SW-Seg: Semi- and Weakly-Supervised Learning for Accurate Medical Image Segmentation Using Inaccurate Scribble Annotations

Zhenghua Xu, Biao Tian, Xiangtao Wang, Shuo Zhang, Yuefu Zhan, Thomas Lukasiewicz

***Abstract*—Insufficient pixel-level annotations have been a challenge for medical image segmentation. Semi-supervised learning has been proposed as a solution, training deep segmentation models on a combination of labeled and unlabeled data. However, despite utilizing unlabeled data efficiently, these methods still require a high ratio of high-quality annotations for optimal performance and may not achieve satisfactory results when the ratio is low. Therefore, in this work, we propose a novel semi- and weakly-supervised medical image segmentation method, called SW-Seg, where weak labels are used to partially replace pixel-wise labels, thereby greatly reducing the number of pixel-wise labeled data required for the model to achieve a satisfactory medical image segmentation performance. Specifically, we design a mixed consistency training strategy that constrains the consistency of the prediction results for the same input under different decoders. Secondly, we also design a mixed supervised learning method that generates labels as supervision by dynamically mixing the two prediction results. Experimental results on two public benchmark segmentation datasets show that, by using only 5% (resp., 10%) of pixel-wise annotated data together with 5% (resp., 10%) of weakly annotated data, the proposed SW-Seg can greatly outperform the state-of-the-art semi-supervised methods using 10% (resp., 20%) of pixel-wise annotated data and also achieve performances that are very close to the fully supervised results with 100% pixel-wise labels, which thus proves that our work can achieve superior segmentation performances with very low labeling workload.**

***Index Terms*—Semi-Supervised Learning, Weakly-Supervised Learning, Medical Image Segmentation, Mix Consistency Learning**

I. Introduction

Deep learning has achieved an extraordinary success in medical image segmentation tasks [1, 2, 3, 4]. The performance of fully supervised learning depends on the quality and quantity of pixel-level annotations. To guarantee good segmentation results, it is necessary to have a large amount of medical image data with accurate segmentation annotations for training the deep learning model. However, obtaining high-quality segmentation and annotation of medical images is a complex and specialized task that requires annotators with exceptional professional skills and experience. The judgment of the lesion edge in the image is very dependent on the subjective experience of professional doctors, and different doctors may have different perceptions of the edge of the same lesion. To ensure the accuracy and consistency of labeling, obtaining high-quality medical image segmentation annotations often requires cross-validation by multiple professionals with extensive experience. At the same time, the high intensity and homogenization of labeling work also tend to cause doctor fatigue and affect the accuracy of labeling. Therefore, the difficulty of obtaining a large number of high-quality segmentation annotations in the clinical setting is one of the key factors limiting the clinical implementation of intelligent medical image segmentation systems.

Considering that unlabeled data are usually abundant and easily available, many researchers focus on implementing segmentation tasks in semi-supervised learning (SSL) [5, 6]. The main idea of semi-supervised segmentation is to learn from a limited amount of labeled data and a large amount of unlabeled data to improve the accuracy of segmentation. Studies on semi-supervised learning indicate that using substantial amounts of unlabeled data can enhance model performance. Specifically, according to the training manner, common semi-supervised segmentation methods include especially consistency regularization [7, 8], self-training [9], collaborative training [10, 11], adversarial training [12, 13], and entropy minimization [14]. However, although these methods achieve a better segmentation performance, they still require a large proportion of high-quality annotations, and the effect is not satisfactory when the ratio is lower. For example, some semi-supervised segmentation methods train models using 10% high-quality labeled data and 90% unlabeled data to achieve a near fully supervised performance [15, 9], but when trained at lower labeling ratios (e.g., 5% pixel-level labeled data), their performance results show a significant difference from fully supervised performance. Acquiring 10% pixel-level annotations remains excessively costly for medical image data. Therefore, efficiently utilizing a lower ratio of high-quality segmentation annotations to improve the performance of medical image segmentation remains a challenge.

This work was supported by the National Natural Science Foundation of China under the grants 62276089 and 61906063, by the Natural Science Foundation of Hebei Province, China, under the grant F2021202064, by the Key Research and Development Project of Hainan Province, China, under the grant ZDYF2022SHFZ015, by the Natural Science Foundation of Hainan Province, China, under the grant 821RC1131, and by the Hainan Province Clinical Medical Center, China, under the grant QWYH202175. This work was also supported by the AXA Research Fund. (Corresponding author: Zhenghua Xu, e-mail: zhenghua.xu@hebut.edu.cn.)

Zhenghua Xu, Biao Tian, Xiangtao Wang and Shuo Zhang are with the State Key Laboratory of Reliability and Intelligence of Electrical Equipment, School of Health Sciences and Biomedical Engineering, Hebei University of Technology, China.

Yuefu Zhan is with the Department of Radiology, Hainan Women and Children's Medical Center, Haikou, China.

Thomas Lukasiewicz is with the Department of Computer Science, University of Oxford, United Kingdom and the institute of Logic and Computation, Vienna University of Technology, Vienna, Austria.

Inspired by the success of semi-supervised learning, we propose a semi- and weakly-supervised framework for medical image segmentation, called SW-Seg, which uses weak labels to partially replace pixel-wise labels, ensuring the segmentation accuracy and full utilization of existing mixed-label data. By using weak labels, our approach reduces the reliance on pixel-wise annotations and lowers annotation costs. Weak annotations are easily accessible, such as image-level, bounding box, scribble, and point annotations. Scribble annotations, which consist of masks for a small fraction of pixels in images, are particularly useful for annotating complex lesions and organs. They are more general than bounding box annotations, point annotations, and image-level labels, and have been shown to be effective in segmentation tasks [16]. In our approach, we use a combination of pixel-wise annotations, scribble annotations, and unlabeled data to perform mixed supervised segmentation tasks. Specifically, SW-Seg is a mixed supervised framework that combines a shared encoder for feature extraction with two different decoders for segmentation and auxiliary training. The shared encoder extracts features that are then used by the two decoders to perform segmentation tasks. We have designed two different training strategies to make full use of the mixed data. First, to further improve the consistency of network predictions, with the help of auxiliary decoders, we design a mixed consistency training strategy by performing consistent constraints on prediction results of the same input under different decoders. In addition to the mixed consistency training strategy, we introduce a mixed supervised learning approach where a mixed label is generated as supervision by dynamically mixing the two of predictions. This approach allows us to make full use of the available labeled and unlabeled data in order to improve the performance of our medical image segmentation framework.

This work's major contributions are summarized as follows:

- We identify a common issue among existing semi-supervised medical image segmentation methods, namely, the unsatisfactory performance when there is a low ratio of high-quality annotations. We propose a novel semi- and weakly-supervised medical image segmentation framework, called SW-Seg, that uses weak labels to partially replace pixel-wise labels. This method improves segmentation accuracy while reducing the required number of pixel-level annotated data.
- To optimize the SW-Seg network by consistency learning, we propose a new mixed consistency training strategy by dynamically mixing different prediction results from different labeled data for consistency constraints. Additionally, we employ mixed supervised training by dynamically mixing the two predictions to generate mixed labels. These further improve the performance of our semi- and weakly-supervised medical image segmentation method.
- Experiments on the ACDC and MSCMRSeg datasets show that, by using only 5% pixel-level annotated data and 5% weakly annotated data, our proposed SW-Seg greatly outperforms the state-of-the-art semi-supervised methods using 10% pixel-level annotated data; furthermore, the performances of SW-Seg are also very close to the fully supervised results using 100% annotated data. These findings prove that our work can achieve excellent segmentation performances with very low labeling workload.

II. Related Work

A. Semi-Supervised Semantic Segmentation

Semi-supervised learning utilizes a small amount of pixel-wise annotated data along with a large amount of unlabeled data to train models with strong generalization capabilities. Due to the scarcity of labeled data in medical image segmentation datasets, various semi-supervised learning methods have been extensively studied in recent years, including self-training [17, 9], adversarial training [12, 13, 18, 19], entropy minimization [14], and consistency regularization [5, 7, 20, 21, 22, 23]. Self-training models iteratively retrain using pseudo-labels generated by previously trained models as supervision. [19] proposes a new adversarial dual-student framework and differentiable geometric warping for unsupervised data augmentation to improve MeanTeacher [5]. [24] proposes an uncertainty-aware and multi-granularity consistent constrained semi-supervised hashing method to alleviate the negative effects of noisy supervised signals and enlarge the inter-class distance. Several methods have been proposed to generate better pseudo-labels. FixMatch [25] uses pseudo-segmentation from weakly perturbed images to supervise the predictions of strongly perturbed images on a single network. Additionally, Chen et al. [9] introduce the Cross Pseudo-Supervision (CPS) method, which initializes two different segmentation networks on the same input image and uses the prediction of one network to supervise the other, and vice versa. Consistency regularization methods utilize unlabeled data, assuming that a robust model should generate consistent predictions for similar inputs. Cheng et al. [24] propose an uncertainty-aware and multi-granularity consistent constrained semi-supervised hashing method to alleviate the negative effects of noisy supervised signals and enlarge the inter-class distance. Ouali et al. [23] propose the CCT model to further improve the consistency between the prediction of the main decoder and the prediction of the auxiliary decoder for unlabeled data. Typical approaches achieve this by enforcing prediction consistency with perturbed inputs, such as using Cutout [26], CutMix [27], or Gaussian noise as weak perturbations. Verma et al. [22] adopt the concept of MixUp [28] and propose Interpolation Consistency Training (ICT), which encourages predictions on mixed inputs to be consistent with the mixed predictions.

