npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-026-02345-7

PrysmNet a polyp refining system using salience and multimodal guidance for reproducible cross domain segmentation

Junbo Xiao[1,2], Yi Han[1,2], Lei Wang[3], Ying Li[4], Xiaotong Wang[1,2], Shizhe Li[1,2], Jun Yi[1,2], Yu Wu[5] ✉ & Xiaowei Liu[1,2,6,7] ✉

Colorectal cancer prevention benefits from accurate and reproducible polyp segmentation, yet cross-domain generalization and boundary precision remain challenging in real-world deployments. We propose Prysm-Net, a ViT-based framework designed to address these issues through architectural innovation and advanced training guidance. Our model is augmented with a biologically inspired salience module (BSM) that dynamically sharpens boundary-relevant features. To further enhance robustness without increasing inference costs, we introduce two training-only strategies: (i) foundation-model distillation from SAM, which transfers knowledge at the output, boundary, and feature levels, and (ii) multi-modal guidance that injects auxiliary structural and textural cues via gated cross-attention. Extensive experiments on standard in-domain benchmarks and challenging cross-domain datasets demonstrate that Prysm-Net achieves superior segmentation accuracy and robust generalization compared to state-of-the-art methods, all while maintaining a lightweight inference process by disabling auxiliary guidance at test time.

Colorectal cancer (CRC) is a leading cause of cancer mortality worldwide, and most CRC cases develop from benign colorectal polyps over time[1]. Early detection and removal of polyps during colonoscopy can drastically reduce CRC incidence and mortality[2,3]. Colonoscopy is the gold standard for polyp screening and intervention, but it remains a highly operator-dependent procedure. In practice, the diversity of polyp shapes, sizes, and appearances makes some lesions (especially diminutive or flat polyps < 10 mm) easy to miss or difficult to delineate accurately[4,5]. Studies report that roughly 17–28% of polyps may be missed even by experts during colonoscopy exams[2,5]. Precise segmentation of polyps can assist endoscopists by highlighting lesion boundaries for complete resection and by reducing inter-observer variability. However, producing pixel-accurate polyp masks is labor-intensive and costly, as it requires expert manual annotation[3,5]. This creates a strong motivation for automated, high-accuracy polyp segmentation methods to improve early CRC prevention and alleviate clinical workload[6].

Over the past few years, deep-learning-based polyp segmentation methods have achieved remarkable in-domain performance on standard datasets. In controlled evaluations, many models report high mean Dice scores (often 0.85–0.95) on popular benchmarks[7,8]. Nevertheless, a critical gap remains between such in-domain success and real-world generalization. Most existing approaches are trained and tested on data from the same domain and implicitly assume similar distributions, an assumption that breaks down in clinical deployment. In practice, factors like different endoscopy centers, imaging devices, patient populations, and bowel preparation protocols lead to distribution shifts that can significantly degrade model performance[9,10]. For example, models that approach near human accuracy on seen datasets often plummet to Dice scores around 0.6–0.7 on unseen datasets such as the recent multi-center PolypGen collection[2,4]. Moreover, certain challenging polyp cases remain problematic: small polyps occupying only a tiny fraction of the frame and polyps with faint or blurred boundaries are frequently segmented poorly by current models[8,11]. These observations underscore the need for new research focusing not only on

[1]Department of Gastroenterology, Xiangya Hospital, Central South University, Changsha, Hunan, China. [2]National Clinical Research Center for Geriatric Disorders, Xiangya Hospital, Central South University, Changsha, Hunan, China. [3]School of Computer Science and Engineering, Central South University, Changsha, Hunan, China. [4]Department of Infectious Diseases, Xiangya Hospital, Central South University, Hunan Key Laboratory of Viral Hepatitis, Changsha, Hunan, China. [5]Teaching and Research Section of Clinical Nursing, Xiangya Hospital, Central South University, Changsha, Hunan, China. [6]Hunan International Scientific and Technological Cooperation Base of Artificial Intelligence Computer-Aided Diagnosis and Treatment for Digestive Disease, Changsha, Hunan, China. [7]Gut-Liver Axis and Intestinal Barrier Research Center, Xiangya Hospital, Central South University, Changsha, Hunan, China. ✉e-mail: wuyu527@csu.edu.cn; liuxw@csu.edu.cn

maximizing in-domain metrics, but also on ensuring robust out-of-domain performance and explicit handling of hard cases (e.g., diminutive or weak boundary lesions)[12].

In this paper, we address these gaps by proposing PrysmNet (a Polyp Refining sYstem with Salience and Multi-modal guidance), a comprehensive solution for reproducible cross-domain polyp segmentation. Our approach combines architectural innovation with advanced training strategies to improve generalization, especially for small and boundary-challenged polyps. Our contributions are presented along two main axes:

1. *An architectural innovation*: A biologically inspired salience module (BSM) that mimics the feature amplification mechanisms of the human visual cortex to dynamically enhance features along lesion boundaries. The BSM employs a learnable multi-scale edge and texture filter bank to compute saliency maps that gate and amplify features, forcing the network to allocate more computational resources to critical boundaries and resulting in significantly sharper segmentation masks.
2. *An advanced training guidance strategy*: This strategy combines two synergistic mechanisms: (a) a multi-level foundation model distillation module (FMDM) that transfers both output-level and feature-level knowledge from SAM using temperature-scaled KL divergence and representational similarity, and (b) a multi-modal guidance module (MGM) that uses self-generated auxiliary structural and textural information to create a more invariant feature representation through cross-attention mechanisms.

Early research on automatic polyp segmentation was constrained by limited data, with algorithms often evaluated on small single-center datasets. Notable early benchmarks include CVC-ClinicDB (612 colonoscopy images with polyp masks)[13] and Kvasir-SEG (1000 images)[14], which provided a valuable testbed but cover a narrow distribution of imaging conditions. Models trained and tuned on one benchmark sometimes showed decreased performance when tested on another, hinting at dataset bias. To improve comparability, some studies adopted combined training sets or reported cross-dataset results (e.g., testing a model trained on Kvasir-SEG/CVC on the ETIS-LaribPolypDB dataset)[8]. However, until recently, there was no standardized protocol for cross-domain evaluation, making it difficult to assess true generalizability. As a result, many publications have reported only in-domain metrics, which do not guarantee real-world clinical reliability.

Recognizing these issues, the community has moved toward more comprehensive evaluation frameworks. Bernal et al.[15] emphasized the lack of a common validation standard and released a complete benchmark covering detection, segmentation, and classification tasks, including considerations of execution speed and robustness. Meanwhile, multiple challenge competitions (MICCAI EndoScene, GIANA, etc.) and the EndoCV series have encouraged participants to test algorithms on hidden multi-center test sets. Ali et al.[4] recently organized a polyp detection and segmentation challenge using a large multi-center dataset, demonstrating that methods with top in-lab accuracy often struggled on diverse "unseen" colonoscopy data. In parallel, new datasets have emerged aimed at assessing generalization. PolypGen, introduced by Ali et al.[2], is a prime example: it consists of 3762 images from six different centers with expert annotations, expressly designed to evaluate algorithm performance across varying patient populations and imaging devices. The introduction of PolypGen and similar multi-center datasets is a significant step toward benchmarking real-world performance. In this work, we build on these efforts by proposing a cross-domain evaluation protocol that uses combined source training and held-out diverse test sets, enabling a reproducible assessment of how segmentation models generalize beyond their training distribution.

Convolutional neural networks (CNNs) have formed the backbone of most polyp segmentation methods to date. The seminal U-Net architecture[16] and its extensions (such as U-Net++[17] and ResUNet++[18]) were early choices, owing to their ability to capture multi-scale context through encoder-decoder designs. Numerous specialized CNN variants have been proposed to boost segmentation accuracy. For instance, Fan et al.'s PraNet[7] introduced a parallel reverse attention mechanism to progressively refine predictions and achieved state-of-the-art results on Kvasir-SEG and CVC-ClinicDB. Similarly, DoubleU-Net and other cascaded architectures stack multiple decoder stages to improve segmentation quality on difficult regions[19]. These CNN-based approaches effectively leverage local spatial features and have demonstrated high precision under favorable conditions.

A key consideration for practical deployment, however, is model efficiency. Colonoscopy is a real-time procedure, so inference speed and model size are critical. Many high-performing CNN models use complex decoder paths or heavy backbones (e.g., ResNet-101), which can hinder real-time performance. Recent research has thus explored lightweight networks that sacrifice minimal accuracy while greatly improving efficiency. For example, Yu et al.[8] proposed HarDNet-CPS, which uses a Harmonic Densely Connected backbone known for its low computational cost, combined with multi-scale feature fusion. Their model surpassed 0.90 mean Dice on CVC-ClinicDB and Kvasir-SEG while being faster and more compact than many predecessors. Dumitru et al.[3] presented DUCK-Net, an efficient CNN with a custom convolutional kernel and residual downsampling; notably, it achieved competitive accuracy even when trained on relatively small datasets. In general, there is a trend toward lighter architectures that maintain strong segmentation performance. Such models, often integrating attention or efficient blocks, are better suited for clinical integration where memory and processing time are limited[20]. Our work follows this trend by proposing a new model enhanced for generalization, ensuring that improved robustness does not come at the cost of impractical model complexity.

Vision transformers have recently gained popularity in medical image segmentation, including for colonoscopic polyps, due to their strength in modeling long-range dependencies. Several transformer-based polyp segmentation models have been introduced, often demonstrating top-tier accuracy. For example, researchers have applied the Pyramid Vision Transformer (PVT) as an encoder in a cascade architecture[21], and others developed specialized transformer networks like CTNet[22] and Polyp-Transformer[23] to capture global context. These transformer-driven approaches have achieved new state-of-the-art results on standard polyp datasets, frequently outperforming pure CNN models in terms of Dice and IoU metrics. The ability of transformers to attend globally is particularly useful for colonoscopy images, which may contain complex backgrounds and widely separated object regions.

