Collaborative and Adversarial Learning of Focused and Dispersive Representations for Semi-supervised Polyp Segmentation

Huisi Wu[1] Guilian Chen[1] Zhenkun Wen[1] Jing Qin[2]
[1]Shenzhen University [2]The Hong Kong Polytechnic University

Abstract

Automatic polyp segmentation from colonoscopy images is an essential step in computer aided diagnosis for colorectal cancer. Most of polyp segmentation methods reported in recent years are based on fully supervised deep learning. However, annotation for polyp images by physicians during the diagnosis is time-consuming and costly. In this paper, we present a novel semi-supervised polyp segmentation via collaborative and adversarial learning of focused and dispersive representations learning model, where focused and dispersive extraction module are used to deal with the diversity of location and shape of polyps. In addition, confidence maps produced by a discriminator in an adversarial training framework shows the effectiveness of leveraging unlabeled data and improving the performance of segmentation network. Consistent regularization is further employed to optimize the segmentation networks to strengthen the representation of the outputs of focused and dispersive extraction module. We also propose an auxiliary adversarial learning method to better leverage unlabeled examples to further improve semantic segmentation accuracy. We conduct extensive experiments on two famous polyp datasets: Kvasir-SEG and CVC-Clinic DB. Experimental results demonstrate the effectiveness of the proposed model, consistently outperforming state-of-the-art semi-supervised segmentation models based on adversarial training and even some advanced fully supervised models.

1. Introduction

Automatic segmentation of polyp plays a key role in computer-aided diagnosis for Colorectal cancer (CRC), which is one of the most common type of cancer around the world [26]. As for polyps, colonoscopy is an significant detection way, which can help in the removal of the polyps and greatly prevent them from developing into the CRC.

Recent years, convolutional neural networks (CNNs) have exhibited excellent performance in the image segmentation tasks. Semantic segmentation aims to assign each pixel of an image with a label so that the pixels with the same label can be used to infer the target of wanted. For semantic segmentation tasks, most approaches based on deep learning methods can be utilized in different medical tasks [6, 29, 33, 34, 36] like segmentation of blood vessel, skin lesion, lung nodule, and cell nuclei. Compared to manual segmentation by physicians during the diagnosis, which is time-consuming and subjective, medical automatic segmentation has great advantages and huge potential in computer-aided diagnosis (CAD). However, lacking of a large number of pixel-wise annotations for training is a great challenge for automatic medical segmentation task. Annotating medical data such as polyps, which vary in shape, texture and appearance location, always needs a lot of time and effort, results in a challenging task for polyp segmentation.

Commonly in some medical image segmentation tasks, allocating each pixel with a correct label in blurred images is very hard. However, obtaining the global information of blurred images can effectively resolve the problem about distinguishing target and background in a segmentation task. For this purpose, the non-local network [31] is proposed to model the long-range dependencies using self-attention mechanism [28]. Furthermore, Cao et al. [3] proposed a new instantiation of the general network, called global context (GC) block by a combination of the optimal implementation of non-local (NL) block and squeeze-and-excitation (SE) block [11] at each step. In addition, convolutional block attention module (CBAM) provides a simple yet effective attention mechanism for feed-forward CNNs [32]. But these methods for distinguishing target and background is in view of paired medical images and its ground truth images.

In recent years, different kinds of semi-supervised method are proposed for dealing with the shortage of labeled data. Hung et al. [13] adopted adversarial learning by using a segmentation network and a discriminator network to produce the confidence map based on the segmentation predictions of unlabeled images as supervisory signals to train the segmentation network. Another adversarial training method can be found in [22] by using a virtual adversarial regularization. Besides, the methods of consistency are widely used in semi-supervised training for both classification task and segmentation task [15, 24, 27]. The augmentation methods of perturbations on images have also proven their effectiveness for evaluating semi-supervised regularizers[9].

In this paper, considering the characteristics of polyp and the limited of pixel-wise annotations, we propose a novel semi-supervised polyp segmentation method, which uses collaborative segmentation networks with two different feature extraction modules to capture the location and edge information of polyp and uses an adversarial learning method similar to [13] and another method for auxiliary adversarial learning to fine tune the segmentation networks with unlabeled images. We perform extensive experiments to evaluate our method on the Kvasir dataset [14] and CVC-Clinic DB [2] that the experiments results show the effectiveness of our method.

Our contributions are summarized as follows:

- We propose a novel semi-supervised method for polyp segmentation, named collaborative and adversarial learning of focused and dispersive representations learning model. We propose two extraction modules, FEM and DEM, in the encoding path of two segmentation networks respectively. FEM allows our network to capture the focused information of the input feature maps like location information and spatial information, while DEM attempts to aggregate the scattered boundary information of the inputs.

- We simultaneously train two segmentation networks and a discriminator network with labeled images through an adversarial training method. With the help of consistency constraint, we can take the advantage of the two feature maps of FEM and DEM to produce confidence maps by the trained discriminator network with high credibility. The confidence maps generated based on unlabeled images in semi-supervised training stage can be used as supervised signals to fine-tune the segmentation networks.

- Another adversarial training method, named auxiliary adversarial learning (AAL), is proposed to improve the quality of segmentation predictions from unlabeled images in the semi-supervised training stage. We adopt a new discriminator to assign the true label for the segmentation results of labeled images and fake label for the prediction of unlabeled images. With AAL, we can obtain confidence maps with higher credibility which can be better utilized for the segmentation networks.

