ConDistFL: Conditional Distillation for Federated Learning from Partially Annotated Data

Pochuan Wang[1], Chen Shen[2], Weichung Wang[1], Masahiro Oda[2],
Chiou-Shann Fuh[1], Kensaku Mori[2], and Holger R. Roth[3(✉)]

[1] National Taiwan University, Taipei City, Taiwan
[2] Nagoya University, Nagoya, Japan
[3] NVIDIA Corporation, Santa Clara, USA
hroth@nvidia.com

Abstract. Developing a generalized segmentation model capable of simultaneously delineating multiple organs and diseases is highly desirable. Federated learning (FL) is a key technology enabling the collaborative development of a model without exchanging training data. However, the limited access to fully annotated training data poses a major challenge to training generalizable models. We propose "ConDistFL", a framework to solve this problem by combining FL with knowledge distillation. Local models can extract the knowledge of unlabeled organs and tumors from partially annotated data from the global model with an adequately designed conditional probability representation. We validate our framework on four distinct partially annotated abdominal CT datasets from the MSD and KiTS19 challenges. The experimental results show that the proposed framework significantly outperforms FedAvg and FedOpt baselines. Moreover, the performance on an external test dataset demonstrates superior generalizability compared to models trained on each dataset separately. Our ablation study suggests that ConDistFL can perform well without frequent aggregation, reducing the communication cost of FL. Our implementation will be available at https://github.com/NVIDIA/NVFlare/tree/main/research/condist-fl.

Keywords: Federated learning · Partially labeled datasets · Multi-organ and tumor segmentation · Abdominal CT

1 Introduction

Accurately segmenting abdominal organs and malignancies from computed tomography (CT) scans is crucial for clinical applications such as computer-aided diagnosis and therapy planning. While significant research has focused on segmenting individual organs [1,7] and multiple classes of organs without malignancies [4,6], a generalized model capable of handling multiple organs and diseases simultaneously is desirable in real-world healthcare scenarios. Traditional

supervised learning methods, on the other hand, rely on the amount and quality of the training data. Regrettably, the cost of high-quality medical image data contributed to a paucity of training data. For many anatomies, only trained professionals can produce accurate annotations on medical images. On top of this, even experts often only have specialized knowledge for a specific task, making it challenging to annotate the organs and corresponding malignancies of various anatomies and imaging modalities.

The lack of sufficient annotated datasets for multiple organs and tumors poses a significant challenge in developing generalized segmentation models. To address this issue, several studies have explored partially annotated datasets, where only a subset of targeted organs and malignancies are annotated in each image, to build generalized segmentation models [5,10,11,21,23]. However, sharing private medical datasets among institutions raises privacy and regulatory concerns. To overcome these challenges, federated learning (FL) was introduced [16]. FL enables collaborative training of a shared (or "global") model across multiple institutions without centralizing the data in one location.

FL has emerged as a promising technology to enhance the efficiency of medical image segmentation [19,24,25]. In FL, each client trains a local model using its data and resources while only sending model updates to the server. The server then combines these updates into a global model using "FedAvg" [16]. Recent studies have utilized FL to develop unified multi-organ segmentation models using partially annotated abdominal datasets [15,26] as illustrated in Fig. 1. However, these approaches often neglect lesion areas. Only a few studies attempt to generate to segment the various organs and their cancers simultaneously [20,29]. The model aggregation in FL is a major hurdle because of the data heterogeneity problem brought on by data diversity [30]. Merging models from different sources with non-IID data can lead to performance degradation. This issue is further exacerbated when clients use data annotated for different tasks, introducing more domain shifts in the label space. Additionally, unbalanced dataset sizes among clients may affect the global model's performance on tasks with limited data.
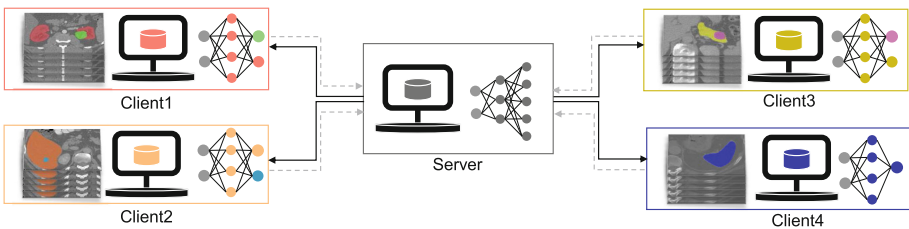

