ASPS: Augmented Segment Anything Model for Polyp Segmentation

Huiqian Li[1,2], Dingwen Zhang[2,3] (✉), Jieru Yao[2,3], Longfei Han[4,1] (✉), Zhongyu Li[5], and Junwei Han[2,3]

[1] University of Science and Technology of China, Hefei, China
[2] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei, China
[3] Northwestern Polytechnical University, Xi'an, China
[4] Beijing Technology and Business University, Beijing, China
[5] Xi'an Jiaotong University, Xi'an, China
zhangdingwen2006yyy@gmail.com, draflyhan@gmail.com

Abstract. Polyp segmentation plays a pivotal role in colorectal cancer diagnosis. Recently, the emergence of the Segment Anything Model (SAM) has introduced unprecedented potential for polyp segmentation, leveraging its powerful pre-training capability on large-scale datasets. However, due to the domain gap between natural and endoscopy images, SAM encounters two limitations in achieving effective performance in polyp segmentation. Firstly, its Transformer-based structure prioritizes global and low-frequency information, potentially overlooking local details, and introducing bias into the learned features. Secondly, when applied to endoscopy images, its poor out-of-distribution (OOD) performance results in substandard predictions and biased confidence output. To tackle these challenges, we introduce a novel approach named **A**ugmented **S**AM for **P**olyp **S**egmentation (**ASPS**), equipped with two modules: Cross-branch Feature Augmentation (CFA) and Uncertainty-guided Prediction Regularization (UPR). CFA integrates a trainable CNN encoder branch with a frozen ViT encoder, enabling the integration of domain-specific knowledge while enhancing local features and high-frequency details. Moreover, UPR ingeniously leverages SAM's IoU score to mitigate uncertainty during the training procedure, thereby improving OOD performance and domain generalization. Extensive experimental results demonstrate the effectiveness and utility of the proposed method in improving SAM's performance in polyp segmentation. Our code is available at https://github.com/HuiqianLi/ASPS.

Keywords: Polyp Segmentation · Segment Anything Model · Domain Adaptation.

1 Introduction

Automated polyp segmentation stands as a pivotal tool in the diagnosis of colorectal cancer, to aid effective interventions and timely treatment strategies.

Studies like Polyp-PVT[5], SSFormer[22] used Pyramid Vision Transformer for polyp segmentation; CFANet[35] integrated boundaries with a Cross-level Feature Aggregation Network; Endo-FM[23] captured spatial-temporal dependencies to build a foundation model. However, limited by the model's size, the existing methods still lack sufficient capabilities for feature representation and extraction, making it challenging to fully capture the morphology and characteristics of polyps. Furthermore, the limited scale of the dataset may limit the diversity and generalization of the existed methods. Recently, the Segment Anything Model[12] (SAM) was introduced. SAM is pre-trained on the largest segmentation dataset SA-1B, demonstrating remarkable segmentation capabilities across various downstream tasks. With its significant model size and data size, this innovative approach has introduced novel perspectives to the field of polyp segmentation. It also possesses enhanced representation and feature extraction capabilities, surpassing existing methods.

However, SAM's performance in the polyp segmentation task is unsatisfactory [34], due to the domain gap between the training data and endoscopy images. This results in two primary issues: firstly, SAM fails to adequately capture the distinctive features of polyp images, leading to a bias in its learned representations. Secondly, it produces erroneous predictions with inaccurate confidence estimates for out-of-distribution (OOD) data. In addition, because it relied on prompts, SAM has significantly impeded its convenience in clinical applications. Despite several methods improving SAM, such as MedSAM[15], these approaches either rely on prompts or directly fine-tune substantial models. SAMUS[14] effectively integrates CNN and ViT, but its design is quite complex and is particularly suited for processing small images. Consequently, the efficacy of these methods is somewhat constrained. Various methods have been proposed to tackle the challenge of unsupervised domain adaptation in semantic segmentation. MIC[8] proposed a Masked Image Consistency module for target domain context learning; Context-Aware Domain Adaptation[26] improved context transfer via cross-attention. Yet, domain-specific information integration and uncertainty reduction are still unexplored.

