# Joint Multi Organ and Tumor Segmentation from Partial Labels Using Federated Learning

Chen Shen[1], Pochuan Wang[2], Dong Yang[3], Daguang Xu[3], Masahiro Oda[1], Po-Ting Chen[4], Kao-Lang Liu[4], Wei-Chih Liao[4], Chiou-Shann Fuh[2], Kensaku Mori[1], Weichung Wang[2](✉), and Holger R. Roth[3]

[1] Nagoya University, Nagoya, Japan
[2] National Taiwan University, Taipei, Taiwan
wwang@math.ntu.edu.tw
[3] NVIDIA Corporation, Santa Clara, USA
[4] National Taiwan University Hospital, Taipei, Taiwan

**Abstract.** Segmentation studies in medical image analysis are always associated with a particular task scenario. However, building datasets to train models to segment multiple types of organs and pathologies is challenging. For example, a dataset annotated for the pancreas and pancreatic tumors will result in a model that cannot segment other organs, like the liver and spleen, visible in the same abdominal computed tomography image. The lack of a well-annotated dataset is one limitation resulting in a lack of universal segmentation models. Federated learning (FL) is ideally suited for addressing this issue in the real-world context. In this work, we show that each medical center can use training data for distinct tasks to collaboratively build more generalizable segmentation models for multiple segmentation tasks without the requirement to centralize datasets in one place. The main challenge of this research is the heterogeneity of training data from various institutions and segmentation tasks. In this paper, we propose a multi-task segmentation framework using FL to learn segmentation models using several independent datasets with different annotations of organs or tumors. We include experiments on four publicly available single-task datasets, including MSD liver (w/ tumor), MSD spleen, MSD pancreas (w/ tumor), and KITS19. Experimental results on an external validation set to highlight the advantages of employing FL in multi-task organ and tumor segmentation.

**Keywords:** Federated learning · Segmentation · Partial labels

## 1 Introduction

Fully automated segmentation of organs and tumors from computed tomography (CT) volumes is essential for medical image analysis. Numerous studies have

C. Shen and P. Wang—Equal contribution.

S. Albarqouni et al. (Eds.): DeCaF 2022/FAIR 2022, LNCS 13573, pp. 58–67, 2022.
https://doi.org/10.1007/978-3-031-18523-6_6

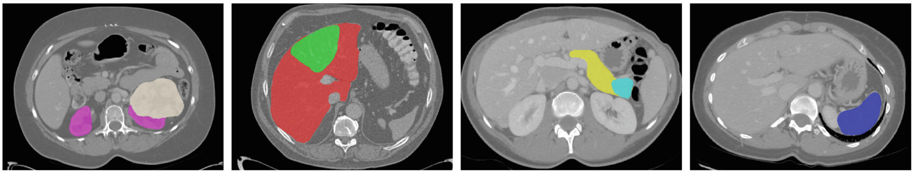
concentrated on single specialized task segmentation throughout the last few decades [1,4,15,17]. For instance, the pancreas regions and pancreatic tumors will be included in the annotations if we want to develop an automated segmentation model for pancreatic cancer. However, this model cannot segment other organs and pathologies, like the liver and liver tumors. In a real-world clinical scenario, a generalized segmentation model for various organ types and associated malignancies is desired to develop comprehensive computer-aided diagnostic (CAD) systems.

The main challenge for achieving such generalized models is the lack of substantial datasets for multi-task organ segmentation. Most datasets are solely intended for a few very specialized segmentation tasks [5,18]. It is also tough to get annotated datasets for multi-task scenarios from multiple institutions to cover a large and diverse patient population and different scanner types and acquisition protocols. To simultaneously annotate various organ and tumor types demands extensive medical expertise as well as time.

In order to address these issues, several studies have attempted to build a generalized segmentation using multiple partially annotated datasets [2,6,21]. However, they centralized all training datasets locally. In real-world clinical situations, sharing the datasets among different institutions presents numerous technological, legal, and privacy concerns and might be therefore infeasible.

