Partially Supervised Multi-organ Segmentation via Affinity-Aware Consistency Learning and Cross Site Feature Alignment

Qin Zhou, Peng Liu, and Guoyan Zheng(✉)

Institute of Medical Robotics, School of Biomedical Engineering, Shanghai Jiao Tong University, No. 800, Dongchuan Road, Shanghai 200240, China
guoyan.zheng@sjtu.edu.cn

Abstract. Partially Supervised Multi-Organ Segmentation (PSMOS) has attracted increasing attention. However, facing with challenges from lacking sufficiently labeled data and cross-site data discrepancy, PSMOS remains largely an unsolved problem. In this paper, to fully take advantage of the unlabeled data, we propose to incorporate voxel-to-organ affinity in embedding space into a consistency learning framework, ensuring consistency in both label space and latent feature space. Furthermore, to mitigate the cross-site data discrepancy, we propose to propagate the organ-specific feature centers and inter-organ affinity relationships across different sites, calibrating the multi-site feature distribution from a statistical perspective. Extensive experiments manifest that our method generates favorable results compared with other state-of-the-art methods, especially on hard organs with relatively smaller sizes.

Keywords: Multi-organ Segmentation · Partially Supervised · Affinity Relationship · Consistency Learning

1 Introduction

Automatic multi-organ segmentation (MOS) plays a vital role in computer-aided diagnosis and treatment planning. Recently, deep learning based methods have made remarkable progress in solving MOS tasks. However, they typically require a large amount of expert-level accurate, densely-annotated data for training, which is laborious and time consuming to collect. Therefore, existing fully labeled datasets (termed as FLDs) are very few and often low in sample size [1]. While there exist many publicly available partially labeled datasets (PLDs) [2,3], each with one or a few out of the many organs annotated. This has motivated the development of various Partially-Supervised Multi-Organ Segmentation (PSMOS) methods that aim to learn a unified model from a union of such datasets. For example, Dmitriev and Kaufman proposed the conditional

This study was partially supported by the National Natural Science Foundation of China via projects U20A20199 and 62201341.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14221, pp. 671–680, 2023.
https://doi.org/10.1007/978-3-031-43895-0_63

U-Net to enable PSMOS using a single unified network [4]. Co-training between two models with consistency constraints on soft pseudo labels [6], and multi-scale features learned in a pyramid-input and pyramid-output network [7] were both explored for PSMOS. Other researchers resorted to prior knowledge to guide the training process. In PaNN [8], the average organ size distributions on the PLDs were constrained to resemble the prior statistics obtained from the FLD. Another method was introduced in [9] where the non-overlapping characteristics between different organs were exploited to design the exclusion loss.

Although witnessed great progress in PSMOS, existing methods are faced with the following challenges: 1) Shortage in sufficiently labeled samples for supervised learning, since voxel-level labels are only available for a subset of organs in PLDs; 2) Significant cross-site appearance variations caused by different imaging protocols or subject cohorts. Different from existing methods, we propose a novel framework to explicitly tackle the above-mentioned challenges.

To handle the label-scarcity problem in PLDs, we propose a novel Affinity-aware Consistency Learning (ACL) scheme to incorporate voxel-to-organ affinity in the embedding space into consistency learning. Although consistency learning is frequently used for leveraging unlabeled data in label-efficient learning [10–12], it is mostly deployed in the label space [13–15], while little attention has been paid to exploring consistency in the latent feature space. Zheng et al. [16] proposed to adopt auxiliary student-teacher networks to utilize the features for consistency learning, which introduced more parameters, thus were computationally expensive. By incorporating voxel-to-organ affinity in the embedding space into consistency learning, our ACL scheme is plug-and-play and can capture rich context information in the embedding space.