Fig. 1. An illustration of the ConDistFL framework for multi-organ and tumor segmentation from partial labels. Each client has only a subset of the targeted organs and malignancies annotated in their local datasets.

In this work, we suggest a framework to tackle data heterogeneity in FL for multi-class organ and tumor segmentation from partially annotated abdominal CT images. The main contributions of this work are as follows:

1. Our proposed conditional distillation federated learning (ConDistFL) framework enables joint multi-task segmentation of abdominal organs and malignancies without additional fully annotated datasets.
2. The proposed framework exhibits stability and performance with long local training steps and a limited number of aggregations, reducing data traffic and training time in real-world FL scenarios.
3. We further validate our models on an unseen fully annotated public dataset AMOS22 [13]. The robustness of our approach is supported by both the qualitative and quantitative evaluation results.

2   Method

ConDistFL extends the horizontal FL paradigm [27] to handle partially annotated datasets distributed across clients. An illustration of our ConDistFL framework for multi-organ and tumor segmentation from partial labels is shown in Fig. 1. In the client training of ConDistFL, we combine supervised learning on ground truth labels and knowledge distillation learning [9] using the global model's predictions. During supervised learning, we adopt the design of marginal loss [21] to avoid knowledge conflicts caused by missing labels. To improve the knowledge distillation in FL settings, we proposed a conditional distillation loss to maximize the agreement between the global model and local model predictions on unlabeled voxels.
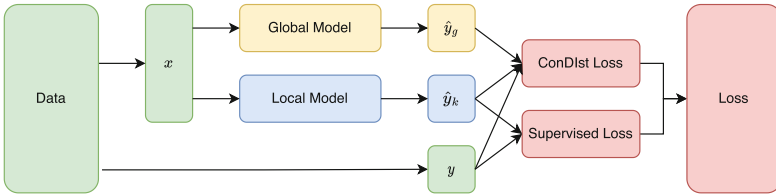

Fig. 2. ConDistFL data flow diagram for client k; x is a batch of image patches from the local dataset; y is the corresponding label; $\hat{y}_g$ is the output of global model; and $\hat{y}_k$ is the output of the local model.

2.1   Conditional Distillation for Federated Learning

In ConDistFL, the client keeps the latest global model as the teacher model for knowledge distillation and uses the local model as the student model. Figure 2 illustrates the training data flow in client k and the relationship between the global, local, and loss functions.

2.2   Supervised Loss

We adopt the design of marginal loss [21] for the supervised loss $\mathcal{L}_{sup}$. Let N be the total number of classes across all datasets, F_k be a collection of foreground classes on client k, B_k be the background class, and all unlabeled classes on client k, and $\hat{y}_{k,i}$ be the output logits of client k's model for class i. By applying softmax normalization on the output logits $\hat{y}_{k,i}$ for each $i \in N$, we can derive the output probability $\hat{p}_{k,i}$ for each class i as

$$\hat{p}_{k,i} = \frac{e^{\hat{y}_{k,i}}}{\sum_{j=0}^{N} e^{\hat{y}_{k,j}}} \text{ for } i = 0, \ 1, \ 2, \ \ldots, \ N-1. \tag{1}$$

Similar to the marginal loss, all probabilities in the background B_k are merged into one for a new non-foreground class. The probabilities remain the same as $\hat{p}_{k,i}$ for all $i \in F_k$. Then we apply Dice loss [17] with cross-entropy loss [28] (DiceCELoss) for supervised learning of the segmentation model. The final loss term $\mathcal{L}_{sup}$ is defined as

$$\mathcal{L}_{sup} = \text{DiceCELoss}(\hat{p}'_k, y'), \tag{2}$$

where the background merged probability is $\hat{p}'_k$, and the corresponding one-hot label is y'.

