CD-PolypNet: Cross-Domain Polyp Segmentation Network with Internal Feature Distillation and Dual-Stream Boundary Focus via Large Vision Model

Changpeng Yue[1], Jianxiang Zhao[1], Chao Wang[1], Xinglun Zhao[1], Axiu Mao[1], Jia Hou[3], Chenggang Yan[1], Kai Zhao[2], and Shuai Wang[1] (✉)

[1] Hangzhou Dianzi University, Hangzhou, China
[2] First Medical Center, Chinese PLA General Hospital, Beijing, China
[3] Lishui Institute of Hangzhou Dianzi University, Lishui, China
shuaiwang.tai@gmail.com

Abstract. Leveraging large vision models (LVMs), such as the Segment Anything Model (SAM), in medical image analysis presents significant potential to enhance diagnostic efficiency. Existing SAM-based medical segmentation methods inadequately address two critical challenges: rapidly adapting LVMs to medical tasks through few-shot fine-tuning, and the inherent difficulty in distinguishing lesions from anatomically similar background regions in medical images. To overcome these limitations, we propose CD-PolypNet, a novel framework integrating a Semantic Supervision via Feature Distillation (SSFD) and an Edge-Guided Feature Branch (EFB). The SSFD module leverages feature distillation to transfer knowledge from SAM's strongly supervised features into early-stage feature learning, enabling efficient domain adaptation of large vision models under data scarcity. Concurrently, EFB enhances boundary discrimination in lightweight decoder through a hybrid strategy combining the Canny operator and Edge-Frequency Gated Convolution (EFG-Conv), thereby prioritizing edge-aware feature extraction. Extensive experiments across five challenging medical imaging datasets demonstrate that our method not only surpasses state-of-the-art approaches in accuracy and robustness but also establishes a new paradigm for cross-domain adaptation of large vision models in specialized medical applications. The codes are available at https://github.com/ChangpengYue/CD-PolypNet.

Keywords: Polyp segmentation · SAM · Feature distillation.

1 Introduction

In medical diagnostics, manual analysis of medical images by pathologists remains time-consuming and labor-intensive. The emergence of machine learning, particularly convolutional neural networks (CNNs), has driven significant

(✉) Corresponding author.

progress in automated lesion segmentation. CNNs demonstrate notable capabilities in capturing local image patterns, achieving high precision through supervised learning on large annotated datasets. However, medical image annotation requires specialized expertise, resulting in severe data scarcity that fundamentally limits further improvements in supervised learning approaches.

Recent advancements in Transformer architectures have shifted research focus from CNNs by leveraging superior global contextual understanding, outperforming CNNs in natural image processing. Vision Transformer[6] (ViT) -based models exhibit exceptional segmentation accuracy and generalization in natural and remote sensing imagery. However, in specialized domains like medical imaging, such models struggle due to insufficient domain-specific knowledge. In medical imaging, current research adapts Segment Anything Model[2] (SAM) through parameter fine-tuning[7,1], automated prompting[8,33], and framework enhancements[5,10]. Nevertheless, existing approaches fail to address the critical challenge of enabling large vision models to rapidly learn discriminative features from extremely limited annotated data.

Medical image segmentation faces persistent challenges, exemplified by polyp segmentation in endoscopic imaging. Endoscopy, widely used in gastrointestinal examinations, generates images where polyps exhibit ambiguous boundaries due to overlapping textures, colors, and contrast with surrounding tissues. This characteristic severely compromises segmentation accuracy, leading to frequent misidentification or omission of lesion boundaries.