To address these issues, we introduce a novel method based on SAM from a domain adaptation perspective, designed to augment the feature extraction capability and generalization without relying on prompts. We propose the Cross-branch Feature Augmentation Module (CFA) and the Uncertainty-guided Prediction Regularization Module (UPR). CFA incorporates an additional trainable convolutional neural network (CNN) encoder branch, which complements the frozen vision transformer (ViT) encoder, to capture multi-scale and multi-level features. UPR adjusts the normalization layer to promote the adaptation in the endoscopy field and leverages hints to ensure accurate confidence estimation, so as to improve the OOD performance of SAM.

In summary, our primary contributions are as follows: (1) We build a novel SAM-based model named ASPS, to enhance the feature learning capability and domain generalization for polyp segmentation, demonstrating strong performance without the need for prompts. (2) We introduce the Cross-branch

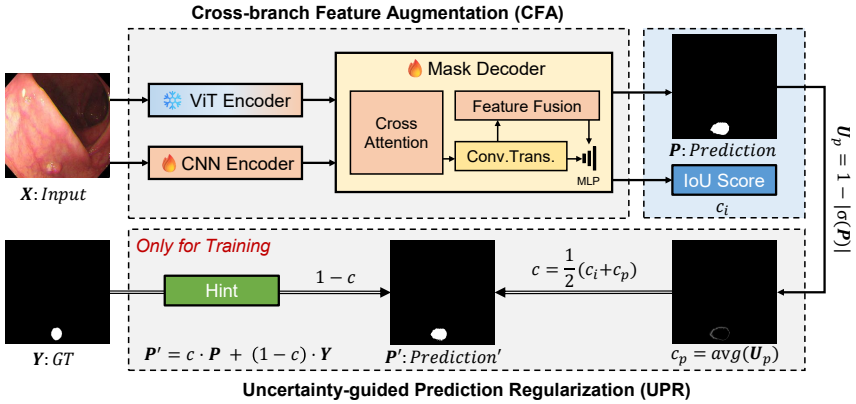

Fig. 1: An overview of our Augmented Segment Anything Model for polyp segmentation. The Cross-branch Feature Augmentation module is encouraged to learn multi-scale features and multi-level representations. The Uncertainty-guided Prediction Regularization module is designed to minimize the uncertainty of the prediction to improve the domain generalization ability of the model.

Feature Augmentation Module (CFA), which introduces an additional CNN encoder branch as a supplement to the ViT encoder. Furthermore, we propose the Uncertainty-guided Prediction Regularization Module (UPR), leveraging hints to reduce uncertainty during training and improve the domain generalization of SAM. (3) Extensive experiments on five common polyp datasets demonstrate the effectiveness and superiority of our method.

2 Method

Overview. Our proposed network is illustrated in Fig. 1. To address the domain degradation issue of SAM, we leverage two modules to enhance its original feature extraction capabilities and domain generalization. The CFA module integrates the CNN encoder feature with global ViT information, leading to generalized feature representation learning. This integration facilitates refined segmentation outputs by aggregating deep information to the superficial layers and incorporating positional information from the shallow levels. Meanwhile, the UPR module is designed to minimize uncertainty and calibrate confidence during training. UPR utilizes a training strategy based on uncertainty, leveraging the ground truth as a guiding 'hint'. The proposed network follows end-to-end training without prompts, jointly optimizing two modules to achieve optimal performance.

