Dual-Branch Differentiated Similarity Network for Semi-supervised Medical Image Segmentation

Weixian Yang, Jing Lin, Wentian Cai, and Ying Gao(✉)

School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, People's Republic of China
gaoying@scut.edu.cn

Abstract. Consistency learning has been successfully applied in semi-supervised medical image segmentation, as it enforces consistency in model prediction across different perturbations or transformations of the same input data. The accuracy of consistent learning is challenged by predictive diversity and model training stability, which are often neglected by existing researches. Meanwhile, the potential inter-class similarities between labeled and unlabeled training data are not effectively mined. To address these issues, we propose a semi-supervised framework based on dual decoders, namely Dual-Branch Differentiated Similarity Network (DBDSNet). First, dual-branch network and cross pseudo supervision can enable model to learn more invariant and meaningful representations from the data. Second, we proposed a Differentiated Similarity Loss (DSL) to encourage dual branches to focus on capturing the semantic information of the data, rather than relying on the noisy pseudo label. Last, we propose a Inter-Class Consistency Module (ICCM) to mine the inter-class similarity between labeled data and unlabelled data. Extensive experiments conducted on two public medical image datasets demonstrate that our method reaches state-of-the-art performance compared with former methods.

Keywords: semi-supervised medical segmentation · consistency learning · medical MRI dataset

1 Introduction

Accurate segmentation of medical images provides clinicians with salient and easily discernible information to help them make appropriate diagnoses, treatment plans and prognosis predictions. With the development of deep learning, deep learning models have demonstrated remarkable accuracy in various medical

Supported by organization Guangdong Provincial Key Laboratory of Artificial Intelligence in Medical Image Analysis and Application, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou 510080, China.

V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 574, pp. 264–278, 2024.
https://doi.org/10.1007/978-3-031-65123-6_19

tasks. Convolutional Neural Networks (CNNs) [14,15,20,24,27] have achieved impressive results in tasks like detecting tumors, identifying abnormalities in medical images, and classifying diseases. Fully supervised segmentation models require large amounts of pixel-level labels for training. However, annotating large-scale medical datasets at the pixel level can be particularly challenging due to the high-resolution nature of medical images and the need for specialized domain expertise. Recent advancements in weakly supervised [3,8,29] and semi-supervised learning approaches [25,32] are being explored to reduce the reliance on pixel-level annotations and enable training with less extensive supervision.

Semi-supervised segmentation provides a effective method to learn from a smaller set of labeled data while leveraging a larger set of unlabeled data. The unlabeled data serves as additional information that can help the model generalize better and learn more representative features. Mean teacher [28] is a widely-used solution in semi-supervised segmentation to improve performance by making the prediction and intermediate features remain consistent under different perturbations. However, single prediction decoder is easily overly sensitive to small perturbations in the input and lacks sufficient generalization performance. Thus, dual branch and cross prediction supervision enforce consistency in model predictions across different perturbations or transformations of the same input data. In order to obtain different views of dual branch prediction, we proposed a differentiated similarity loss to encourage dual branches to focus on capturing the underlying structure or semantics of the data rather than relying on incidental details or noise.

Meanwhile, copy and paste is one of the well-known data augmentations in the semi-supervised field for unlabeled data. But it applies random crop and paste to enforce stability, then the potential inter-class similarities between labeled and unlabeled training data are not effectively mined. Medical image usually contains only several class that means the voxels in same class are more likely to be similar to each other. Some methods [1,19,30,38] apply contrastive learning to encourage model to discover feature representation of each class. Insufficient labeled data restricts the stability of training class prototypes from scratch, it is still a challenge for medical image. We proposed a inter-class consistency loss to find the inter-class similarity in a different way. The loss combines the similar feature between labeled image and unlabelled image in order to mine the potential information.

In short, contributions of our work can be summarized as follows:

- a cross-supervised learning framework based on dual classifiers (DBDSNet) is proposed to enable models to learn more invariant and meaningful representations from the data.
- A differentiated similarity loss (DSL) is proposed to encourage dual branches to focus on capturing the underlying structure or semantics of the data rather than relying on incidental details or noise.
- An inter-class consistency module is proposed to mine the inter-class similarity between labeled data and unlabelled data.

– Experiments demonstrate that our proposed model achieves state-of-the-art performance on the LA2018 and ACDC datasets compared with several baselines.

