Journal of Science and Technology – Data Science and Artificial Intelligence
Volume 36, Issue 1, October 2026, 049–059

Cross-Task Consistency-Guided Mean Teacher with Cross Pseudo Supervision for Semi-Supervised Polyp Segmentation

Thao Tran Thi Phuong[1], Lam Tran Tung[2], Tu Nguyen Ba[2]
Thao Tran Thi[2], Truong Pham Van[*2]

[1]*School of Information and Communications Technology, Hanoi University of Science and Technology, Ha Noi, Vietnam*
[2]*School of Electrical and Electronic Engineering, Hanoi University of Science and Technology, Ha Noi, Vietnam*
[*]*Corresponding author email: truong.phamvan@hust.edu.vn*

Abstract

Colorectal polyp segmentation is crucial for computer-aided diagnosis of colorectal cancer. However, the scarcity of pixel-level annotations remains a major challenge for supervised learning, as manual mask annotation is time-consuming and labor-intensive. To address this limitation, we propose CTC-MT, a semi-supervised framework that integrates Mean Teacher and Cross Pseudo Supervision for colorectal polyp segmentation with limited labeled data. The framework adopts a single-teacher, dual-student architecture, in which the teacher is updated using an exponential moving average of one student's parameters, while the two students exploit complementary supervision signals to learn from labeled and unlabeled images. To enhance spatial feature representation, we extend PMFNet with a Spatial Mixture-of-Experts block and a Mask-Guided Spatial Mixture-of-Experts block, enabling spatially adaptive expert routing. In addition, an auxiliary regression branch predicts signed distance maps and is coupled with the segmentation branch through a cross-task consistency loss to improve boundary representation by exploiting unlabeled data. We evaluate the proposed framework against five representative semi-supervised segmentation methods on five public polyp segmentation datasets under 5%, 15%, and 25% labeled-data settings, covering both seen and unseen benchmarks. On Kvasir-SEG, CTC-MT achieves Dice similarity coefficient scores of 87.72%, 90.35%, and 91.12% at the respective annotation ratios, obtaining the best Dice similarity coefficient and intersection over union results among the compared methods across all three settings. On the unseen ETIS dataset, CTC-MT achieves a Dice similarity coefficient of 75.39% with 15% labeled data, exceeding the Mean Teacher baseline by 9.40 percentage points. These results demonstrate the effectiveness of CTC-MT under limited-annotation conditions and indicate its potential to generalize across datasets with diverse polyp characteristics.

Keywords: Cross pseudo supervision, mean teacher, mixture of experts, polyp segmentation, semi-supervised learning.

1. Introduction

In the field of medical image analysis, image segmentation is a fundamental task that involves delineating regions of interest (ROI) such as lesions, organs, and tissues. Among these fields, polyp segmentation is a vital component as polyps are abnormal tissue growths that form on the inner wall of the colon, which can lead to the risk of colorectal cancer if they are not diagnosed and treated properly. However, precisely detecting and characterizing the polyp faces many challenges due to its variation in size, shape, texture, and its similarity to surrounding tissues. Recently, many deep learning methods have been studied to enhance the accuracy of segmentation models, such as UNet [1] based on Convolutional neural networks and Transformer-based methods [2, 3], which have gradually achieved significant success in medical image segmentation. Nevertheless, the high performance of these methods strongly depends on the well-annotated data and collecting sufficient annotated data for training, which is not always a feasible task. This problem highlights the need for greater focus on studying semi-supervised learning methods to utilize both labeled and unlabeled datasets. Researchers have proposed numerous semi-supervised methods, including adversarial learning [4], pseudo-labeling [5], and consistency regularization [6, 7].

The pseudo-labeling method assigns artificial labels to the unlabeled data without any guidance from the clinician, based on thresholds and softmax probabilities derived from model predictions. In the next step, both labeled and artificial data are combined to train the model, similarly to supervised learning. The consistency regularization method is developed based on the idea that the model must make a consistent prediction for various versions of an image, which can be corrupted by noise or augmentation. Many studies have recently shown promising performance by combining both semi-supervised learning methods. Introduced by Tarvainen *et al.*, Mean Teacher [8] is a semi-supervised learning method that leverages the teacher-student architecture. By combining pseudo-

p-ISSN 3093-3331
e-ISSN 3093-3323
https://doi.org/10.51316/jst.194.dsai.2026.36.1.6
Received: Aug 27, 2026; Revised: Oct 7, 2026;
Accepted: Oct 10, 2026; Online: Oct 15, 2026

labeling and consistency regularization, this method enhances the quality of the pseudo-label for unlabeled data, thereby improving the segmentation performance. Additionally, Cross Pseudo Supervision [9] stems from this concept by ensuring the consistency of two segmentation networks, one for prediction and the other for generating pseudo-label, and vice versa. Inspired by these approaches, we propose CTC-MT, a semi-supervised framework based on Mean Teacher and Cross Pseudo Supervision with enhanced PMFNet [10] architecture as the backbone model, aiming to achieve high accuracy segmentation performance under limited labeled data. Our main contributions are as follows:

- We combine Mean Teacher and CPS in a framework with one teacher and two students, where the teacher is updated using the parameters of a single student.
- We extend PMFNet with an S-MoE block and a mask-guided MGS-MoE block to enable spatial expert routing. We further introduce an SDF regression branch coupled with the segmentation branch through a cross-task consistency loss to learn boundary information from unlabeled images.
- Under scarce labeled data conditions, we evaluate the proposed method on the five public polyp segmentation datasets, including both standard benchmarks and seen-unseen data splits. The experimental results show the effectiveness of the proposed method and its reliable generalization capability across diverse clinical scenarios

2. Related Work

Polyp segmentation in colonoscopy images plays an important role in early colorectal cancer screening. Although supervised deep learning models have achieved remarkable segmentation performance, their effectiveness relies heavily on dense pixel-level annotations, which are labor-intensive and time-consuming to obtain in clinical settings. To alleviate this annotation bottleneck, Semi-Supervised Learning (SSL) has emerged as a promising paradigm for leveraging abundant unlabeled data alongside a limited amount of labeled data. Among various SSL approaches, consistency-based frameworks have demonstrated considerable effectiveness in medical image segmentation. Meanwhile, dynamic representation mechanisms, such as Mixture of Experts (MoE), provide adaptive model capacity for handling substantial variations in polyp morphology. This section reviews representative studies related to these foundational techniques.

2.1. Mean Teacher Framework

The Mean Teacher (MT) framework [8] is a foundational consistency-based semi-supervised learning paradigm. It maintains two networks with the same architecture: a Student model optimized through backpropagation and a Teacher model whose parameters are updated as the Exponential Moving Average (EMA) of the Student's parameters. For unlabeled data, a consistency loss is imposed between the predictions of the Teacher and Student under different perturbations. Extensions such as UA-MT [11] further incorporate uncertainty estimation to filter unreliable pseudo-labels. Although MT provides stable pseudo-label supervision, the EMA-based parameter coupling progressively reduces the diversity between the Teacher and Student representations, which may limit the exploration of complementary feature representations during prolonged training.