In this paper, we present CD-PolypNet, a novel framework that combines Semantic Supervision via Feature Distillation (SSFD) and an Edge-Guided Feature Branch (EFB). The SSFD employs feature distillation to transfer knowledge from SAM's strongly supervised features to early-stage feature learning, facilitating efficient domain adaptation of large vision models, especially when data is scarce. Simultaneously, the EFB enhances edge detection in lightweight decoder by integrating the Canny operator with Edge-Frequency Gated Convolution (EFGConv), prioritizing edge-aware feature extraction. Extensive experiments on five challenging medical imaging datasets demonstrate that our method outperforms current state-of-the-art approaches in both accuracy and robustness, setting a new benchmark for cross-domain adaptation of large vision models in medical applications.

2 Method

2.1 Overall Architecture

Our proposed architecture builds upon the SAM framework, which utilizes a Masked Autoencoder[14] -pretrained ViT as its image encoder alongside a prompt encoder and mask decoder. The overall framework is shown in Fig. 1. To address the domain-specific challenges of medical image segmentation, we implement two synergistic innovations. First, we freeze SAM's image encoder parameters and introduce a SSFD. This module hierarchically transfers semantically rich features

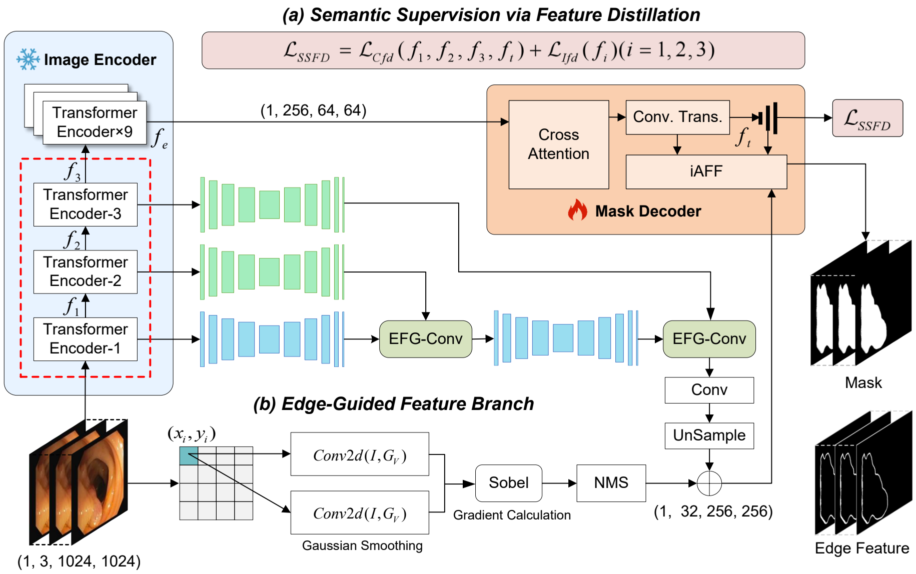

Fig. 1. Overview of the proposed CD-PolypNet, which consists of (a) Feature Distillation via Semantic Supervision (b) Edge-Guided Feature Branch.

from the decoder layers to supervise early-stage encoder features, compelling the encoder to prioritize target regions with minimal training data. By aligning encoder feature learning with decoder-level semantic priors, SSFD bridges the generalization gap between natural and medical imaging domains.

Concurrently, we design an EFB to resolve the persistent weak boundary problem in endoscopic images. The pipeline begins with a Canny[11] operator extracting gradient-based edge maps from raw images. These maps are fused with encoder-derived features through EFGConv, which dynamically gates spatial activations to amplify boundary-related signals. To further reinforce edge sensitivity, multi-scale edge cues are recursively integrated into the encoder through iterative Attentional Feature Fusion[15] (iAFF).