2 Related Work

2.1 Medical 3D Segmentation

Medical images, such as CT or MRI scans, provide volumetric data in three dimensions (3D). These images consist of a stack of 2D slices, with each slice representing a cross-sectional view of the patient's anatomy. UNet [27] firstly proposed a encoder-decoder framework to segment medical MRI image. It use downsample in encoder and upsample in decoder, deriving its name from its U-shaped design. The skip connection plays a pivotal role in U-Net, connecting corresponding encoder and decoder layers at the same spatial resolution. VNet [24] follows a similar concept as UNet but is specifically tailored for volumetric medical image segmentation. V-Net operates directly on 3D medical image data, considering spatial dependencies across the entire volume. [14] introduces a cascaded architecture, where multiple U-Net models are trained in a cascade manner. It contributes to the field of medical image segmentation by providing a comprehensive framework that combines various improvements to the original U-Net architecture.

In recent years, while the Transformer architecture is primarily known for its success in natural language processing tasks, it has also been adapted for medical image segmentation [5,11,18,31]. TransUnet [5] combines the U-Net architecture with Transformer modules to perform medical image segmentation. It leverages the self-attention mechanism of Transformers to capture global contextual information and long-range dependencies within the images, improving segmentation accuracy. However, transformer-based model need more labeled data for training to enhance model stability and robustness, so it is not compatible for semi-supervised segmentation task due to the rare labeled data.

2.2 Semi-supervised Learning

Semi-supervised learning methods discover feature representation from a smaller set of labeled data while leveraging a larger set of unlabeled data. There are various strategies employed by in semi-supervised segmentation, including Contrastive Learning [1,6,12], Consistency Regularization [7,26], Pseudo-labeling [13,16,35,37], etc.

Pseudo-labeling. Pseudo-labeling methods are one of the most well-known methods and the earliest semi-supervised methods. The basic idea behind the method is extending the labeled dataset: generate pseudo-labels for unlabeled images based on predictions made by a model trained on labeled data, then extend the labeled dataset with confident pseudo-labels and corresponding unlabelled image, and train a new model on this new dataset. Lee *et al.* [16] applied

self-training strategy for the first time with deep neural networks and outperforms traditional semi-supervised learning methods in the case of Denoising Auto-Encoder (DAE) and Dropout module. Yang *et al.* [35] applied data augmentation techniques to unlabeled images during self-training. At each iteration of the self-training process, those images with reliable pseudo-labels are prioritized and those with a high probability of errors in the pseudo-labels are discarded.

Contrastive Learning. Contrastive learning focuses on high-level features to differentiate between classes in the absence of ground truth. In most contrastive learning methods, the target samples to be compared are called *queries*, while similar and dissimilar samples are defined as *positive* and *negative* keys, respectively. He *et al.* [12] firstly build the basic framework in semi-supervised segmentation, and proposed to use memory bank to store encoded sample. Chen *et al.* [6] proposed a simple framework to encode the input image and considered negative samples to calculate the loss. Wang *et al.* [30] proposed a uncertainty guided contrastive learning method on medical MRI image in order to effectively alleviate noise sampling from pseudo-labels of unlabeled data.

2.3 Consistency Regularization

Consistency Regularization promotes consistent predictions when the same input is presented with perturbations. [4] proposed the assumption of smoothness that for two nearby voxels in the input space, their labels must be the same. As a result, the main idea of Consistency Regularization is to create augmented versions of the unlabeled data and encourage the model to produce similar segmentations for each version. The consistency loss penalizes discrepancies between predictions on the different augmented versions.

The most commonly used method is Mean-teacher [28], forced the consistency between the predictions of *teacher network* and *student network*. Exponential moving average (EMA) is proposed to update the weights of *teacher network* by the weight sum of *student network* and *teacher network*. Based on Mean-teacher framework, Cutmix [9] applied cutmix data augmentation on dataset in order to generate input perturbations. It combines images and their corresponding labels by randomly cutting and pasting patches of one image onto another image, creating a new training example. Another famous method is proposed by [26], aimed to use some auxiliary decoders to enforce consistency between the outputs of the auxiliary decoders. [7] includes two independent network and emphasizes the importance of enforcing diversity across networks, allows the model to leverage the information present in the unlabeled data to improve segmentation accuracy.

3 Method

Our proposed network is shown in Fig. 1, which is divided into three main module. In the first module, we use labeled training images (follow the partition principle) and unlabeled images to train the backbone network guided by

