Semi-Supervised Medical Image Segmentation Using Adversarial Consistency Learning and Dynamic Convolution Network

Tao Lei, *Senior Member, IEEE*, Dong Zhang, Xiaogang Du, Xuan Wang, Yong Wan, and Asoke K. Nandi, *Life Fellow, IEEE*

***Abstract*— Popular semi-supervised medical image segmentation networks often suffer from error supervision from unlabeled data since they usually use consistency learning under different data perturbations to regularize model training. These networks ignore the relationship between labeled and unlabeled data, and only compute single pixel-level consistency leading to uncertain prediction results. Besides, these networks often require a large number of parameters since their backbone networks are designed depending on supervised image segmentation tasks. Moreover, these networks often face a high over-fitting risk since a small number of training samples are popular for semi-supervised image segmentation. To address the above problems, in this paper, we propose a novel adversarial self-ensembling network using dynamic convolution (ASE-Net) for semi-supervised medical image segmentation. First, we use an adversarial consistency training strategy (ACTS) that employs two discriminators based on consistency learning to obtain prior relationships between labeled and unlabeled data. The ACTS can simultaneously compute pixel-level and image-level consistency of unlabeled data under different data perturbations to improve the prediction quality of labels. Second, we design a dynamic convolution-based bidirectional attention component (DyBAC) that can be embedded in any segmentation network, aiming at adaptively adjusting the weights of ASE-Net based on the structural information of input samples. This component effectively improves the feature representation ability of ASE-Net and reduces the overfitting risk of the network. The proposed ASE-Net has been extensively tested on three publicly available datasets, and experiments indicate that ASE-Net is superior to state-of-the-art networks, and reduces computational costs and memory overhead. The code is available at: https://github.com/SUST-reynole/ASE-Nethttps://github.com/SUST-reynole/ASE-Net.**

***Index Terms*— Semi-supervised learning, medical image segmentation, dynamic convolution, adversarial learning.**

I. Introduction

MEDICAL image segmentation plays a significant role in computer-aided diagnosis and treatment research since it can extract important organs or lesions in abnormal images. In recent years, many supervised-learning based encoder-decoder networks for medical image segmentation have achieved remarkable results such as U-Net [1], U-Net++ [2], H-DenseUNet [3], etc. However, the success of these techniques relies heavily on a large amount of pixel-level labeled data but it is usually very expensive to annotate medical images in practice. One of the reasons is that medical images usually show poor visual effects due to low contrast and noise interference. Moreover, medical image annotation requires much more professional knowledge than natural images. Therefore, it is almost impossible to build a large number of medical image datasets with high-precision labels. Compared to supervised learning, semi-supervised learning is a new learning paradigm to solve the problem of incomplete supervision of data in weakly supervised learning [4]. It mainly uses a small amount of labeled data and a large amount of unlabeled data to achieve joint training. Obviously, semi-supervised learning is of great importance and more in line with the requirement of actual clinical scenes than supervised learning in medical image segmentation.

The main semi-supervised medical image segmentation methods can be roughly classified as consistency

Manuscript received 23 September 2022; revised 21 November 2022; accepted 26 November 2022. Date of publication 30 November 2022; date of current version 2 May 2023. This work was supported in part by the National Natural Science Foundation of China under Program 62271296, Program 61871259, and Program 61861024; in part by the Natural Science Basic Research Program of Shaanxi under Program 2021JC-47; in part by the Key Research and Development Program of Shaanxi under Program 2022GY-436 and Program 2021ZDLGY08-07; and in part by the Shaanxi Joint Laboratory of Artificial Intelligence under Grant 2020SS-03. *(Corresponding author: Xuan Wang.)*

Tao Lei is with the Shaanxi Joint Laboratory of Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710026, China, and also with the Department of Geriatric Surgery, First Affiliated Hospital, Xi'an Jiaotong University, Xi'an 710021, China (e-mail: leitao@sust.edu.cn).

Dong Zhang and Xiaogang Du are with the Shaanxi Joint Laboratory of Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710021, China (e-mail: 201611021@sust.edu.cn).

Xuan Wang is with the Department of Electrical and Computer Engineering, University of Wisconsin–Madison, Madison, WI 53706 USA (e-mail: xwang554@wisc.edu).

Yong Wan is with the Department of Geriatric Surgery, First Affiliated Hospital, Xi'an Jiaotong University, Xi'an 710061, China (e-mail: docwanyong@xjtu.edu.cn).

Asoke K. Nandi is with the Department of Electronic and Electrical Engineering, Brunel University London, UB8 3PH Uxbridge, U.K., and also with the College of Electronic and Information Engineering, Tongji University, Shanghai 201804, China (e-mail: asoke.nandi@brunel.ac.uk).

Digital Object Identifier 10.1109/TMI.2022.3225687

learning [5], [6], [7], [8], [9], [10] adversarial learning [11], [12], [13], self-training [14], [15], [16], contrastive learning [17], [18], [19], and collaborative training [20], [21]. In this paper, we focus on consistency learning and adversarial learning. Consistency learning usually employs consistency regularization with different perturbations to train a network. One of the most representative methods is self-ensembling Mean Teacher (MT) [5], which utilizes perturbation-based consistency loss between the self-ensembling teacher model and the student model on unlabeled data, along with the supervised loss on labeled data. Depending on MT, subsequently improved methods focus on choosing different data perturbations as well as feature perturbations to achieve performance gains. Precisely, the quality of a segmentation network in generating consistent pseudo labels determines the knowledge mining ability of the network for unlabeled data. For adversarial learning, the generative adversarial networks (GAN) [11], [12], [13] for medical image segmentation mainly involve two subnetworks, namely a discriminator and a generator. The discriminator aims to identify whether the input samples are from the ground truth or the generator. The generator aims the discriminator not to be able to distinguish between the ground truth and the output from the segmentation network. Once the discriminator cannot determine where the input is from, the generative samples are considered to be close enough to the ground truth. The two networks update alternately and promote each other.

Albeit those methods mentioned above have achieved great success, they still face the following challenges. First, in consistency learning, a typical Mean Teacher method acquires consistency loss only depending on different data perturbations, which does not effectively exploit the prior relationship between unlabeled and labeled data, resulting in slow feature learning on unlabeled data and weak model generalization ability. Second, in adversarial learning, popular methods only employ a single segmentation network and a single discriminator network to mine the potential knowledge from unlabeled data. Unfortunately, the two networks can often mislead each other, leading to the problem of error accumulation during the training process. Third, it is usually inappropriate for semi-supervised methods to use directly a segmentation network with fixed parameters from supervised learning. On the one hand, the segmentation network with fixed parameters is better to fit labeled data but has poor feature representation for unlabeled data. On the other hand, different samples share the same model weights in the networks with fixed parameters, which easily causes network overfitting for small labeled datasets, leading to poor quality of generating pseudo labels for unlabeled data.

In order to solve the above problems, in this paper, we propose a novel adversarial self-ensembling network using dynamic convolution (ASE-Net) for semi-supervised medical image segmentation. ASE-Net effectively exploits the prior relationship between unlabeled and labeled data as well as pixel-level and image-level consistency by using consistency learning and adding two discriminator networks to an MT framework. In addition, we propose a dynamic convolution-based bidirectional attention component that can be easily embedded in a segmentation network to avoid the overfitting problem. The main contributions of this paper are summarized as follows:

(1) We propose an adversarial consistency training strategy (ACTS) using double discriminators. The first discriminator learns the prior relationship between labeled and unlabeled data and the second one learns the image-level consistency of a segmentation network on the same data with different data perturbations. Both discriminators aim to improve the knowledge transfer ability of the segmentation network from labeled data to unlabeled data.

(2) We design a dynamic convolution-based bidirectional attention component (DyBAC), which can sufficiently mine the prior knowledge of samples and dynamically adjust the parameters of convolutional kernels depending on different input samples. The DyBAC can effectively improve the feature representation ability of our proposed network and avoid network overfitting.

(3) We extensively validate the performance of the proposed method in three challenging medical image segmentation tasks, and the experiments demonstrate that the proposed network is very competitive compared to the state-of-the-art methods. It is worth mentioning that our proposed network is a lightweight network that requires fewer parameters and has a faster inference speed than comparative networks.

II. Related Work

A. Semi-Supervised Medical Image Segmentation

To solve the problem of lacking a large number of labeled data, researchers proposed many semi-supervised learning methods for medical image segmentation. Since traditional semi-supervised medical image segmentation methods usually employ manually designed shallow features with limited representation ability, they cannot provide good segmentation results for medical images with low contrast and serious noise interference. Compared with those methods mentioned above, deep learning-based semi-supervised methods can provide excellent segmentation results since they have powerful feature representation and modeling abilities [22]. Currently, popular semi-supervised medical image segmentation methods often use a regular encoder-decoder segmentation network as the backbone [1], [2], [3], [23], [24], [25], [26]. Aiming to utilize unlabeled data better, more methods focus on improvements in learning strategies. In this paper, we focus on employing consistency learning [27] and adversarial learning [28] to improve network performance.

For consistency learning, the state-of-the-art technique is Mean Teacher (MT) [5], [29], which performs consistency learning under different data perturbations by accumulating the weights of the student model. Specifically, the MT is first conducted in the way of supervised learning on labeled data. After that, the teacher model of MT is used to provide pseudo labels for unlabeled data, and the prediction consistency of the teacher and student model for unlabeled data is maintained through different regularization methods. Finally, the student model is updated through feedback on supervision and consistency loss. Among them, the teacher model is the exponential moving average (EMA) of the student model weights. This operation enables the teacher model to accumulate continuously the historical prediction information of unlabeled data. The subsequent improvements [6], [7], [8], [9], [10], [30] use different consistency regularization strategies to improve the prediction quality of unlabeled data and avoid network overfitting.