2.2 Feature Distillation via Semantic Supervision

Analyze the Features Learned in SAM To investigate the feature learning dynamics within SAM, we apply class activation mapping (CAM) [12] to visualize features extracted from selected encoder and decoder layers, as shown in Fig. 2. The locations of the visualized features f_1, f_2, f_3, f_e, and f_t are annotated in Fig. 1. The visualization results reveal that early encoder features f_1, f_2 and f_3 exhibit strong global feature representation capabilities and contain rich semantic information. However, during few-shot fine-tuning in LVM, these features struggle to rapidly learn beneficial patterns. In contrast, the feature f_t from the decoder, which benefits from stronger supervision during training,

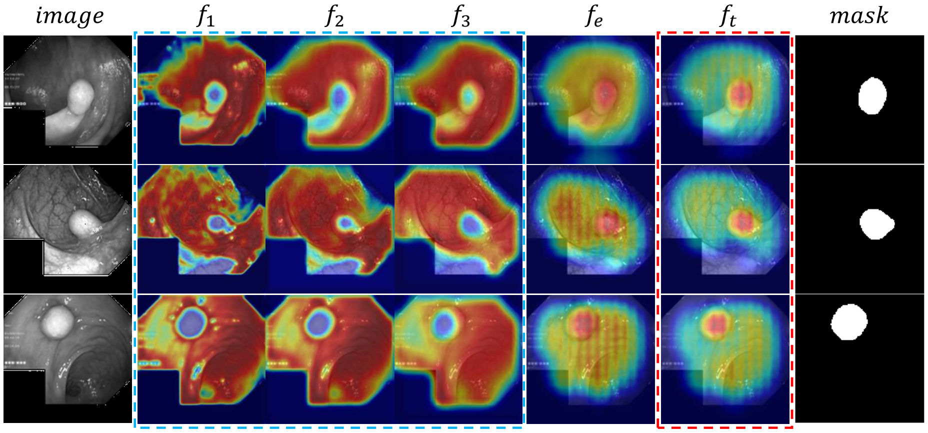

Fig. 2. Visualization of SAM's hierarchical features using CAM. All feature positions are marked in the framework diagram, where f_t denotes the target features, f_1-f_3 represent the supervised features in the SSFD and f_e represents the high-dimensional feature output by the SAM encoder.

demonstrates a higher focus on the target region and contains information that is more aligned with the desired lesion mask.

Some past studies have shown that by adopting strategies such as knowledge distillation and feature alignment, accurate features can be utilized to guide those with less information[3,13]. Inspired by these works, we introduce a feature distillation paradigm that leverages Cross-layer Feature Alignment. Specifically, the encoder feature f_{t}, which benefits from strong supervision, is utilized as the target feature to guide the optimization of the shallow encoder features f_1, f_2 and f_3. Additionally, Intra-layer Feature Decorrelation is applied to f_1, f_2 and f_3 to reduce feature interference and redundancy, thereby enhancing the discriminative quality of the learned representations.

Cross-layer Feature Alignment The framework establishes semantic correspondence between shallow encoder features and deep representations through adaptive spatial matching. Given encoder feature $f_s \in \mathbb{R}^{B\times C_s\times H\times W}$ and target feature $f_t \in \mathbb{R}^{B\times C_t\times H'\times W'}$, we first align their spatial dimensions:

$$\tilde{f}_s = \mathcal{P}(f_s) = \begin{cases} \text{AdaptiveAvgPool}(f_s, (H', W')) & H \neq H' \\ f_s & \text{otherwise} \end{cases}, \tag{1}$$

where $s \in \{1, 2, 3\}$, representing the three early-stage features of the encoder. $\mathcal{P}(\cdot)$ implements resolution matching through adaptive pooling. We then stochastically select $\min(C_s, C_t)$ channels via uniform sampling without replacement, yielding paired channel indices $\mathcal{I}_s \subset [0, C_s)$ and $\mathcal{I}_t \subset [0, C_t)$. The cross-layer distillation loss is computed as:

$$\mathcal{L}_{\mathrm{Cfd}} = \frac{1}{B|\mathcal{I}|} \sum_{b=1}^{B} \sum_{i=1}^{|\mathcal{I}|} \left\| \tilde{f}_s^{(b,\mathcal{I}_s[i])} - f_t^{(b,\mathcal{I}_t[i])} \right\|_2^2 . \tag{2}$$

This process dynamically updates the attention and MLP layers in the mask decoder during training, promoting cross-network depth spatial-semantic consistency between shallow and deep features.

