2024 IEEE International Symposium on Biomedical Imaging (ISBI) | 979-8-3503-1333-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISBI56570.2024.10635274

SEMI-CONTRANS: SEMI-SUPERVISED MEDICAL IMAGE SEGMENTATION VIA MULTI-SCALE FEATURE FUSION AND CROSS TEACHING OF CNN AND TRANSFORMER

Weiren Zhao[1], *Lanfeng Zhong*[2,3], *Guotai Wang*[2,3*]

[1]School of Computer Science and Engineering, [2]School of Mechanical and Electrical Engineering
University of Electronic Science and Technology of China, Chengdu, China
[3]Shanghai Artificial Intelligence Laboratory, Shanghai, China

ABSTRACT

Convolutional Neural Networks (CNNs) and Transformers have achieved promising results in fully supervised medical image segmentation. However, acquiring high-quality annotations for medical images is prohibitively expensive, making semi-supervised learning a promising way to reduce the annotation cost by leveraging both labeled and unlabeled images for training. In this work, we propose a novel model named **Semi-ConTrans** that unifies the advantages of CNNs and Transformers through multi-scale feature fusion and cross teaching for semi-supervised segmentation. Specifically, to leverage localization capability from CNNs and global context modeling of self-attention in Transformers in a unified framework, we adaptively fuse them at multiple scales in the encoder. Furthermore, we use a CNN decoder and a Transformer decoder with different decision boundaries for cross teaching, obtaining more holistic pseudo labels for dealing with unlabeled images. Experiments on the ACDC dataset of cardiac images demonstrate that our approach greatly improves the performance with only 10% or 20% labeled images by exploiting unlabeled images, outperforming eight state-of-the-art semi-supervised segmentation methods.

***Index Terms*—** Semi-supervised learning, CNN, Transformer, Attention.

1. INTRODUCTION

Medical image segmentation plays an important role in quantitative measurement of lesions and organs for disease diagnosis and treatment planning. Under the main frameworks of Convolutional Neural Networks (CNNs) [1] and Vision Transformer (ViT) [2], deep learning methods in medical image segmentation have achieved remarkable performance when trained on large-scale pixel-level annotated datasets. However, obtaining full annotation with pixel-level labels is labor-intensive and extremely expensive. In order to reduce annotation costs, Semi-Supervised Learning (SSL) methods [3] that utilize a large amount of unlabeled images and limited labeled images have become an attractive solution.

In semi-supervised medical image segmentation, a common method is pseudo-labeling, which generates pseudo labels for unlabeled images to provide additional supervision [4]. Another popular technique is consistency regularization, such as teacher-student consistency [5, 6] and Cross Consistency Training (CCT) [7]. However, most existing SSL frameworks rely only on CNNs that have limitations in capturing long-range dependencies of dense prediction tasks. In contrast, the Transformer architecture leveraging self-attention shows superior performance in modeling long-range dependencies. Therefore, effectively combining Transformers and CNNs has become an emerging technology in the field of computer vision. However, combining the advantages of CNNs and Transformers have rarely been investigated for semi-supervised medical image segmentation. Though Luo et al. [8] proposed cross supervision between a CNN and a ViT, there is no any interaction between their feature extraction stage, which leads to a limited ability to combine local and long-range features.

In this work, we propose Semi-ConTrans, a network that deeply integrates CNN and ViT encoders along with dual CNN and ViT decoders for semi-supervised medical image segmentation. The Semi-ConTrans encoder fuses CNN and ViT features at multiple scales to obtain both local details and global contexts. The fused representations are decoded through separate CNN and ViT branches, which provide complementary pseudo labels on unlabeled images for cross teaching. This unified CNN-ViT architecture enables joint multi-scale feature learning on labeled and unlabeled images through the integrated encoder dual-decoder design. Our main contributions can be summarized as follows:

- We propose an end-to-end SSL framework Semi-ConTrans that uses a reciprocal fusion strategy to integrate the complementary feature learning ability of CNNs and ViTs.
- An integrated encoder with dual decoders is proposed for leveraging unlabeled images, where cross teaching between CNN and ViT decoders is introduced.
- We demonstrate state-of-the-art semi-supervised segmen-

