Boost the Inference with Co-training: A Depth-guided Mutual Learning Framework for Semi-supervised Medical Polyp Segmentation

Yuxin Li[1] Zihao Zhu[1] Yuxiang Zhang[1] Yifan Chen[1] Zhibin Yu[1,2*]
[1]College of Electronic Engineering, Ocean University of China
[2]Key Laboratory of Ocean Observation and Information of Hainan Province,
Sanya Oceanographic Institution, Ocean University of China

Abstract

Semi-supervised polyp segmentation has made significant progress in recent years as a potential solution for computer-assisted treatment. Since depth images can provide extra information other than RGB images to help segment these problematic areas, depth-assisted polyp segmentation has gained much attention. However, the utilization of depth information is still worth studying. The existing RGB-D segmentation methods rely on depth data in the inference stage, limiting their clinical applications. To tackle this problem, we propose a semi-supervised polyp segmentation framework based on the mean teacher architecture. We establish an auxiliary student network with depth images as input in the training stage, and we propose a depth-guided cross-modal mutual learning strategy to promote the learning of complementary information between different student networks. Meanwhile, we use the high-confidence pseudo-labels generated by the auxiliary student network to guide the learning progress of the main student network from different perspectives. Our model does not need depth data in the inference phase. In addition, we introduce a depth-guided patch augmentation method to improve the model's learning performance in difficult regions of unlabeled polyp images. Experimental results show that our method achieves state-of-the-art performance under different label conditions on five polyp datasets. The code is available at https://github.com/pingchuan/RD-Net.

1. Introduction

Colorectal cancer (CRC) is a common gastrointestinal malignancy and has become the third most common cancer [20]. Various screening methods (such as colonoscopy) are proposed for early detection and treatment to reduce its incidence [18]. Intestinal polyps are often considered a precursor to colorectal cancer [25]. Therefore, the precise segmentation of polyps plays a key role in the early diagnosis and treatment of colorectal cancer. Recently, supervised deep learning methods have achieved remarkable success in the field of polyp segmentation [9, 23, 24, 29, 33]. However, these methods rely on high-quality pixel-wise labeled data. Collecting sufficient labeled data in the medical field is not always feasible. To alleviate this limitation, some studies have explored semi-supervised learning (SSL) [3, 16, 34, 36, 41] methods for medical image segmentation, using unlabeled data to improve the segmentation performance of the model. These semi-supervised learning (SSL) methods mainly rely on RGB images for training, neglecting the rich spatial information in the image. Studies have shown that combining RGB images with depth data can efficiently complement the contour information that can not be easily extracted in RGB images [6, 11, 14, 40]. As shown in Figure 1, the pseudo-labels predicted by the network based on depth images and RGB images can complement each other during training. However, these RGB-D segmentation methods usually require complex feature splicing and fusion in the training phase. This factor leads to reliance on depth data input during the inference phase, greatly limiting their application in clinical practice. To this end, we propose a semi-

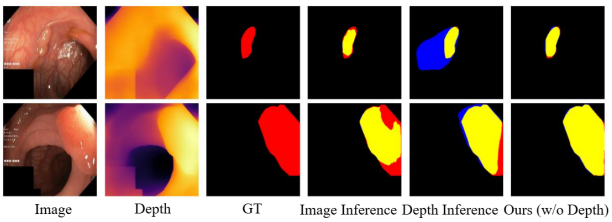

Figure 1. To illustrate the complementary role of RGB and depth images, we present segmentation results on unlabeled images trained with 10% labeled data in the Kvasir-SEG dataset. The results indicate that our method effectively learns richer information from unlabeled data. The red, blue, and yellow areas represent the ground truth, prediction, and their overlap, respectively.

*Corresponding author: yuzhibin@ouc.edu.cn.

supervised polyp segmentation framework RD-Net based on mean teacher, which effectively takes advantage of RGB and depth images during training and only needs the RGB images without depth information during inference. Specifically, in the training phase, we introduce an auxiliary student network with depth images as input into the traditional teacher-student network framework and adopt the proposed depth-guided cross-modal mutual learning strategy to effectively promote the learning of complementary information between different student networks. Our model uses the high-confidence pseudo-labels generated by the auxiliary student network to guide the training of the main student network from different perspectives. In addition, we propose depth-guided patch augmentation to improve the model's ability to learn difficult areas of polyp images. We summarize our main contributions as follows:

- A novel semi-supervised polyp segmentation framework is proposed to promote the learning of complementary information between student networks through a depth-guided cross-modal mutual learning strategy. No depth information is required in the inference stage.
- The pseudo-labels generated by the auxiliary student network are additionally used to guide the learning of the main student network from different perspectives, making the most of pixel information.
- Depth-guided patch augmentation is proposed to boost model learning on challenging regions in unlabeled data. Our method outperforms state-of-the-art methods on five challenging polyp datasets.

2. Related Work

2.1. RGB-D semantic segmentation

RGB-D information contains RGB features such as color texture and depth features such as distance boundaries [8], which are complementary to some extent. Some studies have explored how to integrate RGB-D information [1, 8, 11, 14, 39] to improve the performance of semantic segmentation. ACNet [14] effectively improved the semantic segmentation performance by leveraging the proposed attention complementary modality and using the channel attention mechanism to select bimodal features for fusion at multiple stages. CMX [1] achieved the calibration of different modal features by utilizing the designed cross-modal feature correction module to correct the features of another modality using the features of one modality and effectively fuses the features of different modalities through the proposed feature fusion module. However, these methods often perform complex feature splicing and fusion in the encoder stage. They still rely on depth images in the inference stage, greatly limiting their clinical applications. Although MaskMentor [43] adopted self-training token-pixel joint reconstruction to narrow the intermediate representation between the missing modality and the complete modality, thereby enhancing the model's ability to handle the modality missing to some extent. The performance of MaskMentor would decrease if modality was missing. Instead, our method can effectively learn the information between different modalities during the training phase without involving feature fusion between different modalities. Notably, our model does not require depth information during the inference phase.