To tackle the data discrepancy problem [17], based on the assumption that a well trained joint model should generate consistent feature distributions across different sites, we propose a novel Cross-Site Feature Alignment (CSFA) module, where two terms are introduced to attend to both the organ-specific and inter-organ statistics in the latent feature space. Concretely, for each PLD, we restrain the organ-specific prototypes calculated in each mini-batch to be close to the corresponding prototypes generated on the small-sized FLD. To further reduce the data discrepancy problem, we constrain the affinity relationships across different organ-specific prototypes to be consistent among different sites. By doing this, we transfer not only the single-class centroid, but also the inter-organ affinity learned from the small-sized FLD to PLDs, allowing for knowledge propagation at multiple granularity levels. Our contributions can be summarized as follows:

- We propose a novel affinity-aware consistency learning scheme to incorporate voxel-to-organ affinity in the embedding space into a consistency learning framework, which can capture semantic context in the latent feature space.
- We design a novel cross site feature alignment module to calibrate feature distributions of PLDs with distribution priors learned from a small-sized FLD, alleviating the cross-site data discrepancy.
- We demonstrate on five datasets collected from different sites that our method can effectively learn a unified MOS model from multi-source datasets, achieving superior performance over the state-of-the-art (SOTA) methods.

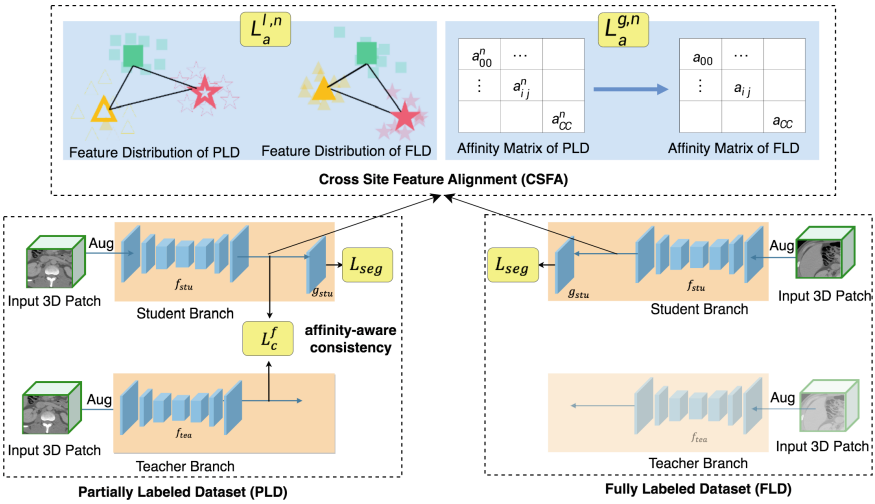

Fig. 1. A schematic illustration of our framework. "Aug" refers to perturbations with data augmentations. In the CSFA module, hollow shapes refer to the features belonging to unlabeled organs in the PLDs, while solid ones refer to labeled organs. The affinity matrix is calculated according to Eq. 10 and Eq. 11. L_{seg} is the segmentation loss.

2 Methodology

To learn a unified model from a small-sized FLD and a number of PLDs, we propose a novel framework to address the issues of label-scarcity and cross-site data discrepancy. The overall workflow of our method is presented in Fig. 1. During training, in each batch, we sample 3D patches from both the FLD and one of the PLDs, where the teacher-student scheme [14] is adopted to impose consistency constraints on the unlabeled voxels of the PLD. In our method, apart from the label space consistency, we introduce the ACL scheme to explore consistency in the embedding space. We further leverage the CSFA module to perform feature alignment between the FLD and the PLD. Please note that consistency constraints are only imposed on the unlabeled voxels of PLDs. The label space consistency loss is omitted in Fig. 1 for brevity.

