SCDL: Synergistic Confidence-Dispersion Learning for Semi-Supervised Video Polyp Segmentation

Yuanqin He[1], Yuhua Zhang[1], Huisi Wu[1(✉)], and Jing Qin[2,3]

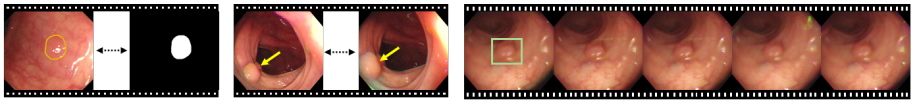
[1] College of Computer Science and Software Engineering, Shenzhen University
[2] Centre for Smart Health, The Hong Kong Polytechnic University
[3] CAS SIAT-PolyU Multi-modal Medical Molecular Imaging Joint Laboratory, The Hong Kong Polytechnic University
hswu@szu.edu.cn

Abstract. The early detection and prevention of colorectal cancer (CRC) is highly dependent on the accurate polyp segmentation from endoscopic videos. However, existing video polyp segmentation (VPS) methods are largely hampered by the scarcity of large-scale, high-quality pixel-level annotations. While some semi-supervised VPS models have been proposed, their performance is limited by the unreliability of pseudo-labels produced in the training process. In addition, inherent challenges of endoscopic videos, such as low contrast, significant inter-frame variations, and consecutive low-quality frames, further exacerbate the above issues. To address these shortcomings, we propose a unified semi-supervised video polyp segmentation framework, dubbed synergistic confidence–dispersion learning (SCDL), which aims to simultaneously improve pseudo-label reliability, reduce regional supervision bias, and alleviate ambiguities caused by low-contrast features. Our framework has three innovative components: (1) a new spectral convex optimization separation (SCOS) algorithm, which employs convex optimization in a confidence–dispersion space for robust pseudo-label selection, (2) contextual consistency perturbation (CCP), a novel augmentation strategy that reduces regional supervision bias by forcing semantic reconstruction from unreliable areas; and (3) a prototype-guided adaptive refinement (PAR) module that leverages prototype-driven aggregation and object-aware fusion to sharpen feature discrimination, suppress background clutter, and strengthen temporal coherence. Extensive experiments demonstrate that our model consistently achieves SOTA performance on the SUN-SEG dataset across all labeled-data ratios, and exhibits zero-shot generalization capability on the unseen CVC-ClinicDB dataset. Code are available at https://github.com/Yuanqin-He/SCDL.

Keywords: Semi-Supervised Video Polyp Segmentation · Pseudo-Label Reliability · Spectral Convex Optimization · Prototype-guided Adaptive Refinement

1 Introduction

Colorectal cancer (CRC) [41] is among the most prevalent malignancies worldwide and poses a substantial threat to human health and life. Since most CRCs develop from benign polyps, automatic and accurate segmentation of polyps in endoscopic videos plays a crucial role in early detection and prevention of CRC [2]. Although



(a) Low Contrast (b) Significant Variations (c) Consecutive Low-quality Frames

Fig. 1: Challenges of VPS including (a) low contrast between polyps and background, (b) significant variations between adjacent frames, and (c) consecutive low-quality frames.

fully-supervised deep-learning methods have achieved remarkable success in polyp segmentation in recent years, most of them are developed for static colonoscopy images [9, 38, 40, 51], and thus largely ignore temporal signals embedded in videos. To this end, these models cannot achieve satisfactory performance on colonoscopy videos. To sufficiently exploit temporal cues, some video polyp segmentation (VPS) models have been proposed [3, 15, 17, 45]. Despite some progress, these models are still limited by the scarcity of costly high-quality pixel-level annotations, making them difficult to deploy in clinical settings.

To circumvent this bottleneck, some semi-supervised learning (SSL) models have been proposed to strategically leverage abundant unlabeled data with limited labeled samples for polyp segmentation. Traditionally, SSL typically relies on mechanisms, such as self-training [7, 42, 55] and consistency regularization [18, 28, 43] to enhance generalization capability. Most recently proposed SSL methods [34, 37, 48] adopt pseudo-label supervision strategies, treating high-confidence predictions on unlabeled frames as reliable self-training targets. However, these strategies are highly dependent on the assumption that high confidence predictions imply high-quality labels. Such an assumption is, however, often violated in VPS from endoscopic videos.

The VPS has some unique challenges compared with natural video analysis tasks, such as (a) intrinsically low contrast between polyps and surrounding tissue, (b) large inter-frame variations, and (c) frequent consecutive low-quality frames (see Figure 1). These challenges exacerbate network overconfidence, leading to high-confidence but incorrect pseudo-labels, particularly in low-contrast regions that lack discriminative cues. Furthermore, temporal continuity amplifies this issue: the erroneous high-confidence labels propagate across frames, resulting in accumulated errors, unstable temporal consistency, and ultimately, degraded segmentation quality. Recently, some semi-supervised VPS models [21, 50] have been proposed to meet these challenges by modeling temporal interactions between labeled and unlabeled frames. However, they still cannot effectively suppress the propagation of noisy pseudo-labels, which are a major cause of degraded performance. Therefore, currently, a key challenge for semi-supervised VPS is how to leverage unlabeled frames robustly while explicitly mitigating pseudo-label overconfidence, so that learning is not misled by confident yet incorrect patterns, especially in low-contrast regions.

In this paper, we propose SCDL, a semi-supervised VPS framework that integrates PAR, SCOS, and CCP into a closed-loop learning process to jointly improve pseudo-label reliability, temporal consistency, and ambiguous-region learning. The three components are structurally and functionally interdependent: the prototype-guided adaptive refinement (PAR) module provides base spatio-temporal representations, whose predic-

tion statistics (confidence and dispersion) are mathematically exploited by the spectral convex optimization separation (SCOS) algorithm to derive sample-adaptive selection boundaries—without heuristic thresholds. Based on these boundaries, the contextual consistency perturbation (CCP) module masks reliable regions, forcing the network to reconstruct semantics from ambiguous areas. The resulting gradients in turn drive PAR to better aggregate temporal cues from low-contrast boundaries via implicit/explicit object-aware fusion and prototype-guided historical frame memory. Through this continuous forward data flow and backward gradient feedback, SCDL jointly mitigates pseudo-label overconfidence, regional supervision bias, and low-contrast feature ambiguity. Extensive experiments demonstrate that SCDL consistently achieves state-of-the-art performance on SUN-SEG across all labeled-data ratios, and exhibits robust zero-shot generalization on the unseen CVC-ClinicDB dataset. We summarize our main contributions as follows:

- We propose a novel semi-supervised VPS model, dubbed as SCDL; its core component is SCOS, an innovative algorithm to ensure robust pseudo-label selection by modeling predictions in a confidence-dispersionn feature space and maximizing separability via convex optimization.
- We further introduce a new mask-based augmentation strategy, called CCP, that alleviates regional supervision bias by compelling semantic reconstruction from unreliable regions via targeted masking of reliable predictions.
- We design a PAR module, which leverages historical object prototypes to adaptively refine features. This effectively resolves low-contrast ambiguity and strengthens spatio-temporal consistency in video sequences.
- Comprehensive experiments on the SUN-SEG dataset demonstrate that our model consistently outperforms SOTA methods. Furthermore, it exhibits exceptional zero-shot generalization capabilities on the unseen CVC-ClinicDB dataset.