2.2. Semi-supervised medical segmentation

In recent years, a large number of excellent methods have emerged in the field of semi-supervised medical image segmentation [3, 12, 16, 32, 34, 41, 44]. SCPNet [41] proposed a cross-sample prototype learning method to enhance the diversity of predictions in consistency learning by integrating rich semantic information from multiple inputs. BCP [3] encourages unlabeled data to learn richer language information from labeled data by bidirectionally copying and pasting labeled and unlabeled data. ACL-Net [34] proposed a semi-supervised segmentation framework that combines affinity contrast learning and self-training learning, which can adaptively refine pseudo-labels with optimized affinity to improve the performance of semi-supervised polyp segmentation. PH-Net [16] evaluated difficult regions by calculating the hardness scores of image patches and guides the model's attention to difficult regions, thereby improving segmentation performance. However, this method used the model's prediction results to identify difficult regions, which are susceptible to network performance. In contrast, we propose a depth-guided patch enhancement that does not rely on the network and can accurately identify difficult regions, thereby strengthening the learning of difficult regions. In addition, we introduce an auxiliary student network trained with depth images and promote the learning of complementary information between different modal networks through a deep-guided cross-modal mutual learning strategy. At the same time, the pseudo-labels of the auxiliary student network are used to adaptively guide the main student network, thereby utilizing more pixel information.

3. Method

3.1. Preliminary

In semi-supervised polyp segmentation, the dataset consists of two parts: labeled images and unlabeled images. For the labeled images $\mathcal{D}_l = \{(X_i, Z_i, Y_i)\}_{i=1}^{N}$, the unlabeled images $\mathcal{D}_u = \{X_i, Z_i\}_{i=N+1}^{N+M}$, where $X_i \in \mathbb{R}^{H_i \times W_i \times 3}$ is the image, $Z_i \in \mathbb{R}^{H_i \times W_i \times 3}$ is the corresponding depth image, and $Y_i \in \{0,1\}^{H_i \times W_i \times 1}$ represents the ground truth. For the images X_i, we use a monocular depth estimation network to infer grayscale depth maps of the same size as the original images. Specifically, we use the Depth Anything model [37] and its pre-trained model Depth-Anything-

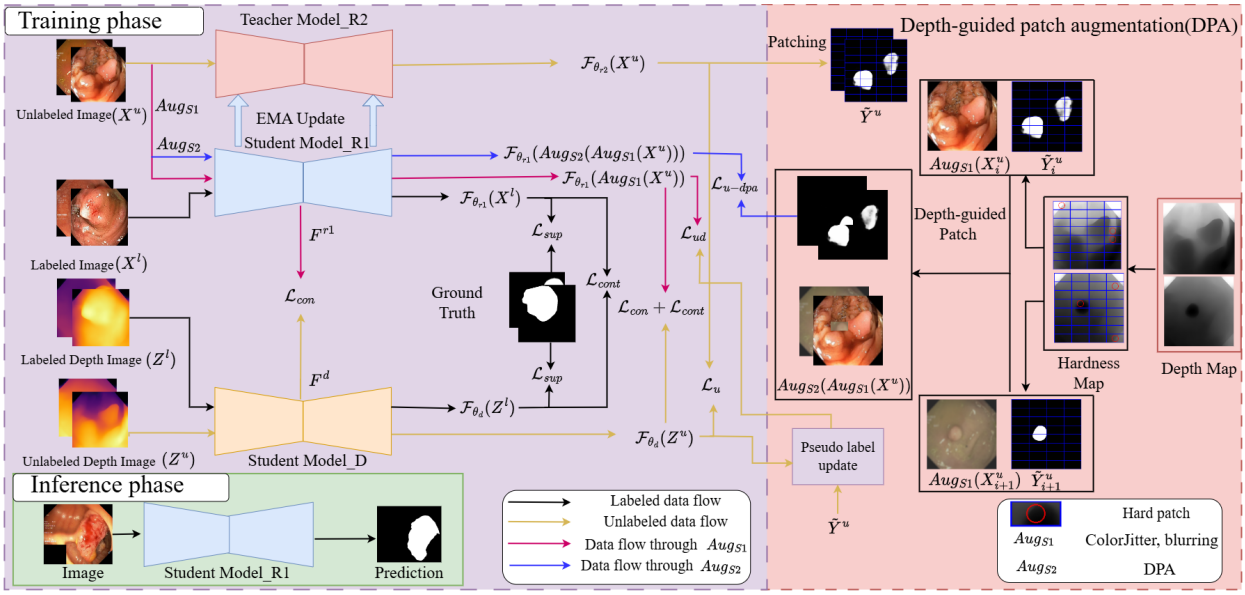

Figure 2. The overall framework of the proposed RD-Net. $\mathcal{L}_{\text{sup}}$ is used to guide the student to learn from labeled data; $\mathcal{L}_{\text{con}}$ and $\mathcal{L}_{\text{cont}}$ are used to promote different student networks to learn from each other from consistent and different predictions; the high-confidence pseudo-labels generated by Model_D guide Model_R1 to learn through $\mathcal{L}_{\text{ud}}$, and the student network Model_R1 can strengthen the learning of difficult areas through depth-guided patch augmentation.

Large to infer about the dataset used in our work directly. Subsequently, the grayscale depth maps are converted by applying color mapping into three-channel depth images, resulting in the final corresponding depth image.

Our RD-Net framework is shown in Figure 2, which is based on the classic semi-supervised learning framework mean teacher (MT) [28]. It consists of the teacher network Model_R2 and the main student network Model_R1, both based on RGB image input, and an auxiliary student network Model_D based on depth image input. The three networks use the same UNet network architecture, and the network output is realized through the Sigmoid function. Two student networks will be used for training and optimization, and the exponential moving average (EMA) will be used to update the teacher network [28]. Since the two student networks learn independently during training, and only constraints are imposed at the loss level, only Model_R1 is used for inference without depth information. We define weak image augmentation (flipping, scaling, and cropping) operations as Aug_W and strong image augmentation (color jitter, blurring) operations as Aug_{S1}. By default, all image data have been subjected to Aug_W operations.

