Hindawi
International Journal of Intelligent Systems
Volume 2024, Article ID 9928155, 11 pages
https://doi.org/10.1155/2024/9928155

Research Article

Semisupervised Medical Image Segmentation through Prototype-Based Mutual Consistency Learning

Xinqiang Wang,[1,2] **Wenhuan Lu**,[1] **Si Li**,[1] **Ke Zheng**,[1] **Junhai Xu**,[1] **and Jianguo Wei**[1]

[1]*College of Intelligence and Computing, Tianjin Key Lab of Cognitive Computing and Application, Tianjin University, Tianjin 300350, China*
[2]*School of Software and Communication, Tianjin Sino-German University of Applied Sciences, Tianjin 300350, China*

Correspondence should be addressed to Junhai Xu; jhxu@tju.edu.cn

Received 27 June 2023; Revised 24 January 2024; Accepted 7 February 2024; Published 15 March 2024

Academic Editor: Yu-An Tan

Medical image segmentation is a critical task in the healthcare field. While deep learning techniques have shown promise in this area, they often require a large number of accurately labeled images. To address this issue, semisupervised learning has emerged as a potential solution by reducing the reliance on precise annotations. Among these approaches, the student-teacher framework has garnered attention, but it is limited in its reliance solely on the teacher model for information. To overcome this limitation, we propose a prototype-based mutual consistency learning (PMCL) framework. This framework utilizes two branches that learn from each other, incorporating supervision loss and consistency loss to adapt to minor data perturbations and structural differences. By employing prototype consistency learning, we are able to achieve reliable consistency loss. Our experiments on three public medical image datasets demonstrate that PMCL outperforms other state-of-the-art methods, indicating its potential in semisupervised medical image segmentation. Our framework has the potential to assist medical professionals in enhancing their diagnoses and delivering improved patient care.

1. Introduction

Automatic and accurate segmentation of tumors, organs, or lesions is the premise of designing computer-aided diagnosis and detection systems. Deep convolutional neural networks have performed well at many medical image segmentation tasks [1–3]. However, these methods require a large number of high-quality labeled images to achieve very good results. It is laborious and time-consuming for experienced experts to make reliable and accurate annotations. We study semisupervised methods to fully utilize a small number of labeled images and a large number of unlabeled images to solve this problem.

Semisupervised methods have developed rapidly, especially in the field of medical image segmentation. Temporal ensembling and the II model [4] are proposed to accomplish semisupervised learning tasks by adding noise to the unlabeled data and then minimizing the difference between the prediction results of the source data and the noised data. The mean teacher framework [5] utilized the *exponential moving average* (EMA) of the temporal ensembling method. The network consists of a teacher model and a student model. The student model is trained by gradient descent, and the teacher model is obtained by using the parameters of the student model. The mean teacher framework has a simple structure and excellent experimental results, so many subsequent methods [6–8] make full use of this framework and extend this framework. Xie et al. [6] added a confidence module to the mean teacher framework to predict the confidence of the model and improve the performance of the network. Li et al. [7] introduced more perturbations to both the data and model of the mean teacher framework to construct the consistency loss. Yu et al. [8] encouraged the model to learn more reliable goals by adding uncertainty awareness to the mean teacher framework. Adversarial learning is also used for semisupervised segmentation [9, 10]. Zhang et al. [10] proposed a deep adversarial network to encourage consistency between the predicted segmentation of unlabeled data. More recently, there have been some multitask network structures for semisupervised medical image segmentation tasks [11, 12]. Li et al. [11] performed image segmentation and signed distance map regression tasks at the same time and used the discriminator as a regularization item. Luo et al. [12] built a multitask network that builds the consistency from the difference of segmentation tasks and the level set function regression task.

However, in the mean teacher framework, the parameters of the student network are obtained by the combination of the segmentation and consistency loss, the exponential moving average is calculated to obtain the parameters of the teacher network, and the total loss is updated to guide the student network in turn. We want to build a framework that consists of two student models, which we encourage to learn from each other, combining their learned information to improve network performance. We propose a prototype-based mutual consistency learning framework (PMCL) for medical image segmentation tasks, which is divided into two branches, which we can regard as two student models. To make them learn different information, the two student models are slightly different. The two branches use prototype learning to obtain the segmentation predictions of unlabeled images under different disturbances, and we obtain the consistency loss by comparing the segmentation predictions of the two branches. The prediction difference between the two branches can be considered as a complex area. By applying the consistency loss to the output of each decoder, high-confidence regions can be learned. For labeled images, the two branches obtain different pieces of information through slightly different decoders. The framework learns more reliable information through a combination of two supervision losses. The two branches learn from each other, allowing the network to train end-to-end.The main contributions of this work are as follows:

(1) We propose a semisupervised 2D medical image segmentation framework, PMCL, which allows two networks to learn from each other for semisupervised segmentation tasks. The proposed framework can also be applied to other 2D and 3D semisupervised medical image segmentation tasks.

(2) We use prototype consistency learning to generate high-quality pseudolabels specifically for unlabeled images, which are more reliable than those generated by other methods. The performance of the network can be significantly improved by using labels obtained specifically from unlabeled images.

(3) Comprehensive experiments on three public medical image datasets demonstrate the superiority of PMCL to other semisupervised methods. Ablation experiments confirm the effectiveness of each submodule of the proposed method.

2. Related Work

We introduce related work on semisupervised medical image segmentation, mutual learning, and prototype consistency learning.

2.1. Semisupervised Medical Image Segmentation. Semisupervised learning plays an increasing role in the field of medical image segmentation. It can be roughly divided into regularization methods based on data or model disturbances, adversarial learning methods, and consistency methods based on multitask levels.

There are many pseudolabel methods [14, 15], which utilize labeled data to train the model, generate pseudolabels for unlabeled data, and add these to the training set to continue training. The most important task is finding high-quality soft labels. Hung et al. [16] designed a discriminator to provide supervisory signals to perform semisupervised medical image segmentation tasks. It can learn to distinguish between ground-truth label maps and probability maps for segmentation prediction. Combining spatial cross-entropy loss, this paper uses adversarial loss to encourage segmentation networks to generate prediction probability maps that are close to the real label map in high-order structures. Temporal ensembling and the II model [4] were proposed to complete semisupervised learning by minimizing the difference between the predicted results of the original unlabeled data and the noised unlabeled data. Virtual adversarial training (VAT) [17] proposed a regularization method based on virtual adversarial loss: a new measure of local smoothness of label distribution given input conditions. The virtual adversarial loss is defined as the robustness of the conditional label distribution around each input data point to local disturbances. It replaces random perturbations with adversarial perturbations designed to deceive the trained model, enabling the network to effectively learn the local smoothness a priori and become more resilient to various noises.