B. Weakly Supervised Semantic Segmentation

To reduce the labor-intensive pixel-level annotation, recent work has explored different forms of weak annotations for semantic segmentation [29], such as scribbles [30, 31, 32], bounding boxes [33, 34, 35], points [36], and image-level labels [37, 38], among others. These methods have been shown to effectively improve the performance of medical image segmentation while minimizing annotation burden. Dong et al. [38] proposed a novel weakly supervised lesion propagation

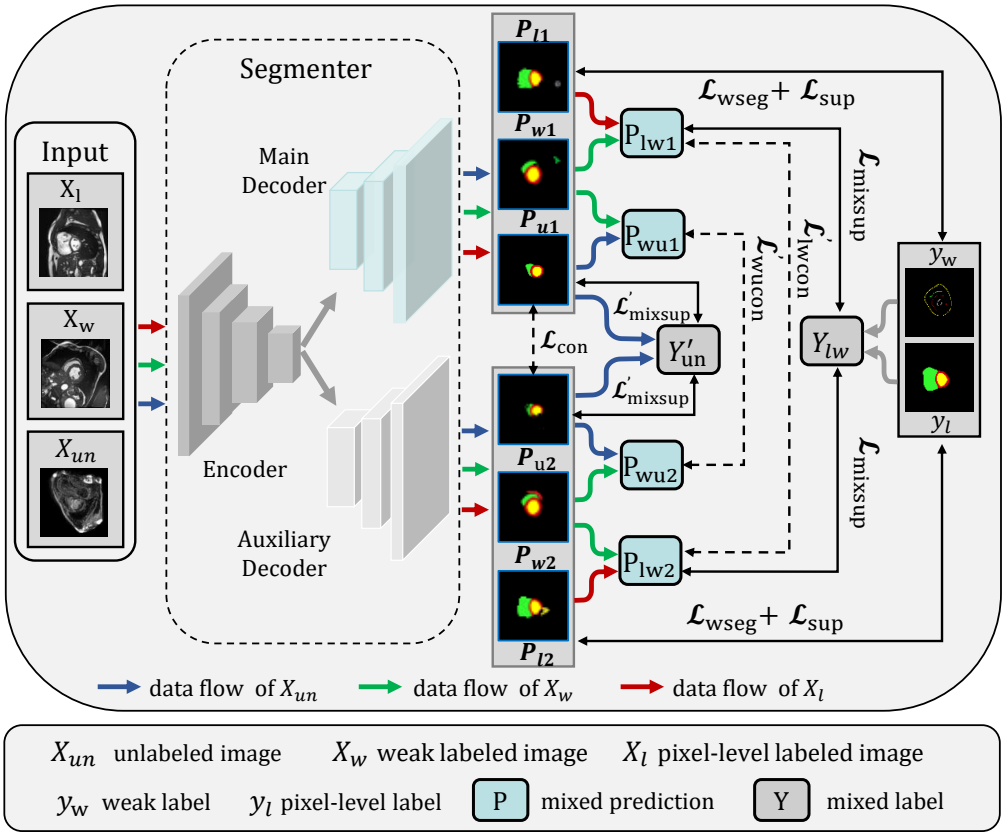

Fig. 1. Overview of the proposed segmentation network model based on semi- and weakly-supervised learning. The solid lines of different colors represent the forward processing of the corresponding data. The red line indicates pixel-level labeled data, the blue line indicates weakly labeled data, and the green line indicates unlabeled data.

framework based on image-level labels, which not only explores transferable domain-invariant feature knowledge across different datasets.

Scribble annotation is a flexible form of weak annotation that accurately reflects the distribution of semantic classes and is easy to control during the annotation process. Given its potential in medical image segmentation, we focus on the case of scribble-supervised semantic segmentation in this paper. Previous studies have demonstrated the effectiveness of weakly supervised methods with scribble annotations in medical imaging, achieving impressive results in various tasks. This work emphasizes on scribble-based annotations, which serve as an intuitive and effective substitute for segmentation annotations. During training, scribbles are provided for the background and each object, while the basic facts of other pixels remain unknown. The main challenge lies in the lack of supervision on object structures and uncertainty at boundaries. Therefore, [39] proposes to effectively learn from end-to-end scribble annotations by combining scribbles with auxiliary pseudo-label supervision. Lee and Jeong (2020) introduced the Scribble2Label architecture, which improves the labeled-filtered pseudo-labels by iteratively averaging predictions to leverage prediction consistency. In medical image segmentation, [40] proposed the USTM model, which adds uncertainty-aware segmentation methods to scribble-based segmentation approaches. Valvano et al. [41] proposed the use of attention mechanisms to generate segmentation masks in adversarial training. [31] proposed a method that utilizes mixed data augmentation and cycle-consistency constraints for scribble annotation data, significantly improving segmentation performance. This approach highlights the effectiveness of combining multiple techniques to maximize the utility of scribble annotations in medical image segmentation.

C. Mixed Supervised Semantic Segmentation

Mixed supervised learning (MSL) typically combines a small amount of pixel-level annotated data with a large amount of weakly annotated data. Pixel-level annotations provide precise and detailed information, while weak annotations include forms such as scribbles, bounding boxes, or image-level labels, providing a broader understanding of the data. For example, Sun et al. [42] proposed a hierarchical attention module to utilize pixel-level annotated data and weakly annotated data for liver and tumor segmentation. Wang et al. [43] proposed a mixed-supervision dual network that uses detection and segmentation networks to utilize weakly annotated bounding boxes and pixel-level labeled annotations. Dolz et al. [44] proposed a dual-branch architecture where the teacher receives strong supervision while the student is supervised by weakly annotated data. In contrast to MSL, we consider a semi-weak mixed-supervision setting in this paper, which trains a network using a small amount of pixel-level annotations, a small number of scribble annotations, and a large amount of unlabeled data.

III. METHOD

Figure 1 illustrates the overall structure of our semi- and weakly-supervised segmentation model. To facilitate information sharing and address supervision inconsistencies, we have implemented a dual-branch structure consisting of a shared encoder and two different decoders to handle different types of supervision. This allows us to efficiently use data of different annotation types to improve the performance of our model. First, the encoder is responsible for extracting features from the input data, the main decoder generates the prediction results directly using the extracted features, and the auxiliary decoder adds a feature perturbation to generate the prediction results. Then, after generating prediction results, supervised training is performed for the labeled data and consistency regularity constraints are used for the unlabeled data. Meanwhile, we perform mixed consistency training using the outputs of different decoders on the prediction results of mixed weakly labeled data and unlabeled data. This helps to optimize our model for an enhanced performance. Finally, to improve the interactive learning of different labeled data, we propose a mixed supervision training strategy to efficiently learn mixed labeled data.

A. Semi- and Weakly-Supervised Segmentation

In our setting, we have a small set of labeled training examples, a small set of scribble labeled examples, and a large set of unlabeled training examples. Let $\mathcal{D}_l = (x_l^1, y_l^1), \ldots, (x_l^n, y_l^n)$ represent the n labeled examples, where x_l^n is the n-th labeled input image with spatial dimensions $H \times W$, and its corresponding label $y_l^n \in \mathbb{R}^{C \times H \times W}$ is the pixel-wise mask. $\mathcal{D}_w = (x_w^1, y_w^1), \ldots, (x_w^m, y_w^m)$ the m weak labeled examples, x_w^m is the m-th weak labeled input image with spatial dimensions $H \times W$, and its corresponding label $y_w^m \in \mathbb{R}^{C \times H \times W}$ is the scribble mask, where c is the number of classes. $\mathcal{D}_u = (x_u^1, \ldots, x_u^v)$ the v unlabeled examples, and x_u^v is the v-th unlabeled input image with spatial dimensions $H \times W$. To take full advantage of pixel-level labeled data, weakly labeled data, and unlabeled data, we design a semi- and weakly-supervised segmentation network ($f(\theta_e, \theta_{d1}, \theta_{d2})$) by a shared encoder θ_e for feature extraction and two separate and distinct decoders (θ_{d1}, θ_{d2}) for segmentation. Specifically, to enhance the performance of our model, we have incorporated an auxiliary decoder θ_{d2} into a general U-Net [45], where we have introduced a feature-based perturbation [23]. This configuration allows SW-Seg to generate prediction results from two different decoders, enabling the encoder to benefit from separate supervision and improve feature extraction. The two slightly different decoders used in our approach increase the diversity of the segmentation models, and the features of the different decoders reduce overfitting and improve performance. During training, these decoders update the entire network simultaneously, and only the main decoder is used for inference.

The segmentation network is initially trained in a supervised manner using the available labeled data. The auxiliary decoder receives different prediction maps under a feature perturbation, and the encoder and the two decoders are trained under supervised training. This approach allows us to improve the performance of our model and increase its robustness. Specifically, the two decoders are each able to output a prediction map for pixel-level labeled data and weakly labeled data, with each prediction map being supervised by the corresponding labeled mask. To optimize our model, we define the pixel-level segmentation loss as follows:

$$\mathcal{L}_{\text{sup}}(y, y_l) = \mathcal{L}_{\text{bce}}(y, y_l) + \mathcal{L}_{\text{Dice}}(y, y_l), \tag{1}$$

where y is the prediction result, $\mathcal{L}_{\text{bce}}$ is the binary cross-entropy loss, and $\mathcal{L}_{\text{Dice}}$ is the Dice loss. The pixel-level segmentation loss is a measure of the difference between the prediction map produced by the decoder for pixel-level labeled data and the corresponding labeled mask. This loss is used to guide the training of the model and improve the accuracy of the model.

For the training with weakly labeled data, previous research has shown that CNNs can be directly trained using scribbles by minimizing partial cross-entropy loss [12, 33]. Specifically, the partial cross-entropy loss is a measure of the difference between the prediction map produced for weakly labeled data and the corresponding labeled mask. We define partial cross-entropy loss as follows:

$$\mathcal{L}_{\text{pCE}}(y, y_w) = -\sum_{c} \sum_{i \in y_w} log y_i^c, \tag{2}$$

where y_w represents a scribble annotation, and y_i^c is the predicted probability that pixel i belongs to class c. To make more efficient use of weakly labeled data, we employ traditional random walks to generate pseudo labels y_{rw} from scribble annotations [46]. These pseudo labels are then used as supervision to guide the training of the model. We define the weak supervision loss as follows:

$$\mathcal{L}_{\text{wseg}}^{'} = \mathcal{L}_{\text{bce}}(y, y_{rw}) + \mathcal{L}_{\text{Dice}}(y, y_{rw}). \tag{3}$$

Finally, we use the generated pseudo labels as supervision signals. We define the total weak supervision loss as follows:

$$\mathcal{L}_{\text{wseg}} = \mathcal{L}_{\text{pCE}} + \mathcal{L}_{\text{wseg}}^{'}, \tag{4}$$

where $\mathcal{L}_{\text{wseg}}$ is the weakly supervised loss. The weak supervision loss is used to optimize the training of the encoder and decoder and to improve the accuracy of the model when weakly labeled data is available.