2 Related work

2.1 Polyp Segmentation

Deep learning has driven substantial advances in image polyp segmentation (IPS). Early IPS relied on convolutional neural networks for localized feature extraction [8, 40], but their limited global view often produces blurred boundaries and suboptimal predictions. To mitigate this, Transformer-based or hybrid CNN–Transformer architectures were adopted to strengthen global context modeling [19, 32, 49], often combined with explicit boundary-aware constraints for edge refinement [8, 11, 52]. However, static IPS methods ignore temporal information, motivating the development of VPS to better handle dynamic scenes. Initial VPS approaches fused spatial and short-term temporal cues using hybrid 2D/3D convolutions [26], but their limited receptive field hampers capture of long-range temporal dependencies. This motivated attention-based designs [16], and methods such as PNS+ [17] that model global temporal relations across sequences. Despite their effectiveness, global attention mechanisms can be computationally heavy and vulnerable to noise; consequently, more efficient keyframe-guided schemes that use high-quality frames as anchors have been proposed [15, 46]. However, the scarcity of pixel-level annotations in clinical settings limits the scalability of supervised VPS

methods, making SSVPS a more practical choice. Consequently, leveraging abundant unlabeled frames becomes essential for easing the annotation bottleneck.

2.2 Semi-supervised Segmentation

The core challenge of semi-supervised segmentation (SSS) lies in effectively leveraging unlabeled images alongside limited labeled data. Current research in semi-supervised semantic segmentation mainly focuses on self-training [7, 42, 55] and consistency regularization [18, 28, 43]. While effective, these methods often suffer from distribution mismatches, leading to inaccurate supervisory signals. To address this, two main directions have been explored. One improves threshold-based pseudo-labeling, such as SoftMatch [5], which models confidence with a Gaussian distribution for adaptive thresholds, though such methods rely on restrictive assumptions. The other introduces additional confidence metrics, including ensembles [35] and auxiliary discriminators [20], but at higher computational cost. The core challenge remains generating accurate pixel-level pseudo-labels. Methods such as PS-MT [22] and UniMatch [47] improve consistency or separate features but rarely enhance pseudo-label quality directly. Recent approaches address this using AEL [13] to emphasize reliable predictions, DAW [30] to adapt thresholds based on labeled distributions, ESL [23] with multi-hot labels, and CorrMatch [29] to propagate labels via correlation maps, though biased supervision remains a challenge. Recent work in semi-supervised video polyp segmentation aims to make efficient use of limited annotations. TCCNet [21] and SSTAN [50] use temporal modules to enable feature interaction between labeled and unlabeled frames, supervising only the labeled positions. PSDNet [14] employs a dual-teacher framework to generate and fuse temporal pseudo-labels with stable semantic guidance. More recent methods, such as VPSentry [4] and SSL-CPCD [44], further explore video-specific temporal modeling and clustering-based pseudo-label refinement. Despite these advances, completely eliminating pseudo-label bias remains challenging. Unlike previous approaches, SCDL introduces additional metrics to address the low discriminability of confidence scores and explicitly mitigates supervision bias caused by pseudo-labels.

3 Method

3.1 Preliminaries

In semi-supervised video polyp segmentation (SSVPS), the training set comprises a small labeled set $D_l = \{(x_t^l, y_t^l)\}_{t=1}^{N_l}$ and a much larger unlabeled set $D_u = \{x_t^u\}_{t=1}^{N_u}$, where $N_l \ll N_u$. Each input $x_t \in \mathbb{R}^{3\times H\times W}$ has height H and width W. The pixel-wise label map $y_t \in \{0,1\}^{H\times W}$ encodes the binary polyp/background segmentation. For the segmentation network $\mathcal{F}$, the supervised loss is:

$$L_S = \frac{1}{B_L}\sum_{t=1}^{B_L} \mathrm{CE}\big(y_t^l,\, \mathcal{F}(A_\omega(x_t^l))\big), \tag{1}$$

where B_L is the labeled batch size, $\mathrm{CE}(\cdot,\cdot)$ denotes cross-entropy, and $A_\omega(\cdot)$ is weak augmentation. For unlabeled data we adopt a weak-to-strong consistency scheme [47].

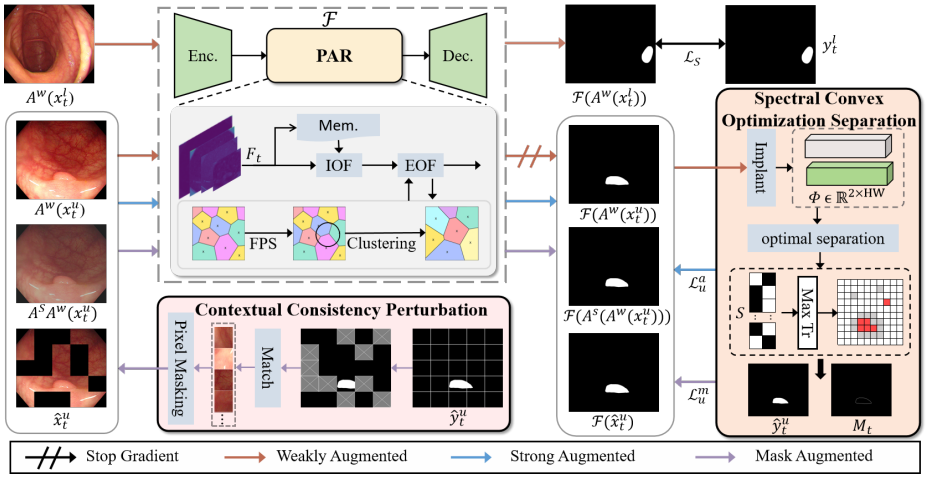


Fig. 2: Overview of our proposed SCDL. PAR addresses low-contrast ambiguity and enhances temporal consistency to improve feature representations, thereby facilitating more reliable pseudo-label generation. SCOS embeds residual dispersion into a confidence–distribution space and, through spectral relaxation, formulates the construction of M_t as a convex optimization problem to derive sample-adaptive selection thresholds. CCP provides auxiliary supervision by encouraging the network to reconstruct semantic context in regions with unreliable pseudo-labels, compensating for the loss of reliable supervisory signals.

Given an unlabeled batch of size B_U, the unsupervised loss is:

$$L_u^a = \frac{1}{B_U} \sum_{t=1}^{B_U} M_t \odot \mathrm{CE}\big(\hat{y}_t^u,\ \mathcal{F}(A_s(A_\omega(x_t^u)))\big), \tag{2}$$

where $\hat{y}_t^u$ is the pseudo-label produced from $\mathcal{F}(A_\omega(x_t^u))$, $A_s(\cdot)$ is strong augmentation, and M_t is the pseudo-label indicator map. In the baseline, M_t is obtained by simple confidence thresholding:

$$M_t = \mathbb{I}\big(\max(\mathcal{F}(A_\omega(x_t^u))) > \tau\big), \tag{3}$$

with $\mathbb{I}(\cdot)$ the indicator function and τ the confidence threshold. This fixed-threshold strategy, however, leads to cognitive bias and supervision gaps.

