Multi-organ Segmentation via Co-training Weight-Averaged Models from Few-Organ Datasets

Rui Huang[1], Yuanjie Zheng[2(✉)], Zhiqiang Hu[1], Shaoting Zhang[1], and Hongsheng Li[3,4(✉)]

[1] SenseTime Research, Hong Kong, China
huangrui@sensetime.com
[2] School of Information Science and Engineering, Shandong Normal University, Jinan, China
zhengyuanjie@gmail.com
[3] CUHK-SenseTime Joint Laboratory, The Chinese University of Hong Kong, Hong Kong, China
hsli@ee.cuhk.edu.hk
[4] Centre for Perceptual and Interactive Intelligence (CPII), Hong Kong, China

Abstract. Multi-organ segmentation requires to segment multiple organs of interest from each image. However, it is generally quite difficult to collect full annotations of all the organs on the same images, as some medical centers might only annotate a portion of the organs due to their own clinical practice. In most scenarios, one might obtain annotations of a single or a few organs from one training set, and obtain annotations of the other organs from another set of training images. Existing approaches mostly train and deploy a single model for each subset of organs, which are memory intensive and also time inefficient. In this paper, we propose to co-train weight-averaged models for learning a unified multi-organ segmentation network from few-organ datasets. Specifically, we collaboratively train two networks and let the coupled networks teach each other on un-annotated organs. To alleviate the noisy teaching supervisions between the networks, the weighted-averaged models are adopted to produce more reliable soft labels. In addition, a novel region mask is utilized to selectively apply the consistent constraint on the un-annotated organ regions that require collaborative teaching, which further boosts the performance. Extensive experiments on three publicly available single-organ datasets LiTS [1], KiTS [8], Pancreas [12] and manually-constructed single-organ datasets from MOBA [7] show that our method can better utilize the few-organ datasets and achieves superior performance with less inference computational cost.

Keywords: Multi-organ segmentation · Co-training · Few-organ datasets

A. L. Martel et al. (Eds.): MICCAI 2020, LNCS 12264, pp. 146–155, 2020.
https://doi.org/10.1007/978-3-030-59719-1_15

1 Introduction

In medical image segmentation, obtaining multi-organ annotations on the same set of images is labor-intensive and time-consuming, where only experienced radiologists are qualified for the annotation job. On the other hand, different medical centers or research institutes might have annotated a subset of organs for their own clinical and research purposes. For instance, there are publicly available single-organ datasets, such as LiTS [1], KiTS [8] and Pancreas [12], each of which only provides a single organ's annotations, as shown in Fig. 1. However, existing methods cannot effectively train a multi-organ segmentation network based on those single-organ datasets with different images.

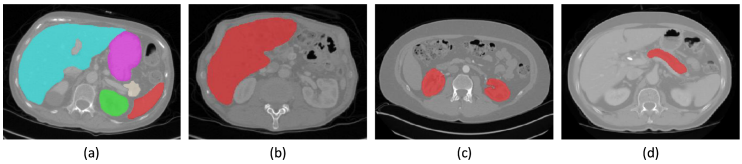

Fig. 1. (a) Multi-organ segmentation is required to achieve more comprehensive computer-aided analysis. (b) LiTS dataset [1] contains only liver annotations. (c) KiTS [8] dataset contains only kidney annotations. (d) Pancreas [12] dataset contains only pancreas annotations.

This work focuses on learning multi-organ segmentation from few-organ datasets for abdominal computed tomography (CT) scans. An intuitive solution is to segment each organ by a separate model using a training dataset. However, this solution is computationally expensive and the spatial relationships between different organs cannot be well exploited. Furthermore, some researches adopt self-training [10], which generates pseudo labels for un-annotated organs in each dataset using a trained single-organ segmentation model, and constructs a pseudo multi-organ dataset. The multi-organ segmentation model can be learned from the pseudo multi-organ dataset. Obviously, the pseudo labels might contain much noise due to the generalization inability of each single-organ segmentation model, as well as the domain gap between different datasets. The inaccurate pseudo labels would harm the training process and limit the performance of self-training.