For the input image $\{X_i, Z_i\}$, the outputs of the teacher network and the two student networks can be expressed as $\mathcal{F}_{\theta_{r2}}(X_i)$, $\mathcal{F}_{\theta_{r1}}(X_i)$, and $\mathcal{F}_{\theta_d}(Z_i)$. For the labeled data, the two student models are trained with cross-entropy and Dice loss, they are formulated as:

$$\mathcal{L}_{\text{sup}} = \frac{1}{2|\mathcal{D}_l|} \sum \left(\mathcal{L}_{ce}(\tilde{Y}_i^l, Y_i) + \mathcal{L}_{dice}(\tilde{Y}_i^l, Y_i) \right) \quad (1)$$

where $\tilde{Y}_i^l$ represents the prediction of the student network for labeled data. For unlabeled data, the output of the teacher network is used as the pseudo-label. Since unreliable pseudo-labels are unsuitable for supervision, we use the confidence threshold method to filter the pseudo-labels. The unsupervised loss can be formulated as:

$$\mathcal{L}_{\text{u}} = \frac{1}{2|\mathcal{D}_u|} \sum \left(\mathcal{L}_{ce}(\tilde{Y}_i^u, \tilde{Y}_i) + \mathcal{L}_{dice}(\tilde{Y}_i^u, \tilde{Y}_i) \right) \cdot A_i \quad (2)$$

$$A_i = \mathbb{1}\left[\tilde{Y}_i \geq \gamma \text{ or } \tilde{Y}_i \leq 1 - \gamma \right] \quad (3)$$

where $\tilde{Y}_i^u$ and $\tilde{Y}_i$ represent the predictions of the student network and the teacher network for the unlabeled data, respectively, and γ represents the set confidence threshold for filtering out unreliable pixels. The teacher weights θ_{r2} are updated by the exponential moving average of the student weights θ_{r1}, with a momentum parameter α set to 0.99:

$$\theta_{r2} \xleftarrow{\text{EMA}} \alpha\theta_{r2} + (1-\alpha)\theta_{r1} \quad (4)$$

3.2. Depth-guided cross-modal mutual learning

To encourage Model_R1 to learn more complementary information from Model_D and prevent Model_R1 from learning some wrong information from Model_D due to the large performance gap between the two student networks in the later stage of training, we propose a novel mutual learning strategy that uses a large amount of unlabeled data to guide different student networks to learn cross-modal complementary knowledge effectively. Specifically, we perform

mutual learning in three aspects based on the consistency and difference of Model_R1 and Model_D predictions.

Consistency learning. For unlabeled data $\{X_i, Z_i\} \in \mathcal{D}_u$, we use consistency constraints to encourage the student network to have consistent prediction. In addition, we also encourage the deep encoding features of the student network outputs of different modalities to be aligned. We use the mean square error (MSE) loss as a constraint for simplicity. The unlabeled data consistency loss can be expressed as:

$$\mathcal{L}_{\text{con}} = \frac{1}{2|\mathcal{D}_u|} \sum_{(X_i,Z_i)\in\mathcal{D}_u} \left(\left\| F_i^{r1} - F_i^d \right\|_2^2 + \left\| \mathcal{F}_{\theta_{r1}}(X_i^{S1}) - \mathcal{F}_{\theta_d}(Z_i) \right\|_2^2 \right) \tag{5}$$

where F_i^{r1} and F_i^d represent the output features of the last layer encoders of the two student networks, respectively, and $X_i^{S1} = \text{Aug}_{S1}(X_i)$.

Complementary knowledge learning. By only constraining the consistency of outputs and features of different student networks, this method easily leads to the premature convergence of the two student networks to consistency, making it difficult for the student networks to learn richer cross-modal complementary information. At present, contrastive learning [2, 5, 31, 34] has shown great potential in semi-supervised methods. Considering the large differences between RGB and depth modalities in the feature space, it may be difficult to capture the cross-modal association between them using feature-level contrastive learning directly. Therefore, we use a contrastive learning strategy based on output-level positive and negative sample pairs to guide the model to learn richer cross-modal complementary information. Considering that medical polyp image segmentation is a two-class task, we regard the different category regions of each sample as the same object and establish positive and negative sample pairs between different modalities and different samples. For each batch of training sample set $\{(X_i, Z_i) \in \mathcal{D}_l\}_{i=1}^{K} \cup \{(X_i, Z_i) \in \mathcal{D}_u\}_{i=K+1}^{2K}$, the network output of the paired modality is used as a pair of positive sample pairs. For the positive sample pair set T_l^+ with labeled data $\{(X_i, Z_i) \in \mathcal{D}_l\}_{i=1}^{K}$, it can be formulated as:

$$T_l^+ = \{(\mathcal{F}_{\theta_{r1}}(X_i), \mathcal{F}_{\theta_d}(Z_i))\}_{i=1}^{K} \tag{6}$$

For the unlabeled data $\{(X_i, Z_i) \in \mathcal{D}_u\}_{i=K+1}^{2K}$, the corresponding positive sample pair set T_u^+ can be expressed as:

$$T_u^+ = \{(\mathcal{F}_{\theta_{r1}}(\text{Aug}_{S1}(X_i)), \mathcal{F}_{\theta_d}(Z_i))\}_{i=K+1}^{2K} \tag{7}$$

We can get the positive sample pair set Γ^+ under each training batch as:

$$\Gamma^+ = T_l^+ \cup T_u^+ \tag{8}$$

For simplicity, we can express Γ^+ as:

$$\Gamma^+ = \{\overline{Y}_i^{r1}, \overline{Y}_i^d\}_{i=1}^{2K} \tag{9}$$

where $\overline{Y}_i^{r1}$ represents the prediction of the student network Model_R1, and $\overline{Y}_i^d$ represents the prediction of the student network Model_D. For the i-th positive sample pair, we take the remaining $4K-2$ predictions as the i-th corresponding negative sample pair T_i^-. The Dice coefficient can measure the overlapping area of two samples and is robust to the imbalance of foreground and background. Inspired by this, we use the Dice coefficient to evaluate the regional similarity of positive and negative samples. The calculation formula for the Dice coefficient can be defined as:

$$S_{\text{Dice}}(\overline{Y}_i^{r1}, \overline{Y}_i^d) = \frac{2\sum_{k=1}^{H\times W} \overline{Y}_{i,k}^{r1} \cdot \overline{Y}_{i,k}^d}{\sum_{k=1}^{H\times W} \overline{Y}_{i,k}^{r1} + \sum_{k=1}^{H\times W} \overline{Y}_{i,k}^d} \tag{10}$$

where $\overline{Y}_{i,k}^{r1}$ and $\overline{Y}_{i,k}^d$ represent the predicted values of the k-th pixel $\overline{Y}_i^{r1}$ and $\overline{Y}_i^d$ predicted by the network respectively. H and W represent the predicted height and width respectively. Since the prediction results with RGB images as the main modality have a higher confidence level, when calculating the similarity between the positive sample pairs and the corresponding negative sample pairs, only the prediction $\overline{Y}_i^{r1}$ obtained by the Model_R1 is used. Then the contrastive learning loss function $\mathcal{L}_{\text{cont}}$ can be expressed as:

$$\mathcal{L}_{\text{cont}} = -\frac{1}{|\Gamma^+|} \sum_{(\overline{Y}_i^{r1}, \overline{Y}_i^d)\in\Gamma^+} \log \frac{e^{S_{\text{Dice}}(\overline{Y}_i^{r1}, \overline{Y}_i^d)}}{e^{S_{\text{Dice}}(\overline{Y}_i^{r1}, \overline{Y}_i^d)} + \sum_{\overline{Z}_i^j \in T_i^-} e^{S_{\text{Dice}}(\overline{Y}_i^{r1}, \overline{Z}_i^j)}} \tag{11}$$

where $\overline{Z}_i^j$ represents the j-th sample in the negative sample pair corresponding to the i-th positive sample pair.

