structured multi-scale fusion, enabling accurate and robust segmentation while maintaining architectural simplicity and scalability.

IV. Experimental Setup and Results

We evaluate the proposed LS-PolypSeg framework on three public polyp segmentation benchmarks: *Kvasir-SEG* [15], *CVC-ClinicDB* [16], and *BKAI-IGH NeoPolyp-Small* [17]. Kvasir-SEG contains 1,000 images with significant variability in polyp morphology, while CVC-ClinicDB (612 images) presents challenges in small polyp detection and boundary delineation. BKAI-IGH NeoPolyp-Small includes 1,200 high-resolution images acquired under WLI and FICE modalities, introducing variations in illumination and imaging conditions. For all datasets, we adopt an 80

The LS-PolypSeg model is implemented in PyTorch and trained on NVIDIA GPUs. The SAM3 encoder is initialized with pretrained weights and remains largely frozen, while LoRA adapters enable targeted adaptation. Given an input image $\mathbf{X} \in \mathbb{R}^{H \times W \times 3}$, it is resized to $\mathbf{X}_{enc} \in \mathbb{R}^{1008 \times 1008}$ for encoding. Optimization is performed using AdamW with learning rates in the range $[1 \times 10^{-5}, 1 \times 10^{-4}]$, batch size 8, and 50 epochs. LoRA is configured with rank $r \in [4, 16]$, scaling factor $\alpha = 16$, and dropout in $[0.0, 0.1]$, applied to attention and feed-forward layers. The training objective combines binary cross-entropy and Dice loss:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{BCE} + \lambda_2 \mathcal{L}_{Dice}, \tag{12}$$

where λ_1 and λ_2 balance their contributions. Performance is evaluated using Dice (DSC), Intersection over Union (IoU), Precision, Recall, and F2-score, ensuring consistent evaluation across all experiments.

A. Quantitative Results

Tables I, II, and III present the quantitative comparison of LS-PolypSeg with state-of-the-art methods across all three datasets. The proposed method consistently achieves superior or highly competitive performance across all evaluation metrics. The results demonstrate several important observations regarding the effectiveness of the proposed framework.

TABLE I
KVASIR-SEG RESULTS.

Method	IoU	DSC	Rec.	Prec.	F2
U-Net [18]	0.7472	0.8264	0.8504	0.8703	0.8353
U-Net++ [19]	0.7420	0.8228	0.8437	0.8607	0.8295
ResU-Net++ [20]	0.5341	0.6453	0.6964	0.7080	0.6576
HarDNet-MSEG [21]	0.7459	0.8260	0.8485	0.8652	0.8358
ColonSegNet [22]	0.6980	0.7920	0.8193	0.8432	0.7999
UACANet [23]	0.7692	0.8502	0.8799	0.8706	0.8626
UNeXt [24]	0.6284	0.7318	0.7840	0.7656	0.7507
TransNetR [25]	0.8016	0.8706	0.8843	0.9073	0.8744
LSSNet [26]	0.8780	0.9261	0.9246	0.8802	0.9156
QueryNet [27]	0.8835	0.9328	0.8264	0.9174	0.8431
FusePolyp [7]	0.8592	0.9243	0.8960	0.9544	0.9062
LS-PolypSeg (ours)	**0.8974**	**0.9404**	**0.9463**	**0.9476**	**0.9416**

First, LS-PolypSeg consistently achieves the highest Dice and IoU scores across all datasets, indicating strong segmentation accuracy and robust overlap with ground truth masks. On Kvasir-SEG, the proposed method achieves a Dice score of 0.9404, outperforming recent transformer-based models such as QueryNet (0.9328) and LSSNet (0.9261). This improvement highlights the effectiveness of combining global contextual representations from SAM3 with structured multi-scale decoding.

TABLE II
CVC-CLINICDB RESULTS.