2.1 Preliminaries

Denote Y_{full} as the full label set, i.e., $Y_{full} = \{0, 1, 2, \cdots, C\}$, where 0 refers to the background class, and $\{1, \cdots, C\}$ are one-to-one mappings to the target organs, C is the number of target organs. Given a small-sized FLD D_f and a number of PLDs $D_p = \{D_p^n, n \in \{1, \cdots, N\}\}$, where N is the number of PLDs. Each dataset can then be formally defined as either $D_f = \{I_{j,i}^f, y_{j,i}^f\}$ or $D_p^n = \{I_{j,i}^n, y_{j,i}^n\}$, where $I_{j,i}^f$ is the i-th pixel of the j-th image in the FLD D_f, and $y_{j,i}^f$ is its corresponding label. Similarly, $(I_{j,i}^n, y_{j,i}^n)$ is the i-th pixel-label

pair of the j-th image in the n-th PLD D_p^n. Please note that each D_p^n contains only a subset of the full label set, i.e., $Y_p^n = unique(\{y_{j,i}^n\}) \subsetneqq Y_{full}$, where $unique(\cdot)$ returns the unique values in the label set. The task of PSMOS aims to learn the mapping function $\varphi = f \circ g$ to project the 3D image patch $I_j \in \mathbb{R}^{h \times w \times z}$ to its corresponding semantic labels, where f is the feature extractor, g is the segmentation head, and $\circ$ means sequentially executing f and g, (h, w, z) are the 3D patch size. Since foreground organ in one PLD may be labeled as background in another dataset, such a background ambiguity brings challenges to joint training on multiple PLDs. To address this issue, we follow [7,9] to calculate the marginal cross entropy and marginal Dice loss as the baseline segmentation loss L_{seg}.

2.2 Prototype Generation

In our proposed framework, the calculation of both the pixel-to-prototype predictions (in ACL) and the feature alignment loss (in CSFA) are based on organ-specific prototypes. In each mini-batch, denote the organ-specific prototypes for the FLD as $\{\boldsymbol{q}_c\}, c \in \{0, \cdots, C\}$ and prototypes for the n-th PLD as $\{\boldsymbol{q}_c^n\}, c \in \{0, \cdots, C\}$, then they are generated as follows. On the FLD, we generate the prototypes in an exponential moving average scheme. Specifically, the feature prototype of the t-th iteration is calculated as (for brevity, we omit the iteration superscript t),

$$\boldsymbol{q}_c = \alpha \boldsymbol{q}_c + (1 - \alpha) \boldsymbol{q}_c^{update}, c \in \{0, \cdots, C\}, \tag{1}$$

where $\boldsymbol{q}_c^{update}$ is the average feature of the c-th class in current mini-batch of the FLD and α is the weighting coefficient. Given the feature maps $\boldsymbol{F} = \{\boldsymbol{f}_i\}$ and their related labels $\{y_i\}$, where $\boldsymbol{f}_i$ represents the i-th pixel in the feature maps of current mini-batch, the feature center of the c-th class is then calculated as,

$$\boldsymbol{q}_c^{update} = \frac{1}{Z_c} \sum_{i, y_i = c} \boldsymbol{f}_i, c \in \{0, \cdots, C\}, \tag{2}$$

where Z_c is the number of pixels belonging to the c-th class in current mini-batch.

On the n-th PLD, we directly adopt the feature centers calculated in each mini-batch as the organ-specific prototypes. In specific, for the labeled organs, the prototypes $\{\boldsymbol{q}_c^n\}, c \in Y_p^n$ are calculated according to Eq. 2, with feature maps generated on 3D patches from the n-th PLD. While on the unlabeled organs, only reliable features are used for calculating the pseudo feature centers as,

$$\boldsymbol{q}_c^n = \frac{1}{Z_c} \sum_{i, y_i = c} \mathbb{1}[p_i^n > \tau] \boldsymbol{f}_i, c \notin Y_p^n, \tag{3}$$

where p_i^n is the normalized prediction score generated from the teacher model, y_i denotes the corresponding pseudo label, τ is the confidence threshold, Z_c is the number of reliable predictions in class c, and $\mathbb{1}[\cdot]$ returns 1 if the inside condition is True, otherwise, returns 0.