For example, Li et al. [6] proposed a transformation-consistent self-ensembling model (TCSM_v2) to utilize effectively unlabeled data by introducing the regularization strategy of data transformation consistency. Chen et al. [8] proposed a cross pseudo supervision (CPS) method based on network perturbation to encourage the high consistency between prediction results from two perturbed networks. However, calculating the consistency between two predictions of unlabeled data may cause some unreliable guidance and thus make the training unstable. In order to solve this problem, Yu et al. [9] proposed an uncertainty-aware framework based on the Mean Teacher structure (UA-MT), which makes the student model gradually learn more reliable targets according to uncertainty estimates after multiple forward propagations. In order to reduce the time and memory overhead, Wu et al. [31] proposed a mutual consistency network (MC-Net). The network includes two decoders and expresses the difference between the two predictions as model uncertainty information to regularize model training, so as to improve the quality of the pseudo labels. Liu et al. [32] proposed a perturbed and strict mean teacher (PS-MT) framework to improve the segmentation accuracy by adding an auxiliary teacher model, designing different loss functions, and using different data perturbation methods. In addition, Luo et al. [33] constructed a dual-task consistency (DTC) regularization method by jointly predicting the pixel-wise segmentation map and the geometry-aware level set representation of targets. DTC focuses on task-level consistency rather than data-level consistency.

Adversarial learning [11], [12], [28], [34], [35], [36], [37] is a popular strategy for improving model robustness by effectively mining potential knowledge from unlabeled data. For example, Zhang et al. [12] proposed a deep adversarial network (DAN) to improve the prediction quality of unlabeled data. However, popular semi-supervised adversarial learning methods, [11], [12], [28] only contain a single generator and a single discriminator, which may lead to low segmentation accuracy due to over-reliance on the result of a single network. Therefore, the knowledge obtained from a model with low segmentation accuracy may produce misguidance during the learning process on unlabeled data. To go a step further, some improved methods [34], [35], [36], [37] give consideration to both consistency learning and adversarial learning to improve the learning ability of models.

B. Dynamic Neural Network

Traditional deep learning networks perform inference in a static manner, in other words, the network parameters are fixed after training. For different input samples, these static networks output different predictions using the same parameters combined with different inputs, which leads to poor predictions for some complex input samples due to weak feature representation ability. Contrary to static networks, dynamic neural network [38] means that the network structure [39], parameters [40], and features maps [41], [42] change according to different inputs in the inference stage. For example, in terms of dynamic feature networks based on attention mechanisms, Gu et al. [42] demonstrated in detail the effectiveness of attention mechanisms and achieved better results in medical image segmentation. Therefore, the dynamic neural network is more compatible with the human visual system. In this paper, we focus on the study of convolution neural networks with dynamic parameters.

The conditionally parameterized convolutions proposed by Yang et al. [41] and the dynamic convolutional neural network (CNN) proposed by Chen et al. [40] mainly dynamically aggregate multiple groups of weights from different convolutional kernels according to input images to achieve dynamic convolution. However, both of them lead to a dramatic increase in the number of parameters and only use the prior knowledge of channels without considering spatial information of feature maps. To solve the problem, Involution [43] and Decoupled Dynamic Filter Networks (DDF) [44] propose the idea of spatial specificity, which makes the values of convolutional kernel parameters vary with the spatial location in a feature map. Involution and DDF skillfully use the spatial prior knowledge of samples to extract the spatial structure information of images and therefore achieve good results. In contrast to the above methods, Li et al. [45] introduced an omni-dimensional dynamic convolution via a parallel strategy to learn more flexible attention to improve the network performance. In general, the dynamic convolution applies soft attention to convolution kernels by adjusting network parameter values depending on different inputs. Thus, dynamic CNNs can effectively exploit the prior knowledge of samples to improve feature representation.

Different from the above methods, first, considering the MT framework, we extend an adversarial consistency training strategy to a semi-supervised learning framework (ACTS), which makes better use of the essential relationship between unlabeled and labeled data. Second, we propose a dynamic convolution-based bidirectional attention component (DyBAC), which aims to reduce the overfitting risk of the network and to reduce the memory overhead while maintaining the segmentation accuracy.

III. Method

In this paper, we propose an adversarial self-ensembling network (ASE-Net) for semi-supervised medical image segmentation. As shown in Fig. 1, our ASE-Net consists of segmentation networks and discriminator networks. The segmentation networks consist of a student model and a teacher model. The student model has the same structure as the teacher model and both of them are based on the encoder-decoder structure; the difference is that the former is trained by the loss function while the latter is the exponential moving average (EMA) of the student model weights. The discriminator networks consist of convolutional layers, the proposed DyBAC, and the global average pooling, whose specific structure of our ASE-Net is shown in Fig. 1.

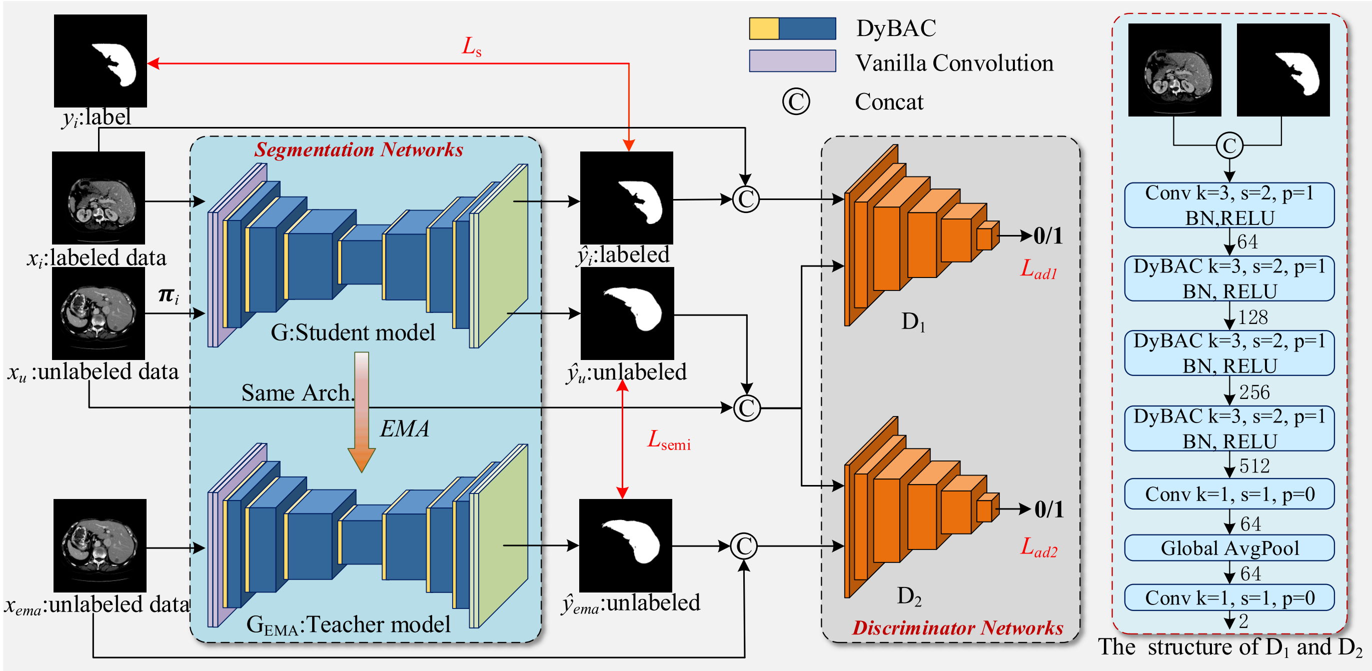


Fig. 1. The framework of the proposed ASE-Net. The ASE-Net consists of two main parts: the segmentation networks (left) and the discriminator networks (right). The segmentation network is based on the encoder-decoder architecture. The right figure shows the detailed structure of the discriminative network, where *k*, *s*, and *p* represent the kernel size, the stride, and the padding of convolutional kernels, respectively. The discriminators are unnecessary in the inference stage.

In our ASE-Net, we propose an adversarial consistency training strategy (ACTS) based on the MT framework to mine the prior knowledge from unlabeled data. We use two discriminators of the same structure for different purposes. The first discriminator learns the predicted quality consistency of the segmentation network for unlabeled data as well as the labeled data. The second discriminator learns the prediction consistency of teacher and student networks using the same inputs but under different perturbations. It is worth mentioning that the input of our discriminator network is a concatenation of the segmentation result after softmax and the original image, rather than just the segmentation result. In this way, the quality of the segmentation results can be further evaluated by using the original image as a benchmark to discriminate the matching relationship between the segmentation results and the benchmark. In terms of network structure, we apply DyBAC to replace all convolutional layers except the first layer in segmentation networks and discriminator networks. The DyBAC can improve the feature representation ability of the network and reduces the risk of overfitting. In addition, the segmentation networks and the discriminators are trained alternately, and the discriminators are unnecessary in the inference stage, which avoids additional computational costs.