Method	IoU	DSC	Rec.	Prec.	F2
U-Net [18]	0.5433	0.6336	0.6982	0.7891	0.6563
U-Net++ [19]	0.5475	0.6350	0.6933	0.7967	0.6556
ResU-Net++ [20]	0.3585	0.4642	0.5880	0.5770	0.5084
HarDNet-MSEG [21]	0.6059	0.6960	0.7173	0.8528	0.7010
ColonSegNet [22]	0.5090	0.6126	0.6564	0.7521	0.6246
UACANet [23]	0.6808	0.7659	0.7639	0.8820	0.7599
UNeXt [24]	0.3901	0.4915	0.6125	0.6609	0.5318
TransNetR [25]	0.8628	0.9107	0.9044	0.9380	0.9065
G-Cascade [28]	0.8910	0.9374	0.9483	0.9349	0.9416
QueryNet [27]	0.8940	0.9421	0.9705	0.9705	0.9705
LSSNet [26]	0.9004	0.9448	0.9430	0.9200	0.9394
FANetv2 [29]	0.9039	0.9481	0.9514	0.9490	0.9496
FusePolyp [7]	0.9039	0.9495	0.9443	0.9548	0.9464
LS-PolypSeg (ours)	**0.9120**	**0.9514**	**0.9545**	**0.9548**	**0.9523**

On CVC-ClinicDB, which is known for challenging boundary delineation, LS-PolypSeg achieves a Dice score of 0.9514 and IoU of 0.9120, surpassing all competing methods, including FANetv2 and LSSNet. Notably, the recall (0.9545) and precision (0.9548) values are well balanced, suggesting that the model effectively captures polyp regions without introducing excessive false positives. This balance is critical in clinical scenarios where both sensitivity and specificity are essential.

On the BKAI-IGH dataset, which introduces variability in imaging modalities, LS-PolypSeg again achieves state-of-the-art performance with a Dice score of 0.9503 and IoU of 0.9112. Compared to FusePolyp, which achieves strong precision (0.9625), our method maintains a more balanced trade-off between precision and recall, leading to improved overall segmentation quality as reflected in the F2-score.

TABLE III
BKAI-IGH RESULTS.

Method	IoU	DSC	Rec.	Prec.	F2
U-Net [18]	0.7599	0.8286	0.8295	0.8999	0.8264
DeepLabV3 [30]	0.8314	0.8938	0.8870	0.9333	0.8882
PraNet [31]	0.8264	0.8904	0.8901	0.9247	0.8885
UACANet [23]	0.8275	0.8945	0.8870	0.9297	0.8882
CaraNet [32]	0.8329	0.8962	0.8939	0.9273	0.8937
LDNet [33]	0.8254	0.8927	0.8867	0.9153	0.8874
TGANet [34]	0.8409	0.9023	0.9025	0.9208	0.9002
TransNetR [25]	0.8474	0.9107	0.8982	0.9396	0.9018
G-Cascade [28]	0.8465	0.9096	0.8872	0.9489	0.8948
FANetv2 [29]	0.8646	0.9186	0.9058	0.9535	0.9096
FusePolyp [7]	0.8928	0.9434	0.9249	0.9625	0.9321
LS-PolypSeg (ours)	**0.9112**	**0.9503**	**0.9540**	**0.9542**	**0.9508**

A key insight from these results is the effectiveness of the proposed hybrid architecture. Unlike purely convolutional methods that may struggle with global context, or transformer-only approaches that may lack fine spatial precision, LS-PolypSeg successfully integrates both aspects. The SAM3 encoder captures rich semantic representations, while the UNet-style decoder ensures accurate spatial reconstruction through hierarchical feature fusion. Furthermore, the consistent performance gains across datasets of varying characteristics demonstrate the generalization capability of the model. The improvements are not limited to a specific dataset or imaging condition, but extend across diverse scenarios, including small polyps, irregular boundaries, and varying illumination conditions.

