Privacy-Preserving Collaboration for Multi-Organ Segmentation via Federated Learning from Sites with Partial Labels

Adway Kanhere Pranav Kulkarni Paul H. Yi Vishwa S. Parekh*
University of Maryland Medical Intelligent Imaging (UM2ii) Center
University of Maryland School of Medicine, Baltimore, MD 21201
{akanhere,pkulkarni,pyi,vparekh}@som.umaryland.edu

Abstract

Manual annotation of 3D medical images is expensive and time-consuming, resulting in datasets focused on segmenting individual organs. This leads to training several specialized models that limit clinical translational utility. To that end, we developed SegViz, a federated learning (FL) framework to aggregate knowledge from heterogeneous datasets with partial annotations into a single multi-organ segmentation model. SegViz uses collaborative 3D-U-Nets, with selective weight synchronization across distributed sites, to consolidate knowledge by averaging shared representation weights while isolating task-specific heads during synchronization. SegViz was compared to conventional FL using FedAvg, single-organ baseline models, and a single centralized model trained using data aggregated from all sites. Four partially annotated datasets were used in this study: Spleen MSD, Liver MSD, Pancreas MSD, and the Kidney Tumor Segmentation dataset. All approaches were evaluated using the independent BTCV dataset for segmentation of liver, spleen, pancreas, and kidneys using the dice similarity metric. Extensive experiments across the two-, three- and four-client FL setups with each client holding a dataset with single-organ annotations demonstrated the effectiveness of SegViz for collaborative multi-task segmentation from distributed sites with partial labels. All our implementations and code are available at https://github.com/UM2ii/SegViz.

1. Introduction

Medical image segmentation is a fundamental task in artificial intelligence (AI)-assisted decision support, enabling applications like diagnosis, treatment planning, and assessing therapy response. However, it is reliant on expensive and time-consuming manual annotations from domain experts like radiologists. As a result, deep learning (DL) segmentation models developed in literature are "narrowly" focused only on a subset of structures that are present in a patient based on the research groups' focus, thereby reducing their clinical translational utility and interoperability. For example, a model trained to only segment pneumonia in lung CTs would fail to segment any additional abnormalities that might be present in a patient (e.g., lung tumors). This siloed approach results in hundreds of models that would need to be deployed in the clinical environment. Therefore, there is a critically unmet need to collaboratively train global models by aggregating knowledge from decentralized datasets curated by different research groups focusing on different tasks within the same domain. Knowledge aggregation would not only save time but also allow different groups to benefit from each other's annotations without explicitly sharing them (Figure 1).

Federated Learning (FL) presents an opportunity to aggregate knowledge from datasets curated by different research groups into unified multi-task models in a privacy-preserving manner. However, aggregating knowledge from diverse datasets curated at different imaging centers using FL is challenging as each imaging center may focus on related but different tasks. For example, as shown in Figure 1, each center is focused on segmenting only one of the four organs (liver, spleen, pancreas, kidneys) despite sharing the same field-of-view, thereby missing annotations for the remaining organs.

To that end, we developed SegViz, an FL framework to aggregate knowledge from heterogeneous medical imaging datasets into a single multi-organ segmentation model (Figure 1). SegViz employs a multi-head 3D-U-Net architecture with selective knowledge aggregation to collaboratively learn a shared representation while preserving task-specific knowledge for each class. Using selective knowledge aggregation at the server, SegViz overcomes the limitations of existing techniques by removing the need for every client to work on the same task (e.g., tumor segmentation) or have knowledge of all the tasks in the network. We evaluated SegViz across different FL setups with different

*Corresponding author.