*Corresponding author (guotai.wang@uestc.edu.cn)
Weiren Zhao and Lanfeng Zhong—Equal contribution

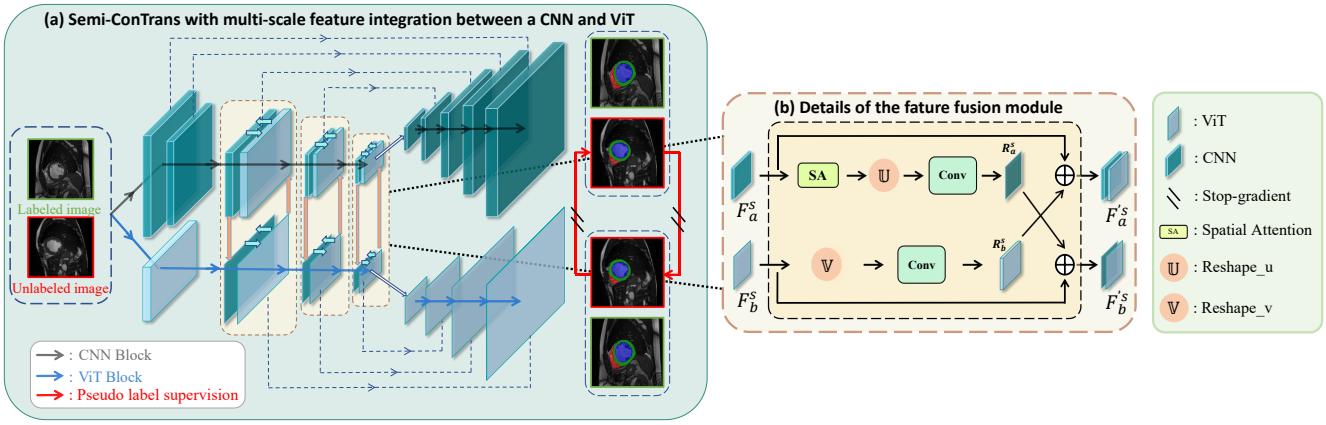

Fig. 1: The overall architecture of our **Semi-ConTrans** for semi-supervised segmentation. We propose multi-scale feature fusion between a CNN and a ViT in the encoder to enhance the feature representation ability, and introduce cross teaching between CNN and ViT decoders to deal with unlabeled images.

tation performance through extensive experiments on a public benchmark dataset for cardiac image segmentation.

2. METHOD

For semi-supervised learning, let D_l denote the set of labeled images, and D_u denote the set of unlabeled images, and the entire training set is $D = D_l \cup D_u$. Our proposed Semi-ConTrans framework is illustrated in Fig. 1, and it deeply integrates a CNN and a ViT for semi-supervised segmentation. In the encoder part, we fuse local features from a U-Net [9] and long-range features from a Swin-UNet [10] at multiple scales. At the same time, we propose cross teaching between Swin-UNet and U-Net decoders by mutually generating pseudo labels to leverage the unlabeled images.

2.1. Multi-scale fusion for CNN and ViT encoders

Aiming to simultaneously utilize the local feature representation of CNNs and the global modeling capability of Transformers, we propose a novel fusion scheme to enhance the encoder's understanding of the input image. As shown in Fig. 1, both the CNN encoder and the ViT encoder extract features at multiple scales or depths. To better leverage both local and global features for the two types of networks, we introduce multi-scale bidirectional feature fusion between the two encoders, and use a fusion module at each of the last S scales of the encoders, i.e., $S = 3$ in this work.

Specifically, we use $F_a^s \in \mathbb{R}^{C_a \times H \times W}$ and $F_b^s \in \mathbb{R}^{N \times C_b}$ to denote the feature map at the last s-th scale of the CNN and ViT encoders, respectively, where C_a and C_b are channel numbers. H and W represent the height and width of the feature map, and $N = HW$ is the number of tokens in the ViT encoder at scale s.

