The following publication Hang, W., Feng, W., Liang, S., Yu, L., Wang, Q., Choi, K. S., & Qin, J. (2020). Local and global structure-aware entropy regularized mean teacher model for 3d left atrium segmentation. In Proceedings of Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, pp. 562-571. Springer International Publishing is available at https://doi.org/10.1007/978-3-030-59710-8_55.

Local and Global Structure-aware Entropy Regularized Mean Teacher Model for 3D Left Atrium segmentation

PaperID: 1867

Abstract. Emerging self-ensembling methods have achieved promising semi-supervised segmentation performances on medical images through forcing consistent predictions of unannotated data under different perturbations. However, the consistency only penalizes on independent pixel-level predictions, making structure-level information of predictions not exploited in the learning procedure. In view of this, we propose a novel structure-aware entropy regularized mean teacher model to address the above limitation. Specifically, we firstly introduce the entropy minimization principle to the student network, thereby adjusting itself to produce high-confident predictions of unannotated images. Based on this, we design a local structural consistency loss to encourage the consistency of inter-voxel similarities within the same local region of predictions from teacher and student networks. To further capture local structural dependencies, we enforce the global structural consistency by matching the weighted self-information maps between two networks. In this way, our model can minimize the prediction uncertainty of unannotated images, and more importantly that it can capture local and global structural information and their complementarity. We evaluate the proposed method on a publicly available 3D left atrium MR image dataset. Experimental results demonstrate that our method achieves outstanding segmentation performance than the state-of-the-art approaches in scenes with limited annotated images.

1 Introduction

Accurate segmentation of left atrium (LA) from 3D magnetic resonance (MR) images is essential for obtaining its morphological information, which provides support for the diagnosis and treatment of various cardiovascular diseases [5]. Deep learning has achieved promising performance on LA segmentation [1], but its performance depends on the availability of abundant annotated images. For 3D medical images, it is difficult to obtain abundant annotated data because manually annotating data slice by slice is costly and time consuming.

Deep semi-supervised learning methods have been proposed by utilizing limited annotated data together with abundant unannotated data to achieve better generalization for medical image segmentation [11,6,12,2,16]. Inspired by the success of self-ensembling method, Li *et al.* [6] embedded the transformation consistency into Π-model (TCSE) to boost the generalization capability of network. The mean teacher (MT) model [12] was extended to an adapted MT

model with soft Dice consistency loss (MT-Dice) [2] for brain lesion segmentation. Subsequently, Yu *et al.* [16] proposed an uncertainty-aware MT model (UA-MT) to generate more reliable predictions by encouraging low uncertainty of the teacher network. However, these methods focus only on the pixel-level consistency of prediction, ignoring the underlying structural information [13]. More recently, Liu *et al.* [7,8] proposed to learn structural relationship between pixels by measuring their spatial distance. Based on this, Kim *et al.* [4] exploited the inter-pixel relationship of prediction to optimize MT model for semantic segmentation. Other approaches [15] leveraged the entropy map of prediction to obtain object border information for pixel-wise image segmentation. Despite the success in some applications, current self-ensembling methods fail to fully exploit the complementarity of rich spatial and geometric structural information in prediction.

Herein, we propose a novel deep semi-supervised learning method named *local and global structure-aware entropy regularized mean teacher* (LG-ER-MT) for 3D LA segmentation. In general, our method inherits the robustness of MT model in the same way that encourages the segmentation consistency under different perturbations. Concretely, since the entropy minimization can adjust network to generate high-confident predictions [9], we design an *entropy regularized mean teacher* (ER-MT) model to penalize voxel-level prediction uncertainty of unannotated data. To further exploit structure-level information of 3D MR images, we calculate inter-voxel similarities within small volumes stochastically sampled from prediction to obtain local spatial structural information, while utilizing weighted self-information (*i.e.*, the disentanglement of Shannon Entropy [14]) to acquire global geometric structure information. We simultaneously introduce local and global structural information to MT model to capture their complementarity. By encouraging structural consistencies between teacher and student networks, we further improves the generalization capacity of network. Experiments on MICCAI 2018 Atrial Segmentation Challenge dataset demonstrate that our method can achieve state-of-the-art performances.

2 Methodology

In this section, we first introduce the proposed entropy regularized mean teacher (ER-MT) model. Then, we elaborate on both local and global structural consistencies. Fig. 1 illustrates the pipeline of our local and global structure-aware entropy regularized mean teacher (LG-ER-MT) model for 3D LA segmentation. The student network leverages the entropy minimization principle to adjust itself to make more precise segmentation of unannotated data. The local and global structural consistencies jointly enhance the generalization capability of the framework.