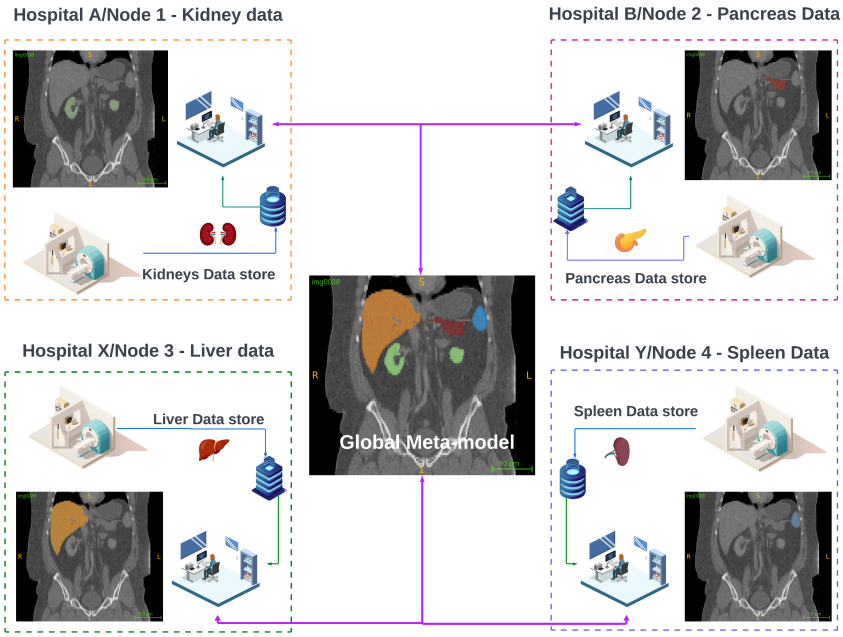

Figure 1. Illustration of SegViz. Suppose there are four research groups, each working on segmenting a single organ using abdominal CT scans. Using our proposed method, they can collaboratively train a multi-organ segmentation model capable of segmenting all four organs.

number of nodes, segmentation tasks, and dataset characteristics. We further compared the performance of SegViz to different FL techniques, individual models trained separately for each task, and centralized models trained by aggregating all the datasets in one place. We hypothesize that SegViz will effectively consolidate knowledge across different datasets and demonstrate equivalent performance to the current state-of-the-art non-FL baselines such as models trained individually on each task.

2. Related Work

Generating manual annotations for medical images is time-consuming, requires high skill, and is an expensive effort, especially for 3D images [14]. 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 FL [3]. 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 distributed heterogeneous datasets with partial, incomplete labels [6, 9, 12, 15, 16].

Xu et al. introduced Fed-MENU [16], an FL setup for segmentation using partial labels where client nodes were trained on specific encoders for their specific tasks using a shared decoder. However, this technique required as many encoder blocks as the number of tasks making it computationally expensive. In addition, this technique requires apriori knowledge of all potential tasks across all participating nodes making it practically challenging to scale. In contrast, Shen et al. [13] trained a single global model for aggregating knowledge from partially labeled nodes. However, the global federated learning framework developed in their work 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. In contrast, the marginal loss implemented in Liu et al. [9] and the conditional knowledge distillation technique implemented in Wang et al. [15] were both able to successfully segment all organs being segmented across all the nodes. Similarly, Jiang et al. [6] also introduced a knowledge distillation based technique for training a global FL model from partially annotated sites. However, a major limitation across all these techniques is the requirement of knowledge of all tasks being tackled across different nodes. This is practically challenging to setup, especially in a class-heterogeneous scenario. Furthermore, these techniques cannot scale to accommodate

scenarios with new nodes with newer tasks joining the federation at different points in time.

Therefore, we developed SegViz to address the shortcomings of current techniques in efficiently aggregating knowledge from heterogeneous datasets with partial annotations. Our method 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, and is able to tackle domain shifts between these datasets.

3. Materials and Methods

3.1. Clinical Data

This study was approved by the institutional review board and compliant with HIPAA regulations. Informed consent was waived given the use of de-identified public datasets. This retrospective study utilized four public datasets for model development and validation, and one independent dataset for testing. The image acquisition, processing, and annotation details for these datasets have been summarized in Table 1.