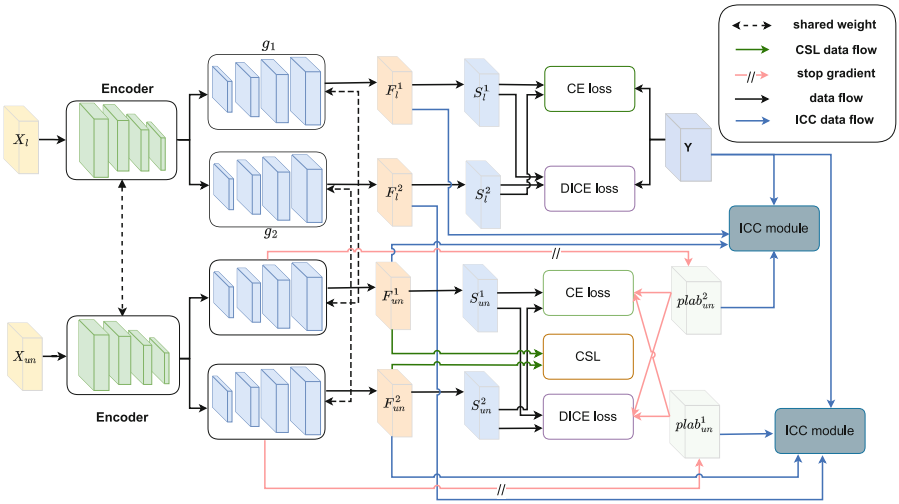

Fig. 1. The whole framework of DBDSNet.

cross-entropy loss and dice loss. In the second module, we design inter-class consistency module that mentioned in Sect. 3.2 to mine the inter-class similarity between labeled data and unlabelled data. In the third module, we introduce a differentiated similarity loss L_{DSL} that mentioned in Sect. 3.3 to encourage two branches to focus on capturing the underlying structure or semantics of the data. Notice that we construct two ICCM for both branch and we draw both module to show the data flow. Best viewed in color.

3.1 Framework

Mathematically, we assume that D_l and D_u denote the labeled set and unlabeled set, respectively. Labelled set contains far more elements than unlabelled set. Then the 3D volume of a MRI image can be denoted as $I^l \in R^{C\times H\times W\times D}$, which will be fed into the following network.

For labeled data, the input 3D MRI data $X_l \in D_l$ ($X_l \in R^{c\times h\times w\times d}$) is fed into the Encoder module f that consists of a series of 3D convolutional layers and 3D max-pooling operations to downsample the spatial resolution while increasing the number of feature channels. Then the decoder module g_1 follows the contracting path and is responsible for upsampling the feature maps and reconstructing the segmentation mask. In this way, we can get the corresponding feature map F_l^1 and prediction map S_l^1 from Decoder g_1. In order to make different predictions, we rebuild another decoder g_2 as same as g_1 to segment the same input data. Similarly, we can obtain feature map F_l^2 and prediction map S_l^2 from decoder g_2.

$$F_l^1, S_l^1 = g_1(f(X_l)) \tag{1}$$

$$F_l^2, S_l^2 = g_2(f(X_l)) \tag{2}$$

Similar to above, for unlabelled data $X_{un} \in D_u$ ($X_{un} \in R^{c \times h \times w \times d}$), we also can obtain feature map F^1_{un}, prediction map S^1_{un} from Decoder g_1, feature map F^2_{un} and prediction map S^2_{un} from Decoder g_2, respectively. To improve the performance of the decoder, we adapt cross-entropy loss (Eq. 3) as a supervision loss and employ cross pseudo supervision [7] strategy to improve the robustness of both decoders. Cross-prediction strategy means that we use the pseudo label generated from Decoder g_1 as the label input for Decoder g_2. And the pseudo label generated from Decoder g_2 is also considered as the label input for Decoder g_1.

$$L_{ce}(X, y) = -\frac{1}{N}\sum_{i}^{N}(y[i] * \log((X[i])) \tag{3}$$

$$L^{un}_{CE} = L_{ce}(S^1_{un}, plab^2_{un}) + L_{ce}(S^2_{un}, plab^1_{un}) \tag{4}$$

$$L^{l}_{CE} = L_{ce}(S^1_{l}, Y) + L_{ce}(S^2_{l}, Y) \tag{5}$$

where Y is the label for labeled data I^l, N is the number of categories, $plab^1_{un}$ and $plab^2_{un}$ are the corresponding pseudo label generated from Decoderg_1 and Decoderg_2 with no gradient.

Dice loss is a commonly used loss function in medical segmentation tasks for evaluating the performance of semantic segmentation models. It measures the similarity between the predicted segmentation mask and the ground truth mask as same as cross-entropy loss. The dice loss can be denoted as Eq. 6.

$$L_{dice}(X, Y) = \frac{2 * (|X \cap Y|)}{|X| + |Y|} \tag{6}$$

where X is prediction, Y is label map, $|X|$ and $|Y|$ means the element of X and Y. Similar to cross-entropy loss, we can get the dice loss L^{un}_{DICE} and L^{l}_{DICE} as Eq. 7 and Eq. 8, respectively.