2.3   ConDist Loss

For the conditional distillation (ConDist) loss $\mathcal{L}_{ConDist}$, we normalize the output logits of both the global and the local model using softmax with temperature τ. The normalized logits from the global model $\hat{p}_g^\tau$ and the k-th client's local model $\hat{p}_k^\tau$ is defined as

$$\hat{p}_k^\tau = \text{softmax}(\hat{y}_k/\tau) \quad \text{and} \quad \hat{p}_g^\tau = \text{softmax}(\hat{y}_g/\tau), \tag{3}$$

where τ is set to 0.5 to enhance the confidence of the model output.

Foreground Merging and Background Grouping. Contrary to the supervised loss, we merge the probabilities for class i for all $i \in F_k$ in $\hat{p}_k^\tau$ and $\hat{p}_g^\tau$. Then we define

$$\hat{p}_{k,F_k} = \sum_{i \in F_k} \hat{p}_{k,i}^\tau \quad \text{and} \quad \hat{p}_{g,F_k} = \sum_{i \in F_k} \hat{p}_{g,i}^\tau, \tag{4}$$

where the $\hat{p}_{k,i}^\tau$ and $\hat{p}_{g,i}^\tau$ are the probabilities for class i in $\hat{p}_k^\tau$ and $\hat{p}_g^\tau$, respectively. Moreover, we group the background classes in client k by the organs in the background B_k. Let M_k be the number of unlabeled organs in client k, and $\mathcal{O}_k = \{G_0, G_1, G_2, \ldots G_{M_k}\}$. G_0 is the set containing the global background class. The probability for each background group can be calculated as

$$\hat{p}_{k,G_i} = \sum_{j \in G_i} \hat{p}_{k,j}^\tau \quad \text{and} \quad \hat{p}_{g,G_i} = \sum_{j \in G_i} \hat{p}_{g,j}^\tau \quad \text{for} \quad i = 0, \ 1, \ \ldots, \ M_k, \quad (5)$$

where G_i is a set containing the class of the healthy part of unlabeled organ i and all the classes of associated lesions to the organ i.

Conditional Probability for Background Organs. We define conditional probabilities $\hat{p}_{k,\mathcal{O}_k|B_k}$ and $\hat{p}_{g,\mathcal{O}_k|B_k}$ as

$$\hat{p}_{k,\mathcal{O}_k|B_k} = \left(\frac{\hat{p}_{k,G_0}}{1 - \hat{p}_{k,F_k}}, \frac{\hat{p}_{k,G_1}}{1 - \hat{p}_{k,F_k}}, \ldots, \frac{\hat{p}_{k,G_{M_k}}}{1 - \hat{p}_{k,F_k}} \right), \quad (6)$$

$$\hat{p}_{g,\mathcal{O}_k|B_k} = \left(\frac{\hat{p}_{g,G_0}}{1 - \hat{p}_{g,F_k}}, \frac{\hat{p}_{g,G_1}}{1 - \hat{p}_{g,F_k}}, \ldots, \frac{\hat{p}_{g,G_{M_k}}}{1 - \hat{p}_{g,F_k}} \right), \quad (7)$$

where $1 - \hat{p}_{k,F_k}$ and $1 - \hat{p}_{g,F_k}$ are the total probabilities of all classes in B_k with respect to $\hat{p}_k^\tau$ and $\hat{p}_g^\tau$. The conditional probability $\hat{p}_{k,\mathcal{O}_k|B_k}$ and $\hat{p}_{g,\mathcal{O}_k|B_k}$ are the probabilities given that the prediction is only in B_k.

Foreground Filtering. To avoid learning from incorrect predictions and reduce the potential conflict with the supervised loss, we filter off undesired voxels with a mask operation $\mathcal{M}$, which removes the union of foreground area in ground truth label y and all the area in $\hat{y}_g$ where the predictions are in F_k.