2.3 Affinity-Aware Consistency Learning

In this paper, we propose to incorporate the voxel-to-organ affinity into consistency learning. Specifically, instance-to-prototype matching is calculated to capture the voxel-to-organ affinity. The affinities are then transformed into normalized scores for calculating the consistency constraint on two perturbed inputs. We adopt the teacher-student scheme [14] for consistency learning on the unlabeled data. Formally, denote I_t, I_s as the perturbed versions of the same sampled 3D patch for the teacher branch and the student branch respectively. In the teacher branch, denote $\boldsymbol{\phi}_i = f_{tea}(I_{t,i}) \in \mathbb{R}^d$ as the extracted feature for the i-th pixel of 3D image patch I_t. Given the prototypes generated on the FLD $\{\boldsymbol{q}_c\}, c \in \{0, \cdots, C\}$, then the pixel-to-prototype classification logit $\boldsymbol{p}_{t,i} = \{p^c_{t,i}\}$ is calculated as,

$$p^c_{t,i} =< \boldsymbol{\phi}_i, \boldsymbol{q}_c >,\ c \in \{0, \cdots, C\} \tag{4}$$

where $< \cdot, \cdot >$ calculates the cosine similarity between the two terms.

Similarly, in the student branch, denote $\boldsymbol{\psi}_i$ as the i-th feature $\boldsymbol{\psi}_i = f_{stu}(I_{s,i}) \in \mathbb{R}^d$, then prototype based predictions $\boldsymbol{p}_{s,i} = \{p^c_{s,i}\}$ can be obtained as,

$$p^c_{s,i} =< \boldsymbol{\psi}_i, \boldsymbol{q}_c >, c \in \{0, \cdots, C\}, \tag{5}$$

Since $\boldsymbol{p}_{t,i}, \boldsymbol{p}_{s,i}$ model the voxel-to-organ affinities in the embedding space, constraining consistency on them introduces rich context information for training on the unlabeled data, which is formulated as,

$$L^f_c = \frac{1}{Z^f_c} \sum_i KL(\boldsymbol{p}_{s,i}, \boldsymbol{p}_{t,i}), \tag{6}$$

where $\frac{1}{Z^f_c}$ is the normalization factor to get the mean KL-Divergence in the feature embedding space. Denote $\varphi_{tea} = f_{tea} \circ g_{tea}$, $\varphi_{stu} = f_{stu} \circ g_{stu}$ as the teacher and student segmentation model respectively, the logits from the student and the teacher branch can be calculated as $\boldsymbol{l}_{s,i} = \varphi_{stu}(I_{s,i}), \boldsymbol{l}_{t,i} = \varphi_{tea}(I_{t,i})$. Then the consistency loss in the label space is calculated as,

$$L^l_c = \frac{1}{Z^l_c} \sum_i KL(\boldsymbol{l}_{s,i}, \boldsymbol{l}_{t,i}), \tag{7}$$

where $\frac{1}{Z^l_c}$ is the normalization factor. The overall affinity-aware consistency loss is finally formulated as,

$$L_c = L^f_c + L^l_c, \tag{8}$$

2.4 Cross-Site Feature Alignment (CSFA) Module

The CSFA module is proposed to calibrate feature distributions across different sites. Specifically, given the learned prototypes from current mini-batch of the n-th PLD ($\{\boldsymbol{q}^n_c\}, c \in \{0, \cdots, C\}$), they can be regarded as the organ-specific

cluster centers in the embedding space. Then, compactness loss is introduced to calibrate D_p^n with the cluster centers learned from the FLD as,

$$L_a^{l,n} = \frac{1}{|Y_p^n|} \sum_{c \in Y_p^n} ||\boldsymbol{q}_c^n - \boldsymbol{q}_c||_2^2, \tag{9}$$

where $|Y_p^n|$ returns the number of labeled organs in D_p^n.