To tackle the challenge, we propose to co-train a pair of weight-averaged models for unified multi-organ segmentation from few-organ datasets. Specifically, to provide supervisions for un-annotated organs, we adopt the temporally weight-averaged model to generate soft pseudo labels on un-annotated organs. In order to constrain error amplification, two models' weight-averaged versions are used to provide supervisions for training each other on the un-annotated organs via consistency constraints, in a collaborative manner. A novel region mask is proposed to adaptively constrain the network to mostly utilize the soft pseudo labels on the regions of un-annotated organs. Note that our proposed framework

with two networks is only adopted during training stage and only one network is used for inference without additional computational or memory overhead.

The contributions of our works are threefold: (1) We propose to co-train collaborative weight-averaged models for achieving unified multi-organ segmentation from few-organ datasets. (2) The co-training strategy, weight-averaged model and the region mask are developed for more reliable consistency training. (3) The experiment results show that our framework better utilizes the few-organ datasets and achieves superior performance with less computational cost.

2 Related Work

Recently, CNNs have made tremendous progress for semantic segmentation. Plenty of predominant approaches have been proposed, such as DeepLab [4], PSPNet [16] for natural images and UNet [11], VoxResNet [2] for medical images. Due to the difficulty of obtaining multi-organ datasets, many approaches are dedicated to the segmentation of one particular organ. Chen et al. [3] proposed a two-stage framework for accurate pancreas segmentation. As these approaches are designed under fully-supervised setting, they cannot be directly applied to train a multi-organ segmentation model from few-organ datasets.

Konstantin et al. [5] firstly present a conditional CNN framework for multi-class segmentation and demonstrate the possibility of producing multi-class segmentations using a single model trained on single-class datasets. However, the inference time of their method is proportion to the number of organs, which is inefficient. Zhou et al. [17] incorporated domain-specific knowledge for multi-organ segmentation using partially annotated datasets. But their training objective is difficult to optimized and it needs some specific optimization methods.

Teacher-student model is a widely used technique in semi-supervised learning (SSL) and model distillation. The key idea is to transfer knowledge from a teacher to a student network via consistency training. Deep mutual learning [15] proposed to train two networks collaboratively with the supervision from each other. Mean-teacher model [13] averaged model weights over different training iterations to produce supervisions for unlabeled data. Ge et al. [6] proposed a framework called Mutual Mean Teaching for pseudo label refinery in person re-ID. Note that these methods are designed under fully-supervised or semi-supervised settings. In this work, we exploit the integration of co-training strategy and weight-averaged models for unifying multi-organ segmentation from few-organ datasets.

3 Method

Single-organ datasets are special cases of few-organ datasets. Without loss of generality, we discuss how to train multi-organ segmentation networks from single-organ datasets in this section. The method can be easily extended to handle few-organ datasets. Formally, given K single-organ datasets $\{\mathbb{D}_1, \cdots, \mathbb{D}_K\}$, where $\mathbb{D}_k = \{(x_i^k, y_i^k) | i = 1, \cdots, N_k, k = 1, \cdots, K\}$, let x_i^k and y_i^k denote the i-th training sample in the k-th single-organ dataset and its associated binary

segmentation mask for organ k out of all K organs. Our goal is to train a unified network that can output segmentation maps for all K organs simultaneously.

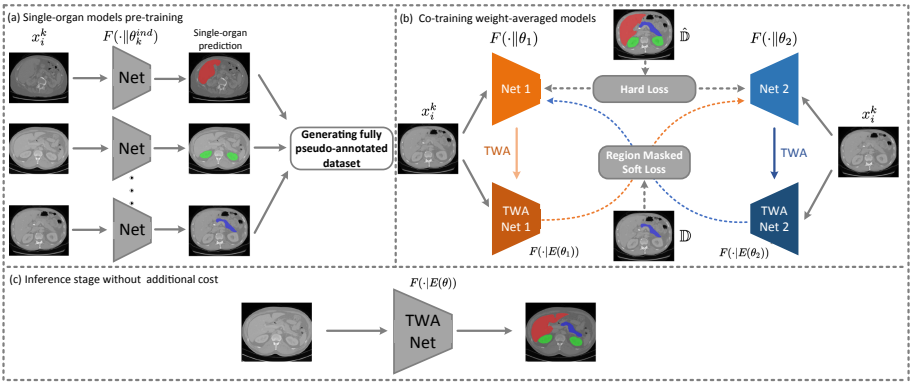

Fig. 2. (a) The pipeline of generating fully pseudo-annotated dataset. (b) The overall framework of our method. (c) In inference phase, only one network is used without requiring additional computational cost.