A. Adversarial Consistency Learning

Although consistency learning and adversarial learning are useful for semi-supervised image segmentation tasks, they still have some limitations. First, regular semi-supervised image segmentation networks usually use consistency strategies under different perturbations to regularize the training of the model. These networks often ignore the prior relationship between labeled and unlabeled data. Besides, they only calculate pixel-level consistency for unlabeled data that may result in uncertain prediction results. Second, the adversarial learning-based approaches rely excessively on a single segmentation network and a single discriminative network, easily causing the problem of misguidance.

To address these problems, we propose a novel training strategy. As shown in Fig.1, we add two discriminators, and the two discriminators have the same structure but different functions. The discriminator D_1 learns the difference between the output quality of labeled data and unlabeled data. The discriminator D_2 learns the difference between perturbed data and unperturbed data in unlabeled data. Finally, through the supervision loss L_s, the consistency loss L_{semi} and the adversarial loss (L_{ad1}, L_{ad2}) encourage the student network to generate high-quality segmentation results on unlabeled data. Actually, the roles of D_2 and L_{semi}, are complementary. The consistency loss L_{semi} is a pixel-level consistency among individual samples, which pays more attention to feature map details. Our D_2 is mainly used for the image-level consistency among perturbed and unperturbed data, which pays more attention to feature map global information.

Specifically, we achieve adversarial consistency learning through alternate training. First, we input medical images into the segmentation networks to obtain the segmentation prediction maps. Then, we concatenate the output feature maps and the corresponding original images into the discriminator networks. The discriminators mainly evaluate the quality of the segmentation results, where 0 means the quality of the segmentation result is poor and 1 means good. During the training process of the segmentation network G, we encourage the segmentation network to generate high-quality segmentation results for unlabeled data x_u, aiming to ensure the results are as close to 1 as possible. During the training process of the discriminative networks, we encourage the discriminative networks to discriminate against different inputs as much as possible. Consequently, the optimization objective function of the student network G and the two discriminative networks D_1, D_2 is defined as:

$$\min_{G} \max_{D_1, D_2} (L_G(\theta) + L_{D_1}(\theta) + L_{D_2}(\theta)), \tag{1}$$

where θ represents the parameter to be optimized. Exactly, the segmentation network and the discriminator networks are trained alternately. The objective function of the segmentation network, $L_G(\theta)$ is defined as:

$$\begin{aligned} L_G(\theta) = {} & L_s(\hat{y}_i, y_i) + \lambda(L_{semi}(\hat{y}_u, \hat{y}_{ema}) \\ & + L_{ad1}(D_1(x_u, \hat{y}_u), 1) + L_{ad2}(D_2(x_u, \hat{y}_u), 1)), \end{aligned} \tag{2}$$

where $L_s(\cdot)$ represents supervision loss, $L_s(\cdot) = L_{ce}(\cdot) + L_{dice}(\cdot)$, $L_{ce}(\cdot)$ is the cross-entropy loss and L_{dice} is Dice loss. $L_{semi}(\cdot)$ is MSEloss, and both $L_{ad1}(\cdot)$ and $L_{ad2}(\cdot)$ are binary-class cross-entropy loss. y_i is the label corresponding to the input x_i, and x_u is unlabeled input data with data perturbations via π_i. π_i denotes random Gaussian noise. $\hat{y}_i$ and $\hat{y}_u$ are segmentation results on labeled and unlabeled data, respectively. $\hat{y}_{ema}$ is the prediction result of the teacher network and λ is the weighting coefficient. According to [6], λ is a Gaussian ramp-up curve, $\lambda = \delta e^{(-5(1-I)^2})$, and I denotes the number of epochs.

At the early stage of the training network, the value of λ is very small and the update of the network mainly relies on supervision loss. Therefore, the network is trained mainly depending on labeled data in the early stage of the training network. As the training proceeds, the value of λ continues to increase, and the network can obtain a reliable segmentation result and generate targets for unlabeled data. This is because other loss functions are in effect. Next, the discriminator networks try their best to distinguish the output of the segmentation network. The objective functions of the discriminators D_1 and D_2 are defined as:

$$L_{D_1}(\theta) = L_{ad1}(D_1(x_i, \hat{y}_i), 1) + L_{ad1}(D_1(x_u, \hat{y}_u), 0), \tag{3}$$

$$L_{D_2}(\theta) = L_{ad2}(D_2(x_{ema}, \hat{y}_{ema}), 1) + L_{ad2}(D_2(x_u, \hat{y}_u), 0), \tag{4}$$

where x_i and x_{ema} represent labeled data and unlabeled inputs, respectively.

The parameters of the teacher model are the EMA accumulation of the parameters of the student model. The teacher model retains the historical information of the student model and can generate higher quality targets for unlabeled data. Its effectiveness has been proved in [5] and [6], and the parameters θ'_t of the current teacher model are defined as:

$$\theta'_t = \alpha\theta'_{t-1} + (1 - \alpha)\theta_t, \tag{5}$$

where the parameters θ'_{t-1} is the historical accumulation of the teacher model. θ_t is the weight of the student model. α is a hyperparameter of the smoothing coefficient, and α determines the dependency relationship between the teacher model and the student model. According to [5], [6], and [9] and experimental experience, when the value of α is 0.999, the performance of networks is the best.

In conclusion, the segmentation networks and the discriminator networks play games against each other. When the discriminator networks cannot distinguish the segmentation result and ground truth, the segmentation networks have high segmentation quality for labeled data, unlabeled data, and data under different perturbations. This adversarial learning approach can effectively utilize unlabeled data to improve the quality of predicted pseudo labels.

B. Dynamic Convolution-Based Bidirectional Attention Component

Overfitting is a common problem in segmentation tasks. To overcome this problem, many segmentation networks based on semi-supervised learning employ different consistency regularization strategies, such as data perturbation [5], [6], network parameters perturbation [8], [46], and feature perturbation [10]. However, these perturbation-specific approaches are only valid for specific tasks and it is usually very difficult to choose effectively a uniform perturbation type for different tasks, resulting in an unsatisfied segmentation effect. Moreover, since these semi-supervised methods still use segmentation networks with fixed convolutional kernels, their own structures have potential risks of overfitting. Segmentation networks with fixed parameter values are effective only on the premise that there is a large amount of pixel-labeled data in the task, but in practice, semi-supervised learning only involves a small amount of labeled data and a large amount of unlabeled data. Therefore, a semi-supervised segmentation network based on standard convolution easily suffers from overfitting and has poor feature representation ability.

To solve the above problems, we start from the data itself and construct supervision information according to its structure for unlabeled data. Specifically, we utilize dynamic convolution to adjust adaptively a set of parameters for each sample, which can make better use of the prior knowledge while reducing the overfitting risk and improving the feature representation ability of our network. Furthermore, to overcome the problems of low contrast and blurred edges in medical images, we add spatial attention before using dynamic convolution. As a result, the final values of convolutional kernels are decided by the combination of spatial attention and dynamic convolution. Therefore, the strategy is named dynamic convolution-based bidirectional attention component (DyBAC).

Specifically, the structure of DyBAC is shown in Fig. 2, for a given input $x_{in} \in \mathbb{R}^{C \times H \times W}$, where C represents the number of input channels, and H and W represent the height and width of the input feature maps. To enhance the significance of important spatial positions, the input feature maps are first proceeded by a spatial attention module. The specific operation is shown in Fig. 2 (a). First, a 1 × 1 convolution is used in the input feature maps for dimensionality reduction. Second, the

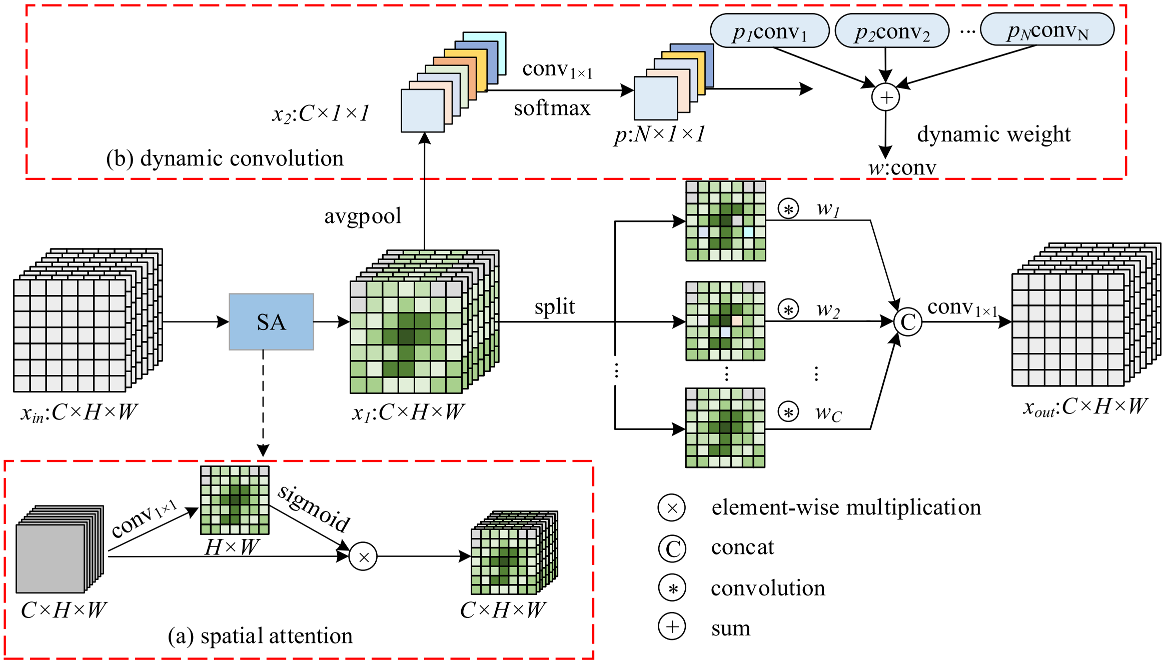


Fig. 2. The structure of DyBAC. (a) Spatial attention, (b) Dynamic convolution. The dynamic convolutional kernels are generated mainly based on the channel and spatial information of samples. For different input samples, the values of convolution kernel parameters change adaptively.

output tensor is normalized by a sigmoid activation function. Finally, the obtained spatial attention weights are multiplied by the input feature maps pixel by pixel to obtain the feature maps $x_1 \in \mathbb{R}^{C\times H\times W}$.