As the ViT encoder's feature F_b^s has an in-design spatial attention, we also apply a spatial attention to the CNN encoder's feature F_a^s to improve its spatial awareness. Inspired by [11, 12], we use average pooling (P_{avg}^C) and max pooling (P_{max}^C) across the channel dimension, and use a convolutional layer followed by sigmoid activation σ to obtain the spatial attention map for feature calibration:

$$\tilde{F}_a^s = F_a^s \cdot \sigma\left(Conv^{7\times7}\left(P_{avg}^C(F_a^s) \oplus P_{max}^C(F_a^s)\right)\right) \quad (1)$$

where $\oplus$ denotes concatenation of feature maps. $\tilde{F}_a^s \in \mathbb{R}^{C_a \times H \times W}$ is the output of spatial attention. $\mathrm{Conv}^{7\times7}$ represents convolution with a kernel size of 7×7 and output channel of 1. Then a reshape operation ($Reshape\_u$) is used to reshape $\tilde{F}_a^s$ from a 2D feature map ($C_a \times H \times W$) to a 1D token ($HW \times C_a$) to match the spatial size of F_b^s. As $\tilde{F}_a^s$ and F_b^s may have different channel numbers, we also use a point-wise convolutional layer ($\mathrm{Conv}^{1\times1}$) to map the channel number from C_a to C_b, and the output is denoted as:

$$R_a^s = Conv^{1\times1}\left(Reshape\_u(\tilde{F}_a^s)\right) \quad (2)$$

Similarly, we use another reshape operation ($Reshape\_v$) to reshape F_b^s from a 1D token ($HW \times C_b$) to a 2D feature map ($C_b \times H \times W$), and use a point-wise convolutional layer to map the channel number to C_a:

$$R_b^s = Conv^{1\times1}\left(Reshape\_v(F_b^s)\right) \quad (3)$$

Finally, we integrate ViT feature to the CNN encoder by adding R_b^s to F_a^s, and integrate CNN feature to the ViT encoder by adding R_a^s to F_b^s, the output features of the fusion module for the CNN and ViT branches is denoted as F'^s_a and F'^s_b, respectively:

$$F'^s_a = R_b^s + F_a^s; \quad F'^s_b = R_a^s + F_b^s \quad (4)$$

The fused features ${F'}_a^s$ and ${F'}_b^s$ are sent to the corresponding decoders through skip connections. Except for the last scale, they are also sent to the next stage of the two encoders, respectively. Based on the fused features, the input of the decoders has a better representation of local details and global contexts, which helps to improve the segmentation performance.

2.2. Cross teaching between CNN and ViT decoders

To further utilize the fused feature information between CNN and ViT, we propose a cross teaching approach between CNN decoder and ViT decoder. By interpreting the fused feature from two different perspectives with different decision boundaries, our method can avoid the bias of a single decoder and alleviate over-fitting to potentially wrong pseudo labels.

The two decoders of our method are implemented with that from U-Net [9] and Swin-UNet [2], respectively. For an input image x, the probability prediction maps from the CNN decoder and the ViT decoder are denoted as p^c and p^t, respectively. We use $argmax$ to convert them into one-hot pseudo labels:

$$y^c = argmax(p^c); \quad y^t = argmax(p^t) \tag{5}$$

For an unlabeled image, each decoder is supervised by pseudo labels from the other decoder, and the cross teaching loss for the unlabeled data is defined as:

$$L_{ct} = \frac{1}{|D_u|} \sum_{x \in D_u} \big(L_{dice}(p^c, y^t) + L_{dice}(p^t, y^c) \big) \tag{6}$$

where L_{dice} is the standard Dice loss function for segmentation. In addition, a supervision loss L_{sup} consisting of cross-entropy loss L_{ce} and Dice loss is used for the labeled images:

$$L_{sup} = \frac{1}{|D_l|} \sum_{(x,y) \in D_l} \sum_{p \in \{p^c, p^t\}} \big(L_{ce}(p, y) + L_{dice}(p, y) \big) \tag{7}$$

where y represents the ground truth of image x. The overall loss function for SSL is defined as:

$$L = L_{sup} + \lambda L_{ct} \tag{8}$$

where λ is a hyper-parameter to control the weight of L_{ct}.