Pre-training Single-Organ and Multi-organ Models. We choose DeepLab [4] with dilated Resnet-50 [14] and IBN modules [9] as our segmentation backbone for its strong capability on different semantic segmentation tasks. Note that other segmentation networks could also be adopted in our proposed framework. We first pre-train K segmentation models on the K single-organ datasets, respectively. Each model is responsible for segmentation of one individual organ, denoted as $F(\cdot\|\theta_k^{ind})$, where θ_k^{ind} denotes network parameters of the k-th single-organ network. With the pre-trained models, for each single-organ dataset $\mathbb{D}_k$, we can generate pseudo labels for those un-annotated organs to create a fully-annotated dataset with pseudo labels $\hat{\mathbb{D}}_k$, where the label of organ k is manually annotated and the others are hard pseudo labels(see Fig. 2(a)). We can then construct a joint fully pseudo-annotated dataset $\hat{\mathbb{D}} = \{\hat{\mathbb{D}}_1, \cdots, \hat{\mathbb{D}}_K\}$.

Based on pseudo segmentation masks, we can pre-train a multi-organ segmentation model $\hat{F}$ on $\hat{\mathbb{D}}$. Obviously, the quality of pseudo labels is vital to the final performance. It is inevitable that some pseudo label maps might be inaccurate due to the generalization inability of each single-organ segmentation model. The noisy pseudo labels would therefore harm the final segmentation accuracy.

Co-training Weight-Averaged Models for Pseudo Label Regularization. Our framework is illustrated in Fig. 2(b). We train a pair of collaborative networks, $F(\cdot\|\theta_1)$ and $F(\cdot\|\theta_2)$, with the same structure as our pre-trained multi-organ network $\hat{F}$ but with randomly initialized parameters. The training utilizes both the fully pseudo-annotated dataset $\hat{\mathbb{D}}$ with hard pseudo labels and the original dataset $\mathbb{D} = \{\mathbb{D}_1, \cdots, \mathbb{D}_K\}$ with online generated soft pseudo labels. Both

the networks are trained with the weighted focal loss and dice loss using hard labels in the fully pseudo-annotated dataset $\hat{\mathbb{D}}$, for handling the high variability of organ size in abdomen. The weighted focal loss is defined as

$$\mathcal{L}_{\text{focal}}(\theta) = -\sum_{k=1}^{K}\sum_{i=1}^{N_k}\sum_{c=1}^{C}\alpha_c\left(1-F(x_i^k|\theta)_c\right)^{\gamma}\log\left(F(x_i^k|\theta)_c\right), \tag{1}$$

where c denotes the c-th organ class and $F(x_i^k|\theta)_c$ is model's estimated probability that a pixel is correctly classified. α_c is the weight of each organ c, which is inversely proportional to each organ's average size. The parameter γ is set as 2 empirically. The dice loss can be formulated as

$$\mathcal{L}_{\text{dice}}(\theta) = \sum_{k=1}^{K}\sum_{i=1}^{N_k}\sum_{c=1}^{C}\left(1-2\frac{\sum \hat{y}_i^{k,c}F(x_i^k|\theta)_c+\epsilon}{\sum \hat{y}_i^{k,c}+\sum F(x_i^k|\theta)_c+\epsilon}\right), \tag{2}$$

where $\hat{y}_i^k$ and $F(x_i^k|\theta)_c$ represent the hard labels and model's predictions for organ c, respectively. ϵ is a small value to ensure numerical stability. Note that the above losses are applied to both networks' parameters, θ_1 and θ_2.

