Computerized Medical Imaging and Graphics 124 (2025) 102591

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

PedSemiSeg: Pedagogy-inspired semi-supervised polyp segmentation ☆

An Wang [a,b], Haoyu Ma [a], Long Bai [a,b], Yanan Wu [c], Mengya Xu [a,b], Yang Zhang [d], Mobarakol Islam [e,f], Hongliang Ren [a,b,*]

[a] Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong Special Administrative Region of China
[b] CUHK Shenzhen Research Institute, Shenzhen, China
[c] School of Health Management, China Medical University, Shenyang, China
[d] School of Mechanical Engineering, Hubei University of Technology, Wuhan, China
[e] UCL Hawkes Institute, University College London, London, UK
[f] Dept of Medical Physics & Biomedical Engineering, University College London, UK

ARTICLE INFO

Dataset link: https://github.com/lofrienger/PedSemiSeg

Keywords:
Semi-supervised learning
Polyp segmentation
Consistency regularization
Negative learning
Pedagogy-inspired learning
Computer-aided diagnosis

ABSTRACT

Recent advancements in deep learning techniques have contributed to developing improved polyp segmentation methods, thereby aiding in the diagnosis of colorectal cancer and facilitating automated surgery like endoscopic submucosal dissection (ESD). However, the scarcity of well-annotated data poses challenges by increasing the annotation burden and diminishing the performance of fully-supervised learning approaches. Additionally, distribution shifts due to variations among patients and medical centers require the model to generalize well during testing. To address these concerns, we present **PedSemiSeg**, a pedagogy-inspired semi-supervised learning framework designed to enhance polyp segmentation performance with limited labeled training data. In particular, we take inspiration from the pedagogy used in real-world educational settings, where teacher feedback and peer tutoring are both crucial in influencing the overall learning outcome. Expanding upon this concept, our approach involves supervising the outputs of the strongly augmented input (the students) using the pseudo and complementary labels crafted from the output of the weakly augmented input (the teacher) in both positive and negative learning manners. Additionally, we introduce reciprocal peer tutoring among the students, guided by respective prediction entropy. With these holistic learning processes, we aim to achieve consistent predictions for various versions of the same input and maximize the utilization of the abundant unlabeled data. Experimental results on two public datasets demonstrate the superiority of our method in polyp segmentation across various labeled data ratios. Furthermore, our approach exhibits excellent generalization capabilities on external unseen multi-center datasets, highlighting its broader clinical significance in practical applications during deployment.

1. Introduction

Polyp segmentation is a vital component in computer-aided diagnosis (CAD) systems used to detect and characterize colorectal polyps (Summers et al., 2002), which are often precursors to colorectal cancer. Identifying and removing polyps at an early stage greatly lowers the risk of developing colorectal cancer, making accurate and efficient polyp segmentation an essential task in clinical practice (Jha et al., 2021). Conventional methods for polyp segmentation heavily depend on time-consuming manual annotation by expert clinicians, making them susceptible to inter-observer differences. Additionally, the escalating quantity of screening colonoscopies has led to an enormous volume of medical imaging data, rendering it unfeasible for clinicians to manually segment each polyp (Zhao et al., 2021; Wang and Zheng, 2024). Therefore, there is a growing need for automated or semi-automated polyp segmentation techniques to assist clinicians in their diagnostic workflow.

The advent of deep learning has revolutionized medical image analysis (Roy et al., 2022; Lin et al., 2022b), including polyp segmentation. Convolutional Neural Networks (CNNs) have demonstrated tremendous success in diverse computer vision tasks, motivating researchers to explore their potential in polyp segmentation (Fan et al., 2020; Shen

☆ This work was supported by Hong Kong Research Grants Council (RGC) Collaborative Research Fund (C4026-21G), General Research Fund (GRF 14211420 & 14203323), Shenzhen-Hong Kong-Macau Technology Research Programme (Type C) STIC Grant SGDX20210823103535014 (202108233000303).

* Corresponding author at: Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong Special Administrative Region of China.
E-mail addresses: wa09@link.cuhk.edu.hk (A. Wang), hkmahaoyu@gmail.com (H. Ma), b.long@ieee.org (L. Bai), wuyanan.cmu@vip.163.com (Y. Wu), mengya@u.nus.edu (M. Xu), yzhangcst@hbut.edu.cn (Y. Zhang), mobarakol.islam@ucl.ac.uk (M. Islam), hlren@ieee.org (H. Ren).

https://doi.org/10.1016/j.compmedimag.2025.102591
Received 6 April 2025; Received in revised form 24 May 2025; Accepted 12 June 2025
Available online 1 July 2025

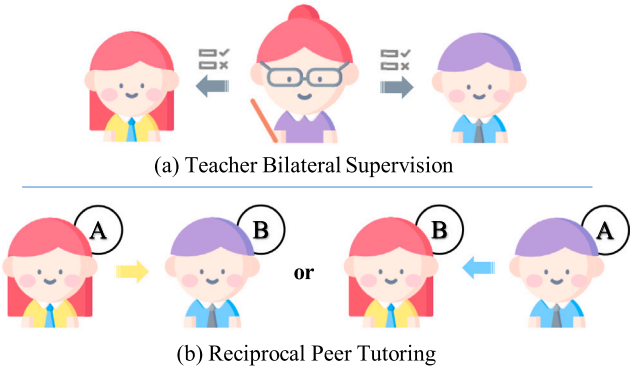


Fig. 1. Two effective pedagogical activities in human education. In (a) Teacher Bilateral Supervision, the teacher provides instructions on both correct and incorrect matters. For (b) Reciprocal Peer Tutoring, the student collaboratively learns from the high-achieving peer. Grade Ⓐis higher than Ⓑ. Arrows indicate the direction of supervision.

et al., 2021; Zhang et al., 2022; Huang et al., 2022). Recently, several works (Dong et al., 2021; Cai et al., 2022; Lin et al., 2022a) have also explored modern transformer-based approaches. However, training deep learning models for polyp segmentation typically demands a substantial amount of annotated polyp images. Unfortunately, such annotated data is often limited in availability due to the labor-intensive and time-consuming nature of the manual annotation process. Besides, variations among data sources, such as variances between patients and discrepancies in imaging protocols, can introduce data shifts that further complicate the generalization of the trained model during deployment (Ji et al., 2024b; Wang et al., 2023a; Hu et al., 2022). Consequently, the scarcity of annotated data and the presence of data shifts pose significant challenges in the development of robust and accurate polyp segmentation algorithms.

In light of these challenges, annotation-efficient learning approaches (Mei et al., 2025; Ji et al., 2024a; Li et al., 2024a) have gained significant attention as promising solutions. These encompass various strategies, including semi-supervised learning (SSL) (Wu et al., 2023b; He et al., 2023; Wang and Li, 2024; Xiong et al., 2024; Du et al., 2025; Zhang and Zhang, 2025), which leverages a small labeled dataset alongside a large volume of unlabeled data; weakly-supervised learning (WSL) (Wang et al., 2023b; Wei et al., 2023; Long et al., 2025; Zhao et al., 2025), which utilizes coarser or less precise forms of annotation; and barely-supervised learning (BSL) (Lucas et al., 2022; Wu et al., 2023a), which operates with even more sparsely annotated data. By capitalizing on partially or sparsely annotated data, these algorithms aim to enhance label efficiency in developing cost-friendly polyp segmentation models. While demonstrating promising results, these methods still exhibit a performance gap compared to fully-supervised models, particularly concerning generalization to out-of-distribution domains.

As depicted in Fig. 1, the concurrent presence of (a) Teacher Bilateral Supervision and (b) Reciprocal Peer Tutoring is frequently observed and effective in real-world educational environments. This approach substantially enhances the overall learning outcomes. Inspired by such classroom practices, we propose a holistic **pedagogy-inspired semi-supervised learning** framework called **PedSemiSeg** for label-efficient polyp segmentation. Specifically, we begin by revisiting the fundamental design principle of consistency regularization with weak-to-strong perturbation in semi-supervised learning and introduce the concept of geometry-to-intensity augmentation to generate diverse variations of the same input. Going beyond traditional methods, PedSemiSeg leverages both positive and negative learning from the weakly augmented branch and implements reciprocal peer tutoring among the strongly augmented branches. This fosters more efficient and effective utilization of the unlabeled data and ultimately boosts overall performance on the in-distribution SUN-SEG (Ji et al., 2022) and Kvasir-SEG (Jha et al., 2020) datasets and out-of-distribution Polyp-Gen (Ali et al., 2023) dataset. Our main contributions are as follows:

- We design the sequential geometry-to-intensity augmentation, implemented in a weak-to-strong manner, to promote more comprehensive consistency regularization on unlabeled data.
- Drawing inspiration from the bilateral guidance in human education, where teachers instruct on both correct and incorrect aspects, we generate pseudo and complementary labels from the weakly augmented branch to facilitate both positive and negative learning for the strongly augmented counterparts.
- Motivated by collaborative learning among students, we introduce reciprocal peer tutoring between two strongly augmented branches with the learning direction decided by their prediction uncertainty.
- Holistically, we propose **PedSemiSeg**, a pedagogy-inspired semi-supervised learning method for label-efficient polyp segmentation. Our method exhibits superior performance on both in-domain and external unseen datasets, thereby demonstrating its applicability for real-world computer-aided diagnosis and intervention.

2. Related works

Semi-supervised learning (SSL) has gained considerable attention in medical image analysis, primarily due to the time-consuming and labor-intensive nature of data annotation, particularly for intricate tasks such as registration (Zhu et al., 2021; Ma et al., 2017), segmentation (Cheplygina et al., 2019; Li et al., 2023a), and 3D reconstruction (Shi et al., 2021, 2023). SSL utilizes unlabeled data to enhance model performance when labeled data is scarce, thereby enhancing annotation efficiency. In this context, various SSL strategies have emerged, including consistency regularization (CR), self-training, adversarial learning, and uncertainty-based methods. These strategies allow for the exploitation of the wealth of unlabeled images while integrating the information from the limited labeled data.

2.1. Consistency regularization

Consistency regularization (CR) has been widely explored and adopted in semi-supervised segmentation (Yang et al., 2023a; Guo et al., 2020; Luo et al., 2022b; Jia et al., 2024). It allows unlabeled data to regularize model training by enforcing invariance to perturbations. In general computer vision, FixMatch (Sohn et al., 2020) is one of the pioneering works focusing on consistency regularization. Noteworthy subsequent approaches include Cross-consistency Training (CCT) (Ouali et al., 2020) and Cross Pseudo Supervision (CPS) (Chen et al., 2021). More recently, ShrinkMatch (Yang et al., 2023b) aims to learn from unlabeled data by excluding confused classes to enhance certainty, while UniMatch V2 (Yang et al., 2025) incorporates a complementary dropout module to unify image-level and feature-level augmentations for improved consistency learning. In the medical domain, numerous approaches have been developed to address specific challenges. These include multiple consistency supervision for OCT images (Lu et al., 2022), co-training between CNN and transformer on MRI cardiac images (Luo et al., 2022a), and cross-level contrastive learning on polyps and skin lesions (Zhao et al., 2022). Recently, CCL-MPC (Du et al., 2025) employed dual-branch networks and dual augmented views to enhance class diversity and facilitate multi-perspective consistency learning for skin lesion and polyp datasets.