To mitigate these problems, we propose SCDL (see Figure 2). Instead of functioning as a sequence of independent stages, the three components in our framework are designed to form an interdependent learning process. First, the PAR module resolves low-contrast ambiguity and enhances temporal consistency, producing stronger features that obtain reliable pseudo-labeling. Next, SCOS incorporates residual dispersion into a confidence–distribution space and, via spectral relaxation, casts the construction of M_t as a convex optimization to obtain sample-adaptive selection boundaries. Finally, CCP supplies additional supervisory signals by compelling the network to reconstruct semantic context from regions with unreliable supervision, thereby compensating for

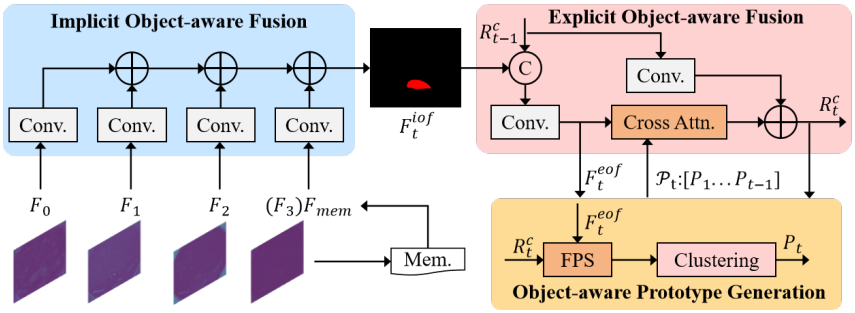


Fig. 3: Structure of the PAR. The PAR fuses high-resolution features with memory-derived prototypes to produce temporally-consistent, object-aware features for reliable pseudo-labeling.

the error guidance in challenging areas. This targeted perturbation provides augmented supervisory signals during training, which iteratively refines the feature representations fed back into the PAR module, thereby establishing a naturally synergistic training loop.

3.2 Prototype-guided Adaptive Refinement

To fundamentally address low-contrast ambiguity and improve temporal consistency, we design a prototype-guided adaptive refinement (PAR) module, as shown in Figure 3, which lays a stronger foundation for subsequent pseudo-label selection and mask refinement. The PAR module extracts fine-grained, high-resolution polyp features from the current frame via Implicit Object-aware Fusion (IOF), mitigating the low-contrast effect of the surrounding mucosa. Concurrently, it employs Object-aware Prototype Generation (OPG) to abstract and store discriminative polyp prototypes from historical frames, and leverages Explicit Object-aware Fusion (EOF) to integrate these prototypes into the current-frame features, thereby strengthening polyp localization across space and time.

Early encoder layers capture detailed textures, while deeper layers encode semantics. Since early features are only implicitly object-aware, we introduce an IOF module to fuse them with memory-conditioned features for better representation. For ResNet, four feature maps from the image encoder are extracted for each frame, i.e., $L = 4$. We have F_t^0, F_t^1, F_t^2 and F_t^3 for the current frame x_t. We denote the memory-conditioned feature as F_t^{mem}, encoded by the memory-attention module on F_t^3. To propagate temporal priors, we maintain a FIFO memory bank ($T = 8$) storing past F^3 features. Following [27], F_t^{mem} is generated via cross-attention, where the current F_t^3 acts as the *Query* and memory features serve as *Keys* and *Values*. This effectively retrieves relevant temporal context, ensuring robust inter-frame consistency. The high-resolution features F_t^0, F_t^1 and F_t^2 are fused with F_t^{mem} via compression modules and point-wise addition to create a refined feature representation $F_t^{iof} \in \mathbb{R}^{c_0 \times h_0 \times w_0}$, where a compression module $C(\cdot)$ consists of two convolutional layers, followed by an upsampling layer, to create compact representations. This process is given by:

$$F_t^{iof} = C_0(F_t^0) + C_1(F_t^1) + C_2(F_t^2) + C_3(F_t^{mem}). \tag{4}$$

After obtaining F_t^{iof}, we further refine it by EOF, which exploits explicit object-aware information from the current frame and previous frames, through employing object prototypes. We have three steps to fuse informative features. First, feature F_t^{iof} is concatenated with the previous contextual representation R_{t-1}^c, followed by a convolutional layer to reduce the channel dimension back to c_0, yielding:

$$F_t^{eof} = \text{Conv}([\, F_t^{iof}; R_{t-1}^c \,]). \tag{5}$$

This step effectively merges historical contextual priors with the current frame's representation, enabling inter-frame consistency propagation. Second, F_t^{eof} goes through a cross-attention layer. Prototypes generated from previous frames, representing discriminative clustered features, serve as informative priors to help distinguish the polyp from its background. A cross-attention mechanism takes F_t^{eof} as a query, and leverages these prototypes as keys and values, effectively exploiting the information within the object prototypes to refine F_t^{eof}. Formally, we update F_t^{eof} by conducting cross-attention with prototypes $\mathcal{P}_t = \{P_0, P_1, \ldots, P_{t-1}\}$ from previous frames, given by:

$$F_t^{attn} = \text{Attn}(F_t^{eof}, \mathcal{P}_t, \mathcal{P}_t). \tag{6}$$

Finally, the attention-refined feature F_t^{attn} is combined with the previous contextual representation R_{t-1}^c. Specifically, R_{t-1}^c is first processed through a convolutional layer and then fused with F_t^{attn} through point-wise addition. This yields the updated contextual representation R_t^c for the current frame:

$$R_t^c = F_t^{attn} + \text{Conv}(R_{t-1}^c). \tag{7}$$

To effectively capture representative polyp object appearances inside the predicted mask, we design the OPG. OPG first applies Farthest Point Sampling (FPS) [24] to select k well-distributed points within the predicted mask as initial cluster centers. A single iteration of k-means is then performed to assign mask pixels to these centers, and each cluster prototype is computed as the mean of the spatial features of pixels in that cluster. Formally, the prototype set for frame x_t is defined as:

$$P_t = \{P_t^i \mid 1 \le i \le k\} = F_p(F_t^{eof}, R_t^c), \tag{8}$$

where $F_p(\cdot)$ denotes the prototype extraction operator that takes high-resolution frame features F_t^{eof} and the mask context R_t^c as input. The prototypes P_t are concatenated and appended to a memory bank for use as object priors in subsequent frames, enabling temporally consistent prototype guidance.

3.3 Theoretical Analysis of Residual Dispersion

Deep networks often exhibit overconfidence on unlabeled data, making maximum confidence alone unreliable for pseudo-label selection. Following the entropy-minimization principle, we theoretically show that reliable predictions should exhibit both high maximum confidence and sufficient residual dispersion.