2. Related Work

2.1. CNN-based Polyp Segmentation

Polyp segmentation and recognition is essential for the patient to prevent the death caused by colorectal cancer. Alexander et al. [21] proposed to identify the polyps in the video sequences based on binary classification and preselection, which required predefined image features. During the last few years, U-Net [25] have been successfully applied to many medical semantic segmentation tasks and applications, which is based on a well-known encoder-decoder architecture to obtain location and context information from input data and infer a relative prediction. Similarly, many modifications and improvements based on U-Net structure were proposed to enhance the segmentation performance [1, 10, 18, 30, 39]. For the automatic polyp segmentation task, several representative networks were also developed to improve the polyp segmentation performance from different aspects, including ResU-Net [14, 37], U-Net++ [40], PraNet [7] and HarDNet-MSEG [12]. ResU-Net applies residual blocks to supplement the location information of polyps, while HarDNet-MSEG consists of the encoder of HarDNet68 [4] and the decoder of Cascaded partial decoder [35] with receptive field block [20] to improve both accuracy and inference speed. Besides, PraNet adopted three reverse attention modules with a parallel partial decoder connection to strengthen the area-boundary constraint for polyp segmentation. However, these methods are based on fully-supervised training strategies. Fully-supervised methods usually require sufficient labeled medical samples for training, but annotating medical data such as polyp images is often expensive and time consuming. In this regard, semi-supervised segmentation method is a better direction to achieve satisfying accuracy for polyp segmentation from limited labeled images.

2.2. Semi-supervised Training

Due to the lack of labeled images for training, semi-supervised methods turn to leverage unlabeled data to obtain useful information. For example, Li et al. [17] proposed a semi-supervised network for the skin lesion segmentation task, which only used 15% labeled images and obtained a similar performance of several fully-supervised methods. Similarly, several pseudo labeling methods [16, 38] also successfully extracted useful information from the unlabeled data to enhance the model training. For semi-supervised segmentation tasks, several representative adversarial learning methods [13, 23] were also proposed to improve the performance of segmentation networks. Hung's method [13] employed the output of a fully convolutional discriminator as supervisory signals, which is combined with self-taught learning framework to provide more useful pseudo labeling information for semi-supervised train-

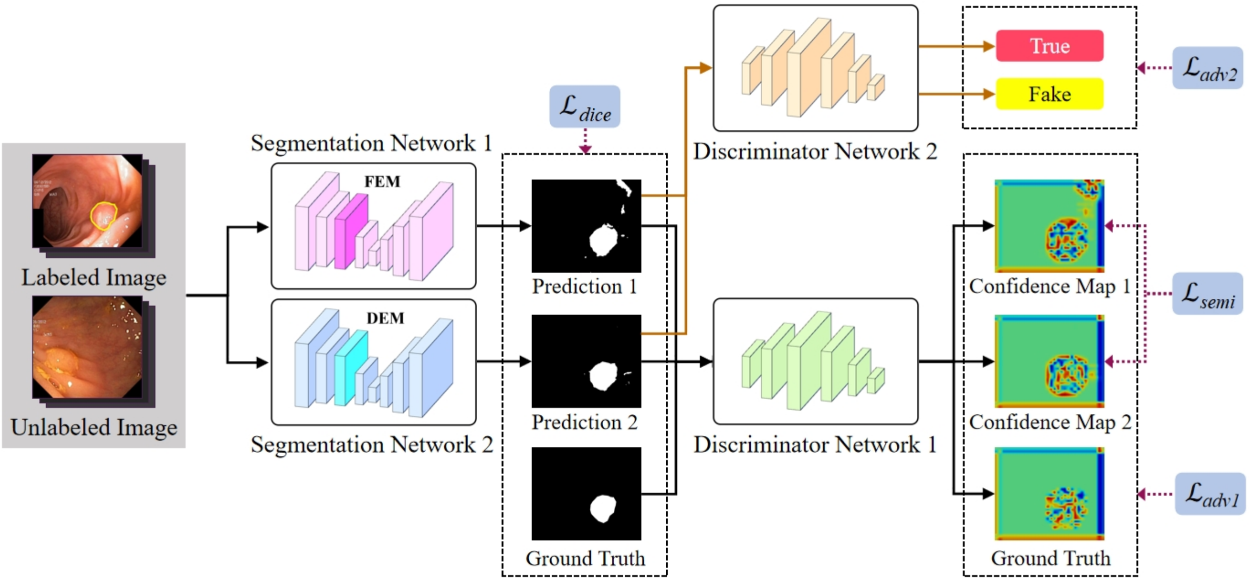

Figure 1. Our proposed collaborative and adversarial learning framework. In supervised training stage, two collaborative segmentation networks equipped with focused extraction module (FEM) and dispersive extraction module (DEM) are applied to generate predictions for the first discriminator network. In semi-supervised training stage, an auxiliary discriminator network is further employed to fine-tune the collaborative segmentation networks by minimizing the bias of the first discriminator due to data imbalance.

ing. Generative Adversarial Network (GAN) method [23] also produced realistic fake examples to prevent over-fitting for segmenting 3D multi-modality images. However, traditional semi-supervised adversarial learning method usually only contains a single segmentation network to implement the generator, which may cause deviation because of over emphasizing targeted aspects. As a result, the obtained information from the model of poor segmentation accuracy may generate wrong guidance from unlabeled data. In this paper, we employ two collaborative segmentation networks to overcome this weakness. On the other hand, traditional adversarial training framework also usually only employed one discriminator trained with limited labeled data, which also may easily suffer from the imbalance between the labeled and unlabeled data, resulting in less valid information from unlabeled data can be used for the semi-supervised training. To overcome this limitation, we further apply another discriminator as auxiliary training module to improve the utilization of unlabeled data.