2.1 Cross-branch Feature Augmentation Module

While SAM has achieved great success in many image segmentation tasks, it still has some limitations in the polyp segmentation task. One of the main reasons is

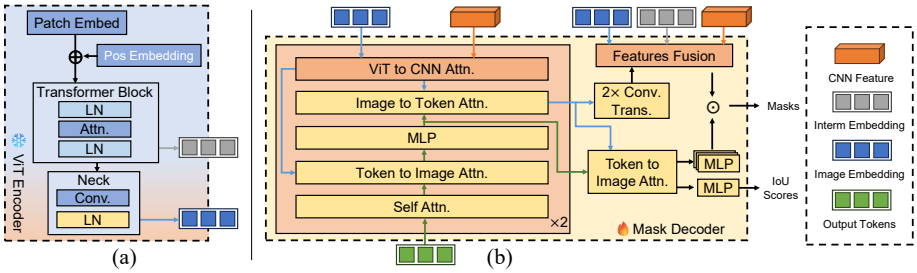

Fig. 2: Detailed architecture of ViT Encoder and Mask Decoder. (a) represents the ViT encoder, while (b) showcases the lightweight decoder of SAM. The CNN feature is derived from the output of the CNN encoder. The yellow and blue modules represent the original SAM structure.

that the image encoder of SAM is not able to capture enough features effectively from unseen endoscopy images. To address this issue, the CFA module is designed to learn multi-scale features and multi-level representations, thereby enhancing the feature extraction capabilities of the encoder.

Firstly, to achieve automatic segmentation, we modified the architecture of SAM by removing its prompt input and prompt encoder components while preserving its image encoder and mask decoder parts. Recent studies[17] have demonstrated that ViT is more focused on low-frequency signals, while CNN is more adept at processing high-frequency signals. Hence, we integrate a parallel CNN-based branch to compensate for the absence of high-frequency and local features. Furthermore, we augment the mask decoder of SAM by proposing an additional multi-head cross-branch attention block to facilitate the integration of features extracted from both the ViT encoder and the CNN encoder. For the features $\mathbf{F}_v$ from the ViT branch and $\mathbf{F}_c$ from the CNN, the cross-branch attention can be formulated as follows:

$$\text{CrossBranchAttention}\left(\mathbf{F}_v, \mathbf{F}_c\right) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\mathsf{T}}}{\sqrt{d}}\right)\mathbf{V}. \tag{1}$$

where $\mathbf{Q} = \mathbf{F}_v\mathbf{W}^{\mathbf{Q}}$, $\mathbf{K} = \mathbf{V} = \mathbf{F}_c\mathbf{W}^{\mathbf{K}}$, and d is the number of channels of each head of $\mathbf{F}_v$. Considering that CNN features offer more precise position information, we substituted SAM's original position embedding in the mask decoder with the final output features from the CNN encoder. Furthermore, we integrated the cross-branch attention mechanism into the attention block of the mask decoder, repeating this process twice to ensure integration of multi-scale features from both the ViT and CNN encoders, as shown in Fig. 2.

Secondly, to obtain more precise segmentation results, we integrate high-level context and low-level boundary information from the encoder with the decoder features of SAM to augment the output information. Specifically, we combine the shallow local features obtained from the intermediate embedding of the ViT encoder, the final global features obtained from the image embedding

of the ViT encoder, and the final features of the CNN encoder, as illustrated in Fig. 2. This approach fully harnesses the rich edge information, extensive global context information, and local position details of each encoder branch. Thus, we can effectively integrate multi-level features from both ViT and CNN.

2.2 Uncertainty-guided Prediction Regularization Module

To augment the generalization capability of SAM, we propose a novel training strategy involving the selective activation of LayerNorm within the encoder. We also employ the ground truth as a 'hint' to further guide the training process by correcting the confidence.

Given that SAM is trained on natural data, its performance may deteriorate in polyp images due to the domain transfer. As previously suggested [13], it is a particularly effective technique for domain transfer by adjusting the normalization layer. Despite the introduction of LayerNorm [1] potentially reducing training time, it fundamentally alters the distribution of the input data. When transferring SAM from natural images to endoscopy images, there is a shift in both the data distribution and the corresponding feature space distribution. These distributional differences can induce internal covariate shifts, thereby influencing the model's performance. To improve the SAM's generalization in the endoscopy field, we fine-tune the normalization layer of the encoder. In this process, the model effectively adapts the data distribution in the target domain and mitigates the effects of internal covariate shifts.