Let $p_n \in \mathbb{R}^C$ denote the predicted class-probability vector for pixel n in an unlabeled image x_t^u, and let $p_n(k^\star) = \max_k p_n(k)$ denote the maximum confidence. Under entropy minimization, the ideal prediction is an approximately unimodal distribution $\tilde{q} = [\tilde{q}_1, \ldots, \tilde{q}_C]$ with $\tilde{q}_{k^\star} = 1 - (C-1)\varepsilon$ and $\tilde{q}_{k \neq k^\star} = \varepsilon$, where $\varepsilon \to 0$. Using cross-entropy decomposition, we have

$$CE(p_n, \tilde{q}) = -\tilde{q}_{k^\star} \log p_n(k^\star) - \varepsilon \sum_{k \neq k^\star} \log p_n(k). \tag{9}$$

For non-maximum probabilities $p_n(k)$ $(k \neq k^\star)$, decompose them as $p_n(k) = \bar{\mu} + \Delta_k$, where $\bar{\mu} = \frac{1-p_n(k^\star)}{C-1}$ and $\sum_{k \neq k^\star} \Delta_k = 0$. A second-order Taylor expansion of $\log p_n(k)$ around $\bar{\mu}$ yields:

$$\log p_n(k) = \log \bar{\mu} + \frac{\Delta_k}{\bar{\mu}} - \frac{\Delta_k^2}{2\bar{\mu}^2} + O(\Delta_k^3), \quad k \neq k^\star. \tag{10}$$

Substituting this expansion into the cross-entropy and simplifying gives:

$$CE(p_n, \tilde{q}) \approx -\log p_n(k^\star) - \varepsilon \frac{(C-1)^3}{2(1-p_n(k^\star))^2} v_n, \tag{11}$$

where the residual dispersion v_n is defined as: $v_n \triangleq -\frac{1}{C-1} \sum_{k \neq k^\star} \Delta_k^2$. As shown in Eq. (11), reliable predictions require both a large $p_n(k^\star)$ and a large v_n. In particular, as $p_n(k^\star) \to 1$, the coefficient of v_n grows rapidly, making v_n a key indicator for identifying reliable predictions under overconfident regimes.

3.4 Spectral Convex Optimization Separation

Sec. 3.3 demonstrates that prediction reliability depends nonlinearly on both $p_n(k^\star)$ and v_n. However, threshold-based methods fail to model this nonlinear complementarity. To address this, we formulate pseudo-label selection as a convex optimization problem via spectral relaxation. For each pixel, we construct an attribute vector $h_n = [p_n(k^\star),\ v_n]^T$ that captures both confidence and dispersion. Our goal is to select predictions with high reliability by maximizing their spectral separability:

$$\max_S \ \mathrm{Tr}(S^T \Phi^T \Phi S), \quad \text{s.t. } S \in \{0,1\}^{HW \times 2}, \tag{12}$$

where $\Phi = [h_1, h_2, \ldots, h_{HW}]$ is the feature matrix of x_t^u, and $\mathrm{Tr}(\cdot)$ denotes the trace operator. To relax the non-convex combinatorial constraint on S, we apply spectral relaxation [25]. According to the Ky Fan theorem [31], the optimal relaxed solution is approximated as:

$$S_{n,c}^* = \Delta\Big(\arg\max_{i \in \{1,2\}} |u_i(n)|,\ c \Big), \tag{13}$$

where u_1, u_2 are the eigenvectors of $\Phi^T \Phi$, and $\Delta(\cdot,\cdot)$ denotes the Kronecker delta.

Without loss of generality, let $c = 2$ denote the subset of predictions with high confidence and dispersion. We define the projection $Z = \Phi S^*_{:,2}$ as the reliable component. A Gaussian weighting function is then applied to construct smooth loss weights:

$$\omega_n = \prod_c \exp\Big(- \frac{(h_n(c) - \mu_c)^2}{\alpha \sigma_c^2} \Big), \tag{14}$$

where μ_c and σ_c are the mean and variance of the reliable subset Z, and $\alpha = 4$ is a smoothing factor. Predictions with both high confidence and dispersion are assigned a weight of 1. Thus, the unsupervised indicator map M_t for image x_t^u is defined as:

$$M_t(n) = \begin{cases} 1, & \text{if } (h_n(c) - \mu_c) > 0,\ \forall c \in \{1, 2\}, \\ \omega_n, & \text{otherwise.} \end{cases} \tag{15}$$

3.5 Contextual Consistency Perturbation

While semi-supervised training significantly benefits from high-quality pseudo-labels, consistently overlooking regions with low prediction confidence can restrict the model's ability to learn robust features, potentially entrenching it in simplistic decision boundaries. To counter this limitation and enhance the model's capacity to resolve ambiguous areas, we introduce CCP. The CCP strategically masks high-confidence predictions, compelling the model to infer missing information from the surrounding, more challenging, and less reliable regions. This process implicitly strengthens the mutual information between these difficult areas and their corresponding pseudo-labels.

The transition from the soft weights ω_n in Eq. (15) to the binary indicator $M_t = 1$ in Eq. (16) serves distinct functional roles in the training objective. At the loss optimization level in Eq. (19), the soft weights ω_n are preserved to smoothly scale the gradients for uncertain predictions, thereby preventing noise accumulation. Conversely, at the input augmentation level, CCP applies perturbations directly to the spatial pixels. Applying continuous soft masks to input frames would alter the natural image statistics and introduce artificial boundary textures, which could interfere with the convolutional receptive fields. Therefore, the strict binary condition ensures that the structural integrity of the unmasked regions remains realistic, while forcing the model to infer the entirely masked areas based on surrounding context. Utilizing the reliable pixel indicator M_t, where $M_t = 1$ denotes reliable predictions, we define a patch-based perturbation mask $R_t \in \{0, 1\}^{H \times W}$ generated as follows:

$$R_t(x, y) = \mathbb{I}\big(M_t(x, y) = 1\big) \cdot \mathbb{I}\big(U\big(\big\lfloor \tfrac{x}{s} \big\rfloor, \big\lfloor \tfrac{y}{s} \big\rfloor\big) \geq \theta\big), \tag{16}$$

where s denotes the side length of the square patch, $\theta \in [0, 1]$ is the masking rate controlling the sparsity of the perturbation, and $U(\cdot, \cdot)$ represents a uniform random variable sampled from $[0, 1]$ for each patch. This formulation ensures that only regions identified as reliable by M_t are candidates for masking, and then a random subset of these reliable patches is selected based on θ.

The perturbed image $\tilde{x}_t^u$ is obtained by applying R_t to the input image x_t^u:

$$\tilde{x}_t^u = x_t^u \odot R_t, \tag{17}$$

Table 1: Quantitative comparison with different SOTA methods on SUN-SEG test sets. The fractions denote the percentage of labeled data used for training and the best results are highlighted in bold.