Auxiliary student network guides learning. We believe that the high confidence prediction of Model_D on unlabeled data can also be used as a pseudo-label to guide the learning of Model_R1. For the k-th pixel prediction y_k, if $y_k > 0.5$, it is a polyp category; if $y_k \leq 0.5$, it is a background category. Specifically, we consider the k-th pixel prediction y_k of Model_R2 and the corresponding prediction y'_k of Model_D. If the distance between y'_k and the corresponding category is lower than the threshold distance, and this distance is lower than the distance between y_k and the corresponding category, we replace the k-th pixel prediction value with y'_k, which can be formulated as follows:

$$Y''_{i,k} = \begin{cases} y'_k, & \text{if } y'_k > \gamma \quad \text{and} \quad d'_k < d_k \\ y'_k, & \text{if } y'_k < 1-\gamma \quad \text{and} \quad d'_k < d_k \\ y_k, & \text{otherwise} \end{cases} \tag{12}$$

where $Y''_{i,k}$ represents the pseudo label of the k-th pixel of the i-th image after update. If y is a polyp category, then

the distance $d = 1 - y$; otherwise, $d = y$. Then the corresponding unsupervised loss $\mathcal{L}_{\text{ud}}$ can be formulated as:

$$\mathcal{L}_{\text{ud}} = \mathcal{L}_{\text{u}}(\mathcal{F}_{\theta_{r1}}(X_i^{S1}), Y_i'')) \tag{13}$$

3.3. Depth-guided patch augmentation

The CutMix [38] method has proven its effectiveness [10] in semi-supervised segmentation tasks, and many CutMix-based improvement methods have emerged [16, 17, 42]. PH-Net [16] considers the difference in learning difficulty between difficult and simple regions in an image and proposes an adaptive patch augmentation (APA) method to identify and process difficult regions to improve performance. This method evaluates the difficulty of a patch by calculating the entropy of the model's prediction distribution. However, there are some limitations in evaluating the difficulty of image patches based on model predictions. The model prediction confidence is low when the amount of labeled data is small. Suppose a high-confidence erroneous prediction is made for difficult regions. In that case, these regions may be misjudged as easy regions, thus affecting the overall performance. Instead, we use depth information for difficult patch evaluation. For the RGB image X and its corresponding single-channel depth map D, the hardness score of a patch P of size $h \times w$ can be expressed as:

$$H_P = \frac{1}{h \times w} \sum_{(i,j) \in P} \text{Var}(D_{i,j}) \tag{14}$$

where H_P is the hardness score of patch P, $D_{i,j}$ is the depth value of pixel position (i, j) in the depth map D, and $\text{Var}(D_{i,j})$ represents the variance of pixel depth values in the patch. The larger the variance of the patch, the more obvious the depth change of the patch is, which means that the difficulty of the patch is higher. To enhance the richness of the training samples, we randomly select h and w from $\{10, 20, 40\}$. The patches are sorted in descending order of hardness scores so that patches with higher hardness scores can be shielded to prevent them from being cut off during the cutting process. The number of shielded patches is dynamically set following PH-Net [16]. Then, the depth-guided patch augmentation can be expressed as:

$$\text{Aug}_{S2}(X_i^{S1}) \leftarrow M \odot X_i^{S1} + (1 - M) \odot X_{i+1}^{S1} \tag{15}$$

$$Y_i^{r2} \leftarrow M \odot \tilde{Y}_i^{r2} + (1 - M) \odot \tilde{Y}_{i+1}^{r2} \tag{16}$$

where $X_i \in \mathcal{D}_u$ and $X_{i+1} \in \mathcal{D}_u$ represent the i-th and $(i + 1)$-th unlabeled RGB images in the batch. $\tilde{Y}_i^{r2}$ and $\tilde{Y}_{i+1}^{r2}$ denote the corresponding pseudo labels predicted by the teacher network for X_i and X_{i+1}, and M is a patch mask randomly selected from the unmasked patches. Then, the corresponding loss $\mathcal{L}_{\text{u-dpa}}$ can be formulated as:

$$\mathcal{L}_{\text{u-dpa}} = \mathcal{L}_{\text{u}}(\mathcal{F}_{\theta_{r1}}(\text{Aug}_{S2}(X_i^{S1})), Y_i^{r2})) \tag{17}$$

3.4. Loss function

The training loss of student networks consists of supervised loss and unsupervised loss. The training loss of the student network Model_R1 can be formulated as:

$$\mathcal{L}_{\text{rgb}} = \frac{1}{3} \left(\mathcal{L}_{\text{sup}}(\mathcal{F}_{\theta_{r1}}(X_i), Y_i) + \mathcal{L}_{\text{u-dpa}} + \mathcal{L}_{\text{ud}} \right) \tag{18}$$

where $X_i \in \mathcal{D}_l$, Y_i is the ground truth corresponding to the image X_i. The student network Model_D predicts $\tilde{Y}_i^{d,u} = \mathcal{F}_{\theta_d}(Z_i)$ for the unlabeled data $Z_i \in \mathcal{D}_u$, while the pseudo label predicted by Model_R2 using the corresponding RGB image X_i is $\tilde{Y}_i^{r2}$. The training loss of the student network Model_D can be formulated as:

$$\mathcal{L}_{\text{depth}} = \frac{1}{2} \left(\mathcal{L}_{\text{sup}}(\mathcal{F}_{\theta_d}(Z_i), Y_i) + \mathcal{L}_{\text{u}}(\tilde{Y}_i^{d,u}, \tilde{Y}_i^{r2}) \right) \tag{19}$$

where $Z_i \in \mathcal{D}_l$, Y_i is the corresponding ground truth. The overall training loss can be formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rgb}} + \mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{cont}} + \mathcal{L}_{\text{con}} \tag{20}$$

4. Experiments

4.1. Dataset and evaluation metrics

Datasets. We evaluated our method on five typical polyp segmentation datasets: Kvasir-SEG [15], CVC-ClinicDB [4], CVC-300 [30], CVC-ColonDB [27] and ETIS-Larib [26].

