Medical & Biological Engineering & Computing
https://doi.org/10.1007/s11517-023-02920-0

ORIGINAL ARTICLE

Dual CNN cross-teaching semi-supervised segmentation network with multi-kernels and global contrastive loss in ACDC

Keming Li[1] · Guangyuan Zhang[1] · Kefeng Li[1] · Jindi Li[1] · Jiaqi Wang[1] · Yumin Yang[1]

Received: 14 January 2023 / Accepted: 22 August 2023

Abstract

The cross-teaching based on Convolutional Neural Network (CNN) and Transformer has been successful in semi-supervised learning; however, the information interaction between local and global relations ignores the semantic features of the medium scale, and at the same time, the information in the process of feature coding is not fully utilized. To solve these problems, we proposed a new semi-supervised segmentation network. Based on the principle of complementary modeling information of different kernel convolutions, we design a dual CNN cross-supervised network with different kernel sizes under cross-teaching. We introduce global feature contrastive learning and generate contrast samples with the help of dual CNN architecture to make efficient use of coding features. We conducted plenty of experiments on the Automated Cardiac Diagnosis Challenge (ACDC) dataset to evaluate our approach. Our method achieves an average Dice Similarity Coefficient (DSC) of 87.2% and Hausdorff distance (HD_{95}) of 6.1 mm on 10% labeled data, which is significantly improved compared with many current popular models.

Keywords Semi-supervised segmentation · Cross-teaching · Dual CNN · Contrastive learning

1 Introduction

Early medical image segmentation is mainly based on traditional algorithms, such as edge detection-based, threshold-based, and region-based segmentation algorithms. But the contrast of different tissues and organs and the boundary of medical images make the effect of traditional segmentation on the medical image not ideal.

With the rapid development of computer computing power, the image segmentation method based on deep learning has been paid more attention than ever before. The Fully Convolutional Networks (FCN) [1] and U-Net [2] have laid the foundation of deep learning in the field of image segmentation; subsequently, nn-unet [3] was proposed to achieve state-of-the-art on many medical datasets. However, the most critical aspect of deep learning is the need for massive labeled datasets to the effect of strongly supervised learning, which is difficult to achieve in the real world because medical datasets are often very difficult to obtain and a lot of accurate data annotation is very expensive. The pathological changes of medical images are highly heterogeneous in morphology, which makes the marking process greatly depend on the cognition and experience of medical experts, while considering the uncertainty of marking doctors' subjective criteria and the cognition of different experts objectively, in the process of marking, mismarking is inevitable, and the accuracy of marking is not completely reliable.

To solve this problem, semi-supervised segmentation models for 2D [4, 5] and 3D [6, 7] images are gradually developed. Semi-supervised learning only relies on a small amount of labeled data and a large number of unlabeled data

✉ Kefeng Li
seafrog1984@hotmail.com

Keming Li
leekm3597@163.com

Guangyuan Zhang
xdzhanggy@163.com

Jindi Li
sdjtu_lw@163.com

Jiaqi Wang
Wjq_Jocelyn@163.com

Yumin Yang
1940795297@qq.com

[1] School of Information Science and Electric Engineering, Shandong Jiaotong University, Jinan, China

Published online: 09 September 2023

to achieve the same effect as supervised learning. For semi-supervised learning, the precondition of semi-supervised learning is that the data distribution should be under some assumptions that the data structure remains unchanged. Pseudo-label [8] and consistency regularization [9] are two main research strategies in semi-supervised segmentation. The training method based on pseudo-label mainly uses the data with labels to train initially, then predicts and generates pseudo-labels for the unlabeled data, and does iterative training. To solve the problem that false labels are easily adulterated with noise, d.- H. Lee [8] proposed confidence screening for pseudo-labels, and Wang et al. [10] added a trust module to evaluate the pseudo-labels in the model output and set a threshold to select high confidence values. In addition to increasing the credibility awareness module, there are many other ways to improve the quality of pseudo-labels. Li et al. [11] build up-to-date predictions by exponential moving averages to avoid noise and unstable pseudo-labels. Luo et al. [12] used CNN and transformer to predict each other to generate pseudo-labels for training. The training method based on consistency regularization mainly uses various disturbances to enforce the consistency of prediction, for example, by enhancing the input image, the characteristic disturbance, and the network disturbance to achieve the input disturbance. The π [13] model adopts the self-ensemble strategy, training by applying the consistency regular constraint between two enhanced data of the same sample. The temporal ensemble [14] is another self-ensemble method; it adds exponential moving average (EMA) prediction modules to the π-model to generate low-noise predictions by considering the results of the last epoch. To update the network parameters more quickly, mean teacher (MT) [15] exponentially weighted the weights of the student networks to generate the teacher network weights on average, thus generating a more stable and continuous low-noise prediction. Yu et al. [7] added Monte Carlo uncertainty estimation on MT to select low entropy samples.