2.2. Cross Pseudo Supervision

To alleviate the parameter coupling inherent in conventional teacher-student frameworks, Cross Pseudo Supervision (CPS) [9] employs two independently initialized networks with identical architectures that are trained jointly. Each network generates pseudo-labels to supervise the other network on unlabeled samples, thereby establishing a cross-supervision mechanism. Compared with EMA-based teacher-student frameworks, the independent parameter updates in CPS maintain greater model diversity and encourage complementary representations. However, CPS remains vulnerable to error propagation during the early stages of training, as unreliable predictions from either network may be propagated to its counterpart in the absence of effective confidence-based filtering.

2.3. Mixture of Experts in Medical Imaging

Mixture of Experts (MoE) [12] increases model capacity by dynamically routing input features to specialized sub-networks, or experts, through a trainable gating mechanism. Sparse MoE variants, such as Switch Transformer [13], employ Top-1 routing together with load-balancing objectives to increase parameter capacity while controlling computational costs. In medical image segmentation, MoE-based architectures can dynamically adapt their representations to diverse anatomical structures, multi-scale contextual information, and complex background appearances. This adaptive routing mechanism is particularly appealing for polyp segmentation, where substantial variations in polyp size, shape, texture, and appearance pose significant challenges to a single static representation.

3. Methodology

3.1. The proposed Semi-supervised framework

Our proposed semi-supervised framework adopts a teacher-student [8] architecture combined with CPS [9]-style cross supervision. The backbone of our architecture is a modified PMFNet [10] model, with specific structural modifications detailed in the next section. The framework consists of one teacher model $F(\theta_T)$ and two distinct student models $F(\theta_{S1})$

and $F(\theta_{S2})$, where each model integrates two main functional components: a primary branch for the segmentation task and an auxiliary branch for the regression task. The overall architecture of the proposed framework is illustrated in Fig. 1.

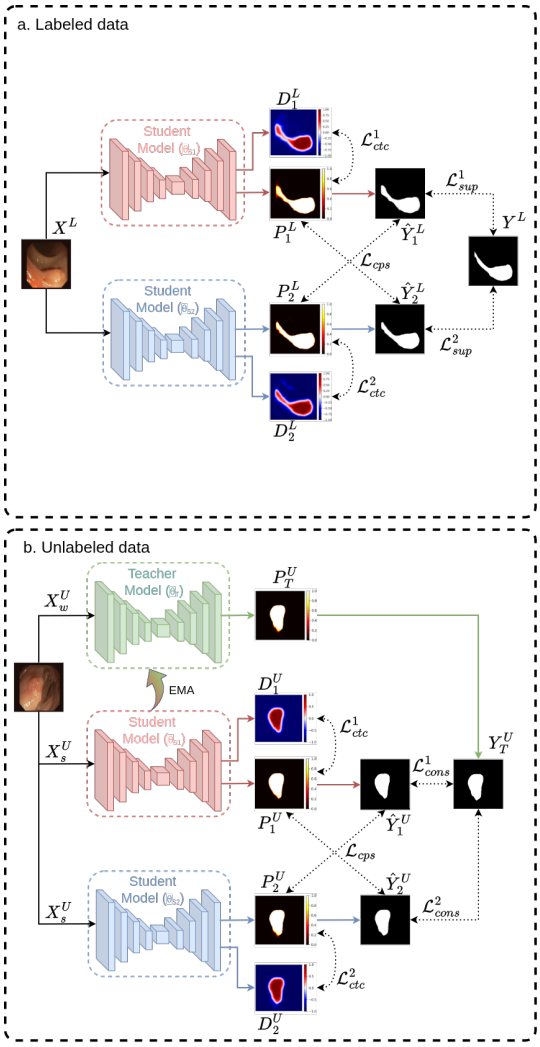


Fig. 1. Overview of the proposed framework

For the labeled dataset X_L, both student models are fully supervised. The primary branch is trained using the ground truth segmentation maps to establish reliable feature representations, while the auxiliary branch is trained on signed distance maps generated from the ground truth maps to determine whether each pixel is inside or outside the polyps.

For the unlabeled dataset X_U, we employ a co-training process guided by the teacher model. An unlabeled image X_U is perturbed into a weakly augmented version X_W and a strongly augmented version X_S. The teacher model uses X_W as input to generate reliable segmentation pseudo-labels (P_T) from its primary branch and pseudo signed distance maps (D_T) from its auxiliary regression branch. Concurrently, X_S is fed into both student models to produce their respective segmentation predictions (P_1, P_2) and signed distance map predictions (D_{S1}, D_{S2}). To enforce consistency and mutual learning, the student models undergo a co-training process, where their predictions are cross-compared and optimized against each other, while being regularized by the respective pseudo-labels generated by the teacher model. Furthermore, a cross-task consistency constraint is enforced within each model to bridge the gap between the segmentation and regression branches by pulling their geometric representations closer. The teacher model's weights θ_T are updated via an Exponential Moving Average (EMA) of Student 1 only. Specifically, $\theta_T^{(t)} = \alpha_t\,\theta_T^{(t-1)} + (1-\alpha_t)\,\theta_{S1}^{(t)}$, where t is the update step and $\alpha_t = \min(1-1/(t+1), 0.99)$. We initially evaluated an EMA update using both student models, in which the teacher receives the weights of both students, for example by replacing $\theta_{S1}^{(t)}$ with the average $(\theta_{S1}^{(t)} + \theta_{S2}^{(t)})/2$; however, immediately after the supervised warm-up, when the EMA update was first activated, the teacher's Dice score dropped to nearly zero, indicating severe instability in the teacher update. This instability is caused by the different initializations of the two students. Their encoders share the same ImageNet-pretrained weights, whereas their decoders are randomly initialized. Due to the permutation symmetry of neural networks, the hidden units of the two decoders are not aligned, so averaging their weights can produce a severely degraded model [14, 15]. Since the EMA decay is close to zero in the first update steps, the teacher is almost immediately replaced by this averaged model. We verified this by averaging the weights of the two trained students. Each student achieves a validation Dice of about 0.90, whereas the averaged model yields 0.00. Averaging only the encoders keeps the Dice at 0.91, while averaging only the decoders reduces it to 0.00. Since using Student 1 or Student 2 individually for EMA resulted in comparable performance, Student 1 was selected as the sole source for the teacher update.

3.2. Spatial Mixture of Experts (SMoE) Modules

We propose an enhancement to the baseline PMFNet [10] architecture by integrating 2 novel variants of the MoE [12] framework: the Spatial MoE (SMoE) block and Mask-Guided Spatial MoE (MGS-MoE) block. Fig. 2 illustrates the detailed architecture of the proposed backbone.