Next, we mainly introduce the generation process of dynamic convolution. Different from the attention mechanism in SE-Net [47], we assign weights to convolutional kernels rather than feature maps. First, through a global average pooling layer, the feature map x_1 is transformed to $x_2 \in \mathbb{R}^{C\times 1\times 1}$, then 1×1 convolution is used to reduce the dimension and we get $p \in \mathbb{R}^{N\times 1\times 1}$ after a softmax activation function, where N is the number of convolutional kernels and it is defined as a hyperparameter in advance. N can be set according to the specific task. In this paper, we empirically set $N = 4$. We multiply the obtained coefficients p to N convolutional kernels respectively and then sum the weights of N convolutional kernels to generate a dynamic convolutional kernel. In this way, we can obtain the most representative convolutional kernel from N convolutional kernels through dynamic aggregation. The weight w of the convolutional kernel is defined as:

$$w = \sum_{i=1}^{N} (p_i \cdot conv_i), \tag{6}$$

where p_i is the i-th coefficient of p, $0 \le p_i \le 1$, $\sum_{i=1}^{N} p_i = 1$, and $conv_i$ is the weight of the i-th convolutional kernel. The number of parameters of a standard dynamic convolution, denoted by Q_s, is defined as:

$$Q_s = C_{in} \times N + N \times C_{in} \times C_{out} \times k \times k, \tag{7}$$

where $k \times k$ is the size of the convolutional kernel, C_{in} and C_{out} indicate the number of channels of the input and output feature maps, respectively. Obviously, the number of parameters is N times more than that of the vanilla convolution.

To reduce the number of parameters, we fully decouple the spatial and channel correlations. Specifically, we define N depthwise convolutions to extract each channel feature and then use pointwise convolution to obtain the information among different channels. We multiply the obtained attention coefficient to the corresponding convolution kernel, and dynamically select a convolutional kernel for the final convolution operation. The number of parameters of our proposed dynamic convolution, denoted by Q_o, is defined as:

$$Q_o = C_{in} \times N + N \times C_{in} \times k \times k + C_{in} \times C_{out}. \tag{8}$$

The ratio r of the number of parameters of our proposed dynamic convolution to the standard convolution is:

$$\begin{aligned} r &= \frac{C_{in} \times N + N \times C_{in} \times k \times k + C_{in} \times C_{out}}{C_{out} \times C_{in} \times k \times k} \\ &= \frac{N + N \times k \times k + C_{out}}{C_{out} \times k \times k} \approx \frac{40 + C_{out}}{9 \times C_{out}} \ll 1. \end{aligned} \tag{9}$$

In practical applications, the convolutional kernel size is usually $k = 3$, the value of C_{out} is larger than 16, and the number of predefined convolutions N is usually 4. Obviously, compared with vanilla convolution and standard dynamic convolution, our DyBAC greatly reduces the number of parameters. Specifically, our operation adaptively adjusts the parameters of the convolutional kernel according to the structure information of each sample, which is different from the vanilla convolution that shares static parameters for all samples.

IV. Experiments

A. Dataset and Pre-Processing

To evaluate our approach, we performed a full evaluation on three different types of medical image datasets, liver Computed Tomography (CT) scans [48], dermoscopy images [49], and 3D left atrium magnetic resonance (MR) image scans [50].

1) Liver Segmentation CT Dataset: In our experiment, we use Liver Tumor Segmentation Challenge (LiTS) [48] as the experimental dataset, which contains 131 labeled CT scans. The size of each image is 512 × 512, and the pixel spacing varied from 0.55 mm to 1 mm. The slice thickness varied from 0.55 mm to 6 mm. To enhance liver contrast and remove interference, we truncate the intensity value of all scans of [−200, 250] Hounsfield Unit (HU). To improve the training efficiency, we resize the images to 256 × 256. In our semi-supervised setting, we randomly select 121 cases as the training set and the remaining 10 cases as the testing set. we perform random data augmentation on the training set, such as flipping, mirroring, and rotating. For better comparison, we randomly select 10% (12 cases) and 20% (24 cases) of the cases in the training set as labeled data, and the rest is used as unlabeled data.

2) Skin Lesion Segmentation Dermoscopy Dataset: The dermoscopy image dataset is from the 2018 International Skin Imaging Collaboration (ISIC) skin lesion segmentation challenge [49]. The training set contains 2,594 images and the validation set contains 100 images. The dataset has different types of skin lesions as well as different resolutions. To improve the computational efficiency of different models, we resize all images to 256 × 192 as in [51]. To perform semi-supervised learning, similarly, we randomly select 10% (259 images) and 20% (519 images) in the training set to be used as labeled data and the rest as unlabeled data, respectively. In the training phase, we perform online random data augmentation.

3) 3D Left Atrium Segmentation MR Dataset: The left atrial (LA) dataset [50] is from the 2018 Left Atrial Segmentation Challenge and consists of 100 3D gadolinium-enhanced MR images with a resolution of 0.625 × 0.625 × 0.625 mm^3. Following [9], [31], and [33], we use 80 scans for training and 20 scans for validation. We adopt a common data preprocessing scheme that randomly crops the left atrial data to the size of 112 × 112 × 80. In this experiment, 10% (8 scans) and 20% (16 scans) are still used as labeled data, and the rest are used as unlabeled data.

B. Experimental Settings and Evaluation Indicators

All the networks in our experiments are implemented on a server with NVIDIA GeForce RTX 3090 24GB, Ubuntu 18.04, and PyTorch 1.7. We choose Adam to optimize the segmentation model. The initial learning rate is 1×10^{-3}. The SGD algorithm with a momentum of 0.9 is used to optimize the discriminator networks. The initial learning rate is 0.01, and the weight decay is 0.0001.

For the liver CT dataset, we use Dice per case score (DI) and average symmetric surface distance (ASD) to evaluate the liver segmentation results based on the 3D volume [48]. For the dermoscopy image dataset, we use Dice coefficient (DI), Jaccard index (JA), Pixelwise Accuracy (AC), Sensitivity (SE), and Specificity (SP) to evaluate the segmentation results according to [6]. For the 3D MR left atrial dataset, we use DI, JA, 95% Hausdorff distance (95HD), and ASD to evaluate the segmentation results. The values of DI, JA, AC, SE, and SP are in the range of 0 to 1. Therefore, better segmentation results imply higher values of DI, JA, AC, SE, and SP as well as lower values of 95HD and ASD. These evaluation indicators are defined as:

$$DI = \frac{2TP}{FP + 2TP + FN}, \tag{10}$$

$$JA = \frac{TP}{TP + FN + FP}, \tag{11}$$

$$AC = \frac{TP + TN}{TP + FP + TN + FN}, \tag{12}$$

$$SE = \frac{TP}{TP + FN}, \tag{13}$$

$$SP = \frac{TN}{TN + FP}, \tag{14}$$

where TP, TN, FP and FN indicate the number of true positives, true negatives, false positives, and false negatives, respectively.

$$HD = max\left\{\underset{s_A\in S(A)}{max}\ d(s_A, S(B)),\ \underset{s_B\in S(B)}{max}\ d(s_B, S(A))\right\}, \tag{15}$$

$$ASD = \frac{\sum\limits_{s_A\in S(A)} d(s_A, S(B)) + \sum\limits_{s_B\in S(B)} d(s_B, S(A))}{|S(A)| + |S(B)|}, \tag{16}$$

where A and B denote the ground truth and the segmentation result, respectively. $S(A)$ and $S(B)$ denote the set of surface voxels corresponding to A and B respectively, and $d(s_B, S(A)) = \underset{s_A\in S(A)}{min} ||s_B - s_A||$ denotes the shortest Euclidean distance of the voxel s_B to the set $S(A)$. Similarly, $d(s_A, S(B)) = \underset{s_B\in S(B)}{min} ||s_A - s_B||$ denotes the shortest Euclidean distance of the voxel s_A to the set $S(B)$. In addition, the 95HD is defined as the $95th$ quantile of Hausdorff distances (HD) instead of the maximum.

C. Ablation Studies

In this paper, we focus on two contributions, ACTS and DyBAC. We use the semi-supervised method MT [5] as the baseline and U-Net [1], U-Net++ [2] and V-Ne [25] as the backbone of the segmentation network respectively. We perform ablation experiments on three datasets including the LiTS [48], the dermoscopy images [49], and the 3D left atrial [50]. Note that the DyBAC is extended to a 3D version of DyBAC when the proposed ASE-Net is used for the 3D left atrial segmentation. To demonstrate the effectiveness of the adversarial consistency learning, we validate two of the discriminators separately.