The idea of contrastive learning, which was first applied to self-supervised learning, is to advance the distance of positive samples in the feature space while pushing the negative samples away. MoCo [16, 17] and SimCLR [18, 19] used the agent method of individual discrimination to train the model of contrastive learning. The performance of the self-supervised pre-training model on downstream tasks has surpassed that of the supervised pre-training model. In order to make full use of unlabeled data to learn data characteristics, Zhong et al. [20] used the spatial consistency of weakly enhanced images to construct positive samples and generated negative samples by cross-image and pseudo-label weighting. The construction of samples using pseudo-labels may deviate from the actual samples' semantics, leading to the introduction of contrastive noise.

The cross-teaching model of CNN and Transformer [12] proves the effectiveness of global information to supplement local information, but too long distance modeling information is often a scale waste for segmentation; for this reason, we design a dual CNN network, and in order to use the coding features efficiently, we use the dual CNN structure to generate contrastive samples for contrastive learning. Our contributions are as follows:

1. We propose a dual CNN cross-teaching approach with different kernel sizes to effectively utilize features at different scales. In contrast to the CNN and Transformer cross-teaching framework, our method specifically emphasizes mid-scale features with higher attention.
2. To learn the global representation, we map the samples to the D-dimensional vector space at the bottom of the CNN network. Then, we regard the different feature representations generated by different scale convolutions of the same sample as positive samples and the other samples as negative samples, construct the global contrastive loss embedded in the total segmentation loss, and jointly optimize the segmentation framework.

2 Related work

2.1 Semi-supervised medical segmentation

Semi-supervised segmentation has received much attention under the condition that only a small amount of labeled data is required, and many semi-supervised network models have been proposed for organ segmentation [31, 32]. For example, Li et al. [8] proposed a semi-supervised skin damage segmentation method. Zhu et al. [33] proposed a hybrid dual mean-teacher model with hybrid, which generates predictions through 2D and 3D networks; Zhang et al. [21] provided a novel semi-supervised contrastive learning segmentation framework, which utilize the cross-modal information and prediction consistency between different modalities to conduct contrastive mutual learning; and Luo et al. [12] introduced Transformer into semi-supervised tasks, the model makes use of the cross-supervision of CNN and Transform to generate pseudo-labels. The advantages of long-distance semantic modeling of Transform make up CNN, and the advantages of CNN in extracting low-level features also make up Transformer. CNN and Transformer describe the complementary information of data from different points of view and carry out cooperative training. Xiao et al. [22] adopted a dual-teacher model, which effectively utilized the explicit and implicit consistency regularization of dual-teacher.

2.2 Contrastive learning

In image-level representation learning, contrastive learning can make full use of unlabeled data to learn effective visual representation; the core idea is to reduce the similarity pairs (positive) and separate different pairs (negative) based on some similar constraints to enhance the distinction of visual representations. The key to contrastive learning is how to construct contrastive samples. He et al. [16] proposed a feasible solution by designing momentum encoders to keep the consistency of samples. Liu et al. [23] used class-level information to construct contrast samples of foreground-background and set the class vector in labeled data as a reference to the class vector in unlabeled data, so as to reduce the noise prediction of unlabeled data. Wang et al. [24] demonstrated the advantages of supervised segmentation using contrastive learning across image pixels. By performing pixel-to-pixel directional contrastive learning, Zhao et al. [28] used data augmentation to construct positive samples and add mapping heads at different locations in the network to perform multi-scale contrastive learning.

3 Methodology

For general semi-supervised learning tasks, the training data usually consists of two parts: N labeled data $D^{\mathrm{L}} = \{X_n, Y_n\}_{n=1}^{N}$ and M unlabeled data $D^{\mathrm{U}} = \{X_n\}_{n=N+1}^{N+M}$, and M >> N, the total data set $D = D^{\mathrm{L}} \cup D^{\mathrm{U}}$.

3.1 Model architecture

Figure 1 illustrates our semi-supervised learning model for heart MRI segmentation. It mainly includes two parts: dual CNN cross-supervised module, which can get prediction by two CNNs with different kernel size and generate pseudo-labels for each other; global contrastive learning module maps the features from dual CNN encoder to the D-dimension vector and carries out the contrastive learning.