To make use of unlabeled samples, we apply a prediction invariance on the decoder output. Specifically, the goal of training on unlabeled data is to minimize the consistency loss $\mathcal{L}_{\text{con}}$, which measures the difference between the output of the primary decoder and the output of the auxiliary decoder. The consistency loss $\mathcal{L}_{\text{con}}$ is calculated as:

$$\mathcal{L}_{\text{con}} = \mathcal{L}_{\text{d}}(P_{u1}, P_{u2}), \tag{5}$$

where $\mathcal{L}_{\text{d}}$ is the distance measure between the two output probability distributions. By adding consistency constraints to the encoder output, we are able to leverage the additional information provided by the unlabeled samples to further improve the performance of the model.

B. Mixed Consistency Training Strategy

To further improve the performance of our model, we use the outputs of the two decoders for mixed consistency based on the dual-branch segmentation network. Specifically, inspired by mixup [28], the prediction P_l of the pixel-level annotation data X_l and the prediction P_w of the weakly-labeled data X_w are dynamically mixed to generate mixed prediction results. This allows us to leverage the complementary information provided by the two decoders to improve the accuracy of the model. The mixed predictions P_{lw1} and P_{lw2} are generated from the output of the separate decoders at the same time. These mixed predictions are defined as follows:

$$P_{lw1} = \alpha \times P_{l1} + (1-\alpha) \times P_{w1}, \tag{6}$$

$$P_{lw2} = \alpha \times P_{l2} + (1-\alpha) \times P_{w2}, \tag{7}$$

where α is randomly generated at (0,1) in each iteration. Our strategy of generating mixed predictions from the output of decoders improves the diversity of the mixture results and forces the network to train through a complex of prediction results. The consistency loss $\mathcal{L}^{'}_{\text{lwcon}}$ is as follows:

$$\mathcal{L}^{'}_{\text{lwcon}} = \mathcal{L}_d \left(P_{lw1}, P_{lw2}\right). \tag{8}$$

To fully utilize the weakly supervised data, we assume that the weakly supervised data $\mathcal{D}_w$ is unlabeled and mix the output of the same decoder where the input is unlabeled data and weakly labeled data.

Similarly, in each training iteration, we sample an equal number of weakly labeled and unlabeled samples and mix the outputs of the two decoders for these samples. Specifically, we obtain the mixed outputs P_{wu1} and P_{wu2} by following Eqs. 6 and 7, and then perform consistency regularization. The consistency loss $\mathcal{L}^{'}_{\text{wucon}}$ is defined as follows:

$$\mathcal{L}^{'}_{\text{wucon}} = \mathcal{L}_{\text{d}} \left(P_{wu1}, P_{wu2}\right), \tag{9}$$

where $\mathcal{L}_{\text{d}}$ is the distance measure between the two output probability distributions. This consistency loss is used to optimize the training of the encoder and decoders and improve the accuracy of the model by leveraging the complementary information provided by the weakly labeled and unlabeled data.

C. Mixed Supervision with Mixed Labels

We leverage the output of the two decoders to further optimize the training of the model. We generate pseudo labels by dynamically mixing the predictions of two models. First, we use the predictions from unlabeled data to generate pseudo labels and the dynamically mixed supervised strategy is defined as:

$$Y'_{un} = argmax[\alpha \times P_{u1} + (1-\alpha) \times P_{u2}]. \tag{10}$$

Then, we use the same strategy of Eq. 6 to generate mixed labels Y_{lw} from pixel-level labels and weak labels to obtain mixed annotations. We use the mixed labels for supervised training, where the mixed labels Y_{lw} are defined as follows:

$$Y_{lw} = \alpha \times y_l + (1-\alpha) \times y_w. \tag{11}$$

By using a dynamically mixed supervised strategy, we are able to obtain an increased diversity of predictions and labels, which helps to improve the robustness and generalization of the model. We define the mixed supervision loss as follows:

$$\mathcal{L}_{\text{mixsup}} = 0.5 \times \left(\mathcal{L}_{\text{Dice}} \left(P_{lw1}, Y_{lw}\right) + \mathcal{L}_{\text{Dice}} \left(P_{lw2}, Y_{lw}\right)\right), \tag{12}$$

$$\mathcal{L}^{'}_{\text{mixsup}} = 0.5 \times \left(\mathcal{L}_{\text{Dice}} \left(P_{u1}, Y^{'}_{un}\right) + \mathcal{L}_{\text{Dice}} \left(P_{u2}, Y^{'}_{un}\right)\right), \tag{13}$$

$$\mathcal{L}_{\text{mix}} = \mathcal{L}_{\text{mixsup}} + \mathcal{L}^{'}_{\text{mixsup}}, \tag{14}$$

where the mixed supervision loss $\mathcal{L}_{\text{mixsup}}$ is used to constrain the mixed prediction results with mixed labels within the respective decoder, while the mixed supervision loss $\mathcal{L}^{'}_{\text{mixsup}}$ is used to constrain the mixed prediction results between different decoders. The total mixed supervision loss $\mathcal{L}_{\text{mix}}$ is the sum of these two losses, and $\mathcal{L}_{\text{Dice}}$ is the widely used Dice loss. This mixed supervision strategy improves the accuracy of the model by constraining the mixed prediction results with mixed labels.

In this work, we use the mean squared error (MSE) as the distance measure, and the unsupervised loss is defined as:

$$\mathcal{L}_{\text{u}} = \mathcal{L}_{\text{con}} + \mathcal{L}^{'}_{\text{lwcon}} + \mathcal{L}^{'}_{\text{wucon}}. \tag{15}$$

Finally, the semi- and weakly-supervised segmentation loss function $\mathcal{L}_{\text{s}}$ is defined as follows:

$$\mathcal{L}_{\text{s}} = \lambda_1 \mathcal{L}_{\text{sup}} + \lambda_2 \mathcal{L}_{\text{mix}} + \lambda_3 \mathcal{L}_{\text{wseg}} + \lambda_4 L_{\text{u}}. \tag{16}$$

This loss function combines the various loss terms discussed above to optimize the training of the semi- and weakly-supervised segmentation model. The final loss function is a weighted sum of these loss terms, with the weights λ_1, λ_2, λ_3, and λ_4 controlling the relative importance of each term.

IV. Experiments

A. Datasets

We evaluated the performance of our propose semi- and weakly-supervised segmentation method on two public medical image segmentation datasets: ACDC and MSCMRSeg. These datasets provide a diverse set of images with various types of annotations, including pixel-level labels, weak labels, and unlabeled data.

ACDC dataset [47]: This dataset contains 200 annotated short-axis cardiac cine-MRI images of 100 patients. For each patient, manual annotations of right ventricle (RV), left ventricle (LV), and myocardium (MYO) are provided for both the end-diastolic (ED) and end-systolic (ES) phase. [48] manually provided scribble annotations for each scan. We randomly divided the dataset with a ratio of 7:1:2 to get the training set, the validation set, and the test set.

MSCMRseg dataset [49, 50]: This dataset provides 45 multi-sequence CMR images from patients who underwent cardiomyopathy, which is more difficult for automated segmentation than unenhanced cardiac MRI. Each patient has been scanned using the three CMR sequences (i.e., LGE, T2, and bSSFP), and there are standard segmentations annotations. [31] manually provide scribble annotations for each scan of

training. We divide the data into training, validation, and test sets according to the [31] setting, where 25 patient data are used for training, 5 patient data for validation, and 15 patient data for testing.

All images of both datasets are resized to 256 × 256 as network input. For training, we first rescale the intensity of each slice to 0-1, then we augment the training set with standard data augmentation methods including random rotation and random flipping. We use 2D slices for training, and generate predictions and stack them into a 3D volume for testing.

B. Implementation Details

We extend the basic U-Net to a dual-branch network by adding an auxiliary decoder, and a feature noise layer before each upsampling layer of the auxiliary decoder to introduce perturbation. We implement and run our proposed method and other comparison experiments using PyTorch [51] on a GeForce RTX 2080 TI GPU. The learning rate is initialized to 0.005, and the learning strategy is the warm-up strategy with a cosine scheduler. The proposed semi- and weakly-supervised segmentation model and all baselines are trained using the SGD optimizer with a mini-batch size of 20, where the weight decay hyperparameter in SGD is set to 0.0001. In our proposed semi- and weakly-supervised segmentation model, the weights λ_1, λ_2, λ_3 and λ_4 in the loss function are 1.0, 0.5, 0.5, and 0.5, respectively.

C. Evaluation Metrics

To evaluate the segmentation performances of our proposed SW-Seg and the state-of-art baselines, four widely used segmentation evaluation metrics, *positive predictive value (PPV)*, *sensitivity (Sens)*, *mean Intersection over Union (mIoU)*, and *dice similarity coefficient (DSC)*, are adopted. The formal definitions of DSC, PPV, Sens, and mIoU are as follows:

$$PPV = \frac{TP}{TP+FP}, \qquad Sens = \frac{TP}{TP+FN},$$
$$DSC = \frac{2 \times PPV \times Sens}{PPV + Sens} = \frac{2TP}{2TP+FP+FN},$$
$$mIoU = \frac{1}{K+1}\sum_{i=0}^{k}\frac{TP}{TP+FP+FN},$$

where TP (i.e., true positive) is the number of positive pixels (i.e., pixels within the annotated organ or tumor areas) that are correctly classified in the segmenting results, FP (i.e., false positive) is the number of negative pixels that are incorrectly classified as positive pixels, and FN (i.e., false negative) is the number of positive pixels that are incorrectly classified as negative pixels. Specifically, sensitivity is the ratio of correctly segmented positive pixels to all pixels annotated as positive in the base fact. The positive predictive value is a statistical measure used to evaluate the accuracy of a diagnostic test. It is calculated as the ratio of true positive results to the total number of positive results, including both true positive and false positive results. The Dice similarity coefficient, which is the summed mean of PPV and Sens, allows a more comprehensive evaluation of the model performance from both PPV and Sens perspectives. The mean intersection over union calculates the mean of the ratio of intersection and concatenation of all categories. A higher mIoU score indicates better performance, as it means that the model is able to produce more accurate segmentation masks that overlap more closely with the ground-truth masks.