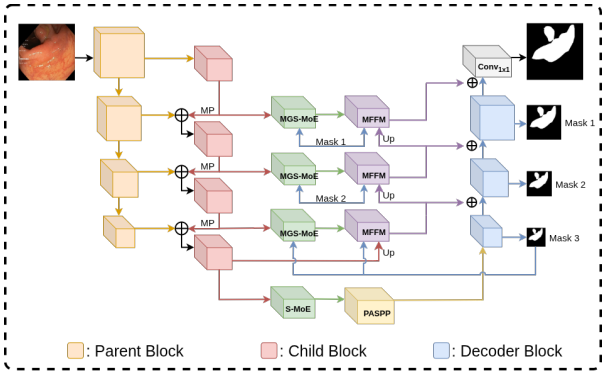


Fig. 2. Overview of the backbone architecture

3.2.1. Top-2 gating

The Top-2 Gating mechanism dynamically assigns each spatial token to at most two experts. Given an input feature X, the gating distribution is computed as $P = \mathrm{Softmax}(XW_g)$, where W_g is the learnable gating matrix. The top two expert indices (e_1, e_2) and their unnormalized probabilities (p_1, p_2) are extracted sequentially via a masked Top-1 selection. The selected probabilities are then normalized as:

$$\tilde{p}_k = \frac{p_k}{p_1 + p_2 + \varepsilon}, \quad k \in \{1,2\} \tag{1}$$

where ε is a small constant for numerical stability.

To prevent expert overload, an expert capacity constraint K limits the maximum number of tokens assigned to each expert. Following [16], the capacity is set to $K = \max(\min(N, \lfloor cN/E \rfloor), 4)$, where N is the number of tokens per image, E is the number of experts, and c is the capacity factor (1.25 for training and 2 for inference). Each token is first assigned to its top-1 expert and then to its top-2 expert, as long as the capacity of the expert is not exceeded. A token that exceeds the capacity is dropped by that expert, and if both of its assignments are dropped, it is passed to the next layer through the residual connection. Using (e_1, e_2) and $(\tilde{p}_1, \tilde{p}_2)$, the gating module constructs the dispatch tensor D and combine tensor C for token routing and expert output aggregation. An auxiliary load-balancing loss is further employed to encourage uniform expert utilization:

$$\mathscr{L}_{\text{aux}} = E^2 \cdot \mathrm{Mean}\left(P_{\text{avg}} \odot f_{\text{avg}}\right) \tag{2}$$

where P_{avg} and f_{avg} denote the average routing probability and the fraction of tokens assigned to each expert, respectively. The overall procedure is summarized in Algorithm 1.

3.2.2. Spatial mixture-of-experts block

The Spatial Mixture-of-Experts (S-MoE) block is positioned between the PASPP module and the bottleneck layer to dynamically refine high-level features. It treats each spatial position as an independent token, enabling fine-grained, position-aware selection across experts.

Algorithm 1: The Top-2 Gating algorithm

Input: Input feature map X, expert capacity K
Output: Dispatch tensor D, combine tensor C, auxiliary loss $\mathscr{L}_{\text{aux}}$

$P \leftarrow \mathrm{Softmax}(XW_g)$;
$(p_1, e_1) \leftarrow \mathrm{Top1}(P)$;
$P' \leftarrow P \odot (1 - \mathrm{OneHot}(e_1, E))$;
$(p_2, e_2) \leftarrow \mathrm{Top1}(P')$;
$\tilde{p}_1, \tilde{p}_2 \leftarrow \frac{p_1}{p_1 + p_2 + \varepsilon}, \frac{p_2}{p_1 + p_2 + \varepsilon}$;
$\mathscr{L}_{\text{aux}} \leftarrow E^2 \cdot \mathrm{Mean}(P_{\text{avg}} \odot f_{\text{avg}})$;
$D, C \leftarrow \mathrm{CapacityRouting}(e_1, e_2, \tilde{p}_1, \tilde{p}_2, K)$;
return $D, C, \mathscr{L}_{aux}$;

Given an input feature tensor $x \in \mathbb{R}^{B \times C \times H \times W}$, spatial dimensions are flattened and normalized before expert routing:

$$x_s = \mathrm{LayerNorm}(\mathrm{Flatten}_{HW}(x)) \tag{3}$$

where $x_s \in \mathbb{R}^{B \times HW \times C}$. The normalized spatial tokens are then routed through the Top-2 Gating mechanism to compute expert assignments and the load-balancing loss:

$$F_s, \mathscr{L}_{\text{aux}} = \mathrm{MoE}(x_s) \tag{4}$$

The processed features F_s are reshaped back to the original spatial dimensions and integrated with the input feature via a residual connection:

$$x_{\text{out}} = x + \mathrm{Reshape}_{HW}(F_s) \tag{5}$$

This design enriches global contextual representations while preserving fine-grained spatial details. The overall S-MoE procedure is detailed in Algorithm 2.

Algorithm 2: The S-MoE Algorithm

Input: Feature map $x \in \mathbb{R}^{B \times C \times H \times W}$
Output: Output feature map x_{out}, auxiliary loss $\mathscr{L}_{\text{aux}}$

$x_s \leftarrow \mathrm{LayerNorm}(\mathrm{Flatten}_{HW}(x))$;
$F_s, \mathscr{L}_{\text{aux}} \leftarrow \mathrm{MoE}(x_s)$;
$x_{\text{out}} \leftarrow x + \mathrm{Reshape}_{HW}(F_s)$;
return $x_{out}, \mathscr{L}_{aux}$;

3.2.3. Mask-guided spatial mixture-of-experts block

At lower intermediate layers, the Mask-Guided Spatial Mixture-of-Experts (MGS-MoE) block is introduced to leverage decoder-derived spatial guidance prior to expert routing. Unlike S-MoE, which performs routing directly on input features, MGS-MoE first enhances feature representations using an attention mask to focus expert selection on salient foreground regions.

Given an input feature map $x \in \mathbb{R}^{B \times C \times H \times W}$ and a decoder attention map M, M is first normalized to M_p and filtered using a hard thresholding operator ($\tau = 0.3$) to suppress background noise. Since M takes values in $[-1.5, 0.5]$, the threshold $\tau = 0.3$ keeps only

the positions where the decoder foreground probability is above 0.95 or below 0.05. This corresponds to a confidence threshold of 0.95, following the confidence threshold used in FixMatch [17], so uncertain predictions around the polyp boundary are not used to guide the expert routing. The filtered mask is resized to (H, W) and encoded via point-wise (PW) convolution followed by a GELU activation:

$$F_M = \mathrm{GELU}(\mathrm{PW}(\mathrm{Resize}(M_p \odot \mathbb{I}(M_p > \tau), H, W))) \tag{6}$$

The spatial guidance feature F_M is then injected into the input feature map x via a learnable scaling factor α, initialized to zero:

$$x' = x + \alpha F_M \tag{7}$$

Finally, the mask-enhanced feature x' is flattened, normalized, and processed through the Top-2 MoE routing mechanism before residual combination:

$$x_{\text{out}} = x' + \mathrm{Reshape}_{HW}\left(\mathrm{MoE}\left(\mathrm{LN}\left(\mathrm{Flatten}_{HW}(x')\right)\right)\right) \tag{8}$$

The overall architecture of the proposed MGS-MoE block is illustrated in Fig. 3, highlighting the mask-guided feature enhancement and subsequent Top-2 expert routing process.