Mean teacher [5] also uses the consistency regularity and is divided into student and teacher models. The student model obtains the parameters through gradient descent, and the teacher model obtains them through the exponential moving average calculation of the student model parameters. The difference between the two model parameters can be regarded as a part of the network disturbance, which, together with the data disturbance, constitutes the total disturbance. Mean teacher has achieved great success in semisupervised image segmentation, and many subsequent networks [6–8] have modified and extended it. Li et al. [7] added more perturbations to the data and model based on the mean teacher framework. Yu et al. [8] used Monte Carlo dropout to add uncertainty awareness to the mean teacher framework to allow the learning of more reliable information.

Multitask network structures for semisupervised medical image segmentation have recently appeared. SASSnet [11] performs signed distance map regression and image segmentation tasks at the same time and uses the discriminator as a regularization item. The stability and robustness of the segmentation results are ensured by introducing prior information of shape and position. DTC [12] also builds consistency from the level of tasks for semisupervised learning and uses a multitask network. Unlike SASSnet, it uses the representation difference between the two tasks to build consistency.

2.2. Mutual Learning. High-performance deep neural networks generally have a huge number of parameters, so sophisticated networks such as MobileNet [18] and ShuffleNet [19] appeared later. Hinton et al. [20] proposed knowledge distillation technology, which uses a more complex teacher model that has been trained to guide a relatively lightweight student model for training. While reducing the model size and computing resource requirements, it tries to maintain the accuracy of the original teacher model. In the semisupervised medical image segmentation tasks, much work [5–8] has used the student-teacher network architecture to improve network performance.

In our work, the entire network framework has a mutual learning framework. In the student-teacher network, the student network can only learn from the teacher network. Unlike the student-teacher network, mutual learning consists of two student networks, which can learn from each other and make progress together. Mutual learning frameworks are widely used in multimodel architectures, and they have achieved good results at various tasks. Zhang et al. [21] first proposed a deep mutual learning strategy. Each network used the sum of its own supervision loss and the interaction loss from other networks to supervise network learning. Wu et al. [22] proposed two decoders in semisupervised medical image segmentation, whose outputs used pseudolabels to guide each other's probability map. This design made the output of the submodel consistent and low entropy, which can better segment edges and isolated parts of the image. Zhang and Zhang [23] designed two networks with the same structure, resulting in segmentation and regression layers. The networks were optimized to learn useful knowledge through mutual learning. Many methods [24–26] have exploited mutual learning methods.

2.3. Prototype Learning. In our method, we generate predicted labels for unlabeled images by using prototype learning in few-shot segmentation learning tasks, where the latter aims to learn transferable knowledge from different tasks with just a few samples. In prototype learning, the labeled data in the training set are used as the model's support set, and the prediction object is used as the network's query set. The network must learn to use the support set to predict the label of the query set.

Many methods, including metric- [27, 28], optimization- [29, 30], and graph-based [31, 32] methods, have been proposed for few-shot learning. Among these, prototype-based methods are widely used in few-shot segmentation, as they reduce computation and perform relatively well. Snell et al. [27] proposed a prototypical network to represent each class with one feature vector in image classification tasks, using the nearest neighbor classifier to predict the category of the query set. Shaban et al. [33] proposed a classical two-branch model for few-shot segmentation tasks, using a conditional branch to extract the prototype features of the support set and a segmentation branch to extract the features of the query set, obtaining a segmentation map through logistic regression. Dong and Xing [34] also used metric learning and prototypical networks to complete few-shot segmentation tasks. SG-One [35] used masked average pooling to generate prototypes for the support set and cosine similarity to establish the relationship between the query set and prototype. Masked average pooling has since been widely used. Wang et al. [36] proposed prototype alignment regularization to make full use of the information of the support set. CANet [37] introduced the attention mechanism in prototype learning, using the middle-level features of the network to compare the query and support sets, and continuously iterating the network to obtain the segmentation results. FWB [38] improved the quality of the prototype by performing the same operations on the support set image as on the query set. AMP [39] considered the support set of the historical state when calculating the prototype and combined prototypes under different feature resolutions. Some methods [40–42] have used superpixels to accomplish few-shot segmentation tasks.

We transfer the prototype learning in few-shot learning to semisupervised learning and use it to generate high-quality pseudolabels for unlabeled images to improve the reliability of network prediction.

3. Proposed Methodology

We present the details of the proposed PMCL method. We introduce the general semisupervised learning framework to make our method more intuitive and easier to understand, and then, we present the prototype consistency and mutual consistency learning modules. In this section, the overall loss composition of the framework is explained first. Then, the process of generating masks generated by prototype learning is explained, and finally, the consistency loss caused by masks is explained.

3.1. Semisupervised Segmentation Framework. Figure 1 shows the PMCL framework, which is trained as follows: The encoders of the two branches have the same structure and share weights. Two decoders from UNet [13] can capture uncertainty information through slight structural differences. A labeled image and an unlabeled image are fed into the two branches. For each branch, a shared backbone encoder is first used to embed the labeled and unlabeled images into deep features. Then, masked average pooling is utilized to obtain prototypes for the foreground and background from the labeled data and corresponding ground-truth, as discussed in Section 3.2. Label each pixel according to the class of the nearest prototype in order to segment the unlabeled images. A mutual learning network framework constrains the outputs of the two branches, as detailed in Section 3.3. Consistency loss and supervised loss constitute the total loss.

In the semisupervised learning setting, we have N labeled and M unlabeled training samples. We denote the respective labeled and unlabeled sets as $D_l = \{x_i, y_i\}_{i=1}^{N}$ and $D_u = \{x_i\}_{i=N+1}^{N+M}$, where $x_i \in \mathbb{R}^{H\times W}$ is the input image, $y_i \in \mathbb{R}^{H\times W}$ is the ground truth of x_i, and H and W are the