Since the hard pseudo labels are quite noisy, to properly regularize the learning process, we also adopt the online generated soft pseudo labels for un-annotated organs when training the networks on the original data $\mathbb{D}$. For training network 1, $F(\cdot\|\theta_1^{(t)})$, at iteration t with image $x_i^k \in \mathbb{D}_k$, the ground-truth labels for organ k are used while other organs' predicted soft labels are generated from the network 2, $F(\cdot\|\mathbb{E}_t(\theta_2))$, with temporally averaged parameters $\mathbb{E}_t(\theta_2)$:

$$\mathbb{E}_t(\theta_2) = \alpha\mathbb{E}(\theta_2^{(t-1)}) + (1-\alpha)\theta_2^{(t)}, \tag{3}$$

where $\alpha \in [0,1]$ controls how fast the parameters are temporally averaged. Similarly, network 2's parameters $\theta_2^{(t)}$ are trained by temporally weight-averaged model of network 1 $F(\cdot\|\mathbb{E}_t(\theta_1))$'s predictions. Intuitively, the temporally weight-averaged version is a temporal ensemble of a network over its past iterations, which can generate more robust online soft pseudo labels for the un-annotated organs than the network at a specific iteration. In addition, we adopt one network's temporal average to supervise the other network. This strategy can avoid each network using its own previous iterations' predictions as supervisions, which might amplify its segmentation errors from previous iterations.

For each image x_i^k, pixels belong to organ-k are with ground-truth annotations. We would avoid adopting the soft pseudo labels for training networks on regions of ground-truth organs. We morphologically dilate the hard pseudo labels for each un-annotated organ to generate a region mask $\mathcal{T}(y_i^k)$:

$$\mathcal{T}(y_i^k) = \begin{cases} 1, & \text{regions without annotations or background,} \\ 0, & \text{regions with organ-k ground-truth annotations.} \end{cases}$$

Therefore, the segmentation loss with soft pseudo labels are formulated as:

$$\mathcal{L}_{\text{soft}}(\theta_1^{(t)}|\theta_2^{(t)}) = -\sum_{k=1}^{K}\sum_{i=1}^{N_k}(\mathcal{T}(y_i^k)\cdot F(x_i^k|\mathbb{E}_t(\theta_2)\cdot\log F(x_i^k|\theta_1^{(t)})), \tag{4}$$

$$\mathcal{L}_{\text{soft}}(\theta_2^{(t)}|\theta_1^{(t)}) = -\sum_{k=1}^{K}\sum_{i=1}^{N_k}(\mathcal{T}(y_i^k)\cdot F(x_i^k|\mathbb{E}_t(\theta_1)\cdot\log F(x_i^k|\theta_2^{(t)})). \tag{5}$$

The key difference between our method and mean teacher [13] is that we use the temporally weight-averaged version of one network to supervise another network. Such a collaborative training manner can further decouple the networks' predictions. In addition, the region masks are important to enforce the soft label supervisions are only applied to un-annotated regions.

Overall Segmentation Loss. Our framework is trained with the supervision of the hard loss and the soft loss. The overall loss function optimizes the two networks simultaneously, which is formulated as:

$$\begin{aligned}\mathcal{L}(\theta_1,\theta_2) =& \lambda_{\text{focal}}(\mathcal{L}_{\text{focal}}(\theta_1)+\mathcal{L}_{\text{focal}}(\theta_2)) + \lambda_{\text{dice}}(\mathcal{L}_{\text{dice}}(\theta_1)+\mathcal{L}_{\text{dice}}(\theta_2)) \\ &+ \lambda_{\text{rampup}}\lambda_{\text{soft}}(\mathcal{L}_{\text{soft}}(\theta_1|\theta_2)+\mathcal{L}_{\text{soft}}(\theta_2|\theta_1)),\end{aligned} \tag{6}$$

where $\lambda_{\text{focal}}, \lambda_{\text{dice}}$ and λ_{soft} are loss weights. Since the predictions at early training stages might not be accurate, we apply a ramp-up strategy to gradually increase λ_{rampup}, which makes the training process more stable.