2.1 Entropy Regularized Mean Teacher (ER-MT)

Supposing that the training set $\mathcal{D}$ consists of L annotated data and U unannotated data, $\mathcal{D}_L = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{L}$ and $\mathcal{D}_U = \{\mathbf{x}_i\}_{i=L+1}^{L+U}$ represent annotated data

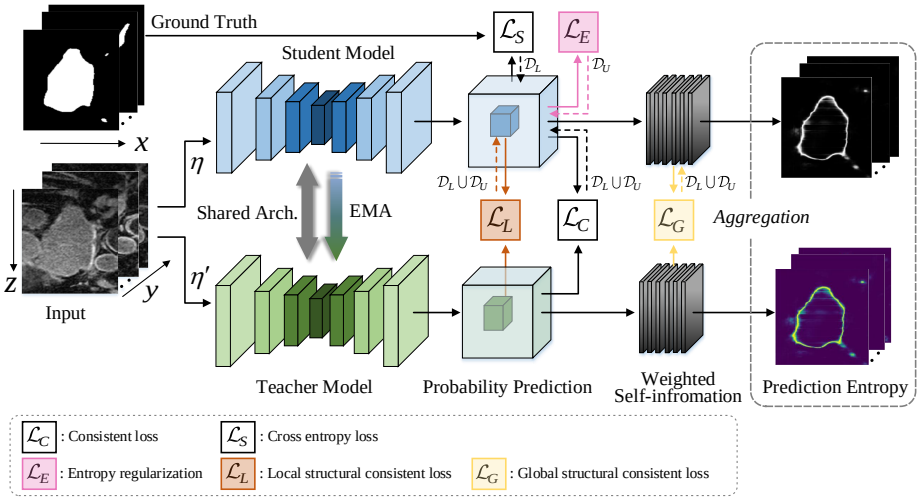

Fig. 1. Overview of our LG-ER-MT framework for semi-supervised segmentation. It jointly encourages local and global structural consistancy on both annotated data $\mathcal{D}_L$ and unannotated data $\mathcal{D}_U$, as well as generates high-confident prediction of $\mathcal{D}_U$. The prediction entropy maps equal to the voxel-wise aggregation of weighted self-information.

set and unannotated data set, where $\mathbf{x}_i$ is a $H \times W \times D$ dimensional input volume and $\mathbf{y}_i \in \{0,1\}^{H \times W \times D}$ is the corresponding ground truth. Let $\mathbf{p}^s_{i,v}$ and $\mathbf{p}^t_{i,v}$ represent prediction vectors of the v-th voxel of the i-th input volume from teacher and student networks, respectively. We denote $p^s_{i,v,c}$ as the predicted probability score of the class c, $c \in \{0,1\}$.

For image segmentation tasks, low-density regions are generally distributed in the pixels of an object border [3]. Considering that the entropy minimization principle can adjust the network passing through the low-density regions to make high-confident predictions [9], we introduce this principle to the student network for the first time in order to achieve more accurate segmentation of unannotated images. The entropy loss $\mathcal{L}_E$ is formulated as follows:

$$\mathcal{L}_E = -\frac{1}{U \cdot \log |c|} \sum_{i \in \mathcal{R}} \sum_{v \in \mathcal{V}} \sum_{c \in \{0,1\}} p^s_{i,v,c} \log p^s_{i,v,c} \tag{1}$$

where $\mathcal{R} = \{1, 2, \ldots, U\}$ and $\mathcal{V} = \{1, 2, \ldots, H \times W \times D\}$.

By incorporating the entropy minimization principle with MT model, we design an entropy regularized mean teacher (ER-MT) model. The optimization problem of ER-MT can be represented as:

$$\mathcal{L}_{ER-MT} = \mathcal{L}_S + \lambda_c \cdot \mathcal{L}_C + \lambda_e \cdot \mathcal{L}_E \tag{2}$$

where $\mathcal{L}_S$ is the cross-entropy loss used to evaluate the segmentation performance of annotated images. $\mathcal{L}_C$ denotes the conventional consistency loss that