3.2 Cross-teaching between dual CNNs

Due to the complementary advantages of long-distance semantic modeling information of transformer and CNN, the cross-teaching of CNN and transformer has achieved success in semi-supervised segmentation, and cross-supervision improves the generalization performance of the model by introducing network framework-level perturbation. However, this information interaction mode is established between small-scale and global information, too large information difference between which will affect the learning performance of the model. As medium-scale semantic information and small-scale information are more consistent, they are more suitable for cross-supervised learning. Considering that transformer has a large number of parameters and it relies heavily on pre-training weights, we replace the transformer with a large-kernel CNN network and use the medium-scale semantics extracted from the large-convolution kernel instead of the global information of transformer to achieve

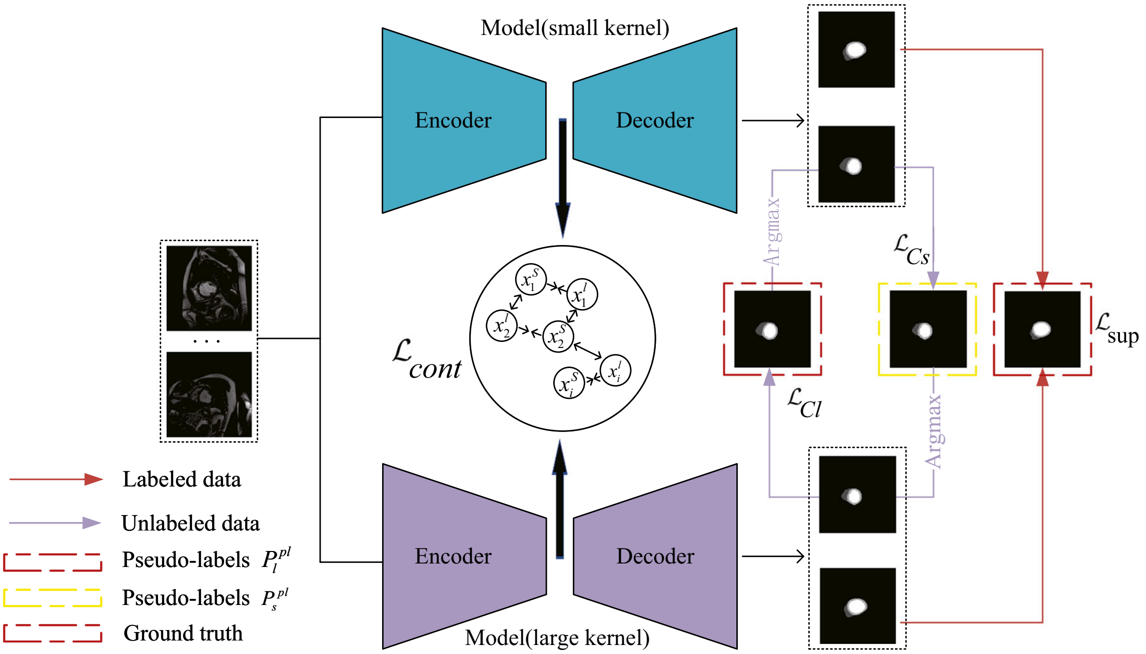


Fig. 1 Illustrating the structure of the model. In the figure with subscript labels, “l” represents the large kernel network, “s” represents the small kernel network, and “C” represents the cross supervision loss, “sup” represents the supervision loss, and “cont” represents the contrastive loss

cross-supervision that does not require pre-training weights and is not sensitive to data volume differences.

Specifically, we use dual CNN networks with the same architecture but different convolution kernel sizes. The training data are transferred to the CNN networks with a small kernel and a large kernel to obtain their segmented prediction maps based on different receptive fields, which are considered as pseudo-labels for each other's data. The dual CNN network avoids the data pressure brought by a transformer and reduces the dependence on pre-training weight. The loss of the dual CNN module includes supervised loss and unsupervised loss, supervised loss through supervised training of labeled data, and loss function using the sum of cross-entropy loss and DICE loss. The formula is as follows:

$$\mathcal{L}_{\text{sup}} = \mathcal{L}_{ce}\left(p_i, y_i\right) + \mathcal{L}_{\text{dice}}\left(p_i, y_i\right) \tag{1}$$

For the unsupervised loss of unlabeled data, small-kernel CNN and large-kernel CNN generate pseudo-label sums by using the prediction graphs generated by the other side, respectively, and their formulas are as follows:

$$P_s^{pl} = \arg\max\left(P_{\text{Cl}}\right), \; P_l^{pl} = \arg\max\left(P_{\text{Cs}}\right) \tag{2}$$

In the equation, the label of the small-kernel network is obtained by the maximum index operation of the prediction graph generated by the large-kernel network, and the pseudo-label of the large-kernel network is obtained in the same way. The unsupervised loss consists of the loss of a small-kernel network and the loss of a large-kernel network. The formula is as follows:

$$\mathcal{L}_{\text{ctl}} = \mathcal{L}_{Cs} + \mathcal{L}_{Cl} \tag{3}$$

As shown in the formula, the unsupervised losses are calculated using DICE losses, and the DICE losses of the dual networks are added to obtain unsupervised losses for unlabeled data.