The Kvasir-SEG dataset contains 1000 polyp images, of which 900 are randomly divided into training sets and the remaining 100 into test sets. The training sets are partitioned into labeled and unlabeled subsets according to the 1/40 and 1/10 partitioning protocols. The CVC-ClinicDB dataset contains 612 polyp images, of which 550 are randomly divided into training sets and the remaining 62 into test sets. The training sets are partitioned into labeled and unlabeled subsets according to the 1/40 and 1/10 partitioning protocols. In addition, we followed the experimental settings of the ACL-Net [34] method. We combined 900 training images from the Kvasir-SEG dataset and 550 training images from the CVC-ClinicDB dataset into a new training set of 1450 images as a semi-supervised training dataset. The training sets are partitioned into labeled and unlabeled subsets according to the 1/10 partitioning protocols. Under this experimental setting, only the CVC-300, CVC-ColonDB, and ETIS-Larib datasets are tested to further evaluate the generalization ability of the model.

Evaluation metrics. We adopt three widely used metrics for quantitative evaluation, including the mean dice similarity coefficient (Dice(%)), the mean pixel-wise accuracy (Acc(%)), and the mean intersection over union (IoU(%)).

Table 1. Quantitative comparison using different state-of-the-art methods on Kvasir-SEG and CVC-ClinicDB datasets. Boldface and underlined numbers represent the best and second-best results of each setting, respectively.

Method	Kvasir-SEG					CVC-ClinicDB				
	Labeled	Unlabeled	Dice(%)↑	IoU(%)↑	Acc(%)↑	Labeled	Unlabeled	Dice(%)↑	IoU(%)↑	Acc(%)↑
U-Net	22(2.5%)	0	76.49	70.48	94.61	14(2.5%)	0	67.60	58.45	95.90
U-Net	90(10%)	0	82.76	77.09	95.23	54(10%)	0	77.84	72.14	97.24
U-Net	900(100%)	0	**89.43**	**84.35**	**97.07**	550(100%)	0	**88.34**	**83.54**	**98.59**
MT [28]			64.27	67.77	93.86			53.19	58.79	95.67
CCT [21]			44.08	44.83	84.65			31.02	37.14	84.43
DTC [19]			73.93	70.99	94.38			63.25	61.76	95.76
SCP-Net [41]			73.53	62.87	92.99			63.93	53.03	94.16
BCP [3]	22(2.5%)	878(97.5%)	82.42	75.19	95.39	14(2.5%)	536(97.5%)	76.65	71.90	96.75
PH-Net [16]			83.12	75.69	95.65			73.37	67.42	96.17
CML [35]			83.21	76.59	95.43			77.60	71.67	97.02
AD-MT [44]			82.29	74.51	95.22			74.36	66.14	96.07
Ours			**86.35**	**80.10**	**96.05**			**82.60**	**75.81**	**97.32**
MT [28]			84.81	79.27	96.13			80.66	74.13	97.51
CCT [21]			70.02	68.39	93.09			65.43	61.89	94.63
DTC [19]			83.84	79.74	96.11			79.32	76.52	97.86
SCP-Net [41]			85.70	78.46	96.33			82.08	74.93	97.70
BCP [3]	90(10%)	810(90%)	87.30	80.92	96.25	54(10%)	496(90%)	85.59	78.70	97.99
PH-Net [16]			87.35	80.67	96.54			82.11	75.06	97.93
CML [35]			86.89	80.46	95.72			85.49	77.92	97.97
AD-MT [44]			86.38	79.91	96.23			83.66	76.22	97.42
Ours			**89.07**	**84.21**	**97.07**			**87.84**	**81.84**	**98.15**

4.2. Implementation details

All experiments in this study were implemented on a single NVIDIA GeForce RTX 3060 using the PyTorch framework and used a fixed random seed. Our baseline uses the UNet [22] network and uses the ResNet34 [13] pretrained on ImageNet [7] as the UNet backbone. The image and label sizes are resized to 320 × 320. The SGD optimizer is used with an initial learning rate of 0.001, momentum of 0.9, and weight decay of 0.00001. The learning rate is decayed using a polynomial strategy: $\eta = \eta_{\text{init}} \cdot \left(1 - \frac{\text{iter}}{\text{iter}_{\text{total}}}\right)^{0.9}$. In our experimental setup, we adopt a batch size of 4 and train for 50,000 iterations, with the batch size for labeled and unlabeled data set to 2.

4.3. Comparisons with SOTA methods

We compare our method with eight state-of-the-art semi-supervised segmentation methods, including MT [28], CCT [21], DTC [19], SCP-Net [41], BCP [3] , PH-Net [16], CML [35], and AD-MT [44]. To ensure a fair comparison, for all experiments, we use UNet with ResNet-34 as the backbone of the segmentation network, adopt the same experimental environment and data augmentation, and the data partitioning protocol of these methods is the same.

As shown in Table 1, our proposed method achieves the highest scores on the three evaluation metrics on two typical polyp segmentation datasets, Kvasir-SEG and CVC-ClinicDB, with 2.5% and 10% labeled data settings. On the Kvasir-SEG dataset, using only 22 labeled images, our method achieves a Dice score of 86.35% and an IoU score of 80.10%, which is 3.14% higher in Dice score and 3.51% higher in IoU score than the previous state-of-the-art method CML [35]. To more intuitively compare the performance of semi-supervised methods, we also conducted experiments in a fully supervised manner. Under the 10% experimental setting, our method approaches the performance of full supervision. Our method achieves the best segmentation under all partitions of the Kvasir-SEG dataset and CVC-ClinicDB dataset, indicating that our method enables the model to be better trained on unlabeled images.