Segmentation ConDist Loss. Combining the conditional probability $\hat{p}_{k,\mathcal{O}_k|B_k}$, $\hat{p}_{g,\mathcal{O}_k|B_k}$, and the foreground filtering mask $\mathcal{M}$, we define the ConDist loss $\mathcal{L}_{ConDist}$ for segmentation task by applying soft Dice loss as

$$\mathcal{L}_{ConDist} = DiceLoss(\mathcal{M}(\hat{p}_{k,\mathcal{O}_k|B_k}), \mathcal{M}(\hat{p}_{g,\mathcal{O}_k|B_k})). \quad (8)$$

To handle meaningless global model predictions in the initial FL rounds, we incorporate an adjustable weight w for the ConDist loss, gradually increasing it as the FL round number increments. The total loss $\mathcal{L}$ for ConDistFL is defined as

$$\mathcal{L} = \mathcal{L}_{sup} + w * \mathcal{L}_{ConDist}. \quad (9)$$

In practice, we schedule the weight w from 0.01 to 1.0 linearly.

3   Experiments

We conducted our experiments on the Medical Segmentation Decathlon (MSD) [22] and the KiTS19 Challenge [8] datasets. In the MSD dataset, we only used the liver, pancreas, and spleen subsets. Except for the spleen dataset, each above

dataset includes annotations for the organs and tumors. We split the dataset into training, validation, and testing subsets by 60%, 20%, and 20%, respectively. For the non-FL standalone training, each model only uses a single dataset. For FL training, we distributed the four datasets to four independent clients. In addition, we evaluated our models on the multi-modality Abdominal Multi-Organ Segmentation Challenge 2022 (AMOS22) dataset [13], which consists of 300 CT volumes with 15 abdominal organ annotations. To accommodate the labeling format of AMOS22, where healthy organs and lesions are not distinguished, we merged the tumor predictions with associated organs from our model output before computing the metrics using ground truth labels.

The nnU-Net [12] data preprocessing pipeline was adopted with minor modifications. We first resampled the images to a median spacing of [1.44, 1.44, 2.87] millimeters and clipped the intensity to the range [−54, 258]. Then we applied z-score normalization by assuming the mean intensity under ROI to be 100 and its standard deviation to be 50 since the complete multi-organ ROI is unavailable. We set the input patch size to [224, 224, 64] and the training batch size to 4.

Our neural network backbone was built using the 3D DynU-Net from MONAI [3], an effective and flexible U-Net implementation. Deep supervision was also enabled to speed up the training process and enhance model performance. The deep supervision loss is identical to the supervised loss with an extra exponential weight decay. We trained our models using stochastic gradient descent (SGD) and cosine annealing schedule. The initial learning rate was set to 10^{-2} and decreased gradually to 10^{-7}.

The loss function for the non-FL standalone baselines is Dice loss with crossentropy. For the FL experiments, we evaluated FedAvg [16], FedOpt [2], FedProx [14], and ConDistFL. To assess the effectiveness of the marginal loss in Sect. 2.2, we trained two sets of FedAvg models: one using the standard Dice loss and the other employing the marginal loss. The FedProx model was trained with the FedAvg aggregator and $\mu = 0.01$. For FedOpt and ConDistFL, we utilized the Federated Optimization (FedOpt) aggregation method, with an additional SGD optimizer with momentum $m = 0.6$ implemented on the server.

We employed a single NVIDIA V100 GPU for each standalone experiment and FL client. The FL experiments were implemented using NVIDIA FLARE [18].

4   Results and Discussion

Our experiments encompass standalone baselines using a single dataset, an ablation study of standard Dice loss on FedAvg (FedAvg*), marginal loss on FedAvg, FedProx, and FedOpt, and the combined marginal loss and ConDist loss on ConDistFL. To establish a fair comparison between related works, we trained a ConDistFL (Union) model with the same learning targets as [15,26], i.e., only to segment the union of the organs and tumors. Additionally, we evaluated the proposed method on the unseen AMOS22 dataset to demonstrate its generalizability.