3. Method

The framework of our proposed collaborative and adversarial learning method is illustrated in Figure 1, where two collaborative segmentation networks are trained under two adversarial learning stages.

3.1. Collaborative Segmentation Networks

Unlike traditional fully supervised networks, which are usually trained with sufficient labeled images, the generator in our semi-supervised framework for polyp segmentation only utilizes limited labeled images. We observed that single segmentation network may easily incur biases by over emphasizing targeted aspects in designing the network, especially for the insufficient condition of limited labeled images. Due to no reference like ground truth in the semi-supervised training stage, the method with single segmentation network has to accept the information from unlabeled data no matter it is right or wrong. Obviously, inaccurate confidence maps from unlabeled data based on bias or wrong predictions may have misleading guidance in the following semi-supervised training stage.

To achieve a more accurate and stable generator, we employ two collaborative segmentation networks in our semi-supervised framework, which are optimized under a mutual consistency constraint loss to minimize the bias. As shown in Figure 1, we apply two collaborative segmentation networks to segment the polyp images, and one discriminator to create confidence maps according to the segmentation results or ground truth.

Focused Extraction Module (FEM). Accurately identifying localization and position features of polyps is an essential aspect in the task of polyp segmentation. Given the feature map extracted in each layer of the original encoder in U-Net, we introduce a FEM to further extract focused location features of polyps from colonoscopy images.

As shown in Figure 2, our FEM first aggregates spatial and global context features based on average pooling [19] and global attention pooling (GAP) [3]. Based on the at-

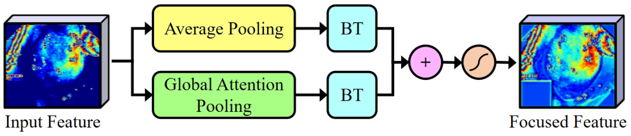

Figure 2. Focused extraction module. By aggregating spatial and global context features based on average pooling and global attention pooling, FEM can effectively extract focused location and shape features of polyps from colonoscopy images. In addition, bottleneck transform (BT) is also applied to reduce the number of parameters for capturing the channel-wise dependencies, where we not only can reduce the risk of overfitting, but also obtain a better optimization efficiency.

tention weights [32] calculated from average-pooling and GAP, we can further perform attention pooling operations to obtain global context features, which mainly focus on the location features of polyps. Furthermore, to minimize the risk of overfitting and achieve a better optimization efficiency, we also introduce a bottleneck transform (BT) after average-pooling and GAP to reduce the number of parameters for capturing the channel-wise dependencies, where a layer normalization is applied inside the bottleneck transform before the ReLU operation. Obviously, the BT block has a similar effect with the excitation operation in squeeze-and-excitation (SE) block [11]. Finally, we obtain the focused features by merging the two feature vectors and aggregating the global context features based on a broadcast element-wise summation and sigmoid activation.

Dispersive Extraction Module (DEM). Accurately capturing boundary features of polyps is another essential aspect in the task of polyp segmentation. Different from FEM, which obviously conceals more scattered boundary features for polyp segmentation, dispersive attention is emphasized in the implementation of DEM. Similarly, we also employ a DEM to the given feature map extracted in each layer of the original encoder in U-Net. As shown in Figure 3, our DEM first apply three dilated convolutions to extract dispersive information from the input feature map, where dilated convolutions have the same small kernel size of 3×3 but with different dilated rates ($r = 1, 2, 5$). In addition, max-pooling operation is also adopted to infer finer channel-wise attention inside three dispersive feature maps. To encode channel-wise representations, we further apply fully connected layer (FC) on each extracted dispersive feature map. Finally, we can obtain the aggregated dispersive feature map based on a broadcast element-wise multiplication, where important scattered boundary feature points are reinforced in the output of our DEM.

Collaborative Learning. Different from most of existing semi-supervised frameworks, which usually rely on a single segmentation network in the generator and easily incur biases under insufficient guidance of limited labeled images, our generator can reduce the biases and enhance the segmentation robustness via collaborative learning between two collaborative segmentation networks. Specifically, we first apply a discriminator based on FCN [13] to produce confidence maps according to the segmentation results of two collaborative networks. In the supervised training stage, we apply a consistency constraint to the confidence maps generated in both FEM network and DEM network, making sure that the targeted regions of interest are similar, which not only strengthens the feature representation, but also reduces the biases single segmentation network via collaborative learning. Besides, dice loss on different confidence maps is used to prompt the results of two collaborative segmentation networks equaling to the ground truth. Therefore, with the help of collaborative segmentation networks, the discriminator can produce confidence maps with higher credibility than training a single segmentation network in the supervised training stage. In the semi-supervised learning, inaccurate confidence maps from unlabeled data based on bias or wrong predictions may easily produce misleading guidances due to no reference like ground truth in the semi-supervised learning. Obviously, our consistency constraint on confidence maps still provide reinforcing signals to fine tune the segmentation networks according to sharing weights of targeted regions of interest between two collaborative segmentation networks.

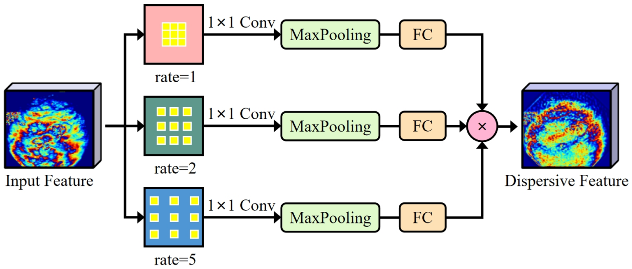

Figure 3. Dispersive extraction module. By extracting dispersive features based on three different dilated convolutions followed by a max-pooling operation and a fully connected layer (FC), DEM pays more attentions on aggregating scattered boundary features for polyp segmentation.