3. EXPERIMENTS AND RESULTS

3.1. Dataset and evaluation metrics

We validated our proposed approach using the ACDC dataset of cardiac cine-Magnetic Resonance Images (MRI) [17] for segmentation of three cardiac substructures: Left Ventricle (LV), Right Ventricle (RV), and Myocardium (MYO). It comprises 200 scans from 100 patients, and was randomly divided at patient level into 70%, 10% and 20% for training, validation and testing, respectively. We used 14 and 28 labeled scans (corresponding to annotation ratio of 10% and 20%) for semi-supervised learning, respectively. Due to the large inter-slice spacing (5-10 mm), 2D networks were used for slice-by-slice segmentation. For preprocessing, all the slices were resized to 256 $\times$ 256. During the inference phase, the slice-level predictions were stacked into a 3D volume for quantitative evaluation in terms of Dice Similarity Coefficient (DSC) and 95% Hausdorff Distance (HD_{95}).

3.2. Implementation details

The CNN structure and ViT structure in our Semi-ConTrans were implemented with U-Net [9] and Swin-UNet [2], respectively. In order to make the ViT have the same computational efficiency as the CNN and to allow them to complement each other, we made the following settings: Patch size: 224 × 224, embedding dimension: 96, Number of heads in self-attention: 3, 3, 6, and 12, Window size: 7, and each resolution level of encoder/decoder in Swin-UNet has two Swin-Transformer-based blocks. All the methods were implemented using the PyTorch framework and PyMIC [3] on a Ubuntu desktop with a GTX2080TI GPU. We used Stochastic Gradient Descent (SGD) for training with 1300 epochs, and the batch size was 24 where 12 images were labeled. We employed the poly learning rate policy to adjust the learning rate that was initialized to 0.01. The weighting factor λ was defined by the commonly used time-dependent Gaussian warm-up function: $\lambda(t) = 0.1 \cdot e^{-5(1-\frac{t}{t_{total}})^2}$, where t represents the current training epoch and t_{total} is the total number of epochs. After training, we used the CNN decoder for inference, due to its higher efficiency and accuracy than the ViT decoder.

3.3. Comparison with existing semi-supervised methods

To validate our proposed method, we conducted comparative experiments under the same settings. First, we set the baseline as training U-Net [9] or Swin-UNet [10] with fully supervised learning from D_l. We then compared our method with eight state-of-the-art SSL methods: 1) Mean Teacher (MT) [5] that employs self-ensembling to provide pseudo labels for supervision, 2) Uncertainty Aware Mean Teacher (UAMT) [6] that incorporates uncertainty estimation into the mean teacher framework to weight predictions, 3) Interpolation Consistency Training (ICT) [15] that promotes consistency between predictions of interpolated images, 4) Cross Pseudo Supervision (CPS) [4] that utilizes pseudo labels from two independent networks for cross supervision, 5) Uncertainty Rectified Pyramid Consistency (URPC) [16] that enforces multi-scale prediction consistency with uncertainty rectification, 6) CCT [7] that encourages consistency between a primary decoder and multiple auxiliary decoders, 7) Deep Co-Training (DCT) [13] that minimizes the expected Jensen-Shannon divergence between two networks, and 8)

Table 1: Quantitative comparison of different SSL methods with two different annotation ratios. * denotes significant improvement (p-value < 0.05) from the best existing method using a paired Student's t-test.

