AIMS Bioengineering, 13(4): 453–482.
DOI: 10.3934/bioeng.2026022
Received: 06 August 2026
Revised: 17 September 2026
Accepted: 13 October 2026
Published: 19 October 2026

https://www.aimspress.com/journal/Bioengineering

Research article

Cross-view feature similarity reweighting and stability-aware dynamic perturbation for semi-supervised polyp segmentation

Yujun Niu[1,*], **Yunqing Cong**[1,*], **Yalei Yin**[2], **and Wanshu Fan**[1]

[1] School of Software Engineering, Dalian University, Dalian, Liaoning 116622, China

[2] Department of Central Laboratory, Affiliated Zhongshan Hospital of Dalian University, Dalian, Liaoning 116001, China

* **Correspondence:** Email: niuyujun@dlu.edu.cn; congyunqing@s.dlu.edu.cn.

Abstract: Pixel-level annotation of colonoscopy images is costly and requires substantial expert effort, motivating the development of semi-supervised methods for polyp segmentation. A common weak-to-strong consistency strategy restricts supervision to high-confidence predictions, which can leave useful information in uncertain regions insufficiently exploited. Moreover, feature perturbations are often applied with a uniform strength across samples, without explicitly accounting for differences in representation stability. We have presented CFSAP, a semi-supervised polyp segmentation framework that addresses these limitations through cross-view similarity-guided reweighting and stability-aware feature perturbation. The proposed cross-view feature similarity reweighting (CFSR) strategy measures spatial feature similarity between weakly and strongly augmented views and constructs class-conditional weights for consistency learning in low-confidence regions. The stability-aware dynamic perturbation (SADP) strategy estimates sample-level representation stability from cross-view similarity and accordingly adjusts the magnitudes of feature noise injection and spatial masking. CFSAP was evaluated on the Kvasir-SEG and CVC-ClinicDB datasets under labeled-data ratios of 1/4, 1/8, and 1/16. It achieved the highest Dice and intersection over union (IoU) scores among all compared methods across the six experimental settings. Compared with the strongest competing method in each setting, CFSAP improved Dice by 0.30–1.38 percentage points on Kvasir-SEG and by 0.47–1.29 percentage points on CVC-ClinicDB. Ablation results further confirmed the complementary contributions of the two components and showed that stability-aware perturbation assignment consistently outperformed fixed and randomly assigned perturbation strengths.

Keywords: semi-supervised learning; polyp segmentation; colonoscopy; weak-to-strong consistency regularization; cross-view feature similarity; dynamic feature perturbation

1. Introduction

Colorectal cancer (CRC) remains a substantial global disease burden. According to GLOBOCAN 2022, more than 1.9 million new CRC cases and approximately 904,000 related deaths were estimated worldwide in 2022, making CRC the third most commonly diagnosed cancer and the second leading cause of cancer-related death [1]. Colonoscopy plays an important role in CRC screening and prevention because it enables the detection and removal of precancerous lesions, such as adenomas. A large community-based cohort study showed that higher adenoma detection rates (ADRs) were associated with lower risks of interval CRC, advanced-stage interval CRC, and fatal interval CRC, further highlighting the importance of reliable detection of adenomatous lesions during colonoscopy [2].

Compared with lesion detection or coarse localization alone, polyp segmentation provides pixel-level information on lesion extent and boundaries. Accurate segmentation results can provide more detailed spatial information for lesion visualization, quantitative morphological analysis, and subsequent computer-aided diagnostic tasks [3]. Therefore, improving the accuracy and robustness of polyp segmentation in challenging endoscopic scenes has potential clinical value. In recent years, deep learning methods have been widely applied to polyp segmentation in colonoscopy images. Classical semantic segmentation networks, including fully convolutional network (FCN), U-Net, and DeepLabV3+, perform target-region segmentation through end-to-end pixel-wise prediction [4–6]. Subsequent fully supervised polyp segmentation methods have introduced improvements in multi-scale feature fusion, boundary modeling, attention mechanisms, and global context modeling to enhance the segmentation of polyps across different scales and in regions with blurred boundaries or complex backgrounds [7–10]. Meanwhile, transformer-based and convolutional neural network (CNN)–transformer hybrid architectures have also been introduced into polyp segmentation to strengthen the modeling of long-range dependencies and global semantic information [11, 12]. Although these methods have achieved favorable segmentation performance, they generally rely on large quantities of high-quality pixel-wise annotations. Compared with image-level annotation, pixel-level annotation for polyp segmentation requires precise delineation of lesion regions and therefore demands substantially greater expert effort and annotation time. In clinical data collection, large numbers of colonoscopy images may be available, whereas obtaining high-quality pixel-level annotations for all of them is considerably more costly. Therefore, effectively exploiting unlabeled colonoscopy images under limited expert annotation has both methodological and practical relevance, as it can reduce the annotation burden in the development of segmentation models and improve the utilization of available clinical imaging data. Improving polyp segmentation performance with limited annotations has consequently become an important research problem in colonoscopy image analysis.

Semi-supervised learning jointly utilizes a small amount of labeled data and a large amount of unlabeled data for model training, providing an effective approach to alleviating annotation scarcity [13]. In the development of semi-supervised segmentation, early studies employed adversarial learning to align the output-space distributions of labeled and unlabeled data [14, 15]. However, the inherent instability of adversarial training limited its direct application to semantic segmentation. With subsequent advances, pseudo-labeling and consistency regularization gradually became two predominant paradigms. Pseudo-labeling methods focus on generating reliable supervision targets for unlabeled samples, whereas consistency regularization methods emphasize the

stability of model predictions under different perturbations. Within the latter category, weak-to-strong consistency regularization generates relatively reliable reference targets from weakly augmented views and constrains predictions from strongly augmented views to remain consistent with these targets. Owing to its simple framework and stable training, this paradigm has been widely applied to semi-supervised segmentation [16, 17].

Nevertheless, when weak-to-strong consistency regularization is applied to semi-supervised polyp segmentation, two issues require further consideration. First, existing weak-to-strong consistency methods commonly use high-confidence predictions from weakly augmented views as supervision signals for unlabeled data [16, 18]. Although this strategy can reduce the interference caused by noisy supervision to some extent, it also concentrates the learning signals from unlabeled samples primarily in regions where the model is relatively confident. In colonoscopy images, regions with blurred boundaries, low contrast, or complex local textures often produce unstable predictions and may fail to pass high-confidence filtering, even though they can still contain valuable pixel-level information [19, 20]. If these regions are consistently ignored during training, the model's ability to learn from difficult regions will be limited. Second, subsequent extensions of weak-to-strong consistency regularization have expanded perturbations from image augmentation to the feature level [21, 22]. However, existing feature perturbation strategies commonly employ sample-independent fixed perturbation strengths and therefore insufficiently account for differences among unlabeled samples. Because polyp images differ considerably in morphology, texture, and imaging quality, the consistency of their feature representations between weakly and strongly augmented views also varies. For samples with relatively stable feature representations, an excessively weak perturbation may fail to provide sufficiently broad training constraints in the feature space. In contrast, for samples with unstable feature representations, an excessively strong perturbation may severely disrupt semantic information.

Recent semi-supervised polyp segmentation methods have explored polyp-aware data augmentation [23], cross-pseudo supervision and contrastive learning [24], pedagogy-inspired training [25], and auxiliary depth guidance [26]. These approaches improve the use of unlabeled colonoscopy images through task-specific augmentation, alternative pseudo-supervision, or additional training signals. Meanwhile, recent semi-supervised segmentation studies have also begun to investigate the utilization of uncertain predictions and more adaptive forms of feature-level regularization. Nevertheless, within weak-to-strong consistency learning, it remains important to determine how low-confidence predictions can provide supplementary supervision without indiscriminately introducing unreliable signals, and how appropriate feature perturbation strengths can be assigned to unlabeled samples with different levels of representation stability. Simply discarding low-confidence regions may lose useful information from difficult regions, whereas treating these predictions indiscriminately may introduce unreliable supervision. Similarly, a uniform feature perturbation strength may not be suitable for unlabeled samples with different representation stability. Our work examines these two issues within a weak-to-strong consistency framework.

To address these issues, we propose CFSAP, a semi-supervised polyp segmentation framework that improves weak-to-strong consistency learning through cross-view similarity-guided supervision and sample-specific dynamic feature perturbation. CFSAP exploits cross-view feature information at two complementary granularities. CFSR uses spatial feature discrepancies between weakly and strongly augmented views to reweight supplementary consistency supervision in low-confidence regions,

whereas SADP aggregates the cross-view feature relationship at the sample level and dynamically adjusts the strengths of feature noise injection and masking according to representation stability. Their integration improves the utilization of unlabeled colonoscopy images from both low-confidence supervision and dynamic feature-perturbation perspectives under limited annotation. From an application perspective, the framework is intended to improve the utilization of unlabeled colonoscopy images when pixel-level annotations are limited, thereby providing a practical approach to reducing the annotation requirements for developing semi-supervised polyp segmentation models.

The main contributions are summarized as follows:

1. We propose CFSAP, a semi-supervised polyp segmentation framework that extends weak-to-strong consistency learning by exploiting cross-view feature information at both spatial and sample levels, thereby improving the utilization of unlabeled data through low-confidence supervision reweighting and sample-specific dynamic feature perturbation.
2. We design two complementary training strategies, CFSR and SADP. CFSR assigns class-conditional weights according to spatial feature discrepancies between weakly and strongly augmented views to selectively reweight supplementary consistency supervision in low-confidence regions. SADP uses sample-level cross-view representation stability to dynamically determine the strengths of feature noise injection and spatial masking for individual unlabeled samples.
3. Experiments on the Kvasir-SEG and CVC-ClinicDB datasets under three labeled-data ratios demonstrate consistent improvements over the compared semi-supervised methods. Ablation studies further validate the complementarity of CFSR and SADP and the effectiveness of matching perturbation strength to sample stability.

2. Related work

2.1. Segmentation-related studies and semi-supervised semantic segmentation

Recent studies in image segmentation and related dense-prediction tasks have explored more effective feature representations through global-context modeling, structural information interaction, and complementary feature integration. The collaborative compensative transformer network (CCTNet) jointly models patch and superpixel tokens to strengthen the interaction between global semantic information and fine-grained structural representations for salient object detection [27]. Salient object detection with non-local feature enhancement and edge reconstruction combines non-local feature enhancement with edge reconstruction to improve long-range contextual representation and local structural details [28]. These studies provide complementary perspectives on representation learning for dense prediction in complex visual scenes.

