Medical Image Analysis 95 (2024) 103156

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Federated learning with knowledge distillation for multi-organ segmentation with partially labeled datasets

Soopil Kim [a,b,1], Heejung Park [a,1], Myeongkyun Kang [a,b], Kyong Hwan Jin [c], Ehsan Adeli [b], Kilian M. Pohl [b], Sang Hyun Park [a,*]

[a] Department of Robotics and Mechatronics Engineering, Daegu Gyeongbuk Institute of Science and Technology, Republic of Korea
[b] Department of Psychiatry and Behavioral Sciences, Stanford University, CA 94305, USA
[c] School of Electrical Engineering, Korea University, Republic of Korea

ARTICLE INFO

MSC:
41A05
41A10
65D05
65D17

Keywords:
Federated learning
Partially labeled datasets
Knowledge distillation
Multi-organ segmentation

ABSTRACT

The state-of-the-art multi-organ CT segmentation relies on deep learning models, which only generalize when trained on large samples of carefully curated data. However, it is challenging to train a single model that can segment all organs and types of tumors since most large datasets are partially labeled or are acquired across multiple institutes that may differ in their acquisitions. A possible solution is Federated learning, which is often used to train models on multi-institutional datasets where the data is not shared across sites. However, predictions of federated learning can be unreliable after the model is locally updated at sites due to 'catastrophic forgetting'. Here, we address this issue by using knowledge distillation (KD) so that the local training is regularized with the knowledge of a global model and pre-trained organ-specific segmentation models. We implement the models in a multi-head U-Net architecture that learns a shared embedding space for different organ segmentation, thereby obtaining multi-organ predictions without repeated processes. We evaluate the proposed method using 8 publicly available abdominal CT datasets of 7 different organs. Of those datasets, 889 CTs were used for training, 233 for internal testing, and 30 volumes for external testing. Experimental results verified that our proposed method substantially outperforms other state-of-the-art methods in terms of accuracy, inference time, and the number of parameters.

1. Introduction

The state-of-the-art in automatically segmenting organs from abdominal CT images are supervised deep learning approaches (Gibson et al., 2018; Kim et al., 2021; Ma et al., 2022). However, training a segmentation model for all abdominal organs and tumors is challenging, as there are no public data sets containing a large number of CT scans accompanied with complete segmentation labels. Existing datasets are only *partially labeled*, i.e., only a few organs or types of tumors are segmented (Bilic et al., 2023; Simpson et al., 2019). Accordingly, models specifically suitable for being trained on partially labeled CT data from multiple institutes have been proposed (Zhang et al., 2021b; Dmitriev and Kaufman, 2019).

A straightforward approach for segmenting multiple organs in CT is to use segmentation models that are separately trained on each partially labeled CT dataset (Hu et al., 2016). However, this strategy is computationally inefficient and accuracy is limited when the dataset includes a small number of samples. An alternative is incremental learning (Li and Hoiem, 2017; Elskhawy et al., 2020; Vu et al., 2021), where a single model is updated by training it sequentially on the datasets. Despite recent advances, these models often struggle to maintain accuracy on the initial data sets, since they easily forget the knowledge gained from the previously used data sets (Xiao et al., 2023). One promising training strategy is federated learning (Li et al., 2020a), which trains a model at each site and then merges the parameters of the model across sites via a central server.

Though federated learning has been applied to medical image segmentation, most existing implementations assume that all sites (a.k.a. client nodes) have labels for the same set of organs (Li et al., 2019; Wang et al., 2020; Xia et al., 2021). In practice, however, the CT acquisitions and the organs annotated differ across sites (or nodes). To account for this difference, one can train a network of nodes (Xu et al., 2023) using federated averaging (McMahan et al., 2017) (FedAvg), i.e., each node has its own encoder and a shared decoder is used across all nodes. However, this approach requires a large number of

* Corresponding author.
E-mail address: shpark13135@dgist.ac.kr (S.H. Park).
[1] Equal contribution.

https://doi.org/10.1016/j.media.2024.103156
Received 10 July 2023; Received in revised form 11 March 2024; Accepted 20 March 2024
Available online 25 March 2024

parameters to be tuned and the accuracy is limited since each encoder is trained on a relatively small number of samples. In addition, the knowledge for segmenting organs is lost during the 'local' training at a node (a.k.a. catastrophic forgetting), which is specific to certain organs. Finally, the isolated updating of local models can result in degraded accuracy of the overall model across sites.

To overcome this limitation, we regularize training at each site with knowledge from the global model and pre-trained organ-specific segmentation models. We do so using global and local knowledge distillation (KD) (Gou et al., 2021), that effectively mitigate forgetting by imposing constraints to retain segmentation results for unlabeled organs when trained with partially labeled data. Furthermore, we propose a new baseline structure that consists of a shared encoder–decoder, similar to U-Net (Ronneberger et al., 2015), and lightweight segmentation heads with just 162 parameters for each target organ. As our model shares most parts of the encoder and decoder across sites, the representations from multiple datasets can be accomplished in a single model and the inference can be quickly performed without repeating the feed-forward process for multiple target organs (Zhang et al., 2021a; Dmitriev and Kaufman, 2019; Wu et al., 2022). We evaluate our proposed method using eight public CT datasets (Bilic et al., 2023; Heller et al., 2019; Simpson et al., 2019; Landman et al., 2015). The datasets differ with respect to the segmented organs and pathology. Besides achieving significantly higher accuracy than several state-of-the-art methods, our method is efficient with respect to inference time, contains a relatively small number of parameters, and is robust against catastrophic forgetting.