$$L^{un}_{DICE} = L_{dice}(S^1_{un}, plab^2_{un}) + L_{dice}(S^2_{un}, plab^1_{un}) \tag{7}$$

$$L^{l}_{DICE} = L_{dice}(S^1_{l}, Y) + L_{dice}(S^2_{l}, Y) \tag{8}$$

3.2 Inter-class Consistency Loss

Medical image usually contains only several class that means the voxels in same class are more likely to be similar to each other. In order to utilize the inter-class similarity, we propose a ICC module which bridges the feature distribution of labeled data and feature prediction of unlabelled data, as shown in Fig. 2. Firstly, notice that we already have ground truth label for labeled data during the training step. Therefore, by multiplying the label Y by labeled data feature $F^i_l (i \in 1, 2)$, we can obtain the foreground feature vector $V^i_l \in R^{N \times D}$ of class i by Eq. 9.

$$V^i_l[j] = GAP(F^i_l \times (Y == j))(j \in 1, ..., N-1) \tag{9}$$

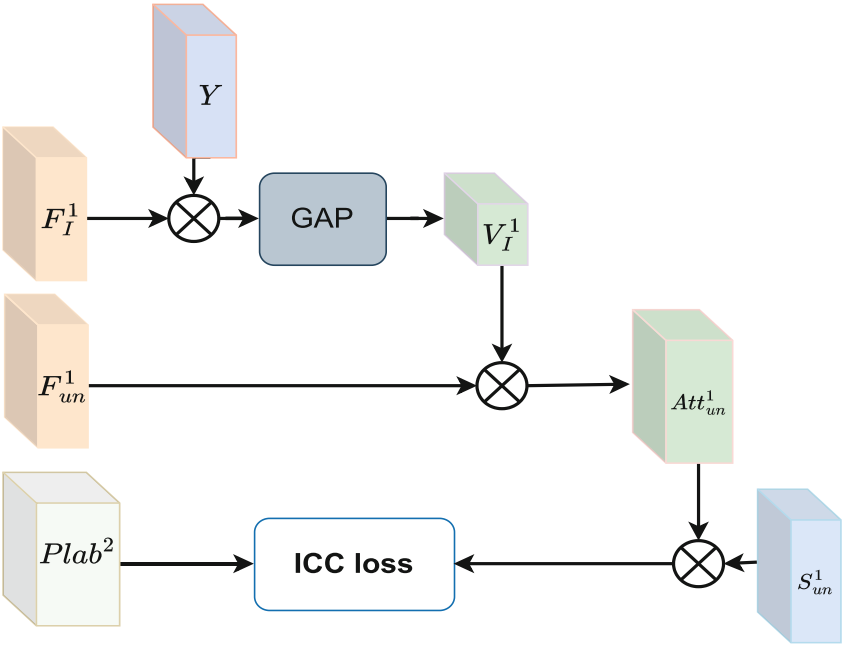

Fig. 2. Inter-class consistency module. It is worth noting that both branches use ICC modules, here we only draw a data flow as a demonstration.

where N is the number of class and GAP means global average pooling. $Y == j$ means that we use label Y to select the correct voxels for each classes. In this way, we can get the initial class feature vector to represent initial average feature for each class. Then we compute the similarity between feature vector V_l^i and F_{un}^i to obtain similarity matrix $Sim(u, v, k) \in R^{c \times h \times w \times d}$, as shown in Eq. 10. The matrix

$$Sim(u, v, k) = \frac{V_l^i \cdot F_{un}^i}{||V_l^i|| \times ||F_{un}^i||} \tag{10}$$

where u,v,k point out the voxel in Sim matrix. Finally, the ICC loss is designed as Eq. 11. Notice that we utilize ICC module for both branch. And for branch i, we consider the internal similarity relationships of all classes individually and finally aggregate to get the final L_{ICC}^i.

$$L_{ICC}^i = \sum_{j=1}^{N-1} \frac{1}{hwd} \sum(\mid (plab_{un}^{2-i} == j) \; - \; Sim \mid) \times L_{ce}(S_{un}^i, plab_{un}^{2-i})) \tag{11}$$

where $|.|$ represent that we take absolute values for the target value. Assuming the voxel (u, v, w) does not belong to the target category, $(plab_{un}^{2-i} == j)(u, v, w) = 0$, then $|0 - Sim(u, v, w)|$ indicates the tendency that the voxel needs to be optimised towards 0. Otherwise, $(plab_{un}^{2-i} == j)(u, v, w) = 1$, $1 - Sim(u, v, w)$

encourages the model to be more precise in assigning higher confidence to the target category.

3.3 Differentiated Similarity Loss

In Sect. 3.1, we employ cross pseudo supervision to encourage same between the predictions of two perturbed networks for the same input image. Cross pseudo supervision also encourages the model to maintain consistent predictions, which affects the feature map of different decoders and is detrimental to the learning of model perturbations. In semi-supervised segmentation, due to the lack of labelled training data, it is of little value and a waste of effort to obtain similar features from different encoders.