2.2. Pseudo Labeling

Pseudo labeling (PL), like consistency regularization, is another essential SSL technique where the model's own high-confidence predictions on unlabeled data serve as training targets. For instance, Chaitanya et al. (2023) proposes a pseudo-label-based self-training framework and incorporates the local contrastive loss to efficiently explore the partially annotated datasets of cardiac and prostate anatomies. For fundus and prostate MRI segmentation, FSSL-DPL (Qiu et al., 2023) explores cross-domain scenarios and proposes a federated approach for generating and denoising pseudo labels. Recently, with the emergence of the foundation models (Bommasani et al., 2021), some works (Li et al., 2023b; Rahman et al., 2024) have also attempted to utilize the pretrained Segment Anything Model (SAM) (Kirillov et al., 2023) as a strong pseudo label generator.

2.3. Teacher–student framework

Teacher–student frameworks are a prominent and effective strategy in SSL, designed to improve learning from unlabeled data by having a "teacher" model guide the training of a "student" model (Tarvainen and Valpola, 2017). This paradigm often integrates principles from consistency regularization and pseudo-labeling. Several core methodologies characterize teacher–student learning in SSL. A foundational approach is the Mean Teacher model (Tarvainen and Valpola, 2017; Wang et al., 2022). In this setup, the teacher model's weights are an exponential moving average (EMA) of the student model's weights. The student is trained on perturbed inputs, while the teacher, with its more stable, averaged weights, provides reliable pseudo-labels or consistency targets. This helps to smooth the learning trajectory and reduce noise from self-generated supervision. Concepts from knowledge distillation (Hinton et al., 2015) are frequently adapted in teacher–student semi-supervised learning (Wang et al., 2024; Zhao et al., 2024; Shen et al., 2023). The student model can be trained to mimic the teacher's soft probability outputs (logits) or its intermediate feature representations on unlabeled data. This allows the student to capture richer, more nuanced information than what hard pseudo-labels alone can provide. Some frameworks extend the teacher–student concept to dual or multiple networks that act as peers, effectively teaching each other (Chen et al., 2021; Zheng et al., 2022; Yang et al., 2025). For instance, in Cross Pseudo Supervision (CPS) (Chen et al., 2021), two student networks generate pseudo-labels for each other, encouraging the learning of diverse features and improving overall robustness. These methodologies aim to enhance the quality and stability of supervision signals derived from unlabeled data, making the SSL process more effective, particularly when labeled data is limited. By leveraging a more knowledgeable or stable teacher, the student can learn more efficiently and generalize better.

2.4. More SSL strategies

In addition to consistency regularization, pseudo labeling, and the teacher–student framework, several other SSL strategies have been developed. Negative learning involves identifying the least likely classes to provide indirect supervision signals (Kim et al., 2019; Chen et al., 2020; Yao et al., 2022). For example, ACTION (You et al., 2023) leverages negative samples with global-local contrastive pre-training and anatomical contrast fine-tuning. Other strategies, such as affinity learning (Wu et al., 2023b), contrastive learning (Wang et al., 2022; Basak and Yin, 2023), confidence learning (Xie et al., 2021), adversarial learning (Lei et al., 2022), multi-modality (Li et al., 2024b), and multi-task learning (Luo et al., 2021), have also been developed to mitigate label scarcity issues. Existing SSL works often hybridize several strategies to achieve optimal learning outcomes.

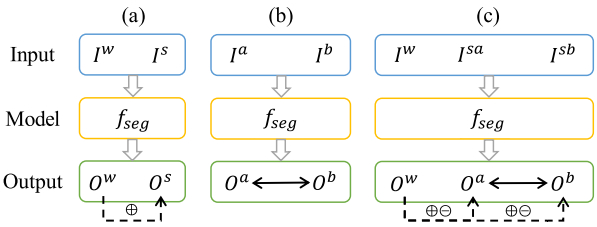


Fig. 2. Comparison of consistency regularization-based architectures. I and O represent the input image and output prediction. The superscripts w and s indicate weak and strong augmentation, while a and b represent different variants. The arrow "⇢" represents supervision from the weakly to the strongly augmented branches, with "⊕" and "⊖" denoting positive and negative learning, respectively. The bidirectional arrow "↔" stands for supervision between the strongly augmented branches. Our framework, as depicted in (c), extended upon (a) and (b), incorporating more comprehensive supervision and resulting in more effective consistency regularization.

2.5. Our pedagogy-inspired SSL framework

In this paper, drawing inspiration from principles of human teaching and learning, we introduce a novel pedagogy-inspired framework, dubbed **PedSemiSeg**, for label-efficient polyp segmentation. As demonstrated in Fig. 2, our framework extends upon previous consistency regularization-based architectures (Sohn et al., 2020; Yang et al., 2023a) by incorporating more comprehensive and multi-faceted forms of supervision to maximize the value extracted from unlabeled data. Specifically, common SSL architectures like FixMatch (Sohn et al., 2020) (Fig. 2(a)) primarily enforce consistency from a weakly augmented branch (teacher) to a strongly augmented branch (student) using pseudo-labels. More advanced methods like UniMatch (Yang et al., 2023a) (Fig. 2(b)) may introduce mutual supervision between dual branches. In contrast, as illustrated in Fig. 2(c), our PedSemiSeg incorporates supervision from the weakly augmented branch to the strongly augmented counterpart, as well as supervision between the strongly augmented variants. To generate multiple views of the unlabeled data, we devise a series of weak-to-strong augmentations using sequential geometry-to-intensity perturbations. Then, we derive artificial labels, i.e., pseudo and complementary labels from the "teacher" branch, to facilitate positive and negative learning of the "student" branch. Furthermore, we incorporate reciprocal peer tutoring between the "student" branches, where the learning direction is determined by respective prediction uncertainty. By combining teacher guidance, peer collaboration, and curriculum-style perturbations, our pedagogical framework represents the first holistic integration of educational theory with semi-supervised medical image analysis, offering a robust and effective approach for label-efficient segmentation tasks.

3. Methodology

The primary objective of this study is to develop a label-efficient semi-supervised learning methodology that can effectively accomplish precise and robust polyp segmentation. To enhance clarity, we first give an overall introduction of our **PedSemiSeg** framework in Section 3.1, including problem formulation, necessary notations, and relevant preliminaries. Subsequently, we explain three fundamental components of our method, namely, **Geometry to Intensity Perturbation** in Section 3.2, **Positive and Negative Learning from the Teacher** in Section 3.3, and **Uncertainty-guided Reciprocal Peer Tutoring** in Section 3.4. Finally, a detailed elaboration on the holistic loss supervision employed for model optimization is presented in Section 3.5.

3.1. Preliminaries and overview

In semi-supervised learning, only a small portion of the training dataset is well annotated, while the majority remains unlabeled. The

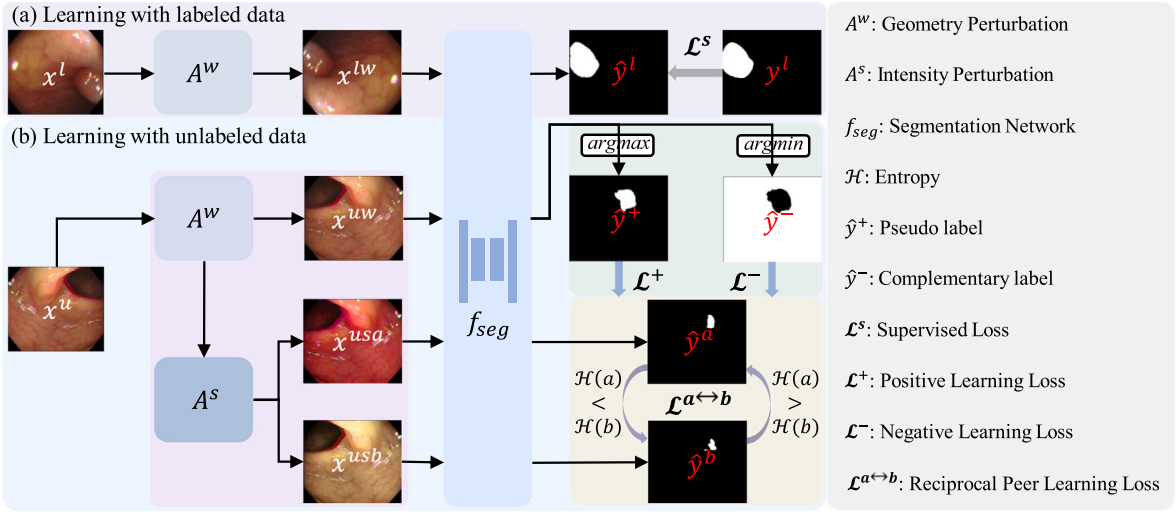


Fig. 3. Overall diagram of our proposed **PedSemiSeg**, a pedagogy-inspired semi-supervised approach for polyp segmentation. As depicted in (a), the labeled data undergoes conventional supervised training to optimize the segmentation model f_{seg} with the supervised loss $\mathcal{L}^s$. Meanwhile, the unlabeled input, as shown in (b), first goes through sequential weak-to-strong (geometry-to-intensity) perturbations to generate three variants. These variants are then fed into the model, resulting in one "teacher" branch and two "student" branches. Inspired by pedagogical activities, we obtain pseudo and complementary labels with argmax and argmin, respectively, to facilitate positive learning $\mathcal{L}^+$ and negative learning $\mathcal{L}^-$ from the teacher branch. Additionally, the dual student branches engage in reciprocal peer tutoring $\mathcal{L}^{a\leftrightarrow b}$ between themselves, where their prediction entropy $\mathcal{H}$ determines the learning direction.

complete training dataset can be represented as $D = \{D^l, D^u\}$, where $D^l = \{x^l, y^l\}$ represents the labeled data with input samples x^l and corresponding labels y^l, and $D^u = \{x^u\}$ represents the unlabeled data. The primary objective of semi-supervised training is to effectively exploit the abundant unlabeled data D^u in conjunction with the limited labeled data D^l to obtain a performant segmentation model, referred to as f_{seg}. Generally, the overall training loss $\mathcal{L}$ consists of two components, namely the supervised loss $\mathcal{L}^s$ on D^l and the unsupervised loss $\mathcal{L}^u$ on D^u.

As depicted in Fig. 3, our **PedSemiSeg** deals with labeled and unlabeled data separately. Specifically, for the labeled input x^l, the model generates predictions $\hat{y}^l$ on its weakly augmented version x^{lw} and computes supervised loss $\mathcal{L}^s$ using the corresponding ground truth mask y^l. On the other hand, our framework adopts the prevailing design principle of consistency regularization in semi-supervised learning, when training with unlabeled data. Concretely, we utilize the model's prediction on the unlabeled input x^u subjected to weak augmentations A^w to generate the pseudo label $\hat{y}^+$ and complementary label $\hat{y}^-$. These artificial labels then supervise the model's prediction on the same input but with strong augmentations A^s in positive and negative learning regimes, respectively. Additionally, we apply dual-stream strong augmentations on the corresponding input x^u, resulting in two parallel predictions $\hat{y}^a$ and $\hat{y}^b$. Subsequently, we establish reciprocal supervision between these predictions, with the direction determined by the prediction uncertainty. Consequently, we construct three learning mechanisms: **positive learning** ($\mathcal{L}^+$) aligns student predictions with high-confidence teacher pseudo-labels, **negative learning** ($\mathcal{L}^-$) enforces divergence from error-indicative complementary labels, and **reciprocal peer tutoring** ($\mathcal{L}^{a\leftrightarrow b}$) facilitates uncertainty-guided knowledge exchange between student branches.

