

SegViz: A federated-learning based framework for multi-organ segmentation on heterogeneous data sets with partial annotations

Adway Kanhere^{1,2}, Pranav Kulkarni², Paul H. Yi², and Vishwa S. Parekh²

¹ Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD
akanher1@jhu.edu

² University of Maryland Medical Intelligent Imaging (UM2ii) Center, Baltimore, MD
{akanhere,pkulikarni,pyi,vparekh}@som.umd.edu

Abstract. Segmentation is one of the most primary tasks in deep learning for medical imaging, owing to its multiple downstream clinical applications. However, generating manual annotations for medical images is time-consuming, requires high skill, and is an expensive effort, especially for 3D images. One potential solution is to aggregate knowledge from partially annotated datasets from multiple groups to collaboratively train global models using Federated Learning. To this end, we propose SegViz, a federated learning-based framework to train a segmentation model from distributed non-i.i.d datasets with partial annotations. The performance of SegViz was compared against training individual models separately on each dataset as well as centrally aggregating all the datasets in one place and training a single model. The SegViz framework using FedBN as the aggregation strategy demonstrated excellent performance on the external BTCV set with dice scores of 0.93, 0.83, 0.55, and 0.75 for segmentation of liver, spleen, pancreas, and kidneys, respectively, significantly ($p < 0.05$) better (except spleen) than the dice scores of 0.87, 0.83, 0.42, and 0.48 for the baseline models. In contrast, the central aggregation model significantly ($p < 0.05$) performed poorly on the test dataset with dice scores of 0.65, 0, 0.55, and 0.68. Our results demonstrate the potential of the SegViz framework to train multi-task models from distributed datasets with partial labels. All our implementations are open-source and available at <https://anonymous.4open.science/r/SegViz-B746>

Keywords: Federated Learning · Partial annotations · Segmentation.

1 Introduction

Medical image segmentation is one of the most fundamental tasks in automated medical image analysis as it forms the basis for many downstream applications, including diagnosis, prognosis and treatment planning, image reconstruction, and treatment response assessment [5]. As a result, many large-scale datasets have been curated and released for the segmentation of different organ types and tumor structures [1]. However, each of these datasets has been curated for a

specific use case and therefore, focuses on segmenting only a particular organ or tumor subset in the body. Consequently, developing and deploying algorithms for each use case would potentially result in hundreds of models, thereby limiting their clinical utility – imagine deploying a different algorithm for every type of cancer, injury, and other diseases. Considering the above limitations, the situation is further amplified by the time-consuming and expensive manual annotations required to build large-scale fully annotated multi-organ datasets.

The challenge of training several individual models separately and the need for large-scale multi-organ datasets can be addressed by training multi-task segmentation models from distributed datasets using collaborative learning. Federated learning (FL) has gained importance in recent years for solving this challenge by collaboratively training one global model from several local models without data sharing. However, the capability of FL in aggregating knowledge from datasets curated at different imaging centers is challenging as each imaging center may focus on related but different tasks; suppose one center is training a liver segmentation model while another center is training a spleen segmentation model from CT scans. These two datasets would contain images with a similar field of view but different, incomplete annotations, as illustrated in 1. Such a situation, where one dataset has only a few organs annotated while another dataset contains no overlapping annotations with the first one is very common in medical imaging.

In this paper, we propose SegViz, a federated learning (FL) based framework for aggregating knowledge from heterogeneous, distributed medical imaging datasets with distinct and partial annotations into a single ‘global’ model. We evaluated the SegViz framework for the task of segmenting four organs - liver, spleen, pancreas, and kidneys on CT scans using distributed nodes each containing one local dataset. We compare the performance of SegViz trained global models to models trained individually on each dataset as well as a model trained by centrally aggregating all the datasets.

2 Related Work

Generating manual annotations for medical images is time-consuming, requires high skill, and is an expensive effort, especially for 3D images [19]. One potential solution is to curate datasets with partial annotations, wherein only a subset of structures is annotated for each image or volume. Furthermore, knowledge from similar partially annotated datasets from multiple groups can be aggregated to collaboratively train global models using Federated Learning [6]. Knowledge aggregation would not only save time but also allow different groups to benefit from each other’s annotations without explicitly sharing them. Consequently, different techniques have been proposed in the literature for aggregating knowledge from heterogeneous datasets with partial, incomplete labels [16,17,2,18].

