VPSentry: Semi-supervised Video Polyp Segmentation via Sentry-guided Long-term Prototype Fusion with Correlation Dynamic Propagation

Guilian Chen[1], Xiaoling Luo[1], Huisi Wu[1]*, Jing Qin[2]

[1]College of Computer Science and Software Engineering, Shenzhen University
[2]Centre for Smart Health, School of Nursing, The Hong Kong Polytechnic University
2450101015@mails.szu.edu.cn, hswu@szu.edu.cn

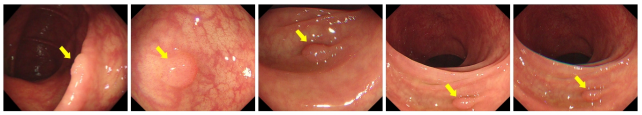
Abstract

Automated polyp segmentation in colonoscopy videos is an essential computer-aided technology for early detection and removal of polyps. However, most existing video polyp segmentation methods are designed with pixel-level temporal learning mechanisms, at the cost of time-consuming frame-wise annotations. In this paper, we present VPSentry, a novel semi-supervised segmentation model with a sentry mechanism. Our model integrates a prototype memory to store the long-term spatiotemporal cues of colonoscopy videos. Moreover, we devise adaptive prototypes to capture and generalize critical representations from individual frames, enabling long-term temporal fusion across labeled and unlabeled frames. In addition, we propose a correlation dynamic propagation module that propagates information from prototypes to features while simultaneously extracting dynamic features to perceive variations in polyp details between adjacent frames. Since colonoscopy scenes may change among consecutive frames, we further employ a sentry mechanism to assess the inter-frame continuity. This mechanism guides the prototype memory updating and the correlation dynamic propagation, further facilitating robust temporal propagation and dynamic detail perception for semi-supervised learning of long-term colonoscopy video sequences. Extensive experiments on the large-scale SUN-SEG dataset demonstrate that our model achieves optimal segmentation performance with real-time inference efficiency.

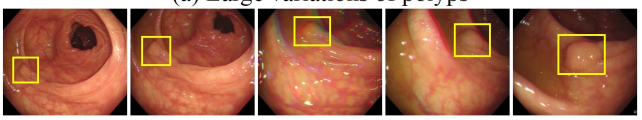
Introduction

Colorectal cancer (CRC) is the third most common cancer worldwide, with high morbidity and mortality (Bray et al. 2024). Early detection and removal of polyps through colonoscopy prevent their progression to CRC. As an essential step in computer-aided diagnosis, automated polyp segmentation in colonoscopy videos significantly reduces polyp miss rates, and hence improves the screening accuracy.

Recently, various deep learning models have been proposed to improve the performance of static image polyp segmentation (IPS) (Zhong et al. 2020; Duc et al. 2022; Sharma et al. 2024). However, video polyp segmentation (VPS) requires temporal consistency to continuously localize polyp targets in colonoscopy videos. It still remains a challenging task due to several factors, including large variations of polyps in terms of scale, shape, contrast, and position across different colonic structures (see Fig. 1a), dramatic scene changes across consecutive frames caused by rapid endoscope movement and jitter (see Fig. 1b), and blurred boundaries between polyps and their surrounding tissues in most cases. Moreover, clinical implementation necessitates accurate segmentation with real-time inference to meet intraoperative endoscopic requirements.



(a) Large variations of polyps



(b) Dramatic scene changes across consecutive frames

Figure 1: Challenges in VPS include (a) large variations of scale, shape, contrast, and position, along with (b) dramatic scene changes in colonoscopy sequences.

To address above challenges, VPS methods harness various spatiotemporal learning techniques, such as normalized self-attention (Ji et al. 2021, 2022) and memory (Hu et al. 2024) mechanisms. These methods, however, require full pixel-wise supervision to model inter-frame correlations, demanding labor-intensive frame-wise annotations. In cases with limited annotations, their accuracy for long-term colonoscopy videos may degrade significantly. Thus, semi-supervised learning (Wu et al. 2023; Ren et al. 2023; Jia et al. 2024) offers a viable alternative in clinic to improve polyp localization and segmentation. To model spatiotemporal dependencies in VPS, SSTAN (Zhao et al. 2022) employs a hybrid transformer architecture, while TCCNet (Li et al. 2022) and STCSR-Net (Li et al. 2025a) utilize a dual-branch co-training schema to mutually constrain segmentation and propagation branches for temporal consistency on unlabeled data. These methods are trained in a label-consistent manner, ignoring temporal variations across frames and generating suboptimal predictions in long-term sequences. There-

*Corresponding Author

fore, PSDNet (Hu et al. 2025) propagates first-frame annotation via a propagative teacher, and generates time-invariant pseudo-labels with a semantic teacher. Though built on a unified memory mechanism, it roughly fuses features without assessing inter-frame continuity. This risks introducing unexpected noise into the long-term representation storage when colonoscopy scene changes dramatically. In addition, these pseudo-label based VPS methods may accumulate misleading information from low-quality pseudo-labels and propagate such inaccuracies into long-term sequences, adversely affecting the segmentation results.

In this paper, we propose a novel semi-supervised segmentation model with a sentry mechanism, namely VPSentry. Specifically, we design adaptive prototypes to capture and generalize critical features from individual frames, complemented by a prototype memory tailored for long-term information storage. Our model ensures temporal propagation on unlabeled frames via adaptive prototypes and enhances the prototype fusion by focusing on critical features, avoiding pixel-level misleading information accumulation in long-term sequences. Although long-term temporal consistency enables consistent polyp localization across frames, it struggles to perceive dynamic variations, particularly the boundary details. In this regard, we propose a correlation dynamic propagation (CDP) module to jointly perform efficient long-term information propagation and dynamic detail exploration. CDP propagates the long-term information based on the correlation map computed between current-frame features and enhanced prototypes, while simultaneously exploring frame-wise dynamic details through divergence calculation between current-frame features and the prototypes of adjacent frames. More crucially, we innovate a sentry network to assess colonoscopy scenes continuity, generating judgment scores to guide the prototype memory updating and the correlation dynamic propagation. Therefore, our model can produce more stable and accurate segmentation predictions. For semi-supervised training, we supervise the first frame to provide basic prototype areas and the last frame to enhance the temporal propagation robustness. Extensive experiments on publicly available SUN-SEG (Ji et al. 2022) dataset demonstrate that the proposed VPSentry achieves a better efficiency-accuracy balance versus state-of-the-art methods.