Intra-layer Feature Decorrelation To eliminate channel redundancy within encoder blocks, features undergo normalized self-distillation. For each feature map $f \in \mathbb{R}^{B\times C\times H\times W}$, we first compute channel importance scores through spatial L2-normalization:

$$\hat{f}^{(b,c)} = \frac{f^{(b,c)}}{\|f^{(b,c)}\|_2}, \quad \|f^{(b,c)}\|_2 = \sqrt{\frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} (f^{(b,c,h,w)})^2}. \tag{3}$$

Channels are ranked by their mean activation magnitude across the batch:

$$\tau = \operatorname{argsort} \left(\frac{1}{B} \sum_{b=1}^{B} \|\hat{f}^{(b,c)}\|_2 \right)_{c=1}^{C}, \quad \text{descending order.} \tag{4}$$

The intra-layer loss enforces similarity between top- and bottom-ranked channel groups:

$$\mathcal{L}_{\mathrm{Ifd}} = \frac{1}{BC} \sum_{b=1}^{B} \sum_{c=1}^{C/2} \left\| \hat{f}^{(b,\tau_c)} - \hat{f}^{(b,\tau_{c+C/2})} \right\|_2^2 . \tag{5}$$

This creates a compressed yet discriminative feature representation by penalizing activation disparities between high- and low-saliency channels.

$$\mathcal{L}_{\mathrm{SSFD}} = \mathcal{L}_{\mathrm{Cfd}} + \mathcal{L}_{\mathrm{Ifd}}. \tag{6}$$

Through Equations (2) and (5), our framework achieves simultaneous inter-layer semantic transfer and intra-layer feature compaction, crucial for medical image analysis under limited supervision.

2.3 Edge-Guided Feature Branch

The intrinsic limitations of Transformer architectures in local feature extraction compared to CNNs have been extensively documented [16]. Prior studies [9,17] substantiate that incorporating CNNs branches into Transformer-based frameworks significantly enhances model accuracy for detail-sensitive tasks. To address the issue of weak edges in lesion regions in endoscopic polyp segmentation, we introduce an EFB. The edge gradient of the original image is computed using the Canny operator, while the image features mapped by the image encoder are filtered for EFGConv. The resulting edge features are then combined with the edge gradient. In the mask decoder, we use iAFF for feature fusion to enhance the decoder's focus on edges.

Canny-Based Edge Gradient Extraction The Canny operator [11] computes gradient intensity maps (G_σ) from raw endoscopic images:

$$G_\sigma = \sqrt{(\mathcal{H}_x * I_\sigma)^2 + (\mathcal{H}_y * I_\sigma)^2}, \tag{7}$$

where $\mathcal{H}_x$, $\mathcal{H}_y$ denote horizontal/vertical Sobel kernels, and I_σ represents Gaussian-filtered input images.

Edge-Frequency Gated Convolution The EFGConv enhances edge discrimination through frequency-domain feature recombination. Given encoder features $f_e \in \mathbb{R}^{C\times H\times W}$ and edge map $g_c \in \mathbb{R}^{1\times H\times W}$, the computation proceeds as:

$$\begin{aligned} f_{\text{low}} &= \mathcal{G}_\sigma * f_e, \\ f_{\text{hf}} &= f_e - f_{\text{low}}, \end{aligned} \tag{8}$$

where $\mathcal{G}_\sigma$ denotes a 7×7 Gaussian kernel with standard deviation $\sigma = 3$. The high-frequency component f_{hf} captures fine boundary details suppressed in conventional encoder features.

