FSD-Net: Foundation-Guided Spatiotemporal Distillation for Video Polyp Segmentation

Yahang Leng[1], Shiquan Min[1*], and Chengzhou Li[2†]

[1] College of Computer Science, Sichuan University, Chengdu, China
[2] College of Software, Dalian University of Technology, Dalian, China
lichengzhou@mail.dlut.edu.cn

Abstract. Video polyp segmentation plays a pivotal role in early colorectal cancer diagnosis, yet its clinical deployment is heavily bottlenecked by the prohibitive cost of dense pixel-level annotations. To address this challenge, we propose a foundation-guided spatiotemporal distillation framework named FSD-Net for sparsely annotated video polyp segmentation. Our approach introduces a Semantic Flow Distillation (SFD) module that leverages a frozen foundation model as a teacher to extract robust semantic features and generate continuous supervision signals for a lightweight student network. To tackle the frequent target disappearance caused by complex intestinal folds, we design an Occlusion-Aware Cycle Consistency (OACC) mechanism that dynamically blocks erroneous gradient propagation based on visibility estimation. Furthermore, we introduce a Prototype Re-identification Module (PRM) to maintain long-term memory and successfully recapture target identities upon reappearance. Extensive experiments on three standardized datasets, including SUN-SEG, CVC-612, and CVC-300, demonstrate that our framework significantly outperforms existing weakly supervised methods and even surpasses state-of-the-art fully supervised models. Notably, on the highly challenging SUN-SEG-Hard dataset, our method achieves a mDice 87.04%, mIoU of 79.20%, and mHD of 20.64 *mm*, proving its exceptional robustness and label efficiency in complex endoscopic environments.

Keywords: Video Polyp Segmentation · Spatiotemporal Distillation · Weakly Supervised Learning · Foundation Models · Cycle Consistency

1 Introduction

Colorectal cancer remains a leading cause of cancer-related mortality worldwide, yet early detection and removal of polyps via colonoscopy can significantly improve overall survival rates [28]. While automatic polyp segmentation has achieved remarkable progress, recent studies indicate that the majority of existing approaches still treat video data as a collection of independent static images [8]. This frame-by-frame paradigm ignores crucial temporal correlations

† Corresponding author.
∗ Equal contribution.

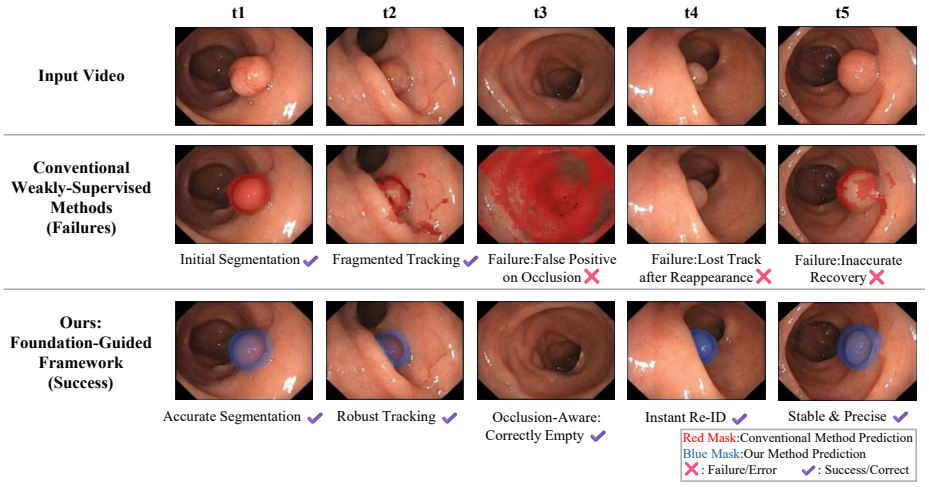


Fig. 1: Challenges in Weakly-Supervised Video Polyp Segmentation: Handling Dynamic Occlusion and Reappearance. Top Row (Input Video): A challenging endoscopic sequence where a polyp is temporarily occluded by a mucosal fold (t_3) and reappears (t_4). **Middle Row (Conventional Methods):** Existing approaches relying on low-level optical flow fail to perceive the occlusion, leading to hallucinated false-positive masks on the fold at t_3 and losing the target ID upon reappearance at t_4. **Bottom Row (Ours):** Our Foundation-Guided framework utilizes semantic visibility estimation to correctly suppress segmentation during occlusion (yielding a clean empty mask at t_3) and successfully re-identifies the target at t_4, ensuring robust long-term temporal consistency.

and often leads to inconsistent predictions, particularly when polyps undergo rapid deformation or severe motion blur. Consequently, Video Polyp Segmentation (VPS) has emerged as a superior alternative by leveraging spatiotemporal information to suppress noise and ensure coherent detection results across consecutive frames [19].

However, the widespread adoption of VPS models is severely hindered by their heavy reliance on data annotation. Training high-performance networks typically necessitates dense pixel-level ground truth masks for every single video frame [14]. Acquiring such comprehensive labeling is extremely labor-intensive and requires expert clinical knowledge, which is often impractical for large-scale datasets as noted by Ji et al. [15]. This annotation bottleneck creates a significant barrier to the deployment of deep learning models in clinical settings, underscoring the urgent need for label-efficient learning strategies where only a small fraction of keyframes are annotated [7, 26].

Conventional weakly-supervised methods predominantly rely on optical flow estimation or local feature matching to propagate supervisory signals from annotated keyframes to unlabeled frames [1, 26]. Although these strategies have proven effective in natural scene video segmentation, their reliance on low-level motion cues makes them inherently fragile in the complex endoluminal environ-

ment. As demonstrated by Yang et al. [34], the presence of specular reflections, fluid interference, and the non-rigid peristalsis of the colon wall frequently introduces severe noise, causing optical flow algorithms to fail. This instability results in the accumulation of errors during the label propagation process, leading to boundary drift and incomplete segmentation masks [15]. A more critical yet often overlooked challenge is the phenomenon of dynamic occlusion, where polyps disappear behind mucosal folds and reappear after a few seconds. As explicitly illustrated in Fig. 1 (Middle Row), recent studies have highlighted a fundamental limitation in existing cycle-consistency frameworks [31]: they inherently lack the advanced semantic awareness required to reliably distinguish between a temporarily disappeared object and a difficult, visually deceptive background. Consequently, during dynamic occlusion events (t_3), these models remain entirely oblivious to the target's absence. Instead of suspending supervision, they blindly enforce erroneous temporal consistency constraints across frames, actively forcing the network to hallucinate false-positive polyp masks directly onto surrounding healthy mucosal tissue [2]. Furthermore, this accumulation of misguided gradients severely corrupts the model's feature memory. Thus, when the polyp eventually reappears (t_4), the network frequently fails to re-identify the target due to the complete loss of tracking continuity.