As shown in Table I, the ablation experiment is performed on the LiTS liver testing set, and the training set is divided into 10% labeled (12 cases) and 90% unlabeled (109 cases). We use U-Net [1] as the backbone network for liver segmentation,

TABLE I
COMPARISON OF ABLATION EXPERIMENTS ON THE LITS-LIVER TESTING SET BY UTILIZING 10% LABELED DATA OF THE TRAINING SET. THE BEST VALUES ARE IN BOLD

Method	Labeled/Unlabeled	MT	D_1	D_2	DyBAC	DI(%)↑
Supervised+U-Net	12/0					88.17
Supervised+U-Net	12/0				✓	89.65
Semi-supervised+U-Net	12/109		✓			92.11
Semi-supervised+U-Net	12/109	✓				92.39
Semi-supervised+U-Net	12/109	✓	✓			93.11
Semi-supervised+U-Net	12/109	✓		✓		93.14
Semi-supervised+U-Net	12/109	✓			✓	93.36
Semi-supervised+U-Net	12/109	✓	✓	✓		93.39
Semi-supervised+U-Net	12/109	✓	✓	✓	✓	**94.12**

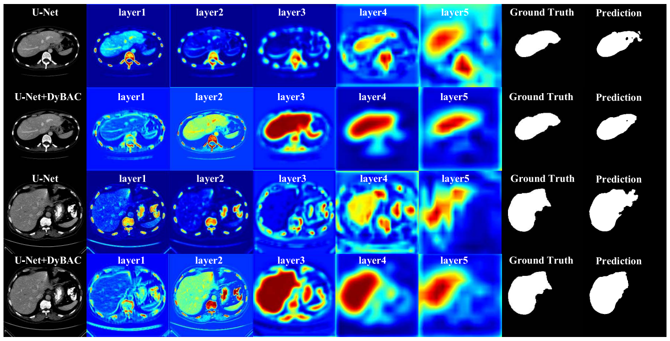
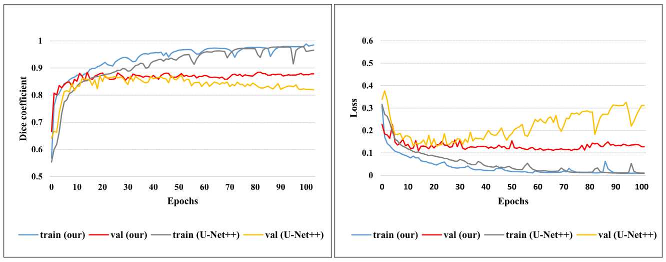


Fig. 3. Visualization of the feature heat maps for each convolutional layer in the encoding phase. The first and third rows are feature heat maps of U-Net employing the standard convolution, and the second and fourth rows are feature heat maps of U-Net employing DyBAC. The encoding of U-Net has five stages, and we replace the convolution after the first layer with the proposed dynamic convolution-based bi-directional attention component (DyBAC). From left to right, the feature maps are shown from shallow to deep layers respectively, and different colors indicate different spatial weights.

TABLE II
COMPARISON OF ABLATION EXPERIMENTS ON THE DERMOSCOPY IMAGE VALIDATION SET UTILIZING DIFFERENT PROPORTIONS OF LABELED DATA FROM THE TRAINING SET. THE BEST VALUES ARE IN BOLD

Method	Labeled/Unlabeled	MT	D_1	D_2	DyBAC	DI(%)↑
Supervised+U-Net++	2594/0					87.67
Supervised+U-Net++	2594/0				✓	88.45
Supervised+U-Net++	519/0					84.36
Semi-supervised+U-Net++	519/2075	✓				85.83
Semi-supervised+U-Net++	519/2075	✓	✓			86.41
Semi-supervised+U-Net++	519/2075	✓		✓		86.36
Semi-supervised+U-Net++	519/2075	✓			✓	86.34
Semi-supervised+U-Net++	**519/2075**	✓	✓	✓	✓	**87.21**



(a) (b)

Fig. 4. The learning curves on the dermoscopy image training and validation sets by utilizing 2,594 labeled data, the blue and red curves represent U-Net++ employing DyBAC and the gray and yellow curves represent U-Net++ employing the standard convolution. (a) The accuracy curve of training and validation sets on the dermoscopy image dataset and (b) The loss curve of training and validation sets on the dermoscopy image dataset.

and the results in Table I demonstrate the effectiveness of our contributions. The semi-supervised adversarial learning method using a single segmentation network and a single discriminator obtains a lower DI of 92.11% compared to other semi-supervised methods. Compared to the supervised U-Net, the semi-supervised method MT gets an improvement (4.22% for DI) and our proposed ASE-Net improves by 5.95%, benefiting from the MT framework that provides pseudo labels for the student model through the teacher model, which can better utilize the unlabeled data and effectively improve the performance of the network. To demonstrate the effectiveness of the proposed ASE-Net, we add the proposed discriminators D_1, D_2, and dynamic convolution-based bidirectional attention component (DyBAC) to MT, respectively. It can be seen that the discriminators D_1, D_2 and DyBAC achieve the increase of DI by 0.72%, 0.75%, and 0.97% based on MT.

In addition, as shown in Fig. 3, we visualize the feature heat maps generated by the standard convolution and the proposed DyBAC. The first and third rows are feature heat maps of U-Net employing standard convolution, and the second and fourth rows are feature heat maps of U-Net employing DyBAC. The encoding of U-Net has five stages, and we replaced all the convolution layers except for the first layer with the proposed DyBAC. From left to right, the feature maps are shown from shallow to deep layers respectively, and different colors indicate different spatial attention weights. It can be seen that our proposed DyBAC can effectively improve the liver segmentation in medical images.

As shown in Table II, the ablation experiment is performed on the dermoscopy image validation set, and the training set is divided into 20% labeled (519 images) and 80% unlabeled (2075 images). We use U-Net++, [2] as the backbone network for skin lesion segmentation and the results are in Table II. The supervised U-Net++ obtains 84.36% of DI, while the semi-supervised method MT obtains 85.83% of DI. It can be seen that the DI of our proposed discriminators D_1, D_2, and DyBAC are 0.58%, 0.53%, and 0.51% higher, respectively, compared to the baseline MT. Moreover, Fig. 4 shows the Dice values and loss curves for U-Net++ and U-Net++ with DyBAC on the training and validation sets under the condition of using 2,594 labeled data. To make an effective analysis, we do not use the semi-supervised regularization strategy during the process of experiments. As shown in Fig. 4, after the 40th epochs, the loss curve on the validation set of U-Net++ shows a large oscillation, which makes it difficult to converge, and the corresponding validation Dice curve is also suffered a drop. In contrast, by adding the DyBAC to U-Net++, the new loss curve becomes relatively stable, and the validation Dice curve has an upward trend. Therefore, the proposed DyBAC can reduce overfitting risk for small datasets.

TABLE III

COMPARISON OF ABLATION EXPERIMENTS ON THE LEFT ATRIUM VALIDATION SET BY UTILIZING 10% LABELED DATA OF TRAINING SET. THE BEST VALUES ARE IN BOLD

Method	Labeled/Unlabeled	MT	D_1	D_2	DyBAC	DI(%)↑
Supervised+V-Net	8/0					79.99
Semi-supervised+V-Net	8/72		✓			76.15
Semi-supervised+V-Net	8/72	✓				84.24
Semi-supervised+V-Net	8/72	✓	✓			85.82
Semi-supervised+V-Net	8/72	✓		✓		86.17
Semi-supervised+V-Net	8/72	✓			✓	85.75
Semi-supervised+V-Net	8/72	✓	✓	✓		86.94
Semi-supervised+V-Net	8/72	✓	✓	✓	✓	**87.83**

In addition, we extend the proposed ASE-Net to the 3D MR left atrium image segmentation task. We use V-Net, [25] as the backbone of the segmentation network. The ablation experiments are performed on the 10% labeled and 90% unlabeled of the training set. As shown in Table III, the semi-supervised adversarial learning method using a single segmentation network and a single discriminator achieves the lowest 76.15% of DI, while the supervised V-Net achieves 79.99% of DI and the MT method achieves 84.24% of DI. Based on MT, our proposed discriminators D_1, D_2, and DyBAC improve the DI values by 1.58%, 1.93% and 1.51%, respectively.

Overall, the additional discriminator D_1 allows the network to obtain effectively the prior relationship between unlabeled data and labeled data. The additional discriminator D_2 enables the network to learn effectively the prediction consistency when performing different perturbations on the same input, which further increases the consistency constraint based on MT. The proposed DyBAC effectively enhances the network for image feature representation and improves the segmentation accuracy.

D. Comparative Experiments on Different Datasets

In order to verify the effectiveness of our proposed method, we compare with supervised methods U-Net [1] U-Net++ [2] and V-Net [25] as well as seven state-of-the-art semi-supervised methods DAN [12] MT [5] UA-MT, [9] TCSM_v2 [6] CP [8], DTC, [33] and MC-Net [31] on three publicly available datasets LiTS [48] ISIC dermoscopy image dataset [49], and 3D MR Left atrial dataset [50]. In addition, for the semi-supervised experimental setup, we perform comparison experiments on 10% labeled and 90% unlabeled, as well as 20% labeled and 80% unlabeled of training sets, respectively.