D. Main Results

We present the results for the proposed semi- and weakly-supervised segmentation model on the ACDC and MSCMRSeg datasets in Table I, which shows the average performance of the three classes in the test set. To evaluate the performance of our proposed method, we compare it with fully supervised baselines. We train a U-Net model using 100%, 20%, and 10% of the training data as the upper bound and two baselines, and present the results of these fully supervised baselines for comparison. To further evaluate the performance of our semi- and weakly-supervised segmentation model, we compare it with eight state-of-the-art semi-supervised methods: DAN [52], mean teacher (MT) [53], uncertainty-aware mean teacher (UAMT) [20], cross-consistency training (CCT) [23], EM [14], interpolation consistency training (ICT) [22], cross pseudo supervision (CPS) [9], and URPC [15].

Semi-supervised methods demonstrate their effectiveness in utilizing unlabeled data to achieve a good performance compared to the baseline methods. This suggests that by combining the strengths of both labeled and unlabeled data, these methods are able to achieve impressive results in training. For example, when using only 10% of the data with pixel-level labels, the URPC method achieved a DSC of 85.37% on the ACDC dataset and a DSC of 73.67% on the MSCMRSeg dataset. These results represent a significant improvement of 10.74% and 19.13%, respectively, over the U-Net method. They demonstrate the effectiveness of the semi-supervised method in leveraging a small amount of labeled data to achieve improved performance. However, we note that while semi-supervised methods achieve a good performance, they still rely on a certain proportion of pixel-wise annotations. The cost of labeling 10% of these pixel-level labeled data is equivalent to the cost of labeling the entire weakly labeled dataset, which is still very time-consuming. Our proposed approach presents a promising solution to the challenge of lacking pixel-level annotations by allowing the replacement of some of these annotations with scribble annotations. This substitution not only reduces the quantity of pixel-level annotations needed, but also lowers the overall cost of annotation, making it a more efficient and cost-effective approach for training a model. The SW-Seg method demonstrates a superior performance compare to traditional semi-supervised methods using a 10% pixel-level annotation scale when using a combination of 5% pixel-level annotations, 5% weak annotations, and 90% unlabeled data on both datasets. This suggests that our method is able to effectively leverage the strengths of both weak and unlabeled data to improve model performance. For example, in the ACDC dataset, using 5% pixel-wise annotations and 5% weak annotations exceeds the performance of most of the methods using 10% pixel-wise labeled data, and the 3D DSC reaches 85.66%, exceeds the performance of the state-of-the-art semi-supervised methods. The cost of 5% weak

TABLE I
RESULTS OF SEMI- AND WEAKLY-SUPERVISION WITH DYNAMICALLY MIXED CONSISTENCY AND MIXED PSEUDO LABELS AND THE STATE-OF-THE-ART FULLY-SUPERVISED AND SEMI-SUPERVISED SEGMENTATION METHODS. THE SEGMENTATION PERFORMANCE IS EVALUATED ON THE ACDC AND MSCMRSEG DATASETS, USING DSC, SENS, PPV, AND MIOU AS EVALUATION METRICS.

Methods		ACDC				MSCMR			
		DSC	Sens	PPV	mIoU	DSC	Sens	PPV	mIoU
10%	U-Net [45]	0.7463	0.6977	0.8505	0.6295	0.5454	0.5418	0.6414	0.4100
	DAN [52]	0.7613	0.7430	0.8254	0.6501	0.6839	0.6685	0.7588	0.5483
	MT [53]	0.7855	0.7592	0.8592	0.6823	0.6607	0.6374	0.7494	0.5203
	UAMT [20]	0.8046	0.7939	0.8515	0.6967	0.6801	0.6626	0.7534	0.5378
	EM [14]	0.8250	0.8174	0.8556	0.7213	0.6924	0.6876	0.7323	0.5551
	CCT [23]	0.8524	0.8627	0.8532	0.7541	0.7258	0.7046	0.7896	0.5912
	ICT [22]	0.8389	0.8452	0.8557	0.7397	0.6796	0.6829	0.7267	0.5397
	CPS [9]	0.8436	0.8397	**0.8661**	0.7440	0.6909	0.6822	0.7561	0.5523
	URPC [15]	0.8537	0.8542	0.8649	0.7554	0.7367	0.6979	**0.8209**	0.6081
	Ours (5%-5%)	**0.8566**	**0.8805**	0.8497	**0.7590**	**0.7490**	**0.7172**	0.8177	**0.6159**
20%	U-Net [45]	0.8161	0.7822	0.8818	0.7122	0.7224	0.6983	0.8095	0.5904
	DAN [52]	0.8256	0.8174	0.8505	0.7194	0.7443	0.7506	0.7821	0.6137
	MT [53]	0.8536	0.8340	0.8853	0.7596	0.7599	0.7260	0.8344	0.6365
	UAMT [20]	0.8574	0.8389	**0.8936**	0.7676	0.7401	0.7465	0.7871	0.6075
	EM [14]	0.8353	0.8068	0.8817	0.7316	0.7438	0.7159	0.8265	0.6227
	CCT [23]	0.8468	0.8270	0.8804	0.7532	0.7639	0.7255	0.8431	0.6378
	ICT [22]	0.8453	0.8253	0.8787	0.7465	0.7713	0.7735	0.7892	0.6493
	CPS [9]	0.8609	0.8462	0.8865	0.7702	0.7681	0.7288	0.8433	0.6474
	URPC [15]	0.8640	0.8552	0.8919	0.7773	0.7528	0.7348	0.8183	0.6271
	Ours (10%-10%)	**0.8727**	**0.8950**	0.8659	**0.7821**	**0.7981**	**0.7754**	**0.8440**	**0.6760**
100%	UNet [45]	**0.9091**	**0.9289**	**0.8958**	**0.8393**	**0.8215**	**0.8495**	**0.8108**	**0.7063**

annotation is much lower than the cost of 5% pixel-wise annotation. In the case of medical images, it takes only a few minutes to label a scribble annotation, while it can take hours to complete a pixel-wise annotation. Therefore, our proposed method not only improves segmentation accuracy but also reduces labeling costs. On the MSCMRSeg dataset, our proposed semi- and weakly-supervised method demonstrate a superior performance compared to traditional semi-supervised methods.

For example, our method achieves the DSC of 74.90%, PPV of 81.77%, and mIoU of 61.59% under the ratio of 5% pixel-wise annotations, 5% weakly annotations, and 90% unlabeled data, surpassing all semi-supervised methods, and compared with the URPC in the semi-supervised methods, DSC (74.90%) improves by 1.23%, and Sens (71.72%) improves by 1.93%.

Furthermore, by increasing the scale of annotations to 10% pixel-level and 10% weak labeled scans, our method can outperform all semi-supervised methods using 20% pixel-level labeled scans on ACDC and MSCMRSeg. For example, our method achieves DSC of 87.27%, Sens of 89.50%, and mIoU of 78.21% on the ACDC dataset, outperforming all semi-supervised methods using 20% pixel-level labeled scans. This demonstrates the effectiveness of our proposed method in utilizing mixed labeled data to improve the performance of medical image segmentation.

Additionally, Fig. 2 presents a visualization of the segmentation results of all methods when trained with 10% labeled data on the ACDC and MSCMRSeg datasets. It can be observed that these methods generally achieve satisfactory segmentation results for foreground regions in the ACDC dataset. However, on the more challenging MSCMRSeg dataset, our method exhibits a superior prediction accuracy and fewer misclassifications. We find no existing methods in similar to our segmentation results on both datasets. For example, the results of CPS have good segmentation in ACDC but slightly worse segmentation results in the MSCMRSeg dataset. CCT performs slightly worse compared to our method on both datasets.

Different training strategies have a significant impact on the segmentation results. Specifically, the segmentation results of ACDC and MSCMRSeg show that (i) the segmentation results of MT using simple consistency constraints outperform the models trained by adversarial strategies, (ii) the segmentation results of methods trained using pseudo-supervised training methods and uncertainty rectified pyramid cconsistency outperform others, and (iii) the segmentation performance of SW-Seg outperforms the semi-supervised learning method baseline.

Although there are some algorithms with similar results to ours on specific datasets, as our work is based on segmentation with the addition of partially weakly labeled data, our goal is to obtain better segmentation results while further reducing the labeling dependency. Compared with the current state-of-the-art semi-supervised methods, our results are able to retain more details while reducing the proportion of pixel-level labeled data, further demonstrating the effectiveness of our approach.