The spatial attention mechanism combines edge priors with learned semantics:

$$\alpha = \sigma\left(\mathcal{W}_2\left(\text{ReLU}\left(\mathcal{W}_1\left([f_{\text{hf}} \parallel g_c]\right)\right)\right)\right), \tag{9}$$

where $\mathcal{W}_1 \in \mathbb{R}^{(C+1)\times(C+1)}$ and $\mathcal{W}_2 \in \mathbb{R}^{(C+1)\times 1}$ represent 1×1 convolutional transformations, $\parallel$ denotes channel concatenation, and σ is the sigmoid activation.

The final feature map integrates amplified boundary signals with original semantics through:

$$f_{\text{out}} = iAFF(\underbrace{(1+\alpha)\odot f_{\text{hf}}}_{\text{boundary enhancement}}, \underbrace{\mathcal{C}_{3\times 3}(f_e)}_{\text{semantic preservation}}), \tag{10}$$

where $\mathcal{C}_{3\times 3}$ denotes a 3×3 convolution maintaining dimensional consistency. The iAFF introduces a multi-scale channel attention module, which resolves the problem of feature fusion resulting from inconsistent scales and semantics. The use of iAFF to fuse edge features with the original semantics avoids the model's excessive attention to edges.

3 Experiments and Results

3.1 Datasets and Evaluation Metric

Datasets The performance of CDPolypNet is evaluated using five widely recognized benchmark datasets for polyp segmentation: Kvasir-SEG[34], CVC-ClinicD

B[35], CVC-ColonDB[36], EndoScene-CVC300, and ETIS-LaribPolypDB[37], which are extensively adopted in polyp segmentation research. The datasets were randomly partitioned into three subsets: training, validation, and test sets, with a ratio of 80%, 10%, and 10% respectively. To ensure fair comparison with prior SOTA methods, we adopted the fixed test dataset configuration provided by PraNet. Consequently, only Kvasir-SEG and CVC-ClinicDB contain both training and test splits, while CVC-ColonDB, EndoScene-CVC300, and ETIS-LaribPolypDB were exclusively employed for testing purposes.

Evaluation Metric Performance evaluation is performed using two widely used metrics: Dice and IoU. Dice measures the degree of overlap between predicted segmentation and real segmentation. The IoU calculates the ratio of intersection and union between the predicted region and the real region.

3.2 Implementation Details

The proposed framework is implemented in PyTorch and trained on an NVIDIA RTX 3090 GPU. We adopt the Adam optimizer with an initial learning rate of 0.01 and batch size of 4, decaying the rate by 20% every 5 epochs over 30 total epochs. All endoscopic images are resized to 1024×1024 resolution using bicubic interpolation and normalized to [0,1] range. To enhance model robustness, we apply real-time data augmentation including random horizontal flipping and ±15° rotation during training.

3.3 Results

Comparative Results As demonstrated in Table 1, our method surpasses the SOTA polyp segmentation approaches in all benchmark datasets. Notably, CD-PolypNet achieves an average Dice coefficient of 0.917 and an IoU of 0.867, outperforming the previous best method, Polyp-PVT (with an average Dice coefficient of 0.870 and an average IoU of 0.804), by 4.6 and 5.2 percentage points, respectively. Compared to SAM-based methods, our framework exhibits superior performance on all datasets except ClinicDB and EndoScene. Moreover, we observed that ASPS also utilized a CNN branch to improve local feature extraction, achieves strong performance. This further proves that dual-stream architectures work effectively in SAM-based medical frameworks.

It's important to note that our model shows improved generalization on the ColonDB, ETIS, and EndoScene datasets. Our model achieves substantial gains of 9 percentage points in Dice coefficient and 12.9 percentage points in IoU on ColonDB, along with 4.2 percentage points in Dice and 10.8 percentage points in IoU improvements on ETIS compared to SOTA. Additionally, on the EndoScene dataset, our model achieves a Dice score of 0.913, which is very close to the SOTA, with an IoU improvement of 0.2 percentage points. These results confirm that our framework successfully adapts SAM's generalization capabilities to medical imaging domains while preserving its inherent robustness.