3.1.1 Spleen MSD

The Spleen dataset was obtained from the Medical Segmentation Decathlon (MSD) challenge [1]. It contains 61 abdominal CT volumes collected from a mix of patients and healthy volunteers. Manual spleen segmentations were obtained for all scans. The original data includes both training (N=41) and test sets (N=20); only the training subset was used in this study for model development. Scans varied in spatial dimensions and voxel spacing. While multiple organs (e.g., liver, kidneys, pancreas) are visible within the imaged anatomy, only the spleen segmentation labels were provided.

3.1.2 Liver MSD

The Liver dataset was also sourced from the MSD challenge [1]. It comprises 201 contrast-enhanced abdominal CT scans collected at varying imaging sites and protocols. Manual delineations of the liver structure and lesions are included; only liver organ annotations were retained for this study. The original data includes both training (N=131) and test sets (N=70); only the training subset was used in this study for model development. Scans exhibit heterogeneity in dimensions and resolutions. Both patient and healthy volunteer cases are covered. Only the liver organ is annotated although other visible structures (spleen, pancreas, kidneys) exist without labels.

3.1.3 Pancreas MSD

This dataset was collected from the pancreas subsection of the MSD challenge [1]. It consists of 420 abdominal CT volumes acquired from multiple institutions using differing scanners and protocols. Manual pancreas segmentation masks are provided for all scans. Similar to the previous datasets, the original data includes both training (N=282) and test sets (N=139); only the training subset was used in this study for model development. As in the other MSD subsets, multiple organs appear within the imaged anatomy, but annotations cover solely the pancreas gland.

3.1.4 Kidney Tumor Segmentation (KiTS19)

The Kidney dataset was obtained from the Kidney Tumor Segmentation (KiTS19) challenge [5]. It encompasses 210 contrast-enhanced abdominal CT volumes collected from multiple institutions using various scanners and protocols. Manual delineations of kidney structures and kidney tumors are included; only organ annotations were used in this work. Consistent with the other datasets leveraged, numerous organs and tissues are visible within the CT scan field of view but labels are only provided for the kidney structures. Preprocessing retained the kidney organ segmentation masks while discarding the kidney tumor labels, yielding a dataset with incomplete annotations covering only the kidneys.

3.1.5 Beyond the Cranial Vault (BTCV)

The BTCV dataset was used as the independent test set across all our experiments [7]. The BTCV dataset consisted of 50 scans of portal-venous phase contrast-enhanced abdominal CT volumes, out of which 30 scans from the training set with annotations for thirteen different organ annotations, including the four organs of interest: liver, spleen, kidneys, and pancreas were considered as part of our test set.

3.2. Image Pre-processing

Each of the four training datasets exhibited heterogeneity in annotations, voxel spacing, dimensions, and formats as shown in Table 1. To harmonize the data, we first discarded any tumor or lesion annotations and retained only the organ segmentations present in each dataset. This resulted in heterogeneous multi-organ data with incomplete disjoint labels across datasets for the liver, spleen, pancreas, and kidneys. To enable efficient harmonization for model training, all volumes were resampled to isotropic 1.5 x 1.5 x 2.0 mm voxel spacing and reshaped to uniform 256 x 256 x 128 voxel dimensions using trilinear interpolation. Normalization scaled intensities for each volume to a range of [0, 1]. To augment data, random 128 x 128 x 32 foreground

Table 1. Dataset description and availability of organ annotations for the Medical Segmentation Decathlon (MSD) Liver, MSD Spleen, MSD Pancreas, Kidney Tumor segmentation (KiTs-2019) and BTCV datasets.

Set	Annotated Organ	Dataset	Modality	Imaging Protocol	Median Shape	Median Spacing (mm)	Sample Size	Intensity Range
Training + Validation	Liver	MSD	CT	Portal Venous Phase	432 x 512 x 512	1 x 0.77 x 0.77	131	[-200 , 200]
Training + Validation	Spleen	MSD	CT	Portal Venous Phase	90 x 512 x 512	5 x 0.79 x 0.79	41	[-57 , 164]
Training + Validation	Kidneys	KiTs-2019	CT	Preop. Late Arterial Phase	107 x 512 x 512	3 x 0.78 x 0.78	206	[-79 , 304]
Training + Validation	Pancreas	MSD	CT	Portal Venous Phase	93 x 512 x 512	2.5 x 0.8 x 0.8	282	[-87 , 199]
Testing	Liver, Spleen, Kidneys, Pancreas	BTCV	CT	Portal Venous Phase	128 x 512 x 512	3 x 0.76 x 0.76	30	[-175, 250]

patches were extracted from each volume, centered on voxels belonging to the labeled organs.