3.2. Primary and Auxiliary Adversarial Learning

Primary Adversarial Learning. In supervised training stage, the first discriminator in our framework is not only trained to generate the confidence maps, but also in charge of performing primary adversarial learning. During this adversarial training process, the primary discriminator considers that the confidence map generated based on ground truth is credible, and the confidence map generated from the results of two collaborative networks is unbelievable. Through continuous training, collaborative networks try to fool the discriminator by producing the confidence maps very close to the ground truth. Therefore, the

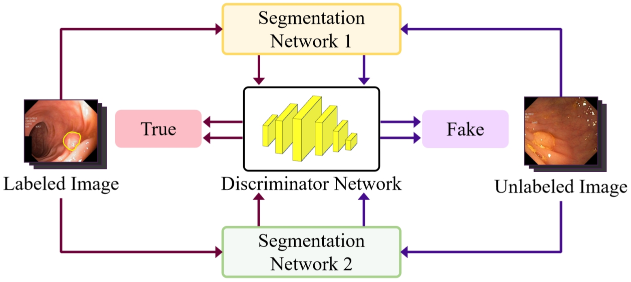

Figure 4. Auxiliary adversarial learning (AAL). To improve the segmentation confidence of unlabeled images in the semi-supervised stage, auxiliary discriminator network is trained to distinguish the segmentation results between the labeled images and the unlabeled images.

confidence maps generated based on collaborative networks can be more acceptable after primary adversarial learning, which also conversely improve the accuracy of the segmentation networks by producing more valid and stable segmentation results. For the unlabeled data in the semi-supervised training stage, we can also obtain corresponding confidence maps from the trained discriminator network according to the segmentation predictions of the two collaborative networks, as shown in Figure 1. However, if labeled images used for training the discriminator network is very limited and the amount of unlabeled data is much more than labeled data, the primary discriminator may generate confidence maps for the unlabeled images with insufficient confidence, which potentially produces misleading guidance in the following semi-supervised training stage. To minimize the impact of imbalance problem between labeled and unlabeled data, we further introduce an auxiliary discriminator and utilize auxiliary adversarial learning (AAL) to make up the insufficient training of the primary discriminator due to limited labeled polyp images.

Auxiliary Adversarial Learning (AAL). Inspired by the principle of generative adversarial network(GAN), we propose an AAL in semi-supervised training stage to improve the performance of segmentation of unlabeled data and relieve the insufficient trained primary discriminator due to limited labeled images. As shown in Figure 4, AAL servers as an auxiliary discriminator, which is trained to distinguish the segmentation results by judging the segmentation of labeled images to be true and the segmentation of unlabeled images to be fake respectively. According to the adversarial training for the auxiliary discriminator, the collaborative FEM network and DEM network are tend to optimize the quality and confidence of segmentation results for unlabeled images, and produce a similar quality and confidence of segmentation results as the labeled images. By promoting the collaborative segmentation networks to generate results with high quality, the corresponding confidence maps for both labeled and unlabeled images can be more credible. Therefore, we can finally obtain a better trained primary discriminator and further fine-tune the collaborative segmentation networks in semi-supervised training stage.

3.3. Loss Functions

Collaborative Segmentation Loss. Given an input images X_i with a resolution of $H \times W \times 3$, we can obtain two focused feature vectors V_{avg} and V_{glb} for FEM network through the average-pooling and global attention pooling respectively. By applying BT operations to capture channel-wise dependencies and utilizing broadcast element-wise addition for feature fusion, we can finally obtain the focused feature map f_i written as

$$f_i = \delta(\alpha_{b2} ReLU(LN(\alpha_{b1}(V_{avg}, V_{glb})))) \tag{1}$$

where δ denotes the sigmoid function; α_{b1} and α_{b2} denote a 1×1 convolution. Similarly, we can obtain the dispersive map d_i written as

$$d_i = \prod_{j=1}^{3} FC(V_{r_j}) \tag{2}$$

where V_{r_j} represents one of the three dispersive feature vectors obtained via three different dilated convolutions and max-pooling.

In supervised training stage, the Dice loss for training the collaborative segmentation networks can be written as

$$\mathcal{L}_{dice} = \left(1 - \frac{2|S(f_i) \bigcap Y_i|}{|S(f_i)| + |Y_i|}\right) + \left(1 - \frac{2|S(d_i) \bigcap Y_i|}{|S(d_i)| + |Y_i|}\right) \tag{3}$$

where $S(f_i)$ and $S(d_i)$ denote the outputs of FEM network and DEM network respectively. Y_i denotes the ground truth images.

For adversarial training in supervised training stage, the loss L_{adv} is formulated as

$$\mathcal{L}_{adv} = -\Big(\sum_{h,w} log\big(D_1(S(f_i))^{(h,w)}\big) + \sum_{h,w} log\big(D_1(S(d_i))^{(h,w)}\Big) \tag{4}$$

where $D_1(\cdot)$ denotes a fully convolutional discriminator network. In addition, we also apply a consistency constraint on the $D_1(S(f_i))$ and $D_1(S(d_i))$ to train our collaborative segmentation networks, and formulate the consistency loss L_C as

$$\mathcal{L}_C = \|D_1(S(f_i)) - D_1(S(d_i))\|^2 \tag{5}$$

obviously, above mean square error (MSE) loss can be used as a consistency loss in the supervised training of labeled images, as well as in semi-supervised training of unlabeled images. In addition, we also reshape the confidence maps of unlabeled images to with the same resolution of the segmentation prediction, and denote the signal maps S' as