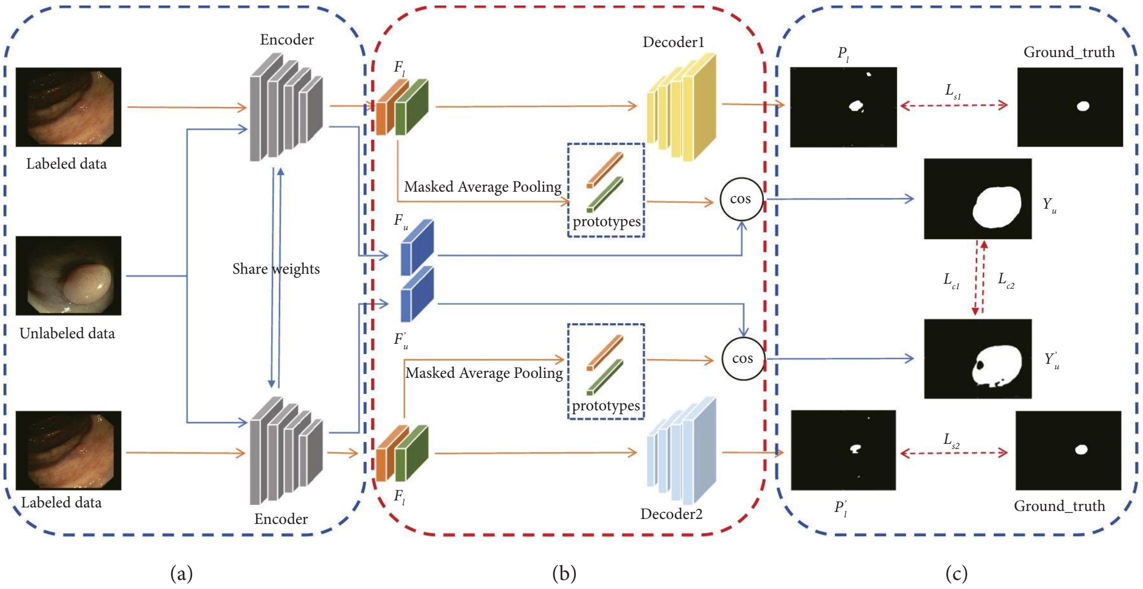

Figure 1: Overview of the proposed prototype-based mutual consistency learning network (PMCL) for semisupervised medical image segmentation. The entire framework consists of two branches, inspired by the Unet [13] method. The orange line represents the flow of labeled images, and the blue line represents the flow of unlabeled images. The entire framework can be divided into three parts. (a) Embed depth features into the original image. (b) Perform segmentation tasks on student models. (c) The composition of the framework loss we proposed.

image height and width, respectively. So, we can train our semisupervised medical image segmentation framework by minimization:

$$\min_{\theta,\xi,\xi'} \sum_{i=1}^{N} \mathscr{L}_s\left(\theta,\xi,\xi',D_l\right) + \lambda \sum_{i=1}^{N+M} \mathscr{L}_c\left(\theta,\xi,\xi',D_u\right), \quad (1)$$

where $\mathscr{L}_s$ and $\mathscr{L}_c$ are supervised loss and consistency loss, θ, ξ, and ξ' are the weights of the encoder, decoder1, and decoder2, and λ is a ramp-up weighting coefficient that controls the trade-off between the supervised and consistency loss and can prevent the network from learning meaningless consistency goals at the beginning of training.

The total loss of our prototype-based mutual consistency learning network is a weighted combination of supervised loss $\mathscr{L}_s$ and consistency loss $\mathscr{L}_c$, which are calculated only from labeled and unlabeled images, respectively. The total loss is

$$\mathscr{L} = \mathscr{L}_s + \lambda \mathscr{L}_c. \quad (2)$$

3.2. Prototype Mutual Learning. Previous semisupervised methods have usually directly used the encoding and decoding structure to generate segmentation predictions for unlabeled images, which does not efficiently utilize the information in the labeled images and corresponding labels. We want to efficiently generate pseudolabels for unlabeled images, which can be accomplished with the prototype learning method in few-shot learning. We use the labeled image and its ground truth as the support set and the unlabeled image as the query set to train the network. Our model is based on a prototypical network [27] that uses the mask annotations of the support set to learn prototypes for the foregrounds and backgrounds of images. To maintain input consistency, we adopt a late fusion strategy that uses a shared feature extractor to generate feature maps for the foregrounds and backgrounds of images [35, 43]. Specifically, we have a support set $S_i = (X_{l(k)}, Y_{l(k)})$, and $F_{l(k)}$ is a feature map extracted by the encoder for the labeled image $X_{l(k)}$, where $k = 1, \ldots, K$ indexes the support images. We can obtain the prototype of the foreground by masked average pooling [35]:

$$p_l(\mathrm{fg}) = \frac{1}{K} \sum_{K} \frac{\sum_{x,y} F_{l(k)}^{(x,y)} \mathbb{1}\left[Y_{l(k)}^{(x,y)} \in \mathscr{C}_{\mathrm{fg}}\right]}{\sum_{x,y} \mathbb{1}\left[Y_{l(k)}^{(x,y)} \in \mathscr{C}_{\mathrm{fg}}\right]}, \quad (3)$$

where (x, y) indexes the spatial locations, $\mathbb{1}(\cdot)$ is an indicator function that returns 1 if the condition is true and otherwise outputs 0, and $\mathscr{C}_{\mathrm{fg}}$ is the foreground segmentation target. We can also obtain the prototype of the background as

$$p_l(\mathrm{bg}) = \frac{1}{K} \sum_{K} \frac{\sum_{x,y} F_{l(k)}^{(x,y)} \mathbb{1}\left[Y_{l(k)}^{(x,y)} \notin \mathscr{C}_{\mathrm{fg}}\right]}{\sum_{x,y} \mathbb{1}\left[Y_{l(k)}^{(x,y)} \notin \mathscr{C}_{\mathrm{fg}}\right]}. \quad (4)$$

Nonparametric metric learning is used to learn the optimal prototype and complete segmentation. Since segmentation can be thought of as a classification of each spatial position, we calculate the distance between the query feature vector for each spatial position and each computed prototype. We introduce a distance function d and apply the softmax function over distances to produce a probability map M_j over classes. Let $\mathscr{P}_l = \{p_l(\mathrm{fg})\} \cup \{p_l(\mathrm{bg})\}$ and denote the feature map extracted from unlabeled data as F_u. For each $p_j \in \mathscr{P}_l$, we have

$$M_j^{(x,y)} = \frac{\exp\left(-\alpha \mathrm{d}\left(F_u^{(x,y)}, p_j\right)\right)}{\sum_{p_j \in P_l} \exp\left(-\alpha \mathrm{d}\left(F_u^{(x,y)}, p_j\right)\right)}, \quad (5)$$

where the distance function $d(\cdot)$ adopts the cosine distance (i.e., cos in Figure 1) to measure the similarity between the unlabeled feature map F_u and the labeled prototypes $\mathscr{P}_l$, and the multiplier α is set as 20, as used in PANet [36].

Then, we can obtain the predicted segmentation mask of unlabeled data as follows:

$$Y_u^{(x,y)} = \underset{j \in \{fg, bg\}}{\operatorname{argmax}} M_j^{(x,y)}. \quad (6)$$

Similarly, we can get the predictive segmentation mask of the unlabeled image of the other branch by performing the above operations (i.e., Y_u'). The two branches generate different pseudolabels for unlabeled data under different data perturbations. We add Gaussian noise on unlabeled data of the second branch (i.e., noise ε). The network can focus on high-confidence areas through the different pseudolabels generated by the two branches and obtain more reliable and robust results through consistency learning. We verify the role of prototype consistency learning in the network through an ablation experiment, as described in Section 4.4.