To further take into consideration the inter-organ affinity relationships during feature distribution alignment, we first model inter-organ affinity relationships on the FLD by calculating the affinity matrix $\boldsymbol{A} = \{a_{ij}\} \in \mathbb{R}^{(C+1)\times(C+1)}$ as shown in Fig. 1,

$$a_{ij} =< \boldsymbol{q}_i, \boldsymbol{q}_j >, i \in \{0, \cdots, C\},\ j \in \{0, \cdots, C\}, \tag{10}$$

Similarly, we can obtain the affinity matrix $\boldsymbol{A}_p^n = \{a_{ij}^n\} \in \mathbb{R}^{(C+1)\times(C+1)}$ on partially labeled dataset D_p^n as,

$$a_{ij}^n =< \boldsymbol{q}_i^n, \boldsymbol{q}_j^n >, i \in \{0, \cdots, C\},\ j \in \{0, \cdots, C\}, \tag{11}$$

Then the affinity relationship aware feature alignment loss is calculated as,

$$L_a^{g,n} = \frac{1}{C+1} \sum_c KL(\boldsymbol{a}_c, \boldsymbol{a}_{p,c}^n), \tag{12}$$

where $\boldsymbol{a}_c, \boldsymbol{a}_{p,c}^n$ refer to the c-th row of $\boldsymbol{A}$ and $\boldsymbol{A}_p^n$ respectively.

The overall cross-site alignment loss is then calculated as the sum of the compactness loss and the affinity relationship aware calibration loss,

$$L_a = L_a^{l,n} + L_a^{g,n}, \tag{13}$$

The overall training objective is finally formulated as,

$$L = L_{seg} + L_c + \lambda_a L_a. \tag{14}$$

where λ_a is the tradeoff parameter.

3 Experiments and Results

Datasets and Implementation Details. We use five abdominal CT datasets (MALBCVWC [1], Decathlon Spleen [3], KiTS [2], Decathlon Liver [3] and Decathlon Pancreas [3] datasets respectively) to evaluate the effectiveness of our method [1–3]. The spatial resolution of all these datasets are resampled to $(1 \times 1 \times 3)mm^3$. We randomly split each dataset into training (60%), validation (20%) and testing (20%). We adopt 3D U-Net [18] as our backbone model. The patch size (h, w, z) is set to $(160, 160, 96)$. The hyper-parameters α and τ are empirically set to 0.9, and 0.8, respectively. λ_a is initialized as 0.01 and linearly decreased to $1e-3$ at 20000 iterations. We use SGD optimizer to train the model and the initial learning rate is set to 0.01. We adopt Dice similarity coefficient (DSC) as metric to evaluate the performance of different methods.

Table 1. Results of the ablation study on the effectiveness of each component in our method (Metric: DSC (%)).

Settings	Liver	Spleen	Pancreas	RK	LK	Overall
baseline	94.7	91.9	77.5	94.1	93.5	90.3
baseline + ACL	94.5	**94.1**	77.4	95.5	95.3	91.4
baseline + ACL + CSFA	**95.2**	93.8	**79.0**	**96.0**	**95.5**	**91.9**

Table 2. Analysis on the effectiveness of CSFA in mitigating cross-site data discrepancy. Please note D_0 - D_4 refer to the MALBCVWC [1], Decathlon-Spleen [3], KiTS [2], Decathlon-Liver [3] and Decathlon-Pancreas [3] datasets respectively, where D_0 is the FLD, and others are PLDs. Please note small MMD indicates small data discrepancy.

Settings	D_0 vs D_1	D_0 vs D_2	D_0 vs D_3	D_0 vs D_4	Overall
Ours wo/CSFA	0.3030	0.3187	0.2818	0.3577	0.3153
Ours w/CSFA	0.1589	0.2036	0.1358	0.2925	0.1977

Ablation Study. In this subsection, we carry out experiments to investigate effectiveness of each component in the proposed framework. Concretely, the baseline results are trained with only the L_{seg} loss. In Table 1, the "baseline+ACL" setting reports the results with our proposed affinity-aware consistency learning scheme. Comparing to the baseline, it brings a 1.1% performance gain in terms of DSC. By introducing the CSFA module, the "baseline+ACL+CSFA" setting can further boost the performance by 0.5% in terms of DSC.