Table 1 compares the average Dice score of each task between standalone baseline models and the best performance server models for FedAvg, FedProx, FedOpt, and ConDistFL. All the models are trained for a total of 120,000 steps to allow for a fair comparison. For the best FedAvg*, FedAvg, and ConDistFL models, we utilized 60 FL aggregation rounds with 2000 local steps per round. As for the FedProx and FedOpt best model, we conducted 120 FL rounds and 1000 local steps per round.

The results of FedAvg* and FedAvg demonstrate that the marginal loss effectively resolves conflicts between inconsistent labels and yields reasonable performance. ConDistFL stands out as the top-performing method among all experiments utilizing the marginal loss. FedAvg, FedProx, and FedOpt show similar performance overall, with FedAvg and FedOpt delivering acceptable results for most tasks, except for the pancreas and tumor. In contrast, FedProx performs well for the pancreas and tumor, but there is a notable drop in performance on other tasks. This suggests that although the FedProx loss can regularize the models for heterogeneous clients like the proposed ConDist loss, its task-agnostic nature harms the performance when training on multiple tasks with different partial labels.

Table 1. Comparison between non-FL results and each FL result. The average Dice score of each organ for the standalone model is computed separately from four distinct models. FedAvg* indicates the model trained with FedAvg and standard Dice loss.

	Kidney	Tumor	Liver	Tumor	Pancreas	Tumor	Spleen	Average
Standalone	0.9563	0.8117	0.9525	0.7071	0.7974	0.5012	0.9632	0.8102
FedAvg*	0.7707	0.4894	0.4937	0.3202	0.5403	0.1396	0.0000	0.3934
FedAvg	0.9419	0.6690	0.9381	0.6500	0.6933	0.2985	0.9059	0.7281
FedProx	0.9247	0.6799	0.8972	0.6244	0.7419	**0.4033**	0.7060	0.7111
FedOpt	0.9473	0.7212	0.9386	0.6087	0.6734	0.2390	0.9394	0.7239
ConDistFL	**0.9477**	**0.7333**	**0.9446**	**0.6944**	**0.7478**	0.3660	**0.9562**	**0.7700**

The ablation study in Fig. 3 investigates the impact of the number of local training steps on the global model performance. Increasing the number of local training steps from 100 to 1000 for all tested methods improved performance. However, when more local training steps were used, both FedAvg and FedOpt encountered model divergence issues, with FedAvg experiencing a more significant performance drop than FedOpt. In contrast, ConDistFL consistently delivered better performance across different local step experiments. This can be attributed to ConDistFL providing a common task, preventing model divergence in FL, and the inherent complexity of tumor segmentation requiring more local training steps. By maintaining consistent representations of unknown classes, ConDistFL allows for the use of larger local step sizes to learn the tumor segmentation task effectively.

Table 2 compares the average Dice scores of standalone baselines, ConDistFL, and ConDistFL (Union) on the unseen AMOS22 dataset. ConDistFL demon-
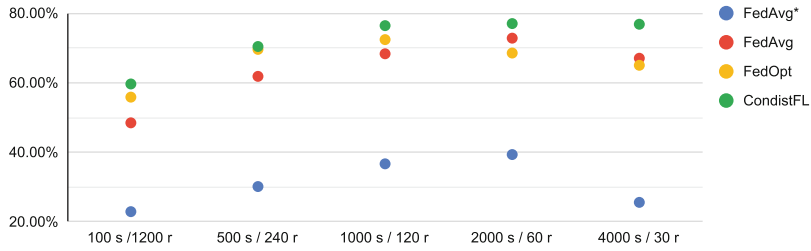
Fig. 3. The ablation study results on the test set. The x-axis is the number of local training steps (s) and rounds numbers (r), while the y-axis is the average Dice score of all organs and tumors.

strates significant generalizability improvements over the standalone baselines, while ConDistFL (Union) further enhances performance. This highlights the challenge of segmenting tumors and organs together compared to considering them a single class.