3.3 Contrastive learning modules

To make efficient use of data features, we add a contrastive learning module to the dual CNN networks. A projection header is introduced at the bottom of the dual CNN networks to map the data to D-dimension vector; we consider two different coding results of the same sample in the batch as positive pairs and the coding results of other samples as negative pair. Compared with the data enhancement, we make full use of the characteristics of the dual CNN network structure. The samples mapped by CNN encode can be regarded as data enhancement. We record the vector of data X_n through $f_{\text{Cs}}(\cdot)$ and $f_{\text{Cl}}(\cdot)$ goes through $proj(\cdot)$ as x_n^s and x_n^l, and we define the contrastive loss as

$$L_{\text{cont}} = -\frac{1}{B}\sum_{i}^{B} \log \frac{e^{\cos\left(x_i^s, x_i^l\right)/\tau}}{\sum_j e^{\cos\left(x_i^s, x_j^l\right)/\tau}} \tag{4}$$

where B is the number of samples in the current batch, τ represents the temperature coefficient, and cos is the cosine similarity; they are as close as possible to each other in the high-dimensional feature space after the same sample is coded at different scales. The contrastive loss can close the similar sample pairs and push away the dissimilar sample pairs.

3.4 Total loss

In our work, if $X_n \in D^{\text{L}}$, then supervised training is performed on the data, and if $X_n \in D^{\text{U}}$, then dual CNN ($f_{\text{Cs}}(\cdot)$ and $f_{\text{Cl}}(\cdot)$) cross-teaching supervised learning is performed on the unlabeled data. All data would be mapped by the two CNNs to generate features, which are used for contrastive learning to optimize the parameters, Overall, our total loss function consists of three components, the supervised loss $\mathcal{L}_{\text{sup}}$ of labeled data, the cross-supervised loss $\mathcal{L}_{\text{ctl}}$ of unlabeled data, and the global contrastive loss $\mathcal{L}_{\text{cont}}$ of all data.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sup}} + \alpha\mathcal{L}_{\text{ctl}} + \lambda\mathcal{L}_{\text{cont}} \tag{5}$$

4 Experiments

4.1 Dataset and evaluation metrics

We validated the validity of our model on the public benchmark dataset ACDC[1] (Bernard et al. [25]), which contains 200 annotated images from 100 patients. The segmentation masks of the left ventricle (LV), myocardium (Myo), and right ventricle (RV) are used for clinical and algorithm research. To be fair, we used Luo's [12] work as a reference and performed the same data partitioning and preprocessing operations: 70 training samples, 10 validation samples, and 20 samples for testing; the sample data is scaled to 256x256 according to slices for 2D segmentation. In the prediction stage, the results of two-dimensional segmentation are stacked to three-dimensional for evaluation. We chose DSC and HD_{95}, which are the most commonly used in

[1] https://www.creatis.insa-lyon.fr/Challenge/acdc/databases.html.

Table 1 Ablation experiment

Model	No loss of contrast			Added the contrastive loss		
	DSC	HD_{95}	Time(s)	DSC	HD_{95}	Time(s)
CNN (3)_CNN (3)	86.87	6.8	**5738**	87.03	6.2	**5531**
CNN (3)_CNN (5)	**86.89**	**6.6**	7612	**87.19**	**6.1**	7627
CNN (3)_CNN (7)	86.86	7.3	8863	87.07	6.3	8757
CNN (3)_CNN (9)	86.29	7.7	10354	86.94	7.6	10214
CNN (3)_CNN (11)	86.51	8.3	12315	86.65	6.7	12188
CNN (3)_CNN (13)	85.86	6.7	14831	86.75	6.5	14600
CNN (3)_CNN (15)	85.04	8.7	17886	86.21	6.8	17653

Boldface means the best result in the same experiment group

semi-supervised segmentation. The Dice similarity coefficient indicates the overlap between the predicted result and the real label map. The HD coefficient explained the maximum mismatch between the predicted results and the real tag images. The evaluation criteria are defined as follows:

$$DSC = \frac{2(A \cap B)}{A + B} \tag{6}$$

$$HD_{95} = \max_{k95\%}[d(A, B), d(B, A)] \tag{7}$$

4.2 Implementation details

In our proposed method, we utilize dual U-Net networks with the same architectures but different convolution kernels. One U-Net network uses 3*3 kernel size for all convolutional layers the same as in Luo's work, and all convolutional layers in the other U-Net are replaced with large kernels as the large kernel network. All models of our ablation analysis and comparison experiments were implemented using the PyTorch library and trained on an Nvidia A100 GPU. The batch size for all models was set to 16, with 8 labeled and 8 unlabeled data, for 30k iterations using the SGD optimizer. The Poly strategy is used to adjust the learning rate, where the initial learning rate is set to 0.01. The mapping head of the contrastive module consists of one layer of global average pooling and two layers of full connection, and finally maps to a 256-dimension vector. The temperature coefficient in the contrastive loss is set to 0.07, and the weight of the contrastive loss is described in a later experiment. The coefficient of cross-supervision loss is set as follows, using the experience of Luo's work for reference.