3.3. Mutual Consistency Learning. In a mutual learning framework, multiple untrained branches learn at the same time to solve tasks together. Each branch is guided by traditional supervised learning loss and consistency loss from other branches.

At the beginning of training, each branch can quickly segment images relatively correctly because of the traditional supervised loss. At this point, the predictions of the same pixels may differ according to initial conditions and network structures. The framework encourages consistent predictions from each branch. The consistency loss from other branches fine-tunes the model to perform better in complex segmentation areas. In the end, mutual learning helps to obtain a more robust and generalized network.

Our mutual learning framework consists of prototype mutual consistency learning for unlabeled data and mutual supervision learning for labeled data.

For prototype mutual consistency learning, to measure the segmentation predictions of the two branches, Kullback Leibler (KL) [21] divergence is used as the consistency loss. The consistency loss from Y_u to $Y_{u'}$ is computed as

$$\mathscr{L}_{c1} = \mathscr{D}_{KL}\left(Y_u' \| Y_u\right) = \sum_j Y_{u(j)}' \log \frac{Y_{u(j)}'}{Y_{u(j)}}. \quad (7)$$

We can similarly obtain the consistency loss from Y_u' to Y_u as

$$\mathscr{L}_{c2} = \mathscr{D}_{KL}\left(Y_u \| Y_u'\right) = \sum_j Y_{u(j)} \log \frac{Y_{u(j)}}{Y_{u(j)}'}. \quad (8)$$

In this way, each branch learns to correctly predict segmentations of training data and to match the probability estimate of its peer. We balance the two consistency losses to obtain the final consistency loss:

$$\mathscr{L}_c = 0.5 * \left(L_{c1} + L_{c2}\right). \quad (9)$$

For mutual supervision learning, decoder1 and decoder2 perform upsampling through bilinear interpolation and deconvolution, respectively. The different decoder structures prompt the model to learn more information. We combine cross-entropy and Dice loss to calculate the supervised loss. The two branches are calculated as follows:

$$\mathscr{L}_{s1} = 0.5 * \left(L_{ce}\left(P_l, Y_l\right) + L_{Dice}\left(P_l, Y_l\right)\right), \quad (10)$$

$$\mathscr{L}_{s2} = 0.5 * \left(L_{ce}\left(P_l', Y_l\right) + L_{Dice}\left(P_l', Y_l\right)\right). \quad (11)$$

To fully utilize the information of both branches and let the model train end-to-end, we combine the two supervised losses:

$$\mathscr{L}_s = 0.5 * \left(L_{s1} + L_{s2}\right). \quad (12)$$

Hence, the network obtains more reliable information from the labeled data through the mutual learning framework.

4. Experiments and Results

We discuss the implementation and compare the performance of PMCL and other semisupervised medical image segmentation algorithms on three public datasets. We performed ablation experiments to validate each part of our method.

4.1. Datasets and Evaluation Metrics. We evaluated our method on three public polyp segmentation datasets: CVC-ClinicDB [44], CVC-ColonDB [45], and Kvasir-SEG [46]. CVC-ClinicDB contains 612 images of size 384 × 288 pixels. CVC-ColonDB contains 380 images of size 574 × 500 pixels. Kvasir-SEG contains 1000 images, which we scaled to 256 × 256 pixels before training, as they vary in size from 332 × 487 to 1920 × 1072 pixels. In our experiments, we follow the training settings of [3, 47, 48]. The division of the three datasets was the same, with random selections of 80% of the images for training, 10% for validation, and 10% for testing. Each image was normalized to unit variance and zero mean. For training images, only 10% and 20% were used as labeled, and the remaining data were used as unlabeled data. Table 1 shows the image size, scale of training set, validation set, and testing set of these datasets.

We evaluated segmentation performance using the Dice similarity coefficient (DSC), Jaccard index (JI), sensitivity (SE), accuracy (AC), 95% Hausdorff distance (95HD), and average surface distance (ASD). We combine the experimental protocols in [6, 8] to calculate these metrics.

Table 1: The medical datasets used in our experiments.

Datasets	Images	Input size	Train	Valid	Test
CVC-ClinicDB [44]	612	384 × 288	490	61	61
CVC-ColonDB [45]	380	574 × 500	304	38	38
Kvasir-SEG [46]	1000	Variable	800	100	100

4.2. Implementation Details. All the networks in our experiments were trained using PyTorch, with an Nvidia GeForce TITAN X GPU. For all the methods, the encoder and the decoders came from UNet [13]. We adopted the SGD optimizer to train the networks, setting the weight decay to 0.0001 and momentum to 0.9. We used no pretrained weights. We set the initial learning rate of the network to 0.01 and reduced it by a factor of 10 every 2500 iterations. The input batch size of the network was set to 4, consisting of two labeled images and two unlabeled images. We set the consistency weight factor λ as a time-dependent Gaussian warming-up function $\lambda(t) = 0.1 * e^{-5(1-t/t_{\max})^2}$, where t and $t_{\max}$ indicate the current and last training step, respectively. Because both branches were trained through mutual learning, we chose the better performance of the two branches as the final test result.

With 1000 or fewer iterations, we let the consistency loss equal 0, because the network parameters did not converge at the beginning, and the consistency loss was meaningless. With greater than 1000 iterations, we added the consistency loss to the total loss.

4.3. Comparison between PMCL and Other Methods. We compared the proposed method with existing methods on CVC-ClinicDB, CVC-ColonDB, and Kvasir-SEG. As shown in Tables 2–4, we implemented several semisupervised segmentation methods for comparison, including mean teacher (MT) [5], deep adversarial network (DAN) [10], entropy minimization (EM) [49], uncertainty aware mean teacher (UAMT) [8], and interpolation consistency training (ICT) [50]. Fully supervised utilized 100% labeled data to obtain an upper bound on performance. For fair comparisons, all methods utilized a UNet [13] backbone network.

Table 2 shows the results of comparative experiments on CVC-ClinicDB under 10% and 20% labeled images, taking the supervised-only method as the baseline. With 10% labeled images, we can see that all semisupervised methods show an improvement over the baseline because they can learn additional information from the unlabeled images by regularization loss. The proposed PMCL method shows steady and obvious improvement over other state-of-the-art semi-supervised learning methods on the six metrics. DSC has increased by 6.64%, 4.89%, 6.56%, 4.9%, and 5.65% compared with [5, 8, 10, 49, 50], respectively, by leveraging 10% labeled images and 90% unlabeled images. When using 20% of labeled images, all semisupervised learning methods improved. Our method still shows a notable performance improvement, as DSC has increased by 2.6%, 1.54%, 1.7%, 2.97%, and 0.79% compared with [5, 8, 10, 49, 50], respectively. The proposed PMCL outperforms the other methods on the DSC, JI, SE, and 95HD metrics.