We also experimentally demonstrate that our method has excellent generalization ability on the unseen CVC-300, CVC-ColonDB, and ETIS-Larib polyp datasets. As shown in Table 2, our method has surpassed competitors significantly. Specifically, compared with the SupOnly method, our method achieves a 10.44% Dice and 8.80% IoU score improvement on the CVC-ColonDB dataset. We also obtained a 28.17% Dice and 23.67% IoU score improvement on the most challenging ETIS-Larib dataset. These experimental results show that by combining the mutual learning strategy of depth images and RGB images, our method can significantly enhance the model's feature representation ability. This approach helps the model better understand the shape and structure of polyps, which is particularly important for delicate tasks such as segmenting polyps. Therefore, our model performs well on different datasets, showing its strong adaptability and effectiveness.

For qualitative analysis, to more intuitively demonstrate our method's superiority, we visually compare challenging cases in Figure 3. Our method produces excellent segmen-

Table 2. Generalization ability comparison with different state-of-the-art methods. We split all the training images (1450) in Kvasir-SEG and CVC-ClinicDB into 1/10 labeled images (144) and 9/10 unlabeled images (1306) for training. "SupOnly" stands for supervised training without using any unlabeled data. Boldface and underlined numbers represent the best and second best results of each setting, respectively.

Method	Kvasir-SEG		CVC-ClinicDB		CVC-300		ETIS-Larib		CVC-ColonDB	
	Dice(%)↑	IoU(%)↑	Dice(%)↑	IoU(%)↑	Dice(%)↑	IoU(%)↑	Dice(%)↑	IoU(%)↑	Dice(%)↑	IoU(%)↑
SupOnly	82.22	77.33	74.71	69.39	79.10	70.47	43.30	39.80	63.85	57.55
MT [28]	84.60	78.52	76.43	69.57	76.84	70.12	48.45	42.17	69.23	62.41
CCT [21]	84.93	79.63	76.68	70.20	81.72	75.66	52.04	47.63	66.14	59.43
DTC [19]	84.61	79.75	77.02	73.60	81.16	75.22	52.89	49.71	65.67	61.49
SCP-Net [41]	85.45	77.75	79.79	72.95	79.79	71.50	48.57	40.35	66.10	57.57
BCP [3]	86.25	<u>80.10</u>	81.75	75.09	80.36	76.12	59.70	51.72	66.19	58.61
PH-Net [16]	<u>87.02</u>	79.95	79.01	72.64	84.35	77.24	59.73	51.85	69.70	61.62
CML [35]	85.39	79.01	<u>84.01</u>	<u>77.77</u>	83.10	75.98	56.00	49.45	63.88	57.51
AD-MT [44]	85.55	78.28	82.13	76.07	<u>84.43</u>	<u>77.52</u>	<u>66.29</u>	<u>58.61</u>	<u>71.46</u>	<u>64.13</u>
Ours	**88.88**	**82.98**	**85.88**	**79.97**	**85.61**	**79.32**	**71.47**	**63.47**	**74.29**	**66.35**

tation results for very small polyps with unclear boundaries and low contrast areas between polyps and their surrounding mucosa. This indicates that our method can effectively deal with challenges such as polyp size, shape, and low contrast.

Table 3. Ablation study of different components on Kvasir-SEG dataset with 10% labeled data. DPA: Depth-guided patch augmentation. "SupOnly" stands for training using only labeled data.

Method	$\mathcal{L}_{ud}$	$\mathcal{L}_{con}$	$\mathcal{L}_{cont}$	DPA	Dice (%)	IoU (%)
SupOnly					82.76	77.09
I	✓	✓			87.96	82.61
II	✓		✓		88.08	82.51
III	✓			✓	88.03	82.68
IV	✓	✓	✓		88.14	82.89
V	✓	✓	✓	✓	**89.07**	**84.21**

Table 4. Ablation study of different enhancement methods on Kvasir-SEG dataset. SA: Aug_{S1}.

Method	# images used		Metrics	
	Labeled	Unlabeled	Dice (%)	IoU (%)
SA	22(2.5%)	878(97.5%)	81.67	74.79
SA+CutMix [38]			84.53	78.44
SA+APA [16]			84.42	77.87
SA+DPA			**86.35**	**80.10**
SA	90(10%)	810(90%)	87.16	81.01
SA+CutMix [38]			88.09	82.27
SA+APA [16]			88.50	82.76
SA+DPA			**89.07**	**84.21**

Table 5. Ablation study of confidence threshold γ on Kvasir-SEG dataset with 10% labeled data.

γ	0.7	0.75	0.8	0.85	0.9	0.95
Dice (%)	88.66	88.74	89.06	**89.07**	89.03	88.72
IoU (%)	83.60	83.55	83.91	**84.21**	83.74	83.67

4.4. Ablation studies

Effectiveness of components. In Table 3, we study the effectiveness of our proposed RD-Net main components on the Kvasir-SEG dataset with 10% labeled data. We use the model trained with supervised learning (SupOnly) using only labeled data as our baseline. The model with depth-guided cross-modal mutual learning ($\mathcal{L}_{ud}$, $\mathcal{L}_{con}$ and $\mathcal{L}_{cont}$) can obtain a performance gain of the Dice of the baseline by +5.38%, verifying that depth-guided cross-modal mutual learning can provide more cross-membrane complementary information for model training. Adding Depth-guided patch augmentation (DPA) improves the Dice of the baseline by +5.27%, verifying that the DPA module can guide the model to learn difficult patch regions. The full model with all components obtains the most accurate segmentation with the highest Dice and IoU scores. Overall, the results demonstrate the importance of each component in our proposed method for semi-supervised polyp segmentation. To more clearly demonstrate the advantages of each component, Figure 4 shows the predicted visualization of our method. Obviously, with the gradual introduction of enhanced components, the predicted lesion boundaries gradually become more refined. This result shows the effectiveness of our method in processing complex polyp images.

Effectiveness of depth-guided cross-modal mutual learning. We evaluated the prediction accuracy of the unlabeled data in the training set, and the results are shown in Figure 5. Notably, by adopting the depth-guided cross-modal mutual learning strategy, the model can improve the segmentation accuracy of the unlabeled training data in each partition of the Kvasir-SEG and CVC-ClinicDB datasets. This result shows that our depth-guided cross-modal mutual learning strategy effectively helps the model better utilize the information of unlabeled images.