We further study the effectiveness of the CSFA module in alleviating cross-site data discrepancy. Concretely, we measure the feature distribution discrepancy between the FLD and each PLD by calculating the Maximum Mean Discrepancy (MMD) using gaussian kernel [19], which was designed to quantify domain discrepancy. We conduct "full vs partial" MMD analysis on the following two settings: "Ours w/CSFA" and "Ours wo/CSFA", where "Ours w/CSFA" is the proposed framework, while "Ours wo/CSFA" setting refers to removing the CSFA module from our framework. In the MMD calculation, for each dataset, we first generate features from the penultimate layer. Then we randomly select 2000 features in each class for MMD calculation. Please note, for each PLD, we adopt the pseudo labels for feature selection. Detailed comparison results are illustrated in Table 2. As shown, by introducing the CSFA module, the feature distribution discrepancy in terms of MMD can be effectively alleviated across all the "full vs partial" dataset pairs.

Comparison with the State-of-the-Art (SOTA) Methods. We compare with four SOTA methods, including PaNN [8], PIPO [7], Marginal Loss [9], and DoDNet [5]. For fair comparison, all the SOTA methods were trained/tested on our own dataset splits. We also implemented our method taking the nnUNet as the backbone to compare with Marginal Loss [9] and PaNN [8]. We reported the

Table 3. Comparison with state-of-the-art methods in terms of DSC. "RK", "LK" refer to "Right Kidney" and "Left Kidney" respectively.

Methods	backbone	Liver	Spleen	Pancreas	RK	LK	Overall	Avg_{hard}
PIPO [7]	3D-UNet	93.01	93.63	76.51	93.50	89.98	89.3	86.7
DoDNet [5]	3D-UNet	95.41	95.09	70.01	94.06	92.00	89.3	85.4
Marginal Loss [9]	nnUNet	95.45	94.88	77.91	94.14	91.52	90.8	87.9
PaNN [8]	nnUNet	95.13	95.14	78.88	96.21	91.02	91.3	88.7
Ours	3D-UNet	95.2	93.82	79.03	96.04	95.49	91.9	90.2
Ours	nnUNet	95.8	95.0	83.1	94.7	93.8	92.5	90.5

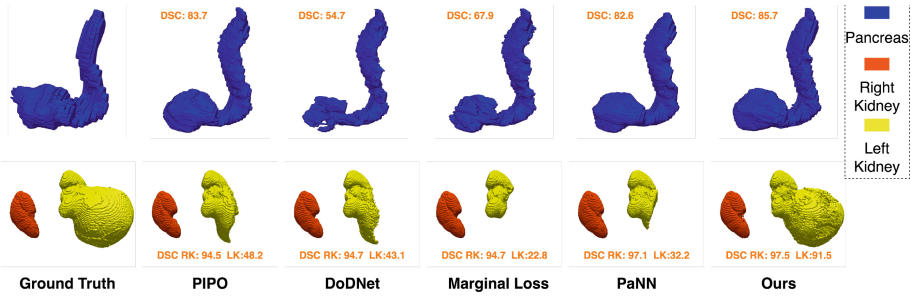

Fig. 2. 3D visualized results of some hard samples.

DSC values for each organ across test sets from all the datasets. For a straightforward comparison with the SOTA, we also recorded the average DSC over all the organs. Detailed results are illustrated in Table 3. As shown, our method achieves the best performance. Specifically, our method outperforms the second-best method PaNN [8] with a 1.2% DSC gain using the same nnUNet backbone. And our method when taking 3D-UNet as the backbone also outperforms the listed SOTA methods. We further conduct paired t-test to compare the difference between ours and other SOTA methods, the p-values are 2E-8 (PIPO), 2E-5 (DoDNet), 2E-4 (Marginal Loss), 0.037 (PaNN), respectively. As all p-values are smaller than 0.05, the differences between ours and other SOTA methods are statistically significant.