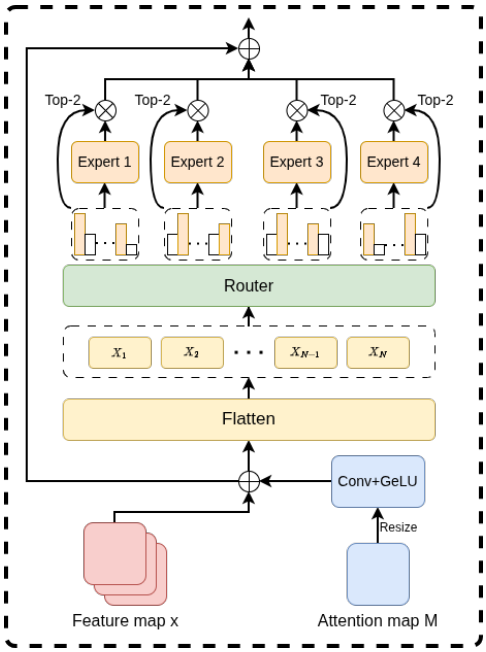


Fig. 3. MGS-MoE module

By incorporating task-specific guidance prior to gating, MGS-MoE encourages expert routing to focus on foreground semantics. The complete execution flow is detailed in Algorithm 3.

Algorithm 3: The MGS-MoE Algorithm

Input: Feature map $x \in \mathbb{R}^{B\times C\times H\times W}$, attention map M
Output: Output feature map x_{out}, auxiliary loss $\mathscr{L}_{\text{aux}}$

$M_p \leftarrow \mathrm{Normalize}(M)$;
$M_c \leftarrow M_p \odot \mathbb{I}(M_p > \tau)$;
$M_r \leftarrow \mathrm{Resize}(M_c, H, W)$;
$F_M \leftarrow \mathrm{GELU}(\mathrm{PW}(M_r))$;
$x' \leftarrow x + \alpha F_M$;
$x_s \leftarrow \mathrm{LN}(\mathrm{Flatten}_{HW}(x'))$;
$F_s, \mathscr{L}_{\text{aux}} \leftarrow \mathrm{MoE}(x_s)$;
$x_{\text{out}} \leftarrow x' + \mathrm{Reshape}_{HW}(F_s)$;
return $x_{out}, \mathscr{L}_{aux}$;

3.3. Consistency Loss Functions

The proposed model is optimized using a multi-task objective consisting of supervised and consistency regularization losses as:

$$\begin{aligned}\mathscr{L} = \mathscr{L}_{\text{s}} &+ \lambda_{\text{CPS}}\mathscr{L}_{\text{CPS}} + \lambda_{\text{mt}}\mathscr{L}_{\text{mt}} \\ &+ \lambda_{\text{CTC}}\mathscr{L}_{\text{CTC}} + \lambda_{\text{MoE}}\mathscr{L}_{\text{MoE}}\end{aligned} \tag{9}$$

where λ_{CPS}, λ_{mt}, λ_{CTC}, and λ_{MoE} are trade-off hyperparameters that balance the contributions of the corresponding loss terms.

For labeled data X_L, the supervised loss combines Cross-Entropy, Dice, and Smooth L_1 losses for the segmentation and regression branches, defined as:

$$\begin{aligned}\mathscr{L}_{\text{s}} = \mathscr{L}_{\text{ce}}(X_{\text{logit}}, Y_{\text{L}}) + \mathscr{L}_{\text{dice}}(P_{\text{prob}}, Y_{\text{L}}) \\ + 2 \cdot \mathscr{L}_{\text{sdf}}(D_{\text{pred}}, D_{\text{L}})\end{aligned} \tag{10}$$

where $P_{\text{prob}} = \mathrm{Softmax}(X_{\text{logit}})$ denotes the predicted polyp probability, while D_{L} and D_{pred} denote the target and predicted Signed Distance Maps (SDFs), respectively. Since the SDF targets are normalized to $[-1, 1]$, the Smooth L_1 loss is three to four times smaller than the Cross-Entropy and Dice losses (about 0.02 compared with 0.06 and 0.07 for the trained model on the labeled images of the 15% setting). The factor of 2 increases its share in $\mathscr{L}_{\text{s}}$ from about 13% to about 23%, so the regression branch receives more supervision while the segmentation losses still account for most of the objective. The regression branch shares the final decoder feature F_{dec} with the segmentation branch and uses an independent prediction head to regress the single-channel SDF map. The regression branch is summarized in Algorithm 4.

For unlabeled data X_U, consistency is enforced through Cross-Pseudo Supervision (CPS) and Mean Teacher (MT) losses based on Cross-Entropy. Specifically, pseudo-labels are generated from strongly augmented views of the unlabeled inputs. The CPS loss is then defined as:

$$\mathscr{L}_{\text{CPS}} = 0.5 \cdot \left(\mathscr{L}_{\text{ce}}(X^1_{\text{logit}}, \hat{Y}^U_2) + \mathscr{L}_{\text{ce}}(X^2_{\text{logit}}, \hat{Y}^U_1)\right) \tag{11}$$

Algorithm 4: The SDF Regression Branch

Input: Final decoder feature F_{dec}
Output: Predicted SDF map D_{pred}
$F_r \leftarrow \text{GELU}(\text{BN}(\text{Conv}_{3\times3}(F_{\text{dec}})))$;
$D_{\text{pred}} \leftarrow \text{Up}_{\times4}(\text{Conv}_{1\times1}(F_r))$;
return D_{pred};

where $\hat{Y}_i^U$ $(i \in \{1,2\})$ denotes the pseudo-labels generated by Student i under strong augmentation.

The Mean Teacher consistency loss is defined as:

$$\mathscr{L}_{\text{mt}} = \mathscr{L}_{\text{ce}}(X^1_{\text{logit}}, Y^U_{\text{T}}) + \mathscr{L}_{\text{ce}}(X^2_{\text{logit}}, Y^U_{\text{T}}) \tag{12}$$

where Y_{teacher} is the pseudo-label generated by the Teacher network.