Effectiveness of Depth-guided patch augmentation. As shown in Table 4, we compare the performance of APA augmentation, ordinary CutMix augmentation, and our DPA

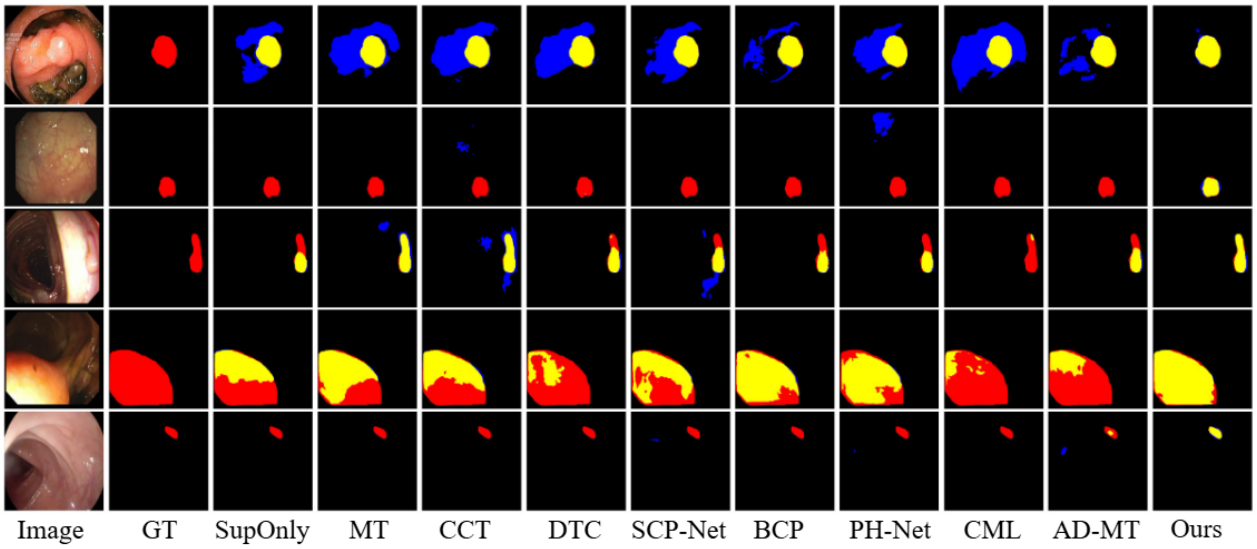

Figure 3. Visual comparison with different state-of-the-art methods on five public polyp datasets. The SupOnly method is trained on only 1/10 labeled data. The red, blue, and yellow areas represent the ground truth, predictions, and their overlap regions, respectively.

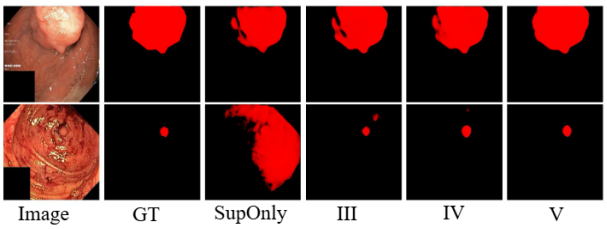

Figure 4. Visual comparison of component ablation studies. Darker red indicates higher confidence in the prediction.

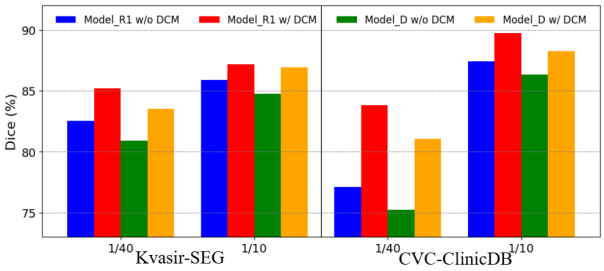

Figure 5. Comparing predictions on unlabeled training data under all partitioning protocols. DCM: Depth-guided cross-modal mutual learning.

augmentation. When the labeling ratio is 2.5%, the Dice value of DPA augmentation is 1.93% higher than that of APA augmentation, which indicates that in the case of very little labeled data, the prediction confidence of the model is low, resulting in poor performance of APA based on prediction for masking patches. In contrast, our method can more effectively select difficult patches for masking, thereby effectively identifying and alleviating the problem of insufficient training of challenging patches.

Ablation study on hyper-parameters. Table 5 shows the ablation study of the confidence threshold γ used to filter pseudo labels in Eq.(3) on the Kvasir-SEG dataset. As the threshold increases from 0.7 to 0.95, the Dice and IoU scores of our method increase first and then decrease. This result indicates that the model is prone to learning wrong prediction information when the threshold is low. Additionally, too much correct prediction information may be filtered out when the threshold is too high. We set $\gamma = 0.85$.

5. Conclusion

This paper proposes a novel semi-supervised polyp segmentation framework. The proposed depth-guided cross-modal mutual learning strategy promotes the learning of complementary information between different student networks in the training phase, while no depth information is required in the inference phase. At the same time, the pseudo-labels generated by the auxiliary student network are additionally used to guide the learning of the main student network from different angles, making the most of more high-confidence pixels. In addition, we introduce a depth-guided patch augmentation method to promote the model's learning in difficult areas of polyp images. Experimental results show that our method significantly outperforms the state-of-the-art methods on five polyp datasets.

Acknowledgement. This work was supported by the National Natural Science Foundation of China (Grant No. 62171419), the Hainan Province Science and Technology Special Fund of China (Grant No. ZDYF2022SHFZ318) and the Natural Science Foundation of Shandong Province of China (Grant No. ZR2021LZH005).