Semi-supervised semantic segmentation aims to improve pixel-wise prediction by leveraging a small number of images with pixel-level annotations together with a large number of unlabeled images. Existing approaches can generally be categorized into pseudo-labeling-based methods, consistency-regularization-based methods, and hybrid methods that incorporate mechanisms such as adversarial learning and contrastive learning [13]. Pseudo-labeling methods generate additional supervision signals for unlabeled images from model predictions [29]. For example, cross-pseudo-supervision employs a dual-network cross-supervision mechanism, in which each

network generates pseudo-labels to supervise the predictions of the other network [30]. Although these methods can directly exploit unlabeled data, their performance is susceptible to pseudo-label noise and confirmation bias. Consistency regularization methods improve the utilization of unlabeled data by encouraging the model to produce consistent outputs under different perturbations, network branches, or predictive views. The mean teacher approach improves training stability by enforcing prediction consistency between the teacher and student models [31]. Cross-consistency training utilizes unlabeled data through consistency constraints between the main decoder and auxiliary decoders [32]. Uncertainty-rectified pyramid consistency enhances multi-scale prediction consistency through uncertainty-rectified pyramid consistency constraints [33]. Evidential inference learning (EVIL) introduces evidence theory into a dual-network consistency regularization framework for uncertainty estimation, thereby improving the reliability of supervision derived from unlabeled data [34]. Other approaches further exploit unlabeled data by incorporating adversarial learning or contrastive learning. For example, consistency and adversarial semi-supervised learning improves segmentation performance through prediction-consistency and distribution-alignment constraints [35], whereas the approach of using unreliable pseudo-labels (U^2PL) incorporates unreliable pseudo-labels into contrastive learning as negative samples [36]. Compared with pseudo-labeling-based methods, consistency regularization centers on prediction stability and has a relatively clear framework, making it convenient to integrate different forms of perturbation and unlabeled supervision signals within a unified framework. Therefore, this study follows the consistency regularization paradigm.

2.2. Weak-to-strong consistency regularization

Weak-to-strong consistency regularization is an important consistency-training paradigm in semi-supervised learning. Its basic principle is to generate relatively reliable supervision targets from weakly augmented views and constrain predictions from strongly augmented views to remain consistent with these targets. FixMatch generates high-confidence pseudo-labels from weakly augmented views and uses them to supervise predictions from strongly augmented views, promoting the widespread adoption of the weak-to-strong consistency paradigm in semi-supervised learning [16]. As this paradigm was extended to semantic segmentation, researchers began to further investigate strong augmentation strategies, pixel-wise pseudo-label quality, and prediction stability. Augmentation Matters demonstrates that strong data augmentation plays an important role in semi-supervised semantic segmentation [18]. UniMatch further extends the weak-to-strong consistency framework by introducing multi-view perturbation-consistency constraints to improve the utilization of unlabeled samples [17]. In addition, methods such as Rethinking Copy-Paste and bidirectional copy-paste (BCP) further explore copy-paste-based mixed augmentation strategies. These methods construct more complex strongly perturbed samples to improve the utilization of unlabeled data in consistency training [37, 38]. Feature-level extensions further combine consistency constraints under image augmentation and feature perturbation to regularize both representation learning and prediction [39].

In addition to augmentation design, the treatment of uncertain or low-confidence supervision derived from weakly augmented predictions has become another important direction in recent semi-supervised segmentation. Different approaches exploit such information in different ways. Class probability space regularization (CPSR) regularizes the class probability space to improve the

usability of uncertain predictions [40]; CorrMatch exploits pixel correlations for pixel- and region-level label propagation [41]; and RankMatch constructs additional consistency constraints by modeling inter-pixel and inter-agent correlations [42]. In medical image segmentation, S^2Match rectifies pixel-wise predictions by modeling intra-image pairwise semantic relationships and further exploits semantic similarity between labeled and unlabeled data for class representation learning [19]. Uncertainty-participation context consistency learning (UCCL) incorporates uncertain regions into consistency learning through context-based supervision [20], while dynamic uncertainty-aware consistency and contrastive learning (DyCON) dynamically adjusts voxel contributions to consistency learning according to uncertainty and combines this constraint with contrastive learning [43]. These studies demonstrate that low-confidence information can be exploited through probability rectification, label propagation, semantic-relation modeling, uncertainty-aware weighting, and complementary learning objectives.

CFSR assigns different roles to confidence and cross-view feature similarity within the weak-to-strong framework. Conventional high-confidence weak-to-strong consistency remains the primary supervision for unlabeled data, while confidence identifies the regions not covered by this primary constraint. Within these low-confidence regions, feature similarity between corresponding spatial locations of the weakly and strongly augmented views determines the relative contribution of each pixel to the supplementary consistency loss. The weights are further normalized separately according to the predicted classes in the weakly augmented view, preventing pixels with different predicted semantics from sharing the same weight-normalization space. In this way, CFSR constructs supplementary low-confidence supervision through discrepancy-guided, class-conditional consistency reweighting, rather than modifying the pseudo-label selection rule or performing label propagation.

Feature-level perturbations have also been used to enlarge the perturbation space of consistency learning. Recent studies further explore adaptive feature regularization from different perspectives. Density-descending feature perturbation (DDFP) determines feature perturbation directions according to estimated feature density and its gradient, shifting confidently predicted features toward lower-density regions [44]. Attention inverted feature perturbation adaptively constructs feature perturbations from attention responses derived from the current feature representations [45]. Multi-constraint approaches also combine feature consistency under image augmentation with prediction consistency under instance-dependent feature intervention [39]. Within this landscape, SADP introduces sample-specific dynamic assignment of perturbation strength based on cross-view representation stability. The mean feature similarity between weakly and strongly augmented views estimates the stability of the current sample and dynamically determines the strengths of feature noise injection and spatial masking. This design places the emphasis on dynamic perturbation-strength assignment across samples while retaining the same feature perturbation operations.

On this basis, CFSAP coordinates the use of cross-view feature information at two granularities within the same weak-to-strong consistency framework. CFSR retains the spatial structure of the cross-view relationship to guide low-confidence consistency reweighting, whereas SADP aggregates this relationship into a sample-level stability measure to dynamically control feature perturbation strength. The two strategies therefore use the same cross-view feature relationship for distinct purposes, respectively improving the utilization of low-confidence supervision and the dynamic assignment of feature-perturbation strength across unlabeled samples.

3. Materials and method

This section presents the proposed CFSAP method in detail. We first describe the overall framework, then introduce the cross-view feature similarity reweighting strategy and the stability-aware dynamic perturbation strategy, and finally present the overall loss function used to optimize the model parameters.

3.1. Overall framework

The overall framework of CFSAP is illustrated in Figure 1. The method is built upon a weak-to-strong consistency training framework and comprises two basic training paths: supervised learning on labeled data and weak-to-strong consistency learning on unlabeled data. During training on unlabeled samples, cross-view feature similarity reweighting (CFSR) and stability-aware dynamic perturbation (SADP) are further incorporated. The encoder and decoder parameters are shared across all training paths.

Let the labeled and unlabeled datasets be denoted by $\mathcal{D}_l = \left\{\left(x_i^l, y_i^l\right)\right\}_{i=1}^{N_l}$, $\mathcal{D}_u = \left\{x_j^u\right\}_{j=1}^{N_u}$, where x_i^l denotes the ith labeled input image, y_i^l denotes the corresponding pixel-level annotation, and x_j^u denotes the jth unlabeled colonoscopy image. N_l and N_u denote the numbers of labeled and unlabeled samples, respectively. For notational simplicity, the following description considers one sample at a time.

For a labeled image, the model obtains the prediction P^l through the shared encoder $E(\cdot)$ and decoder $D(\cdot)$. The supervised loss $\mathcal{L}_{\text{sup}}$ is then computed using the corresponding ground-truth annotation y^l. Its formulation is given in Section 3.4.

For an unlabeled image u, weak and strong augmentations, denoted by $\mathcal{A}_w(\cdot)$ and $\mathcal{A}_s(\cdot)$, are first applied to obtain a weakly augmented view $u^w = \mathcal{A}_w(u)$ and a strongly augmented view $u^s = \mathcal{A}_s(u)$. The two views are subsequently fed into the shared encoder $E(\cdot)$, producing the intermediate features

$$Z^w = E\left(u^w\right), \qquad Z^s = E\left(u^s\right). \tag{3.1}$$

The features Z^w and Z^s are then passed through the shared decoder $D(\cdot)$ to obtain the corresponding output logits $O^w = D(Z^w)$ and $O^s = D(Z^s)$. The probability maps for the weakly and strongly augmented views are obtained using the softmax operation:

$$P^w = \text{Softmax}\left(O^w\right), \qquad P^s = \text{Softmax}\left(O^s\right). \tag{3.2}$$

For a pixel at spatial location (h, w), the pixel-level confidence of the weakly augmented prediction is defined as

$$C^w(h, w) = \max_{c \in \{0,1\}} P_c^w(h, w), \tag{3.3}$$

where c denotes the class index. For binary polyp segmentation, $c \in \{0, 1\}$. Given a confidence threshold τ, the set of pixels involved in high-confidence consistency supervision is defined as

$$\Omega_h = \left\{(h, w) \mid C^w(h, w) \geq \tau\right\}. \tag{3.4}$$

Following conventional weak-to-strong consistency training, the high-confidence regions of the weakly augmented prediction are used to constrain the strongly augmented prediction, yielding the high-confidence consistency loss $\mathcal{L}_h$. Its formulation is given in Section 3.4.

On this basis, CFSR uses Z^w, Z^s, and information from the weakly augmented prediction to construct a reweighted constraint guided by cross-view similarity. This constraint is imposed on the strongly augmented prediction P^s, resulting in the cross-view feature similarity reweighting loss $\mathcal{L}_{\text{cfsr}}$. SADP estimates feature stability from the cross-view relationship between Z^w and Z^s, and applies stability-aware dynamic perturbations to Z^s. The perturbed features are passed through the shared decoder to obtain the perturbed predictions P_n^s and P_d^s, which are constrained to remain consistent with the weakly augmented prediction, forming the stability-aware dynamic perturbation loss $\mathcal{L}_{\text{sadp}}$. The detailed constructions of CFSR and SADP are described in Sections 3.2 and 3.3, respectively.

The methodological core of CFSAP is to exploit the same weak/strong cross-view feature relationship at two complementary granularities. At the spatial level, CFSR uses local cross-view feature discrepancies to redistribute supplementary consistency supervision among class-conditional low-confidence pixels. At the sample level, SADP uses the overall cross-view representation stability to determine the dynamic strength of feature perturbations. Thus, the two strategies address two distinct decisions in unlabeled-data learning—where supplementary supervision should be emphasized and how strongly each unlabeled sample should be perturbed—while using cross-view feature relations as a common basis.

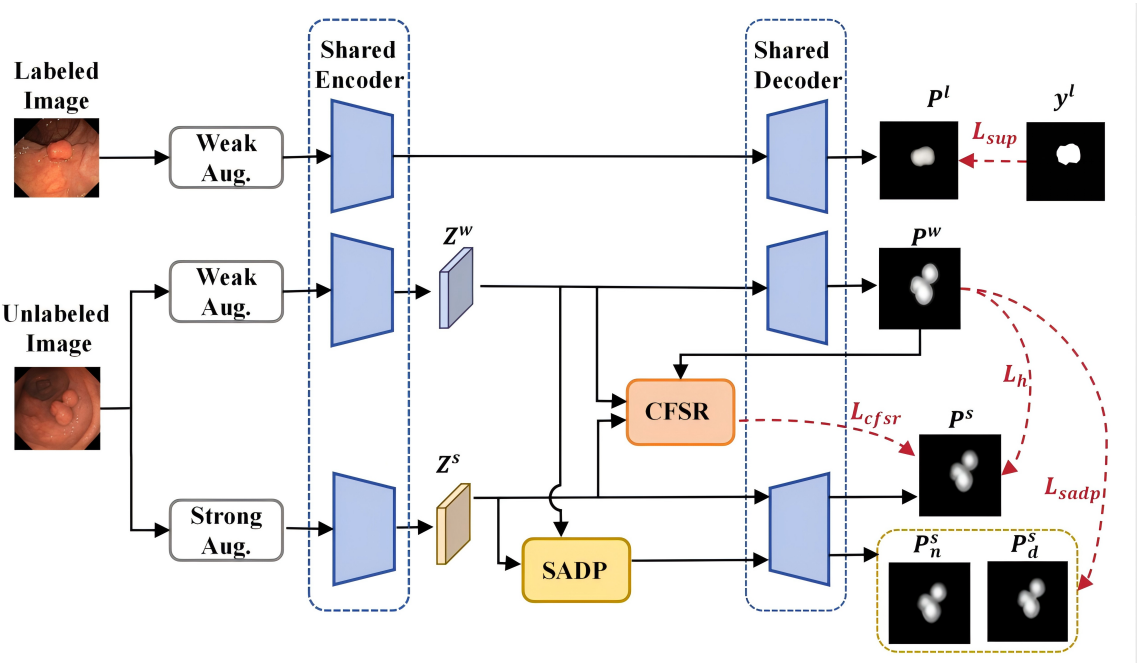


Figure 1. Overall framework of CFSAP. The red dashed lines indicate the loss constraints.