Cosine similarity is a famous measure of similarity between two feature map, indicating the degree of alignment or similarity between their orientations. It is sensitive to angles between feature maps and rewards features with similar orientations regardless of their position. This makes cosine similarity suitable for feature orientation gating. And cosine similarity ranges from -1 to 1, which is easily to gated and discard useless signal. In this way, we can make the two feature maps focus on different perspectives of the original input. This will help us train the model. Therefore, we propose a differentiated similarity loss to minimize the cosine similarity between F^2_{un} and F^1_{un}, as shown in Eq. 12. To avoid producing negative values, we use the Relu [10] function to activate the calculations.

$$L_{DSL} = Relu(\frac{F^2_{un} \cdot F^1_{un}}{||F^2_{un}|| \times ||F^1_{un}||}) \tag{12}$$

3.4 Total Loss Function

The architecture of our classifier contains three main components. First module is a dual-branch VNet backbone and cross prediction supervision to segmentation medical 3D MRI image. The second module is inter-class consistency module to mine the inter-class similarity between labeled data and unlabelled data. In the third module, a differentiated similarity loss L_{DSL} is mentioned to encourage two branches to focus on capturing the underlying structure or semantics of the data. The total loss is formulated as Eq. 13.

$$L_{total} = L^l_{CE} + L^l_{DICE} + L^{un}_{CE} + L^{un}_{DICE} + \alpha L_{DSL} + L_{ICCL} \tag{13}$$

4 Experiment

4.1 Datasets and Preprocessing

Our code is built in Pytorch 1.7.1 and cuda version is 11.2. A RTX2080 12G GPU is used for training with fixed random seed. We evaluate proposed method DBDSNet on two benchmark datasets: LA dataset [34] and ACDC dataset [2].

Table 1. Evaluation of the segmentation result in terms of Dice (%), Jaccard (%), 95% Hausdorff Distance (95HD) and ASD on the LA dataset. Best results are highlighted.

Method	labeled	Unlabelled	Dice(%)↑	Jaccard(%)↑	95HD↓	ASD↓
VNet	4(5%)	0	52.55	39.60	47.05	9.87
VNet	8(10%)	0	82.74	71.72	13.35	3.26
VNet	80(all)	0	91.47	84.36	5.48	1.51
UA-MT [36]	8(10%)	72(90%)	87.79	78.39	8.68	2.12
SASSNet [17]	8(10%)	72(90%)	87.54	78.05	9.84	2.59
DTC [21]	8(10%)	72(90%)	87.51	78.17	8.23	2.36
URPC [23]	8(10%)	72(90%)	86.92	77.03	11.13	2.28
MC-Net [33]	8(10%)	72(90%)	87.62	78.25	10.03	1.82
SSNet [32]	8(10%)	72(90%)	88.55	79.62	**7.49**	1.90
Ours	8(10%)	72(90%)	**89.30**	**80.81**	9.02	**1.68**

LA Dataset. The Atrial Segmentation Challenge dataset [34] comprises a total of 100 3D gadolinium-enhanced magnetic resonance imaging scans (GE-MRIs) with accompanying labels. We strictly adhere to the settings that are used in the SSNet [32] and UA-MT [36]. In terms of data preparation, we use rotations and flips to enhance the input. We utilize an SGD optimizer, initially setting the learning rate at 0.01 and decreasing it every 2.5K iterations. From the whole 3D volume of the MRI image, $112 \times 112 \times 80$ patches are randomly selected during the training step. The number of iterations of the pre-training and self-training are fixed to be 2k and 15k, respectively.

ACDC Dataset. The ACDC dataset [2] comprises four categories, including the background, left ventricle, right ventricle as well as myocardium. It includes imaging scans from 100 patients, with a fixed data split of 70, 10 and 20 patients for training, validation and testing, respectively. For ACDC dataset, we also follow the setting of SSNet [32] and [22]. In this experiment, we employ 2D-UNet as the foundational backbone of our method. During the training phase, we crop the size of the input patch to be 256×256. In addition, we set other parameters such as the batch size, the number of pre-training iterations, and the number of self-training iterations to 24, 10,000, and 30,000, respectively.

Evaluating Metrics. We plan to use four evaluation metrics to compare with state-of-the-art methods and to verify performance on the datasets. Two object regions can be analysed using the Dice and Jaccard metrics, which calculate the overlap percentage. Average Surface Distance(ASD) calculates the median distance between the boundaries of the two regions, while 95% Hausdorff Distance(95HD) measures the nearest distance between their respective points.