3.3. Segmentation Architecture

We implemented a 3D U-Net architecture [4] with 5 levels and 2 residual convolutional blocks per level for receptive field depth. The same architecture was used across all the experiments and methods. We used batch normalization to ensure stable convergence. For every experiment, the network was trained for 500 epochs on combined extracted patches using the Adam optimizer with initial learning rate 1e-4, batch size 2, and cosine annealing schedule [10]. Data augmentation applied random affine transformations including scaling, rotation, and elastic deformation to improve generalization under distribution shifts between datasets. All our models were implemented using the MONAI framework [2].

3.4. SegViz

SegViz adapts U-Net into an FL framework optimized for learning from heterogeneous data with partial incomplete labels. The U-Net architecture is divided into two parts – the representation block and the task block. The representation block consists of subset architecture withing U-Net that encodes task agnostic features that are aggregated across all the nodes every 10 epochs to align the representation block across all the datasets. We utilized a cosine annealing learning rate schedule with FedAvg to account for non-iid distribution across the datasets [8]. The task block corresponds to the final two convolutional layers in the network that encode task specific features for every node's task and are not aggregated at the server to preserve task-specific features. At the end of the training, the task-specific blocks across all the nodes are attached to the same representation block to create the final multi-task model, as shown in Figure 2 (D). Furthermore, conventional FL models lack privacy protection and may cause data leakage as shown in the literature [18, 19]. In contrast, the selective aggregation strategy employed in SegViz provides a privacy-preserving way to aggregate knowledge across different nodes [17].

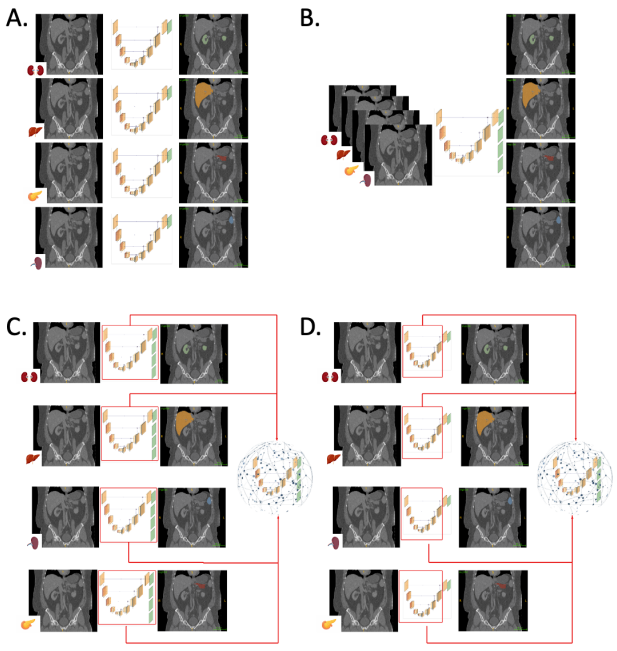

Figure 2. Illustration of the various setups that are trained in this study. (A) Baseline: Independent models trained on each dataset's labels. (B) Central aggregation: Single model trained on combined datasets. (C) Conventional federated learning: Tasks pre-defined, full weight aggregation. (D) SegViz: Selective aggregation preserving task specificity, no prerequisite task knowledge required

Table 2. Comparison of the experimental results between SegViz, single-dataset models, central aggregation, and conventional federated learning (FL) models for each of four organs across all three experimental setups.