3.2. Weak-to-strong image perturbation

The fundamental idea of consistency regularization-based approaches in semi-supervised segmentation is to produce multiple perturbed versions of a given input and encourage the model to generate consistent predictions across these variations. As illustrated in Fig. 3, in our approach, we implement a sequential augmentation strategy that transitions from geometric to intensity transformations, emulating the pedagogical principle of curriculum learning, i.e., progressively increasing task complexity to stabilize model training.

3.2.1. Geometry-based weak perturbations

In the first step, we utilize several widely-used geometric transformations as the weak perturbations A^w, including *Resize*, *Crop*, *HorizontalFlip*, and *VerticalFlip*. Such transformations mimic viewpoint changes caused by endoscope movement while preserving lesion morphology and intensity distributions, yielding minimally perturbed input variants that maintain reliable spatial relationships for pseudo-label generation.

3.2.2. Intensity-based strong perturbations

Following the weak perturbations, we apply strong perturbations A^s using the nonlinear intensity-based *ColorJitter* augmentations to simulate photometric variations across clinical environments. These perturbations involve random changes in *brightness*, *contrast*, *saturation*, and *hue*. By replicating challenges such as uneven illumination, specular highlights, and color calibration discrepancies across endoscopy systems, these augmentations generate input variants that broaden the effective input distribution while maintaining the integrity of lesion geometry.

3.2.3. Assignment of the teacher and the students

As shown in Fig. 3, for the unlabeled input x^u, our framework generates one weakly augmented variant $x^{uw} = A^w(x^u)$ and two strongly augmented variants $x^{usa}, x^{usb} = A^s(A^w(x^u))$ through weak-to-strong perturbations. Intuitively, the segmentation network f_{seg} exhibits higher confidence in making predictions for the weakly augmented input than for the strongly augmented variants. Therefore, we designate the weakly augmented branch as the "teacher" branch, while the strongly augmented branches serve as the peer "students" branches. For clarity, we denote the direct model outputs, also known as the logits, as $l^{uw} = f_{seg}(x^{uw})$, $l^{usa} = f_{seg}(x^{usa})$, and $l^{usb} = f_{seg}(x^{usb})$, respectively. These logits can be further converted to probability maps p^{uw}, p^{usa}, and p^{usb} with the *Softmax* operation.

This sequential perturbation strategy is designed to address two key clinical realities. First, anatomical spatial relationships – such as shape, size, and position – remain relatively stable under endoscope motion, making geometric perturbations a reliable source for generating pseudo-labels. Second, photometric variability is the primary factor driving cross-center domain shifts in endoscopic imaging, necessitating aggressive intensity augmentations to enhance model robustness. By strategically combining these approaches, our framework ensures both stability in geometric features and adaptability to photometric variations, aligning closely with real-world clinical conditions.

3.3. Positive and negative learning from the teacher

To boost the overall learning outcomes in practical educational settings, an effective teacher should not only instruct students on what is correct, but also inform them about what is erroneous. Drawing inspiration from this notion, our framework incorporates a "teacher" role which generates pseudo-labels for the "students" to engage in positive learning regarding the appropriate class assignment for each pixel. It also provides complementary labels to facilitate negative learning by emphasizing the classes that are least likely to be associated with each pixel.

3.3.1. Positive learning with pseudo labels

Pseudo-label generation is a widely embraced technique in the realm of semi-supervised learning, enabling the acquisition of artificial labels that serve as supervision signals during the training process with unlabeled data that lacks ground truth annotations. This strategic approach aligns with the established positive learning paradigm in image segmentation, where labels conventionally indicate the most probable class affiliation for each pixel.

To derive the pseudo label $\hat{y}^+$ for an unlabeled input image x^u, we employ a threshold-based approach utilizing the probability map p^{uw} provided by the "teacher". This approach can be mathematically formulated as follows:

$$\hat{y}^+ = \text{argmax}_C \, \mathbb{1}(\text{Norm}(p^{uw}) > \tau) p^{uw}. \tag{1}$$

Here, $\tau \in [0, 1]$ represents the confidence threshold used to truncate the min–max normalized probability map p^{uw}, thereby producing a mask through the application of the indicator function $\mathbb{1}(\cdot)$ to filter out unreliable pixel predictions. We set $\tau = 0.8$ in our implementation according to extensive ablation studies discussed in Section 5.3.2. C denotes the total number of classes, which in the context of the polyp segmentation task is equal to 2. The crafted pseudo labels from the "teacher" can subsequently be utilized to supervise the two outputs of the "students" in a positive learning manner. Specifically, the positive learning loss can be expressed as follows:

$$\mathcal{L}^+ = (\hat{y}^+ \xrightarrow{\oplus} \hat{y}^a) + (\hat{y}^+ \xrightarrow{\oplus} \hat{y}^b), \tag{2}$$

where "$\xrightarrow{\oplus}$" denotes positive supervision, indicating that the pseudo label $\hat{y}^+$ is used to guide the training of both $\hat{y}^a$ and $\hat{y}^b$ in a positive fashion.

3.3.2. Negative learning with complementary labels

The utilization of positive learning with pseudo labels enables the segmentation model to acquire knowledge from pixel predictions that exhibit higher confidence. However, it is also valuable to leverage the information provided by unreliable pixel predictions, as they can offer complementary guidance regarding what should not be classified as the target class. In our framework, we can obtain the low-confidence complementary labels from the probability map of the "teacher" using the following expression:

$$\hat{y}^- = \text{argmin}_C \, p^{uw}. \tag{3}$$

These complementary labels can be employed to guide the negative learning process with:

$$\mathcal{L}^- = (\hat{y}^- \xrightarrow{\lambda\ominus} \hat{y}^a) + (\hat{y}^- \xrightarrow{\lambda\ominus} \hat{y}^b). \tag{4}$$

Here, "$\xrightarrow{\ominus}$" denotes negative supervision. The coefficient $\lambda \in [0, 1]$ represents a dynamic weighting factor for each sample, which adjusts the strength of negative learning based on prediction entropy. It can be formulated as follows:

$$\lambda = 1 - \frac{\mathcal{H}(p^{uw})}{\log(H \cdot W)}, \tag{5}$$

where H and W denote the height and weight of the input image, respectively. $\mathcal{H}(p^{uw})$ represents the entropy of the probability map of the "teacher" and can be calculated as:

$$\mathcal{H}(p^{uw}) = -\sum_{c=1}^{C} p^{uw}(c) \log(p^{uw}(c)). \tag{6}$$

Note that higher entropy corresponds to lower confidence or larger uncertainty in prediction.

3.4. Uncertainty-guided reciprocal peer tutoring

In addition to the positive and negative learning from the "teacher", our framework introduces a novel learning paradigm known as "peer tutoring". This paradigm facilitates the bidirectional transfer of knowledge between the dual "students", where the learning direction in this paradigm is not fixed but determined based on the performance of each student in every training iteration. Specifically, the learning direction is from the less proficient student to the higher-achieving one, mimicking the process of knowledge exchange in a classroom setting. In terms of semi-supervised learning, we derive the pseudo label from the high-confidence prediction to guide the training of the low-confidence counterpart.

Concretely, for the two probability maps of the dual "students", namely p^{usa} and p^{usb}, we can quantify their uncertainty by calculating their entropy using a similar formulation as Eq. (6). Then we compare the computed entropy values and select the probability map with lower entropy to provide supervision for the counterpart with higher entropy. Our "peer tutoring" process shares similarities with positive learning from the "teacher", as described in Section 3.3.1. In this context, our directional supervision is expressed as:

$$\mathcal{L}^{a \leftrightarrow b} = \begin{cases} \hat{y}^{a+} \xrightarrow{\oplus} \hat{y}^b & \text{if } \mathcal{H}(p^{usa}) < \mathcal{H}(p^{usb}), \\ \hat{y}^{b+} \xrightarrow{\oplus} \hat{y}^a & \text{if } \mathcal{H}(p^{usa}) \geq \mathcal{H}(p^{usb}), \end{cases} \tag{7}$$

where $\hat{y}^{a+}$ and $\hat{y}^{b+}$ represent the pseudo labels derived from p^{usa} and p^{usb} according to Eq. (1).

Our uncertainty-guided bidirectional peer supervision promotes a collaborative learning framework where the "students" mutually benefit from each other's insights. It allows the dual "students" to leverage each other's strengths and compensate for their weaknesses.

3.5. Holistic loss supervision

To establish comprehensive and efficient loss supervision, we have utilized two commonly employed loss functions in image segmentation: the Dice Similarity Coefficient (DSC) loss ℓ^{DSC} and the Cross-Entropy (CE) loss ℓ^{CE}, as the foundational building blocks of our holistic loss function. The incorporation of the DSC loss, which emphasizes spatial overlap and overall segmentation quality, along with the CE loss, which ensures accurate pixel-wise classification, enables us to achieve more precise and visually coherent segmentation outcomes.

To improve data efficiency in semi-supervised training, our overall loss function includes two components: the supervised loss $\mathcal{L}^s$ on the labeled data and the unsupervised loss $\mathcal{L}^u$ on the unlabeled data. With two foundational loss functions, we can formulate the supervised loss as follows:

$$\mathcal{L}^s = \ell^{CE}(l^l, y^l) + \ell^{DSC}(p^l, y^l), \tag{8}$$

where l^l and p^l denote the logit and probability map of the labeled input. In addition, the positive supervision, referred to as "$\xrightarrow{\oplus}$" in Eqs. (2) and (7), is also the direct sum of ℓ^{DSC} and ℓ^{CE}. Hence, the positive learning loss $\mathcal{L}^+$ can be expanded as follow:

$$\mathcal{L}^+ = \underbrace{\ell^{CE}(l^{usa}, \hat{y}^+) + \ell^{DSC}(p^{usa}, \hat{y}^+)}_{\hat{y}^+ \xrightarrow{\oplus} \hat{y}^a} + \underbrace{\ell^{CE}(l^{usb}, \hat{y}^+) + \ell^{DSC}(p^{usb}, \hat{y}^+)}_{\hat{y}^+ \xrightarrow{\oplus} \hat{y}^b}. \tag{9}$$

Similarly, the bidirectional peer tutoring loss $\mathcal{L}^{a\leftrightarrow b}$ can be expressed as

$$\mathcal{L}^{a\leftrightarrow b} = \begin{cases} \underbrace{\ell^{CE}(l^{usb}, \hat{y}^{a+}) + \ell^{DSC}(p^{usb}, \hat{y}^{a+})}_{\hat{y}^{a+}\xrightarrow{\oplus}\hat{y}^{b}} & \text{if } \mathcal{H}(p^{usa}) < \mathcal{H}(p^{usb}), \\ \underbrace{\ell^{CE}(l^{usa}, \hat{y}^{b+}) + \ell^{DSC}(p^{usa}, \hat{y}^{b+})}_{\hat{y}^{b+}\xrightarrow{\oplus}\hat{y}^{a}} & \text{if } \mathcal{H}(p^{usa}) \geq \mathcal{H}(p^{usb}). \end{cases} \tag{10}$$