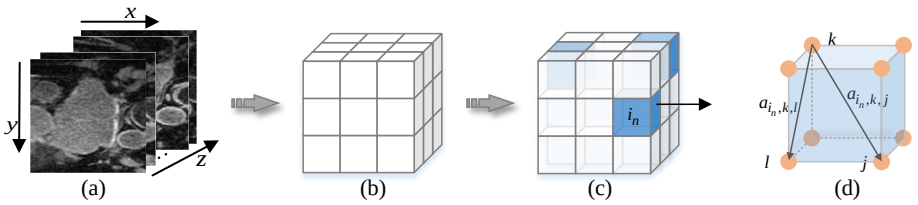

Fig. 2. Schematic of stochastic sampling. (a) 3D MR image. (b) dividing network prediction into sub-volumes. (c) stochastically sampled volumes (indicated in blue). (d) inter-voxel similarity, *e.g.*, $a_{i_n,k,l}$.

is defined as the expected distance between the predictions of teacher and student networks. We adopt mean squared error (MSE) to calculate $\mathcal{L}_C$ on both annotated and unannotated data. λ_c and λ_e are the regularization parameters for balancing the trade-off with other terms. $\mathcal{L}_E$ enables the student network more precisely segment unannotated images, which ultimately improves the generalization capability of MT model.

2.2 Local and Global Structure-aware Entropy Regularized Mean Teacher Model (LG-ER-MT)

In the field of medical image, MT model and its variants treat 3D segmentation as a classification task performed voxel by voxel, thus they employ the original consistency loss generally used in semi-supervised classification. However, image segmentation differs from general classification problems in that network prediction has structure-level characteristics. If ignored, it is difficult for MT models to achieve higher performance. To this end, we exploit local and global structural consistencies to jointly learn spatial and geometric structural information to promote the generalization capability of MT model for 3D LA segmentation.

Local Structural Consistency Loss. To obtain spatial structural information, one possible way is to calculate the similarity (distance) of each pairwise voxels in prediction. However, it is unrealistic to apply it in 3D MR image segmentation due to its high computational cost $\mathcal{O}\left((H \times W \times D)^2\right)$. Accordingly, we propose to stochastically sample few local volumes from prediction in each mini-batch and use them to compute the inter-voxel similarity. Fig. 2 illustrates the schematic of stochastic sampling and similarity computing.

After sampling N local volumes, we calculate the local structural consistency loss, which can be formulated as:

$$\mathcal{L}_L = \frac{1}{L+U} \sum_{i \in \mathcal{R}} \sum_{n \in \mathcal{N}} \frac{1}{(H_n \times W_n \times D_n)^2} \sum_{k,l \in \mathcal{V}} \left\| a^s_{i_n,k,l} - a^t_{i_n,k,l} \right\|^2 \quad (3)$$
$$\text{and} \quad a_{i_n,k,l} = (\mathbf{p}_{i_n,k})^T \mathbf{p}_{i_n,l} \big/ \left(\|\mathbf{p}_{i_n,k}\|_2 \|\mathbf{p}_{i_n,l}\|_2 \right).$$

Here, $\mathcal{R} = \{1, 2, \ldots, L+U\}$, $\mathcal{N} = \{1, 2, \ldots, N\}$, $\mathcal{V} = \{1, 2, \ldots, H_n \times W_n \times D_n\}$. All the local volumes are with the same dimension $H_n \times W_n \times D_n$. For the i-th prediction from student network s and teacher network t , $\mathbf{p}_{i_n,k}$ represents the prediction vector of the k-th voxel located in the n-th local volume, and $a_{i_n,k,l}$ denotes the cosine similarity (distance) between the k-th voxel and l-th voxel in the same local volume. $\mathcal{L}_L$ is designed to encourage inter-voxel similarities of local volumes between teacher and student networks to be consistent. By this way, notably, the computational cost can be driven down to $\mathcal{O}\left(N \cdot (H_n \times W_n \times D_n)^2\right)$, where $(H_n, W_n, D_n) \ll (H, W, D)$ and N is often small (*e.g.*, 4).

Global Structural Consistency Loss. Recall that the local structural consistency merely focuses on the spatial structural relationship within small volumes of prediction. The dependencies of local structural information have been previously ignored but not trivial. Since the prediction entropy of an image can reflect the results similar to the object border detection, it is reasonable to encourage the prediction entropy between teacher and student networks to be consistent, which is conductive to match the global geometric structural information. Given that the prediction entropy equals to the voxel-wise linear aggregation of weighted self-information [14], we intuitively use the latter (in higher dimensional space) to exploit the global geometric structural information.