There has been considerable research in the past on developing multi-task segmentation models using partial labels. The works of [21,11] show how to create subsets of the partially labeled datasets to create fully labeled subsets. However,

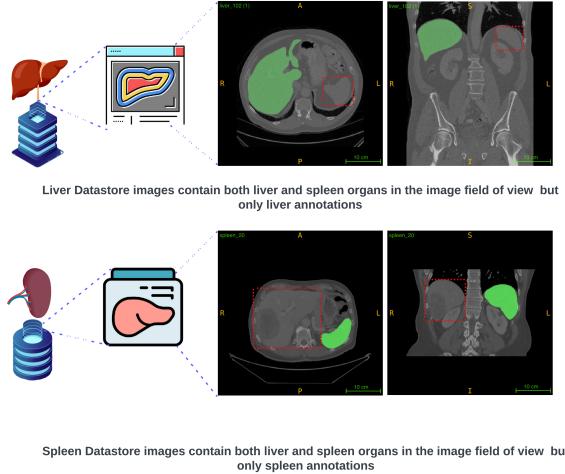


Fig. 1. Illustration of an example federated learning setup with nodes containing datasets with a similar field of view but different and incomplete annotations.

this strategy requires very heavy computational resources. Another approach as described by [10,4] is to design a multi-task head with a common encoder and task-specific decoders that are trained separately. Similarly, the work of [22,8] has shown promise in developing multi-task segmentation models using multi-scale feature abstraction. However, these approaches require all the data to be hosted locally and is not realistic in a medical scenario not only because of privacy and data sharing restrictions but also because it is impossible to anticipate in advance how many distinct activities should the model be trained for.

In [2], the authors developed a multi-task multi-domain deep segmentation model for the segmentation of pediatric imaging datasets with excellent performance. However, the proposed technique was developed and evaluated for different anatomical regions in the body with no overlapping field of view or incomplete annotations. Similarly, the cross-domain medical image segmentation technique developed in [16] was focused on segmentation of the same anatomical structure and the proposed technique was not developed to tackle incomplete annotations.

The work of [20] introduced a real FL setup for segmentation using partial labels where client nodes were trained on specific sub-networks for their specific tasks using a shared decoder. However, this method is not scalable and again, needs knowledge of all the tasks to be trained. It was for the first time in the work of [18] that knowledge aggregation was introduced using a single network in a federated manner. The global federated learning framework developed in their work, however, failed to accurately segment different anatomical structures on

the external test set. For optimal performance, the authors used an ensemble of multiple local federated learning models, making it computationally expensive and practically challenging.

Therefore, we developed SegViz to address the shortcomings of current techniques in efficiently aggregating knowledge from heterogeneous datasets with partial annotations. Our method does not rely on any heavy model or specific feature engineering methods and utilizes the intrinsic similarities between the different imaging datasets to learn a general representation across multiple tasks. Moreover, it does not require knowledge of all the tasks in the participating datasets and is able to tackle domain shift between these datasets.

3 Methods

We developed SegViz as a multi-task federated learning framework to learn a diverse set of tasks from distributed nodes with incomplete annotations, as illustrated in Figure 2. The global SegViz model is initialized at the server with two distinct blocks - a representation block and a task block. The goal of the representation block is to learn a generalized representation of the underlying dataset while the goal of the task block is to learn individual tasks distributed across different nodes. Every client is initialized with a subset of the SegViz model, comprising the representation block and a subset of the task block representing the client’s tasks. During training, the weights of the representation block are always aggregated by the server and redistributed back to the client nodes. On the other hand, the weights of the task block are directly copied from the corresponding client nodes containing the corresponding task, thereby preserving the task-related information for each node in their task block.

3.1 SegViz model architecture

The backbone of the SegViz model architecture was constructed using a modified version of the multi-head 3D-UNet [7] configuration for all our experiments. Each U-Net has 5 layers with down/up-sampling at each layer by a factor of 2. Unlike how U-Net implementations typically operate, these down or up-sampling operations happen at the beginning of each block instead of at the end. The U-Net also contains 2 convolutional residual units at the layers and uses Batch Normalization at each layer. The task block comprised a multi-head architecture with each head consisting of two layers, including the final classification layer. The SegViz model was implemented using the MONAI [3] framework and the pre-processing and training were done using Pytorch. The SegViz model architecture has been illustrated in the supplementary material.