The negative supervision, denoted by "$\xrightarrow{\ominus}$", is adapted from the Cross-Entropy loss. The negative learning loss $\mathcal{L}^{-}$ can be defined as:

$$\mathcal{L}^{-} = \underbrace{\ell^{CE}(1 - l^{usa}, \hat{y}^{-})}_{\hat{y}^{-}\xrightarrow{\oplus}\hat{y}^{a}} + \underbrace{\ell^{CE}(1 - l^{usb}, \hat{y}^{-})}_{\hat{y}^{-}\xrightarrow{\oplus}\hat{y}^{b}}. \tag{11}$$

Herein, the unsupervised loss $\mathcal{L}^{u}$, comprising of $\mathcal{L}^{+}$, $\mathcal{L}^{-}$, and $\mathcal{L}^{a\leftrightarrow b}$, can be formalized as

$$\mathcal{L}^{u} = \mathcal{L}^{+} + \mathcal{L}^{-} + \mathcal{L}^{a\leftrightarrow b}. \tag{12}$$

Collectively, the total objective loss function can be expressed as

$$\mathcal{L} = \mathcal{L}^{s} + w^{u} \cdot \mathcal{L}^{u}. \tag{13}$$

Here, w^{u} serves as a weighting coefficient that adjusts the influence of the unsupervised loss. Notably, w^{u} exponentially increases from 0 to its maximum value, w^{u}_{max}, ensuring a progressively amplified incorporation of the unsupervised loss. This strategy facilitates the gradual acquisition of knowledge from the more reliable supervised learning at the earlier training stage, when the model is still relatively weak in generating high-quality pseudo labels for learning from unlabeled data. As the training process continues, the model improves and can generate less noisy pseudo labels. Consequently, the model can effectively exploit the abundant unlabeled data without encountering substantial misguidance.

4. Experimental settings

4.1. Dataset

In this work, we employ the SUN-SEG (Ji et al., 2022) and Kvasir-SEG (Jha et al., 2020) datasets to conduct our semi-supervised polyp segmentation experiments. The recently curated SUN-SEG (Ji et al., 2022) dataset, derived from the SUN Colonoscopy Database (Misawa et al., 2021), comprises 100 distinct polyp cases encompassing diverse and challenging scenarios. This dataset provides an ideal foundation for the development and evaluation of segmentation models in realistic clinical environments. To reduce data redundancy, the original dataset is downsampled by a factor of 5. The downsampled dataset is then randomly divided into training, validation, and testing sets, consisting of 70, 10, and 20 cases, respectively. This division yields 6677, 1240, and 1993 frames in each corresponding split, ensuring a balanced and representative evaluation of model performance. The Kvasir-SEG (Jha et al., 2020) dataset, widely recognized and extensively used in the field, consists of 1000 images of gastrointestinal polyps accompanied by meticulously annotated and verified segmentation masks. The dataset exhibits significant variability in image resolution, ranging from 332 × 487 to 1920 × 1072 pixels. The dataset is partitioned for training and testing, following the methodology established in prior work (Fan et al., 2020), ensuring consistency and comparability with existing research.

In addition to evaluating the model's performance on in-distribution data, we assess its generalization capabilities on external, unseen datasets. For this purpose, we utilize the PolypGen (Ali et al., 2023) dataset, an expanded version of the EndoCV2021 (Ali et al., 2021) Challenge, comprising data collected from six centers across Europe and Africa. The dataset includes 256, 301, 457, 227, 208, and 88 samples from each center, respectively. Spanning diverse populations, endoscopic systems, and surveillance experts from Norway, France, the United Kingdom, Egypt, and Italy, this dataset exhibits significant domain gaps, as illustrated in Fig. 4. As a comprehensive open-access resource, PolypGen (Ali et al., 2023) serves as a robust benchmark for evaluating the generalizability of polyp segmentation methods across varied clinical settings.

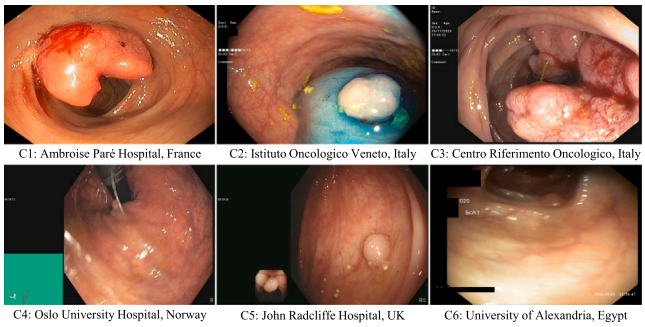


Fig. 4. Sample images from six medical centers of the PolypGen (Ali et al., 2023) dataset, illustrating significant appearance variations that lead to notable domain gaps.

4.2. Implementation details

We implement PedSemiSeg using PyTorch (Paszke et al., 2019), adopting UNet (Ronneberger et al., 2015) as the segmentation backbone due to its proven effectiveness in medical imaging and compatibility with established semi-supervised frameworks. To ensure a fair comparison with baseline methods (CCT (Ouali et al., 2020), CPS (Chen et al., 2021), URPC (Luo et al., 2022b), FixMatch (Sohn et al., 2020), and EVIL (Chen et al., 2024)), we utilize the SSL4MIS[1] codebase with consistent hyperparameters across all experiments. Additionally, to provide a comprehensive performance context, we include two fully-supervised variants of the vanilla UNet (Ronneberger et al., 2015) model. The first, denoted as Full-UNet, is trained with 100% of the labeled training data and serves as the upper-bound performance reference. The second, termed Part-UNet, is trained with the same partial labeled data fractions (1/2, 1/4, 1/8, 1/16) as the semi-supervised methods, establishing the lower-bound performance for each respective labeled ratio. The training protocol employs SGD optimization with a momentum of 0.9 and weight decay of 0.0001, initializing the learning rate at 0.01 followed by polynomial decay over 20,000 iterations. A batch size of 16 balances computational efficiency with gradient stability, containing equal proportions of labeled and unlabeled samples.

Weak augmentations include geometric transformations such as random resizing (224 × 224 crop), horizontal/vertical flipping, and border cropping (maximum 7 pixels), while strong augmentations apply photometric perturbations via *ColorJitter* with randomized brightness (±0.4), contrast (±0.4), saturation (±0.4), and hue (±0.2) adjustments. The unsupervised loss weight w^{u} follows a ramp-up schedule from 0 to $w^{u}_{max} = 1$ during the first 5000 iterations, allowing gradual integration of pseudo-labels as model predictions stabilize. Confidence threshold $\tau = 0.8$ for pseudo-label generation is determined through ablation studies presented in Section 5.3.2, filtering uncertain predictions while retaining sufficient supervision signals.

For evaluation, we report the mean and standard deviation of Dice Similarity Coefficient (DSC), Intersection-over-Union (IoU), Precision,

[1] https://github.com/HiLab-git/SSL4MIS

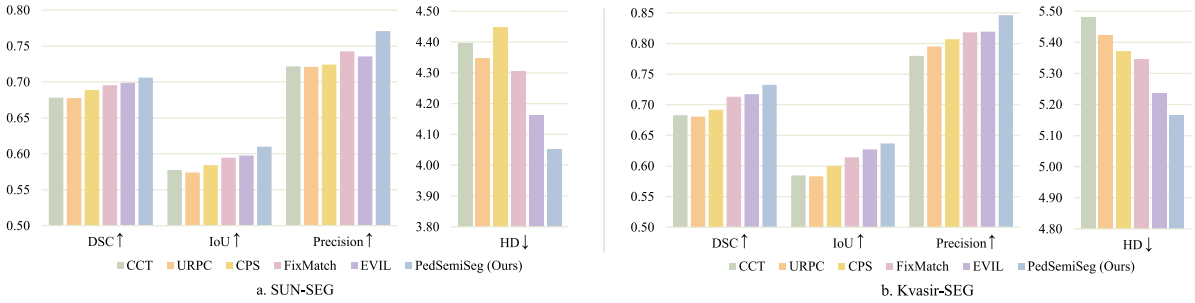


Fig. 5. Average performance across multiple labeled data ratios on SUN-SEG (Ji et al., 2022) and Kvasir-SEG (Jha et al., 2020) datasets. Our proposed PedSemiSeg consistently maintains higher DSC, IoU, and Precision, and lower HD values, demonstrating its robustness in various labeling settings.

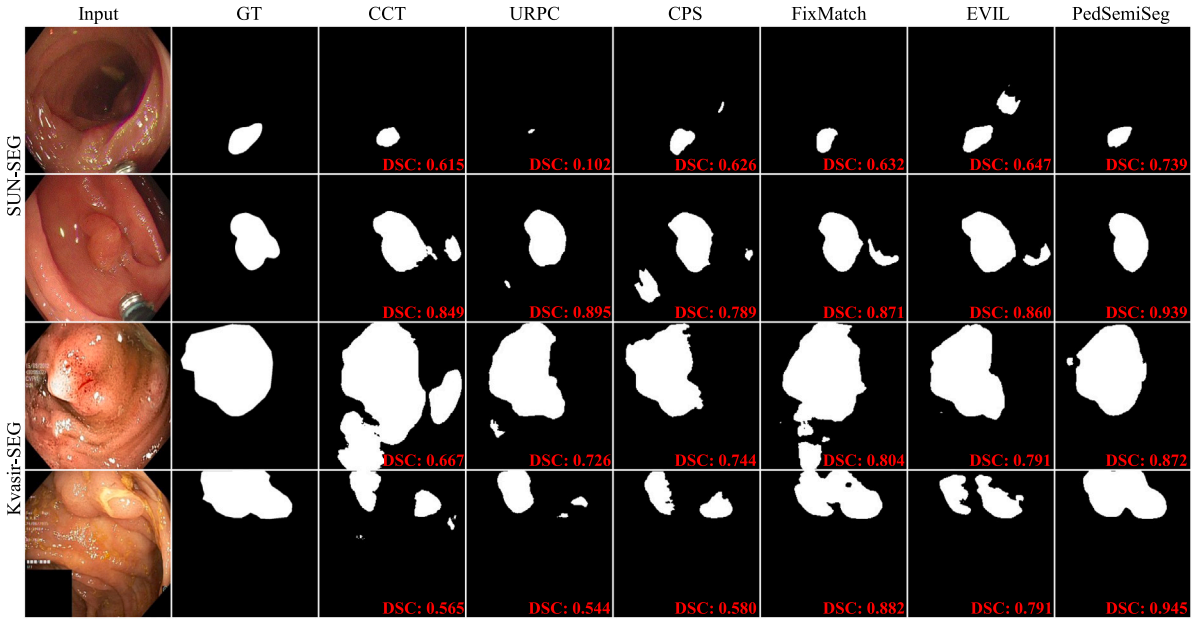


Fig. 6. Qualitative comparison of polyp segmentation performance on the in-distribution dataset SUN-SEG (Ji et al., 2022) and Kvasir-SEG (Jha et al., 2020). DSC values of all predictions are overlaid for direct numerical comparison. Our proposed PedSemiSeg can generate more precise segmentation masks for polyps for both datasets.

and Hausdorff Distance (HD) across three independent training runs. All experiments are conducted on an NVIDIA RTX3090 GPU. Code is available for reference at https://github.com/lofrienger/PedSemiSeg.