Formally, we denote $\mathbf{I}^s_{i,v}$ as the weighted self-information of the v-th voxel of the i-th input from the student network, which equals to $-\mathbf{p}^s_{i,v} \circ \log \mathbf{p}^s_{i,v}$. The notation $\circ$ is Hadamard product and log is the logarithmic operation on each element. We formulate the global structural consistency loss $\mathcal{L}_G$ as minimizing MSE of the weighted self-information between teacher and student networks.

$$\mathcal{L}_G = \frac{1}{L+U} \sum_{i \in \mathcal{R}} \frac{1}{H \times W \times D} \sum_{v \in \mathcal{V}} \left\| \mathbf{I}^s_{i,v} - \mathbf{I}^t_{i,v} \right\|^2 \tag{4}$$

where $\mathcal{R} = \{1, 2, \ldots, L+U\}$ and $\mathcal{V} = \{1, 2, \ldots, H_n \times W_n \times D_n\}$. $\mathcal{L}_G$ encourages the global geometric structure between the two networks to be consistent, which can further improve the generalization capability of our model.

LG-ER-MT Framework. Based on the above discussions, we integrate Eq. (2), Eq. (3) and Eq. (4) in a unified local and global structure-aware entropy regularized MT (LG-ER-MT) framework as:

$$\mathcal{L}_{LG-ER-MT} = \mathcal{L}_{ER-MT} + \lambda_l \cdot \mathcal{L}_L + \lambda_g \cdot \mathcal{L}_G \tag{5}$$

where λ_l and λ_g are the tradeoff parameters for local and global structural consistency loss, respectively. With the proposed LG-ER-MT framework, we can obtain more accurate segmentation results through suppressing voxel-level prediction uncertainty of unannotated data as well as capturing local and global structural information and their complementarity.

3 Experiments and Results

Dataset and Pre-processing. The proposed LG-ER-MT method was extensively evaluated on the Atrial Segmentation Challenge dataset[1]. It provides 100 3D gadolinium-enhanced MR imaging scans with an isotropic resolution of 0.625mm $\times$ 0.625mm $\times$ 0.625mm and their corresponding ground truth. In the experiment, 80 scans were used for training and the remaining 20 scans for testing. All the scans were cropped centering at the heart region and normalized as zero mean and unit variance.

Implementation Details. The framework was implemented using PyTorch and trained on two RTX 2080Ti GPUs. According to [16], we randomly cropped $112 \times 112 \times 80$ sub-volumes and used the standard data augmentation techniques. V-Net [10] was used as our network backbone for both teacher and student networks. After *L-Stage* 5 layer and *R-Stage* 1 layer in V-Net, we added two dropout layers with the dropout rate of 0.5. Following [12], the parameter value of exponential moving average (EMA) was set to 0.99. The network parameters were trained for a total of 6000 iterations and updated by the SGD optimizer. The initial learning rate was set to $1e$-2 and then divided it by 10 every 2500 iterations. We set the batch size to 4, containing 2 annotated scans and 2 unannotated scans. The time-dependent Gaussian warming up function $\lambda(t) = \exp\left(-5(1 - t/t_{\max})^2\right)$ was used to ramp up hyper-parameters λ_c, λ_e,λ_l and λ_g from 0 to 0.1, 0.01, 0.01 and 0.01, respectively. Here, t denotes the current step and $t_{\max}$ is the maximum training step. In particular, it is critical to determine the size of stochastic sampled volumes when computing the local structural consistency loss. In each mini-batch, we stochastically sample 4 local volumes with a size of $16 \times 16 \times 16$ to balance the efficiency of the local structural information and the computational complexity.

Evaluation of 3D LA Segmentation. All methods were evaluated on four metrics, i.e., Dice, Jaccard, average surface distance (ASD), and 95% Hausdorff Distance (95HD). The 20% of training scans (*i.e.*, 16) were used as annotated data and the remaining training scans (*i.e.*, 64) as unannotated data. Table 1 lists the segmentation results of comparison methods on 20 test scans. We compared our LG-ER-MT with the baseline method V-Net and the original MT model [12], as well as several state-of-the-art semi-supervised segmentation methods, including adversarial learning based semi-supervised method (ASDNet) [11], TCSE [6], MT-Dice [2], and UA-MT [16]. These methods all used V-Net as the network backbone.