Model	Backbone	Task	1/16				1/8				1/4				1/2			
			Easy		Hard		Easy		Hard		Easy		Hard		Easy		Hard	
			S_α	Dice	S_α	Dice	S_α	Dice	S_α	Dice	S_α	Dice	S_α	Dice	S_α	Dice	S_α	Dice
Corrmatch	Res2Net-50	NIS	80.92	71.25	79.85	69.98	80.85	72.68	80.82	72.45	83.95	75.58	84.52	76.02	85.27	77.98	84.95	77.42
ACL-Net	Res2Net-50	IPS	81.69	73.43	81.38	71.76	83.21	75.13	81.53	73.21	84.93	76.41	86.04	77.32	85.83	78.61	86.08	78.93
PedSemiSeg	Res2Net-50	IPS	81.42	71.79	80.73	70.84	81.57	73.23	81.93	73.74	84.41	76.02	85.24	76.74	86.92	78.64	85.63	78.35
IFR	Res2Net-50	NVS	81.05	71.38	80.12	70.25	81.23	72.95	81.05	72.68	84.07	75.42	84.73	76.23	85.41	78.12	85.18	77.65
TDC	Res2Net-50	NVS	81.39	71.72	80.51	70.64	81.48	73.20	81.29	72.92	84.32	75.81	85.09	76.59	85.68	78.46	85.43	77.90
TCCNet	Res2Net-50	VPS	80.63	74.18	79.87	73.42	82.95	77.17	81.43	76.62	84.15	78.85	84.78	79.02	85.37	80.25	85.64	80.83
SSTAN	Res2Net-50	VPS	81.93	75.27	81.37	74.51	84.15	78.22	82.89	77.70	85.63	79.96	86.18	80.14	86.91	81.35	87.16	81.96
PSDNet	PVTv2-b2	VPS	87.09	83.07	86.76	81.29	88.01	83.89	87.52	82.54	88.39	84.98	87.64	83.36	89.17	85.71	88.10	83.77
Ours	Res2Net-50	VPS	87.52	83.48	87.17	81.66	88.48	84.34	88.05	82.98	88.90	85.55	88.13	83.94	89.62	86.20	88.62	84.28
Ours	PVTv2-b2	VPS	**89.37**	**85.91**	**88.59**	**83.76**	**89.83**	**86.74**	**89.04**	**84.62**	**90.15**	**87.32**	**89.31**	**85.44**	**90.56**	**87.82**	**89.51**	**86.09**

where $\odot$ denotes element-wise multiplication. A potential concern with masking highly confident regions is that the network might struggle to reconstruct the missing semantics, leading to unconstrained predictions. To address this, CCP is designed to leverage two existing priors: the explicit temporal context preserved in the PAR's memory bank (F_t^{mem}), and the implicit spatial correlations from the unmasked surrounding mucosa. By forcing the model to infer the masked reliable regions, the network must propagate these valid spatiotemporal priors into the neighboring ambiguous areas. This explicitly penalizes under-fitting in difficult regions and bridges the semantic gap between reliable and unreliable pixels, improving its ability to handle challenging polyp segmentation cases. The unsupervised loss for the masked inputs is formulated as:

$$L_u^m = \frac{1}{B_U} \sum_{t=1}^{B_U} M_t \odot \mathrm{CE}\big(\hat{y}_t^u, \mathcal{F}(\tilde{x}_t^u)\big). \tag{18}$$

The final unsupervised loss term combines both the standard and masked losses:

$$L_U = \lambda_1 L_u^a + \lambda_2 L_u^m, \tag{19}$$

where λ_1, λ_2 are loss weights (set to $[0.5, 0.5]$ by default).

4 Experiments

4.1 Experiments Setup

Datasets. We conduct experiments on SUN-SEG [17], the largest video polyp segmentation dataset. Specifically, the SUN-SEG dataset comprises 100 positive cases with a total of 49,136 polyp frames. Among them, 40% of the data are used to construct the SUN-SEG-Train set, which contains 112 sequences (19,544 frames). The remaining data are divided into two test subsets according to difficulty levels: SUN-SEG-Easy, including 119 sequences (17,070 frames), and SUN-SEG-Hard, consisting of 54 sequences (12,522 frames). Each subset is further split into "seen" and "unseen" partitions, where

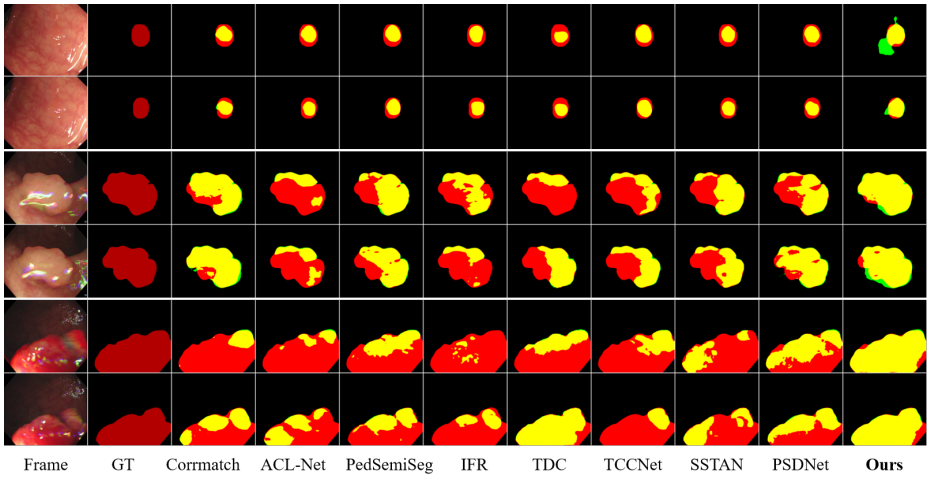


Fig. 4: Visual comparison with SOTA methods on SUN-SEG test sets. Red, green, and yellow represent ground truth, prediction, and their overlapping regions, respectively.

Table 2: Ablation study with performance–efficiency comparison on the SUN-SEG-Hard set under a 1/2 labeled-data partition.

Method	Dice (%)	GFlops	Param. (M)
Baseline	80.21	92.82	41.38
w/ PAR	82.36	107.29 (**+14.47**)	42.43 (**+1.05**)
w/o PAR	83.24	94.90 (**+2.08**)	41.38 (**+0**)
w/ All	84.28	109.37 (**+16.55**)	42.43 (**+1.05**)

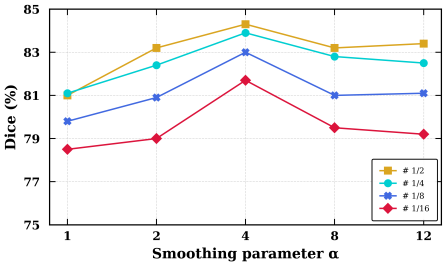


Fig. 5: Ablation study of smoothing parameter α.

"unseen" samples contain scenes absent from training, enabling more rigorous generalization evaluation. In addition, CVC-ClinicDB is a simpler polyp dataset with a lower resolution of 288 × 384, including 612 frames from 29 video sequences.

Evaluation Metrics. For comprehensive comparison, we employ two metrics to evaluate the segmentation results, including Dice, and structure-measure (S_α) [10], following previous studies [15, 17]. These metrics provide a balanced evaluation of pixel-level accuracy and structural integrity.