$$\alpha(t) = 0.1 \cdot e^{\left(-5\left(1-\frac{t_i}{t_{\text{total}}}\right)^2\right)} \tag{8}$$

4.3 Results and analysis

To explore the effectiveness of the proposed modules, we have done a wealth of ablation experiments. Based on 3x3 small kernel U-Net and 10% labeled data, we explore the cross-teaching of different large kernel networks; on this basis, a contrastive learning module is added to further prove the generalization performance of contrastive loss. The results are as follows.

As shown in Table 1, DSC and HD_{95} both achieve good performance when the convolutional kernel of a large-kernel network is set to 5. The effect of the model decreases with the increase of the convolution kernel because the attention to detail of the large-kernel model becomes less and less as the convolution kernel increases, and the learning performance of the small-kernel network is further affected by cross-teaching. However, when the convolutional kernel reaches 9x9, there is an obvious drop, which is because the features extracted by 3x3 and 9x9 convolutional kernels produce redundancy, and the network cannot learn more abundant semantic information. When the convolution kernel reaches 15x15, the model degradation effect is very obvious. After adding a contrastive learning module to the model, the results of all models are improved. The experimental data show the effectiveness and generalization performance of the proposed model, which is based on the loss of contrast and the pull-in of positive samples and the push-out of negative samples.

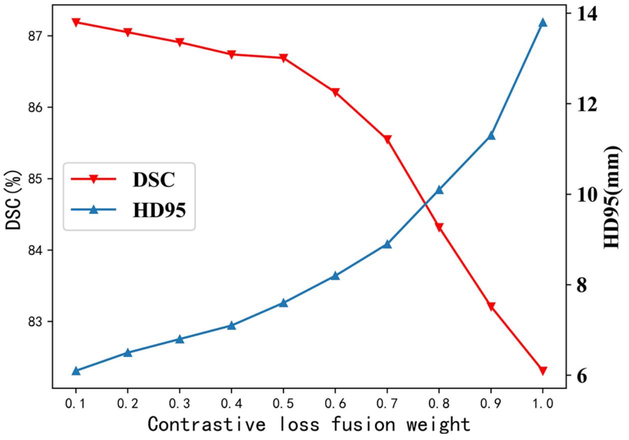


Fig. 2 Compares the impact of loss weights on model performance

Table 2 Comparison experiment

Labeled data	Method	RV		Myo		LV		Mean	
		DSC	HD_{95}	DSC	HD_{95}	DSC	HD_{95}	DSC	HD_{95}
	MT	0.469	40.5	0.63	15.8	0.713	30.3	0.604	28.9
	DAN	0.464	23.03	0.533	27.7	0.63	26.3	0.542	25.7
	EM	0.447	32.4	0.628	19	0.731	20.9	0.602	24.1
5%	UAMT	0.508	35.4	0.615	19.3	0.707	22.6	0.61	25.8
	CPS	0.438	35.8	0.652	18.3	0.72	22.2	0.603	25.5
	CTCT	0.577	21.4	0.628	11.5	**0.763**	15.7	0.656	16.2
	Ours	**0.679**	**18**	**0.66**	**4.9**	0.757	**8.7**	**0.699**	**10.6**
	MT	0.816	5.4	0.821	10.3	0.871	15.2	0.836	10.3
	DAN	0.822	7.7	0.792	12.7	0.87	17.1	0.828	12.5
	EM	0.818	4	0.826	7.8	0.87	19.5	0.838	10.5
10%	UAMT	0.843	**3.5**	0.822	9.8	0.884	13.1	0.849	8.8
	CPS	0.854	3.8	0.837	5.1	0.882	13.6	0.858	7.5
	CTCT	**0.867**	6.4	0.834	8.5	0.893	**10.5**	0.865	8.5
	Ours	**0.867**	4.5	**0.843**	**3.1**	**0.905**	10.7	**0.872**	**6.1**