$$S'(f_i, d_i) = \sum_{h,w} \Big(D_1(S(f_i, d_i))^{(h,w)} > T_{semi}\Big) \cdot \widehat{Y}_i^{(h,w)} \tag{6}$$

where $\widehat{Y}_i$ is element-wise set with $\widehat{Y}_i^{(h,w)} = 1$ if the location of current point is same as $argmaxS(f_i, d_i)^{(h,w)}$. In our experiment, $T_{semi} = 0.2$ is a threshold to control what kinds of points can be believed. So far, we can apply the Dice loss and consitentcy loss on the signal maps and confidence maps respectively to fine tune the FEM network and DEM network, forming a semi-supervised loss as

$$\begin{aligned}\mathcal{L}_{semi} = \Bigg(& \Big(1 - \frac{2|S(f_i) \bigcap S'(f_i)|}{|S(f_i)| + |S'(f_i)|}\Big) + \\ & \Big(1 - \frac{2|S(d_i) \bigcap S'(d_i)|}{|S(d_i)| + |S'(d_i)|}\Big)\Bigg) + \mathcal{L}_C^{unlabel}\end{aligned} \tag{7}$$

To enhance the quality of confidence maps for unlabeled data in semi-supervised training, our auxiliary adversarial loss is written as

$$\mathcal{L}'_{adv} = -\Big(log\big(D_2(S(f_i))\big) + log\big(D_2(S(d_i))\big)\Big) \tag{8}$$

Finally, we obtain an overall loss for training the collaborative segmentation networks as follow

$$\mathcal{L}_{seg} = \mathcal{L}_{dice} + \mathcal{L}_C^{label} + \lambda_{adv}\mathcal{L}_{adv} + \lambda'_{adv}\mathcal{L}'_{adv} + \lambda_{semi}\mathcal{L}_{semi} \tag{9}$$

where λ_{adv} and λ'_{adv} are weights for the two kinds of adversarial loss in supervised training stage and semi-supervised training stage respectively. λ_{semi} is the weight for the semi-supervised loss. In our implements, we set $\lambda_{adv} = 0.01$, $\lambda'_{adv} = 0.001$ and $\lambda_{semi} = 0.1$ respectively.

Discriminator Loss. Similar to Hung's [13], we also train the primary discriminator by minimizing the spatial cross entropy loss written as

$$\begin{aligned}\mathcal{L}_{D_1} = -\frac{1}{2}\Big(& \sum_{h,w}(1 - y_i)log\big(1 - D_1(S(f_i))^{(h,w)}\big) + \\ & \sum_{h,w}(1 - y_i)log\big(1 - D_1(S(d_i))^{(h,w)}\big)\Big) + \\ & y_i log\Big(D_1(Y_i)^{(h,w)}\Big)\end{aligned} \tag{10}$$

where $y_i = 0$ if the sample is drawn from the segmentation networks, while $y_i = 1$ if the sample from the ground truth label. Besides, D_1 denotes discriminator to produce the confidence maps.

Instead of producing confidence maps, the auxiliary discriminator D_2 produces values either 0 (for fake) or 1 (for true). We can formulate its cross entropy loss as

$$\begin{aligned}\mathcal{L}_{D_2} = -\frac{1}{2}\lambda_{D_2}\Bigg(& \bigg((1 - y_i)log\Big(1 - D_2(S(f_i))\Big) + \\ & (1 - y_i)log\Big(1 - D_2(S(d_i))\Big)\bigg)\Bigg) + \\ & \frac{1}{2}\lambda'_{D_2}\bigg(y_i log\Big(D_2(S(f_i))\Big) + y_i log\Big(D_2(S(d_i))\Big)\bigg)\end{aligned} \tag{11}$$

where $y_i = 0$ if the segmentation results is drawn from the unlabeled images, while $y_i = 1$ if the segmentation results from the labeled images. λ_{D_2} and λ'_{D_2} are two weights for the cross entropy loss of the auxiliary discriminator network. λ_{D_2}, λ'_{D_2} is set as 0.01 and 0.05 separately in our implements.

4. Experiments

4.1. Datasets and Evaluation Metrics

Datasets. We evaluate the proposed method on the two famous polyp datasets, including Kvasir-SEG [14] and CVC-Clinic DB [2]. Kvasir-SEG contains 1000 polyp images with ground truth manually segmented by experienced gastroenterologists. CVC-Clinic DB also contains 612 polyp images with corresponding annotations. Each image contains different amount of polyps with different size and shape. In our experiments, we also performed 3 kinds of data augmentations to avoid overfitting, including horizontal flipping, vertical flipping, and random rotation between [-30, 30] degrees. In our ablation studies, we randomly split the dataset into 30% / 50% / 10% / 10% for supervised training, semi-supervised training, validation and testing respectively.

Metrics. We employ the three most commonly used metrics for medical image segmentation tasks to evaluate our method, including Dice coefficient, IoU and MAE. MAE can accurately reflect the actual prediction error of the networks based on mean absolute deviation. Dice coefficient calculates the similarity of the input and target, while IoU calculates the intersection area, which can reflect the authenticity of segmentation predictions to a certain degree.

4.2. Implementation Details

We implemented our proposed model using PyTorch framework on two NVDIA GeForce RTX 2080TI with 24 GB memory. We first trained our model with labeled images for 100 epochs, and then performed the primary adversarial training with both labeled and unlabeled images. After training for 200 epochs, the confidence maps from unlabeled images were converted to supervisory signals and auxiliary adversarial learning was used to train our model for 300 epochs to fine tune collaborative segmentation networks. The batch size is set to 8 and all the inputs were uniformly resized to 384×384 during both supervised training and semi-supervised training. For training the segmentation networks, we employed the Adam optimizer to reduce the overall parameters with the learning rate of $lr = 7e - 4$. Similarly, for training the discriminator networks, we also adopt Adam optimizer with a learning rate $lr = 3e - 4$. The learning rate of segmentation networks and discriminator networks is annealed following the polynomial decay with power of 0.9 [5].