References

[1] Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on intelligent transportation systems*, 2023. 2

[2] Inigo Alonso, Alberto Sabater, David Ferstl, Luis Montesano, and Ana C Murillo. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8219–8228, 2021. 4

[3] Yunhao Bai, Duowen Chen, Qingli Li, Wei Shen, and Yan Wang. Bidirectional copy-paste for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11514–11524, 2023. 1, 2, 6, 7

[4] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015. 5

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 4

[6] Yanhua Cheng, Xin Zhao, Rui Cai, Zhiwei Li, Kaiqi Huang, Yong Rui, et al. Semi-supervised multimodal deep learning for rgb-d object recognition. In *IJCAI*, pages 3345–3351, 2016. 1

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[8] Siqi Du, Weixi Wang, Renzhong Guo, Ruisheng Wang, and Shengjun Tang. Asymformer: Asymmetrical cross-modal representation learning for mobile platform real-time rgb-d semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7608–7615, 2024. 2

[9] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 263–273. Springer, 2020. 1

[10] Geoff French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. *arXiv preprint arXiv:1906.01916*, 2019. 5

[11] Xiangyu Guo, Wei Ma, Fangfang Liang, and Qing Mi. Dual-modal non-local context guided multi-stage fusion for indoor rgb-d semantic segmentation. *Expert Systems with Applications*, 255:124598, 2024. 1, 2

[12] Along He, Tao Li, Yanlin Wu, Ke Zou, and Huazhu Fu. Frcnet: Frequency and region consistency for semi-supervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 305–315. Springer, 2024. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[14] Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang. Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In *2019 IEEE international conference on image processing (ICIP)*, pages 1440–1444. IEEE, 2019. 1, 2

[15] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen. 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*, pages 451–462. Springer, 2020. 5

[16] Siyao Jiang, Huisi Wu, Junyang Chen, Qin Zhang, and Jing Qin. Ph-net: Semi-supervised breast lesion segmentation via patch-wise hardness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11418–11427, 2024. 1, 2, 5, 6, 7

[17] Jang-Hyun Kim, Wonho Choo, and Hyun Oh Song. Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In *International conference on machine learning*, pages 5275–5285. PMLR, 2020. 5

[18] Su Young Kim, Hyun-Soo Kim, Yun Tae Kim, Jung Kuk Lee, Hong Jun Park, Hee Man Kim, and Dae Ryoung Kang. Colonoscopy versus fecal immunochemical test for reducing colorectal cancer risk: A population-based case–control study. *Clinical and translational gastroenterology*, 12(5): e00350, 2021. 1

[19] Xiangde Luo, Jieneng Chen, Tao Song, and Guotai Wang. Semi-supervised medical image segmentation through dual-task consistency. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8801–8809, 2021. 6, 7

[20] Inés Mármol, Cristina Sánchez-de Diego, Alberto Pradilla Dieste, Elena Cerrada, and María Jesús Rodriguez Yoldi. Colorectal carcinoma: a general overview and future perspectives in colorectal cancer. *International journal of molecular sciences*, 18(1):197, 2017. 1

[21] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12674–12684, 2020. 6, 7

[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 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*, pages 234–241. Springer, 2015. 6

[23] Hao Shao, Yang Zhang, and Qibin Hou. Polyper: Boundary sensitive polyp segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4731–4739, 2024. 1

[24] Jing-Hui Shi, Qing Zhang, Yu-Hao Tang, and Zhong-Qun Zhang. Polyp-mixer: An efficient context-aware mlp-based paradigm for polyp segmentation. *IEEE Transactions on Cir-*

cuits and Systems for Video Technology, 33(1):30–42, 2022. 1

[25] Noam Shussman and Steven D Wexner. Colorectal polyps and polyposis syndromes. *Gastroenterology report*, 2(1):1–15, 2014. 1

[26] Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International journal of computer assisted radiology and surgery*, 9:283–293, 2014. 5

[27] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*, 35(2):630–644, 2015. 5

[28] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 3, 6, 7

[29] Quoc-Huy Trinh. Meta-polyp: a baseline for efficient polyp segmentation. In *2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 742–747. IEEE, 2023. 1

[30] David Vázquez, Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Antonio M López, Adriana Romero, Michal Drozdzal, and Aaron Courville. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of healthcare engineering*, 2017(1):4037190, 2017. 5

[31] Xiaoyang Wang, Bingfeng Zhang, Limin Yu, and Jimin Xiao. Hunting sparsity: Density-guided contrastive learning for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3114–3123, 2023. 4

[32] Yongchao Wang, Bin Xiao, Xiuli Bi, Weisheng Li, and Xinbo Gao. Mcf: Mutual correction framework for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15651–15660, 2023. 2

[33] Jun Wei, Yiwen Hu, Ruimao Zhang, Zhen Li, S Kevin Zhou, and Shuguang Cui. Shallow attention network for polyp segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 699–708. Springer, 2021. 1

[34] Huisi Wu, Wende Xie, Jingyin Lin, and Xinrong Guo. Aclnet: semi-supervised polyp segmentation via affinity contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2812–2820, 2023. 1, 2, 4, 5

[35] Song Wu, Xiaoyu Wei, Xinyue Chen, Yazhou Ren, Jing He, and Xiaorong Pu. Cross-view mutual learning for semi-supervised medical image segmentation. In *ACM Multimedia 2024*, 2024. 6, 7

[36] Yang Xia, Haijiao Yun, Peiyu Liu, and Mingjing Li. A novel parallel cooperative mean-teacher framework (pcmt) combined with prediction uncertainty guide and class contrastive learning for semi-supervised polyp segmentation. *Expert Systems with Applications*, 255:124816, 2024. 1

[37] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 2

[38] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 5, 7

[39] Chen Zhang, Runmin Cong, Qinwei Lin, Lin Ma, Feng Li, Yao Zhao, and Sam Kwong. Cross-modality discrepant interaction network for rgb-d salient object detection. In *Proceedings of the 29th ACM international conference on multimedia*, pages 2094–2102, 2021. 2

[40] Hongyan Zhang, Victor S Sheng, Xuefeng Xi, Zhiming Cui, and Huan Rong. Overview of rgbd semantic segmentation based on deep learning. *Journal of Ambient Intelligence and Humanized Computing*, 14(10):13627–13645, 2023. 1

[41] Zhenxi Zhang, Ran Ran, Chunna Tian, Heng Zhou, Xin Li, Fan Yang, and Zhicheng Jiao. Self-aware and cross-sample prototypical learning for semi-supervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 192–201. Springer, 2023. 1, 2, 6, 7

[42] Zhen Zhao, Lihe Yang, Sifan Long, Jimin Pi, Luping Zhou, and Jingdong Wang. Augmentation matters: A simple-yet-effective approach to semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11350–11359, 2023. 5

[43] Zhida Zhao, Jia Li, Lijun Wang, Yifan Wang, and Huchuan Lu. Maskmentor: Unlocking the potential of masked self-teaching for missing modality rgb-d semantic segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1915–1923, 2024. 2

[44] Zhen Zhao, Zicheng Wang, Longyue Wang, Dian Yu, Yixuan Yuan, and Luping Zhou. Alternate diverse teaching for semi-supervised medical image segmentation. In *European Conference on Computer Vision*, pages 227–243. Springer, 2025. 2, 6, 7