Our major contributions are summarized as follows:

- We present a novel semi-supervised VPS model that propagates spatiotemporal information through adaptive prototype. By incorporating a sentry mechanism to assess inter-frame continuity, our model can produce more robust segmentation results for colonoscopy videos.
- We propose a CDP module to enhance long-term information propagation from prototypes to features and perform frame-wise dynamic features extraction. This synergy of temporal coherence and detail perception enables more accurate unlabeled-frame segmentation in long-term sequences under minimal supervision.
- Our method outperforms other state-of-the-art approaches on the large-scale SUN-SEG dataset, achieving a better trade-off between efficiency and accuracy.

Related Work

Automated Video Polyp Segmentation

Remarkable progress has been achieved in automated polyp segmentation with deep learning techniques (Akbari et al. 2018; Fan et al. 2020; Bui et al. 2024), surpassing traditional methods (Mamonov et al. 2014; Tajbakhsh, Gurudu, and Liang 2015) that rely on predefined features. However, these image polyp segmentation (IPS) methods fail to capture temporal information, leading to performance degradation in colonoscopy videos.

For VPS, PNS-Net (Ji et al. 2021) proposes a normalized self-attention module to aggregate temporal features, extended by PNS+ (Ji et al. 2022) through a progressive global-to-local learning strategy to improve short-term dependency. SALI (Hu et al. 2024) integrates short-term alignment and long-term interaction modules with a unified memory for temporal consistency modeling. In addition, Diff-VPS (Lu et al. 2024) employs diffusion model (Ji et al. 2023) to learn complex polyp textures and designs a temporal reasoning module (TRM) for frame reconstruction from previous sequences. VP-SAM (Fang et al. 2024) develops a foundation model (Kirillov et al. 2023) to segment polyps from backgrounds while tracking their motion status. However, these VPS methods require fully supervised training with frame-wise annotations, making them vulnerable in clinical videos where labels are scarce.

Semi-supervised Polyp Segmentation

Given limited annotations, some semi-supervised methods (Basak and Yin 2023; Pan et al. 2025; Lei et al. 2025) improve model segmentation performance by utilizing unlabeled data. RD-Net (Li et al. 2025b) feeds depth images into an auxiliary student network to transfer cross-modal knowledge during training, improving pseudo-labels with a depth-guided patch augmentation manner. For modeling temporal dependencies in colonoscopy videos, SSTAN (Zhao et al. 2022) proposes a proximity frame time-space attention (PFTSA) module to capture long-term contexts, and designs a temporal local context attention (TLCA) module to refine predictions based on the differences between adjacent frames. Additionally, TCCNet (Li et al. 2022) proposes sequence-corrected and propagation-corrected reverse attention (SC-RA/PC-RA) modules to maintain prediction coherence across consecutive frames by correcting the error in saliency maps. However, the segmentation performance of these models still suffers from low-quality pseudo-labels. In this work, we prioritize the long-term temporal propagation and dynamic exploration in colonoscopy video sequences, rather than improving pseudo-labels. By leveraging adaptive prototypes and using a sentry mechanism to judge inter-frame continuity, our VPSentry prevents pixel-level misleading information accumulation and significantly improves semi-supervised segmentation performance.

Methods

VPSentry processes and propagates long-term information across labeled and unlabeled frames using adaptive prototypes, as shown in Fig. 2. First, a four-stage image en-

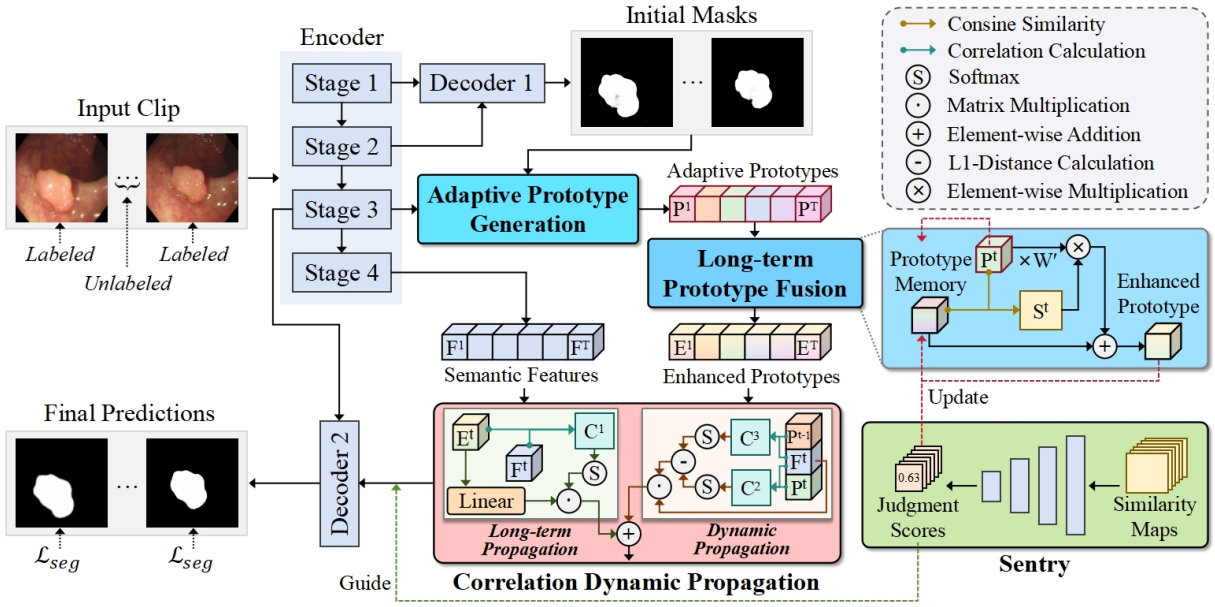


Figure 2: Overview of VPSentry, which processes and propagates long-term information through adaptive prototypes. We employ a prototype memory to store and aggregate information from historical frames. During long-term prototype fusion, frame-wise similarity maps S^t are computed. A sentry leverages these similarity maps to produce judgment scores, which are used to guide the prototype memory update and correlation dynamic propagation.

coder extracts multi-scale features from the input clip. Then, initial masks are generated from low-level features. Adaptive prototypes are created based on the initial masks and third-stage features. Such kind of prototypes efficiently capture and propagate long-term spatiotemporal information across frames, mitigating the impact of limited annotations on temporal propagation. In addition, a discriminator network servers as a sentry to guide prototype memory update and correlation dynamic propagation, facilitating our model to generate more stable and accurate predictions for long-term colonoscopy video sequences.