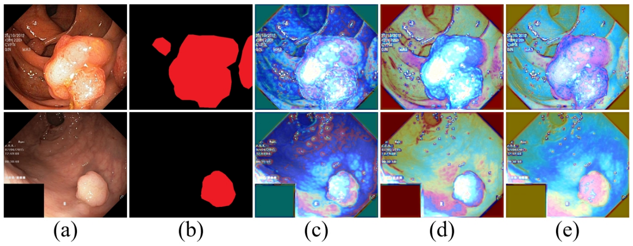

Figure 5. Visual comparison of feature maps extracted with different methods in ablation studies. (a) Input image. (b) Ground truth. (c) Baseline. (d) Baseline+FEM. (e) Baseline+DEM.

Model	Dice	IoU	MAE
Baseline	0.7593	0.6795	0.0824
Baseline+FEM	0.7757	0.6922	0.0771
Baseline+DEM	0.7886	0.7040	0.0725
Baseline+FEM+DEM	0.7965	0.7122	0.0681
Baseline+FEM+DEM+AAL(Ours)	**0.8095**	**0.7163**	**0.0658**

Table 1. Statistical comparison of different models in ablation studies on the Kvasir-SEG dataset.

4.3. Ablation Studies

We employed Hung's adversarial learning method [13] with single segmentation network (U-Net [25]) as the Baseline for our ablation study. By adding FEM and DEM to the baseline respectively, we can obtain two competitors (Baseline+FEM and Baseline+DEM) and observe the impact of the proposed feature enhancement modules. By simultaneously equipping FEM and DEM networks to the baseline, we can obtain another competitor (Baseline+FEM+DEM) to evaluate the effectiveness of collaborative segmentation networks. Finally, our method is implemented by adding AAL to Baseline+FEM+DEM method, where we can further justify the advantages of auxiliary adversarial learning.

To visualize the impact of FEM and DEM on the feature extractions, we apply the Baseline, Baseline+FEM and Baseline+DEM methods to the validation dataset and illustrate their feature maps. Typical feature maps extracted with different competitors are shown in Figure 5. To demonstrate the power of AAL in improving the segmentation accuracy, we also compare the segmentation results between the models with and without AAL. Typical challenging cases are as shown in Figure 6. Based on two collaborative segmentation networks, our method can extract both focused and dispersive features, and obtain much better segmentation accuracy than the Baseline even without using AAL.

In addition, we also performed a statistical comparison among different competitors by collecting the average Dice, IoU and MAE over the validation set. As shown in Table 1, we can see that FEM and DEM improve the performance of segmentation of Baseline with the Mean Dice by 1.64% and 2.93% respectively. Besides, our method obtains a much better accuracy than the competitors only using one segmentation network, implying that our collaborative segmentation networks with a consistency constraint extract more location and boundary information of polyp to improve the prediction of segmentation networks. By comparing the results in the bottom two rows in Table 1, we can also observe the advantage of AAL in improving the segmentation accuracy.

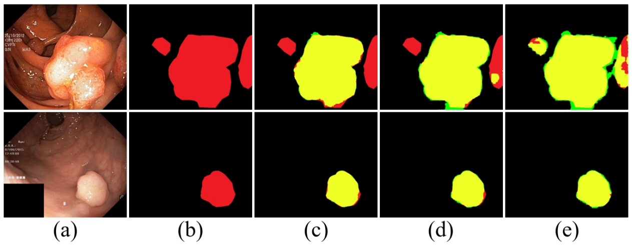

Figure 6. Visual comparisons of segmentation results with and without AAL in ablation studies. (a) Input image. (b) Ground truth. (c) Baseline. (d) Our method without AAL. (e) Our method with AAL. Red, green and yellow regions represent the ground truth, prediction and their overlapping regions respectively.

Model	Year	Labeled Amount	Dice	IoU	MAE
ResU-Net	2018	100%	0.7647	0.6738	0.0886
U-Net++	2018	100%	0.8045	0.7086	0.0679
CE-Net	2019	100%	0.8615	0.7909	0.0493
CPF-Net	2020	100%	0.8415	0.7623	0.0584
PraNet	2020	100%	0.8725	0.7978	0.0458
HarDNet-MSEG	2021	100%	0.8750	0.7908	0.0505
Hung's	2018	15%	0.6839	0.5696	0.1027
Hung's	2018	30%	0.7593	0.6795	0.0824
Ours	2021	**15%**	**0.7676**	**0.6723**	**0.0816**
Ours	2021	**30%**	**0.8095**	**0.7163**	**0.0658**

Table 2. Statistical comparison of different methods on the Kvasir-SEG dataset.

4.4. Comparison with State-of-the-art Methods

To further validate the superiority of our proposed method, we compared it with seven state-of-the-arts methods on Kvasir-SEG and CVC-Clinic DB, including six fully supervised networks (ResU-Net[37], U-Net++[40], CE-Net[10], CPF-Net[8], PraNet[7] and HarDNet-MSEG[12]) and one semi-supervised network (Hung's method [13]). To ensure the fairness of comparison, we implemented all competitors using the same PyTorch framework on two NVDIA GeForce RTX 2080TI with 24 GB memory, with the same augmentation operations. In our experiments, except for six fully supervised competitors, the other two semi-supervised networks are trained with either 15% labeled data or 30% labeled data for comparisons.