E. Ablation Study

In order to assess the relative contribution of each component of our proposed semi- and weakly-supervised image segmentation method, we conduct an ablation study on the ACDC and MSCMRSeg datasets. The results, presented in

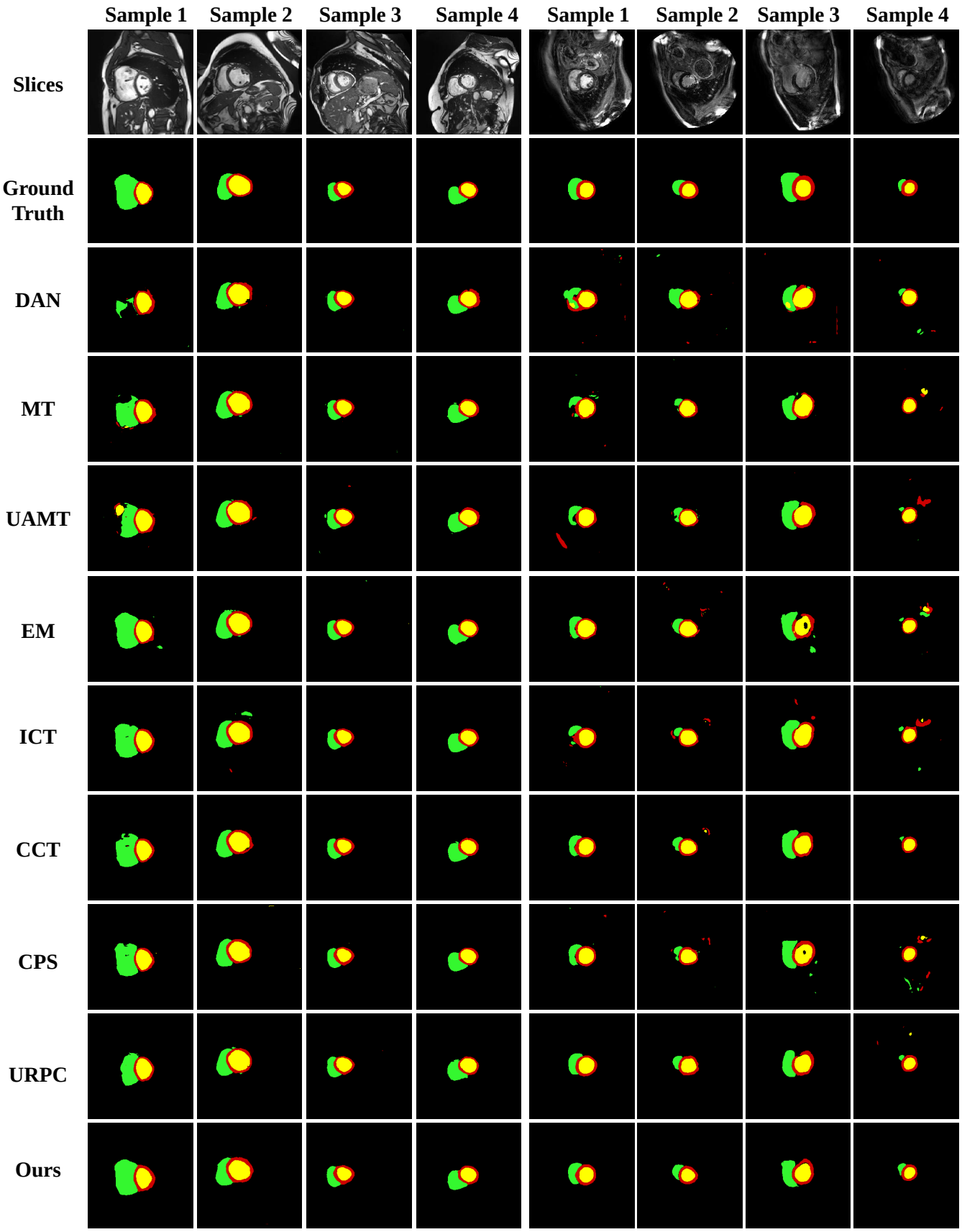

Fig. 2. Visual segmentation results for ACDC and MSCMRSeg with 10% labeled data. In the first four columns are two paired samples from the ACDC dataset. In the last four columns are two paired samples from the MSCMRSeg dataset, which are marked in yellow, red, and green, respectively, in the segmentation results. As observed, our segmentation results are closer to the ground truth and outperform other methods on both datasets.

TABLE II
ABLATION STUDY OF OUR PROPOSED METHOD ON ACDC AND MSCMRSEG WITHOUT AND WITH PSEUDO LABELING FOR WEAK SUPERVISORY LOSS $\mathcal{L}^{'}_{\text{WSEG}}$, MIXED SUPERVISORY LOSS L_{MIXUP}, AND MIXED CONSISTENCY LOSS $\mathcal{L}^{'}_{\text{LWCON}}$ AND $\mathcal{L}^{'}_{\text{WUCON}}$.

Ratios	Methods	ACDC				MSCMRSeg			
	Metrics	DSC	Sens	PPV	mIoU	DSC	Sens	PPV	mIoU
10%	U-Net [45]	0.7463	0.6977	0.8505	0.6295	0.5454	0.5418	0.6414	0.4100
5%-5%	DDH	0.7642	0.7486	0.8514	0.6547	0.5777	0.5221	0.7234	0.4320
	DDH + $\mathcal{L}^{'}_{\text{wseg}}$	0.7817	0.7641	0.8539	0.6650	0.6386	0.6574	0.6835	0.4879
	DDH + $\mathcal{L}^{'}_{\text{wseg}}$ + $\mathcal{L}_{\text{mix}}$	0.8179	0.8193	0.8453	0.7070	0.6869	0.6433	0.7659	0.5428
	DDH + $L^{'}_{\text{wseg}}$ + $\mathcal{L}_{\text{mix}}$ + $\mathcal{L}^{'}_{\text{lwcon}}$	0.8297	0.8342	0.8451	0.7202	0.7105	0.6851	0.7753	0.5698
	DDH + $\mathcal{L}^{'}_{\text{wseg}}$ + $\mathcal{L}_{\text{mix}}$ + $\mathcal{L}^{'}_{\text{lwcon}}$+ $\mathcal{L}^{'}_{\text{wucon}}$	0.8566	0.8805	0.8497	0.7590	0.7490	0.7172	0.8177	0.6159
20%	U-Net [45]	0.8161	0.7822	0.8818	0.7122	0.7224	0.6983	0.8095	0.5904
10%-10%	DDH	0.8465	0.8620	0.8526	0.7463	0.7354	0.6830	0.8466	0.6047
	DDH + $\mathcal{L}^{'}_{\text{wseg}}$	0.8531	0.8660	0.8616	0.7553	0.7665	0.7144	0.8565	0.6393
	DDH + $\mathcal{L}^{'}_{\text{wseg}}$ + $\mathcal{L}_{\text{mix}}$	0.8582	0.8649	0.8662	0.7607	0.7726	0.7298	0.8498	0.6469
	DDH + $\mathcal{L}^{'}_{\text{wseg}}$ + $\mathcal{L}_{\text{mix}}$ + $\mathcal{L}^{'}_{\text{lwcon}}$	0.8613	0.8863	0.8526	0.7670	0.7838	0.7476	0.8489	0.6591
	DDH + $\mathcal{L}^{'}_{\text{wseg}}$ + $\mathcal{L}_{\text{mix}}$ + $\mathcal{L}^{'}_{\text{lwcon}}$+ $\mathcal{L}^{'}_{\text{wucon}}$	0.8727	0.8950	0.8659	0.7821	0.7981	0.7754	0.8440	0.6760

TABLE III
ABLATION STUDY RESULTS WITH 10% LABELED DATA USING AN AUXILIARY DECODER OF DIFFERENT PERTURBATIONS

Methods	ACDC				MSCMR			
	DSC	Sens	PPV	mIoU	DSC	Sens	PPV	mIoU
With FeatureDropout	0.8404	0.8491	**0.8577**	0.7362	0.7421	**0.7228**	0.8017	0.6076
With Dropout	0.8415	0.8508	0.8522	0.7386	0.7220	0.7179	0.7588	0.5831
With FeatureNosie	**0.8566**	**0.8805**	0.8497	**0.7590**	**0.7490**	0.7172	**0.8177**	**0.6159**

Table II, show the effectiveness of each component in improving the segmentation accuracy of our model. In our ablation study, we demonstrate the effectiveness of each component in our proposed method by gradually incorporating them into the basic model, which is a dual-decoder network (DDH). DDH extends U-Net to a dual-branch architecture with an auxiliary decoder θ_2, and uses pixel-wise annotations and weak annotations for supervised training, as well as consistency constraints for unsupervised training. To optimize the model further, we introduce four strategies: $\mathcal{L}^{'}_{\text{wseg}}$, $\mathcal{L}_{\text{mix}}$, $\mathcal{L}^{'}_{\text{lwcon}}$, and $\mathcal{L}^{'}_{\text{wucon}}$.

As depicted in Table II, the DSC of DDH with a ratio of 5% pixel-wise annotations, 5% weak annotations, and 90% unlabeled data outperforms U-Net with the ratio of 10% pixel-wise annotations and 90% unlabeled data on both datasets. This demonstrates that DDH can effectively exploit mixed types of data, while U-Net is restricted by solely using labeled data. Second, we introduce the $\mathcal{L}^{'}_{\text{wseg}}$, and the performance improves by 1.75% compared to DDH on the ACDC dataset under the same annotation ratio (5%-5%). On the MSMCRSeg dataset, it achieves the DSC of 63.86% and mIoU of 48.79%. Because by optimizing $\mathcal{L}^{'}_{\text{wseg}}$, we better use pseudo labels for weakly supervised training.

Thirdly, $\mathcal{L}_{\text{mix}}$ introduces mixed labels to enhance the feature learning capabilities of the segmentation network. On the ACDC dataset, DSC (81.79%) improves by 3.62%, and Sens (81.93%) improves by 5.52%. Subsequently, we conduct further experiments to examine the effect of using the loss function $\mathcal{L}^{'}_{\text{lwcon}}$ on the segmentation performance. As demonstrated in Table II, the performance is enhanced with the incorporation of $\mathcal{L}^{'}_{\text{lwcon}}$. This is due to the fact that the predictions made using labeled data and weakly labeled data are blended after the consistency constraint is applied, thereby augmenting the information mining for labeled data.

Finally, by introducing $\mathcal{L}^{'}_{\text{wucon}}$ leads to a notable improvement in performance. This can be attributed to the reason that the consistency constraint applied to the mixture of predictions made with labeled data and weakly labeled data further enhances the information extracted from labeled data. The introduction of $\mathcal{L}^{'}_{\text{wucon}}$ leads to a 2.69% increase in the DSC with a ratio of 5% pixel-wise annotations and 5% weak annotations on the ACDC dataset, and a 3.85% improvement in DSC on the MSCMRSeg dataset.