However, transformers also introduce challenges. By design, self-attention layers are computationally intensive, and dense global attention can miss subtle local details (e.g., fine texture or edges) that CNN kernels capture well. Polyp segmentation methods purely based on transformers sometimes produce overly smooth masks, lacking the precision on boundaries that CNN-based methods can provide. To address this, a number of hybrid architectures have been proposed that combine CNN and transformer components. These aim to harness the strengths of each: CNNs excel at high-frequency local feature extraction, while transformers contribute global awareness. He et al.[5] designed CTHP, a CNN-Transformer Hybrid Polyp segmentation model, which processes feature maps with both convolution and self-attention in parallel. By introducing a more efficient attention mechanism and a new information propagation module, their hybrid model achieved competitive accuracy with significantly reduced computation. Similarly, another work proposed TransFuse, fusing a transformer branch and a CNN branch in the decoder to improve both global coherence and boundary clarity[24]. The success of such hybrids suggests that local and global feature representations are complementary for polyp segmentation. Recent transformer-based polyp segmentation methods, such as Polyper[25] and Polyp-PVT[21] focus primarily on global context modeling and multi-scale feature fusion, but do not explicitly enforce boundary supervision or leverage training time foundation model distillation. In contrast, our PrysmNet distinguishes itself by: (i) incorporating a dedicated Biologically Inspired Salience Module (BSM) that explicitly

supervises boundary detection and refines features at critical boundary regions, and (ii) employing training-only guidance strategies (SAM distillation and multi-modal fusion) that improve generalization without increasing inference cost. This boundary-focused approach, combined with foundation model knowledge transfer, addresses the specific challenges of cross-domain generalization and small/weak boundary polyp segmentation that remain problematic in pure transformer architectures. Our work leverages a Vision Transformer backbone for its powerful global context modeling, augmented with specialized modules to enhance local details and boundary precision, thereby combining the strengths of both paradigms.

Given that the most clinically significant failures are missing a small polyp or inaccurately segmenting a lesion's boundary (which can lead to incomplete resection), researchers have devoted effort to methods targeting these specific challenges. Boundary-aware segmentation is a recurring theme. Many models have been augmented with explicit edge detection or boundary refinement modules. For instance, PraNet's reverse attention mechanism effectively acts as a boundary refiner by learning from prediction errors at the object edges[7]. Other approaches add a dedicated branch to predict polyp contours or boundary pixels alongside the primary segmentation mask[26], enforcing the network to learn sharp transitions between polyp and background. Techniques like contour loss or boundary-enhanced loss functions have also been shown to improve the delineation of polyps with indistinct borders.

Small polyp segmentation is equally challenging, as tiny polyps often occupy only a few hundred pixels in an image and can be easily overlooked. Data imbalance exacerbates this: large polyps dominate the training signal, potentially biasing models. To combat this, prior works have employed hard negative mining and data augmentation, focusing on small polyps. For example, some studies synthetically enlarge small polyp regions or oversample images containing small polyps during training[27]. Ali et al.[4] note that models in their generalization challenge struggled especially with diminutive polyps, suggesting that specialized handling is needed. A few methods introduce multi-scale feature pyramids or super-resolution modules to ensure even tiny polyps can be recognized in high-resolution feature maps[28]. Yu et al.[8] emphasized edge information to avoid merging nearby small polyps. Despite these advances, small and low-contrast polyps remain a difficult corner case for most algorithms. Our proposed approach addresses this by including small polyp-aware training (through oversampling and hard example emphasis) and by embedding boundary priors directly into the model's design.

The advent of powerful foundation models for segmentation has opened new avenues for medical image analysis. In particular, Meta AI's segment anything model (SAM)[29] demonstrated a prompt-driven segmentation capability on a broad range of natural images. SAM is trained on over a billion masks and is designed to generalize to virtually any image distribution. Early investigations show that SAM's zero-shot outputs on endoscopic images are reasonable for prominent polyps but often miss small or flat lesions, likely due to the domain gap[30]. Simply applying SAM off-the-shelf in colonoscopy can yield under-segmentation or over-segmentation, especially when multiple irregular blobs confuse its prompt mechanism.

Researchers have approached this challenge by adapting or incorporating SAM for polyp tasks. One direction is fine-tuning SAM on medical data. For example, a "MedSAM" model tuned on diverse medical images has been reported to improve segmentation performance on colonoscopy frames[30]. Wang et al. proposed PSF-SAM, an efficient fine-tuning strategy that mitigates catastrophic forgetting while boosting SAM's performance in few-shot polyp segmentation scenarios[31]. Another direction keeps SAM frozen but uses it to assist a smaller polyp-specific model. Dutta et al.[11] introduced SAM-MaGuP, which integrates SAM's image encoder with a lightweight adapter and a boundary-distillation module. Similarly, Zhang et al. proposed SAM-EG, where SAM provides initial masks that are then refined by an edge-guided network tailored for polyps. In weakly supervised settings, Zhao et al. introduced a collaborative learning scheme where SAM-generated masks from scribble inputs guide a student polyp segmentation network[32]. Overall, the integration of foundation models like SAM is a promising trend. Our work aligns with this trend by using SAM in the training loop to provide multi-level guidance, resulting in a more robust segmenter without requiring SAM at inference time.

Robust cross-domain polyp segmentation has been pursued via both data-level and model-level strategies. Data-level approaches aim to augment or modify the training data such that the model learns domain-invariant features. A common technique is style augmentation: randomizing image appearance (colors, contrasts, noise) to mimic the variability between different clinics' data. Poudel and Lee introduced a feature space style conversion module that transfers diverse textures and illuminations into training images[10].

On the model-level side, one line of work is unsupervised domain adaptation (UDA), where a model adapts to an unlabeled target domain. Yang et al.[9] proposed a mutual prototype adaptation framework using self-training with pseudo-labels on target images and introduced a contrastive adaptation approach that pulls the source and target feature distributions together. Beyond UDA, domain generalization (DG) algorithms attempt to train a model that generalizes to any unseen domain. Approaches in DG include using multiple source domains with techniques like domain-specific batch normalization or adversarially learned domain invariant features[33]. Federated learning frameworks have also been explored, where a model is trained collaboratively on data from multiple hospitals without centralizing the data[34].

Finally, hard example mining and curriculum learning have proven useful in the context of generalization. By identifying and up-weighting difficult training examples, the model can progressively become immune to failures on those types of inputs[35]. In polyp segmentation, this might involve mining frames where the model missed a small polyp or had a large false positive. Our proposed training paradigm leverages insights from these works (e.g., style randomness, balanced sampling, and hard example emphasis) to construct a training process that yields robust models.

Results

This section presents quantitative comparisons across a comprehensive benchmark suite. To ensure a fair evaluation, we adopt rigorous protocols tailored to each experimental setting. For in-domain benchmarks (Kvasir-SEG, CVC-ClinicDB, ColonDB, EndoScene, and ETIS), models were trained and evaluated on their respective datasets. In contrast, to rigorously test cross-domain generalization, evaluation on the challenging ETIS-LaribPolypDB dataset was performed using models trained exclusively on the combined Kvasir-SEG and CVC-ClinicDB datasets, without any target-domain tuning.

Reporting protocol and metric choices

Recent polyp segmentation papers consistently report mean Dice (mDice) and mean Intersection-over-Union (mIoU) as the primary metrics on Kvasir-SEG, CVC-ClinicDB, and ETIS. In this work, we report mDice, mIoU, and Boundary F-measure (BF) to provide a comprehensive evaluation of segmentation quality. All metrics are computed with a unified evaluation script (single-scale, no TTA, threshold 0.5). We report 95% bootstrap confidence intervals to ensure statistical reliability.

In-domain performance comparison

We compare PrysmNet against state-of-the-art methods on standard in-domain benchmarks, as summarized in Table 1. Regarding overall segmentation quality, indicated by mDice and mIoU metrics, the proposed framework exhibits strong capability in accurately localizing and segmenting polyp regions. This trend is observed not only on standard datasets like Kvasir-SEG and CVC-ClinicDB but also extends to more challenging scenarios, such as ColonDB and ETIS, which feature complex mucous backgrounds and variable lesion contrasts. In addition to regional overlap, the model shows notable strengths in boundary delineation, as reflected by the Boundary *F*-measure (BF). The favorable results in this metric suggest that the integration of the biologically inspired salience module (BSM) effectively refines the structural details of the segmentation masks.

Table 1 | Comparison with state-of-the-art methods across five datasets (Kvasir-SEG, CVC-ClinicDB, ColonDB, EndoScene, ETIS)