3.2. Cross-view feature similarity reweighting

To alleviate the underutilization of pixels excluded by high-confidence filtering, we introduce the cross-view feature similarity reweighting strategy, whose construction is illustrated in Figure 2. Given the weakly augmented feature Z^w and the strongly augmented feature Z^s, CFSR first computes a cross-view feature similarity map S between the two views and generates a low-confidence region mask from the weakly augmented prediction P^w. Cross-view feature similarity, the low-confidence region mask, and predicted class information are then combined to generate a class-conditional weight map W_c.

This map selectively reweights the pixel-wise consistency constraint within low-confidence regions, yielding the cross-view feature similarity reweighting loss $\mathcal{L}_{\text{cfsr}}$.

The key design of CFSR is the joint use of confidence-based candidate selection, class-conditional partitioning, and discrepancy-based weighting, so that low-confidence pixels are not treated as a single uniform supervision set. Confidence identifies the pixels outside the primary high-confidence supervision, the predicted class defines the corresponding normalization group, and the cross-view feature discrepancy determines the relative supervision contribution within that group.

First, the cross-view feature similarity map S is obtained by computing the cosine similarity between Z^w and Z^s at corresponding spatial locations:

$$S(h,w) = \frac{\langle Z^w(h,w), Z^s(h,w)\rangle}{\|Z^w(h,w)\|_2 \, \|Z^s(h,w)\|_2}. \tag{3.5}$$

Here, $Z^w(h,w)$ and $Z^s(h,w)$ denote the channel-wise feature vectors at spatial location (h,w) in the weakly and strongly augmented features, respectively, and $\langle\cdot,\cdot\rangle$ denotes the vector inner product. A lower cross-view similarity indicates a larger discrepancy between the feature representations at the corresponding location under weak and strong augmentations. Cross-view similarity is therefore used as the basis for reweighting pixels in low-confidence regions, allowing pixels with more pronounced feature discrepancies to receive greater attention in the supplementary consistency constraint.

To spatially align the weakly augmented probability map P^w and confidence map C^w with the cross-view similarity map S, they are resized to the same spatial dimensions as S. For notational simplicity, the resized maps are still denoted by P^w and C^w. Based on the confidence threshold τ defined in Section 3.1, the low-confidence region mask is defined as

$$M^u(h,w) = \mathbb{I}\left(C^w(h,w) < \tau\right), \tag{3.6}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. When $C^w(h,w) < \tau$, $M^u(h,w) = 1$, indicating that the location belongs to a low-confidence region excluded by high-confidence filtering; otherwise, $M^u(h,w) = 0$. This mask identifies the candidate pixels involved in CFSR and provides the spatial basis for the subsequent class-conditional partitioning and reweighting.

To prevent low-confidence pixels with different predicted semantics from sharing the same weight-normalization space, we further partition the low-confidence region according to the class prediction of the weakly augmented view. For each class $c \in \{0,1\}$, the corresponding class-conditional low-confidence region is defined as

$$\Omega_c^u = \left\{ (h,w) \;\middle|\; M^u(h,w) = 1,\ \arg\max_{k\in\{0,1\}} P_k^w(h,w) = c \right\}. \tag{3.7}$$

The set Ω_c^u contains the low-confidence pixels predicted as class c in the weakly augmented view. Constructing an independent region for each class prevents pixels with different predicted semantics from participating in the same normalization process.

Within each class-conditional low-confidence region, the cross-view feature discrepancy is measured by $1 - S(h,w)$. A lower cross-view similarity therefore corresponds to a larger discrepancy score. The class-conditional weight map is obtained by normalizing the discrepancy scores only

within Ω_c^u:

$$W_c(h,w) = \begin{cases} \dfrac{\exp(1 - S(h,w))}{\sum\limits_{(i,j)\in\Omega_c^u} \exp(1 - S(i,j))}, & (h,w) \in \Omega_c^u, \\ 0, & (h,w) \notin \Omega_c^u. \end{cases} \tag{3.8}$$

Consequently, pixels outside Ω_c^u receive zero weight and do not participate in the normalization. Within Ω_c^u, pixels with lower cross-view similarity receive larger normalized weights, allowing the supplementary consistency constraint to place greater emphasis on locations exhibiting more pronounced feature discrepancies.

The exponential form in Equation (3.8) is used as a normalized monotonic weighting function rather than as a probabilistically derived estimator. Defining the feature-discrepancy score as $d(h,w) = 1 - S(h,w)$, Equation (3.8) is equivalent to applying a softmax normalization to the discrepancy scores within each class-conditional low-confidence region. Since cosine similarity satisfies $S(h,w) \in [-1,1]$, the discrepancy score is bounded by $d(h,w) \in [0,2]$. The resulting weights are positive and sum to one within each nonempty class-conditional region, so the total weighting scale does not grow with the number of low-confidence pixels in that region. Moreover, for any two pixels in the same region, the ratio between their unnormalized weights is at most e^2, which emphasizes larger cross-view discrepancies without allowing an arbitrarily large relative weighting ratio. We therefore adopt this simple normalized mapping to emphasize relative feature inconsistency while retaining a controlled class-conditional weighting space.

Let $C_u = \{c \in \{0,1\} \mid \Omega_c^u \neq \varnothing\}$ denote the set of classes containing at least one pixel in the current low-confidence region. The CFSR loss is formulated as

$$\mathcal{L}_{\text{cfsr}} = \frac{1}{|C_u|} \sum_{c\in C_u} \sum_{(h,w)\in\Omega_c^u} W_c(h,w)\, \ell_{\text{ce}}\left(P^s(h,w), P^w(h,w)\right). \tag{3.9}$$

Here, $\ell_{\text{ce}}(\cdot)$ denotes the pixel-wise cross-entropy consistency loss, and $P^s(h,w)$ and $P^w(h,w)$ denote the strongly and weakly augmented predictions at spatial location (h,w), respectively. For notational simplicity, $\ell_{\text{ce}}(P^s(h,w), P^w(h,w))$ denotes the cross-entropy loss computed using the maximum-probability class of $P^w(h,w)$ as the reference target. The weight $W_c(h,w)$ determines the relative contribution of each pixel within the corresponding class-conditional low-confidence region. Since the loss is accumulated only over Ω_c^u, CFSR selectively reweights the consistency constraint within low-confidence regions predicted as class c.

It should be emphasized that CFSR does not assign equal importance to all low-confidence predictions, nor does it replace high-confidence consistency supervision. The high-confidence consistency loss $\mathcal{L}_h$ continues to provide the primary supervision for unlabeled data, whereas CFSR introduces a supplementary constraint within the class-conditional low-confidence regions. The contribution of each pixel is determined by its cross-view feature discrepancy, so pixels exhibiting larger discrepancies receive greater attention, while pixels outside the corresponding region do not contribute to $\mathcal{L}_{\text{cfsr}}$. In this manner, CFSR selectively exploits potentially informative pixels excluded by confidence filtering while reducing the risk associated with indiscriminate supervision of low-confidence predictions. Here, confidence and cross-view feature similarity play different roles. The former identifies candidate regions that are not covered by conventional high-confidence consistency supervision, whereas the latter differentiates the relative contributions of these

low-confidence pixels to the supplementary consistency constraint. Therefore, CFSR does not enlarge the pseudo-label set by simply lowering the confidence threshold or rectify the low-confidence predictions themselves; instead, it selectively reweights low-confidence regions while retaining the original high-confidence supervision.

Because the reference classes in Ω_c^u are obtained from low-confidence predictions, CFSR cannot guarantee that all reweighted pseudo-labels are correct. Cross-view feature discrepancy is used as an importance cue rather than as an estimate of pseudo-label correctness; therefore, a highly discrepant pixel may still have an incorrect weak-view reference prediction. This risk is controlled rather than eliminated. CFSR remains a supplementary constraint instead of replacing the primary high-confidence consistency supervision, the class-conditional softmax normalization fixes the total weight within each nonempty region, and the overall objective scales $\mathcal{L}_{\text{cfsr}}$ by the relatively small coefficient $\lambda_u = 0.015$. The sensitivity analysis in Section 4.4 further shows that increasing the region assigned to CFSR by using a higher confidence threshold does not monotonically improve performance, which supports maintaining a balance between reliable high-confidence supervision and selective low-confidence utilization.

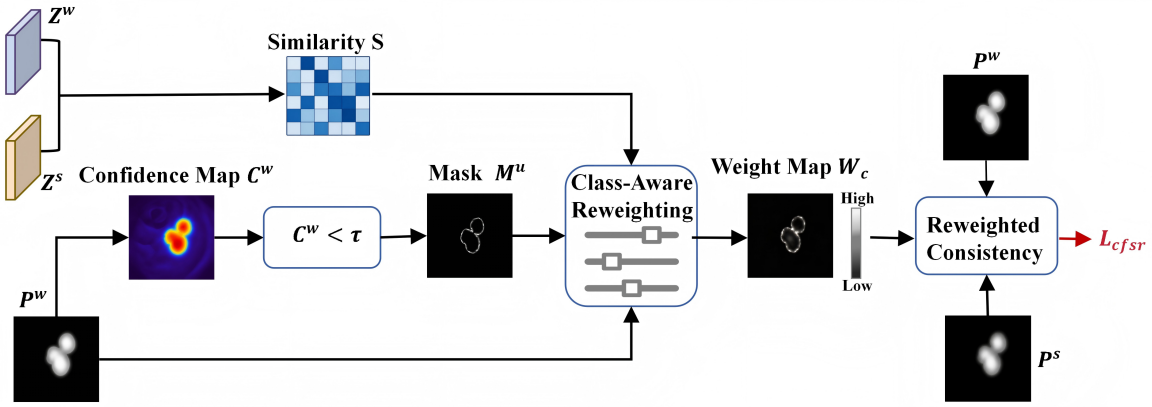


Figure 2. Illustration of cross-view feature similarity reweighting.

3.3. Stability-aware dynamic perturbation

To address the limitations of sample-independent fixed perturbation strengths in accounting for differences among unlabeled samples, we introduce the stability-aware dynamic perturbation strategy, whose construction is illustrated in Figure 3. SADP estimates the feature stability of the current unlabeled sample using cross-view feature similarity between weakly and strongly augmented views and generates a dynamic perturbation strength α accordingly. The value of α controls the noise-amplitude range in the feature noise-injection branch and the spatial-masking range in the feature-masking branch. Stability-aware perturbations are thereby applied to the strongly augmented feature Z^s, and the stability-aware dynamic perturbation loss $\mathcal{L}_{\text{sadp}}$ is constructed through prediction-consistency constraints under these perturbations.