4 Experiments

The proposed framework was evaluated on three publicly available single-organ datasets, LiTS [1], KiTS [8], Pancreas [12] and a manually-constructed single-organ dataset MOBA [7]. LiTS consists of 131 training and 70 test CT scans with liver annotations, provided by several clinical sites. KiTS consists of 210 training and 90 test CT scans with kidney annotations, collected from 300 patients who underwent partial or radical nephrectomy. Pancreas consists of 281 training and 139 test CT scans with pancreas annotations, provided by Memorial Sloan Kettering Cancer Center. Since the annotation is only available for the training set, we use their training sets in our experiments. MOBA is a multi-organ dataset with 90 CT scans drawn from two clinical sites. The authors of [7] provided segmentation masks of eight organs, including spleen, left kidney, gallbladder, esophagus, liver, stomach, pancreas and duodenum. Specifically, the multi-organ segmentation masks are binarized and stored separately, i.e., we have manually constructed eight single-organ datasets. All datasets are divided into training and test sets with a 4:1 ratio. We use the Dice-Score-Coefficient (DSC) and Hausdorff Distance (HD) as the evaluation metric: $\text{DSC}(\mathcal{P},\mathcal{G}) = \frac{2\times|\mathcal{P}\cap\mathcal{G}|}{|\mathcal{P}|+|\mathcal{G}|}$, where $\mathcal{P}$ is the binary prediction and $\mathcal{G}$ is the ground truth. HD measures the largest distance from points in $\mathcal{P}$ to its nearest neighbour in $\mathcal{G}$ and the distances of two directions are averaged to get the final metric: $\text{HD}(\mathcal{P},\mathcal{G}) = (d_H(\mathcal{P},\mathcal{G}) + d_H(\mathcal{G},\mathcal{P}))/2$.

For preprocessing, all the CT scans are re-sampled to $1 \times 1 \times 3$ mm. The CT intensity values are re-scaled to [0, 1] using a window of $[-125, 275]$ HU for better contrast. We then center crop a 352×352 patch as the network input.

4.1 Implementation Details

All models were trained for 10 epochs using synchronized SGD on 8 NVIDIA 1080 Ti GPUs with a minibatch of 24 (3 images per GPU). The initial learning rate is 0.05 and a *cosine* learning rate policy is employed. Weight decay of 0.0005 and momentum of 0.9 are used during training. The hyper-parameters $\lambda_{\text{focal}}, \lambda_{\text{dice}}$ and λ_{soft} are set to 1.0, 0.1 and 0.1, respectively. The smoothing coefficient α is set as 0.999. During inference, only one of the two weight-averaged models with better validation performance is used as the final model.

Table 1. Ablation studies of our proposed methods on the LiTS-KiTS-Pancreas dataset. CT: co-training strategy. WA: weight-averaged model. RM: region mask.

Method	Liver	Kidney	Pancreas	Avg DSC
Individual	95.90	**95.30**	77.05	89.41
Self-training	95.94	94.02	78.54	89.50
CT	95.89	94.42	78.15	89.49
WA	95.90	94.49	78.52	89.63
CT+WA	95.93	94.31	79.12	89.78
CT+WA+RM	**95.96**	95.01	79.25	90.07
Ours (CT+WA+RM+IBN)	95.90	94.98	**79.78**	**90.22**

Table 2. DSC(%) and execution time of conditionCNN [5] and our method.

Method	conditionCNN [5]	Ours
Liver	**95.93**	95.90
Kidney	**95.33**	94.98
Pancreas	77.90	**79.78**
Avg DSC	89.72	**90.22**
Time (s)	12.9	**4.28**

Table 3. DSC (%) and HD (mm) comparison on MOBA dataset.