Prototype Generation and Long-term Fusion

To consistently locate polyps in colonoscopy videos, models need to understand how polyp targets evolve over time in terms of scale, shape and position. Besides, due to the lack of frame-wise annotations, selecting essential feature regions for propagation across consecutive frames, rather than pixel-wise attention calculation, enhances efficiency in semi-supervised learning for long-term colonoscopy videos. This strategy also improves the understanding in unlabeled frames during temporal propagation. Given that similar foreground and background elements among frames strengthen spatiotemporal representations, we devise adaptive prototype to capture and generalize critical features.

Adaptive Prototype Generation. As shown in Fig. 3, we employ a clustering branch to extract latent tokens from the input features, while applying another branch to perform tokens selection and projection on the foreground and background regions. Specifically, the clustering branch adaptively generalizes each frame features along the channel dimension to produce latent tokens, while the other branch highlights and projects critical areas into the selected tokens for long-term attention. Combining both token types yields a compact adaptive prototype that achieves superior spatiotemporal generalization than original features. Thus, subsequent long-term fusion on the adaptive prototype can be efficiently performed at lower computational cost.

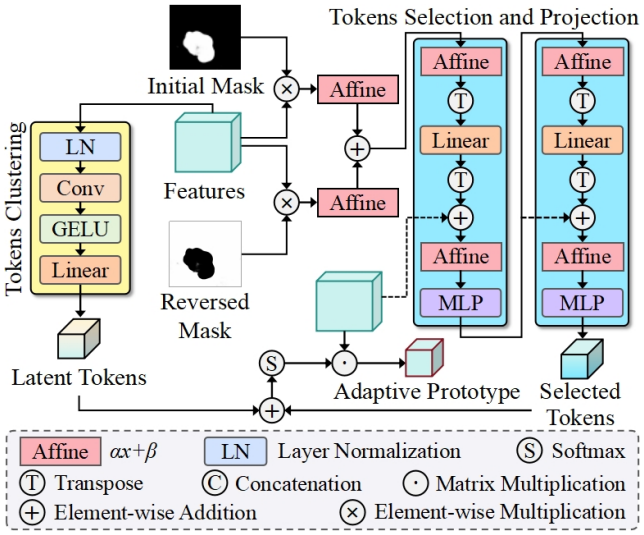


Figure 3: Illustration of Adaptive Prototype Generation.

Given an input clip with T frames, we extract multi-scale features by a four-stage encoder, which is expressed as $f_i^t = \{\text{Stage}_i(I^t), i = 1, 2, 3, 4\}$, where I^t represents

t-th ($1 \leq t \leq T$) frame with a resolution of $H \times W \times 3$. A decoder d_1 is then applied on f_1^t and f_2^t to produce initial mask, defined as $m^t = d_1(f_1^t, f_2^t)$. For prototype generation, the latent tokens $l^t = b_1(f_3^t)$ are produced by the clustering branch b_1, which comprises a layer normalization, a 3×3 convolution, a GELU activation function, and a linear layer. The convolution explores inter-channel correlations and compress the number of channels to $\frac{1}{4}$. Then, the GELU function activates these correlations while the linear layer computes a global weighting matrix over the compressed channels, thereby clustering essential features and maximizing channel-specific responses to obtain latent tokens. Concurrently, we use another branch b_2 to perform tokens selection and projection, which adaptively selects out critical foreground and background regions based on the initial and reversed masks:

$$s^t = b_2(f_3^t, \mathbf{A}(m^t \times f_3^t) + \mathbf{A}((1 - m^t) \times f_3^t)) \quad (1)$$

where s denotes the selected tokens and $\mathbf{A}$ represents an affine layer implementing $\alpha x + \beta$. The affine layer dynamically parameterizes different regions of the foreground and background features. In addition, b_2 adopts a simplified two-stage ResMLP (Touvron et al. 2022), with features selection and projection handled in different stages. ResMLP operates as an efficient transformer-like layer, comprising affine transformations, linear projection and MLP. Such architecture adaptively focuses on critical feature regions during training, facilitating subsequent token-to-prototype conversion. Finally, the adaptive prototype for t-th frame, denoted as P^t, is calculated based on the relevance of different positions. It can be defined as $P^t = f_3^t \odot \mathrm{Softmax}(l^t + s^t)$, where $\odot$ represents the matrix multiplication.

Long-term Prototype Fusion. Effective polyp tracking and localization in colonoscopy videos rely on propagating spatial and temporal information across frames. Therefore, we set a prototype memory to record the spatiotemporal information from past frames for long-term attention. This mechanism reinforces consistent areas across frames and aggregates the corresponding prototypes into the memory. As critical regions are adaptively selected out during the prototype generation, long-term prototype fusion omits the attention-like computation for shielding off irrelevant areas across frames, further improving the spatiotemporal aggregation efficiency in our segmentation model.

Long-term prototype fusion is demonstrated in Fig. 2. A cosine similarity map is calculated between current-frame prototype and the prototype memory, capturing their matching relationship. We use this map and weights allocation to align the current prototype with the prototype memory for subsequent enhancement. As a result, the long-term prototype is progressively enhanced through integration with frame-wise prototypes. We denote the prototype memory as Z and the similarity map as S. Then, the enhanced prototype of t-th frame E^t can be formulated as:

$$E^t = Z^t + S^t \times (W' \cdot P^t) \quad (2)$$

where W' represents the weights allocation achieved by a linear layer. In our correlation dynamic propagation module, the enhanced prototypes can be used to propagate stored long-term information to high-level features.