Federated learning (FL) is inherently suited for solving this problem [10,13]. Recently, combining FL methods with other deep learning techniques has grown in favor. A rising number of studies have been conducted using the FL method in segmentation tasks in the medical field. Li et al. [9] applied FL to brain tumor segmentation in practical for preserving data privacy. Wang et al. [19] carried out the real-world pancreas and pancreatic tumor segmentation using FL between two institutions across different nations. This work shows that FL considerably enhances the model performance of organ and tumor segmentation when compared to local standalone training. Additionally, recent real-world studies have shown that the FL approach is beneficial in many applications such as brain tumor segmentation [16], mammography classification [14], and COVID-19 prediction [3,11]. However, the main goal of these studies is to enhance the effectiveness of a single particular task. Some studies proposed to handle the multiple datasets using FL for classification task [8], but research on segmentation models for medical imaging is lacking.

In this work, we suggested a multi-task segmentation framework that makes use of FL to increase the generalizability of segmentation models using several partially annotated datasets. We explored the efficacy of the FedAvg model aggregation approach across several different segmentation tasks. We employed the MSD liver (w/ tumor), MSD spleen, MSD pancreas (w/ tumor) [18], and KITS19 [5] datasets, which are publicly accessible for single-task segmentation. Examples of axial CT slices of four partial labeled datasets for different segmentation tasks are shown in Fig. 1. Experimental results revealed that the FL framework boosted the segmentation performance of jointly trained task-specific models. Additionally, we evaluated our models on an unseen external dataset, and the segmentation results were satisfactory. To our knowledge, this is the



(a) KITS19 (b) MSD liver (c) MSD pancreas (d) MSD spleen

**Fig. 1.** Samples of (a) KITS19 (b) MSD liver (c) MSD pancreas and (d) MSD spleen. The kidney and kidney tumor are pink and brown, respectively; the liver and liver tumor are red and green, respectively; the pancreas and pancreatic tumor are yellow and aquamarine, respectively; and the spleen is blue. (Color figure online)

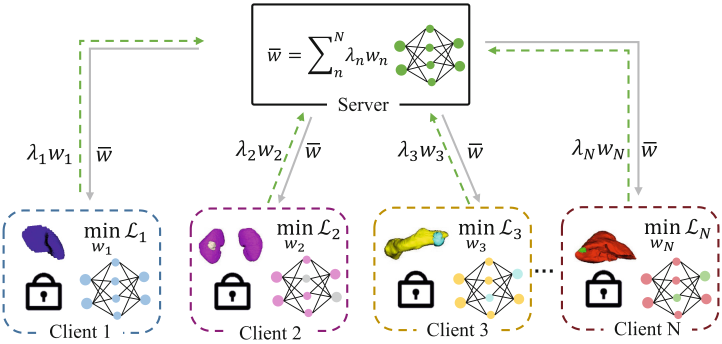


**Fig. 2.** An overview of the federated learning framework for multi-task medical image segmentation from partial labels. The training model is shared by each client for different segmentation tasks, and the model is aggregated by the server.

first work on multi-organ and tumor segmentation from partial labels for medical imaging using FL.

## 2 Methods

### 2.1 Federated Learning

FL [10] is a prominent distributed learning technique applied in many fields. In the area of medical image analysis, there is growing interest in FL techniques [13, 19, 20]. The key advantage of FL is that it can learn from various datasets without the necessity for centralizing all datasets locally. An FL framework consists of a server and several clients. The server manages the whole FL process, and each client tackles their own task independently. The models are trained on the clients using local datasets, and they only exchange the learned parameters with the server; the server does not possess any data. The server only aggregates the model after receiving new parameters from a minimum number of clients specified and then sends an updated global model back to each client. In a new FL round, each client receives the global model from the server and refines it using their local dataset. An overview of our federated learning framework is shown in Fig. 2. In this study, each client trains on a different segmentation task.

### 2.2 Federated Averaging for Learning from Partial Labels

Federated averaging (FedAvg) is an effective aggregation method widely used in FL [10]. In FedAvg, the server aggregates the parameters shared by clients after each client trains trains on their local data using gradient-based optimization. In each round of FL, the following objective is being optimized:

$$\mathcal{L} = min \sum_{k}^{N} \eta_k \mathcal{L}_k, \tag{1}$$

where $\mathcal{L}_k$ represents the $k$-th client's local loss function out of the $N$ clients. Each client trainers to optimize its $\mathcal{L}_k$ independently. The weight of each client is denoted as $\eta_k$, and the total weight of all clients in a round equals to 1. Using all clients for stochastic gradient updates in real-world FedAvg usage is costly in terms of both time and communication. At each round, a subset of $N$ clients, which can be represented as $K$, can be selected for server model updates. Each client's weight is determined by the percentage of training data, with a total of $n$. We have $n = \sum_{k}^{K} n_k$, where $n_k$ is the number of training data in client $k$. The weight of client $\eta_k$ is $\eta_k = \frac{n_k}{n}$. When updating the global model using FedAvg, the client with more training data contributes more during the aggregation.