Organ	DSC (%)					HD (mm)		
	Individual	Combine	Self-training	Ours	ConditionCNN [5]	Individual	Self-training	Ours
Spleen	**96.00**	95.28	95.69	94.95	80.77	**16.81**	32.54	23.15
Kidney(L)	94.51	94.35	94.60	**95.03**	81.51	29.20	26.29	**22.33**
Gallbladder	78.59	79.55	**80.43**	79.65	66.15	54.22	**31.58**	34.35
Esophagus	66.07	62.90	71.87	**72.25**	55.03	26.85	26.58	**24.47**
Liver	**96.61**	96.12	96.23	96.18	94.47	**31.99**	45.89	43.07
Stomach	**91.35**	87.65	90.08	89.68	82.56	57.64	43.71	**42.93**
Pancreas	78.04	73.69	78.10	**79.35**	60.55	28.61	31.04	**29.68**
Duodenum	58.16	54.53	57.72	**61.63**	47.60	41.28	38.43	**35.08**
Avg	82.41	82.02	83.09	**83.60**	62.37	35.82	34.51	**31.88**

4.2 Experiments on LiTS-KiTS-Pancreas Dataset

For the single-organ datasets, LiTS, KiTS and Pancreas, we first train three single-organ models separately for each organ, denoted as "individual" in Table 1. It achieves 95.90%, 95.30%, 77.05% DSC for liver, kidney, pancreas, respectively, and an average DSC of 89.41%.

Ablation Study. In this section, we evaluate each component's effect in our framework. The ablation results are shown in Table 1. We can observe that when training a multi-organ segmentation model directly with hard pseudo labels (denoted as "self-training"), the performance is slightly better than single-organ models (89.41% to 89.50%), which means that even the noisy pseudo labels can improve the segmentation of un-annotated organs. Meanwhile, by applying the co-training scheme, weight-averaged model, region mask and IBN module, our proposed framework achieves a remarkable improvement of 0.81% in terms of average DSC (89.41% to 90.22%) without any additional computational cost for inference. Especially, we observe a significant performance gain of 2.73% for the segmentation of the pancrea, which is more challenging because of its smaller sizes and irregular shapes. Note that when only applying the co-training scheme, the performance is just comparable with self-training, which demonstrates that using the weight-averaged model for supervising the other model can produce more reliable soft labels. The weight-averaged model, region mask and IBN module bring performance gains of 0.29%, 0.29%, and 0.15%, respectively.

Comparison with State-of-the-Art. We compare our method with state-of-the-art conditionCNN [5], which targets at the same task as our work. Since their full dataset is not publicly available and their code is not open-sourced, we re-implement their method using the above mentioned datasets and our baseline model, and tune the hyper-parameters to achieve their best performance for a fair comparison. The results are shown in Table 2. We can see that our method outperforms conditionCNN by a considerate margin (0.5%). In addition, the inference time of conditionCNN is proportion to the number of organs, which makes it inefficient when handling a large number of organs. Some qualitative results are shown in Fig. 3. Our method shows more superior results with less computational cost, compared with existing methods.

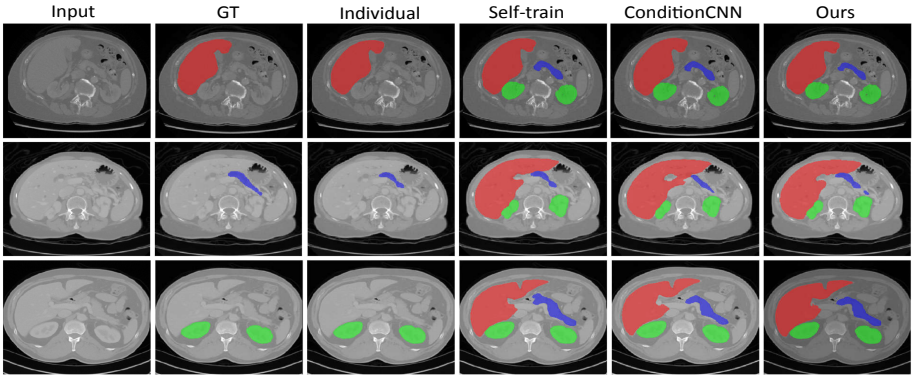

Fig. 3. Qualitative comparison of different methods. Top to bottom: a LiTS dataset image with liver annotation, a Pancreas dataset image with pancreas annotation, and a KiTS dataset image with kidney annotation.