During training, all weights are initialized using LeCun initialization. The batch size was set to 2 and the learning rate was initially set to 1e-4 with the Adam optimizer and CosineAnnealingLR [15] as the scheduler. The Dice Loss was used as the loss function. The average Dice Score was chosen as the final evaluation metric. Each model was trained for a total of 500 epochs.

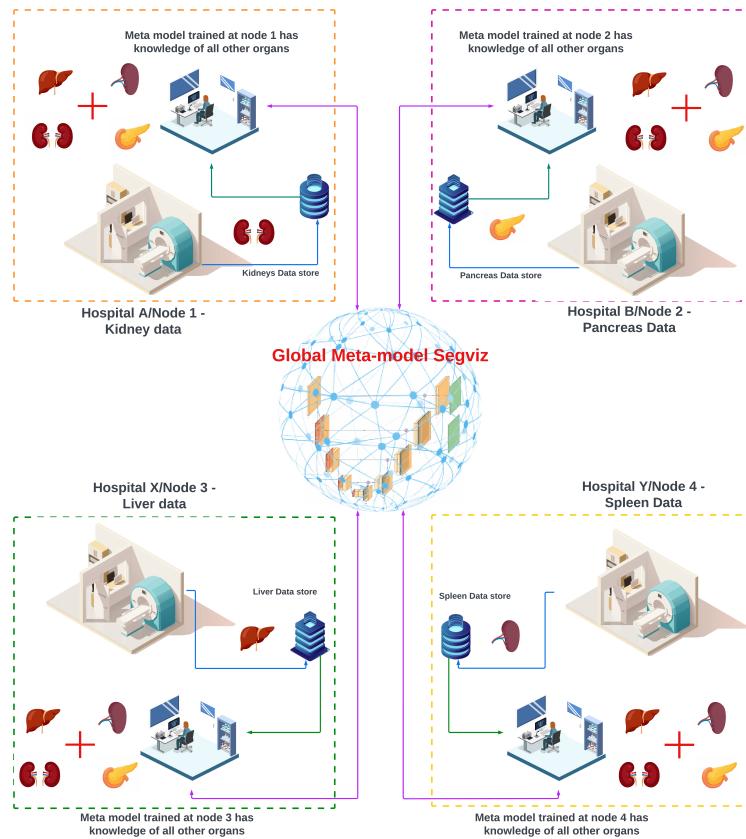


Fig. 2. Illustration of the proposed SegViz framework: Client nodes update the global meta-model where knowledge aggregation occurs after every 10 iterations of the local model. The weights of the global model are then shared with the client models allowing both nodes to share knowledge without sharing data.

3.2 Data

The SegViz framework was evaluated using four publicly available datasets from the Medical Segmentation Decathlon (MSD) challenge [1]. The Spleen MSD dataset consists of 61 3D Computed tomographies (CT) volumes with spleen annotations out of which only the 41 training set volumes were used. The Liver MSD dataset consists of 201 3D CT volumes with liver and liver tumor annotations out of which only 131 training set volumes were considered. Similarly, the Pancreas MSD dataset consists of 420 3D CT volumes of which only 282 training volumes were used. Lastly, from the 2019 Kidney Tumor Segmentation Challenge dataset [9], we used 210 3D CT volumes from the training dataset. For this study, all tumor annotations were discarded and only organ annotations were used. The training and internal validation splits were considered from the overall training data in an 80:20 split. We considered all 30 training image volumes from the Beyond the Cranial Vault (BTCV) dataset [13] as an external test set for all our experiments.

During pre-processing, all the image volumes were first resized to $256 \times 256 \times 128$, and the intensity values normalized between 0 and 1. All the volumes were resampled to a constant spacing of (1.5, 1.5, 2.0). We extract random foreground patches of size $128 \times 128 \times 32$ from each volume such that the center voxel of each patch belonged to either the foreground or background class.