Specifically, the LayerNorm of SAM's ViT encoder is divided into (1) Transformer block norm, and (2) neck layer norm, as illustrated in Fig. 2(a). Given that the features in the neck layer are closer to the output features of the encoder, we ultimately decided to train the neck layer normalization, which is equivalent to re-normalizing the features of the pre-trained ViT encoder. Coincidentally, in this work[32], the straightforward technique of adjusting normalization layers can surprisingly yield comparable or even superior performance to the robust baseline of fine-tuning all parameters.

Moreover, previous research[16] has demonstrated that predictions with lower uncertainty tend to exhibit superior out-of-distribution (OOD) performance, which is also beneficial for domain adaptation. SAM generates an IoU score output, which inherently can represent uncertainty (or, conversely, confidence). However, during the prediction process, SAM may frequently produce incorrect predictions for unseen data with high confidence, which is undesirable. To mitigate this problem, we strive to reduce the uncertainty of the model during training (i.e., to increase confidence). Inspired by [4], we utilize the ground truth as a hint to guide the learning of the model. First, we represent the IoU score of SAM as the image-level confidence c_i. Then we calculate the pixel-level confidence c_p to refine the uncertainty of each pixel using Eq. 2, where $\mathbf{U}_p \in \mathbb{R}^{B\times 1\times H\times W}$.

$$c_p = \left(1 - \frac{1}{H \times W} \sum_{i}^{H} \sum_{j}^{W} \mathbf{U}_p\right). \tag{2}$$

The term $\mathbf{U}_p$ represents the pixel uncertainty, defined as $\mathbf{U}_p = 1 - \sigma(|\mathbf{P}|)$. Here, σ represents the Sigmoid function, while $\mathbf{P}$ represents the output prediction. The final confidence is calculated as the sum of the image-level confidence and the pixel-level confidence, expressed as $c = \frac{1}{2}(c_i + c_p)$. This confidence is determined by the Bernoulli distribution, which decides whether to utilize the ground truth as a hint. In other words, if the confidence is low enough, we believe that the model requires a specific answer hint to learn the correct mask prediction. Thus, the answer is required as a hint, otherwise, it is not necessary. The weight of the hint is determined by the confidence c, which is expressed as follows:

$$\mathbf{P}' = c \cdot \mathbf{P} + (1 - c) \cdot \mathbf{Y}. \tag{3}$$

However, by minimizing the loss function, the model will tend to make $c = 0$ so that $\mathbf{P}'$ will always be GT. This means that the model does not actually learn. Therefore, a confidence loss is introduced to supervise c, which will increase when $c \to 0$, and the confidence loss is defined as follows:

$$\mathcal{L}_c = -log(c). \tag{4}$$

The final loss function is the sum of the segmentation loss $\mathcal{L}_s$ and the confidence loss $\mathcal{L}_c$, as defined in Eq. 5. Here, λ represents a hyperparameter. Specifically, the segmentation loss employed is a combination of CE loss, Dice loss, and MSE loss as $\mathcal{L}_s = \mathcal{L}_{ce} + 0.5 \cdot \mathcal{L}_{dice} + \mathcal{L}_{mse}$.

$$\mathcal{L} = \mathcal{L}_s + \lambda \mathcal{L}_c. \tag{5}$$

3 Experiments

Datasets. We conduct extensive experiments on five polyp segmentation datasets following PraNet[6], including Kvasir-SEG[10], CVC-ClinicDB[2], CVC-ColonDB[20], ETIS[19] and EndoScene[21]. Specifically, the training set consists of 900 images from Kvasir-SEG and 550 images from ClinicDB. The test sets comprise 100 images from Kvasir-SEG, 62 images from CVC-ClincDB, 380 images from CVC-ColonDB, 60 images from EndoScene, and 196 images from ETIS.