To avoid conflict of background labels between each client, we use sigmoid as the output activation function, and $\mathcal{L}_k$ for each client is the average Dice of all output channels except the background.

$$\mathcal{L}_k = \frac{1}{C} \sum_{c=2}^{C} \mathcal{L}_{Dice}(\mathcal{F}(x)_c, y_c) \tag{2}$$

where $\mathcal{F}$ is the model and $C$ is the number of total classes, in this work $C = 8$, the corresponding organ for indices 1 to 8 are background, liver, liver tumor, spleen, pancreas, pancreas tumor, kidney, kidney tumor.

## 3 Experimental Details and Results

### 3.1 Datasets

Four publicly accessible datasets were used in this experiment. The server only collects the model parameters provided by the client and does not possess any data. Each client trained the model with a single dataset among distinct segmentation tasks. There are four types of different segmentation tasks, including the segmentation of the liver and liver tumor (**Task 1**); the pancreas and pancreatic tumor (**Task 2**); the kidney and kidney tumor (**Task 3**); the spleen (**Task 4**).

**Table 1.** Number of images used in the experiments. Each dataset was randomly divided into training, validation, and testing sets in the equal amounts.

| | Training | Validation | Testing | Total |
|---|---|---|---|---|
| **Task 1** (MSD liver) | 79 | 26 | 26 | 131 |
| **Task 2** (MSD pancreas) | 169 | 56 | 56 | 281 |
| **Task 3** (Kits19) | 126 | 42 | 42 | 210 |
| **Task 4** (MSD spleen) | 25 | 8 | 8 | 41 |

The dataset for **Task 1** is from Medical Segmentation Decathlon (MSD) liver task [18]. We only kept the 131 training cases with liver and liver tumor labels here. All the volumes are contrast-enhanced and the resolutions of volumes are (0.5–1.0, 0.5–1.0, 0.45–6.0) mm. For **Task 2**, we used 281 MSD pancreas task cases collected from patients undergoing pancreatic mass resection [18]. These CT volumes are in portal-venous phases. For **Task 3**, we utilized 210 cases from the KITS19 Challenge (Kits19) [5]. The CT volumes are in the late-arterial phase. For **Task 4**, 41 cases of portal venous phase CT from the MSD spleen task were used [18]. We randomly split the datasets into training, validation, and testing sets in the proportions of 60%, 20%, and 20%, respectively. The details of data divisions are shown in Table 1.

We employ another open dataset, MICCAI Multi-Atlas Labeling Beyond the Cranial Vault challenge (BTCV) [7], as an external validation dataset. This dataset contains 30 portal venous phase CT images with segmentation mask of 13 abdomen organs. We only kept the liver, pancreas, kidney, spleen segmentation mask in our testing as they overlap with the partial labels from tasks used during training.

### 3.2 Implementation Details

We use NVIDIA Federated Learning Application Runtime Environment (*NVIDIA FLARE*)[1] [12] as the backend of the FL framework. Our implementation is base on PyTorch Lighting[2]. We use a single NVIDIA GPU (Tesla V100 with 32GB) for each client in all experiments. We resampled all the volumes to $1 \times 1 \times 1mm^3$ isotropic spacing to guarantee the CT volumes had the same resolution. The intensity of the Hounsfield unit (HU) was clipped to the range [–500, 500] and normalized to [0, 1], which encompasses most of the abdominal organs. A random intensity shift augmentation was applied on training volumes with a probability of 0.8. The offset factor of it is 0.1 under the MONAI implementation[3]. We reset the orientation close to RAS+ so that all the CT volumes are in the same orientation. The input size of our model is 96 × 96 × 96 with a batch size of 8.

[1] https://nvidia.github.io/NVFlare/.

[2] https://www.pytorchlightning.ai/.

[3] https://monai.io/.