Implementation Details. To ensure comprehensive evaluation, we adopt two representative backbones: Res2Net-50 [12] and PVTv2-b2 [36]. All ablation studies and most comparisons are conducted with Res2Net-50 to maintain consistency with prior video polyp segmentation works. PVTv2-b2 is used exclusively in the comparison table to enable direct evaluation against the state-of-the-art method PSDNet [14]. DeepLabv3+ [6] is consistently employed as the decoder for both backbones. All training is performed on a single NVIDIA RTX 3090 GPU. We optimize with SGD using a batch size of 8, weight decay 1×10^{-4}, and an initial learning rate of 1×10^{-3}. During training, inputs are

Table 3: Impact of different loss weights on the SUN-SEG-Hard dataset. The best results are highlighted in bold.

Loss Weights		Metrics (Dice)	
λ_1	λ_2	1/16	1/2
0.50	0.50	**81.66**	**84.28**
0.75	0.25	80.73	83.13
0.25	0.75	80.61	82.81
0.30	0.30	81.36	83.92
0.70	0.70	80.92	83.47
1.00	1.00	80.15	82.31

Table 4: Zero-shot cross-dataset generalization on CVC-ClinicDB. Models are trained on the SUN-SEG dataset (1/2 labeled partition).

Model	S_α	Dice
Corrmatch	90.45	87.62
ACL-Net	88.67	86.12
PedSemiSeg	88.75	86.05
IFR	88.84	85.95
TDC	90.13	87.45
TCCNet	90.72	88.06
SSTAN	90.58	87.82
PSDNet	91.84	89.67
Ours	**94.12**	**91.53**

randomly cropped to 352×352 and the network is trained for 60 epochs. To accelerate decoder adaptation, the decoder's learning rate is set to be ten times higher than that of the backbone. To achieve an annotation ratio of $1/x$, our design utilizes x-frame video sequences where the label information is provided solely for the initial frame.

4.2 Comparison with State-of-the-art Methods

To comprehensively evaluate the strengths of our proposed method, we conduct a comparative analysis against eight SOTA approaches on the SUN-SEG. These competing methods are categorized into four groups: (1) Natural Image Segmentation (NIS): Corrmatch [29]; (2) Image Polyp Segmentation (IPS), including ACL-Net [39] and PedSemiSeg [33]; (3) Natural Video Segmentation (NVS), including IFR [53] and TDC [54]; and (4) Video Polyp Segmentation (VPS), which includes TCCNet [21], SSTAN [50], and PSDNet [14]. To ensure a fair comparison, all controllable training parameters were set to identical values across all methods.

Comparison with SOTA Methods. Table 1 presents a comprehensive comparison of our method with SOTA approaches on the SUN-SEG dataset under the semi-supervised setting. Under identical experimental protocols, our method achieves consistently superior performance across different metrics and label ratios, including 1/16, 1/8, 1/4, and 1/2. Notably, the advantage remains clear even under the most annotation-scarce setting, demonstrating that SCDL can effectively exploit unlabeled video frames and maintain strong segmentation performance with limited supervision.

Visual Comparison with SOTA. In Figure 4, we present qualitative comparisons with SOTA methods under three challenging scenarios: (1) similar foreground and background, (2) low-quality frames, and (3) large size variations. Compared with existing methods, SCDL produces more complete masks in low-contrast regions, maintains more coherent predictions on degraded frames, and better preserves object contours across different polyp sizes. These results indicate that SCDL is more robust to appearance ambiguity, frame quality degradation, and scale variation, further confirming its effectiveness in challenging video polyp segmentation scenarios.

Table 5: Ablation study of different components on the SUN-SEG-Hard dataset.

Baseline	PAR	SCOS	CCP	1/16 Dice ↑	1/2 Dice ↑
✓				78.36	80.21
✓	✓			79.83	82.36
✓	✓	✓		81.13	83.72
✓		✓	✓	80.32	83.24
✓	✓	✓	✓	**81.66**	**84.28**

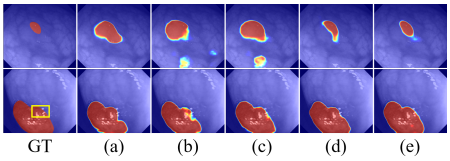


Fig. 6: Visualization of module ablation on the SUN-SEG-Hard dataset under a 1/2 labeled-data partition. (a) baseline. (b) only w/ PAR. (c) w/o CCP. (d) w/o PAR. (e) w/ all.

Table 6: Impact of k-means distance metric and prototype count on PAR on the SUN-SEG-Hard dataset under a 1/2 labeled-data partition.

Distance Metric	Dice ↑	S_α ↑
Euclidean	81.74	86.24
Cosine	**84.28**	**88.62**
# Prototypes (k)	Dice ↑	S_α ↑
3	82.06	86.35
5	**84.28**	**88.62**
7	82.43	86.74

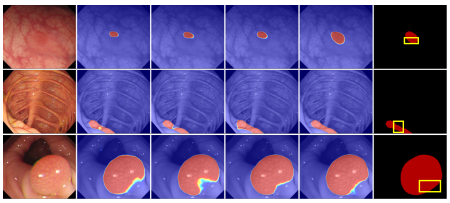


Fig. 7: Visualization of PAR module ablation on the SUN-SEG-Hard dataset under a 1/2 labeled-data partition. (a) baseline. (b) only w/ IOF. (c) w/o OPG. (d) w/ all.

4.3 Ablation Studies

Ablation Studies for Different Modules. Our ablation study (Table 5, Figure 6) verifies the effectiveness of our design. As illustrated in Figure 6, the baseline often suffers from severe background noise and incomplete target segmentation. By introducing PAR, the model achieves better target localization. The addition of SCOS and CCP further eliminates false positives in low-contrast regions and sharpens the polyp boundaries, demonstrating the synergy of improved pseudo-label reliability, bias correction, and low-contrast feature refinement. Furthermore, the t-SNE visualization in Figure 8 confirms that our approach yields more compact and discriminative feature clusters compared to versions without individual modules. Finally, Table 8 and Figure 7 demonstrate positive contributions from each PAR component. The IOF module improves single-frame features by mitigating low-contrast ambiguity. Subsequently, EOF boosts performance by integrating temporal context into the current-frame features. The complete OPG module provides the most stable and precise segmentation by leveraging discriminative historical prototypes, effectively resolving temporal inconsistency.

Impact of CCP. Table 7 shows that credible masking achieves the best performance when $\theta = 0.5$ and the patch size is set to $s = 32$. In contrast, increasing the patch size to $s = 64$ leads to a clear drop in Dice score, likely because overly large patches remove excessive local context and disrupt fine-grained boundary cues. This suggests that a moderate masking scale is important for preserving useful contextual information while still encouraging the model to learn from ambiguous regions.

Effect of hyperparameters. Table 6 presents ablation results on SUN-SEG-Hard with a 1/2 labeled partition, evaluating the k-means design in our OPG module. Using cosine

Table 7: Ablation study with different parameters of CCP on the SUN-SEG-Hard dataset under a 1/2 labeled-data partition.

Patch Size s	**8**	**16**	**32**	**64**
$\theta = 0.3$	82.73	82.55	82.41	81.55
$\theta = 0.5$	82.84	83.23	**84.28**	83.59
$\theta = 0.7$	82.25	83.17	83.92	82.86
$\theta = 0.9$	81.52	81.64	81.63	81.45

Table 8: Ablation study of PAR on the SUN-SEG-Hard dataset under a 1/2 labeled-data partition.