Tables 3 and 4 show the performance of the proposed method and other state-of-the-art methods under 10% and 20% labeled images on CVC-ColonDB and Kvasir-SEG. For CVC-ColonDB, compared with other state-of-the-art semisupervised methods, on all six metrics, our method achieves the best performance under 10% and 20% labeled data. For Kvasir-SEG, our method performs best on five metrics under 10% labeled data and on four metrics under 20% labeled data. Through experiments on these three datasets, we can find that when using a small amount of labeled data, our method improves greatly compared with other methods, which means that it can more efficiently exploit unlabeled images compared with other semisupervised methods.

Figures 2–4 show the predicted segmentation results of the proposed PMCL and other methods under 10% labeled image settings on three datasets. Compared with other semisupervised approaches, the predicted segmentation map of our PMCL has a larger intersection rate with the ground truth, and its segmentation results are smoother in the edge area of the lesion.

Overall, the comparison experiments demonstrate that the PMCL framework can outperform other state-of-the-art methods under different numbers of labeled images, which means that our method is fully capable of learning the rich and effective information from the unlabeled images.

4.4. Ablation Study. To verify the impact of prototype mutual consistency learning and mutual supervision learning on the entire framework, we conducted ablation studies on CVC-ClinicDB. We designed a method to use the MT framework, replacing consistency loss with the proposed prototype mutual consistency learning, referred to as Prototype-MT. The proposed method utilizes mutual learning between the two branches, where both the supervision and consistency losses have two parts. We designed an experiment to explore the impact of the two parts on the overall network, proposing three framework structures based on our network framework for ablation experiments under 10% labeled data settings: PMCL-B1 uses the loss of the branch above, $L = L_{s1} + L_{c1}$; PMCL-B2 uses the loss of the branch below, $L = L_{s2} + L_{c2}$; and PMCL combines the two branch losses.

In Table 5, it can be observed that the performance of Prototype-MT is better than that of MT, which indicates that prototype mutual consistency learning can more effectively utilize unlabeled data and learn more reliable and rich knowledge from it. Moreover, the performance of PMCL significantly exceeds that of PMCL-B1 and PMCL-B2, which means that the two branches obtain better performance through mutual learning. The ablation experiments show

Table 2: Quantitative comparison between our method and other semisupervised methods on CVC-ClinicDB under 10% and 20% labeled data.

Method	Labeled/unlabeled	DSC↑ (%)	JI↑ (%)	SE↑ (%)	AC↑ (%)	95HD↓ (mm)	ASD↓ (mm)
Fully supervised	490/0	84.16	76.00	85.23	96.83	29.76	8.84
Supervised-only	49/0	61.01	50.72	66.96	92.68	80.68	28.56
MT [5]	49/441	62.36	52.38	66.97	93.08	77.20	26.43
DAN [10]	49/441	64.11	53.49	69.05	93.01	73.83	23.84
EM [49]	49/441	62.44	51.47	67.92	92.70	81.96	28.08
UAMT [8]	49/441	64.10	52.83	70.52	93.05	72.67	26.07
ICT [50]	49/441	63.35	53.33	68.38	92.79	68.38	23.27
PMCL (ours)	49/441	**69.00**	**58.50**	**74.86**	**93.55**	**66.68**	**23.08**
Supervised-only	98/392	72.33	61.44	75.65	94.48	60.12	19.23
MT [5]	98/392	73.04	63.33	76.74	94.59	48.86	**13.11**
DAN [10]	98/392	74.10	63.59	76.98	94.87	55.19	17.55
EM [49]	98/392	73.94	63.98	75.93	94.77	55.52	14.83
UAMT [8]	98/392	72.67	62.52	78.84	94.59	53.36	15.16
ICT [50]	98/392	74.85	64.87	76.84	**94.91**	50.35	14.53
PMCL (ours)	98/392	**75.64**	**65.61**	**81.09**	94.76	**48.15**	14.58

The bold values suggest the best performance compared to other state-of-the-art methods.

Table 3: Quantitative comparison between our method and other semisupervised methods on CVC-ColonDB under 10% and 20% labeled data.

Method	Labeled/unlabeled	DSC↑ (%)	JI↑ (%)	SE↑ (%)	AC↑ (%)	95HD↓ (mm)	ASD↓ (mm)
Fully supervised	304/0	81.05	74.16	80.41	98.78	20.57	5.10
Supervised-only	30/0	36.92	28.26	38.31	94.68	150.09	76.87
MT [5]	30/274	42.56	33.74	43.85	93.99	156.51	77.51
DAN [10]	30/274	44.04	35.11	44.23	95.48	133.45	69.00
EM [49]	30/274	43.39	34.66	42.42	95.60	124.81	62.38
UAMT [8]	30/274	43.45	34.84	42.57	95.43	156.10	82.26
ICT [50]	30/274	44.77	36.70	45.63	94.76	135.35	69.11
PMCL (ours)	30/274	**51.14**	**42.81**	**50.29**	**95.82**	**103.89**	**42.62**
Supervised-only	60/244	62.09	53.85	58.55	97.39	97.72	44.42
MT [5]	60/244	65.65	56.87	67.79	97.42	99.50	48.61
DAN [10]	60/244	66.14	58.02	66.65	97.66	91.36	45.87
EM [49]	60/244	65.29	56.74	67.19	97.37	112.44	45.24
UAMT [8]	60/244	64.82	55.81	64.27	97.77	115.44	46.78
ICT [50]	60/244	65.96	57.74	65.66	97.71	93.75	47.81
PMCL (ours)	60/244	**67.42**	**59.97**	**68.80**	**97.91**	**79.39**	**38.87**

The bold values suggest the best performance compared to other state-of-the-art methods.

Table 4: Quantitative comparison between proposed PMCL and other semisupervised methods on Kvasir-SEG under 10% and 20% labeled data.