To maintain cross-task consistency, the Cross-Task Consistency (CTC) loss enforces geometric alignment between the two branches within each model. The CTC loss enforces consistency between the segmentation and regression branches, as formulated as:

$$\mathscr{L}_{\text{CTC}}(X_{\text{seg}}, X_{\text{dis}}) = \frac{\sum_{i,h,w} w_{i,h,w}\big(P_{\text{seg}}(i,h,w) - P_{\text{sdf}}(i,h,w)\big)^2}{\sum_{i,h,w} w_{i,h,w}} \tag{13}$$

where P_{seg} and P_{sdf} denote the probabilistic representations derived from the segmentation and regression branches, respectively. Specifically, P_{seg} is the foreground probability of the segmentation branch, and $P_{\text{sdf}} = \sigma(t \cdot X_{\text{dis}})$, where σ is the sigmoid function, X_{dis} is the predicted SDF, and $t = 15$ matches the normalization radius of the SDF ($R = 15$ pixels). Since the SDF is positive inside the polyp, negative outside, and zero on the boundary, the sigmoid converts it into a soft mask in $[0,1]$, which lies in the same probability space as P_{seg}. This mask is detached from the computational graph, so the CTC loss only updates the segmentation branch. To handle the class imbalance between polyps and background, a pixel-wise weight matrix $w_{i,h,w}$ is dynamically computed as:

$$w_{i,h,w} = \begin{cases} 0.1 & \text{if } P_{\text{sdf}}(i,h,w) < 0.01 \\ 1.0 & \text{otherwise} \end{cases} \tag{14}$$

The condition $P_{\text{sdf}} < 0.01$ corresponds to pixels located more than about 4.6 pixels outside the polyp. These easy background pixels account for 84.6% of the pixels in the training images, while polyps account for only 13.1%. With equal weights, they would dominate the CTC loss (about 85%), and the weight of 0.1 reduces their share to about 35%, so that the loss focuses on the polyp and its boundary.

To guarantee balanced utilization of the experts within the MoE layers during both the supervised and semi-supervised phases, an auxiliary load-balancing loss $\mathscr{L}_{\text{MoE}}$ is applied at a given layer k, as defined in Eq.15:

$$\mathscr{L}_{\text{MoE}}(k) = \mathscr{L}_{\text{aux_moe}}(X_l(k)) + \mathscr{L}_{\text{aux_moe}}(X_s(k)) \tag{15}$$

where $X_l(k)$ and $X_s(k)$ are the layer-k input tokens from the labeled and strongly augmented unlabeled streams. For E experts, the auxiliary function $\mathscr{L}_{\text{aux_moe}}$ is defined as:

$$\mathscr{L}_{\text{aux_moe}} = E \cdot \sum_{e=1}^{E} f_e \cdot P_e \tag{16}$$

where $f_e = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(\text{Expert } e \text{ is chosen for token } i)$ and $P_e = \frac{1}{N}\sum_{i=1}^{N} \text{Gate}(x_i)_e$ represent the empirical fraction of tokens routed to expert e and its average routing probability assigned by the gating mechanism across N batch tokens, respectively.

Minimizing $\mathscr{L}_{\text{aux_moe}}$ promotes a uniform distribution of workloads, effectively preventing gating collapse where only a subset of experts is trained.

4. Experiment

4.1. Datasets

Following the data split released with PraNet [18], 1,450 colonoscopy images, including 900 images from Kvasir-SEG and 550 images from CVC-ClinicDB, are used for training, while the remaining 100 Kvasir-SEG images and 62 CVC-ClinicDB images form the seen test sets. In addition, ETIS (196 images), CVC-300 (60 images), and CVC-ColonDB (380 images) are used as unseen test sets. For each labeling ratio, the first 73, 220, or 363 images of the training set are used as labeled data, and the remaining images are used as unlabeled data. Semi-supervised training uses mini-batches of 4 labeled and 4 unlabeled images. All images are resized to 256 x 256 pixels and normalized with ImageNet statistics. Weak augmentations are applied to labeled data, while weak and strong augmentations are used for unlabeled data to support pseudo-label generation and consistency learning.

4.2. Training and Evaluation Metrics

The model is implemented in PyTorch and trained for 100 epochs using Adam with a learning rate of 1×10^{-4} and weight decay of 1×10^{-5}, together with a ReduceLROnPlateau scheduler. A 5-epoch supervised warm-up is applied before semi-supervised training, after which the Teacher is updated via EMA and the semi-supervised loss weights are gradually ramped up. Specifically, the maximum weights are set to $\lambda_{\text{CPS}} = 0.5$, $\lambda_{\text{mt}} = 0.1$, $\lambda_{\text{CTC}} = 0.1$, and $\lambda_{\text{MoE}} = 0.1$. The CPS loss is assigned the largest weight, since it provides the main supervision on unlabeled data, while the MT and CTC losses act as auxiliary regularizers. The CPS, MT, and CTC weights are increased using sigmoid ramp-up over 40, 20, and 20 epochs, respectively, while λ_{MoE} remains fixed. The CPS loss is ramped up more slowly because the pseudo-labels exchanged between the students are unreliable early in training, whereas the targets of the MT and CTC losses, provided by the EMA [8] teacher

Table 1. Computational cost of the proposed method and the semi-supervised frameworks built on the same PMFNet backbone, measured on a single Tesla T4 GPU with 256×256 inputs, where the training cost is recorded per iteration with 4 labeled and 4 unlabeled images and the inference cost with a batch size of 1

Methods	Networks	Training				Inference (single network)			
	(training / inference)	Total params (M)	Trainable params (M)	Peak mem. (GB)	Time/iter (s)	Params (M)	GFLOPs	Latency (ms)	FPS
Supervised PMFNet	1 / 1	49.5	49.5	2.65	0.20	49.47	21.93	44.1	22.7
MT (PMFNet)	2 / 1	98.9	49.5	4.36	0.47	49.47	21.93	44.1	22.7
CPS (PMFNet)	2 / 1	98.9	98.9	8.07	0.87	49.47	21.93	44.1	22.7
Proposed model	3 / 1	210.3	140.2	9.64	1.07	70.11	25.56	45.2	22.1

and the supervised SDF branch, are more stable. The MoE weight is kept constant, since the load-balancing loss does not depend on pseudo-labels. The sensitivity to these weights is analyzed in Section 4.6. Performance is evaluated using DSC, IoU, Recall, and Accuracy.

4.3. Computational Complexity

Table 1 reports the computational cost of the proposed method and of the semi-supervised frameworks built on the same PMFNet backbone. At inference time, only the teacher network is used. Compared with PMFNet, the MoE blocks and the SDF head increase the number of parameters by 20.6M, whereas the computational cost increases by only 3.63 GFLOPs, since each token is processed by at most two experts. Consequently, the inference speed of the proposed model is close to that of PMFNet (22.1 and 22.7 FPS, respectively). During training, the teacher is updated by EMA and does not require back-propagation, so one iteration of the proposed framework takes 1.07 s, which is approximately 23% longer than that of CPS. With a peak memory of 9.64 GB, the whole training process can be completed on a single Tesla T4 GPU within 2.7 h. These results indicate that the additional cost of the proposed framework mainly lies in the training stage and remains acceptable in practice.

4.4. Results and Comparison

The comparison includes five representative semi-supervised polyp segmentation methods: Mean Teacher (MT) [8], ICT [19], PolypMixNet [20], PedSemiSeg [21], and RD-Net [22]. The evaluation is conducted using the Dice Similarity Coefficient (DSC), Polyp Intersection over Union (IoU), Accuracy and Recall under three annotation ratios (5%, 15%, 25%).