Experimental Setup	Models	Organ Segmented			
		Liver	**Spleen**	**Pancreas**	**Kidneys**
Single-Dataset Models	Liver	0.88 ± 0.15	-	-	-
	Spleen	-	0.79 ± 0.17	-	-
	Pancreas	-	-	0.46 ± 0.20	-
	Kidneys	-	-	-	0.64 ± 0.21
2-node	Central Aggregation	0.00 ± 0.00	0.65 ± 0.14	-	-
	Conventional FL	0.89 ± 0.06	0.84 ± 0.09	-	-
	SegViz	$\mathbf{0.90} \pm 0.04$	$\mathbf{0.84} \pm 0.12$	-	-
3-node	Central Aggregation	0.00 ± 0.00	0.61 ± 0.18	0.00 ± 0.00	-
	Conventional FL	0.75 ± 0.29	0.71 ± 0.17	0.45 ± 0.17	-
	SegViz	$\mathbf{0.91} \pm 0.06$	$\mathbf{0.81} \pm 0.17$	$\mathbf{0.55} \pm 0.19$	-
4-node	Central Aggregation	0.65 ± 0.14	0.00 ± 0.00	$\mathbf{0.55} \pm 0.18$	0.68 ± 0.21
	Conventional FL	0.91 ± 0.03	0.46 ± 0.15	0.53 ± 0.18	0.68 ± 0.12
	SegViz	$\mathbf{0.93} \pm 0.02$	$\mathbf{0.78} \pm 0.14$	0.40 ± 0.20	$\mathbf{0.78} \pm 0.12$

3.5. Comparative Methods

3.5.1 Ensemble of Single-Dataset Models

For comparison, individual 3D U-Net models were trained on each dataset for the corresponding annotated organ as shown in Figure 2 (A). These baseline models represent the conventional setup where there are no incomplete annotations and each model is trained specifically for only the annotations present in a dataset, in this case, single organ annotations. Each model was trained from scratch for 500 epochs on its dataset's extracted patches using a batch size of 2 and cosine annealing learning rate decay.

3.5.2 Central Aggregation

To evaluate multi-organ learning, a central aggregation approach combined all datasets into one unified repository for training a single model on all data as shown in Figure 2 (B). While this does not reflect real-world privacy constraints, it provides a lower bound on performance given the partial incomplete labels across datasets. The aggregated 3D U-Net had the same configuration as the baseline models. It was trained on the combined patch data using the same scheme, learning all tasks simultaneously from the heterogeneous labels.

3.5.3 Conventional Federated Learning

We compare SegViz to a conventional federated learning (FL) approach using Federated Averaging (FedAvg) [11]. Four clients were configured, each initialized with the full 3D U-Net architecture containing segmentation heads for all four organs as shown in Figure 2 (C). Every 10 local epochs, the server aggregated and synchronized all the weights between clients using the FedAvg algorithm. This represents traditional FL without specific optimizations for heterogeneous partial labels. Similar to SegViz, we utilized a cosine annealing learning rate schedule to account for non-iid distribution across the datasets.

3.6. Experiments

We evaluated all four approaches in three experimental setups with two, three, and four nodes. The liver and the spleen datasets were used for the two-node experiment. The three-node experiment added the pancreas dataset and the four-node experiment used all four datasets. For each experiment, the model performance was evaluated on 30 external volumes from the BTCV dataset using the Dice similarity metric. We performed Mann-Whitney non-parametric tests using SciPy 1.5 for statistical comparisons, with $p < 0.05$ defining significance.

4. Results

Table 2 summarizes the performance of all four models on all four organs across all experiments. Our results demonstrated that the SegViz method consistently outperformed the comparative methods for segmentation of all four organs across all experimental setups. Figure 4 provides a visual comparison between the performance of all four models compared to the ground truth for four example cases. Individual comparisons are detailed in the following subsections.