Method	Complexity		Kvasir			ClinicDB			ColonDB			EndoScene			ETIS		
	Params	FLOPs	mDice	mIoU	BF	mDice	mIoU	BF	mDice	mIoU	BF	mDice	mIoU	BF	mDice	mIoU	BF
U-Net[16]	24.56	38.26	81.86 ± 0.42	74.61 ± 0.45	68.42 ± 0.58	82.33 ± 0.40	75.56 ± 0.43	69.16 ± 0.55	51.21 ± 0.85	44.46 ± 0.88	38.58 ± 1.10	71.08 ± 0.52	62.61 ± 0.55	58.28 ± 0.68	39.80 ± 0.95	33.58 ± 0.98	28.48 ± 1.20
UNet++[17]	25.09	84.30	82.11 ± 0.40	74.31 ± 0.43	69.10 ± 0.55	79.49 ± 0.41	72.94 ± 0.44	67.82 ± 0.58	48.39 ± 0.82	41.08 ± 0.85	36.24 ± 1.05	70.76 ± 0.50	62.49 ± 0.53	58.08 ± 0.65	46.18 ± 0.88	34.49 ± 0.92	32.12 ± 1.15
SFA[35]	–	–	72.33 ± 0.55	61.13 ± 0.58	58.38 ± 0.70	70.00 ± 0.52	60.73 ± 0.56	56.48 ± 0.68	46.97 ± 0.90	34.77 ± 0.95	33.18 ± 1.25	46.79 ± 0.75	32.96 ± 0.78	30.58 ± 0.95	29.73 ± 1.10	21.74 ± 1.15	18.91 ± 1.30
ACSNet[27]	46.02	29.45	89.87 ± 0.20	83.03 ± 0.22	79.55 ± 0.30	88.09 ± 0.21	82.64 ± 0.23	78.28 ± 0.32	71.62 ± 0.45	64.99 ± 0.48	59.47 ± 0.60	86.30 ± 0.25	78.79 ± 0.28	75.11 ± 0.35	57.98 ± 0.65	50.05 ± 0.68	45.35 ± 0.80
PraNet[7]	32.50	221.90	89.84 ± 0.18	84.08 ± 0.20	80.22 ± 0.28	89.90 ± 0.19	84.95 ± 0.21	81.51 ± 0.29	70.95 ± 0.42	64.08 ± 0.45	58.83 ± 0.58	87.16 ± 0.22	79.70 ± 0.25	76.44 ± 0.33	62.84 ± 0.60	56.75 ± 0.63	51.24 ± 0.75
SANet[28]	–	–	90.44 ± 0.16	86.47 ± 0.18	82.12 ± 0.25	93.74 ± 0.15	88.96 ± 0.17	85.36 ± 0.24	75.35 ± 0.38	67.08 ± 0.40	62.12 ± 0.52	88.85 ± 0.20	81.57 ± 0.22	78.25 ± 0.30	75.08 ± 0.55	65.42 ± 0.58	60.59 ± 0.70
TransFuse[24]	115.59	38.73	91.89 ± 0.12	86.85 ± 0.14	83.43 ± 0.20	93.48 ± 0.11	88.69 ± 0.13	85.13 ± 0.19	74.43 ± 0.35	67.63 ± 0.38	61.84 ± 0.48	90.42 ± 0.15	83.86 ± 0.18	80.54 ± 0.25	73.72 ± 0.50	66.15 ± 0.53	61.21 ± 0.65
TransUNet[29]	105.28	24.66	91.34 ± 0.14	85.70 ± 0.16	82.51 ± 0.22	93.56 ± 0.12	88.79 ± 0.14	84.98 ± 0.20	78.14 ± 0.32	69.99 ± 0.35	64.29 ± 0.45	89.30 ± 0.18	82.47 ± 0.20	79.11 ± 0.28	73.11 ± 0.48	66.08 ± 0.50	59.82 ± 0.62
HarDNet-MSEG[10]	33.80	192.74	92.12 ± 0.11	85.97 ± 0.13	83.20 ± 0.19	93.27 ± 0.10	88.25 ± 0.12	84.80 ± 0.18	73.16 ± 0.33	66.09 ± 0.36	60.55 ± 0.45	90.88 ± 0.14	83.29 ± 0.16	80.14 ± 0.24	77.78 ± 0.45	69.02 ± 0.48	63.46 ± 0.58
DS-TransUNet[41]	177.44	30.97	93.44 ± 0.09	88.82 ± 0.11	85.61 ± 0.17	93.83 ± 0.09	89.13 ± 0.10	86.22 ± 0.16	79.83 ± 0.28	71.72 ± 0.31	66.31 ± 0.40	90.06 ± 0.12	82.84 ± 0.14	80.45 ± 0.22	77.09 ± 0.42	69.01 ± 0.45	64.13 ± 0.55
SwinE-Net[42]	–	–	92.08 ± 0.10	87.04 ± 0.12	84.54 ± 0.18	93.87 ± 0.09	89.20 ± 0.11	86.59 ± 0.17	80.44 ± 0.26	72.51 ± 0.29	67.17 ± 0.38	90.63 ± 0.13	84.09 ± 0.15	81.27 ± 0.23	75.87 ± 0.40	68.75 ± 0.43	62.88 ± 0.53
Polyp-PVT[21]	–	–	91.77 ± 0.10	86.44 ± 0.12	83.82 ± 0.18	93.73 ± 0.09	88.90 ± 0.11	85.71 ± 0.17	80.69 ± 0.25	72.78 ± 0.28	67.48 ± 0.36	90.08 ± 0.12	83.31 ± 0.15	80.39 ± 0.22	78.77 ± 0.38	70.67 ± 0.41	65.22 ± 0.60
CaraNet[43]	46.64	21.69	91.80 ± 0.11	86.61 ± 0.13	83.55 ± 0.20	93.62 ± 0.10	88.83 ± 0.12	85.50 ± 0.18	77.35 ± 0.29	68.91 ± 0.32	63.26 ± 0.42	90.31 ± 0.14	83.31 ± 0.16	80.64 ± 0.24	74.71 ± 0.42	67.20 ± 0.45	61.56 ± 0.56
ColonFormer[44]	52.94	22.94	92.42 ± 0.08	87.66 ± 0.10	85.11 ± 0.16	94.01 ± 0.08	89.70 ± 0.10	86.84 ± 0.15	81.14 ± 0.24	73.33 ± 0.27	68.51 ± 0.35	90.67 ± 0.11	84.21 ± 0.13	81.05 ± 0.20	80.13 ± 0.35	72.21 ± 0.38	66.81 ± 0.48
SegT[45]	–	–	92.77 ± 0.07	88.01 ± 0.09	85.58 ± 0.15	94.02 ± 0.07	89.74 ± 0.09	87.18 ± 0.14	81.43 ± 0.23	73.07 ± 0.26	68.84 ± 0.34	89.54 ± 0.11	82.88 ± 0.13	79.59 ± 0.21	81.03 ± 0.32	73.26 ± 0.35	67.54 ± 0.45
Polyper (final)[25]	29.11	43.54	94.87 ± 0.05	90.30 ± 0.07	88.24 ± 0.12	94.48 ± 0.05	89.89 ± 0.08	87.51 ± 0.11	83.73 ± 0.18	74.56 ± 0.21	70.12 ± 0.28	92.46 ± 0.09	86.70 ± 0.11	83.50 ± 0.16	86.52 ± 0.22	78.52 ± 0.25	78.23 ± 0.32
Prysm-Net (Ours)	34.72	52.87	**95.14 ± 0.04**	**90.93 ± 0.06**	**89.47 ± 0.11**	**94.98 ± 0.05**	**90.31 ± 0.07**	**88.61 ± 0.10**	**84.61 ± 0.15**	**75.89 ± 0.18**	**71.29 ± 0.24**	**92.96 ± 0.08**	**87.39 ± 0.10**	**84.28 ± 0.13**	**87.26 ± 0.19**	**79.69 ± 0.21**	**79.17 ± 0.29**

All models were re-implemented and evaluated under a unified protocol. We report mDice, mIoU, and BF (%) as well as model complexity (Params, M; FLOPs, G). Bold indicates the best performance. "–" indicates the metric was not reported. Prysm-Net (Ours) follows the same protocol. In inference, only the ViT backbone, decoder, and BSM are active; the training-time FMDM/MGM provide guidance only and are disabled at test.

Table 2 | Cross-domain generalization on ETIS-LaribPolypDB

Method	Dice ↑	IoU ↑	BF ↑
U-Net	39.83 ± 0.92	33.59 ± 0.95	31.57 ± 1.10
U-Net++	40.12 ± 0.88	34.46 ± 0.90	32.84 ± 1.05
PraNet	62.83 ± 0.65	56.78 ± 0.68	55.12 ± 0.75
PVT	78.76 ± 0.42	70.69 ± 0.45	72.36 ± 0.55
HSNet	80.82 ± 0.38	73.44 ± 0.41	74.21 ± 0.50
ISCNet	80.42 ± 0.40	71.67 ± 0.43	73.59 ± 0.52
Polyper	86.55 ± 0.25	78.26 ± 0.28	77.47 ± 0.35
Mamba	82.52 ± 0.32	74.77 ± 0.35	75.47 ± 0.40
DDPM	78.32 ± 0.45	73.49 ± 0.48	71.18 ± 0.58
Ours	**88.12 ± 0.19**	**79.93 ± 0.21**	**79.77 ± 0.26**

All models are trained on Kvasir-SEG ∪ CVC-ClinicDB and evaluated on ETIS with *no* target-domain tuning. We report per-image mean Dice (mDice), mean IoU (mIoU), and Boundary *F*-measure (BF) in % (higher is better). Bold indicates the best overall. Our metrics are computed with a unified script (per-image Dice/IoU at 0.5 threshold, single-scale inference).

Collectively, these results highlight the model's robustness and its ability to generate precise segmentation maps across diverse imaging conditions compared to existing techniques.

Cross-domain generalization on ETIS-LaribPolypDB

We rigorously test the cross-domain generalization capabilities of our model on the challenging ETIS-LaribPolypDB dataset. Following a strict protocol, all models are trained on the combined Kvasir-SEG ∪ CVC-ClinicDB dataset and evaluated on ETIS.

The results are summarized in Table 2. Our proposed method, PrysmNet (listed as "Ours"), achieves SOTA performance with a mean Dice (mDice) of 88.12%, a mean IoU (mIoU) of 79.93%, and a Boundary *F*-measure (BF) of 79.77%. This surpasses the previous best published method, Polyper, which obtained 86.55% mDice, 78.26% mIoU, and 77.47% BF under the same conditions. This represents improvements of +1.57 mDice, +1.67 mIoU, and +2.30 BF over the next-best method, validating our model's ability to generalize to unseen data distributions and delineate sharp boundaries more effectively than existing approaches.

Qualitative analysis

Figure 1 provides a visual comparison of PrysmNet against several state-of-the-art methods on representative samples from both in-domain (Kvasir, CVC-ClinicDB) and cross-domain (ETIS) datasets. We specifically select challenging cases, including (i) diminutive polyps that are easily missed, (ii) lesions with weak or blurry boundaries that blend with the surrounding mucosa, and (iii) images containing specular highlights or other visual artifacts. As shown, earlier methods like PraNet may struggle with precise boundary delineation. While recent Transformer-based models like Polyper show improved structural consistency, our PrysmNet, empowered by the BSM and advanced training guidance, often produces sharper and more complete masks, particularly for the challenging cases of small and faint polyps.

To better understand how different components of PrysmNet contribute to its performance, we visualize the feature maps from various model configurations in Fig. 2. We use Grad-CAM to highlight the regions the network focuses on. As seen in Fig. 2a, the baseline model without BSM or guidance produces diffuse attention maps. Adding the BSM (Fig. 2b) sharpens the focus around the polyp boundary. The inclusion of FMDM and MGM guidance further refines this attention, leading to a more concentrated and accurate localization in the full PrysmNet model (Fig. 2c), which corresponds to more precise segmentation results (Fig. 2d–f).