IOF	EOF	OPG	Dice ↑	S_α ↑
			83.24	87.29
✓			83.64	87.73
✓	✓		83.92	88.10
✓	✓	✓	**84.28**	**88.62**

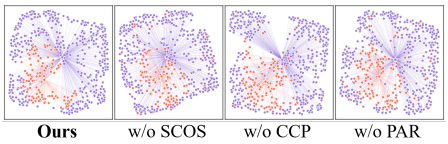


Fig. 8: T-SNE visualization of different module. Purple represents the background, and red represents polyps.

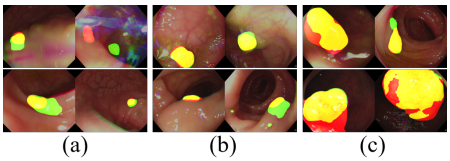


Fig. 9: Failure cases. Red, green and yellow represent the GT, prediction and their overlapping regions, respectively.

distance consistently outperforms Euclidean distance, suggesting that angular similarity is more effective for grouping polyp features. The number of prototypes also affects performance: too few ($k = 3$) or too many ($k = 7$) degrade results, while $k = 5$ achieves the best balance between capturing structural details and maintaining discriminability. Figure 5 shows that an excessively large smoothing factor α leads to a decline in performance, as it overemphasizes unreliable predictions. Accordingly, we set $\alpha = 4$ as the default value to achieve optimal overall performance.

Different loss weights. In Table 3, we investigate the impact of different loss weightings on segmentation accuracy. The imbalance between consistency loss and masking loss affects model performance. Moreover, assigning excessively high or low loss weights to the unlabeled data also leads to a degradation in performance. The results indicate that the optimal performance is achieved when $[\lambda_1, \lambda_2]$ are set to [0.5, 0.5].

Complexity Analysis. For clinical deployment, inference latency is a critical consideration. As shown in Table 2, the SCOS and CCP modules act only as training-time supervision mechanisms and are discarded during inference, introducing no additional cost to the final model. Although the PAR module brings a structural overhead of +14.47 GFLOPs, this increase is relatively modest compared with the overall complexity of the baseline. These results indicate that SCDL improves segmentation performance while maintaining practical inference efficiency.

Cross-dataset Generalization. To evaluate whether the learned features generalize beyond the specific data distribution of the training set, we conduct a zero-shot cross-dataset evaluation. Specifically, models trained solely on the SUN-SEG dataset are directly evaluated on the unseen CVC-ClinicDB [1] without any further fine-tuning. As reported in Table 4, SCDL maintains consistent relative improvements under this zero-

shot setting, indicating that the framework effectively captures robust, domain-agnostic polyp characteristics rather than overfitting to the training domain.

4.4 Discussions and Limitations.

According to the above ablation studies and comparisons, our proposed SCDL framework exhibits a strong ability to resolve low-contrast ambiguity and generate reliable pseudo-labels for semi-supervised video polyp segmentation. The PAR module effectively refines polyp features by modeling prototype-guided spatiotemporal dependency and incorporating historical prototypes for temporal consistency, while the SCOS and CCP components further enhance segmentation accuracy through reliability-aware pseudo-label selection and contextual consistency perturbation. However, our method still has certain limitations. As illustrated in Figure 9, factors such as blurred fields of view affected by specular highlights (a), extremely small and low-contrast polyps (b), and extreme shape variations (c) may still challenge the method and warrant further investigation. Specifically, for exceptionally small targets under low contrast, the spatial resolution loss introduced by standard pooling layers can lead to severe feature entanglement with the surrounding mucosa, making temporal propagation difficult.

5 Conclusion

In conclusion, this paper introduces the SCDL framework, a unified SSVPS approach designed to overcome core challenges such as unreliable pseudo-labels, regional supervision bias, and low-contrast feature ambiguities in endoscopic videos. SCDL integrates three key innovative components: Spectral Convex Optimization Separation (SCOS) for robust pseudo-label selection using convex optimization; Contextual Consistency Perturbation (CCP) to reduce regional supervision bias by enforcing semantic reconstruction from unreliable areas; and a Prototype-guided Adaptive Refinement (PAR) module to sharpen feature discrimination and strengthen temporal coherence using discriminative prototypes. Extensive experiments confirm that SCDL not only achieves state-of-the-art performance on the SUN-SEG dataset but also exhibits zero-shot generalization on the unseen CVC-ClinicDB dataset.

Acknowledgements

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 grant of General Research Fund under Hong Kong Research Grants Council (project no 15217825), and the grant under the Collaborative Research with World-leading Research Groups in The Hong Kong Polytechnic University (project no G-SACF).

References

1. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized medical imaging and graphics **43**, 99–111 (2015)
2. Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R.L., Soerjomataram, I., Jemal, A.: 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 (2024)
3. Chen, G., Yang, J., Pu, X., Ji, G.P., Xiong, H., Pan, Y., Cui, H., Xia, Y.: Mast: Video polyp segmentation with a mixture-attention siamese transformer. arXiv preprint arXiv:2401.12439 (2024)
4. Chen, G., Luo, X., Wu, H., Qin, J.: Vpsentry: Semi-supervised video polyp segmentation via sentry-guided long-term prototype fusion with correlation dynamic propagation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 2850–2858 (2026)
5. Chen, H., Tao, R., Fan, Y., Wang, Y., Wang, J., Schiele, B., Xie, X., Raj, B., Savvides, M.: Softmatch: Addressing the quantity-quality trade-off in semi-supervised learning. arXiv preprint arXiv:2301.10921 (2023)
6. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818 (2018)
7. Chen, S., Tan, X., Wang, B., Hu, X.: Reverse attention for salient object detection. In: Proceedings of the European conference on computer vision (ECCV). pp. 234–250 (2018)
8. Cheng, M., Kong, Z., Song, G., Tian, Y., Liang, Y., Chen, J.: Learnable oriented-derivative network for polyp segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. pp. 720–730. Springer (2021)
9. 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)
10. Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: A new way to evaluate foreground maps. In: Proceedings of the IEEE international conference on computer vision. pp. 4548–4557 (2017)
11. Fan, D.P., Ji, G.P., Sun, G., Cheng, M.M., Shen, J., Shao, L.: Camouflaged object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2777–2787 (2020)
12. Gao, S.H., Cheng, M.M., Zhao, K., Zhang, X.Y., Yang, M.H., Torr, P.: Res2net: A new multi-scale backbone architecture. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(2), 652–662 (2021). https://doi.org/10.1109/TPAMI.2019.2938758
13. Hu, H., Wei, F., Hu, H., Ye, Q., Cui, J., Wang, L.: Semi-supervised semantic segmentation via adaptive equalization learning. Advances in Neural Information Processing Systems **34**, 22106–22118 (2021)
14. Hu, Q., Liu, M., Li, Q., Wang, Z.: 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). pp. 1–5. IEEE (2025)
15. Hu, Q., Yi, Z., Zhou, Y., Peng, F., Liu, M., Li, Q., Wang, Z.: 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. pp. 531–541. Springer (2024)
16. 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: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 142–152. Springer (2021)