5. Results and analysis

5.1. In-distribution evaluation

Our proposed PedSemiSeg demonstrates superior segmentation performance under limited supervision across both SUN-SEG (Ji et al., 2022) and Kvasir-SEG (Jha et al., 2020) datasets. As detailed in Table 1, PedSemiSeg consistently outperforms other semi-supervised methods across all labeled data ratios (1/2, 1/4, 1/8, and 1/16). For instance, with only 1/16 labeled data, PedSemiSeg achieves a DSC of 0.6832 ± 0.0268 on SUN-SEG (Ji et al., 2022) and 0.6532 ± 0.0384 on Kvasir-SEG (Jha et al., 2020), surpassing the strong baseline EVIL (Chen et al., 2024) by 0.65% and 1.18% DSC, respectively. Moreover, PedSemiSeg demonstrates notable robustness, with less performance degradation than other baselines as labeled data diminishes, highlighting its effectiveness in scenarios with minimal supervision.

Table 1 also includes results for Full-UNet (Ronneberger et al., 2015) (trained with 100% labeled data) and Part-UNet (Ronneberger et al., 2015) (trained with the corresponding partial labeled data), serving as upper and lower performance bounds, respectively. The Full-UNet establishes the performance ceiling, achieving a DSC of 0.7254 on SUN-SEG (Ji et al., 2022) and 0.8195 on Kvasir-SEG (Jha et al., 2020). Conversely, the Part-UNet results illustrate the significant performance drop when a fully-supervised model is trained with limited data. For instance, on SUN-SEG (Ji et al., 2022), the Part-UNet DSC degrades from 0.6803 ± 0.0251 (1/2 ratio) to 0.6372 ± 0.0350 (1/16 ratio). A similar, more pronounced trend is observed on Kvasir-SEG (Jha et al., 2020), where the Part-UNet DSC falls from 0.7368 ± 0.0321 (1/2 ratio) to 0.5105 ± 0.0450 (1/16 ratio).

Crucially, our PedSemiSeg consistently and substantially outperforms the Part-UNet across all labeled ratios on both datasets. For example, with only 1/16 labeled data on SUN-SEG (Ji et al., 2022), PedSemiSeg (DSC: 0.6832) surpasses Part-UNet (DSC: 0.6372) by a significant margin of 4.6%. On Kvasir-SEG (Jha et al., 2020) with 1/16 data, PedSemiSeg (DSC: 0.6532) outperforms Part-UNet (DSC: 0.5105) by an even larger margin of 14.27%. This demonstrates our method's superior ability to leverage unlabeled data effectively, achieving strong performance even when supervised signals are scarce. While there is still a gap to the Full-UNet upper bound, PedSemiSeg significantly narrows this gap compared to the Part-UNet lower bound, showcasing its strong label efficiency. For instance, on SUN-SEG (Ji et al., 2022) with 1/2 labeled data, PedSemiSeg (DSC: **0.7225**) nearly matches the Full-UNet performance (DSC: 0.7254) and substantially exceeds the Part-UNet (DSC: 0.6803).

Fig. 5 illustrates the average performance across labeled ratios, where PedSemiSeg maintains superior DSC, IoU, Precision, and lower

Table 1
Quantitative segmentation results on SUN-SEG (Ji et al., 2022) and Kvasir-SEG (Jha et al., 2020) datasets with 1/2, 1/4, 1/8, and 1/16 labeled ratios. Our proposed PedSemiSeg consistently outperforms other semi-supervised baselines across all labeled ratios on both datasets. The blue-shaded rows indicate the upper-bound performance of fully-supervised training with vanilla UNet (Ronneberger et al., 2015) and 100% labeled data, while the gray-shaded rows show the lower-bound performance of fully-supervised training with the corresponding partial labeled data. The best results are highlighted in bold, and runner-ups are underlined.

Dataset: SUN-SEG Ji et al. (2022)

Method	Labeled Ratio	DSC ↑	IoU ↑	Precision ↑	HD ↓	Method	Labeled Ratio	DSC ↑	IoU ↑	Precision ↑	HD ↓
Full-UNet Ronneberger et al. (2015)	1	0.7254±0.0213	0.6305±0.0185	0.7792±0.0229	4.2199±0.1241	Full-UNet Ronneberger et al. (2015)	1	0.7254±0.0213	0.6305±0.0185	0.7792±0.0229	4.2199±0.1241
Part-UNet Ronneberger et al. (2015)	1/2	0.6803±0.0251	0.5750±0.0210	0.7300±0.0275	4.3726±0.1887	Part-UNet Ronneberger et al. (2015)	1/8	0.6430±0.0314	0.5397±0.0258	0.7011±0.0320	4.6601±0.2128
CCT Ouali et al. (2020)		0.7068±0.0236	0.6149±0.0204	0.7492±0.0257	4.1347±0.1612	CCT Ouali et al. (2020)		0.6680±0.0261	0.5631±0.0232	0.7243±0.0284	4.5845±0.2247
URPC Luo et al. (2022b)		0.6796±0.0253	0.5769±0.0215	0.7342±0.0266	4.3149±0.1693	URPC Luo et al. (2022b)		0.6694±0.0279	0.5652±0.0244	0.7089±0.0278	4.3446±0.1704
CPS Chen et al. (2021)		0.7213±0.0212	0.6175±0.0193	0.7440±0.0219	4.1967±0.1644	CPS Chen et al. (2021)		0.6826±0.0267	0.5736±0.0225	0.7244±0.0284	4.6706±0.2290
FixMatch Sohn et al. (2020)		0.7145±0.0210	0.6159±0.0181	0.7690±0.0226	4.1346±0.1620	FixMatch Sohn et al. (2020)		0.6991±0.0274	0.5914±0.0232	0.7388±0.0290	4.6324±0.2270
EVIL Chen et al. (2024)		0.7191±0.0211	0.6233±0.0183	0.7595±0.0223	3.8729±0.1563	EVIL Chen et al. (2024)		0.7043±0.0276	0.6024±0.0236	**0.7448±0.0292**	4.1177±0.1669
PedSemiSeg (Ours)		**0.7225±0.0198**	**0.6288±0.0173**	**0.8272±0.0227**	**3.7899±0.1530**	**PedSemiSeg (Ours)**		**0.7096±0.0243**	**0.6094±0.0209**	0.7432±0.0255	**4.0701±0.1652**
Part-UNet Ronneberger et al. (2015)	1/4	0.6729±0.0291	0.5688±0.0302	0.7150±0.0290	4.4518±0.2021	Part-UNet Ronneberger et al. (2015)	1/16	0.6372±0.0350	0.5307±0.0311	0.6716±0.0352	4.9127±0.2508
CCT Ouali et al. (2020)		0.6839±0.0268	0.5807±0.0228	0.7230±0.0283	4.3338±0.1700	CCT Ouali et al. (2020)		0.6544±0.0301	0.5511±0.0270	0.6899±0.0338	4.5259±0.2218
URPC Luo et al. (2022b)		0.6959±0.0273	0.5940±0.0233	0.7399±0.0290	4.2204±0.1656	URPC Luo et al. (2022b)		0.6658±0.0294	0.5597±0.0247	0.7006±0.0309	4.5022±0.2205
CPS Chen et al. (2021)		0.6967±0.0273	0.5974±0.0234	0.7352±0.0288	4.1300±0.1621	CPS Chen et al. (2021)		0.6545±0.0321	0.5489±0.0269	0.6932±0.0340	4.7933±0.2349
FixMatch Sohn et al. (2020)		0.6935±0.0272	0.5951±0.0234	0.7500±0.0294	4.1907±0.1642	FixMatch Sohn et al. (2020)		0.6749±0.0298	0.5761±0.0254	0.7123±0.0315	4.2575±0.2083
EVIL Chen et al. (2024)		0.6952±0.0272	0.5971±0.0234	0.7372±0.0289	4.2943±0.1684	EVIL Chen et al. (2024)		0.6767±0.0299	0.5672±0.0250	0.7007±0.0309	4.3605±0.2141
PedSemiSeg (Ours)		**0.7094±0.0243**	**0.6125±0.0210**	**0.7694±0.0264**	**4.0991±0.1606**	**PedSemiSeg (Ours)**		**0.6832±0.0268**	**0.5892±0.0231**	**0.7437±0.0292**	**4.2415±0.1661**

Dataset: Kvasir-SEG Jha et al. (2020)

Method	Labeled Ratio	DSC ↑	IoU ↑	Precision ↑	HD ↓	Method	Labeled Ratio	DSC ↑	IoU ↑	Precision ↑	HD ↓
Full-UNet Ronneberger et al. (2015)	1	0.8195±0.0241	0.7355±0.0216	0.8910±0.0262	4.8761±0.1434	Full-UNet Ronneberger et al. (2015)	1	0.8195±0.0241	0.7355±0.0216	0.8910±0.0262	4.8761±0.1434
Part-UNet Ronneberger et al. (2015)	1/2	0.7368±0.0321	0.6412±0.0295	0.8341±0.0350	5.3273±0.2104	Part-UNet Ronneberger et al. (2015)	1/8	0.5418±0.0412	0.4391±0.0378	0.7450±0.0485	5.7815±0.2850
CCT Ouali et al. (2020)		0.7329±0.0288	0.6399±0.0251	0.8020±0.0314	5.3267±0.2090	CCT Ouali et al. (2020)		0.6839±0.0335	0.5874±0.0288	0.8076±0.0396	5.3816±0.2637
URPC Luo et al. (2022b)		0.7458±0.0293	0.6564±0.0257	0.8540±0.0335	5.0111±0.1966	URPC Luo et al. (2022b)		0.6777±0.0332	0.5818±0.0285	0.8141±0.0399	5.3826±0.2636
CPS Chen et al. (2021)		0.7559±0.0296	0.6729±0.0264	0.8425±0.0330	4.9428±0.1939	CPS Chen et al. (2021)		0.6644±0.0326	0.5660±0.0278	0.8014±0.0393	5.4163±0.2653
FixMatch Sohn et al. (2020)		0.7842±0.0307	0.6913±0.0271	0.8754±0.0343	5.1472±0.2018	FixMatch Sohn et al. (2020)		0.6958±0.0341	0.5979±0.0293	0.8005±0.0392	5.3247±0.2608
EVIL Chen et al. (2024)		0.7868±0.0309	0.6987±0.0274	0.8757±0.0343	5.1177±0.2006	EVIL Chen et al. (2024)		0.6908±0.0339	0.6054±0.0297	0.7952±0.0390	5.3095±0.2600
PedSemiSeg (Ours)		**0.7916±0.0271**	**0.7037±0.0241**	**0.8925±0.0306**	**4.8772±0.1913**	**PedSemiSeg (Ours)**		**0.7235±0.0319**	**0.6294±0.0278**	**0.8408±0.0372**	**5.2707±0.2065**
Part-UNet Ronneberger et al. (2015)	1/4	0.6740±0.0417	0.5813±0.0342	0.7716±0.0421	5.7455±0.2857	Part-UNet Ronneberger et al. (2015)	1/16	0.5105±0.0450	0.4138±0.0402	0.6719±0.0520	6.4052±0.3189
CCT Ouali et al. (2020)		0.7089±0.0347	0.6180±0.0303	0.7945±0.0389	5.3771±0.2635	CCT Ouali et al. (2020)		0.6059±0.0386	0.4929±0.0339	0.7158±0.0456	5.8337±0.2859
URPC Luo et al. (2022b)		0.7051±0.0345	0.6052±0.0297	0.7936±0.0389	5.4364±0.2664	URPC Luo et al. (2022b)		0.5937±0.0378	0.4891±0.0336	0.7194±0.0458	5.8642±0.2874
CPS Chen et al. (2021)		0.7369±0.0361	0.6446±0.0316	0.8242±0.0404	5.3979±0.2645	CPS Chen et al. (2021)		0.6098±0.0389	0.5163±0.0354	0.7603±0.0485	5.7274±0.2807
FixMatch Sohn et al. (2020)		0.7367±0.0361	0.6407±0.0314	0.8367±0.0410	5.3097±0.2601	FixMatch Sohn et al. (2020)		0.6354±0.0405	0.5268±0.0361	0.7599±0.0484	5.6030±0.2745
EVIL Chen et al. (2024)		0.7498±0.0367	0.6606±0.0324	0.8238±0.0403	**5.0133±0.1965**	EVIL Chen et al. (2024)		0.6414±0.0408	0.5437±0.0373	**0.7834±0.0499**	5.5043±0.2697
PedSemiSeg (Ours)		**0.7622±0.0336**	**0.6695±0.0295**	**0.8682±0.0383**	5.0357±0.1974	**PedSemiSeg (Ours)**		**0.6532±0.0384**	**0.5445±0.0320**	0.7832±0.0460	**5.4783±0.2147**