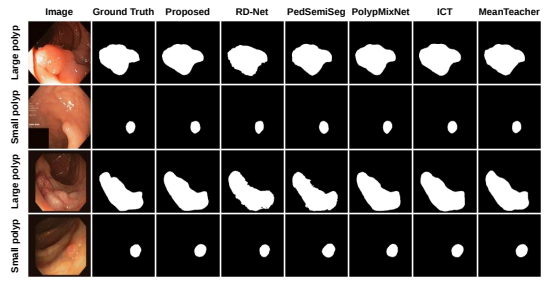


Fig. 4. Visualization results for two seen datasets: Kvasir-SEG and CVC-ClinicDB

Table 2 summarizes the segmentation performance on Kvasir-SEG and CVC-ClinicDB under three semi-supervised settings. The proposed method achieves the best DSC and IoU on Kvasir-SEG across all labeling ratios and the best Accuracy at the 5% and 25% settings, with the most notable gains observed at 5% labeled data, where it outperforms PolypMixNet [20] by 1.11% in DSC and 2.07% in IoU. On CVC-ClinicDB, the proposed method achieves the best DSC, IoU, and Accuracy at 5%, whereas Mean Teacher [8] achieves the best results at 15% and 25%. The DSC gap between Mean Teacher and the proposed method is small at 25% (0.59%) but larger at 15% (3.13%). These results confirm the effectiveness of the proposed method, particularly under low-annotation conditions.

To complement the quantitative evaluation, Fig. 4 presents qualitative segmentation results on representative large and small polyp samples from Kvasir-SEG and CVC-ClinicDB. For large polyps with irregular, lobulated boundaries, the mask produced by the proposed model most closely follows the contour of the ground truth. In contrast, RD-Net and PedSemiSeg tend to generate jagged, less smooth boundaries that slightly deviate from the true lesion shape, and PolypMixNet [20], ICT [19], and MT [8] occasionally over-smooth fine boundary details, losing subtle lobulation cues that are clinically relevant for polyp delineation. For small polyps, all competing methods, including the proposed approach, yield visually consistent, accurate predictions that closely match the ground truth.

Table 3 summarizes the generalization performance on three unseen benchmarks, ETIS, CVC-300, and CVC-ColonDB, under three semi-supervised settings with 5%, 15%, and 25% labeled data. Overall, the proposed method achieves the competitive performance across most evaluation metrics, indicating strong cross-domain generalization. Under the 5% labeled setting, our method achieves the best DSC, IoU, and Accuracy on ETIS and CVC-300. Compared with PolypMixNet [20], it improves the DSC by 0.91% on ETIS and 3.17% on CVC-300. As the labeled ratio increases to 15%, the proposed method achieves the best DSC, IoU, and Accuracy on all three unseen datasets, with the most notable improvement observed on ETIS, where the DSC reaches 75.39%, representing an increase of 9.40% over Mean Teacher [8]. At the 25% labeled setting, our method continues to achieve the best

Table 2. Performance (%) on seen datasets (Kvasir-SEG and CVC-ClinicDB) under 5% (73 labeled/1,377 unlabeled), 15% (220 labeled/1,230 unlabeled), and 25% (363 labeled/1,087 unlabeled) settings. Bold values with gray background indicate the best performance for each metric within each labeling ratio setting

Methods	Data		Kvasir-SEG				CVC-ClinicDB			
	Unlabeled	Labeled	DSC	IoU	Accuracy	Recall	DSC	IoU	Accuracy	Recall
MT [8]	1377	73	86.43	79.32	96.52	86.09	79.27	71.46	97.26	79.69
ICT [19]	1377	73	84.57	76.51	95.55	83.88	77.27	69.40	96.93	78.16
PolypMixNet [20]	1377	73	86.61	79.03	96.28	**89.14**	80.85	73.13	97.32	**84.07**
PedSemiSeg [21]	1377	73	81.01	72.57	94.56	79.51	76.55	67.15	96.90	76.30
RD-Net [22]	1377	73	76.85	65.97	93.83	80.88	69.33	59.55	95.95	74.04
Proposed model	1377	73	**87.72**	**81.10**	**96.61**	86.98	**82.74**	**75.58**	**97.53**	83.41
MT [8]	1230	220	90.07	84.21	**97.39**	89.52	**87.14**	**80.92**	**98.10**	**87.96**
ICT [19]	1230	220	88.71	81.67	96.97	87.94	82.02	74.71	97.43	84.93
PolypMixNet [20]	1230	220	90.28	84.02	97.08	**89.89**	86.64	79.49	97.97	86.37
PedSemiSeg [21]	1230	220	81.22	72.37	94.95	79.69	71.67	62.55	96.78	73.29
RD-Net [22]	1230	220	81.66	72.78	95.10	82.57	77.87	69.35	97.19	75.91
Proposed model	1230	220	**90.35**	**84.81**	97.22	88.09	84.01	77.88	97.84	83.69
MT [8]	1087	363	90.17	84.55	97.70	88.88	**86.82**	**81.66**	**98.39**	**86.81**
ICT [19]	1087	363	89.43	82.91	96.74	87.32	84.15	78.15	97.99	82.43
PolypMixNet [20]	1087	363	90.30	84.62	97.49	89.22	83.57	77.30	97.87	84.17
PedSemiSeg [21]	1087	363	84.98	77.84	96.31	83.19	82.04	74.58	97.72	83.99
RD-Net [22]	1087	363	84.14	75.35	95.76	83.59	80.00	71.13	97.25	80.62
Proposed model	1087	363	**91.12**	**85.90**	**97.89**	**90.30**	86.23	80.61	98.03	86.78

DSC, IoU, and Accuracy on ETIS and CVC-ColonDB, while remaining competitive on CVC-300, where the performance gap from the best-performing method is below 1% in DSC and IoU. Overall, these results demonstrate that the proposed framework learns more transferable feature representations and generalizes effectively to unseen domains.

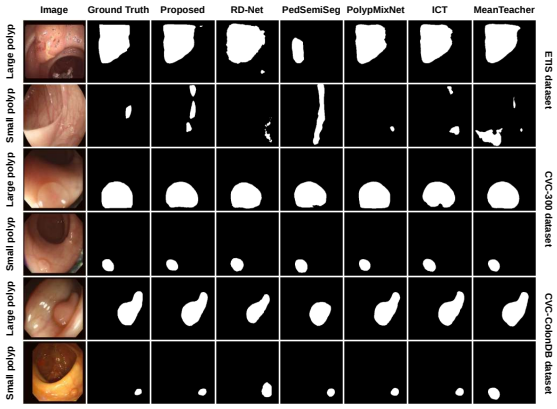


Fig. 5. Visualization results for three unseen datasets: ETIS, CVC-300, and CVC-ColonDB.