1) CT Liver Segmentation: For a fair comparison, we use U-Net as the backbone network for all methods in the liver segmentation task. Table IV shows the comparison results of different methods on the LiTS-liver testing set under the condition of utilizing 10% labeled data. It can be seen that DAN [12] improves DI by 4.01% and ASD by 2.25mm compared to U-Net [1] under the condition of utilizing the same proportion of labeled data. This shows that DAN can effectively use unlabeled data to obtain better segmentation results by using adversarial training methods. MT, [5] and its improved methods UA-MT [9], TCSM_V2, [6] CPS [8] DTC [33], and MC-Net [31] also show some advantages compared to DAN, which indicates that the consistency regularization methods further enhance the utilization of unlabeled data. The proposed ASE-Net reaches 94.12% of Dice and 3.51mm of ASD. Compared to the supervised method, our method records improvements of 5.95% for DI and 3.38 mm for ASD. Compared with the latest semi-supervised method MC-Net [31], our method offers improved performance by 0.5% for DI and 0.21 mm for ASD.

TABLE IV

QUANTITATIVE COMPARISON BETWEEN OUR METHOD AND OTHER COMPARISON METHODS ON THE LITS-LIVER TESTING SET BY UTILIZING 10% LABELED DATA OF TRAINING SET. THE BACKBONE NETWORK OF ALL EVALUATED METHODS IS U-NET. THE BEST VALUES ARE IN BOLD

	LiTS-Liver (10% labeled / 90% unlabeled)				
Method	Labeled/Unlabeled	DI (%)↑	Imp.	ASD (mm)↓	Imp.
U-Net [1]	121/0	96.57	–	2.19	–
U-Net [1]	12/0	88.17	–	6.89	–
DAN [12]	12/109	92.18	4.01	4.64	2.25
MT [5]	12/109	92.39	4.22	3.85	3.04
UA-MT [9]	12/109	93.14	4.97	4.22	2.67
TCSM_v2 [6]	12/109	93.22	5.05	3.91	2.98
CPS [8]	12/109	93.31	5.14	3.83	3.06
DTC [33]	12/109	93.67	5.50	3.64	3.25
MC-Net [31]	12/109	93.62	5.45	3.72	3.17
ASE-Net	12/109	**94.12**	**5.95**	**3.51**	**3.38**

Table V shows the experimental results with 20% labeled and 80% unlabeled conditions, and we can see that our ASE-Net improves 6.02% for DI and 3.32mm for ASD compared to the supervised learning method. Moreover, the experimental results of the proposed ASE-Net under the condition of 20% labeled data are much closer to those of U-Net using 100% labeled data. It can be demonstrated that our ASE-Net effectively utilizes the advantages of consistency learning and adversarial learning, which can further improve the performance of our network.

In addition, Fig. 5 shows the visualization results of different methods under the condition of 10% labeled data, where the green indicates the ground truth, the red indicates the segmentation result and the yellow indicates the overlap of the segmentation result, and the ground truth. Therefore, fewer green and red regions, and more yellow regions represent better segmentation results. The last column in Fig. 5 shows the segmentation results provided by our ASE-Net, it is clear that our ASE-Net provides better segmentation results than other methods used for comparison.

E. Skin Lesion Segmentation

To validate further our proposed ASE-Net, we conducted sufficient experiments on the ISIC dataset. We use U-Net++ [2] as the backbone network for all semi-supervised methods, and we also perform quantitative comparisons using 10% and 20% labeled data, respectively. Table VI shows the segmentation results for the validation set under the condition of 10% labeled data of the training set. Using the same number

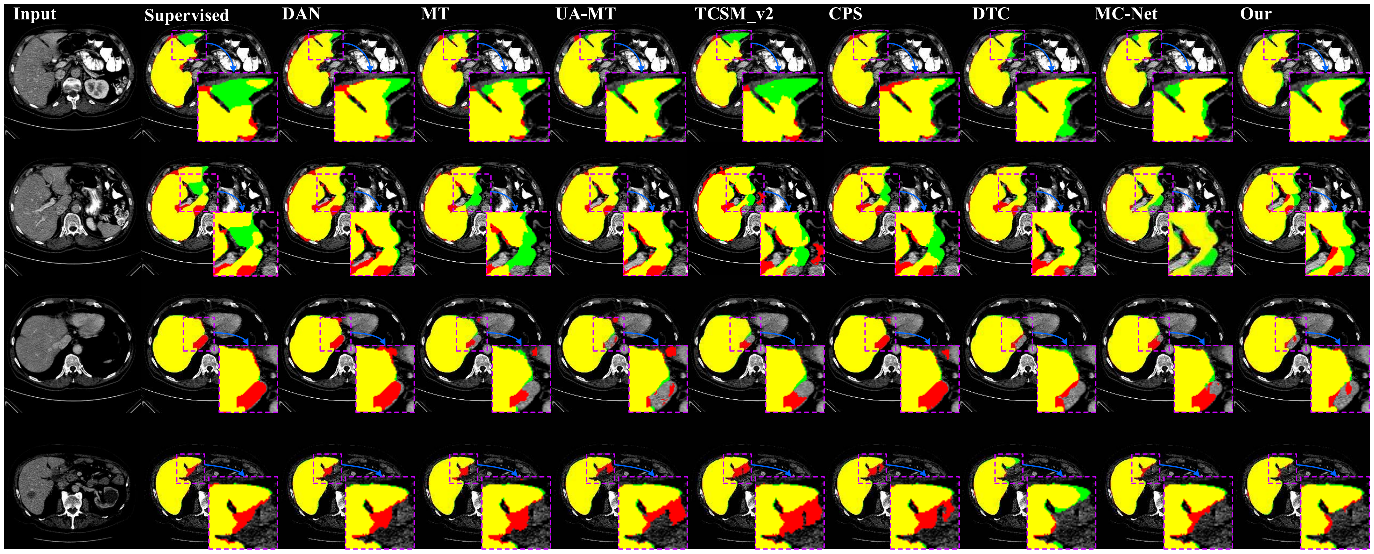


Fig. 5. Visualization result of different methods on the LiTS testing set by utilizing 10% labeled data of training set. Green is the ground truth, red is the segmentation result, and yellow is the overlap region of the segmentation result and ground truth. Therefore, fewer green and red regions imply better segmentation results.

TABLE V

Quantitative Comparison Between Our Method and Other Comparison Methods on the LiTS-Liver Test Dataset by Utilizing 20% Labeled Data of Train Dataset. The Backbone Network of All Evaluated Methods Is U-Net. The Best Values Are in Bold

LiTS-Liver (20% labeled / 80% unlabeled)					
Method	Labeled/Unlabeled	DI (%)↑	Imp.	ASD (mm)↓	Imp.
U-Net [1]	24/0	89.05	–	6.36	–
DAN [12]	24/97	93.01	3.96	3.98	2.38
MT [5]	24/97	93.42	4.37	3.64	2.72
UA-MT [9]	24/97	93.71	4.66	3.75	2.61
TCSM_v2 [6]	24/97	94.30	5.25	3.35	3.01
CPS [8]	24/97	94.23	5.18	3.46	2.90
DTC [33]	24/97	94.36	5.31	3.38	2.98
MC-Net [31]	24/97	94.58	5.53	3.21	3.15
ASE-Net	24/97	**95.07**	**6.02**	**3.04**	**3.32**

TABLE VI

Quantitative Comparison Between Our Method and Other Comparison Methods on the Dermoscopy Image Validation Set by Utilizing 10% Labeled Data of the Training Set. The Backbone Network of All Evaluated Methods Is U-Net++. The Best Values Are in Bold

Skin lesion (10% labeled / 90% unlabeled)						
Method	Labeled/Unlabeled	DI(%)↑	JA(%)↑	SE(%)↑	AC(%)↑	SP(%)↑
U-Net++ [2]	2594/0	87.67	80.06	90.65	93.29	96.78
U-Net++ [2]	259/0	82.57	73.55	88.31	91.01	93.76
DAN [12]	259/2335	84.26	75.15	87.23	91.97	95.75
MT [5]	259/2335	84.58	76.54	87.25	92.02	95.69
UA-MT [9]	259/2335	84.80	78.02	88.63	91.94	95.82
TCSM_v2 [6]	259/2335	84.71	75.55	90.22	91.92	95.77
CPS [8]	259/2335	84.72	76.81	86.87	91.87	95.42
DTC [33]	259/2335	84.56	76.33	87.19	91.79	95.54
MC-Net [31]	259/2335	84.81	76.64	87.41	91.91	95.97
ASE-Net	259/2335	**85.19**	**78.80**	**90.38**	**92.40**	**96.15**

of labeled data, our method shows an overall improvement compared to the supervised method (2.62% for DI, 5.25% for JA, 2.07% for SE, 1.39% for AC, 2.39% for SP). Our method also shows some improvement compared to state-of-the-art semi-supervised methods. Moreover, Table VII shows the results of our ASE-Net compared with other methods under 20% labeled data condition, and it can be seen that our method obtains the highest DI of 87.21%, JA of 79.25%, SE of 91.15% and AC of 93.09%. Therefore, our ASE-Net can effectively utilize the prior relationship between unlabeled and labeled data and possesses a better feature representation ability.

Fig. 6 shows some of the visualization results of the validation set under the condition of 20% labeled data of the skin lesions dataset. We can see that the comparative methods only provide rough boundaries but our ASE-Net obtains high-quality segmentation results with smooth boundaries compared to other methods. One of the main reasons is that the two additional discriminator networks generate additional supervised information for the segmentation network by learning the matching relationship between the original image and segmentation results. It can be further analyzed to demonstrate that the discriminator networks are very sensitive to the boundaries of segmentation results. The main reason is that the segmentation network can roughly predict the location of targets, but the prediction of boundaries is not fine enough. Therefore, the discriminator networks make the segmentation network generate high-quality segmentation results with smooth boundaries by continuously feeding back the segmentation network's prediction quality on boundaries.