F. More Experiments

The effect of adding different perturbations to the auxiliary decoder. In order to further enhance the robustness of our model, we utilize the output of the auxiliary decoder and apply various perturbations to the hidden representations during the consistency training phase. Previous research has suggested several effective perturbation techniques [23], such as adding noise to the features extracted from the encoder, randomly discarding certain activations of the encoder output feature map, or applying spatial dropout [54] as a random perturbation. By introducing these types of perturbations during training, we can help the model to learn more robust and generalizable representations, resulting in an improved performance on the target task. We conduct additional ablation experiments using 10% labeled data on both the ACDC and MSCMRSeg datasets. The results are summarized in Table III. We compare the results of adding three different types of perturbations to the auxiliary decoder, and found that the auxiliary decoder performed best when we added FeatureNoise perturbation. Specifically, on the ACDC dataset, the DSC was 1.43% higher and the mIoU was 1.72% higher when using FeatureNoise perturbation compared to using Featureout

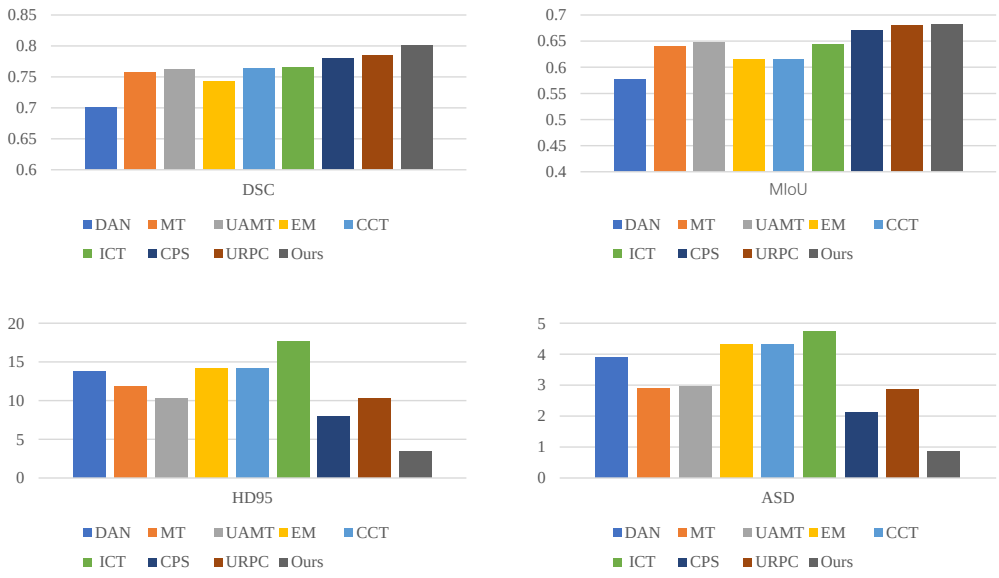

Fig. 3. Results of our proposed SW-Seg method and state-of-the-art semi-supervised segmentation methods on the ACDC and MSCMRSeg datasets. The semi-supervised method uses 5% labeled data and 95% unlabeled data, and our proposed method uses 2% of pixel-level labeled data, 3% of weakly annotated data, and 95% unlabeled data, where DSC, mIoU, HD95, and ASD are used as evaluation metrics.

perturbation. Based on these results, we use FeatureNoise perturbation in all our subsequent experiments.

Effectiveness of our proposed method is verified by further reducing the annotation ratio to 5%. To verify the effectiveness at a lower annotation ratio, where our proposed method uses 2% of pixel-level standard data, 3% of weakly annotated data, and 95% of unlabeled data, the state-of-the-art semi-supervised segmentation method uses 5% of annotated and 95% of unlabeled data, and Fig. 3 illustrates its comparison results. We see that our method achieves a better performance and results at a lower ratio of annotation compared to the semi-supervised method using 5% annotation. This also proves that at a very low annotation ratio, our method also effectively utilizes the information from both pixel-level labeled and unlabeled data to further improve the performance of the model.

Training with other losses to verify the effectiveness of the method. To verify the effectiveness of mixing consistency constraints on predictions with labeled data, i.e., using Eq. 8, a consistency constraint is applied. We directly subject the predictions P_{l1} and P_{l2}, and P_{w1} and P_{w2} of the two decoders to the consistency constraint separately, without the mixed prediction constraint. We define a new consistency loss as follows:

$$\mathcal{L}_{\text{lwcon}}^{''} = \mathcal{L}_{\text{d}}\left(P_{l1}, P_{l2}\right) + \mathcal{L}_{\text{d}}\left(P_{w1}, P_{w2}\right). \tag{17}$$

We compare the segmentation results of applying this consistency loss $\mathcal{L}_{\text{lwcon}}^{''}$, our propose mixed consistency loss $\mathcal{L}_{\text{lwcon}}^{'}$ and without adding consistency loss $\mathcal{L}_{\text{lwcon}}^{'}$ on the ACDC and MSCMRSeg datasets, and the results are shown in Table IV. The results show that training the model directly using the consistency loss $\mathcal{L}_{\text{lwcon}}^{''}$ on labeled data has improved the performance over training the model without consistency loss, indicating that using consistency to constrain the prediction results of labeled data is effective, but the improvement is insufficient, because supervised training using labeled data is an effective and accurate model training method, and the model using accurate labels to learn is a strong constraint. Although training with consistency constraints also improves the model, this improvement tends to be minimal compared to supervised training. Therefore, we build on this by imposing the consistency loss on the mixture predictions of labeled data, and further exploit the diversity of the mixture prediction results to mine the underlying information of the labeled data. Compared with the direct use of the consistency loss $\mathcal{L}_{\text{lwcon}}^{''}$, the proposed $\mathcal{L}_{\text{lwcon}}^{'}$ has a significant advantage in segmentation performance. It shows that our proposed mixed consistency approach effectively utilizes the information from the labeled data, thus improving the performance of the model.

Similarly, to verify the effectiveness of mixed supervision, we deform Eq. 11 by using y_l and y_{rw} to generate mixed annotations Y_{lw} and perform mixed supervised training, i.e., we define mixed annotations $Y_{lw}^{'}$ as follows:

$$Y_{lw}^{'} = \alpha \times y_l + (1-\alpha) \times y_{rw}, \tag{18}$$

and we define the mixed supervised loss $L_{\text{mixsup}}^{''}$ as follows:

$$\mathcal{L}_{\text{mixsup}}^{''} = \mathcal{L}_{\text{Dice}}(P_{lw1}, Y_{lw}^{'}) + \mathcal{L}_{\text{Dice}}(P_{lw2}, Y_{lw}^{'}). \tag{19}$$

We compared the model performance using different mixed label training, and the results are in Table IV. We see that the segmentation results using Y_{rw} to generate mixed labels for training are lower than those using the original scribbled labels. This is because the inclusion of pseudo labels in the mixture training process carries with it the risk of introducing inaccuracies, which can ultimately hinder the performance of the model.

Sensitivity of our method to different values of α. To further verify the effectiveness of our proposed method, we conduct experiments on the MSCMRSeg and ACDC datasets to assess the sensitivity of our method to different values of α,

TABLE IV
RESULTS OF TRAINING WITH OTHER LOSSES ON ACDC AND MSCMR.

Methods	ACDC				MSCMR			
	DSC	Sens	PPV	mIoU	DSC	Sens	PPV	mIoU
Without $\mathcal{L}'_{\text{lwcon}}$	0.8427	0.8438	0.8606	0.7410	0.7220	0.7179	0.7588	0.5831
With $\mathcal{L}''_{\text{lwcon}}$	0.8510	0.8616	0.8600	0.7501	0.7260	0.7189	0.7602	0.5871
With $\mathcal{L}'_{\text{lwcon}}$	0.8566	0.8805	0.8497	0.7590	0.7490	0.7172	0.8177	0.6159
With $\mathcal{L}''_{\text{mixsup}}$	0.8476	0.8660	0.8488	0.7464	0.6948	0.7149	0.7425	0.5545
With $\mathcal{L}_{\text{mixsup}}$	0.8566	0.8805	0.8497	0.7590	0.7490	0.7172	0.8177	0.6159

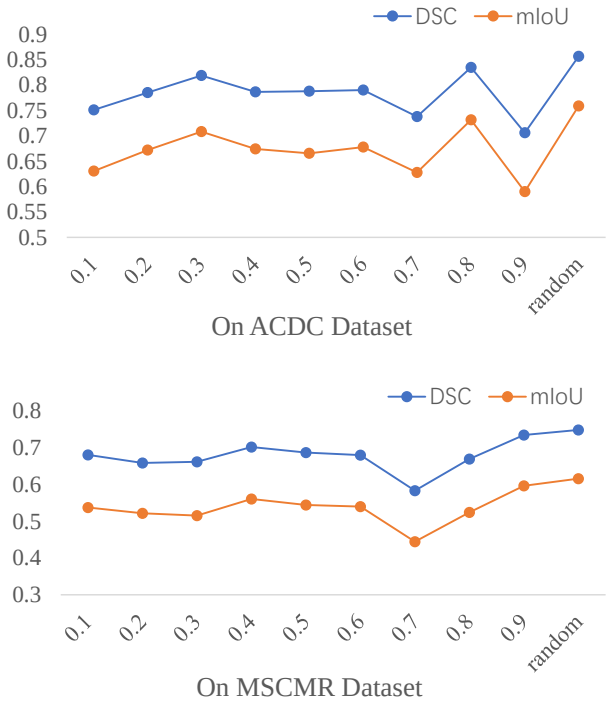

Fig. 4. Effect of α sensitivity on mixed consistency and mixed pseudo label training.

TABLE V
THE HD95 AND ASD RESULTS OF OUR METHOD AND SEMI-SUPERVISED BASELINES ON ACDC AND MSCMR DATASET WITH 10% LABELED DATA

Methods	ACDC		MSCMR	
	HD95	ASD	HD95	ASD
DAN [52]	8.3854	2.1119	35.599	9.6855
MT [53]	**5.2567**	2.5282	51.907	13.352
UAMT [20]	7.3121	2.0560	40.064	11.577
EM [14]	10.346	2.7575	60.743	16.343
CCT [23]	11.604	3.1939	30.285	9.1485
ICT [22]	11.227	3.3037	65.114	16.683
CPS [9]	8.3929	2.2740	63.903	17.606
URPC [15]	9.6594	2.4801	16.111	4.6320
Ours (5%-5%)	6.1444	**1.8958**	**15.331**	**4.0202**

TABLE VI
RESULTS FOR 10% OF THE ANNOTATED DATA WITH DIFFERENT PIXEL-LEVEL ANNOTATION AND WEAK ANNOTATION RATIOS.