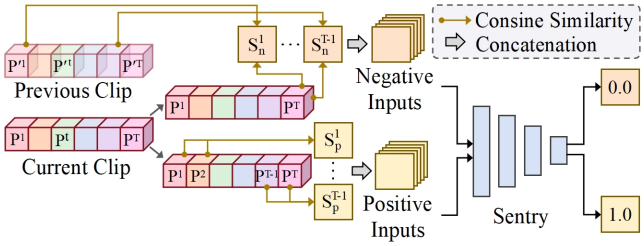


Figure 4: Overview of sentry training. It takes similarity maps between adjacent frames in current clip as positive inputs and cross-clip similarity maps as negative inputs. During training, the positive inputs are assigned with true labels while the negative ones are assigned with false labels.

Sentry

Rough fusion of current prototype with the prototype memory without continuity assessment may introduce unexpected noise during dramatic scene changes, degrading the prediction quality in long-term sequences. To address this, we propose a sentry mechanism to judge whether current-frame scene is consistent with previous frames.

The continuity assessment exploits the generalization superiority inherent to prototypes. Specifically, we calculate the similarity maps based on prototypes to provide an intuitive representation of scene continuity. When current-frame scene is consistent, its prototype may closely align with the prototype memory due to shared focus on similar foreground and background regions. As a result, higher positive values are obtained at corresponding positions in the similarity map. Conversely, scene changes may cause different region focus between the current-frame prototype and the memory, leading to mismatch and lower negative values in the similarity map. Accordingly, the sentry aims to generate judgment scores based on the similarity maps computed from different prototypes: higher scores for consistent scene and lower scores for dramatic scene changes.

For training the sentry, false labels (value as 0.0) are assigned for the negative similarity maps (derived from the last frame of current clip and frames of previous clip), while true labels (value as 1.0) are assigned for the positive pairs (derived from adjacent frames in current clip), as shown in Fig. 4. We denote the prototypes of previous clip as P'^t. The similarity maps are calculated as:

$$S_n^t = \mathrm{Cos}\langle P'^t, P^T \rangle, \quad 1 \leq t < T \quad (3)$$

$$S_p^t = \mathrm{Cos}\langle P^t, P^{t+1} \rangle, \quad 1 \leq t < T \quad (4)$$

where S_p and S_n denote positive and negative similarity maps, respectively. Cos represents cosine similarity calculation. We maximize scene divergence in negative pairs by calculating on the last frame of current clip and the frames of previous clip. Training with these similarity maps enables the sentry to identify inter-frame scene changes and produce corresponding judgment scores. These scores further guide the prototype memory updating following the formula as:

$$Z^t = \begin{cases} E^t, & \text{score} > c \\ P^t, & \text{score} \leq c \end{cases} \quad (5)$$

where c denotes a confidence threshold. When the judgment score exceeds the confidence threshold, the current-frame scene is considered consistent with the scene recorded in the prototype memory, and hence the enhanced long-term prototype will be updated to the memory. Otherwise, it indicates that the scene in current frame have changed comparing to historical frames, and the memory will be replaced by the prototype of current frame for reinitialization.

Correlation Dynamic Propagation

To propagate the long-term information from enhanced prototypes to high-level semantic features, we propose a correlation dynamic propagation (CDP) module. Although long-term information assists model in consistently locating the polyps in consecutive frames, it struggles to perceive detailed shape and boundary variations between adjacent frames. Therefore, dynamic feature extraction becomes essential to learn polyp textural details in unlabeled frames. By explicitly modeling regional attention shifts across prototypes of adjacent frames, we can extract dynamic features to compensate for boundary details.

The CDP module computes prototype-feature correlation maps to facilitate the information propagation from prototypes to features, as shown in Fig. 2. For long-term spatiotemporal propagation, it calculates the correlation maps between current-frame features and enhanced prototypes, transferring the long-term information recorded in prototypes to semantic features. For dynamic propagation, it generates distinct correlation maps calculated from current features and the prototypes of adjacent frames. These correlation maps, C^1, C^2 and C^3, are calculated as:

$$C^1 = \eta(E^t) \odot \eta(\text{Transpose}(F^t)) \tag{6}$$

$$C^2 = \eta(P^t) \odot \eta(\text{Transpose}(F^t)) \tag{7}$$

$$C^3 = \eta(P^{t-1}) \odot \eta(\text{Transpose}(F^t)) \tag{8}$$

where η denotes channel-wise L2 normalization. For simplicity, we use F^t to represent the high-level semantic features f_4^t of t-th frame. The intermediate products of the CDP module can be formulated as:

$${F_l}^t = \text{Softmax}(C^1) \odot \theta(E^t) \tag{9}$$

$${F_d}^t = \left|\text{Softmax}(C^2) - \text{Softmax}(C^3)\right| \odot F^t \tag{10}$$

where ${F_l}^t$ indicates the features enhanced by long-term information, and ${F_d}^t$ denotes the extracted dynamic features. θ represents a linear layer. According to Eq. 9, the enhanced prototype is converted by the linear layer, propagating its inherent long-term information to high-level semantic features based on the computed correspondence. In addition, L1 distance calculation between C^2 and C^3 highlights the variations in regional attention across prototypes, as formalized in Eq. 10. Furthermore, the correlation dynamic propagation is also guided by the judgment scores of the sentry:

$$O^t = \begin{cases} {F_d}^t + {F_l}^t, & \text{score} > c \\ {F_d}^t, & \text{score} \leq c \end{cases} \tag{11}$$

where O denotes the aggregated outputs of CDP, which are then fed into another decoder d_2 to generate refined predictions. We formulate the final prediction as $y^t = d_2(O^t, f_3^t)$.