1) MR Left Atrium Segmentation: In order to demonstrate the effectiveness of the proposed ASE-Net in 3D medical image segmentation tasks, we extend the application of

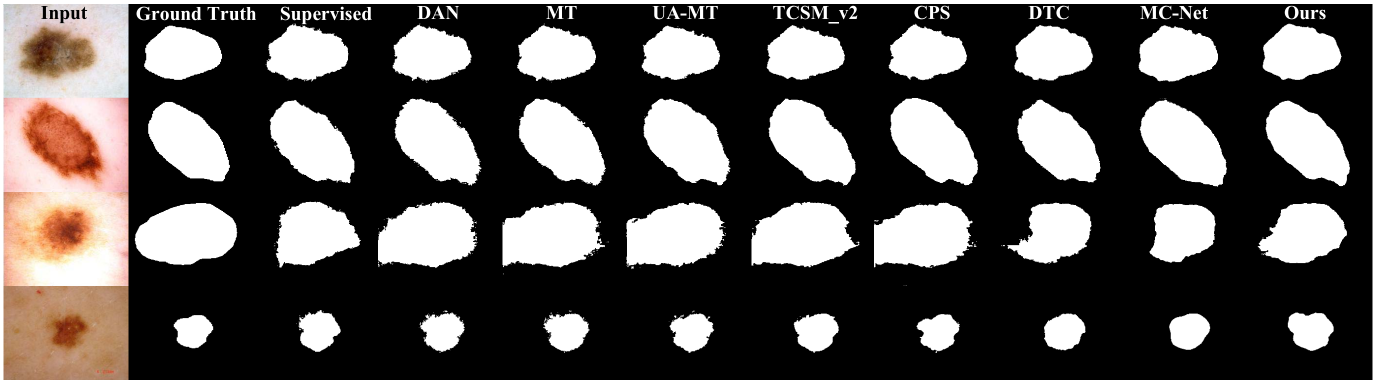


Fig. 6. Visualization result of different methods on the dermoscopy image validation set by utilizing 20% labeled data of training set.

TABLE VII
QUANTITATIVE COMPARISON BETWEEN OUR METHOD AND OTHER COMPARISON METHODS ON THE DERMOSCOPY IMAGE VALIDATION SET BY UTILIZING 20% LABELED DATA OF THE TRAINING SET. THE BACKBONE NETWORK OF ALL EVALUATED METHODS IS U-NET++. THE BEST VALUES ARE IN BOLD

Skin lesion (20% labeled / 80% unlabeled)						
Method	Labeled/Unlabeled	DI(%)↑	JA(%)↑	SE(%)↑	AC(%)↑	SP(%)↑
U-Net++ [2]	519/0	84.36	75.64	88.83	92.15	94.95
DAN [12]	519/2075	85.41	77.16	89.69	92.16	**95.01**
MT [5]	519/2075	85.83	77.48	89.97	92.57	94.46
UA-MT [9]	519/2075	86.19	78.06	90.94	92.71	94.49
TCSM_v2 [6]	519/2075	86.16	77.98	91.07	92.56	94.26
CPS [8]	519/2075	86.34	78.17	90.57	92.72	94.78
DTC [33]	519/2075	85.91	77.63	90.24	92.79	94.40
MC-Net [31]	519/2075	86.37	78.11	90.85	92.61	94.64
ASE-Net	519/2075	**87.21**	**79.25**	**91.15**	**93.09**	94.52

TABLE VIII
QUANTITATIVE COMPARISON BETWEEN OUR METHOD AND OTHER COMPARISON METHODS ON THE LEFT ATRIUM VALIDATION SET BY UTILIZING 10% LABELED DATA OF TRAINING SET. THE BACKBONE NETWORK OF ALL EVALUATED METHODS IS V-NET. THE BEST VALUES ARE IN BOLD

left atrium (10% labeled / 90% unlabeled)					
method	Labeled/Unlabeled	DI(%)	JA(%)↑	95HD(mm)↓	ASD(mm)↓
V-Net [25]	80/0	91.14	83.82	5.75	1.52
V-Net [25]	8/0	79.99	68.12	21.11	5.48
DAN [12]	8/72	75.11	63.47	19.03	3.57
MT [5]	8/72	84.24	73.26	19.41	2.71
UA-MT [9]	8/72	84.25	73.48	13.84	3.36
TCSM_V2 [6]	8/72	84.21	73.19	19.56	3.07
CPS [8]	8/72	84.09	73.17	22.55	2.41
DTC [33]	8/72	86.57	76.55	14.47	3.74
MC-Net [31]	8/72	87.71	78.31	**9.36**	2.18
ASE-Net	8/72	**87.83**	**78.45**	9.86	**2.17**

TABLE IX
QUANTITATIVE COMPARISON BETWEEN OUR METHOD AND OTHER COMPARISON METHODS ON THE LEFT ATRIUM VALIDATION SET BY UTILIZING 20% LABELED DATA OF TRAINING SET. THE BACKBONE NETWORK OF ALL EVALUATED METHODS IS V-NET. THE BEST VALUES ARE IN BOLD

left atrium (20% labeled / 80% unlabeled)					
method	Labeled/Unlabeled	DI(%)↑	JA(%)↑	95HD(mm)↓	ASD(mm)↓
V-Net [25]	16/0	86.03	76.06	14.26	3.51
DAN [12]	16/64	87.52	78.29	9.01	2.42
MT [5]	16/64	88.42	79.45	13.07	2.73
UA-MT [9]	16/64	88.88	80.21	7.32	2.26
TCSM_V2 [6]	16/64	86.26	76.56	9.67	2.35
CPS [8]	16/64	87.87	78.61	12.87	2.16
DTC [33]	16/64	89.42	80.98	7.32	2.10
MC-Net [31]	16/64	**90.34**	82.48	**6.00**	1.77
ASE-Net	16/64	90.29	**82.76**	7.18	**1.64**

ASE-Net to the 3D left atrium for experiments. We perform quantitative comparisons using 10% and 20% labeled data, respectively. All the comparison methods in the experiment employ V-Net [25] as the backbone. The specific experimental results are shown inTables VIII and IX. It can be seen that our ASE-Net obtains a higher Dice value of 87.83% than other semi-supervised methods under the condition of 10% labeled data. However, our ASE-Net achieves a slightly lower value of DI (0.05%) than the latest MC-Net [31] under the condition of 20% labeled data as shown in Table IX. One of the main reasons is that MC-Net [31] employs a double-decoder architecture containing more parameters (12.35 M) than our ASE-Net (3.92 M) to improve the segmentation accuracy. Fig. 7 shows the segmentation results on the left atrium dataset with the latest methods DTC [33] and MC-Net [31] under 10% labeled and 20% labeled data, respectively. It is clear that our results are closer to the ground truth.

In general, our ASE-Net can effectively combine consistency and adversarial learning to make the segmentation network learn consistently for both labeled and unlabeled data. In addition, the proposed two discriminators can effectively learn the segmentation difference between labeled data and unlabeled data, the segmentation difference between perturbed data and unperturbed data. Furthermore, the obtained difference is used to update the segmentation network for achieving better segmentation results.

V. DISCUSSION

A. Model-Size Comparison

Table X shows the comparison of parameters, floating point operations (FLOPs), and model size of different networks in

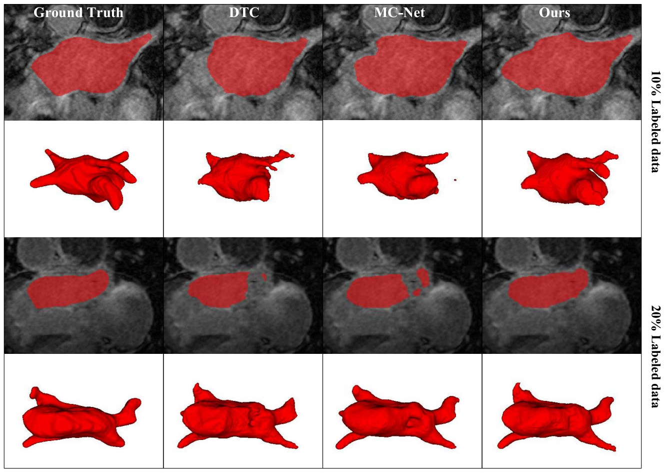


Fig. 7. Visualization result of different methods on the left atrium validation set by utilizing 10% and 20% of the labeled data in the training set, respectively.

TABLE X

COMPARISON OF THE EFFICIENCY OF DIFFERENT NETWORKS, THE BEST VALUES ARE IN BOLD

Model	Operations (GFLOPs)	Parameters (M)	Model Size (MB)
U-Net [1]	65.39	34.52	131.82
ASE-Net(U-Net)	**9.26**	**5.18**	**21.11**
U-Net++ [2]	49.95	11.79	45.08
ASE-Net(U-Net++)	**25.34**	**4.92**	**19.79**
V-Net [25]	46.94	9.44	36.11
ASE-Net(V-Net)	**22.97**	**3.92**	**15.75**

the inference phase. Since our proposed discriminator networks are only used in the training phase, we only test the efficiency of the segmentation network. Specifically, we replace the standard convolution of the segmentation network with a dynamic convolution-based bidirectional attention component (DyBAC) while the first layer is excluded. The computational cost of the 2D networks is estimated with an input size of 1 × 256 × 256, and the computational cost of the 3D networks is evaluated with an input size of 112 × 112 × 80. It can be seen that when the backbone network adopts U-Net, [1], the number of parameters of ASE-Net is only 15.0% of the original U-Net. When the backbone network is U-Net++ [2] with dense skip-connection, the number of parameters of ASE-Net is only 41.7% of the original U-Net++. When the backbone network is V-Net [25], the number of parameters of ASE-Net is only 41.5% of the original V-Net. Obviously our ASE-Net significantly reduces the number of parameters and computational costs.