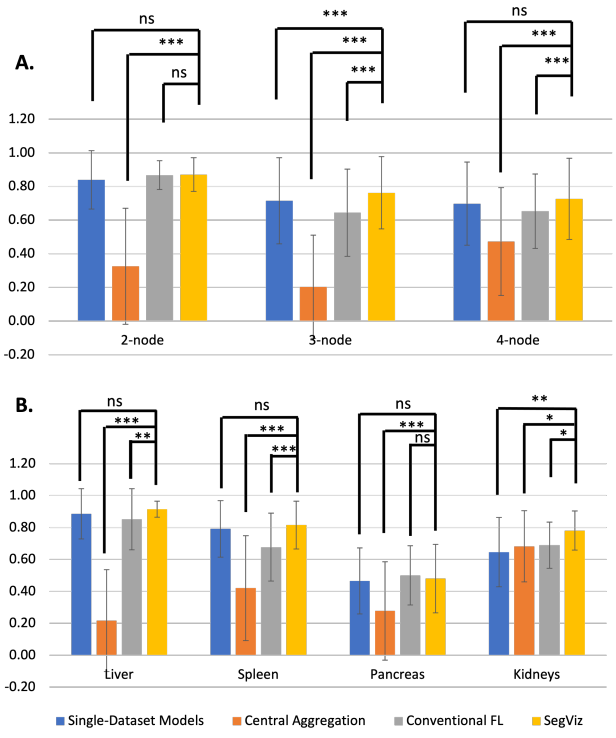

Figure 3. (A) Comparison between all four comparative models in different experimental setups: 2-, 3-, and 4-node. (B) Comparions between the four comparative models across each organ evaluated in this work. (ns: $p \geq 0.05$, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$)

4.1. SegViz vs. Single-Dataset Models

As shown in Figure 3 (A), the overall performance for SegViz across the 2-, 3-, and 4-node setups was 0.87 ± 0.10, 0.76 ± 0.21, and 0.73 ± 0.24, respectively. In comparison, the baseline models individually trained on each dataset had a similar overall performance ($p > 0.05$) of 0.84 ± 0.17 and 0.70 ± 0.25 for 2-, 4-dataset scenarios, respectively. However, baseline model performance in the 3-dataset scenario was 0.71 ± 0.26, significantly ($p < 0.001$) lower than SegViz.

Similarly, there was no significant difference in the overall performance of the two models across liver, spleen, and pancreas as shown in Figure 3 (B). For the kidneys, SegViz again outperformed the baseline model trained only KiTS19 (SegViz: 0.78 ± 0.12 vs. Baseline: 0.65 ± 0.22, $p < 0.01$)

4.2. SegViz vs. Central Aggregation

As shown in Figure 3 (A) and (B), SegViz outperformed central aggregation across all organs and experimental setups. Furthermore, our results demonstrated the inability of conventional segmentation models to effectively train on datasets with missing annotations, even with central aggregation. As shown in Table 2, the central aggregation model failed to segment liver in the 2-node setup, failed to segment both liver and pancreas in the 3-node setup, and failed to segment spleen in the 4-node setup. SegViz, on the other hand successfully segmented all organs across all datasets across all experiments.

4.3. SegViz vs. Conventional FL

As shown in Figure 3 (A), there was no significant difference between the performance of conventional FL and SegViz in a 2-node setup. However, the performance of the conventional FL fails to scale with the increasing number of nodes with a significantly lower performance than SegViz for both the 3-node (SegViz: 0.76 ± 0.21 vs. Conventional FL: 0.64 ± 0.26, $p < 0.001$) and the 4-node (SegViz: 0.73 ± 0.24 vs. Conventional FL: 0.65 ± 0.22, $p < 0.001$) setups.

Similarly as shown in Figure 3 (B), SegViz significantly outperformed conventional FL across all organs, except pancreas where the difference was not significant. As shown in Table 2, the conventional FL model performs consistently well for the segmentation of Liver across all experimental setups (0.89 ± 0.06 for the 2-node to 0.91 ± 0.03 for the 4-node setup). However, the performance of the spleen drops significantly with the increasing number of nodes (0.84 ± 0.09 for the 2-node to 0.46 ± 0.15 for the 4-node setup).