Table 1. Performance comparison of different methods on polyp segmentation datasets. **Bold** indicates the best scores and underline denotes the second best.

Methods	Published	Kvasir		ClinicDB		ColonDB		ETIS		EndoScene	
		mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
U-Net[18]	MICCAI'15	0.818	0.746	0.823	0.755	0.504	0.436	0.398	0.335	0.710	0.627
PraNet[19]	MICCAI'19	0.898	0.840	0.913	0.84	0.709	0.640	0.628	0.567	0.871	0.797
UNet++[27]	TMI'20	0.821	0.743	0.794	0.729	0.504	0.436	0.401	0.344	0.707	0.642
SANet[20]	CVPR'20	0.904	0.847	0.916	0.859	0.753	0.670	0.750	0.654	0.888	0.815
TransFuse[21]	MICCAI'21	0.920	0.870	0.942	0.897	0.781	0.706	0.737	0.826	0.894	0.654
ADSNet[28]	JAG'21	0.920	0.871	0.938	0.890	0.815	0.730	0.798	0.715	0.890	0.819
LDNet[22]	MICCAI'22	0.912	0.855	0.932	0.872	0.794	0.715	0.778	0.707	0.893	0.826
SSFormer[23]	MICCAI'22	0.917	0.864	0.906	0.855	0.802	0.721	0.796	0.720	0.895	0.827
Polyp-PVT[30]	TMI'22	0.917	0.864	0.937	0.889	0.808	0.727	0.787	0.706	0.900	0.833
CFANet[29]	PR'23	0.915	0.861	0.933	0.883	0.743	0.665	0.732	0.655	0.893	0.827
TransUnet[31]	MIA'24	0.913	0.857	0.935	0.887	0.781	0.699	0.731	0.824	0.893	0.660
SAM-H	ICCV'23	0.778	0.707	0.547	0.500	0.441	0.396	0.517	0.477	0.651	0.606
SAM-L	ICCV'23	0.782	0.710	0.579	0.526	0.468	0.422	0.551	0.507	0.726	0.676
SAM-Adapter[24]	ICCV'23	0.847	0.763	0.774	0.673	0.671	0.568	0.590	0.476	0.815	0.725
AutoSAM[33]	ArXiv'23	0.784	0.675	0.751	0.642	0.535	0.418	0.402	0.308	0.829	0.739
SAMPath[32]	MICCAI'23	0.828	0.730	0.750	0.644	0.535	0.418	0.555	0.442	0.844	0.756
SurgicalSAM[25]	AAAI'24	0.740	0.597	0.644	0.505	0.460	0.330	0.342	0.238	0.623	0.472
MedSAM[4]	Nature'24	0.862	0.795	0.867	0.803	0.734	0.651	0.687	0.604	0.870	0.798
ASPS[26]	MICCAI'24	0.920	0.858	**0.951**	**0.906**	0.799	0.701	0.861	0.769	**0.919**	0.852
Proposed		**0.933**	**0.877**	0.945	0.898	**0.889**	**0.830**	**0.903**	**0.877**	0.913	**0.854**

Ablation Results We conducted comprehensive ablation studies across five datasets to validate the efficacy of our proposed components. As summarized in Table 2, the baseline model (fine-tuned SAM) achieves 0.792 Dice coefficient and 0.721 IoU, while separate integration of SSFD and EFB improves these metrics by 9.5 percentage points in Dice and 12.5 percentage points in IoU for SSFD, and 10.1 percentage points in Dice and 13.5 percentage points in IoU for EFB, respectively. The CD-PolypNet, combining both modules, demonstrates synergistic enhancement with 12.5 percentage points improvement in Dice and 14.6 percentage points improvement in IoU compared to the baseline.

Table 2. Ablation study of architectures composed of different modules across five distinct benchmarks. **Bold** indicates the best scores.