4.2 Comparison with State-of-the-Art Methods

LA Dataset. We compare our DBDSNet method with recent semi-supervised semantic segmentation methods, including UT-MA [36], SASSNet [17], SSNet

Table 2. Evaluation of the segmentation result in terms of Dice (%), Jaccard (%), 95% Hausdorff Distance (95HD) and ASD on ACDC dataset. Best results are highlighted.

Method	labeled	Unlabelled	Dice(%)↑	Jaccard(%)↑	95HD↓	ASD↓
UNet	3(5%)	0	47.83	37.01	31.16	12.62
UNet	7(10%)	0	79.41	68.11	9.35	2.70
UNet	70(all)	0	91.44	84.59	4.30	0.99
UA-MT [36]	3(5%)	67(95%)	46.04	35.97	20.08	7.75
SASSNet [17]	3(5%)	67(95%)	57.77	46.14	20.05	6.06
DTC [21]	3(5%)	67(95%)	56.90	45.67	23.36	7.39
URPC [23]	3(5%)	67(95%)	55.87	44.64	13.60	3.74
MC-Net [33]	3(5%)	67(95%)	62.85	52.29	7.62	2.33
SSNet [32]	3(5%)	67(95%)	65.83	55.38	**6.67**	**2.28**
Ours	3(5%)	67(95%)	**86.14**	**76.80**	8.37	2.53
UA-MT [36]	7(10%)	60(90%)	81.65	70.64	6.88	2.02
SASSNet [17]	7(10%)	60(90%)	84.50	74.34	5.42	1.86
DTC [21]	7(10%)	60(90%)	84.29	73.92	12.81	4.01
URPC [23]	7(10%)	60(90%)	83.10	72.41	4.84	1.53
MC-Net [33]	7(10%)	60(90%)	86.44	77.04	5.50	1.84
SSNet [32]	7(10%)	60(90%)	86.78	77.67	6.07	**1.40**
Ours	7(10%)	60(90%)	**87.78**	**79.13**	**4.51**	1.50

Table 3. Ablation study of hyper-parameter for differentiated similarity loss (DSL)

weight	Dice(%)↑	Jaccard(%)↑	95HD↓	ASD↓
0.1	88.05	78.84	11.83	2.22
0.2	89.01	80.37	**9.01**	1.79
0.5	**89.30**	**80.81**	9.02	1.68
1	88.90	80.19	9.89	**1.65**

[32], DTC [21], MC-Net [33] and URPC [23]. In addition, we also include the results of supervised methods that train the model with only labelled data for comparison (denoted as backbone name). For all of the experiments, we follow SSNet and randomly split the datasets on LA datasets, as shown in Table 1. Our model performs state-of-the-art performance in the all sota methods, and outperforms the current SOTA method by 0.75% and 1.09% on Dice and Jaccard when only 8 labelled data are available, respectively.

ACDC Dataset. The Table 2 displays the results of four-class segmentation performance on the ACDC dataset using labeled ratios of 5% and 10%. In comparison to state-of-arts, our model DBDSNet achieves excellent performance, with a large margin of improvement in many metrics. For 10% labeled ratio, our

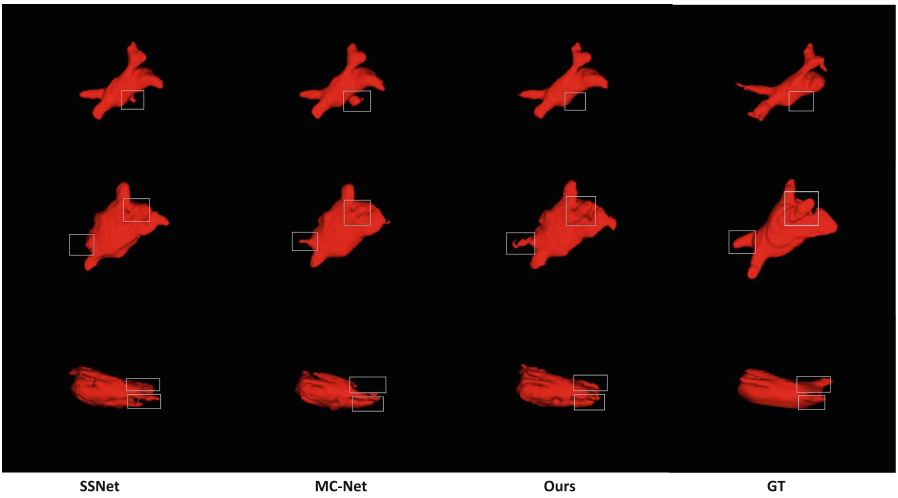

Fig. 3. Visualization for our model on LA dataset compared with some state-of-art methods.