To surmount these limitations, we propose a Foundation-Guided Spatiotemporal Distillation framework that shifts the supervision paradigm from unreliable pixel flow to robust semantic flow. We leverage the Segment Anything Model (SAM) [16] as a frozen teacher to guide a lightweight student network. Although several adaptations like MedSAM [20] and SAM-Med2D have been proposed, they primarily focus on static image segmentation. While Yin et al. recently demonstrated the utility of memory-augmented SAM 2 for surgical video segmentation, their approach focuses on training-free inference rather than distilling knowledge for efficient clinical deployment [35]. In contrast, our method utilizes the teacher's visibility estimation to construct a dynamic occlusion-aware mechanism. This ensures that the student network learns to suppress predictions when the target is occluded and instantly re-identify the polyp upon reappearance. By distilling the heavy computation of the foundation model into an efficient student network, we achieve consistently high-precision segmentation with true real-time clinical feasibility.

The main contributions of this work are summarized as follows:

- We propose a novel weakly-supervised VPS framework that utilizes the semantic tracking capability of SAM to replace fragile optical flow, enabling robust label propagation under sparse supervision.
- We design an Occlusion-Aware Cycle Consistency loss that dynamically weights supervision based on target visibility, effectively eliminating hallucinations caused by temporary occlusions.
- We introduce a Prototype-Based Re-Identification mechanism that maintains a long-term memory of polyp features to ensure consistent segmentation during disappearance and reappearance events.

– Extensive experiments on the SUN-SEG, CVC-612, and CVC-300 datasets demonstrate that our method significantly outperforms state-of-the-art competitors while achieving real-time inference speeds suitable for clinical use.

2 Related Work

2.1 Medical Image Segmentation

To alleviate the high cost of dense pixel-level annotations, weakly supervised paradigms utilizing sparse annotations (*e.g.*, bounding boxes, points, scribbles) have been widely explored [6, 23, 24]. Recently, vision foundation models [16] have been adapted for medical imaging [20, 22]. Methods like MedSAM and SAMed employ low-rank adaptation to fine-tune on diverse medical datasets [33, 37], while other frameworks integrate medical priors to enhance prompt-driven accuracy [4, 18]. These approaches significantly reduce annotation dependency while maintaining high precision [11].

2.2 Polyp Segmentation

Automatic polyp segmentation is crucial for early colorectal cancer diagnosis [13]. Early methods utilized CNNs for localized feature extraction [8,14,39], while subsequent architectures introduced vision transformers to capture global contextual dependencies [5, 29]. Recently, prompt-driven foundation models have been applied to colonoscopy sequences [12, 30], demonstrating robust generalization under complex clinical scenarios (*e.g.*, severe reflections, non-rigid deformations) through extensive pre-training and efficient adaptation strategies [3, 21, 27].

2.3 Spatiotemporal Knowledge Distillation

Knowledge distillation (KD) compresses cumbersome networks into lightweight models [10]. In medical imaging, standard spatial KD typically transfers pixel-level features across independent static frames, ignoring crucial temporal correlations [25]. While Video Object Segmentation (VOS) frameworks model spatiotemporal dependencies across frames [7, 15], low-level motion priors such as optical flow easily fail under the severe deformations and motion blur inherent to colonoscopy [34]. To bridge this gap, we propose a spatiotemporal distillation paradigm leveraging the high-level semantic tracking capabilities of foundation models [9, 16, 35]. By distilling spatial representations and enforcing dynamic temporal consistency [2, 31], our method guides the student network to robustly model the endoluminal environment without relying on fragile motion vectors.

3 Methodology

3.1 Architecture Overview

The core objective of the proposed foundation-guided spatiotemporal distillation framework is to optimize the parameters of a lightweight video polyp segmentation network using sparse annotation data. We define the input colonoscopy

video sequence as a four-dimensional tensor $V = \{I_t\}_{t=1}^{T} \in \mathbb{R}^{T\times C\times H\times W}$. In this tensor representation T denotes the total number of frames in the sequence, C denotes the number of color channels, while H and W denote the spatial height and width of the image respectively. As illustrated in the overall architecture of Fig. 2, the system primarily consists of two core components including a frozen SAM-ViT-B teacher model Φ_T and a lightweight student network $\Phi_S(\cdot;\theta_S)$ containing learnable network weights θ_S. The student adopts a PVTv2-B0 encoder with a lightweight FPN decoder (12.7M parameters, 18.4 GFLOPs, 1.62GB inference memory, 46.7 FPS on a single RTX 4090). Throughout the processing pipeline the system initially utilizes the ground truth annotation of the first frame as guidance and generates continuous pseudo labels for each subsequent frame via the forward propagation of the teacher model.

$$\hat{M}_t = \Phi_T(I_{t-1}, I_t, \hat{M}_{t-1}) \tag{1}$$

During this forward tracking process $\hat{M}_t$ represents the pseudo label generated at the current time t, whereas I_{t-1} and I_t represent two adjacent continuous video frames. Simultaneously the student network independently receives the image of the current frame and outputs the predicted segmentation mask.

$$P_t = \Phi_S(I_t;\theta_S) \tag{2}$$

To establish a closed loop constraint in the temporal dimension, the system subsequently employs the predicted mask P_t of the student network at the current time t as a new prompt signal to perform backward propagation to the historical moment through the teacher model.

$$\tilde{M}_{t-1} = \Phi_T(I_t, I_{t-1}, P_t) \tag{3}$$

The generated $\tilde{M}_{t-1}$ from this backward process represents the mask of the previous time deduced from the current moment. Ultimately the framework jointly updates the parameters of the student network by calculating the spatial difference between the forward pseudo labels and the predicted masks alongside the cycle consistency error between the backward propagation results and the historical predictions.

Fig. 3 details the internal computational graphs of the semantic flow distillation, occlusion-aware cycle consistency and prototype re-identification modules. We will elaborate on the specific calculation processes of these components in the subsequent sections.