Loss Functions

As the prototype generation and long-term propagation are the critical components of our semi-supervised model, we supervise on the first frame to provide basic prototype areas and the last frame to enhance the temporal propagation robustness. Those intermediate unlabeled frames are indirectly trained according to the temporal propagation onto the last frame. Specifically, we supervise initial masks and final predictions using weighted BCE and weighted IoU losses (Wei, Wang, and Huang 2020), improving the learning of global colonic structures and polyp boundaries. Given a prediction $M \in \{m^1, y^1, m^T, y^T\}$ and a corresponding ground truth Y, the segmentation loss is defined as:

$$\mathcal{L}_{seg}(M, Y) = \mathcal{L}_{Wbce}(M, Y) + \mathcal{L}_{WIoU}(M, Y) \tag{12}$$

In addition, we adopt standard BCE loss for sentry supervision, where we assign true labels (denoted as $\hat{\mathbf{1}}$) to positive similarity maps S_p and false labels (denoted as $\hat{\mathbf{0}}$) to negative similarity maps S_n:

$$\mathcal{L}_{sentry} = \mathcal{L}_{BCE}(S_n, \hat{\mathbf{0}}) + \mathcal{L}_{BCE}(S_p, \hat{\mathbf{1}}) \tag{13}$$

Experiments

Datasets and Evaluation Metrics

We evaluate our VPSentry on a public large-scale video polyp dataset, SUN-SEG (Ji et al. 2022). It contains 1,013 colonoscopy video sequences with 158,690 frames, among which 100 positive cases (49,136 frames) are used for both training and testing. SUN-SEG-Train contains 19,544 frames from 112 sequences. SUN-SEG-Easy includes 17,070 frames from 119 sequences while SUN-SEG-Hard includes 12,522 frames from 54 sequences. 'Easy' and 'Hard' are defined referring to the difficulty levels of each pathological category. Both test sets are split into 'Seen' and 'Unseen' sub-datasets based on the case similarity to the training set. For evaluation, we employ four most commonly used metrics to assess the segmentation performance, including mean Dice coefficient (Dice), structure measure ($\mathcal{S}_\alpha$ with $\alpha = 0.5$) (Fan et al. 2017), enhanced-alignment measure (E_ϕ) (Fan et al. 2018) and weighted F-measure (F_β^w) (Margolin, Zelnik-Manor, and Tal 2014).

Implementation Details

We train our VPSentry on a single NVIDIA GeForce RTX 4090 GPU with 24 GB memory. We apply random horizontal and vertical flipping, random zooming, and random rotation for data augmentation. Our encoder adopts the pretrained feature extractor of Res2Net-50 (Gao et al. 2019) and PVTv2-B2 (Wang et al. 2022), while the decoder follows the design in (Ji et al. 2022). During training, the initial learning rates of the segmentation components and the sentry network are set to $2e-4$ and $4e-4$, respectively. We train the model for 30 epochs with a batch size of 2. The input clip length is set to 20 frames when using 10% training data, and 10 frames for 20% data. All input frames are unified to a resolution of 352×352. For optimization, both the segmentation components and the sentry utilize the AdamW optimizer with a cosine annealing scheduler (Loshchilov and

Method	Task	Backbone	SUN-SEG-Easy (20%)				SUN-SEG-Hard (20%)				SUN-SEG-Easy (10%)				SUN-SEG-Hard (10%)			
			$\mathcal{S}_\alpha$	E_ϕ	F_β^w	Dice	$\mathcal{S}_\alpha$	E_ϕ	F_β^w	Dice	$\mathcal{S}_\alpha$	E_ϕ	F_β^w	Dice	$\mathcal{S}_\alpha$	E_ϕ	F_β^w	Dice
CML	MIS	UNet	81.64	83.12	71.93	72.38	82.06	84.86	71.71	73.01	79.38	81.20	67.33	68.56	79.80	83.10	67.34	69.31
PedSemiSeg	IPS	Res2Net-50	84.78	90.09	75.63	78.82	84.05	89.39	74.14	77.73	82.42	86.77	69.65	73.18	82.55	87.01	69.74	73.65
CFANet	IPS	Res2Net-50	85.09	87.54	77.06	77.57	84.83	87.30	76.31	76.99	82.84	85.33	74.30	74.70	82.52	85.36	72.88	73.68
MemSAM	EVS	ViT-B	89.02	92.37	82.42	84.26	87.66	91.98	79.50	82.17	83.38	86.18	72.62	76.80	82.38	85.19	70.65	74.83
SSTAN	VPS	Res2Net-50	87.75	89.84	80.44	80.12	88.03	90.91	80.34	80.14	85.84	87.76	77.91	76.56	85.70	87.87	76.63	75.53
TCCNet	VPS	Res2Net-50	88.07	87.45	78.14	78.85	88.18	88.09	77.51	79.02	86.77	87.27	76.03	77.17	86.71	87.19	75.47	76.62
PSDNet	VPS	PVTv2-B5	89.31	93.86	85.63	86.11	87.97	92.80	82.66	83.85	87.29	93.27	81.65	83.37	86.31	91.60	78.80	81.11
VPSentry	VPS	Res2Net-50	89.08	91.98	82.11	83.73	89.26	92.37	81.60	83.82	87.50	90.66	79.32	80.92	87.35	90.37	78.72	80.52
VPSentry	VPS	PVTv2-B2	**90.87**	**94.48**	**86.12**	**87.15**	**90.13**	**93.90**	**84.33**	**85.84**	**90.17**	**93.43**	**83.97**	**85.41**	**89.43**	**92.65**	**81.95**	**83.95**

Table 1: Quantitative comparisons with different state-of-the-art methods on SUN-SEG test sets (10% and 20% labeled data).

Method (10% labeled data)	$\mathcal{S}_\alpha$	E_ϕ	F_β^w	Dice
All components	**89.43**	**92.65**	**81.95**	**83.95**
w/o LP	88.72	91.93	80.62	81.78
w/o DP	88.65	92.05	81.32	82.67
w/o sentry	88.77	92.42	81.38	83.14
Baseline	86.98	89.90	75.67	77.23

Table 2: Statistical comparisons of ablation studies on the SUN-SEG-Hard test set for different components.