Method	Annotation ratio: 10%					Annotation ratio: 20%				
	DSC (%)				HD_{95} (mm)	DSC (%)				HD_{95} (mm)
	RV	Myo	LV	Mean	Average	RV	Myo	LV	Mean	Average
baseline (ViT)	67.57±18.72	68.41±13.41	77.24±16.85	71.08±16.32	10.41±13.61	75.92±17.27	76.16±9.31	82.57±15.36	78.22±13.98	8.26±11.80
baseline (CNN)	65.69±26.82	76.71±10.57	83.44±14.16	75.28±17.18	7.98±12.46	72.67±26.83	80.78±7.42	86.07±12.25	79.84±15.50	7.98±12.41
DCT [13]	74.21±22.05	76.31±11.97	83.01±16.93	77.84±16.98	7.64±10.32	73.21±26.67	81.23±8.01	87.52±10.49	80.66±15.06	7.64±9.75
CPS [4]	77.48±19.63	79.90±7.90	85.45±11.98	80.92±13.17	6.52±10.78	81.65±17.92	83.27±6.31	88.94±6.94	84.62±10.39	4.89±9.65
UAMT [6]	70.68±26.51	77.39±9.80	83.37±13.34	78.07±16.55	8.24±14.32	75.42±26.85	82.18±6.72	88.16±9.97	81.92±14.51	6.36±12.57
MT [5]	77.53±19.16	76.24±11.02	83.08±15.47	78.95±15.21	7.60±11.69	74.91±25.01	82.00±6.98	87.06±10.61	81.32±14.20	6.74±12.34
R-Drop [14]	72.95±23.04	78.68±9.68	85.01±12.66	78.88±15.12	6.82±13.50	77.72±23.33	82.88±6.15	89.10±8.87	83.23±12.78	6.31±11.73
CCT [7]	78.93±17.91	80.10±6.40	86.40±11.31	81.81±11.90	6.95±10.81	79.98±22.34	81.54±8.27	86.27±12.82	82.59±14.48	5.10±10.95
ICT [15]	77.63±19.40	79.77±9.17	86.54±10.65	81.12±13.07	9.62±10.58	77.21±22.48	82.88±6.01	88.75±8.70	82.95±12.40	9.62±11.29
URPC [16]	68.48±25.85	76.12±14.57	84.31±15.21	76.30±18.54	4.78±12.67	76.91±24.50	81.67±8.72	88.01±9.98	82.19±14.40	4.64±11.92
Ours	**83.13±8.71***	**82.65±6.96***	**89.30±9.63***	**85.02±8.43***	**4.45±7.42***	**87.22±8.45***	**85.65±4.67***	**91.65±7.41***	**88.20±6.84***	**2.54±3.81***

Table 2: DSC (%) of different variants of our method with annotation ratio being 10%.

Method	RV	Myo	LV	Average
Two CNNs	77.48±19.63	79.90±7.90	85.45±11.98	80.92±13.17
Two ViTs	76.79±12.42	75.15±10.15	84.52±12.94	78.82±11.84
CNN&ViT	80.04±9.77	81.27±6.51	88.36±9.80	83.22±8.69
CNN&ViT (ViT output)	75.96±14.98	77.19±9.56	84.28±13.40	79.14±12.64
Ours (ViT output)	79.63±11.61	78.29±9.92	85.81±12.76	81.05±11.43
Ours (one fusion)	81.09±11.06	82.13±8.93	88.97±9.94	84.06±9.97
Ours	**83.13±8.71**	**82.65±6.96**	**89.30±9.63**	**85.02±8.43**

Regularized Dropout (R-Drop) [14] that generates predictions via dropout at test time for consistency regularization.

As demonstrated in Table 1, when the annotation ratio was 10%, the baseline method using ViT and CNN obtained an average DSC of 71.08% and 75.28%, respectively. The best existing method was CCT [7] and it obtained an average DSC of 81.81%. In contrast, our method significantly outperformed it by achieving an average DSC of 85.02%. When the annotation ratio was 20%, CPS [4] achieved the best performance among existing methods, and our method was also significantly better than it in terms of DSC (88.20% vs 84.62%) and HD_{95} (2.54 mm vs 4.89 mm). Fig. 2 shows a visual comparison between our method and CPS and CCT that are top existing methods according to Table 1, and the annotation ratio was 10%. It can be observed that our method exhibits better segmentation performance than CPS [4] and CCT [7].

3.4. Ablation study in architectural design

To validate our network structure, we compared different variants of encoder and decoder design with an annotation ratio of 10%. As shown in Table 2, Two CNNs, Two ViTs and CNN&ViT mean cross teaching without inter-network feature fusion between two CNNs, two ViTs and between a CNN and a ViT, respectively. Their average DSC values were 80.92%, 78.82% and 83.22%, respectively, showing the effectiveness of asymmetric networks for cross teaching. We then added a feature fusion module at the bottleneck for CNN&ViT, which is refereed to as 'ours (one fusion)', and it improves the average DSC to 84.06%. Furthermore, our method with multi-scale feature fusion obtained the highest DSC of 85.02%. Table 2 also shows that inference with the ViT decoder has a lower performance than the using CNN decoder for inference, which is mainly because that ViT has a lower ability to obtain detailed segmentation results.