3.2 Semantic Flow Distillation Module

Most existing weakly supervised video segmentation methods rely on low-level pixel optical flow. We utilize the powerful semantic tracking capability of the teacher model to generate high-quality supervision signals in this module. As specifically illustrated in the left component of Fig. 3, this module takes the raw video frames and initial masks as inputs to construct a forward knowledge

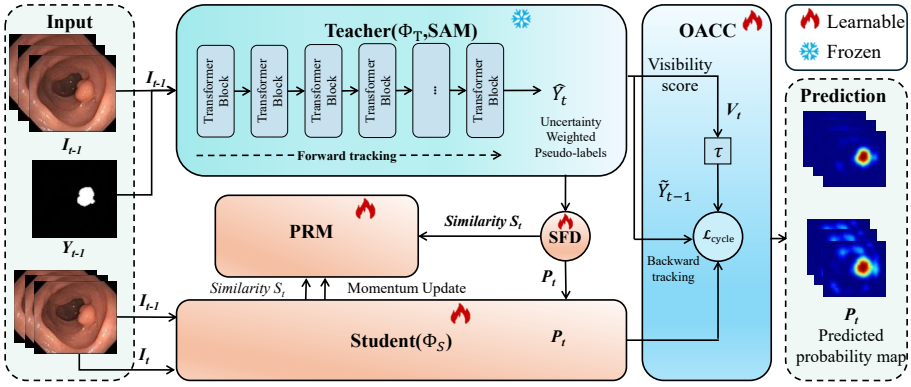


Fig. 2: The overall architecture of the proposed foundation guided spatiotemporal distillation framework.

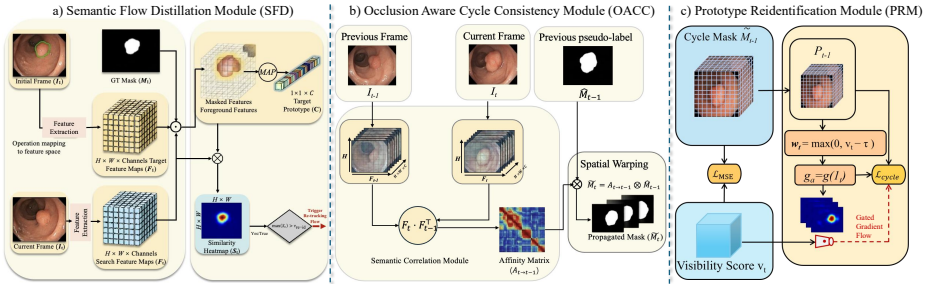


Fig. 3: Detailed internal computational graphs of the semantic flow distillation, occlusion-aware cycle consistency and prototype re-identification modules.

transfer pathway. Under the sparse annotation setting, only the first video frame contains the binarized ground truth segmentation mask $M_1 \in \{0,1\}^{H\times W}$ provided by experts. The numerical zero in this binary mask represents the background region, and the numerical one represents the polyp target region. We treat this as the absolute initial state of tracking and transfer its semantic information to the current frame through the recursive mapping function of the teacher model, as previously defined in Eq. (1). In the complex environment of actual endoscopy, the pseudo-labels generated by the teacher model inevitably contain noise. To filter out these unreliable supervision signals, we introduce a pixel-level uncertainty weighting mechanism. Given SAM-ViT-B mask logits z_t and its predicted-IoU score s_t^{iou}, we compute $p_t = \sigma(z_t)$ and derive a deterministic confidence matrix:

$$Q_t = s_t^{iou}|2p_t - 1|, \tag{4}$$

which reflects the degree of certainty that each pixel belongs to the target class. We first calculate the standard binary cross-entropy loss at each spatial coordinate point:

$$\mathcal{L}_{BCE}(x,y) = -\left[\hat{M}_t(x,y)\log P_t(x,y) + (1-\hat{M}_t(x,y))\log(1-P_t(x,y))\right], \tag{5}$$

where x and y represent the two-dimensional spatial coordinate indices in the image matrix. Subsequently, we embed the confidence matrix as a learnable weight term to calculate the sum of the weighted loss of a single frame image across all spatial locations:

$$\mathcal{L}_{frame}^{t} = \sum_{x=1}^{H}\sum_{y=1}^{W} Q_t(x,y)\mathcal{L}_{BCE}(x,y). \tag{6}$$

We finally perform average aggregation on this single-frame loss over the temporal dimension of the entire video sequence to obtain the complete uncertainty-aware distillation loss:

$$\mathcal{L}_{distill} = \frac{1}{T}\sum_{t=1}^{T}\mathcal{L}_{frame}^{t}. \tag{7}$$

This distillation mechanism incorporating confidence weights can effectively prevent the student network from fitting the erroneous boundaries generated by the teacher model. Consequently, it compels the lightweight student network to implicitly and robustly learn the cross-frame semantic constancy understanding capability of the teacher model under single-frame input conditions.

3.3 Occlusion-Aware Cycle Consistency Module

Polyps are frequently occluded by complex intestinal folds. Lacking visibility perception, traditional cycle consistency blindly enforces temporal alignment, often hallucinating false positives during target disappearance. To address this, we design an occlusion-aware cycle consistency loss. As shown in Fig. 3 (middle), we construct a reverse validation loop by tracking the student's current prediction backward via the teacher. Ideally, this backward-deduced mask should perfectly align with the student's prior prediction P_{t-1}. We formulate this basic temporal stability constraint using the spatial mean squared error:

$$\mathcal{L}_{MSE} = \frac{1}{H \times W}\sum_{x=1}^{H}\sum_{y=1}^{W} ||\tilde{M}_{t-1}(x,y) - P_{t-1}(x,y)||_2^2. \tag{8}$$

Introducing a dynamic weight coefficient is the core innovation to handle the occlusion phenomenon. We derive a normalized global target visibility score $v_t \in [0,1]$ from the same SAM-ViT-B outputs using the foreground region $\Omega_t = \{p_t > 0.5\}$:

$$v_t = \text{mean}_{\Omega_t}(Q_t)\min\left(1, \frac{|\Omega_t|}{|\Omega_{t-1}| + \epsilon}\right), \tag{9}$$

where a score closer to one indicates that the target is more clearly visible, and sudden foreground shrinkage suppresses v_t under occlusion or disappearance. To dynamically block the propagation of erroneous gradients when the target is completely occluded, we introduce a scalar visibility tolerance threshold τ and calculate the adaptive weight coefficient using a maximum function:

$$w_t = \max(0, v_t - \tau). \tag{10}$$

We multiply this adaptive weight coefficient w_t by the basic pixel-level mean squared error term and perform cumulative averaging over the temporal dimension to obtain the final occlusion-aware cycle consistency loss:

$$\mathcal{L}_{cycle} = \frac{1}{T-1}\sum_{t=2}^{T} w_t \left(\frac{1}{H \times W} \sum_{x=1}^{H}\sum_{y=1}^{W} ||\tilde{M}_{t-1}(x,y) - P_{t-1}(x,y)||_2^2 \right). \quad (11)$$

3.4 Prototype Re-identification Module

Conventional tracking frameworks suffer from catastrophic feature drift during prolonged occlusions, irreversibly losing target identities to background noise. To maintain long-term memory, our Prototype Re-identification (PRM) module in Fig. 3 (right) extracts dense visual features to build a robust, dynamically updated global memory bank.