The key design of SADP lies in establishing a sample-specific correspondence between cross-view representation stability and perturbation strength. Rather than applying a uniform perturbation level to all unlabeled samples, the estimated stability of each sample determines a dynamic strength that jointly controls the feature noise-injection and feature-masking branches.

To obtain a sample-level measure of feature stability, the cross-view similarity map S in Equation (3.5) is globally averaged:

$$\bar{s} = \frac{1}{H'W'} \sum_{h=1}^{H'} \sum_{w=1}^{W'} S(h, w), \tag{3.10}$$

where H' and W' denote the height and width of the cross-view feature similarity map S, respectively. A larger $\bar{s}$ indicates that the overall feature representations of the weakly and strongly augmented views are more similar and that the current sample has higher feature stability.

The dynamic perturbation strength is then generated from the sample-level feature similarity:

$$\alpha = \lambda_\alpha (1 + \bar{s}), \tag{3.11}$$

where λ_α is the scaling coefficient for the perturbation strength and is set to 0.15 in the experiments. According to Equation (3.11), α increases with the sample-level cross-view feature similarity, allowing the subsequent perturbation strength to be dynamically adjusted according to the feature stability of each unlabeled sample.

Equation (3.11) is designed as a simple bounded monotonic transformation that links sample-level representation stability to perturbation strength. Since each cosine similarity value lies in $[-1, 1]$, its spatial average also satisfies $\bar{s} \in [-1, 1]$. Therefore, $1 + \bar{s} \in [0, 2]$ and the perturbation strength is bounded by $\alpha \in [0, 2\lambda_\alpha]$. For $\lambda_\alpha > 0$, the mapping is strictly increasing with respect to $\bar{s}$, preserving the intended relationship that samples with more stable cross-view representations receive stronger perturbations, whereas samples with less stable representations receive weaker perturbations. At the same time, λ_α directly controls the upper bound of the perturbation range without introducing an additional shape parameter. Section 4.4 further evaluates the empirical sensitivity to λ_α.

After obtaining α, the strongly augmented feature Z^s is perturbed through feature noise injection and feature masking. In the feature noise-injection branch, SADP samples a noise tensor N from the uniform distribution $\mathcal{U}(-\alpha, \alpha)$, where N has the same dimensions as Z^s. The noise is then injected into the strongly augmented feature to obtain the noise-perturbed feature:

$$Z_n^s = Z^s \odot (1 + N), \tag{3.12}$$

where $\odot$ denotes element-wise multiplication. Because the sampling range of the noise is determined by α, a larger α broadens the noise-amplitude range and produces a stronger noise perturbation, whereas a smaller α narrows the range and weakens the perturbation.

In the feature-masking branch, a spatial activation map is first computed from the strongly augmented feature Z^s:

$$A(h, w) = \frac{1}{C} \sum_{c=1}^{C} Z_c^s(h, w), \tag{3.13}$$

where C denotes the number of feature channels and $Z_c^s(h, w)$ denotes the response of the cth channel at spatial location (h, w). The spatial activation map A describes the mean feature-response intensity at each spatial location.

The dynamic perturbation strength α is then used to control the feature-masking threshold. Specifically, a masking adjustment coefficient η is sampled from $\mathcal{U}(1 - \alpha - \delta, 1 - \alpha + \delta)$, where δ

denotes the boundary offset of the sampling interval for η and is set to 0.1 in the experiments. In this interval, α changes the center according to sample stability, whereas δ controls the half-width of the sampling interval and therefore the stochastic variation of the masking threshold around that stability-dependent center. The masking threshold is calculated as $T = \eta \cdot \max_{h,w} A(h, w)$, and the feature-masking mask is constructed as

$$G(h, w) = \mathbb{I}\left(A(h, w) < T\right). \tag{3.14}$$

When $A(h, w) < T$, $G(h, w) = 1$, and the feature at that location is retained. Otherwise, $G(h, w) = 0$, and the feature is masked. The resulting mask-perturbed feature is

$$Z_d^s = Z^s \odot G. \tag{3.15}$$

The feature-masking branch is also dynamically controlled by α. When the sample-level feature similarity is high, α increases, so the masking adjustment coefficient η and the corresponding threshold T tend to decrease. Consequently, more high-response locations fail to satisfy $A(h, w) < T$ and are masked, resulting in a stronger feature-masking perturbation. When the sample-level feature similarity is low, α decreases, and the threshold T tends to increase. More spatial locations are retained, and the masking perturbation is correspondingly weakened.

The noise-perturbed feature Z_n^s and mask-perturbed feature Z_d^s are subsequently passed through the shared decoder to obtain the corresponding perturbed predictions P_n^s and P_d^s. SADP constrains both predictions to remain consistent with the weakly augmented reference prediction P^w. The loss is defined as

$$\mathcal{L}_{\text{sadp}} = \text{MSE}\left(P_n^s, P^w\right) + \text{MSE}\left(P_d^s, P^w\right), \tag{3.16}$$

where the first term constrains consistency between the prediction from the noise-injection branch P_n^s and the weakly augmented prediction P^w, while the second term constrains consistency between the prediction from the feature-masking branch P_d^s and P^w. Together, these two terms enhance prediction stability under feature perturbations. SADP dynamically associates perturbation strength with the cross-view representation stability of the current sample. When the weakly and strongly augmented views exhibit higher overall feature similarity, the representation of the sample is relatively stable, and a stronger feature perturbation is applied to provide a broader regularization constraint. Conversely, when the cross-view feature similarity is lower, a weaker perturbation is used to reduce excessive disruption of an already unstable semantic representation. In this way, SADP explicitly incorporates differences in sample-level representation stability into dynamic perturbation-strength assignment.

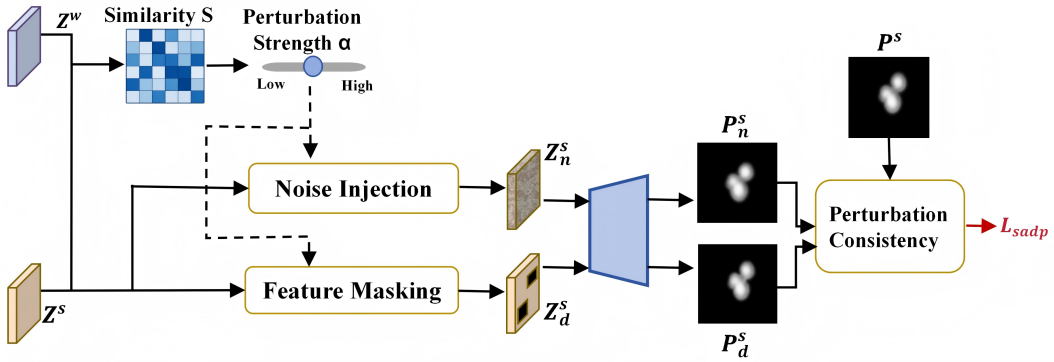


Figure 3. Illustration of stability-aware dynamic perturbation. The dashed lines indicate how the perturbation strength α controls the feature perturbation operations.

3.4. Overall loss function

For a labeled image, let P^l and y^l denote its prediction and ground-truth annotation, respectively, as defined in Section 3.1. The supervised loss is defined as

$$\mathcal{L}_{\text{sup}} = \frac{1}{HW} \sum_{h,w} \ell_{\text{ce}} \left(P^l(h,w), y^l(h,w) \right), \tag{3.17}$$

where H and W denote the spatial dimensions of the input sample, ℓ_{ce} denotes the pixel-wise cross-entropy loss, $P^l(h,w)$ denotes the prediction of the labeled sample at pixel location (h,w), and $y^l(h,w)$ denotes the corresponding ground-truth class label. This loss uses pixel-level annotations from labeled images to provide supervision.

For an unlabeled image, CFSAP adopts the high-confidence consistency constraint commonly used in existing weak-to-strong consistency methods. Given the weakly augmented prediction P^w, the strongly augmented prediction P^s, and the high-confidence pixel set Ω_h defined in Section 3.1, the high-confidence consistency loss is defined as

$$\mathcal{L}_h = \frac{1}{|\Omega_h|} \sum_{(h,w)\in\Omega_h} \ell_{\text{ce}} \left(P^s(h,w), P^w(h,w) \right). \tag{3.18}$$

Here, $\ell_{\text{ce}}(\cdot)$ denotes the cross-entropy consistency loss, and $P^s(h,w)$ and $P^w(h,w)$ denote the strongly and weakly augmented predictions at pixel location (h,w), respectively. Using the weakly augmented prediction as the reference, this loss constrains the strongly augmented prediction only at locations where the weakly augmented prediction has high confidence, thereby forming a high-confidence consistency constraint for unlabeled samples.

In addition to $\mathcal{L}_h$, training on unlabeled images jointly incorporates the cross-view feature similarity reweighting loss $\mathcal{L}_{\text{cfsr}}$ defined in Section 3.2 and the stability-aware dynamic perturbation loss $\mathcal{L}_{\text{sadp}}$ defined in Section 3.3. The overall loss function is therefore defined as

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \mathcal{L}_h + \lambda_u \mathcal{L}_{\text{cfsr}} + \lambda_p \mathcal{L}_{\text{sadp}}, \tag{3.19}$$

where λ_u and λ_p are the weighting coefficients for the cross-view feature similarity reweighting loss and the stability-aware dynamic perturbation loss, respectively. The losses above are formulated for a single sample; during training, each loss term is averaged over the mini-batch dimension.

The supervised loss $\mathcal{L}_{\text{sup}}$ provides pixel-level supervision for labeled samples. The high-confidence consistency loss $\mathcal{L}_h$ uses high-confidence regions from weakly augmented predictions to constrain predictions from strongly augmented views. The CFSR loss $\mathcal{L}_{\text{cfsr}}$ applies a selectively reweighted consistency constraint guided by low-confidence-region responses, thereby alleviating the underutilization of low-confidence regions when supervision relies only on high-confidence consistency. The SADP loss $\mathcal{L}_{\text{sadp}}$ dynamically adjusts the feature perturbation strength according to sample-level cross-view similarity between weakly and strongly augmented views and enhances model robustness to feature perturbations by enforcing consistency between the perturbed predictions and the weakly augmented prediction. By jointly optimizing these losses, CFSAP improves the exploitation of potentially useful supervision signals in unlabeled colonoscopy images from three perspectives: high-confidence consistency supervision, selective utilization of low-confidence regions, and sample-specific dynamic feature-perturbation constraints.

3.5. Datasets and data partitioning

To evaluate the effectiveness of the proposed method, we used two publicly available colonoscopy polyp segmentation datasets: Kvasir-SEG [46] and CVC-ClinicDB [47]. Both datasets are widely used public benchmarks for polyp segmentation in colonoscopy images. They differ in dataset size, image resolution, and scene complexity, enabling a relatively comprehensive evaluation of segmentation performance under different data conditions.