Method	Labeled/unlabeled	DSC↑ (%)	JI↑ (%)	SE↑ (%)	AC↑ (%)	95HD↓ (mm)	ASD↓ (mm)
Fully supervised	800/0	81.79	73.16	84.62	95.13	77.60	23.53
Supervised-only	80/0	73.04	62.32	80.32	92.39	117.23	46.83
MT [5]	80/720	74.09	63.31	82.33	92.90	115.48	41.87
DAN [10]	80/720	75.29	65.14	81.33	92.88	106.65	39.26
EM [49]	80/720	74.73	64.38	**83.75**	92.69	121.12	42.55
UAMT [8]	80/720	74.66	64.32	81.61	92.68	112.18	38.61
ICT [50]	80/720	74.58	64.49	80.83	93.02	100.03	37.46
PMCL (ours)	80/720	**76.02**	**66.40**	81.53	**93.56**	**91.03**	**30.96**
Supervised-only	160/640	77.94	69.26	83.12	94.09	94.98	32.04
MT [5]	160/640	78.14	69.52	77.29	94.40	77.65	21.14
DAN [10]	160/640	78.38	70.04	78.79	94.48	81.64	20.31
EM [49]	160/640	78.59	69.47	83.23	94.28	91.34	31.94
UAMT [8]	160/640	78.42	69.96	80.30	94.56	78.85	23.26
ICT [50]	160/640	78.78	70.55	78.90	94.48	**70.33**	**19.71**
PMCL (ours)	160/640	**79.98**	**70.92**	**84.24**	**94.63**	97.83	31.01

The bold values suggest the best performance compared to other state-of-the-art methods.

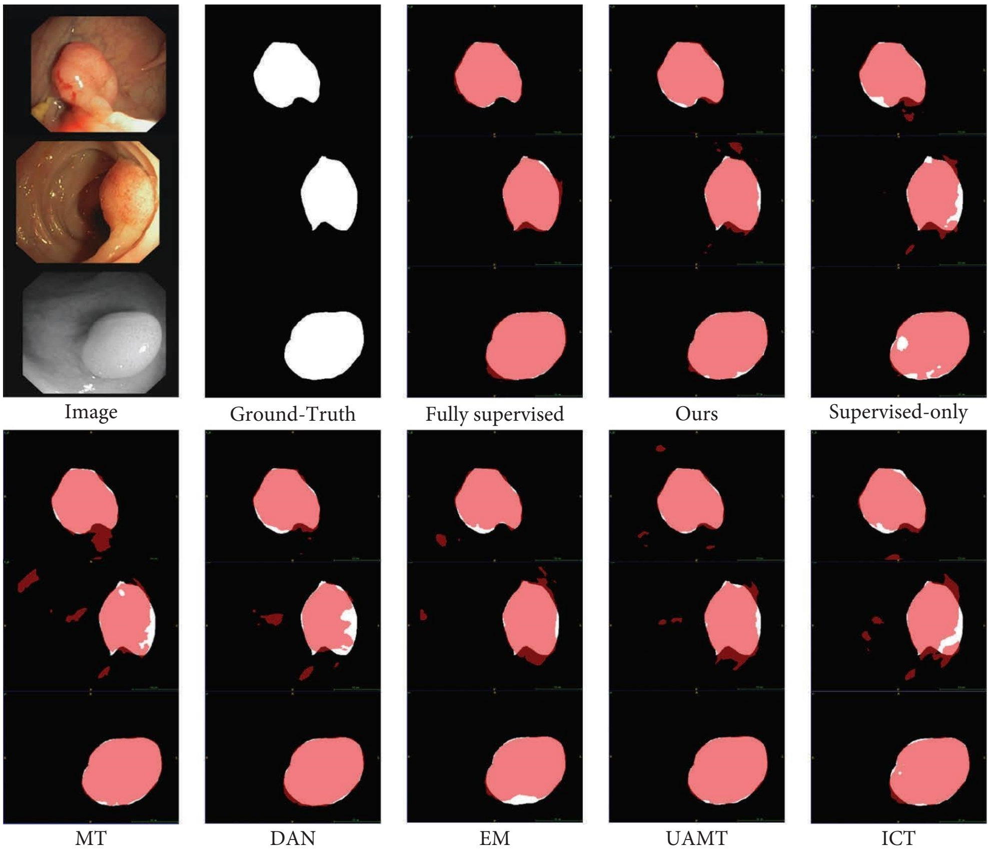

Figure 2: Visual comparison on CVC-ClinicDB under 10% labeled data settings, where the red color indicates predicted polyps.

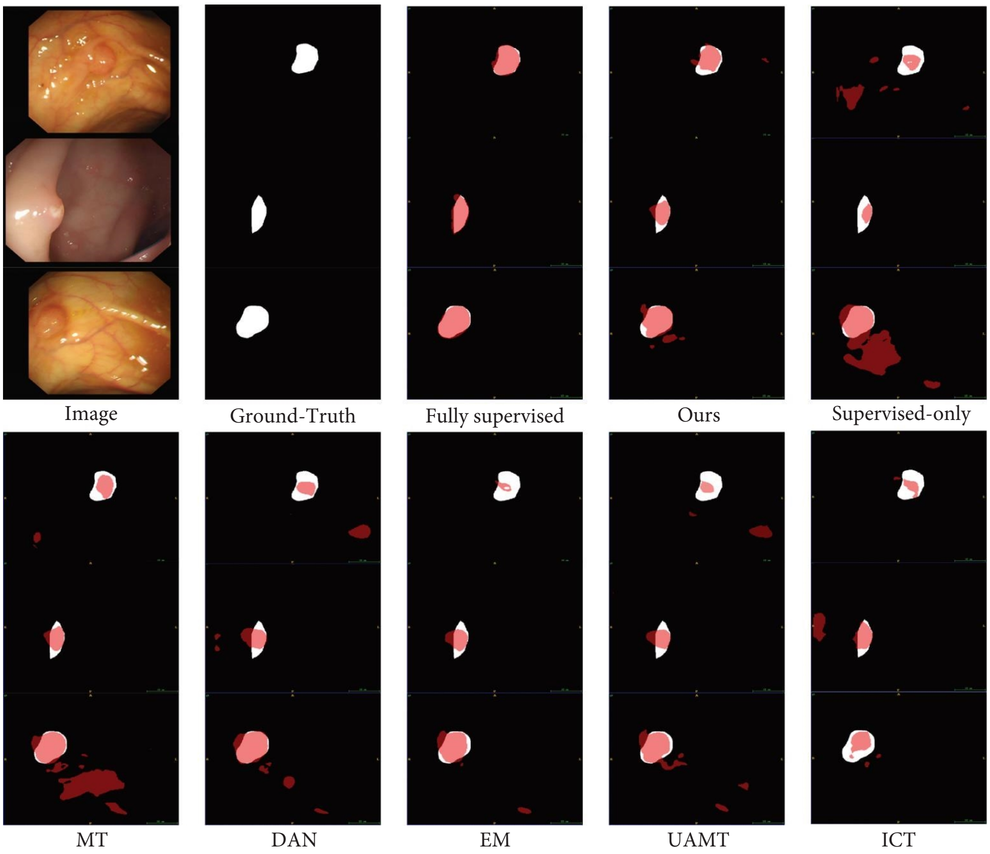

Figure 3: Visual comparison on CVC-ColonDB under 10% labeled data settings, where the red color indicates predicted polyps.