To avoid computationally prohibitive high-resolution storage, a lightweight extractor maps the input to a down-sampled semantic feature $F_t = \mathcal{E}_S(I_t) \in \mathbb{R}^{D \times H' \times W'}$. This bottleneck accelerates similarity computation and enforces translation-invariant morphological learning. We aggregate the initial prototype C_1 using the first frame's ground truth M_1:

$$C_1 = \frac{\sum_{x,y} F_1(x,y) \cdot M_1(x,y)}{\sum_{x,y} M_1(x,y)} \quad (12)$$

To accommodate non-rigid deformations of the polyp occurring in the video sequence, we dynamically update the prototype in high-confidence frames using a momentum coefficient $\alpha = 0.90$:

$$C_t = \alpha C_{t-1} + (1-\alpha)\frac{\sum_{x,y} F_t(x,y) \cdot \hat{M}_t(x,y)}{\sum_{x,y} \hat{M}_t(x,y)} \quad (13)$$

For subsequent frames, we compute the cosine similarity matrix $S_t(x,y)$ between the current pixel features and the latest prototype C_{t-1}:

$$S_t(x,y) = \frac{C_{t-1}^{\top} F_t(x,y)}{||C_{t-1}||_2 ||F_t(x,y)||_2} \quad (14)$$

We then extract the global maximum response $s_t^{max} = \max_{x,y} S_t(x,y)$. If s_t^{max} exceeds a threshold of 0.72, signaling the polyp's reappearance, the model instantly generates a new prompt to reactivate the suspended spatiotemporal tracking flow.

3.5 Overall Loss Function

During the end-to-end joint training phase, we integrate the above multiple supervision objectives defined above into a unified optimization framework. To further suppress the prediction flickering phenomenon generated by the student

network between consecutive continuous frames, we additionally introduce a temporal smoothness regularization term. We calculate the absolute difference between the predicted masks of two adjacent frames across all spatial positions as the smoothing constraint:

$$\mathcal{L}_{smooth} = \frac{1}{T-1} \sum_{t=2}^{T} \sum_{x=1}^{H} \sum_{y=1}^{W} |P_t(x,y) - P_{t-1}(x,y)|. \tag{15}$$

The final overall loss function is linearly composed of the semantic flow distillation loss term, the occlusion-aware cycle consistency loss term, and the temporal smoothness regularization loss term:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{distill} + \lambda_2 \mathcal{L}_{cycle} + \lambda_3 \mathcal{L}_{smooth}, \tag{16}$$

where λ_1, λ_2, and λ_3 represent scalar hyperparameters that balance the respective gradient contribution proportions of different loss terms. By optimizing this overall loss function, the lightweight student network not only acquires precise spatial boundary delineation capabilities but also possesses robust temporal consistency to cope with complex occlusion and deformation environments without dense annotations.

4 Experiments

4.1 Experimental Setup

Datasets To comprehensively evaluate our framework, we conducted experiments on three public colonoscopy video polyp segmentation datasets. The first is SUN-SEG [15], the largest dataset (>100,000 frames), which provides an official benchmark split of 113 training and 45 test video sequences. We train all models on the official SUN-SEG training set. All SUN-SEG quantitative results in Table 1 and the ablation studies (Tables 2–7) are reported on the official test set, evaluated separately on its Easy and Hard subsets. The second, CVC-612 (CVC-ClinicVideo), contains 29 short video sequences with 612 fine pixel-level annotated images and is included in Table 1 as an additional video benchmark. The third, CVC-300, includes 300 clear endoscopic images to assess static-image generalization on unseen data. CVC-300 is used only as a generalization test set and is never used for training or hyperparameter tuning. At evaluation, each CVC-300 image is treated as a one-frame clip ($T = 1$): the SAM teacher is discarded and only the trained student performs a single-frame forward pass, without temporal tracking, cycle-consistency supervision, or prototype update. During preprocessing, to unify input dimensions and balance computational efficiency, all images and corresponding masks were uniformly resized to a fixed resolution of 352×352.

Evaluation Metrics Following standard protocols, we employ three comprehensive metrics. **mean Dice (mDice)** and **mean Intersection over Union (mIoU)** serve as the primary indicators of region-level segmentation accuracy, with mIoU penalizing false positives more severely. Additionally, **mean Hausdorff Distance (mHD)** is utilized as a stringent boundary-based metric to evaluate how accurately the predicted boundaries fit the complex contours of the polyps.

Implementation Details All experiments were strictly implemented using the PyTorch framework on a workstation equipped with a single NVIDIA RTX 4090 (24GB) GPU. The lightweight student network was trained end-to-end for 50 epochs using the AdamW optimizer. We set the initial learning rate to 1×10^{-4} and employed a cosine annealing decay schedule to facilitate smooth convergence. To prevent overfitting given the sparse annotations, we applied robust data augmentation strategies, including random affine transformations, color jittering to simulate different lighting conditions, and random horizontal flipping.

To balance inference speed and feature extraction quality, we adopted the SAM-ViT-B architecture with completely frozen weights as the powerful semantic teacher model. For our sparse annotation strategy, the video sequences were partitioned into 5-frame clips. Only the first frame of each clip utilized a dense mask as the absolute ground truth, while the subsequent four unlabeled frames participated exclusively in the spatiotemporal distillation process. The scalar hyperparameters for the joint loss function were empirically set to $\lambda_1 = 1.0$ for distillation, $\lambda_2 = 0.5$ for cycle consistency, and $\lambda_3 = 0.2$ for temporal smoothness. The visibility tolerance threshold τ for determining target occlusion was fixed at 0.80, the prototype momentum coefficient α at 0.90, and the re-identification threshold at 0.72. Training FSD-Net requires 13.6 GPU-hours with 15.8GB peak memory on one RTX 4090, compared with 9.8 GPU-hours and 8.9GB for the fully supervised student baseline; the additional 1.39× cost comes from frozen SAM-ViT-B teacher processing and is incurred only during training. During the inference phase, the heavy teacher model is completely discarded. Our lightweight student network alone processes the video streams, achieving 46.7 FPS, which fully satisfies the stringent real-time processing requirements of clinical endoscopic interventions.