In practice, some organs are much harder to be well-segmented than others due to their relatively small organ sizes. Therefore, we pay more attention to the performance on those hard organs (in our datasets, Pancreas and Kidneys are deemed to be more difficult due to their relatively small sizes). From the last column of Table 3, we can see that the segmentation performance gains of our method are more pronounced on hard organs (on average a 1.8% DSC gain). Figure 2 demonstrates the qualitative visualization results on some hard samples. As shown in this figure, our method can generate better segmentation results than other SOTA methods. Besides, the reasonable performance on segmenting

kidney with tumors (row 2 in Fig. 2) makes our method promising in clinical practice.

4 Conclusion

In this paper, we designed a novel Affinity-aware Consistency Learning scheme (ACL) to model voxel-to-organ affinity context in the feature embedding space into consistency learning. Meanwhile, the CSFA module was designed to perform feature distribution alignment across different sites, where both organ-specific cluster centers and the inter-organ affinity relationships were propagated from the small-sized FLD to PLDs for cross-site feature alignment. Extensive ablation studies validated effectiveness of each component in our method. Quantitative and Qualitative comparison results with other SOTA methods demonstrated superior performance of our method.

References

1. Bennett, L., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault-workshop and challenge. In: Proceedings of MICCAI Multi-Atlas Labeling Beyond Cranial Vault-Workshop Challenge. vol. 5, pp. 12 (2015)
2. Nicholas, H., 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)
3. Amber L.S., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063 (2019)
4. Konstantin, D., Kaufman, A.E.: Learning multi-class segmentations from single-class datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9501–9511 (2019)
5. Zhang, J., Xie, Y., Xia, Y., Shen, C.: DoDNet: learning to segment multi-organ and tumors from multiple partially labeled datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1195–1204 (2021)
6. Huang, R., Zheng, Y., Hu, Z., Zhang, S., Li, H.: Multi-organ segmentation via co-training weight-averaged models from few-organ datasets. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12264, pp. 146–155. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59719-1_15
7. Xi, F., Yan, P.: Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction. IEEE Trans. Med. Imaging **39**(11), 3619–3629 (2020)
8. Zhou, Y., et al.: Prior-aware neural network for partially-supervised multi-organ segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10672–10681 (2019)
9. Shi, G., Xiao, L., Chen, Y., Zhou, S.K.: Marginal loss and exclusion loss for partially supervised multi-organ segmentation. Med. Image Anal. **70**, 101979 (2021)
10. Jisoo, J., Lee, S., Kim, J., Kwak, N.: Consistency-based semi-supervised learning for object detection. In: Advances in Neural Information Processing Systems. vol. 32 (2019)

11. Abuduweili, A., Li, X., Shi, H., Xu, C.Z., Dou, D.: Adaptive consistency regularization for semi-supervised transfer learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6923–6932 (2021)
12. Ouali, Y., Hudelot, C., Tami, M. Semi-supervised semantic segmentation with cross-consistency training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12674–12684 (2020)
13. Samuli, L., Aila, T.: Temporal ensembling for semi-supervised learning. In: International Conference on Learning Representations (ICLR) (2017)
14. Antti, T., Valpola, H.: Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in Neural Information Processing Systems. vol. 30 (2017)
15. French, G., Laine, S., Aila, T., Mackiewicz, M., Finlayson, G.: Semi-supervised semantic segmentation needs strong, varied perturbations. In: British Machine Vision Conference (2020)
16. Zheng, K., Xu, J., Wei, J.: Double noise mean teacher self-Ensembling model for semi-supervised tumor segmentation. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1446–1450 (2022)
17. Bento, M., Fantini, I., Park, J., Rittner, L., Frayne, R.: Deep learning in large and multi-site structural brain MR imaging datasets. Front. Neuroinformatics **15**(82), 805669 (2022)
18. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., Ronneberger, O.: 3D-UNet: learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention-MICCAI, pp. 424–432 (2016)
19. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. J. Mach. Learn. Res. **13**(1), 723–773 (2012)