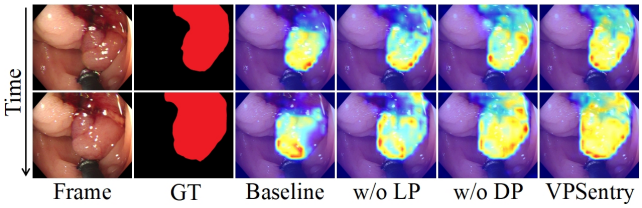


Figure 5: Features visualization of different methods.

Hutter 2016). In addition, the confidence threshold c for assessing the scene continuity is set to 0.5.

Tab. 2 presents the ablation variants of VPSentry. The baseline model adopts the feature encoder from PVTv2-B2 and the decoder from (Ji et al. 2022). We also visualize the feature maps of different methods, as shown in Fig. 5. The results demonstrate that removing long-term propagation (w/o LP) from the CDP module degrades performance, confirming its significance in learning long-term information for colonoscopy videos understanding. Furthermore, model without dynamic propagation (w/o DP) fails to perceive polyp details within each frame, leading to performance decline. In addition, by incorporating the sentry to guide the prototype memory updating and correlation dynamic propagation, the collaboration of all components achieves optimal performance.

To further evaluate the effectiveness of the proposed sentry mechanism, we visualize predictions with/without sentry in Fig. 6. For cases with rapid camera movement, rough long-term prototype fusion without sentry may fuse some unexpected noise in the long-term representations, generating suboptimal predictions in long-term frames when colonoscopy scene changes dramatically. However, the model with sentry is able to provide more stable segmentation results in such cases. According to Tab. 2 and Fig. 6, sentry's crucial value lies in providing robustness for the cases with dramatic scene changes, and it intervenes minimally on most continuous frames, thus yielding a modest performance improvement in average.

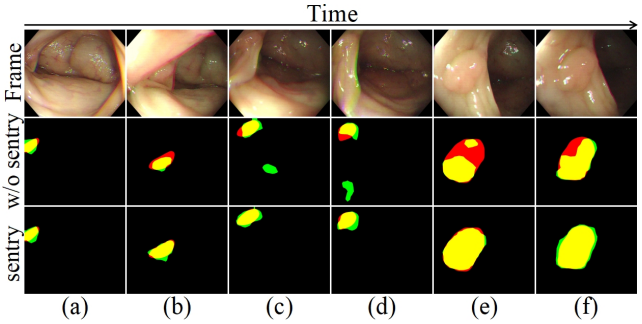


Figure 6: Predictions with/without sentry under fast camera moving and scene changes. Colors: red (ground truth), green (prediction), yellow (overlap regions).

Comparison with State-of-the-art Methods

To further validate the effectiveness of our VPSentry, we conduct comprehensive comparisons with different types of state-of-the-art semi-supervised methods. Among them, CML (Wu et al. 2024) is a medical image segmentation (MIS) model. PedSemiSeg (Wang et al. 2025) is an image polyp segmentation (IPS) model. We also develop CFANet (Zhao et al. 2025) into a semi-supervised IPS approach for comparison. In addition, MemSAM (Deng et al. 2024), designed for echocardiography video segmentation (EVS), has similar supervision strategy with ours by using the first and last frames for constraining the temporal propagation. Both SSTAN (Zhao et al. 2022) and TCCNet (Li et al. 2022) are VPS models. Moreover, we apply PSDNet (Hu et al. 2025) to 5-frame and 10-frame colonoscopy clips, corresponding to use 10% and 20% labeled data respectively for semi-supervised training.

The quantitative comparisons are demonstrated in Tab. 1. It shows that our proposed VPSentry based on the PVTv2-B2 backbone outperforms other competitors in all metrics. We also conduct performance-efficiency comparisons, as shown in Tab. 3 and Fig. 8. Compared to other methods, our

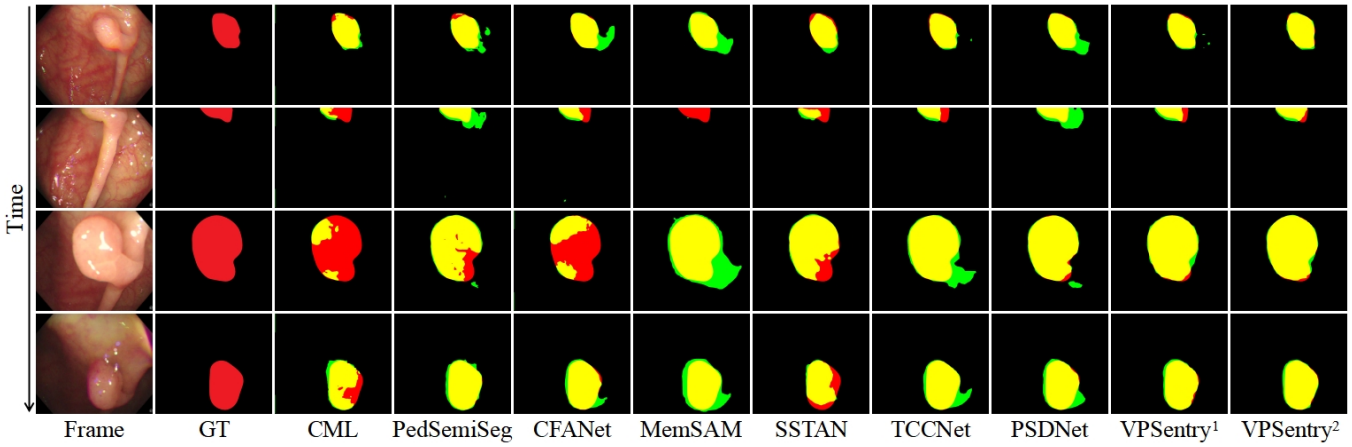


Figure 7: Visual comparisons with different competitors on the SUN-SEG test sets. All methods are trained with 10% labeled data. Red, green and yellow areas represent the ground truth, prediction and their overlapping regions, respectively. Note that VPSentry[1] uses Res2Net-50 as backbone, and VPSentry[2] uses PVTv2-B2 as backbone.