**Table 2.** Comparison of Dice Score for the four different segmentation tasks on *standalone model* trained from scratch on single dataset; on *FL global best model* on server side; and *FL local best model* on client side.

| Dice (%) | Task 1 | | Task 2 | | Task 3 | | Task 4 | Avg. |
|---|---|---|---|---|---|---|---|---|
| | Liver | Tumor | Pancreas | Tumor | Kidney | Tumor | Spleen | |
| SL | 67.4% | 13.3% | 64.2% | 19.6% | 93.1% | 30.2% | 48.7% | 63.5% |
| FL global model (server) | 0.0% | 0.0% | 66.9% | 30.4% | 90.6% | 39.3% | 0.0% | 44.6% |
| FL local model (client) | **84.4%** | **33.8%** | **74.3%** | **38.4%** | **94.9%** | **63.7%** | **72.9%** | **86.7%** |

The network architecture we utilized was obtained from the coarse-to-fine network architecture search (C2FNAS), which already demonstrates strong generalizability on organs and tumors segmentation in multiple different medical image segmentation tasks [20]. We training using the loss formulated in Eq. 2. For validation and testing we threshold the output values by 0.5 and calculate the Dice score for each channel separately. Clients train locally in the FL and communicate the learned parameters to the server at every 500 iterations. A total of 60 rounds were completed on the server, and the minimum client number to aggregate the model is 4.

### 3.3 Experimental Results

Our experimental results include the standalone training model (SL) trained with each partially labeled dataset, the FL client model on the local client, and the FL global best model aggregated by FedAvg on the server.

Table 2 compares the Dice score on four different segmentation tasks with standalone training (SL) models, FL global model on the server, and FL local model on each client. Comparing FL local models to SL models, the average Dice score for each client increases by 23.3%. The highest improvement, which is 33.5%, is in the segmentation of kidney tumors. The Dice score of the FL global model is not ideal. The FL global model fails on the liver, liver tumor, and spleen segmentation.

We present the axial visualizations of the four segmentation tasks in Fig. 4. The segmentation of organs and tumors performs best when using the local client model for specialized to each task during FL.

### 3.4 Validation on External Dataset

To verify the generalizability of our FL model for multi-organ segmentation, we validate the ensemble of the local models on the BTCV [7], which is a completely unseen dataset in this work. Table 3 shows our evaluation Dice scores for the spleen, kidney, pancreas, and liver organs. We compared the results of the FL global model, the ensembled local models and the ensembled results with extra post processing. The ensemble method we used is to combine corresponding output channels of the four local models. Since the final activation is sigmoid, the

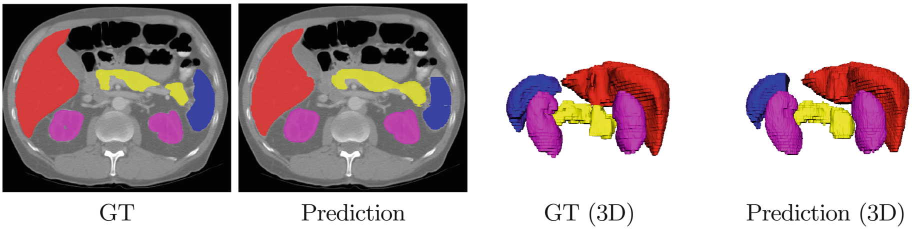


**Fig. 3.** The segmentation visualization of an external dataset in axial slice and 3D rendering using FL local model with post-processing. On this entirely new dataset, the major organ segmentation performs satisfactorily.

**Table 3.** Testing results on external dataset of the ensemble of four FL local models and FL global model.

| Dice (%) | Task 1 | Task 2 | Task 3 | Task 4 | Avg. |
|---|---|---|---|---|---|
| | Liver | Pancreas | Kidney | Spleen | |
| SL | 65.8% | 57.5% | 70.7% | 44.6% | 60.0% |
| FL global model | 0.0% | 62.9% | 54.5% | 0.0% | 29.4% |
| FL local model (ensemble) | 87.7% | 71.8% | **80.5%** | 69.9% | 77.5% |
| FL local model (processed) | **91.6%** | **75.4%** | 80.2% | **73.1%** | **80.1%** |

output of the global model and the ensembled model may overlap. To overcome the overlapping issue we first take the largest connected components from each channel and discard any smaller objects. Then we fuse the output channels from corresponding models in the order of liver, liver tumor, spleen, pancreas, pancreas tumor, kidney and kidney tumor. Note, the later label might override a former label but we did not notice this to be problematic. The visualization of post processed results are presented in Table 3.