The Kvasir-SEG dataset contains 1000 polyp images with pixel-level annotations, with image resolutions ranging from 332×487 to 1920×1072 pixels. The dataset exhibits substantial diversity in polyp size, morphology, texture, and imaging conditions, and is therefore commonly used to evaluate segmentation performance across varied scenarios.

The CVC-ClinicDB dataset consists of 612 polyp images with pixel-level annotations, each with a resolution of 384×288 pixels. Compared with Kvasir-SEG, CVC-ClinicDB has relatively standardized imaging conditions and is also widely used to evaluate segmentation performance in standard clinical settings.

For each dataset, the images were randomly divided into training, validation, and test sets at a ratio of $8 : 1 : 1$. To evaluate semi-supervised learning under limited annotations, labeled and unlabeled subsets were constructed using different labeled-data ratios, and the comparative and ablation experiments used the same data splits. Following the labeled-data settings adopted by an existing semi-supervised segmentation method on Kvasir-SEG [48], experiments were conducted using labeled ratios of 1/4, 1/8, and 1/16.

3.6. Evaluation metrics

Dice coefficient, intersection over union (IoU), precision, recall (sensitivity), and the 95% Hausdorff distance (HD95) were used to evaluate segmentation performance. Dice and IoU quantify the overlap between the predicted and ground-truth regions, while precision and recall characterize foreground prediction from complementary perspectives. Let TP, FP, and FN denote the numbers of true-positive,

false-positive, and false-negative pixels, respectively. The four region-based metrics are defined as

$$\text{Dice} = \frac{2TP}{2TP + FP + FN}, \quad \text{IoU} = \frac{TP}{TP + FP + FN},$$
$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN}.$$

HD95 evaluates the spatial discrepancy between the predicted and ground-truth contours. Let B_p and B_g denote the boundary point sets of the predicted mask and the ground-truth mask, respectively. The minimum Euclidean distance from a point x to a boundary set B is defined as $d(x, B) = \min_{y \in B} \|x - y\|_2$. The bidirectional boundary-distance set and HD95 are then defined as

$$D = \{d(x, B_g) \mid x \in B_p\} \cup \{d(y, B_p) \mid y \in B_g\},$$
$$\text{HD95} = P_{95}(D),$$

where P_{95} denotes the 95th percentile. HD95 was calculated on a common 256×256 pixel grid to ensure a consistent pixel-distance scale across images with different native resolutions. Higher Dice, IoU, precision, and recall values indicate better segmentation performance, whereas a lower HD95 value is preferable.

3.7. Experimental settings and hyperparameters

The proposed method was implemented using PyTorch and trained on a single NVIDIA GeForce RTX 4090 GPU with 24 GB of memory. DeepLabV3+ [6] with an ImageNet-pretrained ResNet-50 [49] backbone was adopted as the segmentation network. The model was optimized using stochastic gradient descent with an initial learning rate of 0.001, a momentum of 0.9, and a weight decay of 1×10^{-4}. The learning rate was updated using a polynomial decay schedule: $\eta = \eta_0(1 - \text{iter}/\text{iter}_{\max})^{0.9}$, where η_0 denotes the initial learning rate and $\text{iter}_{\max}$ denotes the maximum number of training iterations. The batch size was set to 8, and the maximum number of training iterations was set to 20,000. As part of the weak augmentation pipeline, each image and its corresponding mask underwent random scaling and random cropping to a final spatial size of 256×256, together with horizontal flipping. Strong augmentations retained these geometric transformations and additionally included color jittering, random grayscale conversion, and Gaussian blurring.

For the module-specific hyperparameters, the confidence threshold τ was set to 0.8, and the loss weights λ_u and λ_p were set to 0.015 and 0.01, respectively. The perturbation scaling coefficient λ_α was set to 0.15, and the masking-interval offset δ was set to 0.10. All experiments used a fixed random seed and identical data splits to ensure reproducibility. The sensitivity of τ, λ_α, and δ is further evaluated in Section 4.4.

4. Results and discussion

4.1. Comparison with representative semi-supervised methods

To evaluate the effectiveness of CFSAP for semi-supervised polyp segmentation in colonoscopy images, we compared it with several representative semi-supervised segmentation methods on

Kvasir-SEG and CVC-ClinicDB. The compared methods included CPS, FixMatch, UniMatch, URPC, EVIL, and PolypMixNet. Experiments were conducted using labeled ratios of 1/4, 1/8, and 1/16. For each setting, the corresponding fraction of the training samples was treated as labeled data, and the remaining training samples were treated as unlabeled data. SupOnly denotes supervised training using only the labeled samples at the corresponding ratio, without using any unlabeled data, and serves as a reference for limited-label supervised training. To ensure a fair comparison, all methods were evaluated using the same data splits, input settings, and training configurations, with DeepLabV3+ and ResNet-50 uniformly adopted as the segmentation network and backbone, respectively.

Table 1. Quantitative comparison of different methods on the Kvasir-SEG dataset under different labeled ratios.

Method	Labeled ratio	Dice (%) ↑	IoU (%) ↑	Precision (%) ↑	Recall (%) ↑	HD95 (px) ↓
SupOnly	1/4	83.42	71.55	85.78	81.19	37.10
CPS [30]	1/4	84.27	72.82	**90.88**	78.56	**28.68**
FixMatch [16]	1/4	83.78	72.08	86.80	80.96	30.97
UniMatch [17]	1/4	83.71	71.98	88.08	79.75	30.46
URPC [33]	1/4	84.25	72.78	85.31	83.22	31.18
EVIL [34]	1/4	84.41	73.02	87.32	81.69	29.37
PolypMixNet [23]	1/4	83.66	71.91	83.45	**83.87**	36.58
CFSAP (Ours)	1/4	**85.70**	**74.97**	88.50	83.07	29.16
SupOnly	1/8	79.36	65.79	88.22	72.12	36.98
CPS [30]	1/8	82.18	69.76	**90.60**	75.19	36.68
FixMatch [16]	1/8	83.20	71.23	88.13	78.79	34.33
UniMatch [17]	1/8	83.76	72.05	86.93	80.81	**33.83**
URPC [33]	1/8	83.63	71.87	88.85	78.99	39.85
EVIL [34]	1/8	83.28	71.35	86.34	80.43	35.25
PolypMixNet [23]	1/8	83.39	71.51	85.24	81.62	39.37
CFSAP (Ours)	1/8	**85.14**	**74.13**	88.30	**82.20**	35.13
SupOnly	1/16	78.29	64.32	79.16	77.44	48.60
CPS [30]	1/16	78.47	64.56	84.88	72.96	39.87
FixMatch [16]	1/16	81.52	68.80	84.22	78.99	35.27
UniMatch [17]	1/16	82.37	70.03	86.03	79.01	37.99
URPC [33]	1/16	79.73	66.30	79.70	79.76	46.13
EVIL [34]	1/16	81.03	68.11	**86.46**	76.24	**34.88**
PolypMixNet [23]	1/16	79.82	66.42	79.12	**80.53**	42.67
CFSAP (Ours)	1/16	**82.67**	**70.46**	85.00	80.46	37.89

The results on Kvasir-SEG are reported in Table 1. CFSAP achieved the highest Dice and IoU scores under all three labeled ratios. With a labeled ratio of 1/4, CFSAP achieved Dice and IoU scores of 85.70% and 74.97%, respectively, exceeding the second-best method, EVIL, by 1.29 and 1.95 percentage points. With a labeled ratio of 1/8, CFSAP achieved Dice and IoU scores of 85.14% and 74.13%, outperforming the second-best method, UniMatch, by 1.38 and 2.08 percentage points. With

a labeled ratio of 1/16, CFSAP achieved a Dice score of 82.67% and an IoU score of 70.46%, again outperforming the other compared methods. Polyps in Kvasir-SEG vary in scale, morphology, and imaging conditions. Under limited annotations, the model must therefore address segmentation difficulties caused by appearance variations, local reflections, low contrast, and background interference. The results on this dataset suggest that, beyond high-confidence consistency supervision, CFSR and SADP further improve the utilization of unlabeled data by providing supplementary supervision for low-confidence regions and dynamic feature-perturbation constraints, respectively.

Table 2. Quantitative comparison of different methods on the CVC-ClinicDB dataset under different labeled ratios.

Method	Labeled ratio	Dice (%) ↑	IoU (%) ↑	Precision (%) ↑	Recall (%) ↑	HD95 (px) ↓
SupOnly	1/4	90.32	82.35	90.11	90.53	25.48
CPS [30]	1/4	91.95	85.11	**93.55**	90.40	11.79
FixMatch [16]	1/4	92.30	85.70	93.06	91.55	12.36
UniMatch [17]	1/4	92.60	86.23	93.29	91.92	9.75
URPC [33]	1/4	92.89	86.73	92.54	93.24	11.42
EVIL [34]	1/4	91.64	84.57	89.23	94.18	16.19
PolypMixNet [23]	1/4	91.81	84.86	89.12	**94.67**	13.47
CFSAP (Ours)	1/4	**93.50**	**87.79**	93.00	94.01	**9.07**
SupOnly	1/8	88.55	79.45	87.61	89.51	21.13
CPS [30]	1/8	89.33	80.71	**91.61**	87.16	**13.25**
FixMatch [16]	1/8	88.98	80.14	88.63	89.33	13.72
UniMatch [17]	1/8	89.07	80.29	89.86	88.29	14.69
URPC [33]	1/8	90.57	82.76	88.65	92.58	16.01
EVIL [34]	1/8	89.01	80.20	87.38	90.70	15.15
PolypMixNet [23]	1/8	89.28	80.63	86.19	92.60	14.32
CFSAP (Ours)	1/8	**91.04**	**83.55**	89.32	**92.83**	13.78
SupOnly	1/16	82.03	69.54	85.98	78.43	27.61
CPS [30]	1/16	82.47	70.17	**92.41**	74.46	21.16
FixMatch [16]	1/16	86.06	75.53	85.29	86.84	17.62
UniMatch [17]	1/16	85.76	75.08	85.78	85.74	**16.13**
URPC [33]	1/16	86.57	76.32	85.50	87.67	21.41
EVIL [34]	1/16	85.65	74.90	84.84	86.48	19.51
PolypMixNet [23]	1/16	86.17	75.71	84.64	**87.76**	26.90
CFSAP (Ours)	1/16	**87.86**	**78.35**	88.17	87.55	19.25

Table 2 reports the results on CVC-ClinicDB. CFSAP also achieved the highest Dice and IoU scores under all three labeled ratios. With a labeled ratio of 1/4, CFSAP achieved Dice and IoU scores of 93.50% and 87.79%, respectively, exceeding the second-best method, URPC, by 0.61 and 1.06 percentage points. With a labeled ratio of 1/8, CFSAP achieved Dice and IoU scores of 91.04% and 83.55%, outperforming URPC by 0.47 and 0.79 percentage points. With a labeled ratio of 1/16, CFSAP achieved Dice and IoU scores of 87.86% and 78.35%, exceeding URPC by 1.29 and 2.03 percentage points. Compared with Kvasir-SEG, CVC-ClinicDB is smaller and has more uniform image

resolutions. CFSAP nevertheless achieved the best Dice and IoU results on this dataset, indicating that the proposed strategies maintain consistent effectiveness across different data distributions.