4.3 Experiments on MOBA Dataset

To validate the generalization ability of our method, we also conduct experiments on MOBA dataset, which is more challenging with eight target organs. Since MOBA has multi-organ annotations, we can train a multi-organ segmentation model directly for comparison. The results are shown in Table 3. Our method obtains a significant performance gain of 1.19% compared with the baseline "individual" model (82.41% to 83.60%). Similarly, a large improvement can be observed for those organs with smaller size and irregular shape, such as esophagus (66.07% to 72.25%) and duodenum (58.16% to 61.63%), which demonstrate the effectiveness and robustness of our framework. Interestingly, our method even outperforms the fully-supervised results (denoted as "combine"). We speculate that it might result from the MOBA has more organs to segment and there is severe class imbalance among organs, and our framework can alleviate the imbalance problem by the proposed online-generated soft pseudo labels. We further calculate Hausdorff Distance in Table 3. Our method shows much better HD than baseline "individual" and self-training strategy, bring gains of 3.94 and 2.63 respectively.

Additionally, conditionCNN [5] fails to achieve high performance on MOBA. The accuracy drops dramatically, especially for those organs with smaller sizes and irregular shapes. We suspect it's because conditionCNN cannot handle too many organs with high variation by simply incorporating the conditional information into a CNN.

5 Conclusion

We propose to co-train weight-averaged models for achieving unified multi-organ segmentation from few-organ datasets. Two networks are collaboratively trained to supervise each other via consistency training. The weight-averaged models are utilized to produce more reliable soft labels for mitigating label noise. Additionally, a region mask is developed to selectively apply the consistent constraint on the regions requiring collaborative teaching. Experiments on four public datasets show that our framework can better utilize the few-organ data and achieves superior performance on multiple public datasets with less computational cost.

Acknowledgements. This work is supported in part by the General Research Fund through the Research Grants Council of Hong Kong under Grants CUHK14208417, CUHK14239816, CUHK14207319, in part by the Hong Kong Innovation and Technology Support Programme (No. ITS/312/18FX), in part by the National Natural Science Foundation of China (No. 81871508; No. 61773246), in part by the Taishan Scholar Program of Shandong Province of China (No. TSHW201502038), in part by the Major Program of Shandong Province Natural Science Foundation (ZR2019ZD04, No. ZR2018ZB0419).

References

1. Bilic, P., et al.: The liver tumor segmentation benchmark (LITS). arXiv preprint arXiv:1901.04056 (2019)

2. Chen, H., Dou, Q., Yu, L., Qin, J., Heng, P.A.: VoxresNet: deep voxelwise residual networks for brain segmentation from 3D MR images. NeuroImage **170**, 446–455 (2018)
3. Chen, H., Wang, X., Huang, Y., Wu, X., Yu, Y., Wang, L.: Harnessing 2D networks and 3D features for automated pancreas segmentation from volumetric CT images. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11769, pp. 339–347. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32226-7_38
4. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818 (2018)
5. Dmitriev, K., Kaufman, A.E.: Learning multi-class segmentations from single-class datasets. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9501–9511 (2019)
6. Ge, Y., Chen, D., Li, H.: Mutual mean-teaching: pseudo label refinery for unsupervised domain adaptation on person re-identification. arXiv preprint arXiv:2001.01526 (2020)
7. Gibson, E., et al.: Automatic multi-organ segmentation on abdominal CT with dense V-networks. IEEE Trans. Med. Imaging **37**(8), 1822–1834 (2018)
8. Heller, N., et al.: The kits19 challenge data: 300 kidney tumor cases with clinical context, CT semantic segmentations, and surgical outcomes. arXiv preprint arXiv:1904.00445 (2019)
9. Pan, X., Luo, P., Shi, J., Tang, X.: Two at once: enhancing learning and generalization capacities via IBN-net. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 464–479 (2018)
10. Papandreou, G., Chen, L.C., Murphy, K.P., Yuille, A.L.: Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1742–1750 (2015)
11. 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
12. Simpson, A.L., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063 (2019)
13. 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)
14. Zhang, H., et al.: Context encoding for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7151–7160 (2018)
15. Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H.: Deep mutual learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4320–4328 (2018)
16. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890 (2017)
17. Zhou, Y., et al.: Prior-aware neural network for partially-supervised multi-organ segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 10672–10681 (2019)