## 4 Discussion

As seen in Table 2, federated learning considerably improved the segmentation performance on the local model of each client compared to the results of standalone training. The average Dice score of the four segmentation tasks increased by 23.3%. By employing the task-specialized FL local models, there is a noticeable improvement in the segmentation of both tumors and organs. Although the training data on each client only contains annotations for one of the different segmentation tasks, the learned global parameters are beneficial for all other segmentation tasks. The FL local models were improved by adjusting the parameters obtained from the server to fit the particular segmentation task for the local dataset. However, the heterogeneity between different client datasets and annotation tasks causes the averaged global to not perform well, especially on **Task 1** and **Task 3** (liver and spleen).

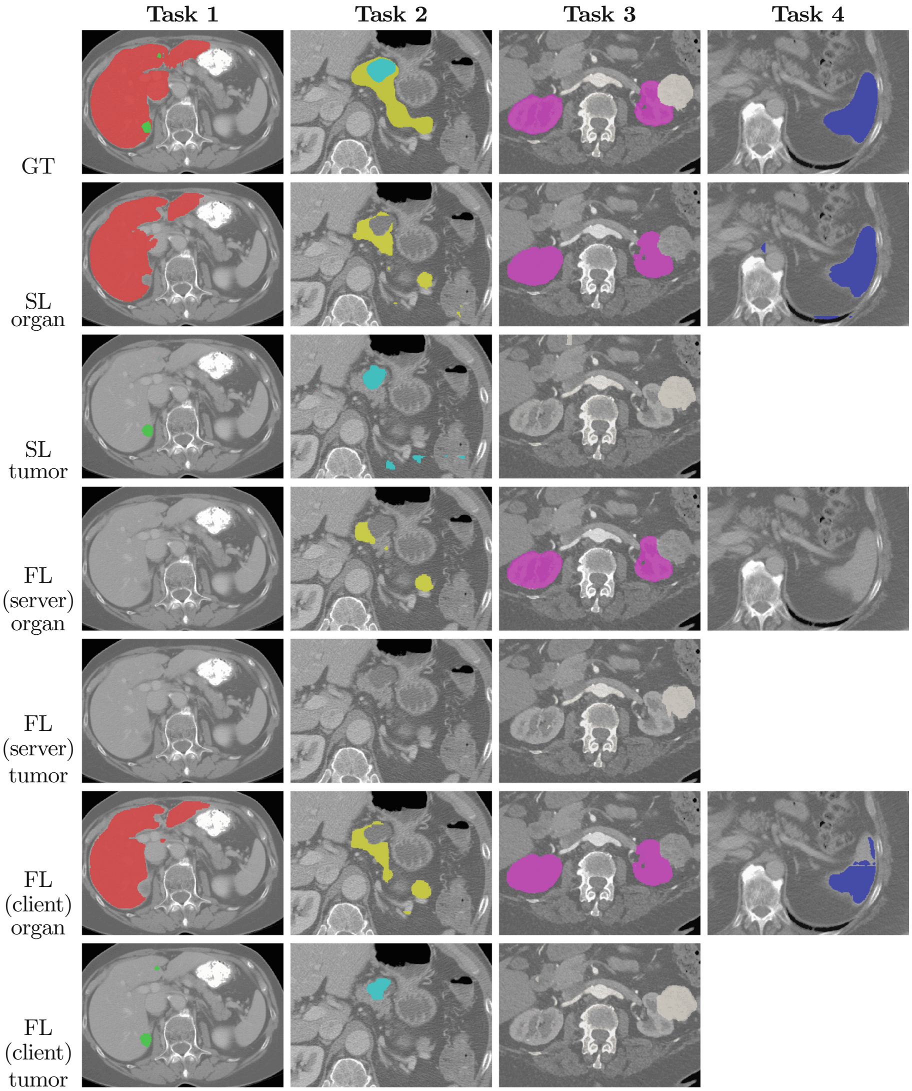


**Fig. 4.** Examples of segmentation results of four tasks including liver and tumor (**Task 1**); pancreas and tumor (**Task 2**); kidney and tumor (**Task 3**); spleen (**Task 4**) on ground truth (GT), standalone model (SL), and FL model on server-site and client-site.