Although CFSAP remained the best-performing method in terms of Dice and IoU on CVC-ClinicDB, its improvement margins over the comparison methods were smaller than those on Kvasir-SEG in several settings. This difference should be interpreted in the context of the distinct characteristics of the two datasets. As described in Section 3.5, Kvasir-SEG contains images with a wide range of native resolutions and substantial variation in polyp size, morphology, texture, and imaging conditions, whereas CVC-ClinicDB has a fixed image resolution and relatively standardized imaging conditions. The greater heterogeneity of Kvasir-SEG therefore corresponds to a broader range of segmentation conditions, while CVC-ClinicDB is associated with relatively higher baseline performance. For example, under the 1/4 labeled setting, SupOnly achieved a Dice score of 90.32% on CVC-ClinicDB, compared with 83.42% on Kvasir-SEG, leaving less numerical room for further improvement on CVC-ClinicDB. Therefore, the differences in improvement margins should be interpreted together with both the dataset characteristics and the baseline performance levels.

Considering all five metrics, CFSAP consistently achieved the highest Dice and IoU across all six dataset–label-ratio settings, while maintaining competitive performance in precision, recall, and HD95. These results provide a more comprehensive evaluation of the proposed method across region-overlap, foreground-prediction, and contour-distance measures.

In addition, CFSAP achieved higher Dice and IoU scores than SupOnly under every labeled ratio on both datasets. In terms of Dice, CFSAP exceeded SupOnly by 2.28, 5.78, and 4.38 percentage points under the 1/4, 1/8, and 1/16 labeled ratios on Kvasir-SEG, respectively. The corresponding improvements on CVC-ClinicDB were 3.18, 2.49, and 5.83 percentage points. These results suggest that, under limited-label training conditions, CFSAP constructs effective training signals from unlabeled samples and thereby helps improve segmentation performance when annotations are limited.

4.2. Qualitative comparison

To further compare the segmentation results qualitatively, representative samples were selected from the Kvasir-SEG and CVC-ClinicDB test sets, as shown in Figure 4. The selected examples exhibit differences in lesion appearance, target extent, and surrounding background, allowing qualitative comparison of local omissions, false-positive regions, and segmentation completeness.

As indicated by the error visualization, several comparison methods still exhibit localized red regions corresponding to missed ground-truth pixels or green regions corresponding to predictions outside the ground-truth regions. In the displayed examples, CFSAP generally produces more complete overlap with the ground-truth masks and fewer localized red or green regions, indicating reduced under-segmentation and over-segmentation in these cases.

These qualitative observations are consistent with the complementary roles of CFSR and SADP. CFSR supplements conventional high-confidence consistency supervision by selectively reweighting supplementary supervision within low-confidence regions according to cross-view feature discrepancies, thereby retaining potentially useful training signals that would otherwise be excluded by confidence filtering. SADP further adjusts feature-perturbation strength according to sample-level cross-view representation stability, providing sample-adaptive feature-level consistency constraints rather than applying a uniform perturbation strength to all unlabeled samples. Their combination

therefore improves the utilization of unlabeled data from both supervision-reweighting and perturbation-regularization perspectives, which is consistent with the more complete and spatially coherent predictions observed in Figure 4. This interpretation is also supported by the ablation results in Table 3, where CFSR and SADP individually improve the baseline and their combination yields a further performance gain.

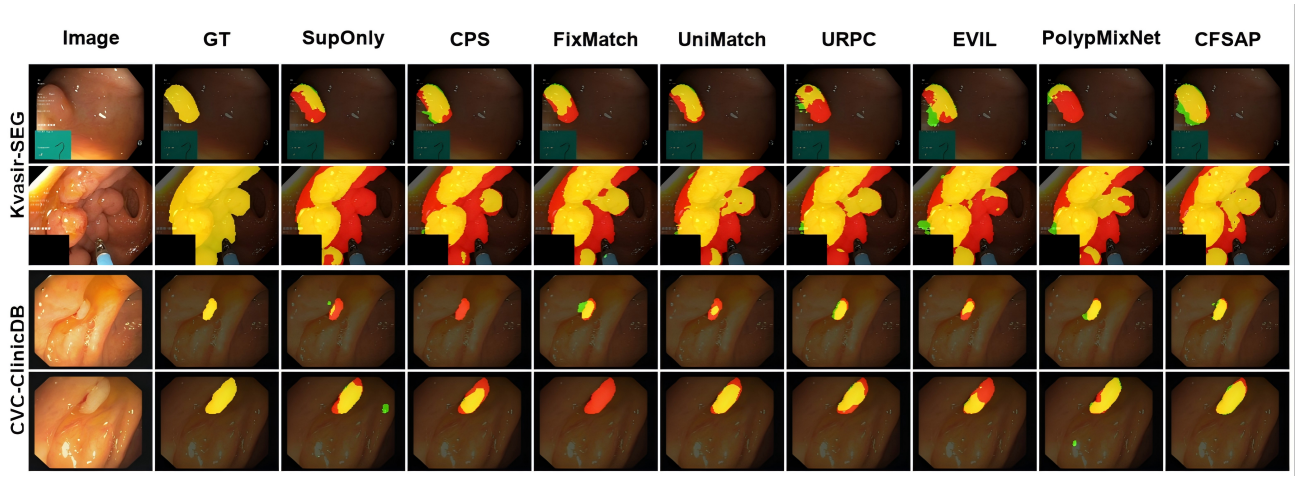


Figure 4. Visual comparison of segmentation results obtained by different methods under the 1/8 labeled ratio. Red regions indicate ground-truth pixels not covered by the predictions, green regions indicate predicted pixels outside the ground-truth regions, and yellow regions indicate the overlap between the predictions and ground-truth annotations.

Table 3. Ablation study of different components under the 1/8 labeled ratio.

Settings			Kvasir-SEG		CVC-ClinicDB	
Baseline	CFSR	SADP	Dice (%)	IoU (%)	Dice (%)	IoU (%)
✓	–	–	83.76	72.05	89.07	80.29
✓	✓	–	84.50	73.16	90.41	82.50
✓	–	✓	84.51	73.18	90.36	82.41
✓	✓	✓	**85.14**	**74.13**	**91.04**	**83.55**

4.3. Ablation study

4.3.1. Ablation study of different components

To evaluate the effectiveness of the individual components in the proposed framework, ablation experiments were conducted on Kvasir-SEG and CVC-ClinicDB using the 1/8 labeled ratio. UniMatch [17] was adopted as the baseline and evaluated under the same training configuration. CFSR, SADP, and their combination were then added to analyze the individual contributions and complementary effects of the two training strategies.

As shown in Table 3, the baseline achieved Dice and IoU scores of 83.76% and 72.05%, respectively, on Kvasir-SEG. Adding SADP alone increased these scores to 84.51% and 73.18%, while adding CFSR alone increased them to 84.50% and 73.16%. Both CFSR and SADP improved performance relative to the baseline, indicating that each strategy contributes to semi-supervised polyp segmentation. When CFSR and SADP were introduced together, the model achieved a Dice score of 85.14% and an IoU score of 74.13%, exceeding the baseline by 1.38 and 2.08 percentage points, respectively, and outperforming either strategy used alone.

A similar trend was observed on CVC-ClinicDB. The baseline achieved Dice and IoU scores of 89.07% and 80.29%, respectively. Adding SADP alone increased the scores to 90.36% and 82.41%, whereas adding CFSR alone increased them to 90.41% and 82.50%. When both strategies were used, the scores further increased to 91.04% and 83.55%, exceeding the baseline by 1.97 and 3.26 percentage points, respectively. These results indicate that both strategies provide effective complementary constraints for weak-to-strong consistency training across different datasets.

A comparison of the ablation results further shows that CFSR and SADP individually yielded gains on both datasets, with similar magnitudes of improvement, indicating that both strategies contribute to segmentation performance. Their combination achieved the best results on both datasets. Importantly, the two strategies intervene at different levels of weak-to-strong consistency learning. CFSR acts on pixel-level supervision by redistributing the contributions of low-confidence regions according to spatial cross-view feature discrepancies, whereas SADP acts on sample-level feature regularization by dynamically adjusting perturbation strength according to overall cross-view feature stability. Their joint improvement therefore reflects complementary effects on low-confidence supervision and dynamic feature perturbation, rather than redundancy between two constraints serving the same role.

4.3.2. Comparison of different perturbation-strength assignment strategies

To further validate the effectiveness of the stability-aware perturbation-strength assignment in SADP, we compared three settings, namely, Fixed, Random, and SADP, under the 1/8 labeled ratio. To eliminate the influence of CFSR, CFSR was disabled in this experiment. Feature noise injection, feature masking, and all other training configurations were kept unchanged, with only the perturbation-strength assignment strategy being varied. Specifically, Fixed applies the same fixed perturbation strength to all unlabeled samples. Random first computes the dynamic perturbation strength of each sample in the current batch using SADP and then randomly shuffles the correspondence between the perturbation strengths and samples. This preserves the values and distribution of the perturbation strengths while removing their association with sample stability. SADP assigns a corresponding perturbation strength to each sample based on its cross-view feature similarity.

As shown in Table 4, SADP achieved higher Dice and IoU scores than both Fixed and Random on the two datasets. It is worth noting that the Random setting preserves the same perturbation-strength values and their within-batch distribution as SADP, while randomly shuffling the correspondence between perturbation strengths and samples. Therefore, the performance difference between Random and SADP cannot be attributed simply to the use of varying perturbation strengths; it also reflects the additional effect of matching perturbation strength to the cross-view feature stability of the corresponding sample. Together with the Fixed setting, these results indicate that the improvement

achieved by SADP is associated not only with dynamic perturbation itself but also with the use of sample-level stability information to guide perturbation-strength assignment.

Table 4. Comparison of different perturbation-strength assignment strategies under the 1/8 labeled ratio.

Strategy	Kvasir-SEG		CVC-ClinicDB	
	Dice (%)	IoU (%)	Dice (%)	IoU (%)
Fixed	84.15	72.64	89.83	81.53
Random	84.28	72.83	90.01	81.84
SADP	**84.51**	**73.18**	**90.36**	**82.41**

4.3.3. Fine-grained ablation of internal design choices

To further isolate the contribution of individual design choices within CFSR and SADP, we conducted fine-grained ablation experiments on Kvasir-SEG under the 1/8 labeled setting. During the internal ablation of CFSR, SADP was disabled, whereas CFSR was disabled when evaluating the internal components of SADP, so that the contribution of each design choice could be examined independently. For CFSR, the "w/o class partition" variant retained the exponential discrepancy weighting but normalized the weights over all low-confidence pixels without class-conditional partitioning, whereas the "w/o exponential weighting" variant retained the class-conditional partitioning but assigned uniform normalized weights within each nonempty class-conditional low-confidence region. For SADP, the "w/o masking" variant retained only the stability-aware feature noise-injection branch, whereas the "w/o noise" variant retained only the stability-aware feature-masking branch, with the same stability-dependent perturbation-strength assignment retained in both cases. For the similarity-metric comparison, both CFSR and SADP were enabled and only the cross-view feature similarity metric was changed. The results are summarized in Table 5.

Table 5. Fine-grained ablation of the internal design choices on Kvasir-SEG under the 1/8 labeled setting.