Implementations. We use PyTorch to implement our method and conduct experiments on a single NVIDIA RTX3090 GPU. The AdamW optimizer is utilized for training 16K iterations with a learning rate of 1e-5, a weight decay of 1e-4, and a batch size of 4. The CNN model we utilize is MSCAN-L, sourced from SegNeXt[7]. The input image size for the ViT branch is 1024×1024, while the input size for the CNN branch is 320×320. In the evaluation stage, we use two common metrics in medical image segmentation, *Dice* and *IoU*.

Results and Analysis. We compare our method with several state-of-the-art polyp segmentation methods and some SAM-based methods in Table 1. It is evident that while the performance enhancement on Kvasir and CVC-ColonDB is

Table 1: Quantitative comparisons with state-of-the-art (SOTA) methods on five public polyps datasets are presented. We mark the best results with **bold** and the second-best scores with underline.

Methods	Published Venue	CVC-ClinicDB		Kvasir		CVC-ColonDB		ETIS		EndoScene	
		Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
UNet[18]	MICCAI'15	0.823	0.755	0.818	0.746	0.504	0.436	0.398	0.335	0.710	0.627
PraNet[6]	MICCAI'19	0.899	0.849	0.898	0.840	0.709	0.640	0.628	0.567	0.871	0.797
SANet[24]	MICCAI'21	0.916	0.859	0.904	0.847	0.752	0.669	0.750	0.654	0.888	0.815
MSNet[33]	MICCAI'21	0.915	0.866	0.902	0.847	0.747	0.668	0.720	0.650	0.862	0.796
UACANet[11]	MM'21	0.916	0.870	0.905	0.852	0.783	0.704	0.694	0.615	0.902	0.837
LDNet[31]	MICCAI'22	0.932	0.872	0.912	0.855	0.794	0.715	0.778	0.707	0.893	0.826
SSFormer[22]	MICCAI'22	0.906	0.855	0.917	**0.864**	0.802	0.721	0.796	0.720	0.895	0.827
DCRNet[27]	ISBI'22	0.869	0.800	0.846	0.772	0.661	0.576	0.509	0.432	0.753	0.670
Polyp-PVT[5]	AIR'23	0.937	0.889	0.917	**0.864**	**0.808**	**0.727**	0.787	0.706	0.900	0.833
CFANet[35]	PR'23	0.933	0.883	0.915	0.861	0.743	0.665	0.732	0.655	0.893	0.827
SAM-H[12]	ICCV'23	0.547	0.500	0.778	0.707	0.441	0.396	0.517	0.477	0.651	0.606
SAM-L[12]	ICCV'23	0.579	0.526	0.782	0.710	0.468	0.422	0.551	0.507	0.726	0.676
SAM-Adapter[3]	ICCV'23	0.774	0.673	0.847	0.763	0.671	0.568	0.590	0.476	0.815	0.725
AutoSAM[9]	ArXiv'23	0.751	0.642	0.784	0.675	0.535	0.418	0.402	0.308	0.829	0.739
SAMPath[29]	MICCAIw'23	0.750	0.644	0.828	0.730	0.632	0.516	0.555	0.442	0.844	0.756
SAMed[30]	ArXiv'23	0.404	0.273	0.459	0.300	0.199	0.115	0.212	0.126	0.332	0.202
SAMUS[14]	ArXiv'23	0.900	0.821	0.859	0.763	0.731	0.597	0.750	0.618	0.859	0.760
SurgicalSAM[28]	AAAI'24	0.644	0.505	0.740	0.597	0.460	0.330	0.342	0.238	0.623	0.472
MedSAM[15]	Nature'24	0.867	0.803	0.862	0.795	0.734	0.651	0.687	0.604	0.870	0.798
Ours	Efficient-SAM[25]	0.942	0.891	0.914	0.849	0.782	0.680	0.854	0.758	0.900	0.819
Ours	ViT-B	0.950	0.905	0.914	0.848	0.792	0.694	0.856	0.764	0.914	0.843
Ours	ViT-H	**0.951**	**0.906**	**0.920**	0.858	0.799	0.701	**0.861**	**0.769**	**0.919**	**0.852**

not particularly noticeable, the improvement on CVC-ClinicDB, ETIS, and EndoScene is quite significant. Especially, Polyp-PVT[5] showed good performance on all datasets with average Dice and IoU of 0.870 and 0.804, respectively, while our method achieved 0.890 and 0.817, which proves the effectiveness of our model.