Table 4. The effectiveness of the main contribution on LA Dataset evaluated with 8 labeled data. Baseline means no additional module for backbone; DSL means that using differentiated similarity loss to improve the performance; ICCL means use inter-class module to supervise the model for better training

baseline	dual+ce	DSL	ICCL	Dice(%)↑	Jaccard(%)↑	95HD↓	ASD↓
√				82.74	71.72	13.35	3.26
√	√			87.30	77.74	11.39	2.10
√	√	√		88.50	79.67	9.61	1.76
√	√	√	√	**89.30**	**80.81**	**9.02**	**1.68**

model improve the second best results by 1.0% on Dice metric, 1.46% on Jaccard and 1.56 on 95HD. For 3 labeled experiment, our model improve the second best results by 20.31% on Dice metric and 21.42% on Jaccard. Following SSNet [32], we use 2-dimensional slices as input data to train the model. With cross pseudo supervision and differentiated similarity loss, our network produces divergent but reasonable multiple predictions, thus mitigating errors due to model error or lack of robustness. In addition, mining for inter-class information can effectively focus on the precision of detail boundaries and confusing voxels.

Visualization on LA Dataset. We visualise the predictions produced by different methods and use this as a comparison to show the effectiveness of our model. The detail can be seen in Fig. 3. Our model successfully avoids generating false positive annotations in some of the regions that are prone to be misclassified, as shown in the first line. As for the second line, our method is capable of uncovering potential voxels in some regions, thus discovering and

completing some true negative pixels. The third row shows that our model is able to achieve a relatively complete segmentation, with smooth boundaries for most of the region.

4.3 Ablation Study

We conduct ablation studies to show the impact of each component in DBDSNet and show how the hyper-parameter influence the performance of our DBDSNet. We choose LA dataset as the main dataset and use 8 labeled data as labeled set. Other parameters are set as same as the state-of-art comparison.

Loss Weight of DSL. We initially set α as the default value at 0.5. We then varied α to 0.1, 0.2, 0.5 and 1 to examine its impact on performance. The data presented in the Table 3 indicates an increase in performance as α is adjusted from 0.1 to 0.5; however, a decline in the performance is observed when α is set to 1. Therefore, for loss weight α selection, we choose 0.5 as the current local optimal solution.

Effectiveness of Components. Recall that our DBDSNet includes a ICC module, a DS loss and dual-branch training. Note that there are two losses, *i.e.*, the differentiated similarity loss L_{DSL} and the inter-class consistency loss L_{ICCL}. We now analyse the individual contributions of these losses and module.

From Table 4, we can observe that if we only apply dual branch and cross pseudo supervision, our method will bring a performance improvement of over 4.56%. But Due to the presence of cross pseudo supervision, the attention of the two branch predictions will be biased towards similar views and the reasoning views of the two sub-nets are kind of correlated, leaving a large space for improvement. When applying our differentiated similarity loss L_{DSL}, the performance is observed to increase 1.2% on Dice metric, indicating that our operation can ensure the two sub-nets reason the input from two differential views, thus improving the robustness of the model. In addition, if we apply the ICC module, there is a 0.8% improvement of Dice metric on dataset, which means our model can mine the potential inter-class similarities and bridge the similar feature representation in this way.

4.4 Limitation and Future Work

In this work, DBDSNet was designed for multiple categories in a small number of organs. One of the limitations is that we did not evaluate our model on a complex dataset containing multiple organs. On this type of dataset, achieving optimal segmentation results can be highly challenging. Besides, the model is limited by traditional CNN backbone, which can be replaced by Transformer. In the future, We intend to expand DBDSNet to address the challenge of segmenting 3D medical images with multiple organs, and explore inter-class attention and similarity mechanisms to effectively handle objects with indistinct boundaries and low contrast.

5 Conclusion

This paper proposes a cross-supervised learning framework based on dual classifiers, namely DBDSNet, that guarantees a level of disagreement through their complementarity. First, dual-branch framework can reduce the wrong prediction due to the average predict of dual branches. Second, differentiated similarity loss effectively maintains the consistency of the two branch predictions and stabilizes the accuracy of the model. Last, we construct a inter-class consistency module to bridge the feature similarity between labeled data and unlabelled data. Extensive experiments conducted on two public medical image datasets demonstrate the superior segmentation performance of DBDSNet compared to other methods.