Despite its strong performance, PrysmNet has limitations, as shown in Fig. 3. Quantitative analysis of failure modes on the test sets reveals that: (i) over-segmentation due to specular highlights occurs in ~2.3% of Kvasir-SEG and 1.8% of CVC-ClinicDB test samples, typically when highlights

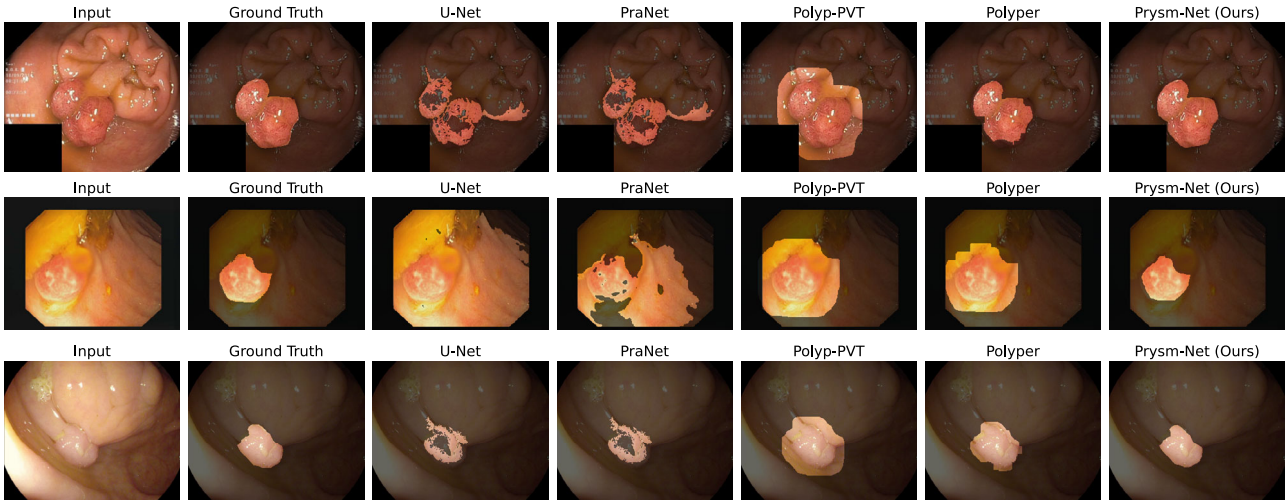

Fig. 1 | Qualitative comparisons on challenging cases from Kvasir-SEG (top row), CVC-ClinicDB (middle row), and the cross-domain ETIS dataset (bottom row). PrysmNet demonstrates superior performance in delineating fine boundaries and capturing diminutive lesions compared to other methods.

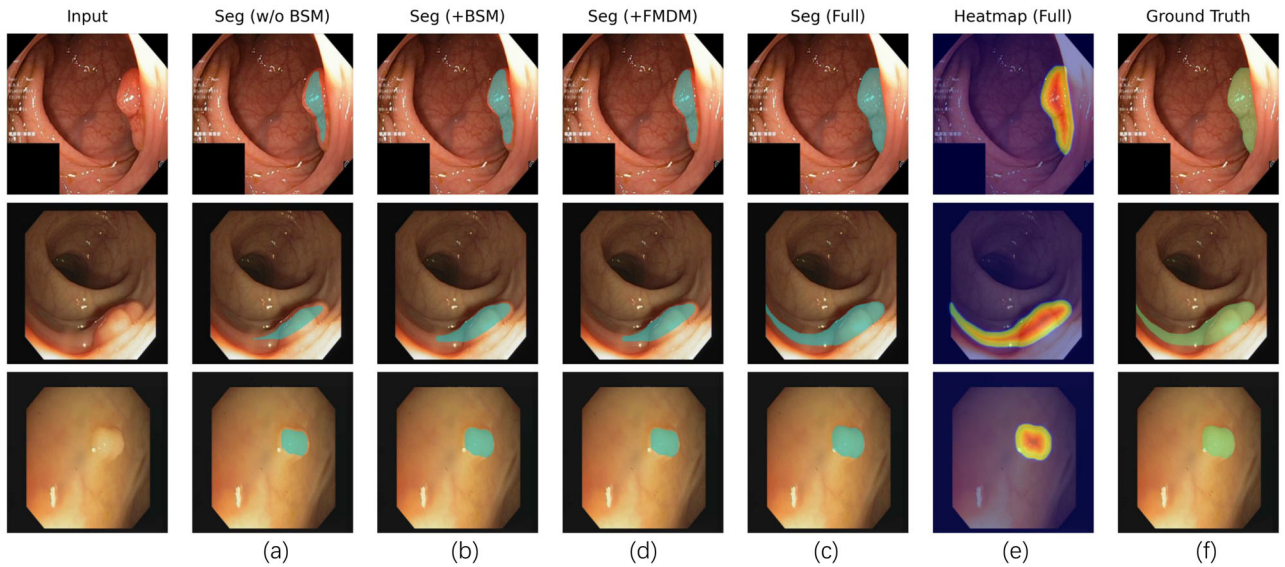

Fig. 2 | Visual ablation analysis with Grad-CAM. a–c show Grad-CAM attention heatmaps for three configurations: **a** baseline (ViT backbone + U-shaped decoder; without BSM/FMDM/MGM), **b** baseline + biologically-inspired salience module (BSM), and **c** the full PrysmNet with BSM plus foundation-model distillation (FMDM) and multi-modal guidance (MGM). **d–f** show the corresponding segmentation masks. Adding BSM sharpens boundary focus; adding FMDM and MGM—used only during training—further consolidates localization, yielding cleaner contours and more accurate masks, especially for diminutive or weak-boundary polyps; the full model (**c**, **f**) performs best.

cover more than 15% of the frame area; (ii) false negatives for extremely flat or isointense lesions affect ~1.5% of cases across all datasets, primarily when polyp to background contrast is below 0.05 (normalized intensity difference); (iii) boundary errors (under-segmentation or over-segmentation) occur in ~3.2% of cases, mostly for polyps smaller than 0.5% of image area. The model can sometimes fail in extreme cases. For instance, it may over-segment when encountering extensive specular highlights that mimic polyp texture. Conversely, it can miss extremely flat or isointense lesions that have virtually no contrast with the background mucosa. These failure cases, while relatively rare, highlight the remaining challenges in automated polyp segmentation and suggest avenues for future work, such as improved handling of visual artifacts and temporal information from video colonoscopies.

Ablation study

To validate the effectiveness of the proposed components in PrysmNet, we conduct a series of ablation studies. We start with a baseline model and incrementally add our proposed modules. All ablation experiments follow the same training protocol and are evaluated on Kvasir-SEG, CVC-ClinicDB (in-domain), and ETIS (cross-domain).

We first investigate the contribution of our main architectural and guidance modules: the biologically inspired salience module (BSM), the foundation-model distillation module (FMDM), and the multi-modal guidance module (MGM). The baseline model consists of the ViT backbone and a standard U-shaped decoder. As shown in Table 3, each component brings a noticeable improvement. The BSM significantly enhances performance, especially on the cross-domain ETIS dataset, underscoring its role in improving boundary detection and generalization. The FMDM and MGM modules further boost the scores by transferring valuable prior knowledge and enforcing feature invariance. The full PrysmNet model, combining all components, achieves the best performance across all datasets.

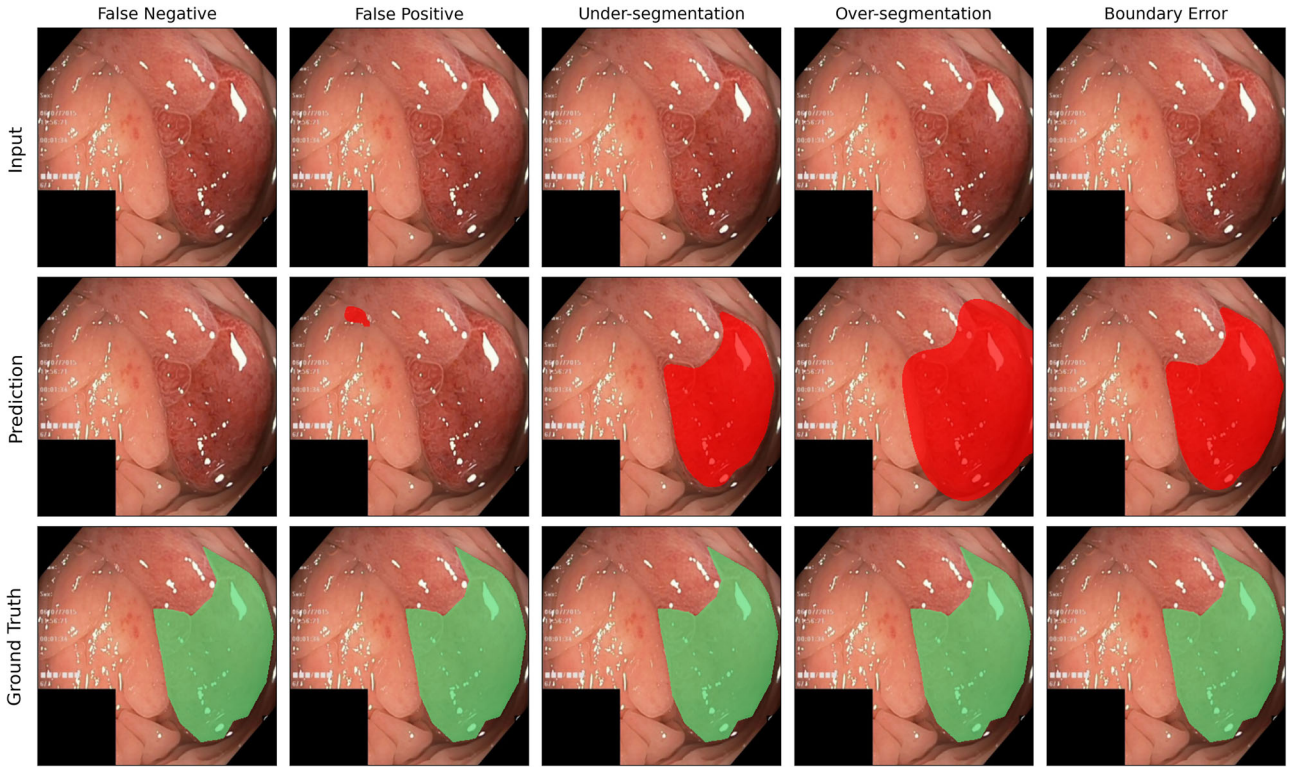

Fig. 3 | Failure modes on challenging colonoscopy frames. Columns (left to right): false negative, false positive, under-segmentation, over-segmentation, and boundary error. Rows: input image, model prediction, and ground truth mask. Over-segmentation is typically triggered by strong specular highlights, whereas the false-negative case corresponds to an extremely flat, low-contrast lesion that blends with the surrounding mucosa.