Compared to SAM-based (ViT-B) methods, our approach outperformed all others across all datasets. It's important to note that methods like MedSAM[15] and SAMUS[14] still incorporated prompts like SAM, but we removed prompts in the comparative experiment. Besides, despite our efforts, we didn't attain satisfactory results with SurgicalSAM[28] and SAMed[30], potentially due to incompatible training hyperparameters. Moreover, we observed that SAMUS[14] also utilized a CNN auxiliary branch and achieved excellent results, respectively, further validating the effectiveness of the CNN branch. Additionally, with the development of the lightweight model of SAM, we successfully combined our method with EfficientSAM[25] and achieved satisfactory results.

In Fig. 3, we use Fourier analysis as a toolkit to show the difference between features from two encoders. The Fourier spectrum and relative log amplitudes of the Fourier transformed feature maps indicate that the CNN branch captures more high-frequency signals than the ViT baseline. We also provide the qualitative results in Fig. 4, where our predictions are closer to the ground truth.

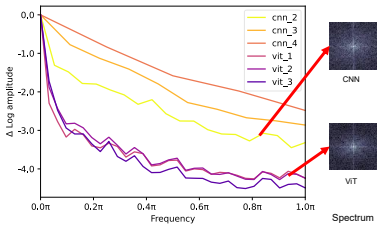

Fig. 3: Relative log amplitudes.

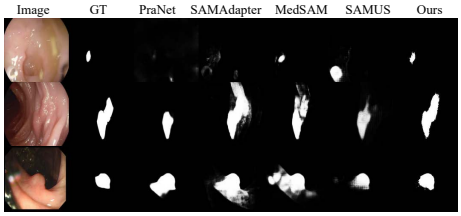

Fig. 4: Several qualitative results.

Table 2: Ablation experiments on five public polyp datasets.

CFA	UPR	CVC-ClinicDB		Kvasir		CVC-ColonDB		ETIS		EndoScene	
		Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
		0.537	0.388	0.652	0.469	0.397	0.268	0.387	0.265	0.525	0.368
✓		0.913	0.843	0.887	0.819	0.775	0.684	0.811	0.708	0.901	0.831
	✓	0.558	0.407	0.672	0.519	0.415	0.282	0.382	0.360	0.610	0.453
✓	✓	0.950	0.905	0.914	0.848	0.792	0.694	0.856	0.764	0.914	0.843

Table 3: Ablation studies on EPA.

CFA			UPR	CVC-ClinicDB		Kvasir	
CA	Fusion	PE		Dice	IoU	Dice	IoU
			✓	0.558	0.407	0.672	0.519
✓			✓	0.838	0.737	0.866	0.772
✓	✓		✓	0.941	0.891	0.911	0.848
✓	✓	✓	✓	0.950	0.905	0.914	0.848

Table 4: Ablation studies on UPR.

CFA	UPR			CVC-ClinicDB		Kvasir	
	TN	NN	Hint	Dice	IoU	Dice	IoU
✓				0.913	0.843	0.887	0.819
✓	✓			0.936	0.885	0.912	0.859
✓		✓		0.944	0.896	0.914	0.850
✓	✓	✓		0.936	0.881	0.875	0.814
✓		✓	✓	0.950	0.905	0.914	0.848

Ablation Study. We conducted ablation experiments to verify the effectiveness of the proposed CFA and UPR. For our baseline, we use ViT-B as the backbone of SAM and remove the prompt encoder. As shown in Table 2, we are confident to assert that the contribution of each module to the overall performance enhancement is significant, and their combination produces the best overall performance.