4 Experiments

4.1 Individual baseline implementation

For every task, we trained a single U-Net model based on the Segviz model architecture on the training dataset after the 80:20 split. Hence we had a single model trained on the training dataset for the liver, spleen, pancreas, and kidneys.

4.2 Central aggregation implementation

As a lower bound for a multi-task segmentation setup, we combine all four datasets together to create a central repository of all the data. We consider this a lower bound because naive aggregation of the data in the case of partial annotations would lead to suboptimal performance compared to an individual model trained for each dataset separately. We setup our centrally aggregated model using the same steps as our baseline implementation.

4.3 SegViz implementation

FedAvg: We use the popular FedAvg algorithm to construct an FL setup where each client node in the setup represents an isolated group having one of the datasets. The same UNet configuration with the Segviz architecture was used at each client. Apart from the same pre-processing steps as the baseline implementation, we also added random affine transformations such as rotation and

scaling. While training the local models, after every 10 epochs, following the FedAvg algorithm, the global model gets all but the last two convolutional layers’ weights and averages them. The updated weights are then shared back to all the local models.

FedBN: We investigate the popular FedBN algorithm in a similar setup as our FedAvg implementation. Making sure that our global model is generalizable to non-i.i.d data is especially important in medical imaging as data from different centers is obtained using different scanners/protocols. FedBN has shown to be successful compared to other FL algorithms such as FedAvg and FedProx in creating a global model that is generalizable well to non-i.i.d data and it does so by not aggregating the batch norm layers during knowledge transfer.

Local fine-tuning In [12,14], the authors demonstrated the need for finetuning in FL models in order to reduce the effect of catastrophic forgetting and stabilize personalized performance. We also finetuned our FedAvg and FedBN models (keeping the representation block frozen) on the local datasets to improve task-specific performance of each task block while keeping the same representation block.

Table 1. Mean Dice score performance of all the experiments on the in-federation validation dataset. The standard deviation values are in parentheses. After a paired t-test, the entries underlined are significant against baseline, in bold against central agg, in italics against our best model (FedBN + FT)

| Models | Mean Dice (SD) | | | |
|-------------------|-----------------------|--------------------|--------------------|--------------------|
| | Liver | Spleen | Pancreas | Kidneys |
| Baseline Liver | 0.94 (0.02) | - | - | - |
| Baseline Spleen | - | 0.94 (0.02) | - | - |
| Baseline Pancreas | - | - | 0.70 (0.15) | - |
| Baseline Kidneys | - | - | - | 0.83 (0.16) |
| Centrally agg | <u>0.27 (0.17)</u> | <u>0</u> | <u>0.64 (0.17)</u> | <u>0.52 (0.27)</u> |
| FedAvg | 0.94 (0.03) | 0.93 (0.02) | <u>0.64 (0.18)</u> | <u>0.86 (0.14)</u> |
| FedAvg + FT | 0.94 (0.03) | 0.92 (0.03) | <u>0.66 (0.17)</u> | 0.85 (0.15) |
| FedBN | <u>0.91 (0.05)</u> | 0.94 (0.02) | 0.69 (0.16) | 0.84 (0.15) |
| FedBN + FT | 0.93 (0.03) | 0.94 (0.02) | 0.69 (0.16) | 0.83 (0.16) |

5 Results

As shown in Figure 3, the FedBN model with fine-tuning performs the best on the in-federation internal validation set as well the out-of-federation BTCV test set. The SegViz framework using FedBN with fine-tuning segmented the BTCV test set with dice scores of 0.93, 0.83, 0.55, and 0.75 for segmentation of liver, spleen, pancreas, and kidneys, respectively, significantly ($p < 0.05$) better (except spleen) than the dice scores of 0.87, 0.83, 0.42, and 0.48 for the baseline models. In contrast, the central aggregation model performed significantly ($p < 0.05$) poorly on the test dataset with dice scores of 0.65, 0, 0.55, and 0.68. We note that the model trained on the centrally aggregated data did not generalize to the spleen label due to the overall model becoming biased toward the liver and pancreas labels, which contain more samples per label. We have included the statistical t-test results between the baseline and the best-performing models in Table 1 and Table 2.