Table 3 | Ablation study: contributions of Prysm-Net's key components

Configuration	Modules			Kvasir-SEG		CVC-ClinicDB		ETIS	
	BSM	FMDM	MGM	mDice	mIoU	mDice	mIoU	mDice	mIoU
ViT Baseline				88.55	87.81	92.10	86.95	81.33	73.15
Baseline + BSM	✓			92.89	89.42	93.65	88.78	84.51	76.52
Baseline + BSM + FMDM	✓	✓		94.52	90.15	94.21	89.60	86.10	78.11
PrysmNet (Full)	✓	✓	✓	95.14	90.93	94.98	90.31	88.12	79.93

All variants are trained under a unified protocol on Kvasir-SEG ∪ CVC-ClinicDB; evaluation is conducted on Kvasir-SEG and CVC-ClinicDB (in-domain) and ETIS-LaribPolypDB (cross-domain). Metrics are mDice/mIoU (%; higher is better). "BSM" denotes the biologically inspired salience module; "FMDM" denotes multi-level foundation-model distillation; Prysm-Net (Full) additionally uses a training-only multi-modal guidance module (MGM), which is disabled at inference.

We also analyze the effect of the auxiliary loss functions used for training: the boundary supervision loss (L_{bnd}) for the BSM module and the structural similarity loss (L_{ssim}). The base loss consists only of the segmentation loss (L_{seg}). Table 4 shows the results. Adding the explicit boundary supervision (L_{bnd}) provides the most significant gain, forcing the BSM to learn semantically meaningful edge features and leading to sharper predictions. The L_{ssim} loss offers a further marginal improvement by encouraging perceptually coherent segmentation masks. The combination of all losses yields the optimal result.

Table 4 | Ablation study on loss components

Loss configuration	Kvasir-SEG		CVC-ClinicDB		ETIS	
	mDice	mIoU	mDice	mIoU	mDice	mIoU
L_{seg} only	91.13	84.65	91.88	86.01	81.34	71.20
$L_{seg} + L_{bnd}$	93.95	88.66	93.70	89.03	84.95	77.25
Full loss	95.14	90.93	94.98	90.31	88.12	79.93

All models use the full Prysm–Net architecture, with identical training protocol and data sampling; we report mDice (%)/mIoU (%) on Kvasir-SEG and CVC-ClinicDB (in-domain) and ETIS (cross-domain). Here, L_{seg} denotes the segmentation term (Soft Dice + two-class cross-entropy), L_{bnd} is the Dice-based boundary supervision computed on a one-pixel-wide ground-truth contour (for BSM), and L_{ssim} is a structural-similarity regularizer; Full Loss corresponds to $L_{seg} + L_{bnd} + L_{ssim}$ (other hyperparameters at defaults). Results indicate that explicit boundary supervision delivers the main gain, while adding L_{ssim} yields a further marginal improvement.

Summary

On in-domain data, recent hybrids (e.g., Polyper) reach ~94–96% Dice on both Kvasir-SEG and CVC-ClinicDB, far above early CNNs (U-Net ~ 82%, PraNet ~ 90%). On ETIS cross-domain tests, classic methods drop sharply (PraNet Dice 62.8%), while modern transformer/SSM/diffusion models achieve 78–87% Dice. Our approach targets this persistent generalization gap with boundary-aware decoding and training-time guidance, and our comprehensive ablations validate the effectiveness of each proposed component in achieving state-of-the-art performance.

Conclusion

We tackled the persistent challenges of cross-domain generalization, small lesion recall, and boundary precision in polyp segmentation by proposing

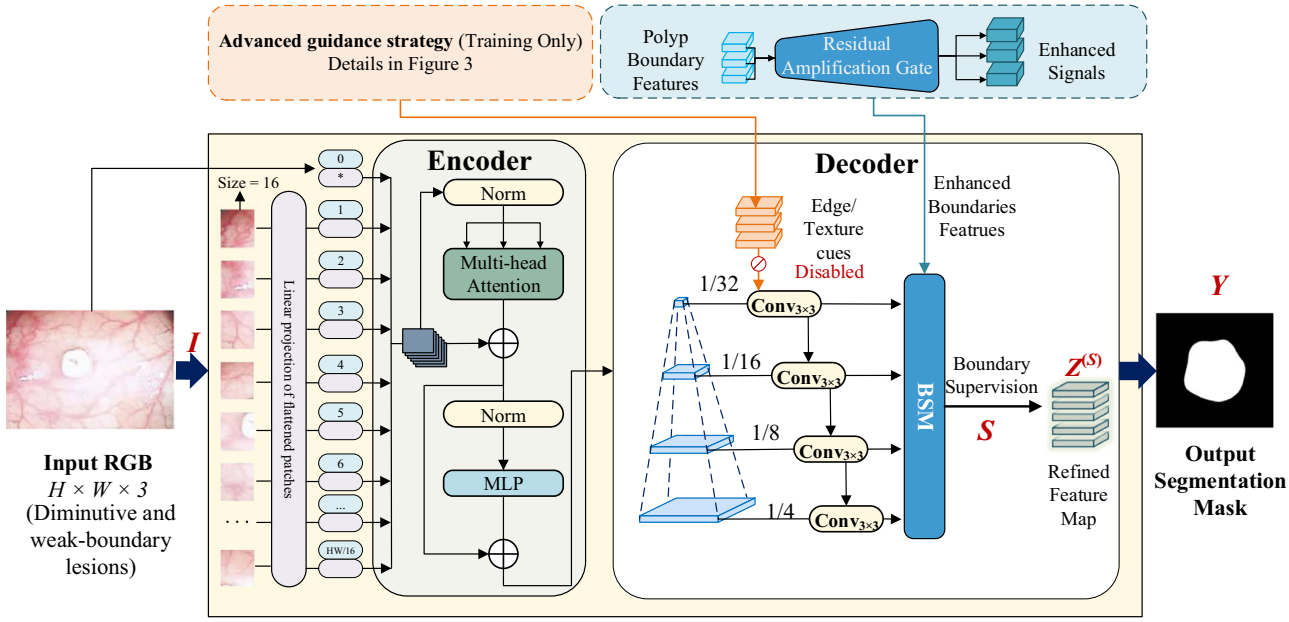

Fig. 4 | Overall architecture of PrysmNet. The model takes an RGB image, potentially containing diminutive and weak-boundary lesions, as input. *(Encoder)* The ViT backbone processes the image by dividing it into patches, which are linearly projected into tokens. These tokens pass through a series of Transformer blocks, each containing Multi-head Attention and an MLP, to capture global context. *(Decoder)* A U-shaped decoder reconstructs the segmentation mask from multi-scale features {1/32, 1/16, 1/8, 1/4}. Skip connections merge features from the encoder. The core innovation, the biologically inspired salience module (BSM), is integrated into the decoder to refine features, especially along polyp boundaries, producing an explicit boundary supervision signal *S*. During training, an advanced guidance strategy provides auxiliary edge/texture cues, which are disabled at inference. The final output is a refined segmentation mask *Y*.

PrysmNet. Our model utilizes a ViT backbone to capture global context and augments the decoder with a biologically inspired salience module to dynamically emphasize boundary-relevant features. To enhance robustness without increasing inference costs, we leveraged multi-level foundation model distillation and multi-modal guidance during training. These strategies effectively transfer structure, texture priors, and foundation-model knowledge, while all guidance branches are removed at inference to maintain a lightweight deployment path. Our comprehensive evaluation across in-domain and cross-domain protocols demonstrates the effectiveness of this approach.

Our experiments provide several key takeaways regarding the model's performance and design. Explicit boundary supervision on the final decoder scales yields sharper contours and consistently improves segmentation metrics for weak boundary and small lesion cases with minimal overhead. Furthermore, training time guidance from foundation models and auxiliary modalities promotes domain-invariant representations, which are crucial for multi-center data. We also found that maintaining train-test consistency via annealing the multi-modal guidance weight to zero avoids modality shift during testing. Additionally, data strategies such as small-polyp oversampling and Fourier amplitude mixing proved to be simple yet reusable tools for enhancing recall and robustness.

Despite its strong performance, the proposed method has certain limitations and risks. Quantitative analysis reveals that extremely tiny or nearly isointense lesions can still be missed, affecting ~1.5% of test cases. Future work will focus on bridging the gap towards clinical integration. Specifically, we aim to leverage the model's precise boundary delineation to assist endoscopists in defining optimal resection margins. We also plan to investigate the deployment of our lightweight model on endoscopic hardware to provide real-time, intraoperative decision support without disrupting clinical workflows.

Methods

Our objective is to develop a polyp segmentation model that not only excels in *in-domain* evaluations but also *generalizes* robustly to unseen clinical environments, with special emphasis on diminutive and weak boundary lesions. We propose PrysmNet (a Polyp-refining system with Salience and Multi-modal guidance), which consists of: (i) a *Vision Transformer (ViT)* backbone plus a boundary-centric refinement block, the biologically inspired salience module (BSM) and (ii) an *advanced training-time guidance* that transfers knowledge from a foundation model (SAM) and injects auxiliary edge/texture cues. The guidance modules are disabled at inference, preserving a lightweight test-time path (see Fig. 4).

Problem setup and notation

Let the training set be $\mathcal{T} = \{(I^{(n)}, Y^{(n)})\}_{n=1}^{N}$, where $I \in \mathbb{R}^{H\times W\times 3}$ is an RGB colonoscopy frame and $Y \in \{0,1\}^{H\times W}$ is a binary polyp mask. Masks are merged into a single foreground channel for training unless specified. We transform Y to two-class one-hot $Y^{1h} \in \{0,1\}^{H\times W\times 2}$, where the second channel denotes foreground. The segmentation head outputs two-channel logits $Z^{(S)} \in \mathbb{R}^{H\times W\times 2}$ with $\widehat{Y} = \mathrm{softmax}(Z^{(S)})$ and the foreground probability $\widehat{Y}_{fg} = \widehat{Y}[:,:,1]$. Symbols used throughout are summarized in Table 5.

Backbone and decoder overview

We employ a ViT-B/16 backbone (patch size 16) pre-trained on ImageNet-21k, chosen for its strong global context modeling. Image tokens are reshaped back to multi-scale feature maps via a lightweight pyramid projection to {1/32, 1/16, 1/8, 1/4} spatial scales.

A U-shaped decoder progressively upsamples features with bilinear interpolation; skip connections fuse {1/32, 1/16, 1/8, 1/4} scales via 1 × 1 and 3 × 3 convolutions. BSM blocks (Fig. 5) are attached to the last two decoder stages to explicitly sharpen boundaries where spatial detail is richest. The segmentation head is a 3 × 3 conv followed by a 1 × 1 conv, producing two-channel logits $Z^{(S)}$.