References

1. Alonso, I., Sabater, A., Ferstl, D., Montesano, L., Murillo, A.C.: Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8219–8228 (2021)
2. Bernard, O., et al.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans. Med. Imaging **37**(11), 2514–2525 (2018)
3. Cai, W., Xie, L., Yang, W., Li, Y., Gao, Y., Wang, T.: DFTNet: dual-path feature transfer network for weakly supervised medical image segmentation. IEEE/ACM Trans. Comput. Biol. Bioinform. **20**(4), 2530–2540 (2022)
4. Chapelle, O., Scholkopf, B., Zien, A.: Semi-supervised learning (Chapelle, O. et al., eds.; 2006)[book reviews]. IEEE Trans. Neural Netw. **20**(3), 542–542 (2009)
5. Chen, J., et al.: Transunet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, pp. 1597–1607. PMLR (2020)
7. Chen, X., Yuan, Y., Zeng, G., Wang, J.: Semi-supervised semantic segmentation with cross pseudo supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2613–2622 (2021)
8. Chen, Z., Tian, Z., Zhu, J., Li, C., Du, S.: C-cam: causal cam for weakly supervised semantic segmentation on medical image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11676–11685 (2022)
9. French, G., Laine, S., Aila, T., Mackiewicz, M., Finlayson, G.: Semi-supervised semantic segmentation needs strong, varied perturbations. arXiv preprint arXiv:1906.01916 (2019)
10. Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier neural networks. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, pp. 315–323. JMLR Workshop and Conference Proceedings (2011)
11. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In: Crimi, A., Bakas, S. (eds.) BrainLes 2021. LNCS, vol. 12962, pp. 272–284. Springer, Cham (2021). https://doi.org/10.1007/978-3-031-08999-2_22

12. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729–9738 (2020)
13. He, R., Yang, J., Qi, X.: Re-distributing biased pseudo labels for semi-supervised semantic segmentation: a baseline investigation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6930–6940 (2021)
14. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
15. Kamnitsas, K., et al.: Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med. Image Anal. **36**, 61–78 (2017)
16. Lee, D.H., et al.: Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on Challenges in Representation Learning, ICML, vol. 3, p. 896. Atlanta (2013)
17. Li, S., Zhang, C., He, X.: Shape-aware semi-supervised 3D semantic segmentation for medical images. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 552–561. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_54
18. Li, Y., Cai, W., Gao, Y., Li, C., Hu, X.: More than encoder: introducing transformer decoder to upsample. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1597–1602. IEEE (2022)
19. Liu, S., Zhi, S., Johns, E., Davison, A.: Bootstrapping semantic segmentation with regional contrast. In: International Conference on Learning Representations (2021)
20. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
21. Luo, X., Chen, J., Song, T., Wang, G.: Semi-supervised medical image segmentation through dual-task consistency. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 8801–8809 (2021)
22. Luo, X., Hu, M., Song, T., Wang, G., Zhang, S.: Semi-supervised medical image segmentation via cross teaching between CNN and transformer. In: International Conference on Medical Imaging with Deep Learning, pp. 820–833. PMLR (2022)
23. Luo, X., et al.: Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12902, pp. 318–329. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_30
24. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571. IEEE (2016)
25. Nie, D., Gao, Y., Wang, L., Shen, D.: ASDNet: attention based semi-supervised deep networks for medical image segmentation. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11073, pp. 370–378. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00937-3_43
26. Ouali, Y., Hudelot, C., Tami, M.: Semi-supervised semantic segmentation with cross-consistency training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12674–12684 (2020)
27. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28

28. Tarvainen, A., Valpola, H.: Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
29. Wang, J., Xia, B.: Bounding box tightness prior for weakly supervised image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12902, pp. 526–536. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_49
30. Wang, T., Lu, J., Lai, Z., Wen, J., Kong, H.: Uncertainty-guided pixel contrastive learning for semi-supervised medical image segmentation. In: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI, pp. 1444–1450 (2022)
31. Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., Li, J.: TransBTS: multimodal brain tumor segmentation using transformer. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 109–119. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_11
32. Wu, Y., Wu, Z., Wu, Q., Ge, Z., Cai, J.: Exploring smoothness and class-separation for semi-supervised medical image segmentation. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13435, pp. 34–43. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_4
33. Wu, Y., Xu, M., Ge, Z., Cai, J., Zhang, L.: Semi-supervised left atrium segmentation with mutual consistency training. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12902, pp. 297–306. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_28
34. Xiong, Z., et al.: A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. Med. Image Anal. **67**, 101832 (2021)
35. Yang, L., Zhuo, W., Qi, L., Shi, Y., Gao, Y.: ST++: make self-training work better for semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4268–4277 (2022)
36. Yu, L., Wang, S., Li, X., Fu, C.-W., Heng, P.-A.: Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 605–613. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_67
37. Yuan, J., Liu, Y., Shen, C., Wang, Z., Li, H.: A simple baseline for semi-supervised semantic segmentation with strong data augmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8229–8238 (2021)
38. Zhao, Z., et al.: MMGL: multi-scale multi-view global-local contrastive learning for semi-supervised cardiac image segmentation. In: 2022 IEEE International Conference on Image Processing (ICIP), pp. 401–405. IEEE (2022)