4.2 Comparison with State-of-the-Art Methods

Table 1 and Fig. 4 summarize quantitative and qualitative comparisons against fully and weakly supervised baselines on the official SUN-SEG test split. Under sparse annotations our method still surpasses strong fully supervised models on SUN-SEG-Hard and achieves competitive or superior performance on CVC-612 and CVC-300 in terms of mDice, mIoU and mHD, indicating good robustness and generalization. The prompt-free row reports inference without any test-time SAM or point prompt: tracking is initialized only from the first-frame dense

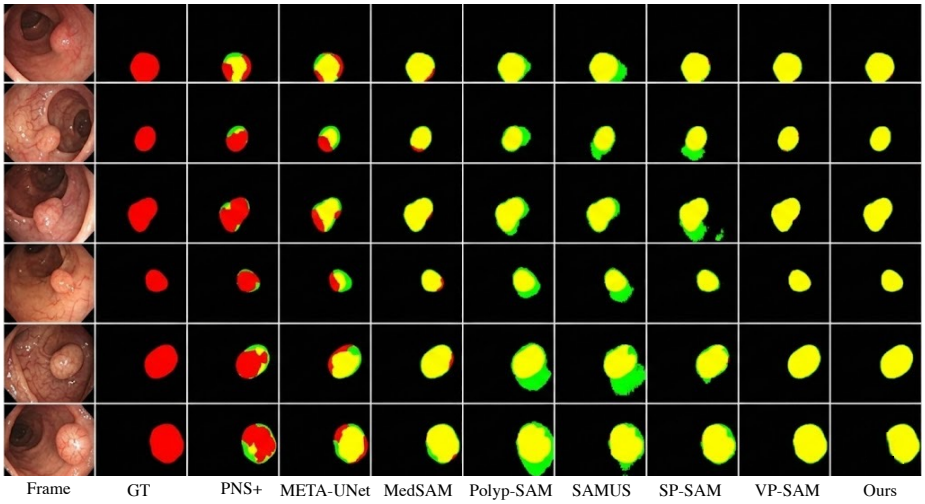


Fig. 4: Qualitative visual comparison of the proposed method against state-of-the-art baseline models across three different datasets. Red, green, and yellow denote predictions, ground-truth labels, and overlap, respectively.

mask during training, and at test time the student alone performs frame-wise segmentation. The prompt-assisted row additionally supplies one point per frame as a test-time prompt for fair comparison with promptable SAM baselines. As

Table 1: Quantitative comparison of the proposed framework against state-of-the-art methods on four challenging dataset splits. All SUN-SEG results are evaluated on the official test set. The best results are highlighted in bold.

Method	Year	Backbone	SUN-SEG-Easy			SUN-SEG-Hard			CVC-612			CVC-300		
			mDice	mIoU	mHD	mDice	mIoU	mHD	mDice	mIoU	mHD	mDice	mIoU	mHD
Fully Supervised Methods														
DCNet [36]	2023	PVT-V2	78.36	68.63	29.24	75.04	65.55	28.48	88.21	80.85	25.77	86.24	78.25	18.26
PNS+ [15]	2022	Res2Net-50	79.26	70.24	26.38	76.51	68.11	28.67	90.06	83.43	21.79	86.59	78.24	15.98
META-UNet [32]	2023	ResNet-34	81.17	72.43	25.71	80.16	70.55	26.27	90.64	84.39	20.49	86.64	78.55	16.02
Polyp-SAM-FS [17]	2023	ViT-B	81.33	73.18	25.19	80.52	71.33	25.82	90.47	84.58	21.29	86.83	78.72	16.53
Semi-Supervised Foundation Methods														
SAM [16]	2023	ViT-B	54.67	46.42	312.97	55.53	47.09	296.93	59.68	49.54	286.37	65.01	56.26	205.38
MedSAM [20]	2024	ViT-B	69.04	60.29	220.35	68.23	58.71	207.92	76.08	66.82	182.58	79.02	70.69	113.32
Polyp-SAM-Prompt [27]	2024	ViT-B	70.80	61.37	151.36	70.42	60.31	151.24	77.76	68.58	135.33	80.96	71.33	91.52
SAMed [37]	2023	ViT-B	78.23	67.99	28.58	76.94	66.78	28.99	89.92	83.33	23.69	86.41	78.98	16.38
SAMUS [18]	2024	ViT-B	81.99	73.37	25.55	80.41	71.37	26.44	90.23	83.68	23.98	86.81	79.23	16.12
SP-SAM [38]	2025	ViT-B	84.34	77.21	22.36	83.25	73.34	24.86	90.47	85.06	21.15	87.09	79.52	14.98
VP-SAM [9]	2024	ViT-B	84.83	77.48	21.72	84.11	75.04	22.52	91.12	85.15	20.24	87.17	79.41	14.33
Ours (prompt-free)	-	PVTv2-B0	85.62	78.16	21.22	85.28	77.16	21.38	92.33	86.79	19.34	88.26	80.17	14.06
Ours (1 pt/frame)	-	PVTv2-B0	**87.56**	**80.04**	**19.80**	**87.04**	**79.20**	**20.64**	**93.54**	**88.83**	**17.86**	**89.93**	**82.38**	**12.38**

shown in Fig. 4, our method produces cleaner boundaries and fewer missed detections than competing approaches across challenging SUN-SEG, CVC-612 and CVC-300 examples.

4.3 Ablation Studies

To deeply analyze the specific contributions of the proposed framework, we conducted detailed ablation experiments across six dimensions on the Easy and Hard subsets of the official SUN-SEG test set. All experiments are based on a unified sparse annotation baseline utilizing a frozen-weight SAM as the encoder.

Effectiveness of Synergistic Enhancement of Core Modules. As can be observed from Table 2, after adding the three modules OACC, SFD, and PRM, our model showed a maximum improvement of 7.89% on mDice, 9.05% on mIoU and a maximum decrease of 5.02 *mm* on mHD compared to adding only a single module. Furthermore, Figure 9 also shows that adding all three modules simultaneously yields better segmentation results than adding only one of them. This further indicates that these three modules are complementary: combining semantic flow, occlusion awareness, and prototype re-identification, they produce consistent, gradual improvements on both simple and difficult subsets, resulting in the most balanced overall performance.