SAM-L	SSFD	EFB	Kvasir		ClinicDB		ColonDB		ETIS		EndoScene	
			mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
✓			0.782	0.710	0.579	0.526	0.468	0.422	0.551	0.507	0.726	0.676
✓	✓		0.926	0.867	0.937	0.884	0.859	0.808	0.824	0.838	0.890	0.832
✓		✓	0.930	0.874	0.943	0.894	0.862	0.813	0.831	0.865	0.900	0.832
✓	✓	✓	**0.933**	**0.877**	**0.945**	**0.898**	**0.889**	**0.830**	**0.903**	**0.877**	**0.913**	**0.854**

4 Conclusion

We proposed CD-PolypNet, a cross-domain polyp segmentation network with internal feature distillation and dual-stream boundary focus via large vision model. The proposed SSFD enables efficient knowledge transfer from foundation models under limited medical annotations. EFB explicitly addresses the persistent challenge of weak boundary discrimination in endoscopic imaging. Extensive evaluations demonstrate the superior capability of our method in capturing clinically critical regions and maintaining robustness across diverse colonoscopy environments. This work demonstrates an effective pathway to tailor general vision models for medical segmentation tasks while preserving their intrinsic generalization strengths.

Acknowledgments. This research was supported by National Natural Science Foundation of China under Grant No. 62201323, Zhejiang Provincial Natural Science Foundation of China under Grant No. LDT23F01015F01 and No. LD25F020005, and Basic Research Program of Jiangsu No. BK20220266.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Li, Y., Hu, M., Yang, X.: Polyp-sam: Transfer sam for polyp segmentation. In: Medical Imaging 2024: Computer-Aided Diagnosis, vol. 12927, pp. 759–765. SPIE (2024)
2. Kirillov, A., Mintun, E., Ravi, N., et al.: Segment anything. In: IEEE/CVF International Conference on Computer Vision, pp. 4015–4026. IEEE (2023)
3. Zhu, W., Chen, X., Qiu, P., et al.: Selfreg-unet: Self-regularized unet for medical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention, pp. 601–611. Springer (2024)
4. Ma, J., He, Y., Li, F., et al.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
5. Huang, X., Yue, C., Guo, Y., et al.: Multidimensional directionality-enhanced segmentation via large vision model. Medical Image Analysis **101**, 103395 (2025)
6. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
7. Hu, M., Li, Y., Yang, X.: Skinsam: Empowering skin cancer segmentation with segment anything model. arXiv preprint arXiv:2304.13973 (2023)
8. Anand, D., Singhal, V., Shanbhag, D.D., et al.: One-shot localization and segmentation of medical images with foundation models. arXiv preprint arXiv:2310.18642 (2023)
9. Valanarasu, J.M.J., Patel, V.M.: Unext: Mlp-based rapid medical image segmentation network. In: Medical Image Computing and Computer-Assisted Intervention, pp. 23–33. Springer (2022)
10. Chai, S., Jain, R.K., Teng, S., et al.: Ladder fine-tuning approach for sam integrating complementary network. Procedia Computer Science **246**, 4951–4958 (2024)