Module / Comparison	Variant	Dice (%)	IoU (%)
CFSR	Baseline	83.76	72.05
	w/o class partition	84.12	72.59
	w/o exponential weighting	84.26	72.80
	Full CFSR	**84.50**	**73.16**
SADP	Baseline	83.76	72.05
	w/o masking	84.09	72.55
	w/o noise	84.24	72.77
	Full SADP	**84.51**	**73.18**
Similarity metric	Euclidean-based similarity	84.57	73.27
	Pearson correlation coefficient	84.89	73.75
	Cosine similarity	**85.14**	**74.13**

For CFSR, removing the class-conditional partitioning strategy reduced the Dice score from 84.50% to 84.12%. In this variant, low-confidence pixels assigned to different predicted classes shared the same normalization space, whereas the exponential discrepancy weighting itself was retained. The performance decrease indicates that separating the normalization according to the predicted classes contributes to the effectiveness of CFSR. When the exponential discrepancy weighting was replaced by uniform weighting while retaining the class-conditional partitioning, the Dice score decreased to 84.26%. This result shows that assigning different relative contributions to low-confidence pixels according to their cross-view feature discrepancies provides an additional benefit beyond class-conditional partitioning alone. The full CFSR achieved the highest Dice score of 84.50%, indicating that both design choices contribute to its overall performance.

For SADP, removing the feature-masking branch decreased the Dice score from 84.51% to 84.09%, whereas removing the feature noise-injection branch resulted in a Dice score of 84.24%. Both reduced variants still outperformed the baseline, indicating that each perturbation branch can provide a useful regularization effect independently. The complete SADP achieved the best performance, with a Dice score of 84.51% and an IoU score of 73.18%. These results show that both feature noise injection and feature masking contribute to SADP and provide complementary empirical benefits when used jointly.

We further compared cosine similarity with the Pearson correlation coefficient and a Euclidean-distance-based similarity formulation while keeping all other training settings unchanged. Pearson correlation was computed between the two channel-wise feature vectors at each spatial location after subtracting the channel mean from each vector. For the Euclidean-distance-based formulation, the feature vectors at each spatial location were first L2-normalized, and the similarity score was defined as

$$S_{\mathrm{euc}}(h, w) = 1 - \left\| \hat{Z}^w(h, w) - \hat{Z}^s(h, w) \right\|_2 ,$$

where $\hat{Z}^w$ and $\hat{Z}^s$ denote the L2-normalized weakly and strongly augmented feature vectors, respectively. Since the Euclidean distance between two unit vectors lies in $[0, 2]$, S_{euc} lies in $[-1, 1]$, keeping its numerical range compatible with the similarity score used by CFSR and SADP. The subsequent reweighting and perturbation-strength estimation procedures were kept unchanged for all metrics.

As shown in Table 5, cosine similarity achieved a Dice score of 85.14%, compared with 84.89% for the Pearson correlation coefficient and 84.57% for the Euclidean-distance-based formulation. All three formulations yielded competitive results, while cosine similarity achieved the best performance among the tested metrics. Cosine similarity directly characterizes the directional agreement between cross-view feature vectors and is insensitive to uniform scaling of their magnitudes, which is suitable for the similarity-guided reweighting and stability estimation used in CFSAP. These results provide empirical support for retaining cosine similarity in the proposed framework.

4.4. Hyperparameter sensitivity analysis

To investigate the sensitivity of CFSAP to its key module-specific hyperparameters, we conducted a one-factor-at-a-time analysis on Kvasir-SEG under the 1/8 labeled setting. Three parameters were examined: the confidence threshold τ, the perturbation scaling coefficient λ_α, and the masking-interval offset δ. Specifically, τ was varied among $\{0.7, 0.8, 0.9\}$, λ_α among $\{0.10, 0.15, 0.20\}$, and δ among $\{0.05, 0.10, 0.15\}$. When one parameter was varied, the remaining parameters and all other training

settings were kept fixed at their default values. The resulting Dice scores are reported in Table 6.

Table 6. Sensitivity analysis of the key module-specific hyperparameters on Kvasir-SEG under the 1/8 labeled setting.

Parameter	Setting 1	Setting 2	Setting 3
τ	0.7 / 83.95%	0.8 / **85.14%**	0.9 / 83.70%
λ_α	0.10 / 84.51%	0.15 / **85.14%**	0.20 / 84.44%
δ	0.05 / 84.61%	0.10 / **85.14%**	0.15 / 84.75%

For the confidence threshold τ, the Dice scores at 0.7, 0.8, and 0.9 are 83.95%, 85.14%, and 83.70%, respectively, with the best performance obtained at $\tau = 0.8$. In CFSAP, τ does not simply determine whether low-confidence predictions are discarded; instead, it controls the partition of unlabeled pixels between the primary high-confidence consistency supervision and the supplementary CFSR constraint. When τ is reduced to 0.7, some moderately uncertain pixels are directly incorporated into the primary consistency branch rather than being selectively reweighted by CFSR. Conversely, increasing τ to 0.9 transfers more relatively reliable pixels from the primary consistency branch to the supplementary CFSR branch. The results therefore suggest that $\tau = 0.8$ provides a more suitable balance between these two complementary supervision mechanisms among the tested settings.

For the perturbation scaling coefficient λ_α, the Dice scores are 84.51%, 85.14%, and 84.44% for $\lambda_\alpha = 0.10$, 0.15, and 0.20, respectively. According to Equation (3.11), λ_α controls the overall range of the dynamically generated perturbation strength. Because $\bar{s} \in [-1, 1]$, the tested values correspond to maximum perturbation strengths of 0.20, 0.30, and 0.40, respectively. Smaller values therefore restrict the perturbation range, whereas larger values allow stronger perturbations. The intermediate setting $\lambda_\alpha = 0.15$ achieves the highest Dice score, while the overall variation across the tested range is limited to 0.70 percentage points, indicating relatively low sensitivity within this interval.

For the masking-interval offset δ, the Dice scores obtained with $\delta = 0.05$, 0.10, and 0.15 are 84.61%, 85.14%, and 84.75%, respectively. The parameter δ determines the half-width of the sampling interval $\mathcal{U}(1 - \alpha - \delta,\ 1 - \alpha + \delta)$ used to generate the masking adjustment coefficient. A smaller δ restricts stochastic variation in the masking threshold, whereas a larger δ broadens the sampling interval and increases the variability of the masking operation. The performance varies by only 0.53 percentage points over the tested range, with $\delta = 0.10$ achieving the highest Dice score, suggesting that the masking mechanism is relatively insensitive to δ within this interval.

Overall, the three parameters exhibit different levels of sensitivity. The confidence threshold τ has a comparatively larger influence because it directly changes the allocation of pixels between the primary high-confidence consistency branch and the supplementary CFSR branch, whereas λ_α and δ mainly control the strength and stochastic range of feature perturbations and show smaller variations over the tested intervals. These results provide empirical support for the default settings $\tau = 0.8$, $\lambda_\alpha = 0.15$, and $\delta = 0.10$ within the examined ranges. They also provide additional evidence regarding the risk of low-confidence supervision: simply enlarging the region handled by CFSR does not continuously improve performance. This observation is consistent with the design of CFSR as a supplementary, selectively weighted constraint rather than a replacement for high-confidence consistency supervision.

4.5. Computational efficiency analysis

To evaluate inference-time computational efficiency, we compared CFSAP with the baseline used in the ablation experiments. The baseline uses the same DeepLabV3+ segmentation network with a ResNet-50 backbone but excludes CFSR and SADP. Table 7 summarizes the model size, theoretical computation, latency, and throughput at an input resolution of 256×256.

Table 7. Computational efficiency comparison at an input resolution of 256×256.

Method	Params (M) ↓	FLOPs (G) ↓	Latency (ms) ↓	FPS ↑
Baseline	40.47	38.29	5.14	194.71
CFSAP	40.47	38.29	5.10	196.19

Note: Inference speed was measured on an NVIDIA GeForce RTX 4090 with a batch size of 1, an input resolution of 256×256, and FP32 precision. After 100 warm-up iterations, 1000 forward passes were timed. The reported latency and throughput in frames per second (FPS) were obtained by averaging three repeated timing measurements using the same trained model. Data loading and preprocessing were excluded. Floating-point operations (FLOPs) were estimated as twice the number of multiply–accumulate operations (MACs).

Table 7 shows that both the Baseline and CFSAP contain 40.47 M parameters and require 38.29 G FLOPs at an input resolution of 256×256. CFSAP achieves an average inference latency of 5.10 ms per image and a throughput of 196.19 FPS, which are comparable to the corresponding Baseline values of 5.14 ms and 194.71 FPS, respectively. The minor numerical difference in runtime reflects normal variation in GPU-based timing rather than a difference in network complexity. This result is consistent with the design of CFSAP: CFSR and SADP are used only during training and do not introduce additional learnable modules or inference-time branches. Consequently, CFSAP improves segmentation performance without increasing the model size or theoretical computational complexity during inference.

5. Conclusions, limitations, and future work

This study presented CFSAP, a semi-supervised framework for polyp segmentation under limited pixel-level annotations. CFSAP extends weak-to-strong consistency learning by exploiting cross-view feature information at both spatial and sample levels. Cross-view feature similarity reweighting uses spatial feature discrepancies between weakly and strongly augmented views to selectively reweight supplementary supervision in low-confidence regions, while stability-aware dynamic perturbation uses sample-level cross-view representation stability to dynamically adjust feature perturbation strength. Through these two uses of the cross-view feature relationship, CFSAP improves the utilization of unlabeled colonoscopy images from both supervision reweighting and dynamic perturbation perspectives.

Experiments on Kvasir-SEG and CVC-ClinicDB under labeled-data ratios of 1/4, 1/8, and 1/16 showed that CFSAP consistently achieved higher Dice and IoU scores than the compared semi-supervised methods. The additional precision, recall, and HD95 results provided a complementary evaluation of foreground prediction and contour discrepancy. The ablation results further confirmed that CFSR and SADP provide complementary improvements. Moreover, assigning perturbation strengths according to cross-view feature stability was more effective than using fixed strengths or randomly matching perturbation strengths to samples. The computational-efficiency analysis further showed that CFSAP preserves the parameter count and FLOPs of the baseline and

exhibits comparable inference latency and throughput, consistent with the training-only use of CFSR and SADP.

The current study still has several limitations, mainly concerning the scope of experimental and clinical validation. First, the evaluation was limited to two public datasets of static colonoscopy images and used DeepLabV3+ with a ResNet-50 backbone as the experimental testbed. Therefore, the cross-architecture generalization of CFSR and SADP, as well as the generalization of CFSAP across centers, imaging devices, and clinical environments, has not yet been systematically evaluated. Second, the current comparison focuses on a representative set of semi-supervised segmentation methods evaluated under a unified protocol and does not constitute an exhaustive benchmark of all recent semi-supervised medical segmentation approaches. The present evaluation is also conducted mainly at the dataset level and does not separately quantify challenging case categories, such as tiny polyps, low-contrast lesions, and lesions with blurred boundaries, under consistently defined case attributes.

Future work will proceed in a staged manner according to the proximity of each validation setting to the current task. As an immediate priority, we will first perform external cross-center validation using colonoscopy data acquired from different institutions, imaging devices, and acquisition conditions. In parallel, the experimental benchmark will be extended under unified settings to include additional convolutional and transformer-based segmentation architectures, a broader set of recent semi-supervised methods, and attribute-based subgroup analyses of challenging cases. At the clinical-validation stage, and subject to appropriate ethical approval and clinical collaboration, reader studies involving gastroenterologists with different levels of experience will be considered to compare lesion delineation with and without CFSAP-generated segmentation assistance, using delineation time and inter-reader agreement as evaluation measures. Beyond polyp segmentation, we will further examine whether the proposed cross-view supervision reweighting and dynamic perturbation strategies remain effective for other endoscopic lesion-segmentation tasks. In the longer term, transferability to other medical imaging modalities will be investigated by re-evaluating the weak/strong augmentation design and the assumptions underlying cross-view feature stability according to modality-specific imaging characteristics.