5. Discussion

This study introduced SegViz, a federated learning framework that achieves excellent multi-organ segmentation performance by effectively aggregating knowledge from unique heterogeneous datasets with partial disjoint annotations. Unlike current approaches in the literature [6, 9, 12, 15, 16], SegViz enables decentralized collaboration between nodes without requiring awareness of all tasks across clients. Furthermore, the selective aggregation approach used in SegViz provides additional security to data leakage as only partial networks are shared and aggregated at the server [17].

SegViz provides a mechanism for research groups and healthcare centers to improve multi-organ capabilities by learning from each other, without sharing raw protected health data by consolidating task knowledge from distinct partial labels into unified federated models. This setup thus allows a flexible federation and provides the possibility of onboarding of new participants to join the federation. This decentralized collaboration has the potential to consolidate learning across isolated datasets thereby reducing the need to manually segment organs and accelerate the translation of radiological research into clinical practice.

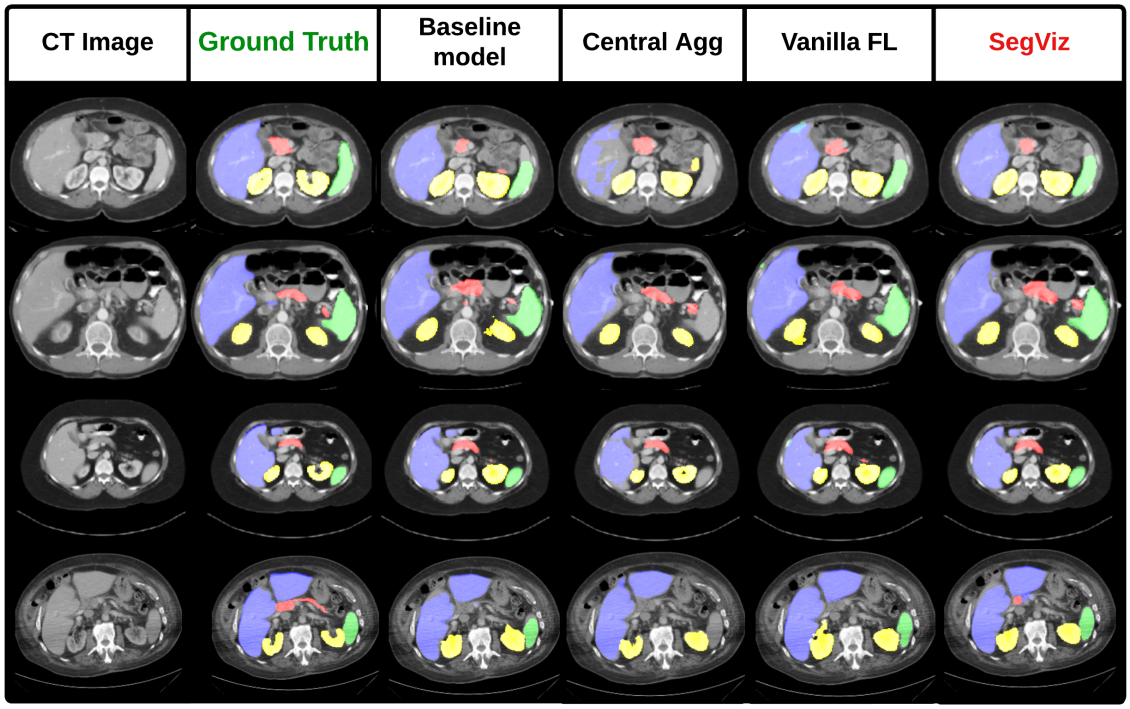

Figure 4. Visualization of the overlay segmentations for four different images from the BTCV test set. Purple masks are for Liver, Green for Spleen, Red for Pancreas, and Yellow for Kidneys

Quantitative results demonstrate SegViz matches or exceed fully supervised baselines; models trained independently on each fully labeled dataset representing an upper bound, despite learning from incomplete heterogeneous data. This highlights effective knowledge transfer, with federated collaboration compensating for partial labels. In contrast, naive central aggregation of all data showed significantly degraded performance. This result was expected given the missing ground truth across datasets. SegViz overcomes this challenge via selective weight synchronization during federated learning, preventing contamination across disjoint partial labels and effectively consolidating complementary knowledge. Further, a conventional FL approach was also inferior in performance to SegViz. Conventional FL demonstrated significantly lower performance with increase in the number of nodes. SegViz outperformed conventional FL as the number of nodes in the federation increased demonstraing SegViz's potential for scaling to several nodes while maintaining knowledge transfer.