17. 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)
18. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:1610.02242 (2016)
19. Li, S., Sui, X., Luo, X., Xu, X., Liu, Y., Goh, R.: Medical image segmentation using squeeze-and-expansion transformers. arXiv preprint arXiv:2105.09511 (2021)
20. Li, S., Jin, W., Wang, Z., Wu, F., Liu, Z., Tan, C., Li, S.Z.: Semireward: A general reward model for semi-supervised learning. arXiv preprint arXiv:2310.03013 (2023)
21. Li, X., Xu, J., Zhang, Y., Feng, R., Zhao, R.W., Zhang, T., Lu, X., Gao, S.: Tccnet: Temporally consistent context-free network for semi-supervised video polyp segmentation. In: IJCAI. pp. 1109–1115 (2022)
22. Liu, Y., Tian, Y., Chen, Y., Liu, F., Belagiannis, V., Carneiro, G.: Perturbed and strict mean teachers for semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4258–4267 (2022)
23. Ma, J., Wang, C., Liu, Y., Lin, L., Li, G.: Enhanced soft label for semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1185–1195 (2023)
24. Moenning, C., Dodgson, N.A.: Fast marching farthest point sampling. Tech. rep., University of Cambridge, Computer Laboratory (2003)
25. Olsson, C., Eriksson, A.P., Kahl, F.: Improved spectral relaxation methods for binary quadratic optimization problems. Computer Vision and Image Understanding **112**(1), 3–13 (2008)
26. Puyal, J.G.B., Bhatia, K.K., Brandao, P., Ahmad, O.F., Toth, D., Kader, R., Lovat, L., Mountney, P., Stoyanov, D.: Endoscopic polyp segmentation using a hybrid 2d/3d cnn. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23. pp. 295–305. Springer (2020)
27. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
28. Sajjadi, M., Javanmardi, M., Tasdizen, T.: Regularization with stochastic transformations and perturbations for deep semi-supervised learning. Advances in neural information processing systems **29** (2016)
29. Sun, B., Yang, Y., Zhang, L., Cheng, M.M., Hou, Q.: Corrmatch: Label propagation via correlation matching for semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3097–3107 (2024)
30. Sun, R., Mai, H., Zhang, T., Wu, F.: Daw: exploring the better weighting function for semi-supervised semantic segmentation. Advances in neural information processing systems **36**, 61792–61805 (2023)
31. Tian, G.: Generalized kkm theorems, minimax inequalities, and their applications. Journal of Optimization Theory and Applications **83**(2), 375–389 (1994)
32. Vaswani, A.: Attention is all you need. Advances in Neural Information Processing Systems (2017)
33. Wang, A., Ma, H., Bai, L., Wu, Y., Xu, M., Zhang, Y., Islam, M., Ren, H.: Pedsemiseg: Pedagogy-inspired semi-supervised polyp segmentation. Computerized Medical Imaging and Graphics p. 102591 (2025)
34. Wang, H., Zhang, Q., Li, Y., Li, X.: Allspark: Reborn labeled features from unlabeled in transformer for semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3627–3636 (2024)
35. Wang, K., Nie, Y., Fang, C., Han, C., Wu, X., Wang, X., Lin, L., Zhou, F., Li, G.: Double-check soft teacher for semi-supervised object detection. In: IJCAI. pp. 1430–1436 (2022)

36. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pvt v2: Improved baselines with pyramid vision transformer. Computational Visual Media **8**(3), 415–424 (2022)
37. Wang, Z., Chen, Z., Liu, C., Zhao, Y., Zhu, X., Wu, Q.J.: Diversity augmentation and multi-fuzzy label for semi-supervised semantic segmentation. Neurocomputing **630**, 129681 (2025)
38. Wei, J., Hu, Y., Zhang, R., Li, Z., Zhou, S.K., Cui, S.: Shallow attention network for polyp segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. pp. 699–708. Springer (2021)
39. Wu, H., Xie, W., Lin, J., Guo, X.: Acl-net: semi-supervised polyp segmentation via affinity contrastive learning. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 2812–2820 (2023)
40. Wu, H., Zhao, Z., Zhong, J., Wang, W., Wen, Z., Qin, J.: Polypseg+: A lightweight context-aware network for real-time polyp segmentation. IEEE Transactions on Cybernetics **53**(4), 2610–2621 (2022)
41. Wu, Z., Lv, F., Chen, C., Hao, A., Li, S.: Colorectal polyp segmentation in the deep learning era: A comprehensive survey. arXiv preprint arXiv:2401.11734 (2024)
42. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves imagenet classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10687–10698 (2020)
43. Xu, H., Liu, L., Bian, Q., Yang, Z.: Semi-supervised semantic segmentation with prototype-based consistency regularization. Advances in neural information processing systems **35**, 26007–26020 (2022)
44. Xu, Z., Rittscher, J., Ali, S.: Ssl-cpcd: Self-supervised learning with composite pretext-class discrimination for improved generalisability in endoscopic image analysis. IEEE Transactions on Medical Imaging **43**(12), 4105–4119 (2024)
45. Xu, Z., Rittscher, J., Ali, S.: Sstfb: Leveraging self-supervised pretext learning and temporal self-attention with feature branching for real-time video polyp segmentation. arXiv preprint arXiv:2406.10200 (2024)
46. Xu, Z., Qiu, D., Lin, S., Zhang, X., Shi, S., Zhu, S., Zhang, F., Wan, X.: Temporal correlation network for video polyp segmentation. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 1317–1322. IEEE (2022)
47. Yang, L., Qi, L., Feng, L., Zhang, W., Shi, Y.: Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7236–7246 (2023)
48. Yang, L., Zhao, Z., Zhao, H.: Unimatch v2: Pushing the limit of semi-supervised semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
49. Zhang, Y., Liu, H., Hu, Q.: Transfuse: Fusing transformers and cnns for medical image segmentation. In: Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, Part I 24. pp. 14–24. Springer (2021)
50. Zhao, X., Wu, Z., Tan, S., Fan, D.J., Li, Z., Wan, X., Li, G.: Semi-supervised spatial temporal attention network for video polyp segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 456–466. Springer (2022)
51. Zhou, T., Zhou, Y., He, K., Gong, C., Yang, J., Fu, H., Shen, D.: Cross-level feature aggregation network for polyp segmentation. Pattern Recognition **140**, 109555 (2023)
52. Zhou, Y., Li, Y., Fu, Y., Benini, L., Konukoglu, E., Sun, G.: Camsam2: Segment anything accurately in camouflaged videos. Advances in Neural Information Processing Systems **38**, 55489–55517 (2026)

53. Zhuang, J., Wang, Z., Gao, Y.: Semi-supervised video semantic segmentation with inter-frame feature reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3263–3271 (2022)
54. Zhuang, J., Wang, Z., Zhang, Y., Fan, Z.: Infer from what you have seen before: Temporally-dependent classifier for semi-supervised video segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3575–3584 (2024)
55. Zoph, B., Ghiasi, G., Lin, T.Y., Cui, Y., Liu, H., Cubuk, E.D., Le, Q.: Rethinking pre-training and self-training. Advances in neural information processing systems **33**, 3833–3845 (2020)