To ensure the reliability of the CFA, we conducted ablation experiments on the introduced cross-branch attention (CA), multi-level features fusion (Fusion), and position embedding replacement (PE). As shown in Table 3, compared with the baseline, our method achieved better performance and improved the mean Dice score by 31.7% and the mean IoU score by 41.4%. In addition, we tried to train (1) the Transformer block norm (TN), (2) the neck layer norm (NN), and (3) both to validate the effectiveness of UPR. The results are shown in Table 4. Due to the limitation of computing power, we set the batch size for experiments involving TN to 2, while the others were set to 4. Currently, only training the neck layer norm demonstrated superior performance.

4 Conclusion

We introduce a novel method called ASPS for polyp segmentation, designed to address the limitations of the SAM model in capturing information and bridging the domain gap between endoscopy images. The CFA module incorporates a trainable CNN encoder branch to supplement the frozen ViT encoder, integrating multi-scale and multi-level features. Additionally, the UPR module reduces uncertainty during training by introducing hints and adjusting the normalization layer, promoting the adaptation of the model in the endoscopy field. Through experiments on five common polyp datasets, we verify the effectiveness and superiority of our method. To extend our work, our future direction focuses on investigating more efficient methods using SAM, enabling us to fully harness the foundation model for effective polyp segmentation.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (No. 62202015, 62322605, 62293543, U21B2048, 62272468, 62306101), Anhui Provincial Key R&D Programmes (2023s07020001), the University Synergy Innovation Program of Anhui Province (GXXT-2022-052), and Key-Area Research and Development Program of Shaanxi Province under Grant 2023-ZDISF-41.

Disclosure of Interests. We declare that we have no competing interests.

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
2. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized medical imaging and graphics **43**, 99–111 (2015)
3. Chen, T., Zhu, L., Ding, C., Cao, R., Zhang, S., Wang, Y., Li, Z., Sun, L., Mao, P., Zang, Y.: Sam fails to segment anything?–sam-adapter: Adapting sam in underperformed scenes: Camouflage, shadow, and more. arXiv preprint arXiv:2304.09148 (2023)
4. DeVries, T., Taylor, G.W.: Learning confidence for out-of-distribution detection in neural networks. arXiv preprint arXiv:1802.04865 (2018)
5. Dong, B., Wang, W., Fan, D.P., Li, J., Fu, H., Shao, L.: Polyp-pvt: Polyp segmentation with pyramid vision transformers. arXiv preprint arXiv:2108.06932 (2021)
6. Fan, D.P., Ji, G.P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L.: Pranet: Parallel reverse attention network for polyp segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 263–273. Springer (2020)
7. Guo, M.H., Lu, C.Z., Hou, Q., Liu, Z., Cheng, M.M., Hu, S.M.: Segnext: Rethinking convolutional attention design for semantic segmentation. Advances in Neural Information Processing Systems **35**, 1140–1156 (2022)
8. Hoyer, L., Dai, D., Wang, H., Van Gool, L.: Mic: Masked image consistency for context-enhanced domain adaptation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11721–11732 (2023)