Ratio(s:w)	ACDC		MSCMR	
	DSC	ASD	DSC	ASD
1% : 9%	0.7833	2.6808	0.6974	6.1868
2% : 8%	0.8166	2.2259	0.6999	8.6161
3% : 7%	0.8403	1.8437	0.7032	7.3469
4% : 6%	0.8437	2.0659	0.7298	5.4139
5% : 5%	0.8566	1.8958	0.7490	4.0202

ranging from 0.1 to 0.9. As shown in Fig. 4, we observed that the performance of our method varied with different values of α. Specifically, we observe that the performance of our method was consistently worse when using fixed values of α, compared to when we randomly chose α values. This is because dynamic mixing increases the diversity of predictions, avoiding the inherent weaknesses of the network [55]. This suggests that our method performs better when we do not set a fixed value for α during training.

Adding Boundary-Based Evaluation Metrics. To further illustrate the superior performance of our proposed SW-Seg method, we add two boundary-based evaluation metrics: Average Surface Distance (ASD) and 95% Hausdorff Distance (HD95). Average Surface Distance (ASD) is a measure of the distance between two surfaces in an image, such as the boundary of a segmentation mask and the ground-truth boundary. It is commonly used as an evaluation metric for image segmentation tasks and is calculated as the average distance between corresponding points on the two surfaces. Hausdorff Distance (HD) is a commonly used distance-based metric for evaluating the performance of image segmentation methods. To mitigate the impact of outliers on the evaluation, we use the variant of HD known as HD95. This helps to accurately reflect the overall performance of the segmentation method by eliminating the influence of a small subset of outliers. These metrics were calculated by averaging the 3D scans of each patient's 2D slices.

To fully demonstrate the superior segmentation performance of our proposed method, we compare its segmentation results to those of all state-of-the-art semi-supervised methods. As can be seen in Table V, our method performs better than the baseline method on both datasets. Our proposed method approaches the state-of-the-art in terms of segmentation performance on the ACDC dataset. On the MSCMRSeg dataset, our method achieves impressive results, with ASD of 4.0202 and HD95 of 15.331. These results outperform the performance of current semi-supervised methods, thus again proving our conclusion that our method outperforms the state-of-the-art semi-supervised segmentation methods and achieves a higher performance at a lower labeling cost.

Training is performed using different ratios of pixel-level and weakly labeled data and unlabeled data. To ensure a comprehensive evaluation of our method, we conduct experiments using various proportions of labeled data and the

results are presented in Table VI, where s and w represent the proportion of pixel-level labeled data and the proportion of weakly labeled data, respectively. These experiments allow us to evaluate the performance of our method under a range of label scales and gain a good understanding of its capabilities. We use 10% labeled data and 90% unlabeled data for training, where the pixel-level labeled data in the labeled data is gradually reduced to 1% and the weakly labeled data is gradually increased to 9%, and test their segmentation performance.

As shown in the results, the performance of our method using a combination of mixed labeled data and unlabeled data decreases gradually as the proportion of pixel-level labeling decreases. However, even at a low proportion of 1% pixel-level labeling, our method still outperforms the supervised training results of U-Net and some semi-supervised algorithms. When 5% pixel-level labeled data and 5% labeled data are used, the performance of the current semi-supervised algorithm is thus exceeded, further validating the effectiveness of our method.

V. Discussion and Future Works

A. Social Benefit of Proposed Framework

The wide range of clinical scenarios in which this model can be applied makes it a valuable solution for the medical image analysis process. We take radiotherapy for cardiovascular as an example: cardiovascular diseases are a pervasive and devastating health problem that afflicts millions of people around the world. For different types of etiologies, imaging physicians are required to segment the heart cavity and determine the area of the lesion. The use of automatic segmentation technology can greatly shorten the diagnosis time of doctors, but the existing semi-supervised segmentation algorithm method needs a considerable part of the labeled data for training to ensure the accuracy of the model, but it is also quite time-consuming to label this part of the data. Our proposed automatic segmentation method requires only a relatively small proportion of pixel-level annotations and scribble annotations for training, and can automatically generate segmentation results in seconds. It greatly reduces the cost of labeling medical images and the workload for doctors, and decreases the clinical application threshold and preparation period of high-precision intelligent medical image segmentation systems, which is of great significance to the clinical practice of computer-aided diagnosis and treatment.

More importantly, our new method of mixed supervised high-precision medical image segmentation based on pixel-level annotation and graffiti annotation data is proposed to break through the key technical bottleneck that restricts the clinical application of intelligent medical image segmentation system and accelerate the application process of artificial intelligence technology in the medical field. Our semi- and weakly-supervised segmentation solution requires the use of only a small amount of pixel-level labeled training data and a small amount of scribbled annotation data. As a result, this significantly reduces application requirements and increases the efficiency of deployment of automated medical image segmentation systems in clinical practice. At the same time, it can effectively assist doctors to complete a diagnosis, surgical simulation, and treatment planning, alleviating the shortage of medical staff and resources, saving time and money for patients, and it is expected to obtain good economic and social benefits.

B. Limitations and Future Works

Existing semi-supervised segmentation methods for medical images require a high proportion of pixel-level labeled data to obtain a better segmentation performance. Due to the lack of pixel-level annotations, their applications in clinical practice are usually limited. Our proposed method alleviates this problem to a great extent, by using weakly labeled data to partially replace the pixel-level labeled data while reducing the labeling costs and achieving better segmentation results. However, we cannot get a better performance with lower annotation costs for now. Therefore, potential future work is to further explore how to use smaller ratio pixel-level annotated data or only partially weakly annotated data [56] to ensure high-accuracy segmentation results. At the same time, the application of the model to unbalanced data conditions is studied [57], and image preprocessing [58] is performed to remove noise and artifacts.

Our experiments utilize scribble labeled data and pixel-level labeled data to train the model. In fact, there are some medical images in clinical practice where other labeling types can exist in clinical practice. For example, annotated data in the form of bounding boxes or in the form of points. We contend that a superior performance can also be attained through the utilization of other types of weak annotations in the training process. Thus, utilizing diverse segmentation annotations, our proposed solution for semi-weak mixed supervised segmentation can be trained, resulting in a substantial reduction in annotation costs.

Although our experiments are conducted on medical data, we believe that this learning model is not only applicable to medical images but can also be used to segment general images in daily life. We believe that it is possible to improve the model performance using our proposed mixed consistency strategy and mixed supervised learning approach, whether for classification, detection, or segmentation. At the same time, it is able to reduce the annotation cost by using pixel-level annotation and weakly annotated data.

Our semi- and weakly-supervised learning approach uses two main strategies to improve the accuracy of the model using pixel-level labeled data and scribble labeled data, which is fundamentally different from other semi-supervised learning strategies. This makes it possible to add other semi-supervised strategies to our approach. Since almost all semi-supervised learning strategies are designed using one data type, these designs can be integrated with our scheme.

VI. Conclusion

In this paper, we proposed a novel semi- and weakly-supervised segmentation framework for medical images, called SW-Seg, which can learn from a small amount of pixel-level labeled data, a small amount of weakly labeled data, as well as a large amount of unlabeled data. Mixed consistency and mixed supervision methods are used to make full use

of limited pixel-level annotations and weak annotations for training, which reduces the inconsistency of supervision and strengthens the consistency of the dual-branch outputs. While maintaining a high segmentation accuracy, partially replacing pixel-level labels with weak labels reduces the reliance on pixel-level annotation data in medical image segmentation tasks, and lowers the labeling costs. Extensive experimental studies are conducted on two publicly available medical image segmentation datasets, and the results show that the proposed semi- and weakly-supervised segmentation model outperforms existing semi-supervised methods and can achieve an exhilarating performance with lower labeling costs.

REFERENCES

[1] L. Wang, B. Wang, and Z. Xu, "Tumor segmentation based on deeply supervised multi-scale U-Net," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine*, 2019, pp. 746–749.

[2] Z. Xu, S. Liu, D. Yuan, L. Wang, J. Chen, T. Lukasiewicz, Z. Fu, and R. Zhang, "ω-net: Dual supervised medical image segmentation with multi-dimensional self-attention and diversely-connected multi-scale convolution," *Neurocomputing*, vol. 500, pp. 177–190, 2022.

[3] J. Zhang, S. Zhang, X. Shen, T. Lukasiewicz, and Z. Xu, "Multi-condos: Multimodal contrastive domain sharing generative adversarial networks for self-supervised medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. Early Access, pp. 1–20, 2023.

[4] S. Mo, M. Cai, L. Lin, R. Tong, Q. Chen, F. Wang, H. Hu, Y. Iwamoto, X.-H. Han, and Y.-W. Chen, "Mutual information-based graph co-attention networks for multimodal prior-guided magnetic resonance imaging segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 2512–2526, 2021.

[5] K. Wang, B. Zhan, C. Zu, X. Wu, J. Zhou, L. Zhou, and Y. Wang, "Tripled-uncertainty guided mean teacher model for semi-supervised medical image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 450–460.

[6] X. Luo, G. Wang, T. Song, J. Zhang, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, and S. Zhang, "MIDeepSeg: Minimally interactive segmentation of unseen objects from medical images using deep learning," *Medical Image Analysis*, vol. 72, p. 102102, 2021.

[7] S. Zhang, J. Zhang, B. Tian, T. Lukasiewicz, and Z. Xu, "Multi-modal contrastive mutual learning and pseudo-label re-learning for semi-supervised medical image segmentation," *Medical Image Analysis*, vol. 83, p. 102656, 2023.

[8] A. Tong, C. Tang, and W. Wang, "Semi-supervised action recognition from temporal augmentation using curriculum learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 3, pp. 1305–1319, 2022.

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

[10] J. Peng, G. Estrada, M. Pedersoli, and C. Desrosiers, "Deep co-training for semi-supervised image segmentation," *Pattern Recognition*, vol. 107, p. 107269, 2020.

[11] Y. Xia, D. Yang, Z. Yu, F. Liu, J. Cai, L. Yu, Z. Zhu, D. Xu, A. Yuille, and H. Roth, "Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation," *Medical Image Analysis*, vol. 65, p. 101766, 2020.

[12] A. Kumar, P. Sattigeri, and T. Fletcher, "Semi-supervised learning with gans: Manifold invariance with improved inference," *Proceedings of the Advances in Neural Information Processing Systems*, vol. 30, 2017.