B. Parameter Efficiency Analysis

To evaluate the parameter efficiency of the proposed LS-PolypSeg framework, we analyze the total and trainable parameters of the model. As shown in Table IV, the model comprises a total of 463.47M parameters, out of which only 9.43M parameters (2.03%) are trainable. This significant reduction is achieved by leveraging LoRA-based adaptation within the SAM3 encoder, where only 4.52M parameters are updated while keeping the remaining 454.04M parameters frozen. The lightweight decoder contributes an additional 4.91M trainable parameters. Compared to conventional full fine-tuning strategies that update all parameters, our approach reduces the number of trainable parameters by over $50\times$ while maintaining superior segmentation performance. These results demonstrate that LS-PolypSeg effectively exploits the representational power of large-scale vision models with minimal parameter updates, making it both computationally efficient and scalable for medical image segmentation tasks.

V. CONCLUSION

In this work, we proposed **LS-PolypSeg**, a parameter-efficient framework that integrates a pretrained SAM3 vision encoder with LoRA-based adaptation and a lightweight UNet-style decoder for accurate polyp segmentation. By selectively

TABLE IV
PARAMETER EFFICIENCY ANALYSIS OF LS-POLYPSEG.

Component	Total Params (M)	Trainable Params (M)
SAM3 Encoder	458.56	4.52
Decoder	4.91	4.91
Overall Model	**463.47**	**9.43**

Metric	Value
Trainable Percentage	2.03%
Frozen Parameters	454.04M
Reduction vs Full Fine-Tuning	$\sim 50\times$

adapting key transformer layers while preserving the majority of pretrained parameters, the proposed method effectively balances representation power and computational efficiency. Extensive experiments on Kvasir-SEG, CVC-ClinicDB, and BKAI-IGH datasets demonstrate that LS-PolypSeg achieves consistent improvements and state-of-the-art performance across multiple evaluation metrics, with particularly strong gains in Dice and IoU while maintaining a balanced precision–recall trade-off. The results highlight the effectiveness of combining global contextual representations from foundation models with structured multi-scale decoding for precise boundary delineation. Despite these promising outcomes, the current framework relies on fixed-resolution inputs and does not explicitly model temporal or cross-domain variations, which may limit performance in real-world clinical deployments. Future work will focus on extending the framework to handle video-based colonoscopy data, improving robustness under domain shifts, and exploring more adaptive feature fusion strategies. Overall, LS-PolypSeg provides a scalable and efficient direction for leveraging foundation models in medical image segmentation.

REFERENCES

[1] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "Doubleu-net: A deep convolutional neural network for medical image segmentation," in *IEEE Symposium on Computer-Based Medical Systems (CBMS)*, 2020, pp. 558–564.

[2] N. K. Tomar, A. Shergill, B. Rieders, U. Bagci, and D. Jha, "Transresu-net: Transformer-based resu-net for real-time colon polyp segmentation," in *IEEE EMBS*, 2023.

[3] A. Srivastava, S. Chanda, D. Jha, U. Pal, and S. Ali, "Gmsrf-net: Global multi-scale residual fusion network for polyp segmentation," in *ICPR*, 2022.

[4] C. S. Raghaw, A. Yadav, J. S. Sanjotra, S. Dangi, and N. Kumar, "Mnet-sat: Multiscale network with spatial-enhanced attention for polyp segmentation," *Biomedical Signal Processing and Control*, 2025.

[5] A. Das, D. Jha *et al.*, "Pam-unet: Progressive attention-based medical image segmentation," in *IEEE EMBC*, 2024.

[6] Y. Zhou, L. Li, L. Lu, and M. Xu, "nnwnet: Rethinking the use of transformers in biomedical image segmentation and calling for a unified evaluation benchmark," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 20 852–20 862.

[7] P. Srikanth, K. Sambanna, S. J. Ganji, and R. S. Rao, "Fusepolyp: Complementary dual encoder design and fusion strategy analysis for medical polyp segmentation," in *International Conference on Distributed Computing and Intelligent Technology*. Springer Nature Switzerland, 2026, pp. 436–451.

[8] J. Ma, Z. Yang, S. Kim, B. Chen, M. Baharoon, A. Fallahpour, and B. Wang, "Medsam2: Segment anything in 3d medical images and videos," *arXiv preprint arXiv:2504.03600*, 2025.