Table 2
Quantitative comparison of generalizability across six data centers of the PolypGen (Ali et al., 2023) dataset. The models are trained with 1/16 labeled SUN-SEG (Ji et al., 2022) dataset. Best and runner-up DSC results are bolded and underlined, respectively. Our proposed PedSemiSeg demonstrates superior performance over other methods.

Center ID (No. frames)	CCT (Ouali et al., 2020)	URPC (Luo et al., 2022b)	CPS (Chen et al., 2021)	FixMatch (Sohn et al., 2020)	EVIL (Chen et al., 2024)	PedSemiSeg (Ours)	Full-UNet (Ronneberger et al., 2015)
1 (256)	0.5712 ± 0.0281	0.6035 ± 0.0305	0.5384 ± 0.0271	0.6223 ± 0.0307	0.5493 ± 0.0278	**0.6725 ± 0.0326**	0.7126 ± 0.0282
2 (301)	0.4306 ± 0.0221	0.4457 ± 0.0216	0.4583 ± 0.0232	0.5238 ± 0.0272	0.4861 ± 0.0239	**0.5681 ± 0.0275**	0.6936 ± 0.0269
3 (457)	0.5721 ± 0.0225	0.6258 ± 0.0246	0.5532 ± 0.0218	0.6071 ± 0.0251	0.6697 ± 0.0261	**0.6754 ± 0.0263**	0.7093 ± 0.0243
4 (227)	0.1563 ± 0.0088	0.1518 ± 0.0087	0.1435 ± 0.0090	0.1516 ± 0.0085	0.1617 ± 0.0090	**0.1673 ± 0.0093**	0.2724 ± 0.0142
5 (208)	0.2782 ± 0.0146	0.3294 ± 0.0157	0.2673 ± 0.0130	0.2504 ± 0.0131	0.3321 ± 0.0158	**0.3722 ± 0.0178**	0.4358 ± 0.0201
6 (88)	0.4285 ± 0.0268	0.4623 ± 0.0270	0.4702 ± 0.0272	0.4305 ± 0.0268	**0.4848 ± 0.0281**	0.4793 ± 0.0273	0.5754 ± 0.0297
Mean ± STD	0.4062 ± 0.1629	0.4364 ± 0.1806	0.4052 ± 0.1602	0.4310 ± 0.1943	0.4473 ± 0.1855	**0.4891 ± 0.1872**	0.5665 ± 0.1682

Table 3
Ablation study on the unsupervised loss components. The combination of $\mathcal{L}^{+}$, $\mathcal{L}^{-}$, and $\mathcal{L}^{a\leftrightarrow b}$ yields optimal performance. The best and runner-up results are in bold and underlined.

$\mathcal{L}^{+}$	$\mathcal{L}^{-}$	$\mathcal{L}^{a\leftrightarrow b}$	DSC ↑	Precision ↑
✓	✗	✗	0.6501 ± 0.0228	0.6911 ± 0.0347
✓	✓	✗	0.6714 ± 0.0331	0.7128 ± 0.0251
✓	✗	✓	0.6792 ± 0.0235	0.7373 ± 0.0254
✓	✓	✓	**0.6832 ± 0.0268**	**0.7437 ± 0.0292**

HD on both datasets. Qualitative results in Fig. 6 further validate our method's ability to generate precise segmentation masks under challenging conditions (e.g., mucosal folds, specular reflections), whereas baseline methods produce fragmented or over-segmented predictions.

5.2. Out-of-distribution evaluation

In medical practice, significant data variability across patients and clinical settings is prevalent. Consequently, it is essential for segmentation models to not only achieve high performance on the training dataset but also demonstrate robust generalization on external, unseen domains. Such generalization is critical for enhancing annotation efficiency and minimizing the need for costly retraining. Considering this, we evaluate our proposed method on the extensive out-of-distribution PolypGen (Ali et al., 2023) dataset, utilizing a model trained with only 1/16 of the labeled data from the SUN-SEG (Ji et al., 2022) dataset.

As shown in Table 2, our PedSemiSeg model demonstrates superior generalization performance, outperforming state-of-the-art methods in 5 out of 6 data centers. Specifically, our approach achieves an average Dice Similarity Coefficient (DSC) across the six data centers that surpasses the second-best method by 4.18%. This remarkable performance can be attributed to two key factors: First, the hybrid weak-to-strong image augmentation strategy enhances the model's ability to learn diverse feature representations. Second, the holistic loss supervision design, which incorporates guidance from the teacher to the student model as well as reciprocal peer tutoring among student models, fosters more effective consistency regularization. These results collectively highlight the robustness and adaptability of our method in diverse clinical settings.

5.3. Ablation study

5.3.1. Unsupervised loss components

In our PedSemiSeg, we introduce three unsupervised loss components to leverage the unlabeled data, namely positive and negative learning from the teacher ($\mathcal{L}^{+}$ and $\mathcal{L}^{-}$) as well as reciprocal peer tutoring between the students ($\mathcal{L}^{a\leftrightarrow b}$). To understand the individual and collective contributions of these components, we conduct a decompositional analysis of these losses. Experiments are performed on the SUN-SEG (Ji et al., 2022) dataset with a labeled ratio of 1/16. The results, presented in Table 3, demonstrate that the combined use of these three loss components achieves optimal DSC and Precision results. This underscores the effectiveness of our comprehensive design, which is inspired by human pedagogical principles and teaching activities. By integrating positive and negative learning from the teacher alongside collaborative peer tutoring, our framework maximizes the potential of unlabeled data, leading to improved generalization and robustness in polyp segmentation.

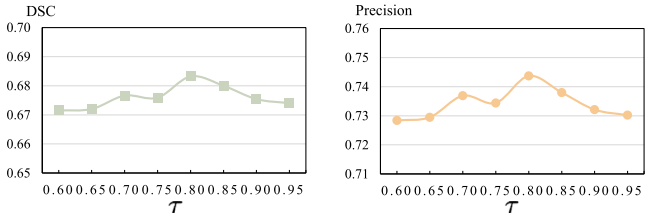


Fig. 7. Ablation study on the confidence threshold τ. Optimal DSC and Precision metrics are obtained when $\tau = 0.8$.

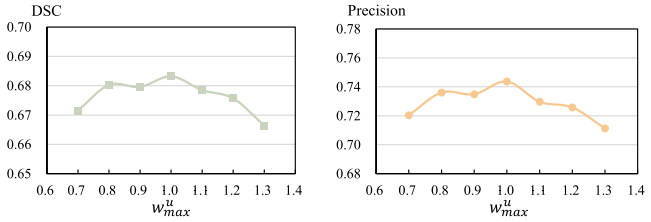


Fig. 8. Ablation study on the maximum loss weighting coefficient w^u_{max}. Optimal DSC and Precision metrics are obtained when $w^u_{max} = 1$.

5.3.2. Confidence threshold

The confidence threshold is a critical parameter for filtering unreliable pseudo-labels and retaining high-quality ones to ensure robust training. To identify the optimal value for the confidence threshold τ, as defined in Eq. (1), we conduct a comprehensive ablation study. As illustrated in Fig. 7, a threshold value of $\tau = 0.8$ achieves the best performance in terms of both DSC and Precision metrics. This finding underscores the importance of carefully selecting the confidence threshold to balance the trade-off between retaining reliable pseudo-labels and minimizing noise in the training process.

5.3.3. Loss weighting coefficient

As formulated in Eq. (13), our holistic loss function integrates a dynamic weighting coefficient w^u to regulate the influence of the unsupervised loss during training. The maximum value of w^u, denoted as w^u_{max}, plays a pivotal role in balancing the contributions of labeled and unlabeled data. An excessively large w^u_{max} overemphasizes the unsupervised loss, while a value that is too small fails to leverage the rich knowledge embedded in the abundant unlabeled data. Both extremes compromise the effectiveness of semi-supervised learning. To address this, we conduct an ablation study to identify the optimal w^u_{max} value that harmonizes the supervised and unsupervised losses. As demonstrated in Fig. 8, a w^u_{max} value of 1 achieves the highest DSC and Precision metrics, respectively. This finding highlights the importance of carefully calibrating w^u_{max} to maximize the synergy between supervised and unsupervised learning components.

Table 4
Performance comparison between Transformer and CNN-based segmentation networks on SUN-SEG (Ji et al., 2022) and Kvasir-SEG (Jha et al., 2020) datasets. Swin-Unet (Cao et al., 2021) demonstrates modest improvements over UNet (Ronneberger et al., 2015). The best results are bolded.