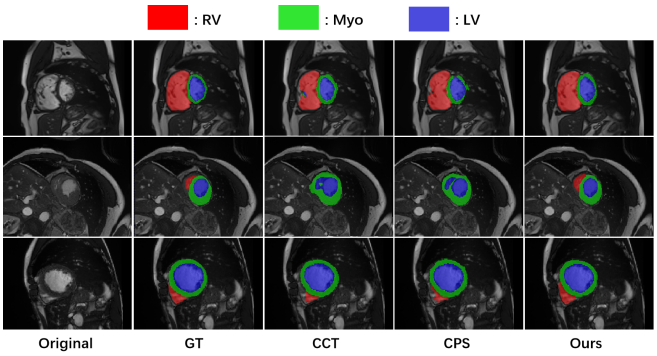

Fig. 2: Visual comparison between different methods at an annotation ratio of 10%.

4. DISCUSSION AND CONCLUSION

In this work, we propose a semi-supervised segmentation model that unifies CNN and ViT via multi-scale fusion of their complementary features in the encoder and cross teaching between CNN and ViT decoders. It can effectively aggregate fine-grained features of CNN with contextual representations of ViT to enhance the feature learning ability for dealing with a partially labeled training set. The cross teaching between the two types of decoders further leverages their complementary information to avoid bias towards a single model, reducing the effect of noisy pseudo labels. Extensive experiments demonstrated that our method achieved state-of-the-art segmentation performance under different annotation ratios, and illustrated the benefits of aggregating complementary CNN and Transformer features and decoders for semi-supervised medical image segmentation. In the future, it is of interest to develop better strategies for dealing with noisy pseudo labels in the cross teaching framework.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research used the ACDC public dataset for experiments. Ethical approval was obtained by the dataset creator from their Institutional Review Board.

6. ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (62271115).

7. REFERENCES

[1] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE TMI*, vol. 39, no. 6, pp. 1856–1867, 2020.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICRL*, 2021.

[3] Guotai Wang, Xiangde Luo, Ran Gu, Shuojue Yang, Yijie Qu, Shuwei Zhai, Qianfei Zhao, Kang Li, and Shaoting Zhang, "PyMIC: A deep learning toolkit for annotation-efficient medical image segmentation," *CMPB*, vol. 231, pp. 107398, 2023.

[4] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *CVPR*, 2021, pp. 2613–2622.

[5] Antti Tarvainen and Harri Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *NeurIPS*, vol. 30, 2017.

[6] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation," in *MICCAI*, 2019, pp. 605–613.

[7] Yassine Ouali, Celine Hudelot, and Myriam Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *CVPR*, 2020, pp. 12674–12684.

[8] Xiangde Luo, Minhao Hu, Tao Song, Guotai Wang, and Shaoting Zhang, "Semi-supervised medical image segmentation via cross teaching between cnn and transformer," in *MIDL*. PMLR, 2022, pp. 820–833.

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241.

[10] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *ECCV 2022 Workshops*, 2023, pp. 205–218.

[11] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu, "Dual attention network for scene segmentation," in *CVPR*, 2018, pp. 3146–3154.

[12] Lanfeng Zhong, Xin Liao, Shaoting Zhang, and Guotai Wang, "Semi-supervised pathological image segmentation via cross distillation of multiple attentions," in *MICCAI*, 2023, pp. 570–579.

[13] Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L Yuille, "Deep co-training for semi-supervised image recognition," in *ECCV*, 2018, pp. 135–152.

[14] Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al., "R-drop: Regularized dropout for neural networks," *NeurIPS*, vol. 34, pp. 10890–10905, 2021.

[15] Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz, "Interpolation consistency training for semi-supervised learning," in *ICCV*, 2019, pp. 5358–5367.

[16] Xiangde Luo, Guotai Wang, Wenjun Liao, Jieneng Chen, Tao Song, Yinan Chen, Dimitris N. Metaxas Zhang, Shichuan, and Shaoting Zhang, "Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency," *MedIA*, vol. 80, pp. 102517, 2022.

[17] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al., "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved?," *IEEE TMI*, vol. 37, no. 11, pp. 2514–2525, 2018.