Firstly, we performed a quantitative evaluation of our methods. The first two rows in Table 1 show the segmentation performances of supervised V-Net using 80 and 16 annotated data, where the former results can be considered as the upper-bound performance. Our LG-ER-MT outperforms ASDNet and TCSE,

[1] https://atriaseg2018.cardiacatlas.org/

Table 1. Comparison results of different segmentation methods.

Method	# scans used		Metrics			
	Annotated	Unannotated	Dice[%]	Jaccard[%]	ASD[voxel]	95HD[voxel]
V-Net	80	0	91.14	83.82	1.52	5.75
	16	0	86.03	76.06	3.51	14.26
ASDNet [11]	16	64	87.90	78.85	2.08	9.24
TCSE [6]	16	64	88.15	79.20	2.44	9.57
MT [12]	16	64	88.12	79.03	2.65	10.92
MT-Dice [2]	16	64	88.32	79.37	2.76	10.50
UA-MT [16]	16	64	88.88	80.21	2.26	7.32
LG-ER-MT	16	64	**89.62**	**81.31**	**2.06**	**7.16**

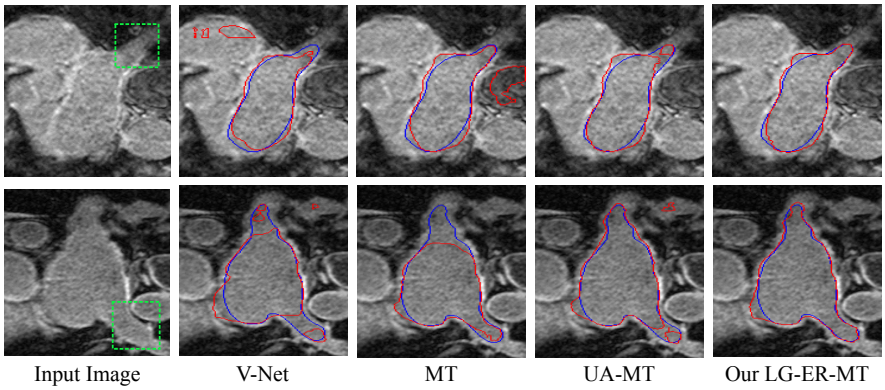

Fig. 3. Visualization of segmentation results. Green box labels PVs. Red and blue colors show the predictions and ground truths, respectively.

both of which have proven to be effective in semi-supervised segmentation. Besides, our method achieves better segmentation results than MT and MT-Dice in almost all cases. Compared to the state-of-the-art method UA-MT, LG-ER-MT improves by 0.74% Dice and 1.1% Jaccard, while reducing the metrics ASD and 95HD. It can be seen that our LG-ER-MT is approaching the results of the supervised V-Net using 80 annotated data.

We next qualitatively evaluated our method. Fig. 3 gives the segmentation examples of the supervised method V-Net (using 16 annotated data) and three semi-supervised methods MT, UA-MT, and our LG-ER-MT. Note that the pulmonary veins (PVs), as indicated by the green box, are difficult to precisely recognize due to the limited MR resolution and ambiguous borders. Our LG-ER-MT method produces segmentation results closer to the ground truth with more accurate borders and shapes, especially for PVs.

To verify the efficacy of entropy loss as well as local and global structural consistency losses, we performed the ablation study as follows: 1) ER-MT; 2) ER-MT with the local structural consistency loss (L-ER-MT), 3) ER-MT with

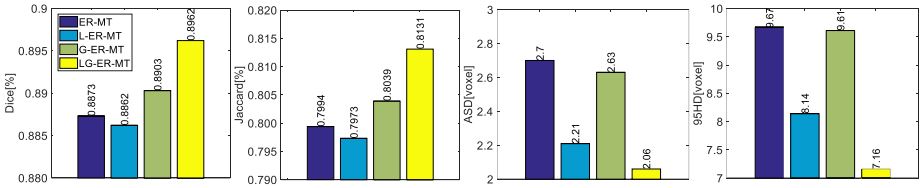

Fig. 4. Ablation study on different components.

Table 2. Comparison results with different amount of annotated data.