Backbone	SUN-SEG (Ji et al., 2022)		Kvasir-SEG (Jha et al., 2020)	
	DSC ↑	Precision ↑	DSC ↑	Precision ↑
UNet (Ronneberger et al., 2015)	0.6832 ± 0.0268	0.7437 ± 0.0292	0.6532 ± 0.0384	0.7832 ± 0.0460
Swin-Unet (Cao et al., 2021)	**0.6954 ± 0.0314**	**0.7683 ± 0.0277**	**0.6785 ± 0.0262**	**0.7926 ± 0.0431**

5.3.4. Backbone architecture

To assess the compatibility of PedSemiSeg with different network architectures, we replaced the default UNet (Ronneberger et al., 2015) backbone with Swin-Unet (Cao et al., 2021), a segmentation model based on Swin transformer (Liu et al., 2021). As shown in Table 4, experiments on the SUN-SEG (Ji et al., 2022) dataset (1/16 labeled ratio) reveal that Swin-Unet (Cao et al., 2021) achieves a DSC of 0.6954, outperforming UNet (Ronneberger et al., 2015) (DSC = 0.6832) by 1.22%. On Kvasir-SEG (Jha et al., 2020), Swin-Unet (Cao et al., 2021) also yields performance gains. This improvement stems from the self-attention mechanism, which captures long-range dependencies and global context—particularly beneficial for segmenting polyps with irregular shapes or diffuse boundaries. However, Swin-Unet (Cao et al., 2021) incurs higher computational costs, reducing inference speed. Despite this trade-off, the results confirm PedSemiSeg's adaptability to diverse architectures, allowing users to prioritize either efficiency (CNN-based UNet (Ronneberger et al., 2015)) or accuracy (transformer-based Swin-Unet (Cao et al., 2021)) based on clinical requirements. This flexibility underscores the framework's generalizability beyond specific network designs.

6. Discussion

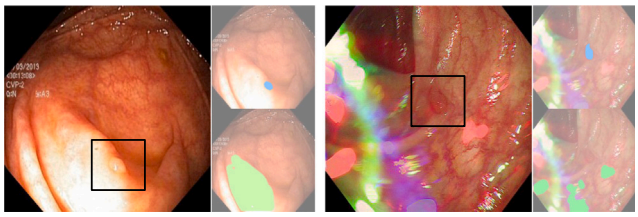
6.1. Methodological and clinical insights

The proposed PedSemiSeg framework addresses the challenges of limited annotated data and domain shifts in polyp segmentation by integrating pedagogy-inspired learning mechanisms. By emulating teacher–student interactions and peer collaboration, our method leverages consistency regularization through sequential geometry-to-intensity augmentations. Weak geometric perturbations preserve spatial relationships critical for pseudo-label reliability, while strong photometric variations simulate cross-center imaging discrepancies. This curriculum-style augmentation strategy aligns with clinical realities, where polyps retain anatomical consistency despite endoscopic viewpoint changes or illumination differences.

The bilateral supervision mechanism – using pseudo-labels for positive learning and complementary labels for negative guidance – reduces overconfidence in erroneous predictions. This approach mirrors clinical training, where experts highlight both pathological features and common diagnostic pitfalls. Qualitative results demonstrate improved boundary precision and reduced false positives near mucosal folds, a persistent challenge in polyp segmentation. The entropy-guided reciprocal peer tutoring further enhances consensus on ambiguous regions, as evidenced by superior performance on the multi-center PolypGen (Ali et al., 2023) dataset.

6.2. Limitations

Despite the promising results achieved by PedSemiSeg, our method exhibits certain limitations that warrant further investigation. As depicted in Fig. 9, the segmentation accuracy declines notably for small polyps and those obscured by clinical artifacts such as specular reflections, blood, or complex tissue textures. These challenges reflect inherent difficulties in polyp segmentation, particularly in a semi-supervised learning context. For small polyps, primary factors contributing to this reduced accuracy include: (1) **Limited Spatial Information**: their small pixel footprint offers scarce data for learning distinctive features. (2) **Class Imbalance in Training Data**: datasets like SUN-SEG (Ji et al., 2022), Kvasir-SEG (Jha et al., 2020), and PolypGen (Ali et al., 2023) often underrepresent small polyps, biasing models towards larger, more frequent ones. (3) **Feature Extraction Difficulties Due to Downsampling**: UNet-based architectures, including ours, can lose critical fine-grained details of small polyps in deeper, downsampled layers. (4) **Annotation Challenges**: the inherent difficulty in accurately annotating small polyps can lead to noisy initial labels, which then propagate and degrade pseudo-label quality within our semi-supervised framework. (5) **Generalization Across Diverse Datasets**: if the limited labeled training subset (e.g., 1/16 of SUN-SEG (Ji et al., 2022)) lacks sufficient examples of small polyps, the model's ability to generalize to new datasets where small polyps are more common or morphologically varied is compromised. Besides, while the framework introduces no additional inference cost, the dual-student design moderately increases training time compared to simpler consistency-based approaches.



(a) (b)

Fig. 9. Visualization of typical challenge cases in polyp recognition and segmentation. The black boxes indicate the target polyps. The blue and green masks, overlaid on the original images, represent the ground truth and inaccurate prediction, respectively.

6.3. Clinical implications

PedSemiSeg's ability to achieve competitive performance with minimal labeled data (e.g., 1/16 labeled ratio) reduces annotation burdens, making it viable for resource-constrained settings. Its robust generalization across diverse datasets, including the multi-center PolypGen (Ali et al., 2023), suggests adaptability to heterogeneous clinical environments without site-specific retraining. This is critical for scalable deployment in global healthcare systems with varying endoscopic equipment and protocols.

7. Conclusion

This work presents **PedSemiSeg**, a pedagogy-inspired semi-supervised framework for label-efficient polyp segmentation. By integrating sequential geometry-to-intensity augmentations, bilateral teacher supervision, and entropy-guided peer tutoring, the method achieves state-of-the-art performance on in-distribution datasets (SUN-SEG (Ji et al., 2022) and Kvasir-SEG (Jha et al., 2020)) and demonstrates strong generalization on the multi-center PolypGen (Ali et al., 2023) benchmark. Grounded in educational theory, the framework's design principles facilitate effective utilization of unlabeled data while mitigating domain shifts, addressing critical challenges in medical image analysis. This approach not only enhances segmentation accuracy but also underscores the potential of pedagogically inspired methodologies in advancing computational tools for medical imaging.

Future work will focus on extending PedSemiSeg to video colonoscopy analysis and integrating shape priors to address rare polyp morphologies. Lightweight architectural adaptations could reduce computational overhead without sacrificing accuracy, enhancing deployability in real-time systems. Collaborative studies with clinicians will evaluate the framework's impact on screening efficiency and lesion characterization accuracy. By bridging the gap between algorithmic innovation and clinical workflows, PedSemiSeg contributes to the development of scalable and generalizable AI tools in gastrointestinal endoscopy, ultimately advancing early detection and treatment of colorectal cancer.

CRediT authorship contribution statement

An Wang: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Haoyu Ma:** Writing – review & editing, Writing – original draft, Visualization, Validation, Formal analysis, Conceptualization. **Long Bai:** Writing – review & editing, Investigation, Conceptualization. **Yanan Wu:** Writing – review & editing, Investigation, Conceptualization. **Mengya Xu:** Writing – review & editing, Investigation, Conceptualization. **Yang Zhang:** Writing – review & editing, Conceptualization. **Mobarakol Islam:** Writing – review & editing, Conceptualization. **Hongliang Ren:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The used datasets are public available. The code is released at https://github.com/lofrienger/PedSemiSeg.

References

Ali, S., Dmitrieva, M., Ghatwary, N., Bano, S., Polat, G., Temizel, A., Krenzer, A., Hekalo, A., Guo, Y.B., Matuszewski, B., et al., 2021. Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy. Med. Image Anal. 70, 102002.

Ali, S., Jha, D., Ghatwary, N., Realdon, S., Cannizzaro, R., Salem, O.E., Lamarque, D., Daul, C., Riegler, M.A., Anonsen, K.V., et al., 2023. A multi-centre polyp detection and segmentation dataset for generalisability assessment. Sci. Data 10 (1), 75.

Basak, H., Yin, Z., 2023. Pseudo-label guided contrastive learning for semi-supervised medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19786–19797.

Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al., 2021. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

Cai, L., Wu, M., Chen, L., Bai, W., Yang, M., Lyu, S., Zhao, Q., 2022. Using guided self-attention with local information for polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 629–638.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M., 2021. Swin-Unet: Unet-like pure transformer for medical image segmentation. arXiv:2105.05537.

Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E., 2023. Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation. Med. Image Anal. 87, 102792.

Chen, J., Shah, V., Kyrillidis, A., 2020. Negative sampling in semi-supervised learning. In: International Conference on Machine Learning. PMLR, pp. 1704–1714.

Chen, Y., Yang, Z., Shen, C., Wang, Z., Zhang, Z., Qin, Y., Wei, X., Lu, J., Liu, Y., Zhang, Y., 2024. Evidence-based uncertainty-aware semi-supervised medical image segmentation. Comput. Biol. Med. 108004.

Chen, X., Yuan, Y., Zeng, G., Wang, J., 2021. Semi-supervised semantic segmentation with cross pseudo supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2613–2622.

Cheplygina, V., de Bruijne, M., Pluim, J.P., 2019. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Med. Image Anal. (ISSN: 1361-8415) 54, 280–296. http://dx.doi.org/10.1016/j.media.2019.03.009.

Dong, B., Wang, W., Fan, D.-P., Li, J., Fu, H., Shao, L., 2021. Polyp-pvt: Polyp segmentation with pyramid vision transformers. arXiv preprint arXiv:2108.06932.

Du, X., Zou, Y., Lei, T., Zhang, W., Wang, Y., Nandi, A.K., 2025. CCL-MPC: Semi-supervised medical image segmentation via collaborative intra-inter contrastive learning and multi-perspective consistency. Neurocomputing 621, 129287.

Fan, D.P., Ji, G.P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L., 2020. Pranet: Parallel reverse attention network for polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 263–273.

Guo, X., Yang, C., Liu, Y., Yuan, Y., 2020. Learn to threshold: Thresholdnet with confidence-guided manifold mixup for polyp segmentation. IEEE Trans. Med. Imaging 40 (4), 1134–1146.

He, A., Li, T., Yan, J., Wang, K., Fu, H., 2023. Bilateral supervision network for semi-supervised medical image segmentation. IEEE Trans. Med. Imaging.

Hinton, G., Vinyals, O., Dean, J., 2015. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Hu, J., Xu, Y., Tang, Z., 2022. DAN-PD: Domain adaptive network with parallel decoder for polyp segmentation. Comput. Med. Imaging Graph. 101, 102124.

Huang, X., Zhuo, L., Zhang, H., Yang, Y., Li, X., Zhang, J., Wei, W., 2022. Polyp segmentation network with hybrid channel-spatial attention and pyramid global context guided feature fusion. Comput. Med. Imaging Graph. 98, 102072.

Jha, D., Ali, S., Tomar, N.K., Johansen, H.D., Johansen, D., Rittscher, J., Riegler, M.A., Halvorsen, P., 2021. Real-time polyp detection, localization and segmentation in colonoscopy using deep learning. IEEE Access 9, 40496–40510.

Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., de Lange, T., Johansen, D., Johansen, H.D., 2020. Kvasir-seg: A segmented polyp dataset. In: MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26. Springer, pp. 451–462.

Ji, G.P., Liu, J., Xu, P., Barnes, N., Khan, F.S., Khan, S., Fan, D.P., 2024a. Frontiers in intelligent colonoscopy. arXiv preprint arXiv:2410.17241.