[13] S. Li, C. Zhang, and X. He, "Shape-aware semi-supervised 3D semantic segmentation for medical images," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2020, pp. 552–561.

[14] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2517–2526.

[15] X. Luo, W. Liao, J. Chen, T. Song, Y. Chen, S. Zhang, N. Chen, G. Wang, and S. Zhang, "Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency," in *Medical Image Analysis*, 2022, p. 102517.

[16] Y. B. Can, K. Chaitanya, B. Mustafa, L. M. Koch, E. Konukoglu, and C. F. Baumgartner, "Learning to segment medical images with scribble-supervision alone," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 2018, pp. 236–244.

[17] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proceedings of the International Conference on Machine Learning Workshops*, vol. 3, no. 2, 2013, p. 896.

[18] Z. Xu, C. Qi, and G. Xu, "Semi-supervised attention-guided CycleGAN for data augmentation on medical images," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine*, 2019, pp. 563–568.

[19] C. Cao, T. Lin, D. He, F. Li, H. Yue, J. Yang, and E. Ding, "Adversarial dual-student with differentiable spatial warping for semi-supervised semantic segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[20] L. Yu, S. Wang, X. Li, C.-W. Fu, and P.-A. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019, pp. 605–613.

[21] X. Li, L. Yu, H. Chen, C.-W. Fu, L. Xing, and P.-A. Heng, "Transformation-consistent self-ensembling model for semisupervised medical image segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 523–534, 2020.

[22] V. Verma, K. Kawaguchi, A. Lamb, J. Kannala, Y. Bengio, and D. Lopez-Paz, "Interpolation consistency training for semi-supervised learning," *ArXiv*, vol. abs/1903.0382, 2019.

[23] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 674–12 684.

[24] S. Cheng, Y. Zhou, W. Zhang, D. Wu, C. Yang, B. Li, and W. Wang, "Uncertainty-aware and multigranularity consistent constrained model for semi-supervised hashing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 6914–6926, 2022.

[25] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in Neural Information Processing Systems*, vol. 33, pp. 596–608, 2020.

[26] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.

[27] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6023–6032.

[28] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," *arXiv*, vol. abs/1710.09412, 2017.

[29] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Medical Image Analysis*, vol. 63, p. 101693, 2020.

[30] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "ScribbleSup: Scribble-supervised convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3159–3167.

[31] K. Zhang and X. Zhuang, "CycleMix: A holistic strategy for medical image segmentation from scribble supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 656–11 665.

[32] G. Valvano, A. Leo, and S. A. Tsaftaris, "Self-supervised multi-scale consistency for weakly supervised segmentation learning," in *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*. Springer, 2021, pp. 14–24.

[33] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele, "Simple does it: Weakly supervised instance and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 876–885.

[34] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3136–3145.

[35] V. Kulharia, S. Chandra, A. Agrawal, P. Torr, and A. Tyagi, "Box2Seg: Attention weighted loss and discriminative feature learning for weakly supervised segmentation," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 290–308.

[36] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei, "What's the point: Semantic segmentation with point supervision," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 549–565.

[37] J. Qin, J. Wu, X. Xiao, L. Li, and X. Wang, "Activation modulation and recalibration scheme for weakly supervised semantic segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, 2022, pp. 2117–2125.

[38] J. Dong, Y. Cong, G. Sun, Y. Yang, X. Xu, and Z. Ding, "Weakly-supervised cross-domain adaptation for endoscopic lesions segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, 2020.

[39] X. Luo, M. Hu, W. Liao, S. Zhai, T. Song, G. Wang, and S. Zhang, "Scribble-supervised medical image segmentation via dual-branch network and dynamically mixed pseudo labels supervision," *arXiv*, vol. abs/2203.02106, 2022.

[40] X. Liu, Q. Yuan, Y. Gao, K. He, S. Wang, X. Tang, J. Tang, and D. Shen, "Weakly supervised segmentation of COVID19 infection with scribble annotation on CT images," *Pattern recognition*, vol. 122, p. 108341, 2022.

[41] G. Valvano, A. Leo, and S. A. Tsaftaris, "Learning to segment from scribbles using multi-scale adversarial attention gates," *IEEE Transactions on Medical Imaging*, vol. 40, no. 8, pp. 1990–2001, 2021.

[42] L. Sun, J. Wu, X. Ding, Y. Huang, Z. Chen, G. Wang, and Y. Yu, "A teacher-student framework for liver and tumor segmentation under mixed supervision from abdominal ct scans," *Neural Computing and Applications*, vol. 34, no. 19, pp. 16 547–16 561, 2022.

[43] D. Wang, M. Li, N. Ben-Shlomo, C. E. Corrales, Y. Cheng, T. Zhang, and J. Jayender, "Mixed-supervised dual-network for medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*. Springer, 2019, pp. 192–200.

[44] J. Dolz, C. Desrosiers, and I. B. Ayed, "Teach me to segment with mixed supervision: Confident students become masters," in *Information Processing in Medical Imaging: 27th International Conference, IPMI 2021, Virtual Event, June 28–June 30, 2021, Proceedings 27*. Springer, 2021, pp. 517–529.

[45] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.

[46] L. Grady, "Random walks for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 11, pp. 1768–1783, 2006.

[47] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester *et al.*, "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?" *IEEE Transactions on Medical Imaging*, vol. 37, no. 11, pp. 2514–2525, 2018.

[48] G. Valvano, A. Leo, and S. A. Tsaftaris, "Learning to segment from scribbles using multi-scale adversarial attention gates," *IEEE Transactions on Medical Imaging*, vol. 40, no. 8, pp. 1990–2001, 2021.

[49] X. Zhuang, "Multivariate mixture model for cardiac segmentation from multi-sequence mri," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2016, pp. 581–588.

[50] ——, "Multivariate mixture model for myocardial segmentation combining multi-source images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 12, pp. 2933–2946, 2018.

[51] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "PyTorch: An imperative style, high-performance deep learning library," *Proceedings of the Advances in Neural Information Processing Systems*, vol. 32, 2019.

[52] Y. Zhang, L. Yang, J. Chen, M. Fredericksen, D. P. Hughes, and D. Z. Chen, "Deep adversarial networks for biomedical image segmentation utilizing unannotated images," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 408–416.

[53] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Proceedings of the Advances in Neural Information Processing Systems*, vol. 30, 2017.

[54] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 648–656.

[55] X. Huo, L. Xie, J. He, Z. Yang, W. Zhou, H. Li, and Q. Tian, "ATSO: Asynchronous teacher-student optimization for semi-supervised image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1235–1244.

[56] F. Gao, M. Hu, M.-E. Zhong, S. Feng, X. Tian, X. Meng, Z. Huang, M. Lv, T. Song, X. Zhang *et al.*, "Segmentation only uses sparse annotations: Unified weakly and semi-supervised learning in medical images," *Medical Image Analysis*, vol. 80, p. 102515, 2022.

[57] J. Wang, T. Lukasiewicz, X. Hu, J. Cai, and Z. Xu, "RSG: A simple but effective module for learning imbalanced datasets," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3784–3793.

[58] D. Yuan, Y. Liu, Z. Xu, Y. Zhan, J. Chen, and T. Lukasiewicz, "Painless and accurate medical image analysis using deep reinforcement learning with task-oriented homogenized automatic pre-processing," *Computers in Biology and Medicine*, p. 106487, 2022.

Zhenghua Xu received a M.Phil. in Computer Science from The University of Melbourne, Australia, in 2012, and a D.Phil in computer Science from University of Oxford, United Kingdom, in 2018. From 2017 to 2018, he worked as a research associate at the Department of Computer Science, University of Oxford. He is now a professor at the Hebei University of Technology, China, and a awardee of "100 Talents Plan" of Hebei Province. He has published more than thirty papers in top AI or database conferences and journals, e.g., NeurIPS, AAAI, IJCAI, ICDE, IEEE TNNLS, Medical Image Analysis, etc. His current research focuses on intelligent medical image analysis, deep learning, reinforcement learning and computer vision.

Biao Tian is currently a master student in the State Key Laboratory of Reliability and Intelligence of Electrical Equipment, Hebei University of Technology, China. He received B.Eng. degree in Electrical Engineering and Automatics from City College of Hebei University of Technology, China, in 2020. His research interests lie in medical image processing using deep learning methods.

Xiangtao Wang is currently a master student at Hebei University of Technology, China. He received the B.Eng. degree in Electrical Engineering and Automation from the North China University of Science and Technology, China, in 2021. His research interests lie in image analysis using deep learning methods.

Shuo Zhang is currently a master student in the State Key Laboratory of Reliability and Intelligence of Electrical Equipment, Hebei University of Technology, China. He received B.Eng. degree in Internet of Things Engineering from Jilin Agricultural University, China, in 2020. His research interests lie in medical image processing using deep learning methods. He has published high-quality papers in top journals, e.g., Medical Image Analysis.

Yuefu Zhan is currently an Associate Chief Physician at Department of Radiology, Hainan Women and Children's Medical Center. He received a Doctor of Medicine degree from the West China Medical School, Sichuan University, China, in 2022, and a Master of Medicine degree from the Xiangya School of Medicine, Central South University, China, in 2010. His current research interests mainly focus on imaging diagnosis, minimally invasive intervention, and AI-based medical image analysis.

Thomas Lukasiewicz is a Professor of Computer Science at the Department of Computer Science, University of Oxford, UK, heading the Intelligent Systems Lab within the Artificial Intelligence and Machine Learning Theme. He currently holds an AXA Chair grant on "Explainable Artificial Intelligence in Healthcare" and a Turing Fellowship at the Alan Turing Institute, London, UK, which is the UK's National Institute for Data Science and Artificial Intelligence. He received the IJCAI-01 Distinguished Paper Award, the AIJ Prominent Paper Award 2013, the RuleML 2015 Best Paper Award, and the ACM PODS Alberto O. Mendelzon Test-of-Time Award 2019. He is a Fellow of the European Association for Artificial Intelligence (EurAI) since 2020. His research interests are especially in artificial intelligence and machine learning.