Our study has several limitations - First, SegViz requires the same model architecture at each client thus making it difficult for multi-task learning at each client site. This was demonstrated by the model's inability to effectively segment pancreas across all experiment setups and models. Next, our study was only evaluated using CT data and further experiments would need to be conducted to include other imaging modalities like MRI. We also note that our experiments were only performed on disjoint data and needs to be evaluated in scenarios with overlapping labels. In the future, we will extend our experiments to other imaging modalities (e.g., MRI) and overlapping labels. In addition, we will also investigate the real-world performance of our FL setup where random client nodes can continue to join the federation at different points in time.

Acknowledgements: This research was supported by the UMMC/UMB Innovation Challenge Award, 2023.

References

[1] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):1–13, 2022. 3

[2] M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022. 4

[3] Alexander Chowdhury, Hasan Kassem, Nicolas Padoy, Renato Umeton, and Alexandros Karargyris. A review of med-

ical federated learning: Applications in oncology and cancer research. In *International MICCAI Brainlesion Workshop*, pages 3–24. Springer, 2022. 2

[4] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016. 4

[5] Nicholas Heller, Niranjan Sathianathen, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg, Paul Blake, Zachary Rengel, Makinna Oestreich, 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

[6] Le Jiang, Li Yan Ma, Tie Yong Zeng, and Shi Hui Ying. Ufps: A unified framework for partially annotated federated segmentation in heterogeneous data distribution. *Patterns*, 5 (2), 2024. 2, 6

[7] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, page 12, 2015. 3

[8] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019. 4

[9] Pengbo Liu, Mengke Sun, and S Kevin Zhou. Multi-site organ segmentation with federated partial supervision and site adaptation. *arXiv preprint arXiv:2302.03911*, 2023. 2, 6

[10] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 4

[11] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data, 2017. 5

[12] Chen Shen, Pochuan Wang, Holger R Roth, Dong Yang, Daguang Xu, Masahiro Oda, Weichung Wang, Chiou-Shann Fuh, Po-Ting Chen, Kao-Lang Liu, 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*, pages 101–110. Springer, 2021. 2, 6

[13] Chen Shen, Pochuan Wang, Dong Yang, Daguang Xu, Masahiro Oda, Po-Ting Chen, Kao-Lang Liu, Wei-Chih Liao, Chiou-Shann Fuh, Kensaku Mori, 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*, pages 58–67. Springer, 2022. 2

[14] Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N Chiang, Zhihao Wu, and Xiaowei Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis*, 63:101693, 2020. 2

[15] Pochuan Wang, Chen Shen, Weichung Wang, Masahiro Oda, Chiou-Shann Fuh, Kensaku Mori, and Holger R Roth. Condistfl: Conditional distillation for federated learning from partially annotated data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 311–321. Springer, 2023. 2, 6

[16] Xuanang Xu, Hannah H Deng, Jamie Gateno, and Pingkun Yan. Federated multi-organ segmentation with inconsistent labels. *IEEE transactions on medical imaging*, 2023. 2, 6

[17] Qian Yang, Jianyi Zhang, Weituo Hao, Gregory P Spell, and Lawrence Carin. Flop: Federated learning on medical datasets using partial networks. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3845–3853, 2021. 4, 6

[18] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. idlg: Improved deep leakage from gradients. *arXiv preprint arXiv:2001.02610*, 2020. 4

[19] Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *Advances in neural information processing systems*, 32, 2019. 4