Compared with fully supervised competitors, we can observe that our method outperforms the ResU-Net and U-

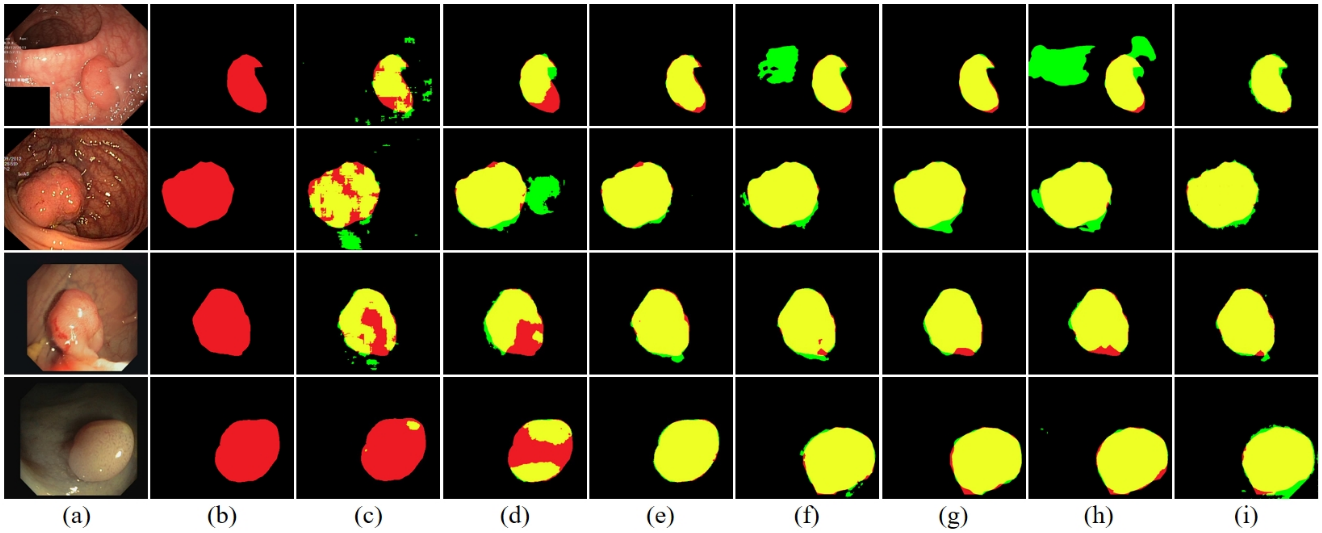

Figure 7. Comparisons with different State-of-the-art methods on the Kvasir-SEG and CVC-Clinic DB. (a) Input image. (b) Ground Truth. (c) ResU-Net. (d) U-Net++. (e) CE-Net. (f) CPF-Net. (g) PraNet. (h) HarDNet-MSEG. (i) Ours. Note that, our semi-supervised polyp segmentation method is trained with only 30% of labeled data, which still achieves comparable performance with other six fully supervised networks. Red, green and yellow regions represent the ground truth, prediction and their overlapping regions respectively.

Model	**Year**	**Labeled Amount**	**Dice**	**IoU**	**MAE**
ResU-Net	2018	100%	0.7762	0.6845	0.0384
U-Net++	2018	100%	0.8769	0.7988	0.0232
CE-Net	2019	100%	0.9314	0.8805	0.0108
CPF-Net	2020	100%	0.9058	0.8417	0.0169
PraNet	2020	100%	0.9520	0.9100	0.0075
HarDNet-MSEG	2021	100%	0.9405	0.8967	0.0086
Hung's	2018	15%	0.5688	0.4761	0.0565
Hung's	2018	30%	0.7609	0.6802	0.0348
Ours	2021	**15%**	**0.8218**	**0.7498**	**0.0301**
Ours	2021	**30%**	**0.8929**	**0.8257**	**0.0226**

Table 3. Statistical comparison of different methods on the CVC-Clinic DB.

Net++ only relying on 30% labeled data. Statistical comparisons for different competitors according to the three metrics are as shown in Table 2 and Table 3. Visual comparisons with different fully supervised competitors on typical challenging cases are as shown in Figure 7.

Compared with the semi-supervised competitor only using single segmentation network in the generator, our method also generally outperforms Hung's method in both visual and statistical comparisons (Figure 6, Table 2 and Table 3), which clearly demonstrates the superiority of our collaborative segmentation networks and auxiliary adversarial learning in the proposed semi-supervised framework.

On the other hand, our method also has some limitations. As shown in Figure 8, our method still cannot handle well the extreme challenging cases where polyps are extremely big and the color contrast around polyps is extremely low. One potential future direction is to extend our collaborative and adversarial learning framework to solve video segmentation tasks with similar limited labeled data.

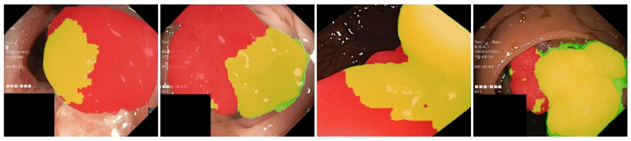

Figure 8. Failure cases. Red, green and yellow maps denote the ground truth, our prediction and overlapping regions respectively. Our model is trained with 30% labeled data on the Kvasir-SEG dataset.