Fig. 5 further illustrates these results through representative qualitative examples on the three unseen datasets. On large polyps from ETIS, RD-Net [22] introduces a spurious disconnected fragment near the main lesion boundary, while PedSemiSeg [21] severely under-segments the polyp, capturing only a small fraction of the true lesion area. In contrast, the proposed model produces a mask that is well-aligned with the ground truth contour. The challenge is most pronounced for small polyps in ETIS, where the lesion occupies only a few pixels and exhibits low contrast with the surrounding mucosa. PolypMixNet [20] produces a very small prediction to detect the lesion, producing a negligibly small prediction, whereas RD-Net [22] and MT [8] generate fragmented, noisy false-positive regions scattered across the image. The proposed method, by comparison, recovers an elongated mask that more closely approximates both the location and shape of the ground truth, illustrating its improved sensitivity to small, low-contrast lesions under cross-domain shift. Overall, these qualitative observations indicate the performance advantage of the proposed framework on unseen data with different shapes of polyp.

4.5. Ablation Study

Table 4 reports the ablation results under the 15% labeled setting. Regarding the training strategy, the MT configuration uses the Mean Teacher framework without CPS, whereas the CPS configuration uses Cross-Pseudo Supervision without the teacher. Both configurations retain the same network components as the proposed model. Combining MT and CPS improves DSC over MT and CPS individually on all datasets except CVC-ClinicDB. The largest improvement is observed on ETIS, where the DSC increases by 9.40 and 3.08 percentage points compared with MT and CPS, respectively. Regarding the network architecture, the MT + CPS (PMFNet) configuration uses the original PMFNet without the proposed S-MoE and MGS-MoE blocks, SDF branch, and CTC loss. Compared with this baseline, the proposed model improves DSC

Table 3. Performance (%) on unseen datasets (ETIS, CVC-300, and CVC-ColonDB) under 5% (73 labeled/1,377 unlabeled), 15% (220 labeled/1,230 unlabeled), and 25% (363 labeled/1,087 unlabeled) settings. Bold values with gray background indicate the best performance for each metric within each labeling ratio setting

Methods	Data		ETIS				CVC-300				CVC-ColonDB			
	Unlabeled	Labeled	DSC	IoU	Accuracy	Recall	DSC	IoU	Accuracy	Recall	DSC	IoU	Accuracy	Recall
MT [8]	1377	73	59.88	51.85	98.09	64.17	84.29	76.36	98.77	91.19	70.57	61.62	96.21	72.14
ICT [19]	1377	73	59.22	50.77	98.20	65.41	86.78	78.55	99.18	92.96	71.82	62.34	96.21	72.09
PolypMixNet [20]	1377	73	63.92	55.17	98.21	**70.97**	84.72	76.20	98.92	93.78	**72.15**	62.99	**96.55**	**75.11**
PedSemiSeg [21]	1377	73	53.76	45.67	97.64	59.14	82.47	72.75	98.78	**94.67**	67.28	57.14	95.76	69.86
RD-Net [22]	1377	73	47.94	38.85	97.00	54.73	79.24	68.20	98.47	87.65	59.33	48.52	95.43	58.67
Proposed model	1377	73	**64.83**	**56.56**	**98.47**	69.21	**87.89**	**80.84**	**99.24**	94.30	72.03	**63.87**	96.37	72.75
MT [8]	1230	220	65.99	58.10	97.58	74.48	86.44	79.45	98.84	74.48	77.36	68.67	96.22	**82.74**
ICT [19]	1230	220	62.14	52.57	97.12	78.30	83.37	74.51	98.57	**95.97**	71.60	61.94	95.77	76.21
PolypMixNet [20]	1230	220	63.41	55.88	97.40	71.23	86.76	79.54	99.01	94.68	74.82	66.26	96.24	78.20
PedSemiSeg [21]	1230	220	51.73	43.31	96.94	58.86	76.88	65.90	98.01	88.90	61.98	51.85	94.98	65.81
RD-Net [22]	1230	220	57.65	49.08	97.77	63.94	85.08	76.20	99.12	85.08	68.69	58.32	96.05	68.65
Proposed model	1230	220	**75.39**	**66.75**	**98.70**	**82.24**	**88.12**	**81.05**	**99.22**	93.22	**78.89**	**70.97**	**96.90**	79.52
MT [8]	1087	363	72.09	63.93	97.93	**80.67**	88.00	80.91	99.05	95.50	78.18	69.99	96.90	80.44
ICT [19]	1087	363	68.99	61.12	98.27	77.00	87.89	80.47	99.13	93.46	73.92	66.12	96.48	74.33
PolypMixNet [20]	1087	363	65.72	58.34	98.42	70.50	**89.08**	**81.89**	**99.36**	94.65	75.18	67.85	96.88	75.80
PedSemiSeg [21]	1087	363	55.94	48.18	95.62	63.86	81.87	71.34	98.60	**97.14**	71.04	61.17	95.54	77.89
RD-Net [22]	1087	363	57.89	49.13	97.42	64.68	84.51	75.36	98.96	84.32	67.30	57.49	95.93	66.52
Proposed model	1087	363	**73.92**	**66.03**	**98.72**	78.49	88.11	81.36	99.17	95.34	**80.26**	**72.21**	**96.92**	**83.08**

Table 4. Ablation study of training strategies and proposed components at 15% labeled setting. Best performance per column is in bold with gray highlight.

Method	S-MoE & MGS-MoE	SDF & CTC	Kvasir-SEG	CVC-ClinicDB	ETIS	CVC-300	CVC-ColonDB
MT	✓	✓	90.07	**87.14**	65.99	86.44	77.36
CPS	✓	✓	89.44	84.59	72.31	85.67	76.19
MT + CPS (PMFNet)			**90.81**	84.56	72.23	87.37	78.74
MT + CPS (Proposed)	✓	✓	90.35	84.01	**75.39**	**88.12**	**78.89**

Table 5. Loss Weight Sensitivity at 15% Labeled Data (DSC, %). Default values are marked with $^*S$

Weight	Value	Kvasir-SEG	CVC-ClinicDB	ETIS	CVC-300	CVC-ColonDB
λ_{CPS}	0	85.09	82.54	61.02	82.83	74.08
	0.5*	**85.80**	83.68	**61.89**	**84.43**	74.71
	1.0	84.73	**84.42**	59.29	81.82	**75.56**
λ_{mt}	0	84.45	**85.72**	**67.20**	83.70	73.50
	0.1*	85.80	83.68	61.89	84.43	74.71
	0.5	**87.19**	85.70	65.38	**87.41**	**76.57**
λ_{CTC}	0	85.77	**85.07**	63.29	**84.91**	**75.25**
	0.1*	**85.80**	83.68	61.89	84.43	74.71
	0.5	85.23	83.43	**65.32**	83.68	74.26

on the three unseen datasets, with the largest gain of 3.16 percentage points on ETIS. However, the original PMFNet baseline performs slightly better on Kvasir-SEG and CVC-ClinicDB.