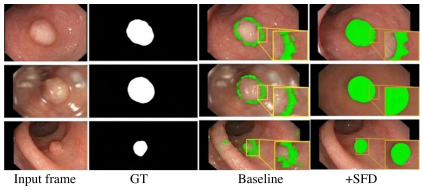


Fig. 5: Visual comparison results verifying the effectiveness of the semantic flow distillation module against baseline optical flow.

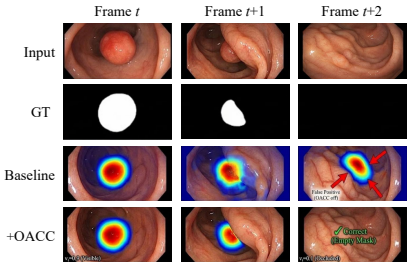


Fig. 6: Visual performance of the effective occlusion-aware cycle consistency module when processing target disappearance scenes.

Effectiveness of SFD. As can be observed in Table 4, simply adding Spatial Knowledge Distillation only brings a 2.64% improvement in mIoU on the SUN-SEG-Hard dataset compared to baseline optical flow propagation, while our proposed SFD brings a 5.27% improvement in mIoU compared to baseline optical flow propagation. This demonstrates that our proposed SFD can generate stronger supervision signals, thereby further improving the segmentation performance of our model. Furthermore, the visualization example in Figure 5 shows that the uncertainty weighting method can effectively suppress noisy pseudo-labels and recover smooth, accurate boundaries.

Effectiveness of OACC. As shown in Table 5, simply enforcing standard cycle consistency resulted in a 0.86% decrease in mDice on the SUN-SEG-Hard dataset, indicating that simply enforcing standard cycle consistency propagates

Table 2: Comprehensive orthogonal ablation study verifying the synergistic enhancement among all proposed modules.

SFD	OACC	PRM	SUN-SEG-Easy			SUN-SEG-Hard		
			mDice	mIoU	mHD	mDice	mIoU	mHD
✓			85.62	79.50	20.85	79.42	70.15	25.66
	✓		85.11	79.05	20.55	79.15	71.04	24.89
		✓	85.26	78.42	20.97	79.56	70.33	24.12
✓	✓		86.65	79.80	20.30	85.31	76.54	21.05
✓		✓	86.20	79.67	20.40	85.11	76.88	21.45
	✓	✓	86.90	79.78	20.90	85.79	76.02	21.11
✓	✓	✓	**87.56**	**80.04**	**19.80**	**87.04**	**79.20**	**20.64**

Table 3: Ablation analysis on the quantity and positional strategy of the provided prompt masks.

Prompt Strategy	SUN-SEG-Easy			SUN-SEG-Hard		
	mDice	mIoU	mHD	mDice	mIoU	mHD
Free	85.62	78.16	21.22	85.28	77.16	21.38
Middle Frame Only	87.53	78.65	20.53	83.20	74.57	24.16
Last Frame Only	86.85	77.23	20.85	82.56	73.86	25.53
First Frame Only	**87.56**	**80.04**	**19.80**	**87.04**	**79.20**	**20.64**
First Middle Frame	86.24	78.56	20.43	86.56	79.04	20.80
First Last Frames	86.83	79.23	20.52	86.86	78.56	21.16
Middle Last Frame	86.26	79.52	21.57	86.56	79.04	21.80
First Middle Last Frames	86.10	79.84	20.80	86.10	78.26	21.50

Table 4: Quantitative evaluation of the semantic flow distillation module using standardized segmentation metrics.

Method Variant	SUN-SEG-Easy			SUN-SEG-Hard		
	mDice	mIoU	mHD	mDice	mIoU	mHD
Baseline (Optical Flow)	79.15	68.22	26.55	75.24	64.88	31.45
+ Spatial Knowledge Distillation	82.30	72.10	24.10	78.37	67.52	28.16
+ SFD	**85.62**	**79.50**	**20.85**	**79.42**	**70.15**	**25.66**

Table 5: Performance evaluation comparing standard cycle consistency and the proposed occlusion-aware cycle consistency.

Method Variant	SUN-SEG-Easy			SUN-SEG-Hard		
	mDice	mIoU	mHD	mDice	mIoU	mHD
Baseline + SFD	85.62	79.50	20.85	79.42	70.15	25.66
+ Standard Cycle Consistency	84.86	78.76	21.52	78.56	69.87	26.36
+ OACC	**86.65**	**79.80**	**20.30**	**85.31**	**76.54**	**21.05**

Table 6: Sensitivity analysis of the model performance concerning variations in the visibility tolerance threshold.

Visibility Threshold	SUN-SEG-Easy			SUN-SEG-Hard		
	mDice	mIoU	mHD	mDice	mIoU	mHD
0.60	87.15	79.56	20.52	84.66	75.24	22.43
0.70	86.56	79.18	20.65	86.12	77.85	20.81
0.80	**87.56**	**80.04**	**19.80**	**87.04**	**79.20**	**20.64**
0.90	87.15	79.55	20.21	86.43	76.55	21.37
0.95	87.20	78.40	20.10	86.12	79.10	21.50

Table 7: Quantitative assessment of the prototype re-identification module showing the impact of momentum updating.

Method Variant	SUN-SEG-Easy			SUN-SEG-Hard		
	mDice	mIoU	mHD	mDice	mIoU	mHD
Baseline + SFD + OACC	86.65	79.80	20.30	85.31	76.54	21.05
+ Static Prototype	86.75	79.96	20.25	85.35	78.10	20.65
+ PRM	**87.56**	**80.04**	**19.80**	**87.04**	**79.20**	**20.64**

errors when the target disappears. Additionally, we observed that using cycle consistency increased our model's mHD by 0.70 *mm* on the SUN-SEG-Hard dataset. In contrast, our proposed OACC improved mDice by 5.89% and reduced mHD by 4.61 *mm* on the SUN-SEG-Hard dataset by suppressing gradients during low-visibility frames, consistent with the qualitative comparison results in Figure 6.

Effectiveness of PRM. Table 7 shows that adding a static prototype improves mDice by 0.04% and reduces mHD by 0.40 *mm* on the SUN-SEG-Hard dataset. The changes in mDice and mHD above indicate that the improvement brought by static prototypes is very limited because they cannot adapt to complex long-term deformations. However, after integrating the momentum update prototype re-identification module, the peak mDice value on the hard subset reaches 87.04%, and the mHD value is reduced by 0.41 *mm* compared to Baseline + SFD + OACC. This demonstrates that our proposed PRM can better remember the long-term morphological changes of polyp lesions. Figure 7 illustrates a key failure case: without this module, the network completely loses its tracking response when the polyp re-enters the field of view after prolonged oc-

clusion. However, our proposed module elicits an extremely high local response peak at the moment the polyp is exposed. The heatmap at the bottom accurately reflects how this strong activation signal guides the network to immediately output an accurate mask. This eliminated long-term missed detections and directly contributed to the improvement of the metrics in Table 7.