9. Hu, X., Xu, X., Shi, Y.: How to efficiently adapt large segmentation model (sam) to medical images. arXiv preprint arXiv:2306.13731 (2023)
10. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., de Lange, T., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26. pp. 451–462. Springer (2020)
11. Kim, T., Lee, H., Kim, D.: Uacanet: Uncertainty augmented context attention for polyp segmentation. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 2167–2175 (2021)
12. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
13. Li, Y., Wang, N., Shi, J., Liu, J., Hou, X.: Revisiting batch normalization for practical domain adaptation. arXiv preprint arXiv:1603.04779 (2016)
14. Lin, X., Xiang, Y., Zhang, L., Yang, X., Yan, Z., Yu, L.: Samus: Adapting segment anything model for clinically-friendly and generalizable ultrasound image segmentation. arXiv preprint arXiv:2309.06824 (2023)
15. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
16. Nguyen, D.M.H., Pham, T.N., Diep, N.T., Phan, N.Q., Pham, Q., Tong, V., Nguyen, B.T., Le, N.H., Ho, N., Xie, P., et al.: On the out of distribution robustness of foundation models in medical image segmentation. arXiv preprint arXiv:2311.11096 (2023)
17. Pan, Z., Cai, J., Zhuang, B.: Fast vision transformers with hilo attention. Advances in Neural Information Processing Systems **35**, 14541–14554 (2022)
18. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
19. Silva, J., Histace, A., Romain, O., Dray, X., Granado, B.: Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. International journal of computer assisted radiology and surgery **9**, 283–293 (2014)
20. Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automated polyp detection in colonoscopy videos using shape and context information. IEEE transactions on medical imaging **35**(2), 630–644 (2015)
21. Vázquez, D., Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., López, A.M., Romero, A., Drozdzal, M., Courville, A., et al.: A benchmark for endoluminal scene segmentation of colonoscopy images. Journal of healthcare engineering **2017** (2017)
22. Wang, J., Huang, Q., Tang, F., Meng, J., Su, J., Song, S.: Stepwise feature fusion: Local guides global. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 110–120. Springer (2022)
23. Wang, Z., Liu, C., Zhang, S., Dou, Q.: Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 101–111. Springer (2023)
24. Wei, J., Hu, Y., Zhang, R., Li, Z., Zhou, S.K., Cui, S.: Shallow attention network for polyp segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. pp. 699–708. Springer (2021)

25. Xiong, Y., Varadarajan, B., Wu, L., Xiang, X., Xiao, F., Zhu, C., Dai, X., Wang, D., Sun, F., Iandola, F., et al.: Efficientsam: Leveraged masked image pretraining for efficient segment anything. arXiv preprint arXiv:2312.00863 (2023)
26. Yang, J., An, W., Yan, C., Zhao, P., Huang, J.: Context-aware domain adaptation in semantic segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 514–524 (2021)
27. Yin, Z., Liang, K., Ma, Z., Guo, J.: Duplex contextual relation network for polyp segmentation. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2022)
28. Yue, W., Zhang, J., Hu, K., Xia, Y., Luo, J., Wang, Z.: Surgicalsam: Efficient class promptable surgical instrument segmentation. arXiv preprint arXiv:2308.08746 (2023)
29. Zhang, J., Ma, K., Kapse, S., Saltz, J., Vakalopoulou, M., Prasanna, P., Samaras, D.: Sam-path: A segment anything model for semantic segmentation in digital pathology. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 161–170. Springer (2023)
30. Zhang, K., Liu, D.: Customized segment anything model for medical image segmentation. arXiv preprint arXiv:2304.13785 (2023)
31. Zhang, R., Lai, P., Wan, X., Fan, D.J., Gao, F., Wu, X.J., Li, G.: Lesion-aware dynamic kernel for polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 99–109. Springer (2022)
32. Zhao, B., Tu, H., Wei, C., Mei, J., Xie, C.: Tuning layernorm in attention: Towards efficient multi-modal llm finetuning. arXiv preprint arXiv:2312.11420 (2023)
33. Zhao, X., Zhang, L., Lu, H.: Automatic polyp segmentation via multi-scale subtraction network. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. pp. 120–130. Springer (2021)
34. Zhou, T., Zhang, Y., Zhou, Y., Wu, Y., Gong, C.: Can sam segment polyps? arXiv preprint arXiv:2304.07583 (2023)
35. Zhou, T., Zhou, Y., He, K., Gong, C., Yang, J., Fu, H., Shen, D.: Cross-level feature aggregation network for polyp segmentation. Pattern Recognition **140**, 109555 (2023)