Method	# scans used		Metrics			
	Annotated	Unannotated	Dice[%]	Jaccard[%]	ASD[voxel]	95HD[voxel]
UA-MT	8	72	84.25	73.48	**3.36**	13.84
LG-ER-MT	8	72	**85.54**	**75.12**	3.77	**13.29**
UA-MT	16	64	88.88	80.21	2.26	7.32
LG-ER-MT	16	64	**89.62**	**81.31**	**2.06**	**7.16**
UA-MT	24	56	90.16	82.18	2.73	8.9
LG-ER-MT	24	56	**90.33**	**82.42**	**2.06**	**6.92**

the global structural consistency loss (G-ER-MT), and 4) LG-ER-MT. Fig. 4 demonstrates that three key components are able to bring performance improvement for semi-supervised 3D LA segmentation. Notably, the better performance of LG-ER-MT reveals the role of structure-level information complementarity.

Impact of the Amount of Annotated Data. To evaluate the effectiveness of LG-ER-MT with different amounts of annotated data, we compared it with UA-MT, as shown in Table 2. Concretely, we investigated two other cases, where 10% training scans (*i.e.*, 8) and 30% training scans (*i.e.*, 24) were used as annotated data, and the remaining training scans as unannotated data, respectively. Compared with UA-MT, LG-ER-MT improves by 1.29% Dice and 1.64% Jaccard using 8 annotated data. The metrics ASD and 95HD are decreased by 0.67 and 1.98, when using 24 annotated data. These promising results further validate the effectiveness of our method for semi-supervised 3D LA segmentation.

4 Conclusion

We propose a novel local and global structure-aware entropy regularized mean teacher model for LA segmentation from 3D MR images. Our method can minimize voxel-level prediction uncertainty of unannotated images, while encouraging structure-level information of predictions between student and teacher networks to be consistent. Extensive experiments verify the effectiveness of our method. The promising semi-supervised segmentation results of our method make it potentially useful for other medical image segmentation tasks.

References

1. Bian, C., Yang, X., Ma, J., Zheng, S., Liu, Y.A., Nezafat, R., Heng, P.A., Zheng, Y.: Pyramid network with online hard example mining for accurate left atrium segmentation. In: International Workshop on Statistical Atlases and Computational Models of the Heart. pp. 237–245. Springer (2018)
2. Cui, W., Liu, Y., Li, Y., Guo, M., Li, Y., Li, X., Wang, T., Zeng, X., Ye, C.: Semi-supervised brain lesion segmentation with an adapted mean teacher model. In: International Conference on Information Processing in Medical Imaging. pp. 554–565. Springer (2019)
3. French, G., Aila, T., Laine, S., Mackiewicz, M., Finlayson, G.: Consistency regularization and cutmix for semi-supervised semantic segmentation. arXiv preprint arXiv:1906.01916 (2019)
4. Kim, J., Jang, J., Park, H.: Structured consistency loss for semi-supervised semantic segmentation. arXiv preprint arXiv:2001.04647 (2020)
5. Lang, R.M., Badano, L.P., Mor-Avi, V., Afilalo, J., Armstrong, A., Ernande, L., Flachskampf, F.A., Foster, E., Goldstein, S.A., Kuznetsova, T., et al.: Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the american society of echocardiography and the european association of cardiovascular imaging. European Heart Journal-Cardiovascular Imaging **16**(3), 233–271 (2015)
6. Li, X., Yu, L., Chen, H., Fu, C.W., Heng, P.A.: Transformation consistent self-ensembling model for semi-supervised medical image segmentation. arXiv preprint arXiv:1903.00348 (2019)
7. Liu, Y., Chen, K., Liu, C., Qin, Z., Luo, Z., Wang, J.: Structured knowledge distillation for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2604–2613 (2019)
8. Liu, Y., Shu, C., Wang, J., Shen, C.: Structured knowledge distillation for dense prediction. arXiv preprint arXiv:1903.04197 (2019)
9. Long, M., Cao, Y., Cao, Z., Wang, J., Jordan, M.I.: Transferable representation learning with deep adaptation networks. IEEE transactions on pattern analysis and machine intelligence **41**(12), 3071–3085 (2018)
10. 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)
11. Nie, D., Gao, Y., Wang, L., Shen, D.: Asdnet: Attention based semi-supervised deep networks for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 370–378. Springer (2018)
12. 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. pp. 1195–1204 (2017)
13. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7472–7481 (2018)
14. Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P.: Dada: Depth-aware domain adaptation in semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 7364–7373 (2019)

15. Wang, S., Yu, L., Li, K., Yang, X., Fu, C.W., Heng, P.A.: Boundary and entropy-driven adversarial learning for fundus image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 102–110. Springer (2019)
16. 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: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 605–613. Springer (2019)