Boldface means the best result in the same experiment group

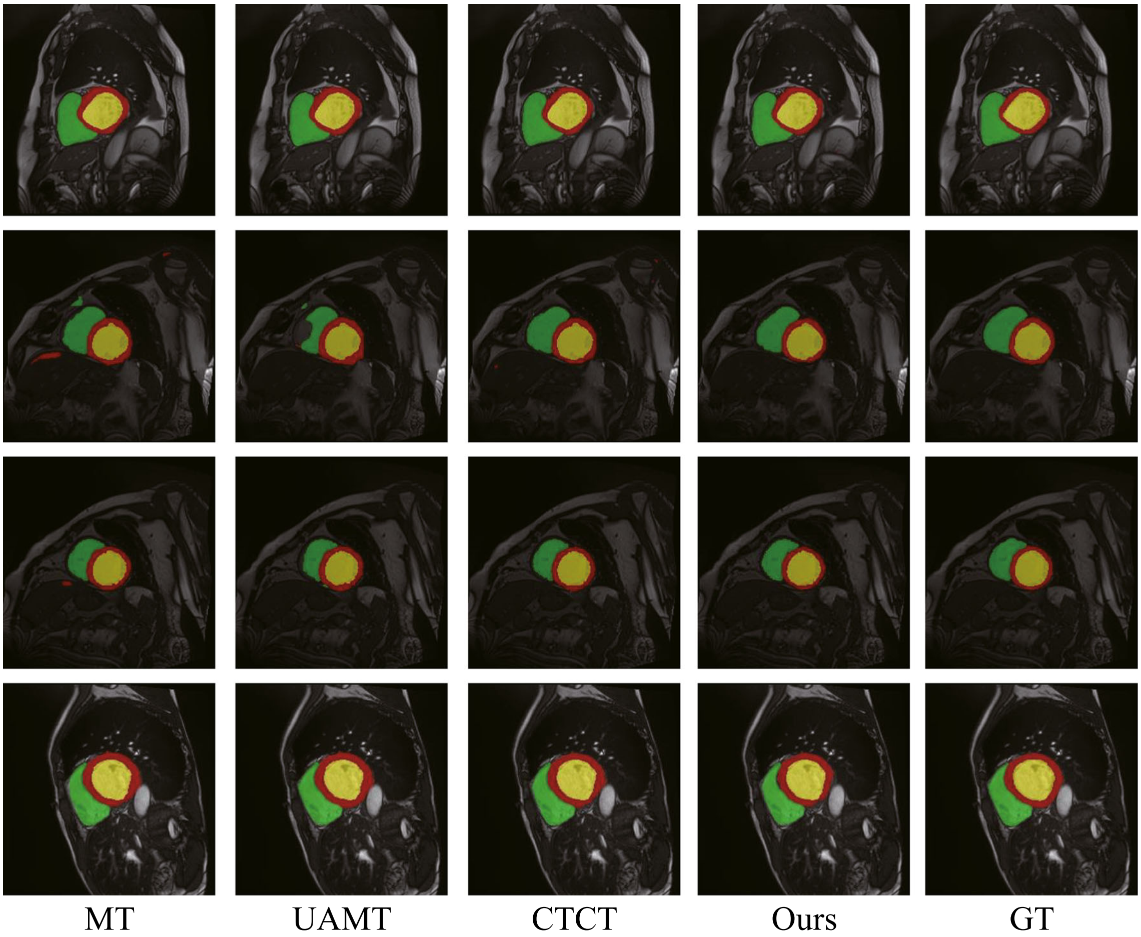


Fig. 3 Visual comparison of different methods on 10% labeled data

4.4 Contrastive loss fusion weight exploration

In order to find out the best weight of the contrastive loss weight, we take the dual CNN network with a large convolutional kernel 5x5 as the basic model, from 0.1 to 1 ran the predictions separately to see how the model worked, as shown in Fig. 2. As can be seen from the graph above, the model works best when the value is 0.1, when the DSC coefficient is the highest, and the HD_{95} coefficient is the lowest; because our loss of contrast is based on the global nature, it plays an auxiliary role in the segmentation model. When the loss of contrast is set too large, the loss of contrast will dominate the total loss; this is considered unreasonable in the pixel-level classification model. Therefore, we set the contrastive loss weight to 0.1, and all model weights for ablation experiments were also based on 0.1.

4.5 Comparison with other methods

To illustrate the effectiveness of our model, we compared the dual-CNN network with a large kernel of 5 with the recently popular six semi-supervised segmentation methods on ACDC datasets. The six methods are (1) mean teacher (MT) [15]; (2) deep adversarial network (DAN) [25]; (3) entropy minimization (EM) [26]; (4) uncertainty-aware mean teacher (UAMT) [7]; (5) cross pseudo supervision (CPS) [27]; and (6) cross-teaching between CNN and trans-former (CTCT) [12]. For fair comparison [30], all comparison method backbones use U-Net as the backbones. As can be seen from the table above, our method and other method were tested on 5% (3 samples) and 10% (7 samples) labeled data, respectively, and our method achieved the best results; in particular, at 5% labeling, all of our metrics greatly outperformed all models, and we achieved a 4.3% and −5.6 mm improvement in DSC and HD_{95}, respectively, compared with the CTCT segmentation method; because the amount of labeled data is too small, the effect of supervised learning is weak, and the contrastive learning module is not dependent on the number of tags, at this time for the model learning help is very obvious (Table 2). Also on 10% of labeled data, our model performed best on most metrics, increasing 0.7% and −2.4 mm on average DSC and HD_{95} compared with CTCT, respectively; the effectiveness of our method is proved by comparative experiments. To visualize the segmentation effect of our method, we visualized our method with other models, and the effect is shown in Fig. 3. The improvement of our proposed method relative to the MT and UAMT methods is obvious, and our model is more accurate for region segmentation and does not have missing regions because our method improves 3.6% and -3.9mm in DSC and HD_{95}, respectively, compared with MT, and 2.3% and -2.7mm in UAMT. Compared with CTCT, our method can handle the edge details better, and the segmentation result is closer to ground truth.