[9] A. Liu, R. Xue, X. R. Cao, Y. Shen, Y. Lu, X. Li, and J. Chen, "Medsam3: Delving into segment anything with medical concepts," *arXiv preprint arXiv:2511.19046*, 2025.

[10] Y. Li, Z. Zhu, Y. Zhang, Y. Chen, and Z. Yu, "Boost the inference with co-training: A depth-guided mutual learning framework for semi-supervised medical polyp segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 10 394–10 403.

[11] A. Konwer, Z. Yang, E. Bas, C. Xiao, P. Prasanna, P. Bhatia, and T. Kass-Hout, "Enhancing sam with efficient prompting and preference optimization for semi-supervised medical image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 20 990–21 000.

[12] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," in *International Conference on Learning Representations (ICLR)*, 2022.

[13] M. M. Rahman, M. Munir, D. Jha, U. Bagci, and R. Marculescu, "Pp-sam: Perturbed prompts for robust adaptation of segment anything model for polyp segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 4989–4995.

[14] T. K. Dutta, S. Majhi, D. R. Nayak, and D. Jha, "Sam-mamba: Mamba guided sam architecture for generalized zero-shot polyp segmentation," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2025, pp. 4655–4664.

[15] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, "Kvasir-seg: A segmented polyp dataset," in *MultiMedia Modeling. MMM 2020*, ser. Lecture Notes in Computer Science, Y. M. Ro, S. Hwang, T. Kim, W.-T. Chu, and M.-C. Hu, Eds., vol. 11962. Springer, Cham, 2020, pp. 451–462.

[16] J. Bernal, F. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized Medical Imaging and Graphics*, vol. 43, pp. 99–111, 2015.

[17] P. N. Lan, N. S. An, D. V. Hang, D. V. Long, T. Q. Trung, N. T. Thuy, and D. V. Sang, "Neounet: Towards accurate colon polyp segmentation and neoplasm detection," in *Proceedings of the 16th International Symposium on Visual Computing*, 2021.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *MICCAI*, 2015.

[19] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," *IEEE TMI*, 2018.

[20] D. Jha *et al.*, "Resunet++: An advanced architecture for medical image segmentation," *IEEE ISM*, 2019.

[21] R. P. Poudel *et al.*, "Hardnet-mseg: A simple encoder-decoder network for polyp segmentation," *MICCAI EndoCV*, 2020.

[22] D. Jha *et al.*, "Colonsegnet: A novel deep learning model for polyp segmentation," *IEEE Access*, 2021.

[23] T. Kim *et al.*, "Uacanet: Uncertainty augmented context attention network," *ACM MM*, 2021.

[24] J. M. J. Valanarasu *et al.*, "Unext: Mlp-based rapid medical image segmentation," *MICCAI*, 2022.

[25] D. Jha *et al.*, "Transnetr: Transformer-based polyp segmentation," *Pattern Recognition*, 2024.

[26] T. Zhou *et al.*, "Lssnet: Local-global semantic segmentation network for polyp segmentation," *IEEE TMI*, 2024.

[27] X. Zhang *et al.*, "Querynet: Query-based polyp segmentation network," in *CVPR*, 2024.

[28] H. Zhou *et al.*, "G-cascade: Global context-aware cascade network," *Medical Image Analysis*, 2023.

[29] W. Li *et al.*, "Fanetv2: Feature aggregation network for medical image segmentation," *IEEE TMI*, 2024.

[30] L.-C. Chen *et al.*, "Encoder-decoder with atrous separable convolution for semantic segmentation," *ECCV*, 2018.

[31] D.-P. Fan *et al.*, "Pranet: Parallel reverse attention network for polyp segmentation," *MICCAI*, 2020.

[32] X. Lou *et al.*, "Caranet: Context axial reverse attention network," *MICCAI*, 2022.

[33] Y. Zhang *et al.*, "Ldnet: Lesion-aware dense network," *IEEE TMI*, 2022.

[34] N. Tomar *et al.*, "Tganet: Transformer guided attention network," *MICCAI*, 2022.