11. Canny, J.: A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **PAMI-8**(6), 679–698 (1986). https://doi.org/10.1109/TPAMI.1986.4767851
12. Zhou, B., Khosla, A., Lapedriza, A., et al.: Learning deep features for discriminative localization. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929. IEEE (2016)
13. Chen, D., Mei, J.-P., Zhang, Y., et al.: Cross-layer distillation with semantic calibration. In: AAAI Conference on Artificial Intelligence, vol. 35(8), pp. 7028–7036. AAAI (2021)
14. He, K., Chen, X., Xie, S., et al.: Masked autoencoders are scalable vision learners. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16000–16009. IEEE (2022)
15. Dai, Y., Gieseke, F., Oehmcke, S., et al.: Attentional feature fusion. In: IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3560–3569. IEEE (2021)
16. Raghu, M., Unterthiner, T., Kornblith, S., et al.: Do vision transformers see like convolutional neural networks?. Advances in Neural Information Processing Systems **34**, 12116–12128 (2021)
17. Wang, Z., Cun, X., Bao, J., et al.: Uformer: A general u-shaped transformer for image restoration. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17683–17693. IEEE (2022)
18. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention, pp. 234–241. Springer (2015)
19. Fan, D.-P., Ji, G.-P., Zhou, T., et al.: Pranet: Parallel reverse attention network for polyp segmentation. In: Medical Image Computing and Computer-Assisted Intervention, pp. 263–273. Springer (2020)
20. Zhong, Z., Lin, Z.Q., Bidart, R., et al.: Squeeze-and-attention networks for semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13065–13074. IEEE (2020)
21. Zhang, Y., Liu, H., Hu, Q.: Transfuse: Fusing transformers and cnns for medical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention 2021, pp. 14–24. Springer (2021)
22. Zhang, R., Lai, P., Wan, X., et al.: Lesion-aware dynamic kernel for polyp segmentation. In: Medical Image Computing and Computer-Assisted Intervention, pp. 99–109. Springer (2022)
23. Wang, J., Huang, Q., Tang, F., et al.: Stepwise feature fusion: Local guides global. In: Medical Image Computing and Computer-Assisted Intervention, pp. 110–120. Springer (2022)
24. Chen, T., Zhu, L., Deng, C., et al.: Sam-adapter: Adapting segment anything in underperformed scenes. In: IEEE/CVF International Conference on Computer Vision, pp. 3367–3375. IEEE (2023)
25. Yue, W., Zhang, J., Hu, K., et al.: Surgicalsam: Efficient class promptable surgical instrument segmentation. In: AAAI Conference on Artificial Intelligence, vol. 38(7), pp. 6890–6898. AAAI (2024)
26. Li, H., Zhang, D., Yao, J., et al.: Asps: Augmented segment anything model for polyp segmentation. In: Medical Image Computing and Computer-Assisted Intervention, pp. 118–128. Springer (2024)
27. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., et al.: Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE Transactions on Medical Imaging **39**(6), 1856–1867 (2019)

28. Wang, D., Chen, X., Jiang, M., et al.: ADS-Net: An attention-based deeply supervised network for remote sensing image change detection. International Journal of Applied Earth Observation and Geoinformation **101**, 102348 (2021)
29. Zhou, T., Zhou, Y., He, K., et al.: Cross-level feature aggregation network for polyp segmentation. Pattern Recognition **140**, 109555 (2023)
30. Dong, B., Wang, W., Fan, D.-P., et al.: Polyp-pvt: Polyp segmentation with pyramid vision transformers. arXiv preprint arXiv:2108.06932 (2021)
31. Chen, J., Lu, Y., Yu, Q., et al.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
32. Zhang, J., Ma, K., Kapse, S., et al.: Sam-path: A segment anything model for semantic segmentation in digital pathology. In: Medical Image Computing and Computer-Assisted Intervention, pp. 161–170. Springer (2023)
33. Shaharabany, T., Dahan, A., Giryes, R., et al.: Autosam: Adapting sam to medical images by overloading the prompt encoder. arXiv preprint arXiv:2306.06370 (2023)
34. Jha, D., Smedsrud, P.H., Riegler, M.A., et al.: Kvasir-seg: A segmented polyp dataset. In: MultiMedia Modeling (MMM 2020), pp. 451–462. Springer (2020)
35. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., et al.: WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized Medical Imaging and Graphics **43**, 99–111 (2015)
36. Bernal, J., Sánchez, F.J., Vilariño, F.: Towards automatic polyp detection with a polyp appearance model. Pattern Recognition **45**(9), 3166–3182 (2012)
37. Silva J, Histace A, Romain O, et al. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer[J]. International journal of computer assisted radiology and surgery, 2014, 9: 283-293.