Architectural innovation: biologically inspired salience module (BSM)

Polyp boundaries can be faint, blurry, or partially occluded. The human primary visual cortex (V1) addresses similar challenges through highly tuned responses to local contrast, orientation, and frequency. We

operationalize this inspiration with a learnable multi-scale filter bank that emphasizes boundary-relevant signals and a residual amplification gate that selectively boosts such signals in the decoder (see Fig. 5, left panel).

Given decoder feature $F \in \mathbb{R}^{H\times W\times C}$, BSM computes a salience map

$$S = \sigma\left(\mathrm{Conv}_{1\times 1}\left(\left[\mathrm{Conv}_{3\times 3}^{(d=1)}(F), \ldots, \mathrm{Conv}_{3\times 3}^{(d=4)}(F)\right]\right)\right), \quad (1)$$

where $\mathrm{Conv}_{3\times 3}^{(d)}$ are dilated convolutions with dilation rates $d \in \{1, 2, 3, 4\}$ using "same" padding, the bracket denotes channel-wise concatenation, and σ is the Sigmoid. The choice of dilation rates {1, 2, 3, 4} was determined through preliminary experiments comparing {1, 2}, {1, 2, 3}, and {1, 2, 3, 4, 5}; the four-scale configuration ($d \in \{1, 2, 3, 4\}$) provided the best balance between boundary sensitivity and computational efficiency, capturing fine edges ($d = 1$) as well as broader context ($d = 4$) relevant for polyp boundaries. Filters can be optionally initialized to approximate Sobel/Gabor banks and then fine-tuned end-to-end, making S adaptive to colonoscopic textures and edges.

Salience gates the feature via

$$F_{\mathrm{ref}} = F + \alpha \cdot (F \odot S), \quad (2)$$

where α is a learnable scalar (initialized to 1.0). The residual pathway stabilizes training by preserving identity when salience is uncertain. Empirically, placing BSM on the last two decoder scales balances capacity with computational overhead.

Table 5 | Key symbols and operations

Symbol	Meaning
$F \in \mathbb{R}^{H\times W\times C}$	Decoder feature map at a given scale
$S \in [0, 1]^{H\times W\times 1}$	BSM salience/boundary probability map
$Z^{(S)} \in \mathbb{R}^{H\times W\times 2}$	Student two-class logits; $\hat{Y} = \mathrm{softmax}(Z^{(S)})$
Y_{SAM}	Pseudo-mask from SAM (binary); $B(\cdot)$ gives its thin boundary
$\psi(\cdot)$	1 × 1 projection for channel alignment
Dilate/Erode	Morphological operators with a 3 × 3 square structuring element
Thin$(\cdot)$	Morphological thinning to unit-pixel width
$\odot$	Hadamard product

Advanced training guidance strategy

To improve generalization without increasing test-time cost, we introduce two *orthogonal, training-only* components (Fig. 5, right): (i) a *foundation-model distillation module (FMDM)*, which *exclusively* transfers knowledge from a frozen foundation model (SAM) via multi-level distillation, and (ii) a *multi-modal guidance module (MGM)*, which leverages *self-derived* auxiliary modalities (HED edges and LBP textures) and injects them through a gated cross-attention branch. Conceptually, FMDM is teacher–student distillation from an external model, whereas MGM is modality-level feature fusion without any teacher. *Both modules are disabled at inference*, leaving an identical test time path.

FMDM primarily draws supervision from a frozen foundation model (SAM). When ground-truth (GT) labels are available during supervised training, we use them for (i) pseudo-mask selection and (ii) quality-aware reweighting in Eq. (7) (this is the default setting in all our main experiments). The GT-free variant, which relies solely on SAM stability scores and area priors, is used only in ablation studies to demonstrate robustness; in practice, we recommend using quality-aware weighting when GT is available, as

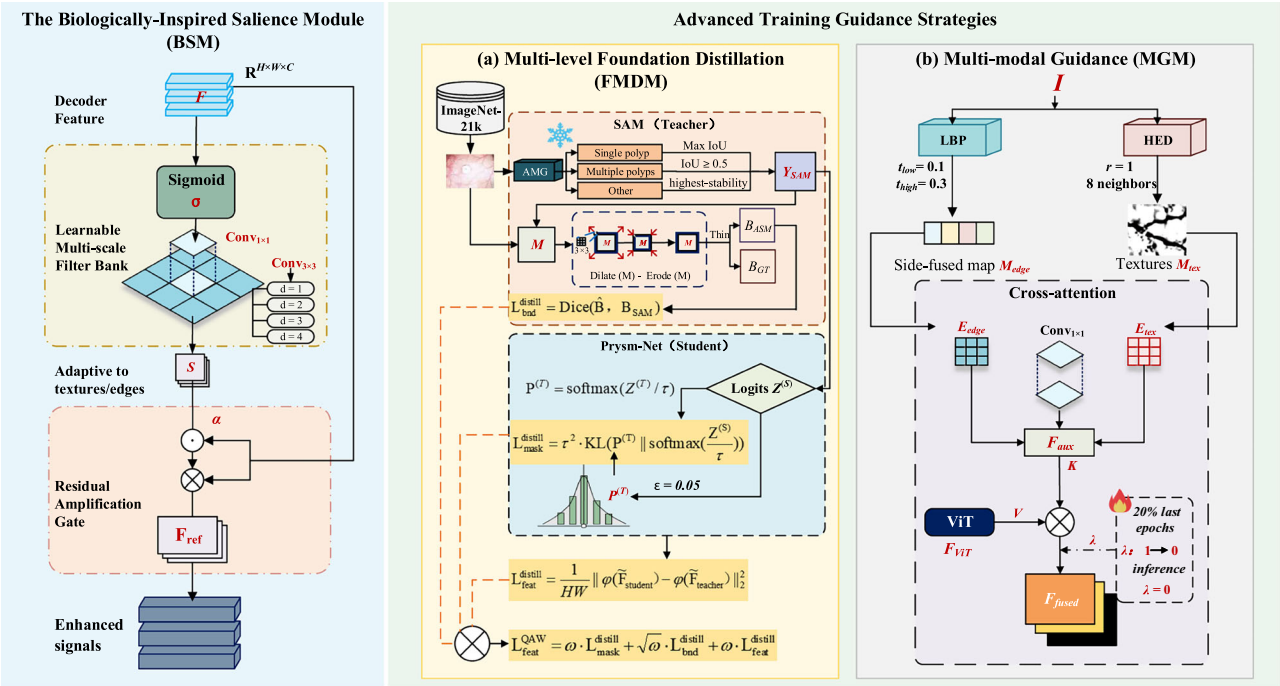

Fig. 5 | Detailed view of the BSM and advanced training guidance strategies. *(Left)* The Biologically Inspired Salience Module (BSM). This module takes a decoder feature map F as input. A learnable multi-scale filter bank, composed of parallel dilated convolutions ($d = 1, 2, 3, 4$), extracts edge and texture information at various receptive fields. The concatenated results are passed through a 1 × 1 convolution and a Sigmoid function to produce a salience map S. This map is then used to gate the original feature map via a residual amplification pathway ($F_{\mathrm{ref}} = F + \alpha \cdot (F \odot S)$), selectively enhancing signals critical to polyp boundaries. *(Right)* Advanced training guidance strategies (training-only). These modules are disabled at inference. **a** Multi-level foundation model distillation (FMDM): A frozen SAM teacher model generates pseudo-masks (Y_{SAM}) and boundaries (B_{SAM}) from the input image. This provides output-level guidance for the student network (PrysmNet) through a distillation loss. Feature-level distillation is also applied by aligning intermediate features between the teacher and student. **b** Multi-modal guidance (MGM): auxiliary modalities, such as texture maps from LBP and edge maps from HED, are generated from the input image. These are processed by lightweight encoders and fused with the main ViT features (F_{ViT}) using a cross-attention mechanism. A gating parameter λ anneals to zero during training to ensure train-test consistency.

it improves distillation quality by down-weighting unreliable SAM predictions (low IoU with GT).

We use the official SAM ViT-H checkpoint trained on SA-1B as the teacher and generate masks with the Automatic Mask Generator (AMG; `points_per_side` = 32, `pred_iou_thresh` = 0.88, `stability_score_thresh` = 0.95, `box_nms_thresh` = 0.7). For each training sample, we obtain a binary pseudo-mask Y_{SAM}: for single polyp frames, we select the mask with the highest IoU to Y (when GT is available); for multi-polyp frames, we take the union of masks with IoU ≥ 0.5; when GT is not available, we select the highest stability mask subject to an area prior (e.g., 50–10^5 pixels). We define a *unit-width* boundary operator

$$B(M) = \mathrm{Thin}\left(\mathrm{Dilate}(M; 3\times 3) - \mathrm{Erode}(M; 3\times 3)\right), \tag{3}$$

with a 3 × 3 square structuring element. Denote $B_{\mathrm{SAM}} = B(Y_{\mathrm{SAM}})$ and $B_{\mathrm{GT}} = B(Y)$.

We adopt temperature-scaled KL from teacher to student. Let the student's logits be $Z^{(\mathrm{S})}$ and the teacher distribution $P^{(\mathrm{T})}$. If teacher logits $Z^{(\mathrm{T})}$ are available, $P^{(\mathrm{T})} = \mathrm{softmax}(Z^{(\mathrm{T})}/\tau)$; otherwise we construct a smoothed two-class distribution from Y_{SAM} with label smoothing $\epsilon = 0.05$:

$$P^{(\mathrm{T})} = \left((1-\epsilon)(1-Y_{\mathrm{SAM}}) + \frac{\epsilon}{2}, (1-\epsilon)Y_{\mathrm{SAM}} + \frac{\epsilon}{2}\right),$$

The loss is

$$L_{\mathrm{mask}}^{\mathrm{distill}} = \tau^2 \cdot \mathrm{KL}\left(\mathrm{P}^{(\mathrm{T})} \,||\, \mathrm{softmax}\left(\frac{\mathrm{Z}^{(\mathrm{S})}}{\tau}\right)\right), \tag{4}$$

with $\tau = 2.0$.