4.6. Hyperparameter Sensitivity

We investigate the sensitivity of the framework to three loss weights, namely λ_{CPS}, λ_{mt}, and λ_{CTC}, under the 15% labeled setting, while keeping all other experimental settings unchanged. Specifically, only one of these three weights is varied at a time, with the remaining weights fixed at their default values. The MoE loss weight, λ_{MoE}, is kept fixed at its default value throughout the experiment. All configurations are retrained under the same protocol; therefore, the absolute results are not directly comparable with those reported in Table 2 and Table 3. As shown in Table 5, removing the CPS loss reduces the DSC on all five test sets, while the default value of $\lambda_{\text{CPS}} = 0.5$ yields the best results on Kvasir-SEG, ETIS, and CVC-300. Varying λ_{CTC} leads to DSC variations of less than 2 percentage points on all test sets except ETIS. In contrast, increasing λ_{mt} to 0.5 improves the DSC on all five test sets, suggesting that stronger teacher consistency may further benefit the framework. We retain the default weights in the main experiments to ensure consistency with the reported results. Overall, the framework exhibits reasonable robustness to moderate changes in the three investigated loss weights, although further tuning of λ_{mt} may yield additional improvements.

5. Conclusion

In this paper, we propose CTC-MT, a novel semi-supervised framework for polyp segmentation that integrates Mean Teacher, Cross-Pseudo Supervision, Cross-Task Consistency, and Mixture of Experts into a unified architecture. The proposed S-MoE and MGS-MoE blocks facilitate feature learning by capturing global contextual information and refining intermediate features through mask guidance. Experimental results on multiple benchmark datasets demonstrate that CTC-MT achieves competitive segmentation performance under limited annotation settings, with Dice scores of 87.72%, 90.35%, and 91.12% on Kvasir-SEG using 5%, 15%, and 25% labeled data, respectively. Moreover, under the 15% labeled setting, CTC-MT achieves a Dice score of 75.39% on the unseen ETIS dataset, outperforming the Mean Teacher baseline by 9.40 percentage points.

Acknowledgements

This research is funded by the Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.05-2025.22.

Reference

[1] O. Ronneberger, P. Fischer, and T. Brox, U-Net: Convolutial networks for biomedical image segmentation, in International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 234–241, Nov. 2015.

[2] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, PVT: A versatile backbone for dense prediction without convolutions, in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 568–578, 2021.

[3] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, PVT v2: Improved baselines with pyramid vision transformer, Computational Visual Media, vol. 8, iss. 3, pp. 415–424, Sep. 2022. https://doi.org/10.1007/s41095-022-0274-8

[4] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang, Adversarial learning for semi-supervised semantic segmentation, in Proceedings of the British Machine Vision Conference (BMVC), Feb. 2018.

[5] Y. Zou, Z. Zhang, H. Zhang, C.-L. Li, X. Bian, J.-B. Huang, and T. Pfister, PseudoSeg: Designing pseudo labels for semantic segmentation, in International Conference on Learning Representations (ICLR), 2021.

[6] V.-T. Nguyen, G.-B. Truong, V.-T. Pham, and T.-T. Tran, 1S-MambaMatch: A semi-supervised and one-shot learning framework with multi-input visual state space model for skin lesion segmentation, Biomedical Signal Processing and Control, vol. 119, Part. B, Jun. 2026, Art. no. 109875. https://doi.org/10.1016/j.bspc.2026.109875

[7] J. Wang, C. H. Q. Ding, S. Chen, C. He, and B. Luo, Semi-supervised remote sensing image semantic segmentation via consistency regularization and average update of pseudo-label, Remote Sensing, vol. 12, iss. 21, pp. 3603, Nov. 2020. https://doi.org/10.3390/rs12213603

[8] A. Tarvainen and H. Valpola, Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, in Advances in Neural Information Processing Systems (NeurIPS), pp. 1195–1204, 2017.

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

[10] T.-L. Tran, M.-T. Pham, T.-T. Tran, and V.-T. Pham, Polyp segmentation with transformer-CNN integration and multi-scale feature fusion, pp. 224–229, Jul. 2025 . https://doi.org/10.1109/ATiGB66719.2025.11142182

[11] L. Yu, S. Wang, X. Li, C.-W. Fu, and P.-A. Heng, Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation, pp. 605–613, Oct. 2019 . https://doi.org/10.1007/978-3-030-32245-8_67

[12] S. Masoudnia and R. Ebrahimpour, Mixture of experts: a literature survey, Artificial Intelligence Review, vol. 42, no. 2, pp. 275–293, May 2014. https://doi.org/10.1007/s10462-012-9338-y

[13] W. Fedus, B. Zoph, and N. Shazeer, Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity, Journal of Machine Learning Research, vol. 23, no. 120, pp. 1–39, 2022.

[14] J. Frankle, G. K. Dziugaite, D. Roy, and M. Carbin, Linear mode connectivity and the lottery ticket hypothesis, in International Conference on Machine Learning (ICML), pp. 3259–3269, 2020.

[15] R. Entezari, H. Sedghi, O. Saukh, and B. Neyshabur, The role of permutation invariance in linear mode connectivity of neural networks, in International Conference on Learning Representations (ICLR), 2022.

[16] D. Lepikhin, H. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, M. Krikun, N. Shazeer, and Z. Chen, GShard: Scaling giant models with conditional computation and automatic sharding, in International Conference on Learning Representations (ICLR), 2021.

[17] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, FixMatch: Simplifying semi-supervised learning with consistency and confidence, in Advances in Neural Information Processing Systems (NeurIPS), pp. 596–608, 2020.

[18] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, PraNet: Parallel reverse attention network for polyp segmentation, in International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 263–273, 2020.

[19] V. Verma, K. Kawaguchi, A. Lamb, J. Kannala, A. Solin, Y. Bengio, and D. Lopez-Paz, Interpolation consistency training for semi-supervised learning, Neural Networks, vol. 145, pp. 90–106, 2022.

[20] X. Jia, Y. Shen, J. Yang, R. Song, W. Zhang, M. Q.-H. Meng, J. C. Liao, and L. Xing, PolypMixNet: Enhancing semi-supervised polyp segmentation with polyp-aware augmentation, Computers in Biology and Medicine, vol. 170, Jan. 2024, Art. no. 108006. https://doi.org/10.1016/j.compbiomed.2024.108006

[21] A. Wang, H. Ma, L. Bai, Y. Wu, M. Xu, Y. Zhang, M. Islam, and H. Ren, PedSemiSeg: Pedagogy-inspired semi-supervised polyp segmentation, Computerized Medical Imaging and Graphics, vol. 124, Sep. 2025, Art.no. 102591.

[22] Y. Li, Z. Zhu, Y. Zhang, Y. Chen, and Z. Yu, Boost the inference with co-training: A depth-guided mutual learning framework for semi-supervised medical polyp segmentation, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10394–10403, 2025.