Nevertheless, we validate the ensemble of local segmentation models on an external dataset. Both qualitative and quantitative evaluation results demonstrate the robustness of our FL local models. The segmentation performance of the ensemble models on the corresponding organs is satisfactory and shows how a successful ensemble model can be trained using FL with only partially annotated datasets.

## 5 Conclusion

In this study, we apply the FL techniques for multi-task organs and tumors segmentation. The experimental results suggest that FL has a favorable impact on the segmentation of organs and tumors, although the datasets on other clients are dissimilar. The FedAvg is not well-suited to address the heterogeneous problems of multi-task datasets. Hence the FL global model underperformed. We confirmed the robustness of the local model ensemble with external validation using an unseen dataset. Future work is required to address the heterogeneity challenge in multi-task organs and tumor segmentation from partially labeled datasets.

## 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). https://doi.org/10.1016/j.neucom.2021.08.157, https://www.sciencedirect.com/science/article/pii/S0925231222003149
2. Chen, S., Ma, K., Zheng, Y.: Med3D: transfer learning for 3d medical image analysis. arXiv preprint arXiv:1904.00625 (2019)
3. Dayan, I., et al.: Federated learning for predicting clinical outcomes in patients with Covid-19. Nat. Med. **27**(10), 1735–1743 (2021)
4. Heller, N., et al.: The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KITS19 challenge. Med. Image Anal. **67**, 101821 (2021). https://doi.org/10.1016/j.media.2020.101821
5. Heller, N., et al.: The KITS19 challenge data: 300 kidney tumor cases with clinical context, CT semantic segmentations, and surgical outcomes (2019). https://doi.org/10.48550/ARXIV.1904.00445, https://arxiv.org/abs/1904.00445
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. Landman, B., et al.: 2015 MICCAI multi-atlas labeling beyond the cranial vault - workshop and challenge (2015). https://doi.org/10.7303/syn3193805
8. Li, Q., He, B., Song, D.: Model-contrastive federated learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10713–10722 (2021)
9. Li, W., et al.: Privacy-preserving federated brain tumour segmentation. In: Suk, H.I., Liu, M., Yan, P., Lian, C. (eds.) Machine Learning in Medical Imaging, pp. 133–141. Springer, Cham (2019)
10. 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)
11. Nguyen, D.C., Ding, M., Pathirana, P.N., Seneviratne, A., Zomaya, A.Y.: Federated learning for covid-19 detection with generative adversarial networks in edge cloud computing. IEEE Internet of Things J. **9**, 10257–10271 (2021)
12. Nvidia Corporation: Nvidia FLARE, June 2022. https://doi.org/10.5281/zenodo.6780567,https://github.com/NVIDIA/nvflare

13. Rieke, N., et al.: The future of digital health with federated learning. NPJ Digit. Med. **3**(1), 1–7 (2020)
14. Roth, H.R., et al.: Federated learning for breast density classification: a real-world implementation. In: Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning, pp. 181–191. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-60548-3
15. Seo, H., Huang, C., Bassenne, M., Xiao, R., Xing, L.: Modified U-Net (mU-Net) with incorporation of object-dependent high level features for improved liver and liver-tumor segmentation in CT images. IEEE Trans. Med. Imaging **39**(5), 1316–1325 (2020). https://doi.org/10.1109/TMI.2019.2948320
16. Sheller, M.J., et al.: Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci. Rep. **10**(1), 1–12 (2020)
17. Shen, C.: Multi-task federated learning for heterogeneous pancreas segmentation. In: Oyarzun Laura, C., et al. (eds.) DCL/PPML/LL-COVID19/CLIP -2021. LNCS, vol. 12969, pp. 101–110. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-90874-4_10
18. Simpson, A.L., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. CoRR abs/1902.09063 (2019). http://arxiv.org/abs/1902.09063
19. Wang, P., et al.: Automated pancreas segmentation using multi-institutional collaborative deep learning. In: Albarqouni, S., et al. (eds.) DART/DCL -2020. LNCS, vol. 12444, pp. 192–200. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-60548-3_19
20. Yu, Q., et al.: C2FNAS: coarse-to-Fine neural architecture search for 3D medical image segmentation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), December 2019
21. 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, June 2021