To transmit SAM's contour prior, we guide BSM's boundary map $\widehat{B} := S$ towards B_{SAM} using Dice:

$$L_{\mathrm{bnd}}^{\mathrm{distill}} = \mathrm{Dice}\left(\widehat{B}, B_{\mathrm{SAM}}\right). \tag{5}$$

We match intermediate features at compatible spatial scales. SAM's image-encoder tokens are reshaped to a 2D grid (≈1/16 resolution), bilinearly resized to the student's decoder scale, and projected via ψ:

$$L_{\mathrm{feat}}^{\mathrm{distill}} = \frac{1}{HW}\left\|\psi\left(\widetilde{F}_{\mathrm{student}}\right) - \psi\left(\widetilde{F}_{\mathrm{teacher}}\right)\right\|_2^2. \tag{6}$$

We use the student's 1/8-scale decoder feature.

When GT is available, per-sample weight $w = \mathrm{clip}(\mathrm{IoU}(Y_{\mathrm{SAM}}, Y), 0.1, 1.0)$ re-weights distillation:

$$L_{\mathrm{distill}}^{\mathrm{QAW}} = w \cdot L_{\mathrm{mask}}^{\mathrm{distill}} + \sqrt{w} \cdot L_{\mathrm{bnd}}^{\mathrm{distill}} + w \cdot L_{\mathrm{feat}}^{\mathrm{distill}}. \tag{7}$$

Scope: MGM does *not* rely on any foundation model; it injects self-derived edge/texture cues via a cross-attention branch whose gating parameter λ anneals to zero.

From each input I, derive edges M_{edge} via HED (side fused map, NMS + hysteresis thresholds $t_{\mathrm{low}} = 0.1$, $t_{\mathrm{high}} = 0.3$) and textures M_{tex} via LBP (radius 1, 8 neighbors; per-tile histograms).

Two shallow encoders E_{edge}, E_{tex} (two 3 × 3 conv blocks, 32 channels, BN + ReLU) transform M_{edge}, M_{tex}; features are concatenated and projected:

$$F_{\mathrm{aux}} = \mathrm{Conv}_{1\times 1}\left(\left[E_{\mathrm{edge}}(M_{\mathrm{edge}}), E_{\mathrm{tex}}(M_{\mathrm{tex}})\right]\right).$$

We employ cross-attention (rather than simple concatenation or element-wise addition) to allow the model to selectively attend to relevant edge/texture cues based on the main feature context. This adaptive fusion mechanism ensures that auxiliary modalities contribute only when they provide complementary information, avoiding interference when edge/texture signals are noisy or redundant. Fuse with ViT features F_{ViT} via

$$F_{\mathrm{fused}} = \mathrm{LayerNorm}\left(F_{\mathrm{ViT}} + \lambda \cdot \mathrm{CrossAttn}(Q = F_{\mathrm{ViT}}, K = F_{\mathrm{aux}}, V = F_{\mathrm{aux}})\right). \tag{8}$$

To avoid distribution shift, λ decays linearly from 1 to 0 over the last 20% epochs; at inference $\lambda = 0$ and MGM is fully disabled.

Overall objective and training recipe

Soft Dice on $\widehat{Y}_{\mathrm{fg}}$ plus two-class CE:

$$L_{\mathrm{seg}} = \left(1 - \mathrm{Dice}(\widehat{Y}_{fg}, Y)\right) + \mathrm{CE}\left(Y^{1h}, \widehat{Y}\right). \tag{9}$$

with

$$\mathrm{Dice}(p, g) = \frac{2\sum_{ij} p_{ij} g_{ij} + \epsilon}{\sum_{ij} p_{ij}^2 + \sum_{ij} g_{ij}^2 + \epsilon}, \ \epsilon = 10^{-6}.$$

Supervise $\widehat{B} := S$ against B_{GT}:

$$L_{\mathrm{bnd}} = \mathrm{Dice}\left(\widehat{\mathrm{B}}, \mathrm{B}_{\mathrm{GT}}\right), \ \mathrm{B}_{\mathrm{GT}} = \mathrm{B(Y)} \text{ via Eq.(3)}. \tag{10}$$

Encourage perceptual fidelity:

$$L_{\mathrm{ssim}} = 1 - \mathrm{SSIM}\left(\widehat{Y}_{\mathrm{fg}}, Y\right) (\mathrm{window} 11 \times 11). \tag{11}$$

$$L_{\mathrm{total}} = L_{\mathrm{seg}} + \lambda_{\mathrm{bnd}} L_{\mathrm{bnd}} + \lambda_{\mathrm{distill}}\left(L_{\mathrm{distill}}^{\mathrm{mask}} + L_{\mathrm{distill}}^{\mathrm{bnd}} + L_{\mathrm{distill}}^{\mathrm{feat}}\right) + \lambda_{\mathrm{ssim}} L_{\mathrm{ssim}}. \tag{12}$$

Default hyperparameters: $\lambda_{\mathrm{bnd}} = 0.8$, $\lambda_{\mathrm{distill}} = 1.0$, $\lambda_{\mathrm{ssim}} = 0.5$, $\tau = 2.0$; $L_{\mathrm{feat}}^{\mathrm{distill}}$ at the 1/8 decoder scale. These values were determined through validation experiments on a held-out subset of the training data. Sensitivity analysis (varying each hyperparameter by ±20%) showed that performance is relatively robust: mDice changes by <0.5% for λ_{bnd} and $\lambda_{\mathrm{distill}}$, and <0.3% for λ_{ssim}, indicating that the chosen values are near optimal within a reasonable range.

Data sampling and augmentation

We employ a domain-balanced sampling strategy in which each mini-batch draws uniformly from the two source datasets; for odd batch sizes, the additional sample alternates between domains across iterations to preserve parity. To strengthen sensitivity to diminutive lesions, we oversample images whose polyp foreground occupies <1% of pixels at training resolution and enforce that exactly 25% of minibatches per epoch consist exclusively of such cases. The augmentation pipeline comprises Fourier amplitude mixing with coefficient $\alpha \sim \mathcal{U}(0.05, 0.15)$, random channel shuffling ($p = 0.2$), color jittering of brightness/contrast/saturation by ± 0.2, random scaling in [0.75, 1.25] followed by a random crop to the training size, and horizontal flipping.

Optimization, inference, and implementation details

Models are trained for 100 epochs with AdamW (learning rate 3×10^{-4}, weight decay 0.05, $\beta = (0.9, 0.999)$) under a cosine schedule with a 5 epoch warm-up, mixed precision arithmetic, gradient norm clipping at 1.0, an EMA decay of 0.999, and synchronized batch normalization across GPUs. We adopt a simple curriculum for the multi-granularity module (MGM): the weight λ is held at 1 for the first 80% of epochs and then annealed linearly to 0; the distillation coefficient $\lambda_{\mathrm{distill}}$ can be halved during the final 10 epochs.

For reproducibility, we fix random seeds, prefer deterministic kernels, and use bilinear resizing with align_corners = false, inputs are 352 × 352 at train and test. At inference, only the ViT backbone, the decoder with the boundary-aware saliency module (BSM), and a two-class head are active; MGM is disabled ($\lambda = 0$) and predictions are arg $\max_{c\in 0,1} \widehat{Y}[:,:,c]$. BSM augments each attached scale with four dilated 3 × 3 convolutions and one 1 × 1 convolution, yielding a per-scale complexity of $\mathcal{O}!\left(HW(4\cdot 9C^2 + C^2)\right) = \mathcal{O}(37HWC^2)$ at 1/8–1/4 resolutions. Compared to the base ViT + decoder complexity (approximately $\mathcal{O}(HW(C^2 + C))$ per decoder stage), BSM adds roughly 5–8% overhead in FLOPs, which is minimal given the significant boundary refinement benefits. MGM and FMDM are training-only, so test-time cost equals ViT + decoder + BSM.

Ablations are enabled by default: `BSM` (Eqs. (1) and (2) with L_{bnd}), `FMDM-mask/bnd/feat` (Eqs. (4)–(6)), `MGM` with the above annealing (Eq. (8), off at test), `SSIM` (Eq. (11)), and `SmallPolypOversample` (Section 7).

Two implementation notes improved stability: (i) the thinning operator Thin(·) in Eq. (3) prevents over-penalizing near-miss boundaries; and (ii) all losses consume *two-class* softmax outputs (two-class CE; Soft Dice on the foreground probability). When SAM logits are unavailable, we substitute the smoothed teacher distribution $P^{(T)}$ derived from Y_{SAM} in the KL context of Eq. (4) for robust training.

Training and inference

Algorithm 1 provides a high-level overview of the training and inference procedures for PrysmNet.

Algorithm 1. High-Level Training and Inference Logic for PrysmNet

procedure TrainingStep(*mini_batch*)
Sample images and masks from source datasets (domain-balanced, small-polyp emphasis).
▷ — *Training-time Guidance* —
// *Multi-modal guidance (MGM)*
Generate auxiliary modalities M_{edge} (HED) and M_{tex} (LBP).
Fuse auxiliary modalities with ViT features F_{ViT} via Eq. (8) using current λ.
// *Foundation model distillation (FMDM)*
Generate pseudo-mask Y_{SAM} and boundary B_{SAM} from frozen SAM teacher.
▷ — *Forward Pass and Loss Computation* —
Decode fused features with BSM to get student logits $Z^{(S)}$ and boundary map $\widehat{S}$.
Compute ground-truth boundary B_{GT} from the ground-truth mask Y.
Calculate total loss L_{total} using ground truth (Y, B_{GT}) and SAM guidance (Y_{SAM}, B_{SAM}) via Eq. (12).
Backpropagate L_{total} and update model weights.
end procedure
procedure Inference *image*
Set $\lambda = 0$ to disable all training-only modules (MGM, FMDM).
Perform forward pass through the inference-only path (ViT backbone, Decoder with BSM).
Obtain final logits $Z^{(S)}$.
$\widehat{Y} \leftarrow$ softmax $(Z^{(S)})$.
return arg max$(\widehat{Y})$ ▷Return the final binary segmentation mask
end procedure

Data availability

All datasets are publicly available for download: Segmented Polyp Dataset for Computer-Aided Gastrointestinal Disease Detection (Kvasir-SEG): https://datasets.simula.no/kvasir-seg/, CVC-ClinicDB: https://www.kaggle.com/datasets/balraj98/cvcclinicdb, ETIS-LaribPolypDB: https://www.kaggle.com/datasets/nguyenvoquocduong/etis-laribpolypdb, ColonDB: https://www.kaggle.com/datasets/longvil/cvc-colondb, and EndoScene: https://github.com/CAMMA-public/Endoscapes?tab=readme-ov-file. The deep learning algorithms and analysis pipelines for this study were implemented using the PyTorch framework. The codebase supports the complete training and inference procedures, including the integration of foundation model distillation and auxiliary guidance modules. To facilitate reproducibility, all custom scripts, model architectures, configuration files, and evaluation tools will be made publicly available following the paper's publication. Detailed documentation regarding the specific software environment, dependencies, and library versions is provided within the repository.