Table 2. Mean Dice score performance of all the experiments on the out-of-federation BTCV dataset. The standard deviation values are in parentheses. The same conventions as Table 1 are followed

| Models | Mean Dice (SD) | | | |
|-------------------|---------------------------|---------------------------|--------------------|--------------------|
| | Liver | Spleen | Pancreas | Kidneys |
| Baseline Liver | 0.87 (0.11) | - | - | - |
| Baseline Spleen | - | 0.83 (0.14) | - | - |
| Baseline Pancreas | - | - | 0.42 (0.24) | - |
| Baseline Kidneys | - | - | - | 0.48 (0.30) |
| Centrally agg | <u>0.65 (0.14)</u> | 0 | <u>0.55 (0.18)</u> | <u>0.68 (0.21)</u> |
| FedAvg | <u>0.93 (0.02)</u> | 0.78 (0.14) | 0.40 (0.20) | 0.78 (0.12) |
| FedAvg + FT | <u>0.93 (0.02)</u> | <u>0.73 (0.15)</u> | 0.44 (0.20) | 0.76 (0.12) |
| FedBN | <u>0.93 (0.01)</u> | 0.83 (0.14) | <u>0.55 (0.17)</u> | <u>0.74 (0.11)</u> |
| FedBN + FT | 0.93 (0.01) | 0.83 (0.15) | <u>0.55(0.17)</u> | 0.75 (0.10) |

6 Discussion

The SegViz framework proposed in this work demonstrated excellent performance in aggregating knowledge from heterogeneous datasets with different, incomplete labels. Our approach successfully aggregated knowledge from all nodes

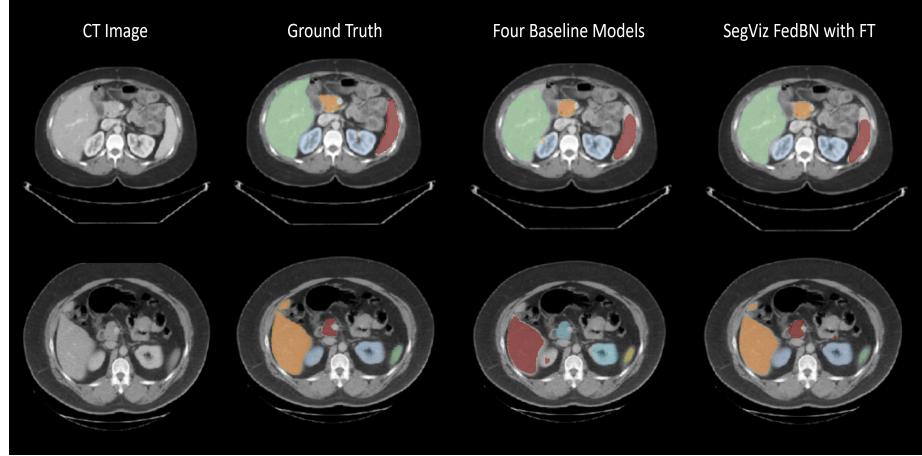


Fig. 3. A comparison of the ground truth segmentation masks with the masks generated by the baseline and SegViz models.

with little to no drop in the performance of the global meta-model in terms of the average dice score. The comparable performance between the SegViz segmentations and multiple baseline model segmentation illustrates a preliminary example of constructing a single global multi-task segmentation model with clinical applicability from dispersed datasets with disjoint partial annotations. It is important to note that the FedAvg global model can be extended to contain a multi-head classifier block while this is not true for the FedBN model.

Image segmentation from heterogeneous datasets with incomplete annotations has many potential benefits. For example, SegViz can potentially reduce labeling time by $1/\eta$ where η is the number of distinct labels in the distributed data sets by allowing the transfer of knowledge between each client. This would not only save time but also allow different research groups to potentially benefit from each others' annotations without explicitly sharing them.

We believe the success of SegViz is attributed to several inherent advantages in its implementations, such as using a learning rate decay and random affine transformations during training which makes it more robust to non-i.i.d data. Moreover, extending our FL implementations with fine-tuning allows for creating stable, high-performing personalized local models. In the future, we would like to extend our experiments using a modality than is less stable than CT such as MRI. We would also like to investigate the real-world performance of our FL setup where client nodes can join and drop contact with the server at any point in time while maintaining no drop in performance.