Use of AI tools declaration

The authors acknowledge the use of AI-based language tools solely for grammatical refinement and sentence polishing. All research design, experiments, data analysis, and conclusions are the authors' own.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (No. 62502064), the Dalian University interdisciplinary project (No. DLUXK-2025-QN-010), and Joint Plan of Liaoning Province Science and Technology Plan (No. 2025JH2/101800422).

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

1. Bray F, Laversanne M, Sung H, et al. (2024) Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 74: 229–263. https://doi.org/10.3322/caac.21834
2. Corley DA, Jensen CD, Marks AR, et al. (2014) Adenoma detection rate and risk of colorectal cancer and death. *N Engl J Med* 370: 1298–1306. https://doi.org/10.1056/NEJMoa1309086
3. Qayoom A, Xie J, Ali H (2025) Polyp segmentation in medical imaging: challenges, approaches and future directions. *Artif Intell Rev* 58: 169. https://doi.org/10.1007/s10462-025-11173-2
4. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation, *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965
5. Ronneberger O, Fischer P, Brox T (2015) U-Net: convolutional networks for biomedical image segmentation, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Springer, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
6. Chen LC, Zhu Y, Papandreou G, et al. (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation, *Computer Vision – ECCV 2018*, Springer, 833–851. https://doi.org/10.1007/978-3-030-01234-2_49
7. Huang Z, Xie F, Qing W, et al. (2024) MGF-net: multi-channel group fusion enhancing boundary attention for polyp segmentation. *Med Phys* 51: 407–418. https://doi.org/10.1002/mp.16584
8. Liu Y, Shen X, Lyu Y (2024) MCA-Net: multi-cascade attention network for polyp segmentation. *Multimed Tools Appl* 83: 33713–33730. https://doi.org/10.1007/s11042-023-16805-9
9. Jiang L, Hui Y, Fei Y, et al. (2024) Improving polyp segmentation with boundary-assisted guidance and cross-scale interaction fusion transformer network. *Processes* 12: 1030. https://doi.org/10.3390/pr12051030
10. Yue G, Li Y, Jiang W, et al. (2024) Boundary refinement network for colorectal polyp segmentation in colonoscopy images. *IEEE Signal Process Lett* 31: 954–958. https://doi.org/10.1109/LSP.2024.3378106
11. Huang X, Wang L, Jiang S, et al. (2024) DHAFormer: dual-channel hybrid attention network with transformer for polyp segmentation. *PLoS One* 19: e0306596. https://doi.org/10.1371/journal.pone.0306596
12. Wang Z, Liu Z, Yu J, et al. (2024) Multi-scale nested UNet with transformer for colorectal polyp segmentation. *J Appl Clin Med Phys* 25: e14351. https://doi.org/10.1002/acm2.14351
13. Ran L, Li Y, Liang G, et al. (2025) Pseudo labeling methods for semi-supervised semantic segmentation: a review and future perspectives. *IEEE Trans Circuits Syst Video Technol* 35: 3054–3080. https://doi.org/10.1109/TCSVT.2024.3508768
14. Souly N, Spampinato C, Shah M (2017) Semi supervised semantic segmentation using generative adversarial network, *2017 IEEE International Conference on Computer Vision (ICCV)*, IEEE, 5689–5697. https://doi.org/10.1109/ICCV.2017.606

15. Hung WC, Tsai YH, Liou YT, et al. (2018) Adversarial learning for semi-supervised semantic segmentation, *Proceedings of the British Machine Vision Conference*, BMVA Press, 65.

16. Sohn K, Berthelot D, Carlini N, et al. (2020) FixMatch: simplifying semi-supervised learning with consistency and confidence. *Adv Neural Inf Process Syst* 33: 596–608.

17. Yang L, Qi L, Feng L, et al. (2023) Revisiting weak-to-strong consistency in semi-supervised semantic segmentation, *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 7236–7246. https://doi.org/10.1109/CVPR52729.2023.00699

18. Yuan J, Liu Y, Shen C, et al. (2021) A simple baseline for semi-supervised semantic segmentation with strong data augmentation, *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, 8209–8218. https://doi.org/10.1109/ICCV48922.2021.00812

19. Xie S, Wang H, Niu Z, et al. (2026) S^2Match: revisiting weak-to-strong consistency from a semantic similarity perspective for semi-supervised medical image segmentation. *IEEE J Biomed Health Inform* 30: 5325–5338. https://doi.org/10.1109/JBHI.2025.3639673

20. Yin J, Chen Y, Zheng Z, et al. (2025) Uncertainty-participation context consistency learning for semi-supervised semantic segmentation, *ICASSP 2025 – 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 1–5. https://doi.org/10.1109/ICASSP49660.2025.10889133

21. Li S, He Y, Zhang W, et al. (2023) CFCG: semi-supervised semantic segmentation via cross-fusion and contour guidance supervision, *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, 16302–16312. https://doi.org/10.1109/ICCV51070.2023.01498

22. Wu Z and Cui J (2024) Image-feature weak-to-strong consistency: an enhanced paradigm for semi-supervised learning, *Computer Vision – ECCV 2024*, Springer, 372–388. https://doi.org/10.1007/978-3-031-72633-0_21

23. Jia X, Shen Y, Yang J, et al. (2024) PolypMixNet: enhancing semi-supervised polyp segmentation with polyp-aware augmentation. *Comput Biol Med* 170: 108006. https://doi.org/10.1016/j.compbiomed.2024.108006

24. Guo P, Liu G, Liu H (2024) SemiPolypSeg: leveraging cross-pseudo supervision and contrastive learning for semi-supervised polyp segmentation. *Appl Sci* 14: 7852. https://doi.org/10.3390/app14177852

25. Wang A, Ma H, Bai L, et al. (2025) PedSemiSeg: pedagogy-inspired semi-supervised polyp segmentation. *Comput Med Imaging Graph* 124: 102591. https://doi.org/10.1016/j.compmedimag.2025.102591

26. Li Y, Zhu Z, Zhang Y, et al. (2025) Boost the inference with co-training: a depth-guided mutual learning framework for semi-supervised medical polyp segmentation, *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 10394–10403. https://doi.org/10.1109/CVPR52734.2025.00972

27. Chen J, Zhang H, Gong M, et al. (2024) Collaborative compensative transformer network for salient object detection. *Pattern Recognit* 154: 110600. https://doi.org/10.1016/j.patcog.2024.110600

28. Xu T, Jiang J, Cai L, et al. (2025) Salient object detection with non-local feature enhancement and edge reconstruction. *Sci Rep* 15: 397. https://doi.org/10.1038/s41598-024-84680-x

29. Lee DH (2013) Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks, *ICML 2013 Workshop on Challenges in Representation Learning*, ICML, 1–6.

30. Chen X, Yuan Y, Zeng G, et al. (2021) Semi-supervised semantic segmentation with cross pseudo supervision, *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2613–2622. https://doi.org/10.1109/CVPR46437.2021.00264

31. Tarvainen A and Valpola H (2017) Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Adv Neural Inf Process Syst* 30: 1195–1204.

32. Ouali Y, Hudelot C, Tami M (2020) Semi-supervised semantic segmentation with cross-consistency training, *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 12671–12681. https://doi.org/10.1109/CVPR42600.2020.01269

33. Luo X, Wang G, Liao W, et al. (2022) Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency. *Med Image Anal* 80: 102517. https://doi.org/10.1016/j.media.2022.102517

34. Chen Y, Yang Z, Shen C, et al. (2024) Evidence-based uncertainty-aware semi-supervised medical image segmentation. *Comput Biol Med* 170: 108004. https://doi.org/10.1016/j.compbiomed.2024.108004

35. Tang Y, Wang S, Qu Y, et al. (2023) Consistency and adversarial semi-supervised learning for medical image segmentation. *Comput Biol Med* 161: 107018. https://doi.org/10.1016/j.compbiomed.2023.107018

36. Wang Y, Wang H, Shen Y, et al. (2022) Semi-supervised semantic segmentation using unreliable pseudo-labels, *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 4238–4247. https://doi.org/10.1109/CVPR52688.2022.00421

37. Huang S, Ge Y, Liu D, et al. (2025) Rethinking copy-paste for consistency learning in medical image segmentation. *IEEE Trans Image Process* 34: 1060–1074. https://doi.org/10.1109/TIP.2025.3536208

38. Bai Y, Chen D, Li Q, et al. (2023) Bidirectional copy-paste for semi-supervised medical image segmentation, *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 11514–11524. https://doi.org/10.1109/CVPR52729.2023.01108

39. Yin J, Chen T, Pei G, et al. (2025) Semi-supervised semantic segmentation with multi-constraint consistency learning. *IEEE Trans Multimedia* 27: 6449–6461. https://doi.org/10.1109/TMM.2025.3586111

40. Yin J, Yan S, Chen T, et al. (2024) Class probability space regularization for semi-supervised semantic segmentation. *Comput Vis Image Underst* 249: 104146. https://doi.org/10.1016/j.cviu.2024.104146

41. Sun B, Yang Y, Zhang L, et al. (2024) CorrMatch: label propagation via correlation matching for semi-supervised semantic segmentation, *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 3097–3107. https://doi.org/10.1109/CVPR52733.2024.00299

42. Mai H, Sun R, Zhang T, et al. (2024) RankMatch: exploring the better consistency regularization for semi-supervised semantic segmentation, *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 3391–3401.

43. Assefa M, Naseer M, Ganapathi II, et al. (2025) DyCON: dynamic uncertainty-aware consistency and contrastive learning for semi-supervised medical image segmentation, *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 30850–30860. https://doi.org/10.1109/CVPR52734.2025.02873

44. Wang X, Bai H, Yu L, et al. (2024) Towards the uncharted: density-descending feature perturbation for semi-supervised semantic segmentation, *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 3303–3312.

45. Yang Y, Wang T, Li S (2026) Attention inverted feature perturbation for semi-supervised medical image segmentation. *Discov Comput* 29: 16. https://doi.org/10.1007/s10791-025-09896-5

46. Jha D, Smedsrud PH, Riegler MA, et al. (2020) Kvasir-SEG: a segmented polyp dataset, *MultiMedia Modeling*, Springer, 451–462. https://doi.org/10.1007/978-3-030-37734-2_37

47. Bernal J, Sanchez FJ, Fernandez-Esparrach G, et al. (2015) WM-DOVA maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians. *Comput Med Imaging Graph* 43: 99–111. https://doi.org/10.1016/j.compmedimag.2015.02.007

48. Kong X, Ren Z, Zhu H, et al. (2025) Semi-CLMT: a semi-supervised framework for medical image segmentation. *IEEE Trans Emerg Top Comput Intell* 9: 3938–3950. https://doi.org/10.1109/TETCI.2025.3554787

49. He K, Zhang X, Ren S, et al. (2016) Deep residual learning for image recognition, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 770–778. https://doi.org/10.1109/CVPR.2016.90



AIMS Press

© 2026 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0)