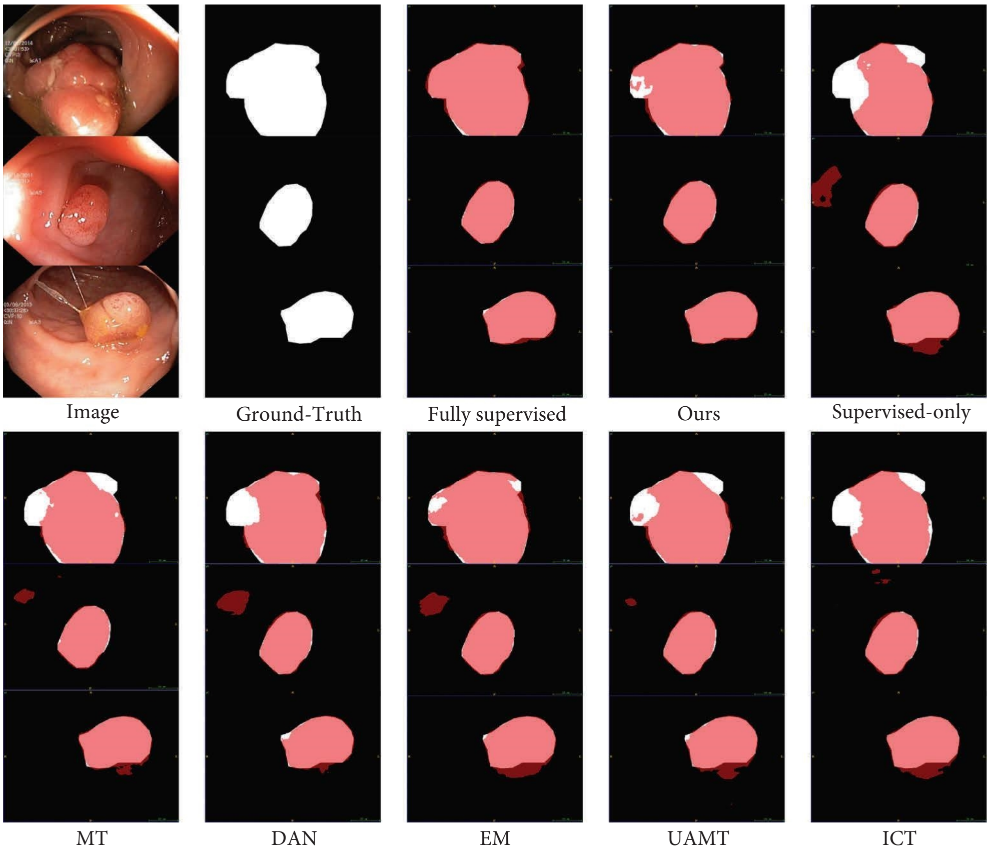

Figure 4: Visual comparison on Kvasir-SEG under 10% labeled data settings, where the red color indicates predicted polyps.

Table 5: Ablation study on CVC-ClinicDB under 10% labeled data settings.

Method	DSC	JI	SE	AC	95HD	ASD
MT [5]	62.36	52.38	66.97	93.08	77.20	26.43
Prototype-MT	63.58	53.25	67.55	93.24	67.16	26.24
PMCL-B1	64.70	54.07	69.97	93.15	84.34	27.57
PMCL-B2	64.25	53.47	72.96	92.66	74.43	27.98
PMCL	**69.00**	**58.50**	**74.86**	**93.55**	**66.68**	**23.08**

The bold values suggest the best performance compared to other models in the ablation study.

that our mutual learning framework can learn rich information from both branches and effectively improve network performance.

5. Conclusion

We investigated common methods for semisupervised medical image segmentation and proposed the PMCL framework. Through experiments on these three datasets, it can be found that when using a small number of labeled images, the PMCL framework has a greater improvement than other methods. This is because the proportion of labeled data is smaller, the semisupervised method can utilize less reliable information, and the proportion of unlabeled data is higher. Therefore, the semisupervised method can extract more information from unlabeled data. At this point, different semisupervised learning methods have significant differences in their ability to extract information, resulting in significant differences in the final results.

From the experiment, it can be seen that the PMCL method can more fully utilize unlabeled images to improve network performance compared to other semisupervised methods. The proposed method makes full use of a mutual learning framework to improve its performance and robustness. We designed prototype mutual consistency learning to obtain more reliable consistency loss for unlabeled images and supervision mutual learning for labeled images. Experiments demonstrated that our method has potential in semisupervised segmentation tasks.

Data Availability

The data that support the findings of this study are available on request from the corresponding author.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (grant number 62176181).

References

[1] A. Srivastava, D. Jha, S. Chanda et al., "Msrf-net: a multi-scale residual fusion network for biomedical image segmentation," 2021, https://arxiv.org/abs/2105.07451.

[2] M. Benčević, I. Galić, M. Habijan, and D. Babin, "Training on polar image transformations improves biomedical image segmentation," *IEEE Access*, vol. 9, 133375 pages, 2021.

[3] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "Doubleu-net: a deep convolutional neural network for medical image segmentation," in *Proceedings of the 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 558–564, Rochester, MN, USA, July 2020.

[4] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 1–13, Austria, May 2017.

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

[6] Z. Xie, E. Tu, H. Zheng, Y. Gu, and J. Yang, "Semi-supervised skin lesion segmentation with learning model confidence," in *Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1135–1139, Toronto, ON, Canada, June 2021.

[7] X. Li, L. Yu, H. Chen, C.-W. Fu, L. Xing, and P.-A. Heng, "Transformation-consistent self-ensembling model for semisupervised medical image segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 523–534, 2021.

[8] 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," in *Proceedings of the Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pp. 605–613, Shenzhen, China, October 2019.

[9] Z. Zhang, C. Tian, and Z. Jiao, "Mutual-and self-prototype alignment for semi-supervised medical image segmentation," 2022, https://arxiv.org/abs/2206.01739.

[10] Y. Zhang, L. Yang, J. Chen, M. Fredericksen, D. P. Hughes, and D. Z. Chen, "Deep adversarial networks for biomedical image segmentation utilizing unannotated images," in *Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI*, pp. 408–416, Quebec City, QC, Canada, September 2017.

[11] S. Li, C. Zhang, and X. He, "Shape-aware semi-supervised 3d semantic segmentation for medical images," in *Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2020*, pp. 552–561, Lima, Peru, October 2020.

[12] X. Luo, J. Chen, T. Song, and G. Wang, "Semi-supervised medical image segmentation through dual-task consistency," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 10, pp. 8801–8809, 2021.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*, pp. 234–241, Munich, Germany, October 2015.