B. Statistical Analysis

Since the statistical significance of an algorithm can indicate that the differences observed in experiments are real but not accidental, we perform the paired t-test with $\alpha = 0.05$ on the LiTS [48] and the dermoscopy image, [49] datasets. As shown in Table XI, we mainly conducted the one-tailed test by Dice metric and calculated the p value between MT [5] and our ASE-Net. We find that the p value is less than 0.05. Generally, if $p < 0.05$, then there is a significant difference. Therefore, through the above analysis, it can be concluded that the proposed ASE-Net is statistically significant.

TABLE XI

STATISTICAL SIGNIFICANCE OF THE PROPOSED ASE-NET AND BASELINE MT METHODS ON DIFFERENT DATASETS

Datasets-Labeled	Dice (%)		p value
	MT	ASE-Net (our)	
LiTS (10%)	92.39	94.12	0.011
LiTS (20%)	93.42	95.07	0.0091
Skin lesion (10%)	84.58	85.19	0.021
Skin lesion (20%)	85.83	87.21	0.0074

VI. CONCLUSION

In this work, we have proposed ASE-Net for semi-supervised medical image segmentation. First, the proposed ACTS effectively combines adversarial learning and consistency learning, using adversarial training to maximize consistency learning. This allows the network to learn quickly the prior relationship between unlabeled and labeled data, and further mines the potential knowledge existing in unlabeled data. Then, our proposed DyBAC adaptively adjusts the parameter values of convolutional kernels according to input samples, which not only effectively avoids network overfitting and improves the feature representation ability of the network but also reduces the memory overhead. Experiments on three publicly available benchmark datasets demonstrate that our proposed ASE-Net outperforms state-of-the-art methods and provides an effective solution for semi-supervised medical image segmentation, significantly reducing network overfitting risk and uncertainty prediction in consistency learning.

REFERENCES

[1] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[2] Z. Zhou et al., “UNet++: Redesigning skip connections to exploit multiscale features in image segmentation,” *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Dec. 2019.

[3] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, “H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes,” *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2663–2674, Dec. 2017.

[4] Z.-H. Zhou, “A brief introduction to weakly supervised learning,” *Nat. Sci. Rev.*, vol. 5, no. 1, pp. 44–53, 2018.

[5] A. Tarvainen and H. Valpola, “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results,” in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 1195–1204.

[6] X. Li, L. Yu, H. Chen, C.-W. Fu, and P.-A. Heng, “Transformation-consistent self-ensembling model for semisupervised medical image segmentation,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 523–534, Feb. 2021.

[7] K. Sohn et al., “FixMatch: Simplifying semi-supervised learning with consistency and confidence,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 596–608.

[8] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2613–2622.

[9] 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 *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 605–613.

[10] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12674–12684.

[11] W. C. Hung, Y. H. Tsai, Y. T. Liou, Y. Y. Lin, and M. H. Yang, "Adversarial learning for semi-supervised semantic segmentation," in *Proc. 29th Brit. Mach. Vis. Conf. (BMVC)*, 2018, pp. 1–12.

[12] 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 *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2017, pp. 408–416.

[13] G. Chen et al., "MTANS: Multi-scale mean teacher combined adversarial network with shape-aware embedding for semi-supervised brain lesion segmentation," *NeuroImage*, vol. 244, Dec. 2021, Art. no. 118568.

[14] Y. Zhu et al., "Improving semantic segmentation via self-training," 2020, *arXiv:2004.14960*.

[15] Z. Feng et al., "DMT: Dynamic mutual training for semi-supervised learning," 2020, *arXiv:2004.08514*.

[16] Y. Shi et al., "Inconsistency-aware uncertainty estimation for semi-supervised medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 41, no. 3, pp. 608–620, Mar. 2022.

[17] J. Xiang, Z. Li, W. Wang, Q. Xia, and S. Zhang, "Self-ensembling contrastive learning for semi-supervised medical image segmentation," 2021, *arXiv:2105.12924*.

[18] X. Hu, D. Zeng, X. Xu, and Y. Shi, "Semi-supervised contrastive learning for label-efficient medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 481–490.

[19] C. You, Y. Zhou, R. Zhao, L. Staib, and J. S. Duncan, "SimCVD: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 41, no. 9, pp. 2228–2237, Sep. 2022.

[20] C. Li et al., "Self-ensembling co-training framework for semi-supervised COVID-19 CT segmentation," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 11, pp. 4140–4151, Nov. 2021.

[21] P. Wang, J. Peng, M. Pedersoli, Y. Zhou, C. Zhang, and C. Desrosiers, "Self-paced and self-consistent co-training for semi-supervised image segmentation," *Med. Image Anal.*, vol. 73, Oct. 2021, Art. no. 102146.

[22] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, and A. K. Nandi, "Medical image segmentation using deep learning: A survey," *IET Image Process.*, vol. 16, no. 5, pp. 1243–1267, Apr. 2022.

[23] T. Lei, R. Wang, Y. Zhang, Y. Wan, C. Liu, and A. K. Nandi, "DefED-Net: Deformable encoder–decoder network for liver and liver tumor segmentation," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 6, no. 1, pp. 68–78, Jan. 2022.

[24] J. Wang, L. Wei, L. Wang, Q. Zhou, L. Zhu, and J. Qin, "Boundary-aware transformers for skin lesion segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 206–216.

[25] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

[26] T. Lei, W. Zhou, Y. Zhang, R. Wang, H. Meng, and A. K. Nandi, "Lightweight V-Net for liver segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 1379–1383.

[27] L. Samuli and A. Timo, "Temporal ensembling for semi-supervised learning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017, vol. 4, no. 5, p. 6.

[28] M. Kozinski, F. Jurie, and L. Simon, "An adversarial regularisation for semi-supervised training of structured output neural networks," in *Proc. Conf. Neural Inf. Process. Syst.*, 2017, pp. 1–7.

[29] J. Dolz, C. Desrosiers, and I. B. Ayed, "Teach me to segment with mixed supervision: Confident students become masters," in *Proc. Int. Conf. Inf. Process. Med. Imag.* Cham, Switzerland: Springer, 2021, pp. 517–529.

[30] S. Li, Z. Zhao, K. Xu, Z. Zeng, and C. Guan, "Hierarchical consistency regularized mean teacher for semi-supervised 3D left atrium segmentation," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 3395–3398.

[31] Y. Wu, M. Xu, Z. Ge, J. Cai, and L. Zhang, "Semi-supervised left atrium segmentation with mutual consistency training," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 297–306.

[32] Y. Liu, Y. Tian, Y. Chen, F. Liu, V. Belagiannis, and G. Carneiro, "Perturbed and strict mean teachers for semi-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4258–4267.

[33] X. Luo, J. Chen, T. Song, and G. Wang, "Semi-supervised medical image segmentation through dual-task consistency," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 10, pp. 8801–8809.

[34] H. Peiris, Z. Chen, G. Egan, and M. Harandi, "Duo-SegNet: Adversarial dual-views for semi-supervised medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Springer, 2021, pp. 428–438.

[35] J. Hou, X. Ding, and J. D. Deng, "Semi-supervised semantic segmentation of vessel images using leaking perturbations," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Jan. 2022, pp. 2625–2634.

[36] C. Li and H. Liu, "Generative adversarial semi-supervised network for medical image segmentation," in *Proc. IEEE 18th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2021, pp. 303–306.

[37] H. Wu, G. Chen, Z. Wen, and J. Qin, "Collaborative and adversarial learning of focused and dispersive representations for semi-supervised polyp segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3489–3498.

[38] Y. Han, G. Huang, S. Song, L. Yang, H. Wang, and Y. Wang, "Dynamic neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7436–7456, Nov. 2022.

[39] Y. Li et al., "Learning dynamic routing for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8553–8562.

[40] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11030–11039.

[41] B. Yang, G. Bender, Q. V. Le, and J. Ngiam, "CondConv: Conditionally parameterized convolutions for efficient inference," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.

[42] R. Gu et al., "CA-Net: Comprehensive attention convolutional neural networks for explainable medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 40, no. 2, pp. 699–711, Feb. 2021.

[43] D. Li et al., "Involution: Inverting the inherence of convolution for visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12321–12330.

[44] J. Zhou, V. Jampani, Z. Pi, Q. Liu, and M.-H. Yang, "Decoupled dynamic filter networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6647–6656.

[45] C. Li, A. Zhou, and A. Yao, "Omni-dimensional dynamic convolution," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–20.

[46] Z. Ke, D. Wang, Q. Yan, J. Ren, and R. Lau, "Dual student: Breaking the limits of the teacher in semi-supervised learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6728–6736.

[47] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[48] P. Bilic et al., "The liver tumor segmentation benchmark (LiTS)," 2019, *arXiv:1901.04056*.

[49] N. Codella et al., "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC)," 2019, *arXiv:1902.03368*.

[50] Z. Xiong et al., "A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging," *Med. Image Anal.*, vol. 67, Jan. 2021, Art. no. 101832.

[51] A. Jungo and M. Reyes, "Assessing reliability and challenges of uncertainty estimations for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 48–56.