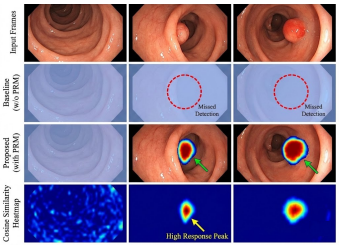


Fig. 7: Cosine similarity heatmap and segmentation recovery results of the proposed prototype re-identification module upon target reappearance.

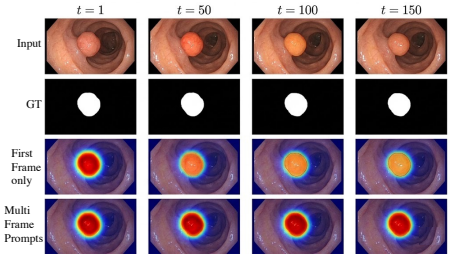
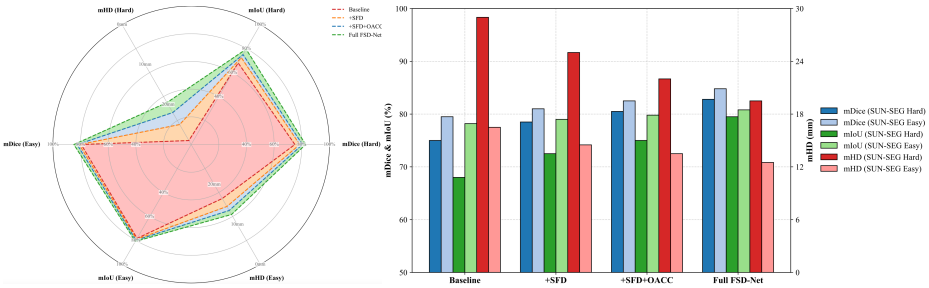


Fig. 8: Visual feature evolution maps showing stable boundary propagation using only the first frame as a prompt mask compared to multi-frame prompts.



(a) (b)

Fig. 9: Quantitative progressive bar chart illustrating the incremental performance gains of various module combinations.

Effectiveness of Prompt Mask Strategy. Table 3 further evaluates various cueing strategies. Using only the first frame as a cue achieves a good balance between annotation cost and accuracy. As shown in Table 3, compared to other cueing methods, using the first frame as a cue can improve accuracy by up to 4.48% on mDice and reduce it by up to 4.89 *mm* on mHD in the SUN-SEG-Hard dataset. Furthermore, Table 3 shows that increasing the number of cue items sometimes does not bring marginal benefits and requires additional labels. Figure 8 shows that our propagation mechanism can still maintain high-quality masks at the end of long sequences.

Effectiveness of Hyperparameter Sensitivity Analysis. Table 6 analyzes the model's sensitivity to the visibility tolerance threshold on simple and hard

subsets. As the threshold increases from 0.60 to 0.95, key metrics such as mIoU, mDice, and mHD show significant fluctuations. Notably, thresholds that are too low or too high lead to severe performance degradation, validating 0.80 as the optimal choice.

5 Discussion and Limitations

While the proposed FSD-Net exhibits state-of-the-art performance under sparse annotations, it is crucial to acknowledge its limitations in certain challenging extreme clinical scenarios. First, our occlusion-aware cycle consistency heavily relies on the semantic visibility score provided by the frozen foundation teacher model. If a polyp shares an identical texture and morphological structure with the surrounding healthy intestinal folds (i.e., highly camouflaged flat polyps), the foundation model might yield inaccurate visibility estimations, potentially causing the student network to prematurely suppress its gradients. Second, the prototype re-identification module maintains a global memory bank updated via a momentum mechanism. If the endoscope undergoes a violent camera withdrawal causing the polyp to disappear from the field of view for an extensively long duration (e.g., several minutes), the stored semantic prototype may lose its matching sensitivity against drastic perspective changes upon the polyp's eventual reappearance. Future work will focus on integrating depth estimation priors and 3D geometric reconstruction of the colon wall to provide a more robust, spatial-aware occlusion reasoning mechanism. Furthermore, we plan to optimize the student network using structural pruning and quantization techniques to facilitate seamless, low-power deployment on resource-constrained edge devices directly within the endoscopy suite, bridging the gap from algorithmic design to real-world surgical application.

6 Conclusion

This paper presents FSD-Net, a novel foundation-guided spatiotemporal distillation framework addressing the prohibitive annotation costs in video polyp segmentation. By integrating semantic flow distillation, occlusion-aware cycle consistency, and prototype re-identification, our model successfully handles complex endoluminal noise, dynamic occlusions, and target reappearance. Extensive experiments on three standardized datasets demonstrate that FSD-Net establishes a new state-of-the-art for weakly-supervised VPS, offering a highly robust, label-efficient, and real-time solution for clinical computer-assisted diagnosis.

References

1. Cao, M., Wei, F., Xu, C., Geng, X., Chen, L., Zhang, C., Zou, Y., Shen, T., Jiang, D.: Iterative proposal refinement for weakly-supervised video grounding. In: CVPR. pp. 6524–6534 (2023)