5 Conclusion

In this paper, we propose a new type of semi-supervised segmentation network, which introduces large kernel convolution into semi-supervised learning. We explore the cross-teaching performance of small kernel convolution networks and different large kernel networks, and we also designed a contrastive learning module, using the characteristics of the dual CNN network structure to carry out contrastive learning, and also carried out a series of studies on the weight of contrastive loss, it is proved that image-level contrastive learning is effective for semi-supervised segmentation. After summing up, we designed the best-performing method. Compared with many models, our method has obvious improvement in the medical segmentation ACDC dataset. It is proved that large receptive field is helpful to model complementary learning to some extent, but when the convolutional kernel is too large, the network complexity will increase.

In this paper, the cross-supervised loss weight of our dual networks is the same, but the actual training is usually faster for one network. Therefore, if the dynamic loss weight can be designed according to the model effect, it will guide the model's loss to decrease faster and achieve better learning effect, which is worth studying.

Declarations

Ethical approval This is a numerical simulation study for which no ethical approval was required.

References

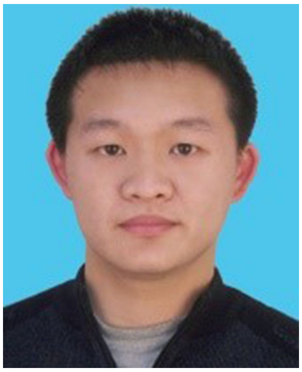
1. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition pp 3431–3440
2. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, pp 234–241
3. Isensee F, Petersen J, Klein A et al (2018) nnu-net: self-adapting framework for u-net-based medical image segmentation. arXiv preprint arXiv:1809.10486
4. Luo X, Liao W, Chen J et al (2021) Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, pp 318–329
5. Bai W, Oktay O, Sinclair M et al (2017) Semi-supervised learning for network-based cardiac MR image segmentation. International

Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, pp 253–260
6. Luo X, Chen J, Song T et al (2021) Semi-supervised medical image segmentation through dual-task consistency. Proc AAAI Conf Art Intell 35(10):8801–8809
7. Yu L, Wang S, Li X et al (2019) Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, pp 605-613
8. Lee DH (2013) Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks. Workshop on challenges in representation learning, ICML 3(2):896
9. Li X, Yu L, Chen H et al (2018) Semi-supervised skin lesion segmentation via transformation consistent self-ensembling model. arXiv preprint arXiv:1808.03887
10. Wang X, Yuan Y, Guo D et al (2011) SSA-Net: spatial self-attention network for COVID-19 pneumonia infection segmentation with semi-supervised few-shot learning. Med Image Anal 79:102459
11. Li C, Dong L, Dou Q et al (2021) Self-ensembling co-training framework for semi-supervised COVID-19 CT segmentation. IEEE J Biomed Health Inform 25(11):4140–4151
12. Luo X, Hu M, Song T et al (2021) Semi-supervised medical image segmentation via cross teaching between CNN and Transformer. arXiv preprint arXiv:2112.04894
13. Sajjadi M, Javanmardi M, Tasdizen T (2019) Regularization with stochastic transformations and perturbations for deep semi-supervised learning. Advances in neural information processing systems 29
14. Laine S, Aila T (2016) Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:1610.02242
15. Tarvainen A, Valpola H (2017) Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems 30
16. He K, Fan H, Wu Y et al (2020) Momentum contrast for unsupervised visual representation learning. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9729-9738
17. Chen X, Fan H, Girshick R et al (2020) Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297
18. Chen T, Kornblith S, Norouzi M et al (2020) A simple framework for contrastive learning of visual representations. International conference on machine learning. PMLR, 1597-1607
19. Chen T, Kornblith S, Swersky K et al (2020) Big self-supervised models are strong semi-supervised learners. Advances in Neural Information Processing Systems 33:22243–22255
20. Zhong Y, Yuan B, Wu H et al (2021) Pixel contrastive-consistent semi-supervised semantic segmentation. Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 7273-7282
21. Zhang S, Zhang J, Tian B et al (2023) Multi-modal contrastive mutual learning and pseudo-label re-learning for semi-supervised medical image segmentation. Med Image Anal 83:102656
22. Xiao Z, Su Y, Deng Z et al (2022) Efficient combination of CNN and Transformer for dual-teacher uncertainty-aware guided semi-supervised medical image segmentation. Available at SSRN 4081789
23. Liu Y, Wang W, Luo G et al (2022) A contrastive consistency semi-supervised left atrium segmentation model. Comput Med Imaging Graph, 2022, 99:102092
24. Wang T, Lu J, Lai Z et al (2022) Uncertainty-guided pixel contrastive learning for semi-supervised medical image segmentation. Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, pp 1444–1450
25. Zhang Y, Yang L, Chen J et al (2017) Deep adversarial networks for biomedical image segmentation utilizing unannotated images. International conference on medical image computing and computer-assisted intervention. Springer, Cham, pp 408-416
26. Vu TH, Jain H, Bucher M et al (2019) Advent: adversarial entropy minimization for domain adaptation in semantic segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 2517-2526
27. Chen X, Yuan Y, Zeng G et al (2021) Semi-supervised semantic segmentation with cross pseudo supervision. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 2613-2622
28. Zhao Z, Hu J, Zeng Z et al (2022) MMGL: multi-scale multi-view global-local contrastive learning for semi-supervised cardiac image segmentation. 2022 IEEE International Conference on Image Processing (ICIP) pp 401-405
29. Bernard O, Lalande A, Zotti C et al (2018) Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans Med Imaging 37(11):2514–2525
30. Wilkinson MD, Dumontier M, Aalbersberg IJJ et al (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3(1):1–9
31. You C, Dai W, Min Y et al (2023) Action++: improving semi-supervised medical image segmentation with adaptive anatomical contrast. arXiv preprint arXiv:2304.02689
32. Wu H, Li X, Lin Y et al (2023) Compete to win: enhancing pseudo labels for barely-supervised medical image segmentation. IEEE Trans Med Imaging
33. Zhu J, Bolsterlee B, Chow BVY et al (2023) Hybrid dual mean-teacher network with double-uncertainty guidance for semi-supervised segmentation of MRI scans. arXiv preprint arXiv:2303.05126

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Keming Li is a master's student at the School of Information Science & Electric Engineering, at Shan dong Jiaotong University. His research interest covers deep learning, semi-supervised image segmentation, and digital image processing. Published an EI paper, participated in more than 10 national, provincial and municipal projects. In 2022, she was selected as a member of the National Innovation and Entrepreneurship Project.

Guangyuan Zhang is dean of School of Information Science & Electric Engineering, Shandong Jiaotong University. He received his PhD degree in Computer software and theory from Northeastern University and finished his post-doctoral study in Tsinghua University. He is deputy director of the Institute of traffic information engineering and control in Shandong Jiaotong University, a member of Shandong Province Higher Education Association Computer Teaching Research Committee and deputy director member of Shandong Information Association Information security Committee. His areas of interest include pattern recognition, digital image processing, intelligent traffic and intelligent vehicle. He led and participated in more than 20 national, provincial and municipal projects; published more than 40 papers at all levels of academic journals and international conferences; published 7 works, including 1 treatise.



Kefeng Li is associate professor of School of Information Science & Electric Engineering, Shandong Jiaotong University. He received his PhD degree in Computer Science from University of Central Lancashire. He is member of Professional Committee of Intelligent Interaction from Chinese Association for Artificial Intelligence. His areas of interest include digital image processing, video analysis and biometric. He led a project from National Natural Science Foundation of Shandong Province and participated in more than 10 national, provincial and municipal projects. He has published more than 10 EI papers, published 1 treatise and applied 3 national patents, 4 utility model patents.

Jindi Li is a master's student at the School of Information Science & Electric Engineering, at Shan dong Jiaotong University. His research interest covers deep learning, pattern recognition, and digital image processing. Published one sci paper, participated in more than 10 national, provincial and municipal projects. In 2022, as the person in charge, was selected for the national innovation and entrepreneurship project.

Jiaqi Wang is a master's student at the School of Information Science & Electric Engineering, at Shan dong Jiaotong University. Her research interests include deep learning, medical image processing and 3D reconstruction of medical images.She published an EI paper and participated in more than 10 national, provincial and municipal projects. In 2022, she was selected as a member of the National Innovation and Entrepreneurship Project.

Yumin Yang is a master's student at the School of Information Science & Electric Engineering, at Shan dong Jiaotong University. Her research interests include deep learning, image processing and gesture recognition. She published an EI paper and participated in more than 10 national, provincial and municipal projects. In 2022, she was selected as a member of the National Innovation and Entrepreneurship Project.