Table 2. External test results for AMOS22 dataset in average Dice scores.

	Kidney	Liver	Pancreas	Spleen	Average
Standalone	0.5916	0.9419	0.5944	0.8388	0.7417
FedAvg	0.5032	0.8718	0.4637	0.5768	0.6039
FedProx	0.4698	0.6994	0.5185	0.7120	0.5999
FedOpt	0.5171	0.6740	0.4113	0.6418	0.5611
ConDistFL	0.7218	0.9191	0.6188	0.8556	0.7788
ConDistFL (Union)	**0.8746**	**0.9471**	**0.7401**	**0.9079**	**0.8674**

Table 3 compares ConDistFL (Union) and the reported performance of FL PSMOS and MENU-Net on the MSD test set. FL PSMOS utilizes a fully annotated dataset for model pre-training, while MENU-Net introduces a fifth client with a fully annotated dataset. The results demonstrate that ConDistFL achieves comparable performance without needing fully annotated data. Additionally, ConDistFL significantly reduces the number of aggregation rounds, leading to substantial savings in data traffic and synchronization overheads.

Figure 4 showcases 3D visualizations of our proposed ConDistFL, demonstrating effective and simultaneous segmentation of multiple organs and tumors without ensembling. Compared to FedAvg and FedOpt, ConDistFL achieves smoother and more continuous segmentations. The comparison with the ground truth of AMOS22 validates the generalizability of our FL framework.

Table 3. Comparing the average Dice scores of ConDistFL to reported performance of related works. "Rounds" is the number of FL aggregation rounds.

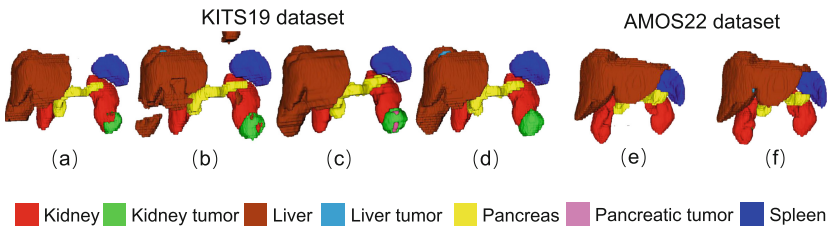|	Rounds	Kidney	Liver	Pancreas	Spleen	Average
FL PSMOS [15]	2,000	**0.966**	0.938	0.788	**0.965**	0.9143
MENU-Net [26]	400	0.9594	0.9407	0.8005	0.9465	0.9118
ConDistFL (Union)	120	0.9657	**0.9619**	**0.8210**	0.9626	**0.9278**

Fig. 4. 3D renderings of segmentation on the best performed FL server model using (a) FedAvg, (b) FedOpt, (c) FedProx, (d) ConDistFL on KITS19 data, and (e) ground truth and (f) the external segmentation using ConDistFL on AMOS22 data.

5   Conclusion

This work offers a promising FL approach for generalized segmentation models from partially annotated abdominal organs and tumors, reducing annotation costs and speeding up model development. Moreover, the proposed method requires less frequent aggregation, making it suitable for real-world FL scenarios with limited communication bandwidth.

References

1. Altini, N., et al.: Liver, kidney and spleen segmentation from CT scans and MRI with deep learning: a survey. Neurocomputing **490**, 30–53 (2022)
2. Asad, M., Moustafa, A., Ito, T.: FedOpt: towards communication efficiency and privacy preservation in federated learning. Appl. Sci. **10**(8), 2864 (2020)
3. Cardoso, M.J., et al.: MONAI: an open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022)
4. Cerrolaza, J.J., et al.: Computational anatomy for multi-organ analysis in medical imaging: a review. Med. Image Anal.**56**, 44–67 (2019). https://doi.org/10.1016/j.media.2019.04.002, https://linkinghub.elsevier.com/retrieve/pii/S1361841518306273
5. Fang, X., Yan, P.: Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction. IEEE Trans. Med. Imaging **39**(11), 3619–3629 (2020)
6. Fu, Y., Lei, Y., Wang, T., Curran, W.J., Liu, T., Yang, X.: A review of deep learning based methods for medical image multi-organ segmentation. Physica Med. **85**, 107–122 (2021)