5. Conclusion

In this work, we present a novel semi-supervised adversarial learning method for polyp segmentation. Specifically, we introduce collaborative segmentation networks with focused extraction module (FEM) and dispersive extraction module (DEM) based on an adversarial training architecture. In addition, we propose an auxiliary adversarial learning method for eliminating the impact of labeled and unlabeled data amount imbalance. Both collaborative and adversarial learning methods are simultaneously applied to fully utilize the abundant unlabeled data to enhance the segmentation performance. The experimental results on Kvasir-SEG and CVC-Clinic DB have shown the effectiveness of our proposed model. In the future, more collaborative and adversarial learning mechanisms will be explored for better leveraging of the unlabeled data.

References

[1] Bhakti Baheti, Shubham Innani, Suhas Gajre, and Sanjay Talbar. Eff-unet: A novel architecture for semantic segmentation in unstructured environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 358–359, 2020.

[2] 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.

[3] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[4] Ping Chao, Chao-Yang Kao, Yu-Shan Ruan, Chien-Hsiang Huang, and Youn-Long Lin. Hardnet: A low memory traffic network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3552–3561, 2019.

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017.

[6] Rahul Duggal, Anubha Gupta, Ritu Gupta, Manya Wadhwa, and Chirag Ahuja. Overlapping cell nuclei segmentation in microscopic images using deep belief networks. In *Proceedings of the Tenth Indian Conference on Computer Vision, Graphics and Image Processing*, pages 1–8, 2016.

[7] 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.

[8] Shuanglang Feng, Heming Zhao, Fei Shi, Xuena Cheng, Meng Wang, Yuhui Ma, Dehui Xiang, Weifang Zhu, and Xinjian Chen. Cpfnet: Context pyramid fusion network for medical image segmentation. *IEEE Transactions on Medical Imaging*, 2020.

[9] Geoffrey French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. In *British Machine Vision Conference*, number 31, 2020.

[10] Zaiwang Gu, Jun Cheng, Huazhu Fu, Kang Zhou, Huaying Hao, Yitian Zhao, Tianyang Zhang, Shenghua Gao, and Jiang Liu. Ce-net: Context encoder network for 2d medical image segmentation. *IEEE Transactions on Medical Imaging*, 38(10):2281–2292, 2019.

[11] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[12] Chien-Hsiang Huang, Hung-Yu Wu, and Youn-Long Lin. Hardnet-mseg: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps. *ArXiv Preprint ArXiv:2101.07172*, 2021.

[13] Wei Chih Hung, Yi Hsuan Tsai, Yan Ting Liou, Yen-Yu Lin, and Ming Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. In *29th British Machine Vision Conference, BMVC 2018*, 2018.

[14] 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 *International Conference on Multimedia Modeling*, pages 451–462. Springer, 2020.

[15] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *ArXiv Preprint ArXiv:1610.02242*, 2016.

[16] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, volume 3, 2013.

[17] Xiaomeng Li, Lequan Yu, Hao Chen, Chi-Wing Fu, and Pheng-Ann Heng. Semi-supervised skin lesion segmentation via transformation consistent self-ensembling model. In *29th British Machine Vision Conference, BMVC 2018*, 2018.

[18] Zhuoying Li, Junquan Pan, Huisi Wu, Zhenkun Wen, and Jing Qin. Memory-efficient automatic kidney and tumor segmentation based on non-local context guided 3d u-net. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 197–206. Springer, 2020.

[19] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *ArXiv Preprint ArXiv:1312.4400*, 2013.

[20] Songtao Liu, Di Huang, et al. Receptive field block net for accurate and fast object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 385–400, 2018.

[21] Alexander V Mamonov, Isabel N Figueiredo, Pedro N Figueiredo, and Yen-Hsi Richard Tsai. Automated polyp detection in colon capsule endoscopy. *IEEE Transactions on Medical Imaging*, 33(7):1488–1502, 2014.

[22] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2018.

[23] Arnab Kumar Mondal, Jose Dolz, and Christian Desrosiers. Few-shot 3d multi-modal medical image segmentation using generative adversarial learning. *arXiv preprint arXiv:1810.12241*, 2018.

[24] 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.

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 234–241. Springer, 2015.

[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(2):283–293, 2014.

[27] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, pages 1195–1204, 2017.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[29] Shuo Wang, Mu Zhou, Zaiyi Liu, Zhenyu Liu, Dongsheng Gu, Yali Zang, Di Dong, Olivier Gevaert, and Jie Tian. Central focused convolutional neural networks: Developing a data-driven model for lung nodule segmentation. *Medical Image Analysis*, 40:172–183, 2017.

[30] Wei Wang, Jiafu Zhong, Huisi Wu, Zhenkun Wen, and Jing Qin. Rvseg-net: An efficient feature pyramid cascade network for retinal vessel segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 796–805. Springer, 2020.

[31] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[32] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[33] Huisi Wu, Junquan Pan, Zhuoying Li, Zhenkun Wen, and Jing Qin. Automated skin lesion segmentation via an adaptive dual attention module. *IEEE Transactions on Medical Imaging*, 40(1):357–370, 2020.

[34] Huisi Wu, Wei Wang, Jiafu Zhong, Baiying Lei, Zhenkun Wen, and Jing Qin. Scs-net: A scale and context sensitive network for retinal vessel segmentation. *Medical Image Analysis*, page 102025, 2021.

[35] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3907–3916, 2019.

[36] Yading Yuan, Ming Chao, and Yeh-Chi Lo. Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance. *IEEE Transactions on Medical Imaging*, 36(9):1876–1886, 2017.

[37] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, 2018.

[38] Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision*, 129(4):1106–1120, 2021.

[39] Jiafu Zhong, Wei Wang, Huisi Wu, Zhenkun Wen, and Jing Qin. Polypseg: An efficient context-aware network for polyp segmentation from colonoscopy videos. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 285–294. Springer, 2020.

[40] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11. Springer, 2018.