Method	Backbone	Dice	Param. (M)	GFLOPs	FPS
MemSAM	ViT-B	74.83	142.14	2848.80	5
SSTAN	Res2Net-50	75.53	118.09	496.60	36
TCCNet	Res2Net-50	76.62	25.03	257.32	11
PSDNet	PVTv2-B5	81.11	82.76	87.08	10
VPSentry	Res2Net-50	80.52	39.05	173.93	30
VPSentry	PVTv2-B2	83.95	27.20	106.50	28

Table 3: Performance-efficiency comparisons with different video segmentation methods on SUN-SEG-Hard test set.

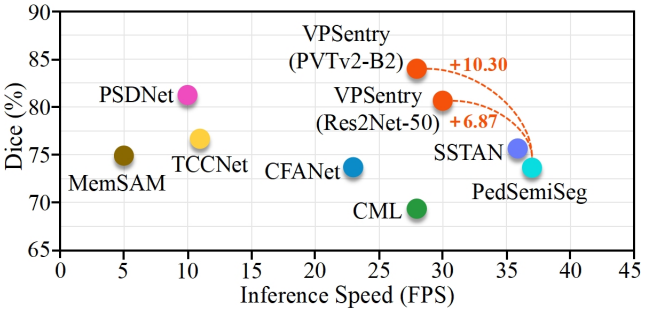


Figure 8: Performance-efficiency comparisons with other state-of-the-art methods on SUN-SEG-Hard test set.

VPSentry achieves a better balance between performance and efficiency. Visual comparisons with different methods on typical challenging cases are illustrated in Fig. 7. The proposed VPSentry demonstrates superior stability during scene variations and captures more edge details through dynamic features extraction.

Discussions and Limitations

We harness the sentry to improve semi-supervised VPS performance in our prototype memory based framework. By assessing inter-frame continuity, the sentry enables more effective long-term and dynamic propagation across labeled and unlabeled frames. Owing to severe annotation scarcity, our model still has some limitations in extreme cases with facula interference (Fig. 9 a-b) or extremely low-contrast scenarios (Fig. 9 c-f).

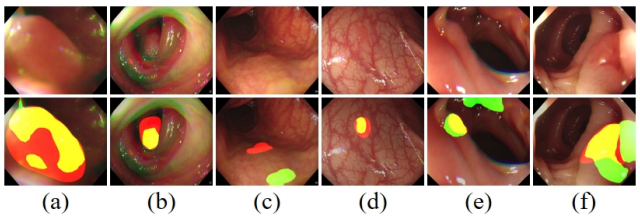


Figure 9: Failure cases. Red, green and yellow represent the GT, prediction and their overlapping regions, respectively.

Conclusion

In this paper, we present VPSentry for semi-supervised video polyp segmentation, leveraging a sentry mechanism to judge whether the scene of current frame is consistent from previous frames. Such kind of judging strategy enables our model to effectively learn from long-term colonoscopy video sequences while avoiding fusion with interference caused by dramatic scene changes. To achieve efficient long-term information fusion, we devise adaptive prototype tailored for capturing and generalizing critical features. Furthermore, we propose a correlation dynamic propagation module to achieve both long-term and dynamic propagation from prototypes to features. Consequently, this enables our model to produce more accurate segmentation for unlabeled frames through temporal propagation and dynamic details perception. Experimental results on SUN-SEG dataset demonstrate the effectiveness of our VPSentry. Future work could explore more prototype variants and apply the sentry mechanism in broader domains. In addition, more robust training strategy for the sentry network could be developed to adapt to diverse video sequence properties.

Acknowledgments

This work was supported partly by National Natural Science Foundation of China (No. 62273241), Natural Science Foundation of Guangdong Province, China (No. 2024A1515011946), the Shenzhen Research Foundation for Basic Research, China (No. JCYJ20250604181940054), the Hong Kong RGC Collaborative Research Fund (project no. C5055-24G) and the Shenzhen-Hong Kong-Macao Science and Technology Plan Project (Category C Project) under Shenzhen Municipal Science and Technology Innovation Commission (project no. SGDX20230821092359002).

References

Akbari, M.; Mohrekesh, M.; Nasr-Esfahani, E.; Soroushmehr, S. R.; Karimi, N.; Samavi, S.; and Najarian, K. 2018. Polyp segmentation in colonoscopy images using fully convolutional network. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 69–72. IEEE.

Basak, H.; and Yin, Z. 2023. Pseudo-label guided contrastive learning for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19786–19797.

Bray, F.; Laversanne, M.; Sung, H.; Ferlay, J.; Siegel, R. L.; Soerjomataram, I.; and Jemal, A. 2024. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 74(3): 229–263.

Bui, N.-T.; Hoang, D.-H.; Nguyen, Q.-T.; Tran, M.-T.; and Le, N. 2024. Meganet: Multi-scale edge-guided attention network for weak boundary polyp segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 7985–7994.

Deng, X.; Wu, H.; Zeng, R.; and Qin, J. 2024. Memsam: Taming segment anything model for echocardiography video segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9622–9631.

Duc, N. T.; Oanh, N. T.; Thuy, N. T.; Triet, T. M.; and Dinh, V. S. 2022. Colonformer: An efficient transformer based method for colon polyp segmentation. *IEEE Access*, 10: 80575–80586.

Fan, D.-P.; Cheng, M.-M.; Liu, Y.; Li, T.; and Borji, A. 2017. Structure-measure: A New Way to Evaluate Foreground Maps. In *Proceedings of the IEEE International Conference on Computer Vision*, 4548–4557.

Fan, D.-P.; Gong, C.; Cao, Y.; Ren, B.; Cheng, M.-M.; and Borji, A. 2018. Enhanced-alignment Measure for Binary Foreground Map Evaluation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 698–704. International Joint Conferences on Artificial Intelligence Organization.

Fan, D.-P.; Ji, G.-P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; and Shao, L. 2020. Pranet: Parallel Reverse Attention Network for Polyp Segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, 263–273. Springer.

Fang, Z.; Liu, Y.; Wu, H.; and Qin, J. 2024. VP-SAM: Taming Segment Anything Model for Video Polyp Segmentation via Disentanglement and Spatio-Temporal Side Network. In *European Conference on Computer Vision*, 367–383. Springer.

Gao, S.-H.; Cheng, M.-M.; Zhao, K.; Zhang, X.-Y.; Yang, M.-H.; and Torr, P. 2019. Res2net: A New Multi-scale Backbone Architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2): 652–662.