2. Cen, Z., Guo, N., Xu, W., Feng, Z., Huang, D.: Static-dynamic class-level perception consistency in video semantic segmentation. arXiv preprint arXiv:2412.08034 (2024)
3. Chen, G., Wu, H., Qin, J.: Stddnet: Harnessing mamba for video polyp segmentation via spatial-aligned temporal modeling and discriminative dynamic representation learning. In: ICCV. pp. 21364–21373 (2025)
4. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., et al.: Sam-med2d. arXiv preprint arXiv:2308.16184 (2023)
5. Dong, B., Wang, W., Fan, D.P., Li, J., Fu, H., Shao, L.: Polyp-pvt: Polyp segmentation with pyramid vision transformers. arXiv preprint arXiv:2108.06932 (2021)
6. Dorent, R., Joutard, S., Shapey, J., Bisdas, S., Kitchen, N., Bradford, R., Saeed, S., Modat, M., Ourselin, S., Vercauteren, T.: Scribble-based domain adaptation via co-segmentation. In: MICCAI. pp. 479–489. Springer (2020)
7. Duke, B., Ahmed, A., Wolf, C., Aarabi, P., Taylor, G.W.: Sstvos: Sparse spatiotemporal transformers for video object segmentation. In: CVPR. pp. 5912–5921 (2021)
8. Fan, D.P., Ji, G.P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L.: Pranet: Parallel reverse attention network for polyp segmentation. In: MICCAI. pp. 263–273. Springer (2020)
9. Fang, Z., Liu, Y., Wu, H., Qin, J.: Vp-sam: Taming segment anything model for video polyp segmentation via disentanglement and spatio-temporal side network. In: ECCV. pp. 367–383. Springer (2024)
10. Hinton, G.E., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
11. Huang, Y., Yang, X., Liu, L., Zhou, H., Chang, A., Zhou, X., Chen, R., Yu, J., Chen, J., Chen, C., et al.: Segment anything model for medical images? Medical Image Analysis **92**, 103061 (2024)
12. Jeong, S.M., Lee, S.G., Seok, C.L., Lee, E.C., Lee, J.Y.: Lightweight deep learning model for real-time colorectal polyp segmentation. Electronics **12**(9), 1962 (2023)
13. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., De Lange, T., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: MMM. pp. 451–462. Springer (2019)
14. Ji, G.P., Chou, Y.C., Fan, D.P., Chen, G., Fu, H., Jha, D., Shao, L.: Progressively normalized self-attention network for video polyp segmentation. In: MICCAI. pp. 142–152. Springer (2021)
15. Ji, G.P., Xiao, G., Chou, Y.C., Fan, D.P., Zhao, K., Chen, G., Van Gool, L.: Video polyp segmentation: A deep learning perspective. Machine Intelligence Research **19**(6), 531–549 (2022)
16. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: ICCV. pp. 4015–4026 (2023)
17. Li, Y., Hu, M., Yang, X.: Polyp-sam: Transfer sam for polyp segmentation. In: Medical Imaging (2023)
18. Lin, X., Xiang, Y., Zhang, L., Yang, X., Yan, Z., Yu, L.: Samus: Adapting segment anything model for clinically-friendly and generalizable ultrasound image segmentation. arXiv preprint arXiv:2309.06824 **4**(11) (2023)
19. Liu, X., Isa, N.A.M., Chen, C., Lv, F.: Colorectal polyp segmentation based on deep learning methods: A systematic review. Journal of Imaging **11**(9), 293 (2025)
20. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature communications **15**(1), 654 (2024)

21. Mansoori, M., Shahabodini, S., Abouei, J., Plataniotis, K.N., Mohammadi, A.: Polyp sam 2: Advancing zero-shot polyp segmentation in colorectal cancer detection. In: 2025 IEEE 5th International Conference on Human-Machine Systems (ICHMS). pp. 1–6. IEEE (2025)
22. Mazurowski, M.A., Dong, H., Gu, H., Yang, J., Konz, N., Zhang, Y.: Segment anything model for medical image analysis: an experimental study. Medical Image Analysis **89**, 102918 (2023)
23. Misera, L., Müller-Franzes, G., Truhn, D., Kather, J.N.: Weakly supervised deep learning in radiology. Radiology **312**(1), e232085 (2024)
24. Pan, Z., Jiang, P., Wang, Y., Tu, C., Cohn, A.G.: Scribble-supervised semantic segmentation by uncertainty reduction on neural representation and self-supervision on neural eigenspace. In: ICCV. pp. 7416–7425 (2021)
25. Qu, Y., Lu, T., Zhang, S., Wang, G.: Scribsd+: Scribble-supervised medical image segmentation based on simultaneous multi-scale knowledge distillation and class-wise contrastive regularization. Computerized Medical Imaging and Graphics **116**, 102416 (2024)
26. Rachavarapu, K.K., Ramakrishnan, K., et al.: Weakly-supervised audio-visual video parsing with prototype-based pseudo-labeling. In: CVPR. pp. 18952–18962 (2024)
27. Rahman, M.M., Munir, M., Jha, D., Bagci, U., Marculescu, R.: Pp-sam: Perturbed prompts for robust adaption of segment anything model for polyp segmentation. In: CVPR. pp. 4989–4995 (2024)
28. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F.: Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **71**(3), 209–249 (2021)
29. Tomar, N.K., Jha, D., Bagci, U., Ali, S.: Tganet: Text-guided attention for improved polyp segmentation. In: MICCAI. pp. 151–160. Springer (2022)
30. Wei, J., Hu, Y., Zhang, R., Li, Z., Zhou, S.K., Cui, S.: Shallow attention network for polyp segmentation. In: MICCAI. pp. 699–708. Springer (2021)
31. Wu, H., Wang, X.: Contrastive learning of image representations with cross-video cycle-consistency. In: ICCV. pp. 10149–10159 (2021)
32. Wu, H., Zhao, Z., Wang, Z.: Meta-unet: Multi-scale efficient transformer attention unet for fast and high-accuracy polyp segmentation. IEEE Transactions on Automation Science and Engineering **21**(3), 4117–4128 (2024)
33. Wu, J., Wang, Z., Hong, M., Ji, W., Fu, H., Xu, Y., Xu, M., Jin, Y.: Medical sam adapter: Adapting segment anything model for medical image segmentation. Medical image analysis **102**, 103547 (2025)
34. Yang, Z., Simon, R., Li, Y., Linte, C.A.: Dense depth estimation from stereo endoscopy videos using unsupervised optical flow methods. In: Annual Conference on Medical Image Understanding and Analysis. pp. 337–349. Springer (2021)
35. Yin, M., Wang, F., Ye, X., Meng, Y., Fu, Z.: Memory-augmented sam2 for training-free surgical video segmentation. In: MICCAI. pp. 328–337. Springer (2025)
36. Yue, G., Xiao, H., Xie, H., Zhou, T., Zhou, W., Yan, W., Zhao, B., Wang, T., Jiang, Q.: Dual-constraint coarse-to-fine network for camouflaged object detection. IEEE Transactions on Circuits and Systems for Video Technology **34**(5), 3286–3298 (2024)
37. Zhang, K., Liu, D.: Customized segment anything model for medical image segmentation. arXiv preprint arXiv:2304.13785 (2023)
38. Zhao, H., Liu, W., Ma, L., Xie, Z.: Self-prompting segment anything model for few-shot medical image segmentation. IET Computer Vision **19**(1), e70040 (2025)

39. Zhong, J., Wang, W., Wu, H., Wen, Z., Qin, J.: Polypseg: An efficient context-aware network for polyp segmentation from colonoscopy videos. In: MICCAI. pp. 285–294. Springer (2020)