References

1. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., et al.: The medical

- segmentation decathlon. *Nature communications* **13**(1), 1–13 (2022)
2. Boutillon, A., Conze, P.H., Pons, C., Burdin, V., Borotikar, B.: Generalizable multi-task, multi-domain deep segmentation of sparse pediatric imaging datasets via multi-scale contrastive regularization and multi-joint anatomical priors. *Medical Image Analysis* **81**, 102556 (2022)
 3. Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., et al.: Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701* (2022)
 4. Chen, S., Ma, K., Zheng, Y.: Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625* (2019)
 5. Chen, X., Sun, S., Bai, N., Han, K., Liu, Q., Yao, S., Tang, H., Zhang, C., Lu, Z., Huang, Q., et al.: A deep learning-based auto-segmentation system for organs-at-risk on whole-body computed tomography images for radiation therapy. *Radiotherapy and Oncology* **160**, 175–184 (2021)
 6. Chowdhury, A., Kassem, H., Padoy, N., Umeton, R., Karargyris, A.: A review of medical federated learning: Applications in oncology and cancer research. In: *International MICCAI Brainlesion Workshop*. pp. 3–24. Springer (2022)
 7. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: *International conference on medical image computing and computer-assisted intervention*. pp. 424–432. Springer (2016)
 8. Fang, X., Yan, P.: Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction. *IEEE Transactions on Medical Imaging* **39**(11), 3619–3629 (2020)
 9. Heller, N., Sathianathan, N., Kalapara, A., Walczak, E., Moore, K., Kaluzniak, H., Rosenberg, J., Blake, P., Rengel, Z., Oestreich, M., 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)
 10. Huang, R., Zheng, Y., Hu, Z., Zhang, S., Li, H.: Multi-organ segmentation via co-training weight-averaged models from few-organ datasets. In: *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV* 23. pp. 146–155. Springer (2020)
 11. Isensee, F., Jäger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: Automated design of deep learning methods for biomedical image segmentation. *arXiv preprint arXiv:1904.08128* (2019)
 12. Jiang, Y., Konečný, J., Rush, K., Kannan, S.: Improving federated learning personalization via model agnostic meta learning. *arXiv preprint arXiv:1909.12488* (2019)
 13. Landman, B., Xu, Z., Iglesias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault—workshop and challenge. In: *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*. vol. 5, p. 12 (2015)
 14. Liang, P.P., Liu, T., Ziyin, L., Allen, N.B., Auerbach, R.P., Brent, D., Salakhutdinov, R., Morency, L.P.: Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523* (2020)
 15. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983* (2016)
 16. Parekh, V.S., Lai, S., Braverman, V., Leal, J., Rowe, S., Pillai, J.J., Jacobs, M.A.: Cross-domain federated learning in medical imaging. *arXiv preprint arXiv:2112.10001* (2021)

17. Shen, C., Wang, P., Roth, H.R., Yang, D., Xu, D., Oda, M., Wang, W., Fuh, C.S., Chen, P.T., Liu, K.L., et al.: Multi-task federated learning for heterogeneous pancreas segmentation. In: Clinical Image-Based Procedures, Distributed and Collaborative Learning, Artificial Intelligence for Combating COVID-19 and Secure and Privacy-Preserving Machine Learning, pp. 101–110. Springer (2021)
18. Shen, C., Wang, P., Yang, D., Xu, D., Oda, M., Chen, P.T., Liu, K.L., Liao, W.C., Fuh, C.S., Mori, K., et al.: Joint multi organ and tumor segmentation from partial labels using federated learning. In: International Workshop on Distributed, Collaborative, and Federated Learning, Workshop on Affordable Healthcare and AI for Resource Diverse Global Health. pp. 58–67. Springer (2022)
19. 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. *Medical Image Analysis* **63**, 101693 (2020)
20. Xu, X., Yan, P.: Federated multi-organ segmentation with partially labeled data. arXiv preprint arXiv:2206.07156 (2022)
21. Yu, Q., Shi, Y., Sun, J., Gao, Y., Zhu, J., Dai, Y.: Crossbar-net: A novel convolutional neural network for kidney tumor segmentation in ct images. *IEEE transactions on image processing* **28**(8), 4060–4074 (2019)
22. 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)