Hu, Q.; Liu, M.; Li, Q.; and Wang, Z. 2025. First-frame supervised video polyp segmentation via propagative and semantic dual-teacher network. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Hu, Q.; Yi, Z.; Zhou, Y.; Peng, F.; Liu, M.; Li, Q.; and Wang, Z. 2024. SALI: Short-Term Alignment and Long-Term Interaction Network for Colonoscopy Video Polyp Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 531–541. Springer.

Ji, G.-P.; Chou, Y.-C.; Fan, D.-P.; Chen, G.; Fu, H.; Jha, D.; and Shao, L. 2021. Progressively normalized self-attention network for video polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 142–152. Springer.

Ji, G.-P.; Xiao, G.; Chou, Y.-C.; Fan, D.-P.; Zhao, K.; Chen, G.; and Van Gool, L. 2022. Video Polyp Segmentation: A Deep Learning Perspective. *Machine Intelligence Research*, 19(6): 531–549.

Ji, Y.; Chen, Z.; Xie, E.; Hong, L.; Liu, X.; Liu, Z.; Lu, T.; Li, Z.; and Luo, P. 2023. Ddp: Diffusion model for dense visual prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21741–21752.

Jia, X.; Shen, Y.; Yang, J.; Song, R.; Zhang, W.; Meng, M. Q.-H.; Liao, J. C.; and Xing, L. 2024. PolypMixNet: Enhancing semi-supervised polyp segmentation with polyp-aware augmentation. *Computers in Biology and Medicine*, 170: 108006.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Lei, T.; Yang, Z.; Wang, X.; Wang, Y.; Wang, X.; Sun, F.; and Nandi, A. K. 2025. Adaptive Learning of High-Value Regions for Semi-Supervised Medical Image Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21450–21459.

Li, F.; Huang, Z.; Zhou, L.; Peng, H.; and Chu, Y. 2025a. Semi-supervised spatial-temporal calibration and semantic refinement network for video polyp segmentation. *Biomedical Signal Processing and Control*, 100: 107127.

Li, X.; Xu, J.; Zhang, Y.; Feng, R.; Zhao, R.-W.; Zhang, T.; Lu, X.; and Gao, S. 2022. TCCNet: Temporally Consistent Context-Free Network for Semi-supervised Video Polyp Segmentation. In *IJCAI*, 1109–1115.

Li, Y.; Zhu, Z.; Zhang, Y.; Chen, Y.; and Yu, Z. 2025b. Boost the Inference with Co-training: A Depth-guided Mutual Learning Framework for Semi-supervised Medical Polyp Segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 10394–10403.

Loshchilov, I.; and Hutter, F. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Lu, Y.; Yang, Y.; Xing, Z.; Wang, Q.; and Zhu, L. 2024. Diff-vps: Video polyp segmentation via a multi-task diffusion network with adversarial temporal reasoning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 165–175. Springer.

Mamonov, A. V.; Figueiredo, I. N.; Figueiredo, P. N.; and Tsai, Y.-H. R. 2014. Automated polyp detection in colon capsule endoscopy. *IEEE transactions on medical imaging*, 33(7): 1488–1502.

Margolin, R.; Zelnik-Manor, L.; and Tal, A. 2014. How to Evaluate Foreground Maps? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Pan, D.; Fan, J.; Zhu, J.; Li, L.; and Pan, X. 2025. Dual-calibrated Co-training Framework for Personalized Federated Semi-Supervised Medical Image Segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 6272–6280.

Ren, G.; Lazarou, M.; Yuan, J.; and Stathaki, T. 2023. Towards automated polyp segmentation using weakly-and semi-supervised learning and deformable transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4355–4364.

Sharma, V.; Kumar, A.; Jha, D.; Bhuyan, M. K.; Das, P. K.; and Bagci, U. 2024. ControlPolypNet: towards controlled colon polyp synthesis for improved polyp segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2325–2334.

Tajbakhsh, N.; Gurudu, S. R.; and Liang, J. 2015. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*, 35(2): 630–644.

Touvron, H.; Bojanowski, P.; Caron, M.; Cord, M.; El-Nouby, A.; Grave, E.; Izacard, G.; Joulin, A.; Synnaeve, G.; Verbeek, J.; et al. 2022. Resmlp: Feedforward networks for image classification with data-efficient training. *IEEE transactions on pattern analysis and machine intelligence*, 45(4): 5314–5321.

Wang, A.; Ma, H.; Bai, L.; Wu, Y.; Xu, M.; Zhang, Y.; Islam, M.; and Ren, H. 2025. PedSemiSeg: Pedagogy-inspired semi-supervised polyp segmentation. *Computerized Medical Imaging and Graphics*, 102591.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2022. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3): 415–424.

Wei, J.; Wang, S.; and Huang, Q. 2020. F^3Net: Fusion, Feedback and Focus for Salient Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12321–12328.

Wu, H.; Xie, W.; Lin, J.; and Guo, X. 2023. Acl-net: semi-supervised polyp segmentation via affinity contrastive learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 2812–2820.

Wu, S.; Wei, X.; Chen, X.; Ren, Y.; He, J.; and Pu, X. 2024. Cross-View Mutual Learning for Semi-Supervised Medical Image Segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 9253–9261.

Zhao, X.; Wu, Z.; Tan, S.; Fan, D.-J.; Li, Z.; Wan, X.; and Li, G. 2022. Semi-supervised spatial temporal attention network for video polyp segmentation. In *International conference on medical image computing and computer-assisted intervention*, 456–466. Springer.

Zhao, Y.; Zhou, T.; Gu, Y.; Zhou, Y.; Zhang, Y.; Wu, Y.; and Fu, H. 2025. WeakPolyp-SAM: Segment Anything Model-driven weakly-supervised polyp segmentation. *Knowledge-Based Systems*, 113701.

Zhong, J.; Wang, W.; Wu, H.; Wen, Z.; and Qin, J. 2020. PolypSeg: An efficient context-aware network for polyp segmentation from colonoscopy videos. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23*, 285–294. Springer.