Received: 30 October 2025; Accepted: 4 January 2026;
Published online: 21 January 2026

References

1. Poon, C. C. et al. Ai-doscopist: a real-time deep-learning-based algorithm for localising polyps in colonoscopy videos with edge computing devices. *NPJ Digit. Med.* **3**, 73 (2020).
2. Ali, S. et al. A multi-centre polyp detection and segmentation dataset for generalisability assessment. *Sci. Data* **10**, 75 (2023).
3. Dumitru, S. et al. Duck-net: a lightweight CNN for colorectal polyp segmentation. Lightweight architecture; competitive accuracy. Preprint at arXiv:2311.02239 (2023).
4. Ali, S. et al. Assessing generalisability of deep learning-based polyp detection and segmentation methods through a computer vision challenge. *Sci. Rep.* **14**, 2032 (2024).
5. Xue, H., Yonggang, L., Min, L. & Lin, L. A lighter hybrid feature fusion framework for polyp segmentation. *Sci. Rep.* **14**, 23179 (2024).
6. Biffi, C. et al. A novel AI device for real-time optical characterization of colorectal polyps. *NPJ Digit. Med.* **5**, 84 (2022).
7. Fan, D. P. et al. PraNet: Parallel Reverse Attention Network for Polyp Segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020.* Vol. 12266 (eds Martel, A. L. et al.) (Lecture Notes in Computer Science, 2020).
8. Yu, T. & Wu, Q. Hardnet-cps: colorectal polyp segmentation based on harmonic densely united network. *Biomed. Signal Process. Control* **85**, 104953 (2023).
9. Yang, C., Guo, X., Zhu, M., Ibragimov, B. & Yuan, Y. Mutual-prototype adaptation for cross-domain polyp segmentation. *IEEE J. Biomed. Health Inform.* **25**, 3886–3897 (2021).
10. Bao, J., Zhou, Z., Li, W. J. & Luo, R. Structure-aware stylized image synthesis for robust medical image segmentation. arXiv preprint arXiv:2412.04296 (2024).
11. Dutta, T. K., Majhi, S., Nayak, D. R. & Jha, D. Ma mba g uided bo u ndary p rior matters: a new perspective for generalized polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 380–391 (Springer, 2025).
12. Wang, S. et al. Gh-unet: group-wise hybrid convolution-vit for robust medical image segmentation. *npj Digit. Med.* **8**, 426 (2025).
13. Bernal, J. et al. Comparative validation of polyp detection methods in video colonoscopy: results from the MICCAI 2015 endoscopic vision challenge. *IEEE Trans. Med. Imaging* **36**, 1231–1249 (2017).
14. Jha, D. et al. Kvasir-SEG: A Segmented Polyp Dataset. In *MultiMedia Modeling (MMM),* Vol. 11962 (eds Ro, Y. et al.) (Lecture Notes in Computer Science, 2020).
15. Vázquez, D. et al. A benchmark for endoluminal scene segmentation of colonoscopy images. *J. Healthc. Eng.* **2017**, 4037190 (2017).
16. Ronneberger, O., Fischer, P., Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015.* Vol. 9351 (eds Navab, N., Hornegger, J., Wells, W., Frangi, A.) (Lecture Notes in Computer Science, 2015).
17. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for*

Clinical Decision Support. Vol. 11045 (eds Stoyanov, D. et al.) (Lecture Notes in Computer Science, 2018).
18. Jha, D. et al. ResUNet++: An Advanced Architecture for Medical Image Segmentation. *International Symposium on Multimedia* 225–2255 (2019).
19. Jha, D., Riegler, M. A., Johansen, D., Halvorsen, P. & Johansen, H. D. DoubleU-Net: A Deep Convolutional Neural Network for Medical Image Segmentation. *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, 558−564 (Rochester, MN, USA, 2020).
20. Rana, D., Pratik, S., Balabantaray, B. K., Peesapati, R. & Pachori, R. B. Gcapseg-net: an efficient global context-aware network for colorectal polyp segmentation. *Biomed. Signal Process. Control* **100**, 106978 (2025).
21. Dong, B. et al. Polyp-pvt: polyp segmentation with pyramid vision transformers. arXiv preprint arXiv:2108.06932 (2021).
22. Xiao, B., Hu, J., Li, W., Pun, C.-M. & Bi, X. Ctnet: contrastive transformer network for polyp segmentation. *IEEE Trans. Cybern.* **54**, 5040–5053 (2024).
23. Ji, G.-P., Zhang, J., Campbell, D., Xiong, H. & Barnes, N. Rethinking polyp segmentation from an out-of-distribution perspective. *Mach. Intell. Res.* **21**, 631–639 (2024).
24. Zhang, Y., Liu, H. & Hu, Q. TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021.* Vol. 12901 (eds de Bruijne, M. et al.) (Lecture Notes in Computer Science, 2021).
25. Shao, H., Zhang, Y. & Hou., Q. Polyper: boundary sensitive polyp segmentation. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'24/IAAI'24/EAAI'24)*, Vol. 38, 4731–4739 (AAAI Press, 2024).
26. Cao, R. et al. Cfanet: context feature fusion and attention mechanism based network for small target segmentation in medical images. *Sensors* **23**, 8739 (2023).
27. Wu, H. et al. Polypseg+: a lightweight context-aware network for real-time polyp segmentation. *IEEE Trans. Cybern.* **53**, 2610–2621 (2022).
28. 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*, September 27–October 1, 2021, Proceedings, Part I (Strasbourg, France, 2021).
29. Kirillov, A. et al. Segment anything. In *Proc. of the IEEE/CVF International Conference on Computer Vision* 4015–4026 (2023).
30. Mazurowski, M. A. et al. Segment anything model for medical image analysis: an experimental study. *Med. Image Anal.* **89**, 102918 (2023).
31. Wang, M., Xu, C. & Fan, K. An efficient fine tuning strategy of segment anything model for polyp segmentation. *Sci. Rep.* **15**, 14088 (2025).
32. Zhao, Y. et al. Segment anything model-guided collaborative learning network for scribble-supervised polyp segmentation. arXiv preprint arXiv:2312.00312 (2023).
33. Zhang, M. et al. Adaptive risk minimization: Learning to adapt to domain shift. *Adv. Neural Inf. Process. Syst.* **34**, 23664–23678 (2021).
34. Cao, X., Fan, K. & Ma, H. Federal learning-based a dual-branch deep learning model for colon polyp segmentation. *Multimedia Tools Appl.* **84**, 10425–10446 (2025).
35. Galdran, A., Carneiro, G., Miguel A. & Ballester, G. Balanced-MixUp for Highly Imbalanced Medical Image Classification. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021: 24th International Conference*, Strasbourg, France, September 27 – October 1, 2021, Proceedings, Part V. 323–333 (Springer-Verlag, Berlin, Heidelberg, 2021).
36. Fang, Y., Chen, C., Yuan, Y. & Tong, Ky. Selective Feature Aggregation Network with Area-Boundary Constraints for Polyp Segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Vol 11764 (eds Shen, D. et al) (Lecture Notes in Computer Science, 2019).
37. Khan, A. M., Ashrafee, A., Khan, F. S., Hasan M. B., & Kabir, M. H. AttResDU-Net: Medical Image Segmentation Using Attention-based Residual Double U-Net. *2023 International Joint Conference on Neural Networks (IJCNN)*, 1−8 (Gold Coast, Australia, 2023).
38. 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*. Vol 12901 (eds de Bruijne, M. et al.) (Lecture Notes in Computer Science, 2021).
39. Chen, J. et al. Transunet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021).
40. Huang, C.-H., Wu, H.-Y. & Lin, Y.-L. Hardnet-mseg: a simple encoder–decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps. arXiv preprint arXiv:2101.07172 (2021).
41. Lin, A. et al. Ds-transunet: Dual swin transformer U-Net for medical image segmentation. *IEEE Trans. Instrum. Meas.* **71**, 1–15 (2022).
42. Park, K.-B. & Lee, J. Y. Swine-net: hybrid deep learning approach to novel polyp segmentation using convolutional neural network and swin transformer. *J. Comput. Design Eng.* **9**, 616–632 (2022).
43. Lou A, Guan S, Loew M. CaraNet: context axial reverse attention network for segmentation of small medical objects. *J Med Imaging (Bellingham)* **10**, 014005 (2023).
44. Duc, N. T., Oanh, N. T., Thuy, N. T., Triet, T. M. & Dinh, V. S. Colonformer: an efficient transformer based method for colon polyp segmentation. *IEEE Access* **10**, 80575–80586 (2022).
45. Chen, F., Ma, H. & Zhang, W. Segt: a novel separated edge-guidance transformer network for polyp segmentation. arXiv preprint arXiv:2306.10773 (2023).

Acknowledgements

This study was supported by the National Research Center of Geriatric Diseases, 2022 (Xiangya Hospital) (XYYYJSTG-11 to Xiaowei Liu), National Research Center of Geriatric Diseases, 2024 (Xiangya Hospital to Xiaowei Liu), and Hunan Provincial Natural Science Foundation of China (2023JJ30947 to Yu Wu).

Author contributions

Junbo Xiao: Study concept and design, drafting of the manuscript, and data analysis. Han Yi: Acquisition of data, study design, and investigation. Lei Wang: Analysis and interpretation of data and conceptualization. Ying Li: Acquisition of data, methodology, and investigation. Xiaotong Wang: Figure preparation, methodology, and Investigation. Shizhe Li: Figure preparation, methodology, and investigation. Jun Yi: Figure preparation, methodology, and investigation. Yu Wu: Critical revision of the manuscript, supervision, funding acquisition, and conceptualization. Xiaowei Liu: Critical revision of the manuscript, project administration, supervision, funding acquisition, and conceptualization. All authors contributed to the manuscript and to the interpretation of the results. All authors approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Yu Wu or Xiaowei Liu.