7. Gul, S., Khan, M.S., Bibi, A., Khandakar, A., Ayari, M.A., Chowdhury, M.E.: Deep learning techniques for liver and liver tumor segmentation: a review. Comput. Biol. Med. **157**, 105620 (2022)

8. Heller, N., et al.: The KiTs19 challenge data: 300 kidney tumor cases with clinical context, CT semantic segmentations, and surgical outcomes. https://doi.org/10.48550/ARXIV.1904.00445, https://arxiv.org/abs/1904.00445 (2019)

9. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network (2015)

10. Hongdong, M., et al.: Multi-scale organs image segmentation method improved by squeeze-and-attention based on partially supervised learning. Int. J. Comput. Assist. Radiol. Surg. **17**(6), 1135–1142 (2022)

11. 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

12. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)

13. Ji, Y., et al.: AMOS: a large-scale abdominal multi-organ benchmark for versatile medical image segmentation. arXiv preprint arXiv:2206.08023 (2022)

14. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. Proc. Mach. learn. Syst. **2**, 429–450 (2020)

15. Liu, P., Sun, M., Zhou, S.K.: Multi-site organ segmentation with federated partial supervision and site adaptation (2023). http://arxiv.org/abs/2302.03911

16. McMahan, H.B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: AISTATS (2017)

17. 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)

18. Roth, H.R., et al.: NVIDIA FLARE: Federated learning from simulation to Real-World. (2022)

19. Sheller, M.J., et al.: Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci. Rep. **10**(1), 1–12 (2020)

20. Shen, C., et al.: Joint multi organ and tumor segmentation from partial labels using federated learning. In: Albarqouni, S., et al. Distributed, Collaborative, and Federated Learning, and Affordable AI and Healthcare for Resource Diverse Global Health, pp. 58–67. Springer Nature, Switzerland (2022). https://doi.org/10.1007/978-3-031-18523-6_6

21. Shi, G., Xiao, L., Chen, Y., Kevin Zhou, S.: Marginal loss and exclusion loss for partially supervised multi-organ segmentation (2020)

22. Simpson, A.L., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. CoRR abs/1902.09063. http://arxiv.org/abs/1902.09063 (2019)

23. Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J.N., Wu, Z., Ding, X.: Embracing imperfect datasets: a review of deep learning solutions for medical image segmentation. Med. Image Anal. **63**, 101693 (2020)

24. Tajbakhsh, N., Roth, H., Terzopoulos, D., Liang, J.: Guest editorial annotation-efficient deep learning: the holy grail of medical imaging. IEEE Trans. Med. Imaging **40**(10), 2526–2533 (2021)

25. Xu, J., Glicksberg, B.S., Su, C., Walker, P., Bian, J., Wang, F.: Federated learning for healthcare informatics. J. Healthc. Inf. Res. **5**, 1–19 (2021)
26. Xu, X., Yan, P.: Federated multi-organ segmentation with partially labeled data. arXiv preprint arXiv:2206.07156 (2022)
27. Yang, Q., Liu, Y., Chen, T., Tong, Y.: Federated machine learning: concept and applications. ACM Trans. Intell. Syst. Technol. (TIST) **10**(2), 1–19 (2019)
28. Yi-de, M., Qing, L., Zhi-Bai, Q.: Automated image segmentation using improved PCNN model based on cross-entropy. In: Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004, pp. 743–746. IEEE (2004)
29. 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 (CVPR), pp. 1195–1204 (2021)
30. Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., Chandra, V.: Federated learning with Non-IID data. arXiv preprint arXiv:1806.00582 (2018)