Ji, G.P., Xiao, G., Chou, Y.C., Fan, D.P., Zhao, K., Chen, G., Van Gool, L., 2022. Video polyp segmentation: A deep learning perspective. Mach. Intell. Res. 1–19.

Ji, G.P., Zhang, J., Campbell, D., Xiong, H., Barnes, N., 2024b. Rethinking polyp segmentation from an out-of-distribution perspective. Mach. Intell. Res. 1–9.

Jia, X., Shen, Y., Yang, J., Song, R., Zhang, W., Meng, M.Q.H., Liao, J.C., Xing, L., 2024. PolypMixNet: Enhancing semi-supervised polyp segmentation with polyp-aware augmentation. Comput. Biol. Med. 170, 108006.

Kim, Y., Yim, J., Yun, J., Kim, J., 2019. Nlnl: Negative learning for noisy labels. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 101–110.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al., 2023. Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026.

Lei, T., Zhang, D., Du, X., Wang, X., Wan, Y., Nandi, A.K., 2022. Semi-supervised medical image segmentation using adversarial consistency learning and dynamic convolution network. IEEE Trans. Med. Imaging.

Li, S., Ren, Y., Yu, Y., Jiang, Q., He, X., Li, H., 2024a. A survey of deep learning algorithms for colorectal polyp segmentation. Neurocomputing 128767.

Li, H., Wu, Y., Bai, L., Wang, A., Chen, T., Ren, H., 2023a. Semi-supervised learning for segmentation of bleeding regions in video capsule endoscopy. Procedia Comput. Sci. 226, 29–35.

Li, N., Xiong, L., Qiu, W., Pan, Y., Luo, Y., Zhang, Y., 2023b. Segment anything model for semi-supervised medical image segmentation via selecting reliable pseudo-labels. In: International Conference on Neural Information Processing. Springer, pp. 138–149.

Li, A., Zeng, X., Zeng, P., Ding, S., Wang, P., Wang, C., Wang, Y., 2024b. Textmatch: Using text prompts to improve semi-supervised medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 699–709.

Lin, A., Chen, B., Xu, J., Zhang, Z., Lu, G., Zhang, D., 2022a. Ds-transunet: Dual swin transformer u-net for medical image segmentation. IEEE Trans. Instrum. Meas. 71, 1–15.

Lin, H.Y., Liu, H.W., et al., 2022b. Multitask deep learning for segmentation and lumbosacral spine inspection. IEEE Trans. Instrum. Meas. 71, 1–10.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022.

Long, J., Lin, J., Liu, D., 2025. W-PolypBox: Exploring bounding box priors constraints for weakly supervised polyp segmentation. Biomed. Signal Process. Control. 103, 107418.

Lu, Y., Shen, Y., Xing, X., Meng, M.Q.H., 2022. Multiple consistency supervision based semi-supervised OCT segmentation using very limited annotations. In: 2022 International Conference on Robotics and Automation. ICRA, IEEE, pp. 8483–8489.

Lucas, T., Weinzaepfel, P., Rogez, G., 2022. Barely-supervised learning: Semi-supervised learning with very few labeled images. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, (2), pp. 1881–1889.

Luo, X., Chen, J., Song, T., Wang, G., 2021. Semi-supervised medical image segmentation through dual-task consistency. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 8801–8809.

Luo, X., Hu, M., Song, T., Wang, G., Zhang, S., 2022a. Semi-supervised medical image segmentation via cross teaching between cnn and transformer. In: International Conference on Medical Imaging with Deep Learning. PMLR, pp. 820–833.

Luo, X., Wang, G., Liao, W., Chen, J., Song, T., Chen, Y., Zhang, S., Metaxas, D.N., Zhang, S., 2022b. Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency. Med. Image Anal. 80, 102517.

Ma, J., Jiang, J., Liu, C., Li, Y., 2017. Feature guided Gaussian mixture model with semi-supervised EM and local geometric constraint for retinal image registration. Inform. Sci. 417, 128–142.

Mei, J., Zhou, T., Huang, K., Zhang, Y., Zhou, Y., Wu, Y., Fu, H., 2025. A survey on deep learning for polyp segmentation: Techniques, challenges and future trends. Vis. Intell. 3 (1), 1.

Misawa, M., Kudo, S.e., Mori, Y., Hotta, K., Ohtsuka, K., Matsuda, T., Saito, S., Kudo, T., Baba, T., Ishida, F., et al., 2021. Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). Gastrointest Endosc. 93 (4), 960–967.

Ouali, Y., Hudelot, C., Tami, M., 2020. Semi-supervised semantic segmentation with cross-consistency training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12674–12684.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. Pytorch: An imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. 32.

Qiu, L., Cheng, J., Gao, H., Xiong, W., Ren, H., 2023. Federated semi-supervised learning for medical image segmentation via pseudo-label denoising. IEEE J. Biomed. Heal. Inform. 1–13. http://dx.doi.org/10.1109/JBHI.2023.3274498.

Rahman, M.M., Munir, M., Jha, D., Bagci, U., Marculescu, R., 2024. PP-SAM: Perturbed prompts for robust adaption of segment anything model for polyp segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4989–4995.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer, pp. 234–241.

Roy, K., Banik, D., Bhattacharjee, D., Krejcar, O., Kollmann, C., 2022. LwMLA-NET: A lightweight multi-level attention-based NETwork for segmentation of COVID-19 lungs abnormalities from CT images. IEEE Trans. Instrum. Meas. 71, 1–13.

Shen, Y., Jia, X., Meng, M.Q.H., 2021. Hrenet: A hard region enhancement network for polyp segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. Springer, pp. 559–568.

Shen, N., Xu, T., Huang, S., Mu, F., Li, J., 2023. Expert-guided knowledge distillation for semi-supervised vessel segmentation. IEEE J. Biomed. Heal. Inform. 27 (11), 5542–5553.

Shi, H., Wang, Z., Lv, J., Wang, Y., Zhang, P., Zhu, F., Li, Q., 2021. Semi-supervised learning via improved teacher-student network for robust 3d reconstruction of stereo endoscopic image. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 4661–4669.

Shi, H., Wang, Z., Zhou, Y., Li, D., Yang, X., Li, Q., 2023. Bidirectional semi-supervised dual-branch CNN for robust 3D reconstruction of stereo endoscopic images via adaptive cross and parallel supervisions. IEEE Trans. Med. Imaging.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L., 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Adv. Neural Inf. Process. Syst. 33, 596–608.

Summers, R.M., Jerebko, A.K., Franaszek, M., Malley, J.D., Johnson, C.D., 2002. Colonic polyps: complementary role of computer-aided detection in CT colonography. Radiology 225 (2), 391–399.

Tarvainen, A., 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.

Wang, A., Islam, M., Xu, M., Ren, H., 2023a. Curriculum-based augmented fourier domain adaptation for robust medical image segmentation. IEEE Trans. Autom. Sci. Eng..

Wang, H., Li, X., 2024. Towards generic semi-supervised framework for volumetric medical image segmentation. Adv. Neural Inf. Process. Syst. 36.

Wang, A., Xu, M., Zhang, Y., Islam, M., Ren, H., 2023b. S^2ME: Spatial-spectral mutual teaching and ensemble learning for scribble-supervised polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 35–45.

Wang, K., Zhan, B., Zu, C., Wu, X., Zhou, J., Zhou, L., Wang, Y., 2022. Semi-supervised medical image segmentation via a tripled-uncertainty guided mean teacher model with contrastive learning. Med. Image Anal. 79, 102447.

Wang, C., Zhao, B., Liu, Z., 2024. DistillMatch: Revisiting self-knowledge distillation in semi-supervised medical image segmentation. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine. BIBM, IEEE, pp. 3778–3783.

Wang, R., Zheng, G., 2024. PFMNet: Prototype-based feature mapping network for few-shot domain adaptation in medical image segmentation. Comput. Med. Imaging Graph. 116, 102406.

Wei, J., Hu, Y., Cui, S., Zhou, S.K., Li, Z., 2023. WeakPolyp: You only look bounding box for polyp segmentation. arxiv:2307.10912. [Cs.CV].

Wu, H., Li, X., Lin, Y., Cheng, K.T., 2023a. Compete to win: Enhancing pseudo labels for barely-supervised medical image segmentation. IEEE Trans. Med. Imaging.

Wu, H., Xie, W., Lin, J., Guo, X., 2023b. ACL-Net: Semi-supervised polyp segmentation via affinity contrastive learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 2812–2820.

Xie, Z., Tu, E., Zheng, H., Gu, Y., Yang, J., 2021. Semi-supervised skin lesion segmentation with learning model confidence. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 1135–1139.

Xiong, X., Li, W., Ma, J., Huang, D., Li, S., 2024. Free meal: Boosting semi-supervised polyp segmentation by harvesting negative samples. IEEE Signal Process. Lett..

Yang, L., Qi, L., Feng, L., Zhang, W., Shi, Y., 2023a. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7236–7246.

Yang, L., Zhao, Z., Qi, L., Qiao, Y., Shi, Y., Zhao, H., 2023b. Shrinking class space for enhanced certainty in semi-supervised learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16187–16196.

Yang, L., Zhao, Z., Zhao, H., 2025. Unimatch v2: Pushing the limit of semi-supervised semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell..

Yao, Y., Shen, J., Xu, J., Zhong, B., Xiao, L., 2022. Cls: Cross labeling supervision for semi-supervised learning. arXiv preprint arXiv:2202.08502.

You, C., Dai, W., Min, Y., Staib, L., Duncan, J.S., 2023. Bootstrapping semi-supervised medical image segmentation with anatomical-aware contrastive distillation. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 641–653.

Zhang, R., Lai, P., Wan, X., Fan, D.J., Gao, F., Wu, X.J., Li, G., 2022. Lesion-aware dynamic kernel for polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 99–109.

Zhang, H., Zhang, S., 2025. Federated semi-supervised polyp image detection based on client feature alignment. Multimedia Syst. 31 (2), 159.

Zhao, X., Fang, C., Fan, D.J., Lin, X., Gao, F., Li, G., 2022. Cross-level contrastive learning and consistency constraint for semi-supervised medical image segmentation. In: 2022 IEEE 19th International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 1–5.

Zhao, B., Wang, C., Ding, S., 2024. CrossMatch: Enhance semi-supervised medical image segmentation with perturbation strategies and knowledge distillation. IEEE J. Biomed. Heal. Inform..

Zhao, X., Zhang, L., Lu, H., 2021. Automatic polyp segmentation via multi-scale subtraction network. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. Springer, pp. 120–130.

Zhao, Y., Zhou, T., Gu, Y., Zhou, Y., Zhang, Y., Wu, Y., Fu, H., 2025. WeakPolyp-SAM: Segment anything model-driven weakly-supervised polyp segmentation. Knowl.-Based Syst. (ISSN: 0950-7051) 113701.

Zheng, M., You, S., Huang, L., Wang, F., Qian, C., Xu, C., 2022. Simmatch: Semi-supervised learning with similarity matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14471–14481.

Zhu, Q., Sun, Y., Wu, Y., Zhu, H., Lin, G., Zhou, Y., Feng, Q., 2021. Whole-brain functional MRI registration based on a semi-supervised deep learning model. Med. Phys. 48 (6), 2847–2858.