[14] Y. Li, J. Chen, X. Xie, K. Ma, and Y. Zheng, "Self-loop uncertainty: a novel pseudo-label for semi-supervised medical image segmentation," in *Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2020*, pp. 614–623, Lima, Peru, October 2020.

[15] S. Sedai, B. Antony, R. Rai et al., "Uncertainty guided semi-supervised segmentation of retinal layers in oct images," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 282–290, Springer, Vancouver, Canada, October 2019.

[16] 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)*, Aberdeen, November 2018.

[17] T. Miyato, S.-I. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1979–1993, 2019.

[18] A. G. Howard, M. Zhu, B. Chen et al., "Mobilenets: efficient convolutional neural networks for mobile vision applications," 2017, https://arxiv.org/abs/1704.04861.

[19] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: an extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6848–6856, Salt Lake City, UT, USA, June 2018.

[20] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, https://arxiv.org/abs/1503.02531.

[21] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4320–4328, Salt Lake City, UT, USA, June 2018.

[22] Y. Wu, M. Xu, Z. Ge, J. Cai, and L. Zhang, "Semi-supervised left atrium segmentation with mutual consistency training," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 297–306, Springer, Strasbourg, France, October 2021.

[23] Y. Zhang and J. Zhang, "Dual-task mutual learning for semi-supervised medical image segmentation," in *Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 548–559, Springer, Xiamen, China, October 2021.

[24] Y. Zhang, J. Yang, J. Tian et al., "Modality-aware mutual learning for multi-modal medical image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 589–599, Springer, Strasbourg, France, October 2021.

[25] Z. Xu, Y. Wang, D. Lu et al., "All-around real label supervision: cyclic prototype consistency learning for semi-supervised medical image segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 7, pp. 3174–3184, 2022.

[26] Z. Zhang, R. Ran, C. Tian et al., "Self-aware and cross-sample prototypical learning for semi-supervised medical image segmentation," 2023, https://arxiv.org/abs/2305.16214.

[27] J. Snell, K. Swersky, and R. Zemel, “Prototypical networks for few-shot learning,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[28] O. Vinyals, C. Blundell, T. Lillicrap, and D. Wierstra, “Matching networks for one shot learning,” *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[29] H. L. Sachin Ravi, “Optimization as a model for few-shot learning,” in *Proceedings of the International Conference on Learning Representations*, pp. 214–222, Austria, May 2017.

[30] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” in *Proceedings of the International conference on machine learning. PMLR*, pp. 1126–1135, Vienna, Austria, July 2017.

[31] J. B. Svictor Garcia, “Few-shot learning with graph neural networks,” in *Proceedings of the International Conference on Learning Representations*, pp. 3916–3924, Austria, May 2018.

[32] Y. Liu, J. Lee, M. Park et al., “Learning to propagate labels: transductive propagation network for few-shot learning,” in *Proceedings of the International Conference on Learning Representations*, Austria, May 2019.

[33] A. Shaban, S. Bansal, Z. Liu, I. Essa, and B. Boots, “One-shot learning for semantic segmentation,” 2017, https://arxiv.org/abs/1709.03410.

[34] N. Dong and E. P. Xing, “Few-shot semantic segmentation with prototype learning,” *British Machine Vision Conference (BMVC)*, vol. 3, no. 4, 2018.

[35] X. Zhang, Y. Wei, Y. Yang, and T. S. Huang, “Sg-one: similarity guidance network for one-shot semantic segmentation,” *IEEE Transactions on Cybernetics*, vol. 50, no. 9, pp. 3855–3865, 2020.

[36] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, “Panet: few-shot image semantic segmentation with prototype alignment,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 9197–9206, Seoul, Korea (South), October 2019.

[37] C. Zhang, G. Lin, F. Liu, R. Yao, and C. Shen, “Canet: class-agnostic segmentation networks with iterative refinement and attentive few-shot learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5217–5226, Long Beach, CA, USA, June 2019.

[38] K. Nguyen and S. Todorovic, “Feature weighting and boosting for few-shot segmentation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 622–631, Seoul, Korea (South), October 2019.

[39] M. Siam, B. N. Oreshkin, and M. Jagersand, “Amp: adaptive masked proxies for few-shot segmentation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5249–5258, Seoul, Korea (South), October 2019.

[40] G. Li, V. Jampani, L. Sevilla-Lara, D. Sun, J. Kim, and J. Kim, “Adaptive prototype learning and allocation for few-shot segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8334–8343, Nashville, TN, USA, June 2021.

[41] Y. Liu, X. Zhang, S. Zhang, and X. He, “Part-aware prototype network for few-shot semantic segmentation,” in *Proceedings of the European Conference on Computer Vision*, pp. 142–158, Springer, Glasgow, UK, August 2020.

[42] C. Ouyang, C. Biffi, C. Chen, T. Kart, H. Qiu, and D. Rueckert, “Self-supervision with superpixels: training few-shot medical image segmentation without annotation,” in *Proceedings of the European Conference on Computer Vision*, pp. 762–780, Springer, Glasgow, UK, August 2020.

[43] K. Rakelly, E. Shelhamer, T. Darrell, A. A. Efros, and S. Levine, “Conditional networks for few-shot semantic segmentation,” in *Proceedings of the International Conference on Learning Representations*, Austria, May 2018.

[44] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, “Wm-dova maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians,” *Computerized Medical Imaging and Graphics*, vol. 43, pp. 99–111, 2015.

[45] N. Tajbakhsh, S. R. Gurudu, and J. Liang, “Automated polyp detection in colonoscopy videos using shape and context information,” *IEEE Transactions on Medical Imaging*, vol. 35, no. 2, pp. 630–644, 2016.

[46] D. Jha, P. H. Smedsrud, M. A. Riegler et al., “Kvasir-seg: a segmented polyp dataset,” in *Proceedings of the International Conference on Multimedia Modeling*, pp. 451–462, Springer, Bergen, Norway, January 2020.

[47] D. Jha, P. H. Smedsrud, M. A. Riegler et al., “Resunet++: an advanced architecture for medical image segmentation,” in *Proceedings of the 2019 IEEE International Symposium on Multimedia (ISM)*, pp. 225–2255, San Diego, CA, USA, December 2019.

[48] D.-P. Fan, G.-P. Ji, T. Zhou et al., “Pranet: parallel reverse attention network for polyp segmentation,” 2020, https://arxiv.org/abs/2006.11392.

[49] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Perez, “Advent: adversarial entropy minimization for domain adaptation in semantic segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, June 2019.

[50] V. Verma, A. Lamb, J. Kannala, Y. Bengio, and D. Lopez-Paz, “Interpolation consistency training for semi-supervised learning,” in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp. 3635–3641, Macao, August 2019.