2. Related works

2.1. Leaning a model with partially labeled datasets

Learning a single model for multiple partially labeled datasets is a challenging problem that arises across various contexts such as classification (Duarte et al., 2021; Durand et al., 2019), detection (Feng et al., 2019; Yan et al., 2020), and segmentation (Verbeek and Triggs, 2008; He and Zemel, 2008). In the medical domain, this is particularly relevant for segmentation tasks where multiple organs are present in a single medical image, making it difficult to annotate all pixels accurately. To address it, various strategies have been proposed.

One can use U-Nets (Ronneberger et al., 2015) predicting outputs for each dataset. Chen et al. (2019) proposed an architecture consisting of a shared encoder and multiple decoders of U-Net for multiple organs. Instead of defining separate encoders or decoders, conditional U-Nets share an encoder and a decoder across all organs (Dmitriev and Kaufman, 2019; Zhang et al., 2021b; Wu et al., 2022). Dmitriev and Kaufman (2019) extracted class-specific feature maps by utilizing class labels in intermediate CNN layers. Zhang et al. (2021b) used a shared encoder and decoder of U-Net and dynamically generated weights in the segmentation head based on the organ ID. TGNet (Wu et al., 2022) introduced task attention modules to each layer in U-Net to extract task-relevant features.

Alternatively, Zhou et al. (2019), Shi et al. (2021) and Fidon et al. (2021) suggested novel loss functions since the ordinary multi-class cross-entropy loss used by U-Nets cannot be applied to partially labeled data. Zhou et al. (2019) proposed a prior-aware loss with the assumption that some fully labeled data samples are available. Their loss function regularizes the distribution of model predictions to follow the fully labeled data. Shi et al. (2021) proposed a marginal loss to merge background and unlabeled organs and an exclusive loss to increase the distance between labeled and unlabeled organs. Fidon et al. (2021) proposed a leaf-dice loss compatible with missing labels. However, the approaches discussed earlier are predominantly designed for a centralized experimental setting where the model can access the entire dataset. Moreover, they did not consider the efficiency when predicting multiple organs. In particular, many conditional U-Net models (Zhang et al., 2021a; Dmitriev and Kaufman, 2019; Wu et al., 2022) must repeat the entire feed-forward process to acquire predictions for other organs. Thus, the computation increases by the number of organs. In contrast, our method can consider predictions for various organs with a single feed-forward process and be applied to both the centralized and federated settings.

2.2. Federated learning

Recently, federated learning (Li et al., 2020a; Kang et al., 2024) has been actively explored to learn a global model without sharing local data across institutions. The most commonly used one is Federated Averaging (McMahan et al., 2017) (FedAvg), which obtains a global model by averaging the parameters of the local models updated in each node (client). FedAvgs have been popularly utilized for medical image segmentation (Lu et al., 2022; Yang et al., 2021). However, most of them use a complete dataset or focus on a single target task, e.g., brain tumor segmentation (Li et al., 2019; Sheller et al., 2018), breast tumor segmentation (Wicaksana et al., 2022), pancreas segmentation (Wang et al., 2020; Shen et al., 2021), or COVID-19 lesion segmentation (Xia et al., 2021). However, when the independent identically distributed (i.i.d.) condition of local data is not guaranteed, the accuracy of FedAvg is often substantially degraded. Several methods have been recently proposed to address the non-i.i.d. datasets in federated learning. FedProx (Li et al., 2020b) utilizes the L2 distance between the global and clients' parameters to regularize the client's local update. FedDyn (Acar et al., 2021) aligns the local and global models using a dynamic regulator for each client. Fedscaffold (Karimireddy et al., 2020) adjusted the local update to reduce the variance of client gradients. Other federated learning models are based on knowledge distillation (KD) leveraging additional unlabeled data in the central server to distill the ensembled knowledge of local models (Chen and Chao, 2020; Li and Wang, 2019). Zhang et al. (2022) generated pseudo data for data-free KD. Lee et al. (2022) also utilized KD to address forgetting the knowledge of the global model. Though successfully improving accuracy, they are not suitable for partially labeled datasets since they focus on multi-class classification using complete datasets. Here, we propose the first KD-based segmentation model accurately working on partially labeled datasets.

Note that there are federated learning models specifically designed to handle partially labeled datasets. Shen et al. (2022) applied FedAvg to a U-Net-like segmentation model (C2FNAS) (Yu et al., 2020). Xu et al. (2023) proposed MENU-Net consisting of separate encoders for each client and an auxiliary generic decoder that predicts binary segmentation from the encoders. MENU-Nets concatenate feature maps of encoders and use them as input for the decoder. However, since most methods simply used FedAvg, the performance was limited as the client forgot the knowledge of the global model during the local update. Apart from the prior works, we carefully design a baseline architecture that can quickly obtain predictions of various tasks without repeating the feed-forward process. We also alleviate the forgetting issue by using global and local KD losses. Finally, we evaluate our method using eight datasets including various organs and tumors, which are larger than the prior related work (i.e., Shen et al. (2022) and Xu et al. (2023)).

3. Methods

3.1. Problem setup

Our goal is to learn a single model from multiple datasets $\{D^1, D^2, \ldots, D^{N_D}\}$ distributed across various clients (or nodes) $\{c^1, c^2, \ldots, c^{N_D}\}$ where N_D is the number of data groups or clients. Assuming each client c^k is assigned one partially labeled dataset D^k, we use k as an index for both the client and dataset. Each dataset D^k may have a different distribution compared to other datasets and consists of N_I^k CT images $X_i^k \in \mathbb{R}^{Z\times H\times W} (i = 1, \ldots, N_I^k)$ and corresponding voxel-level

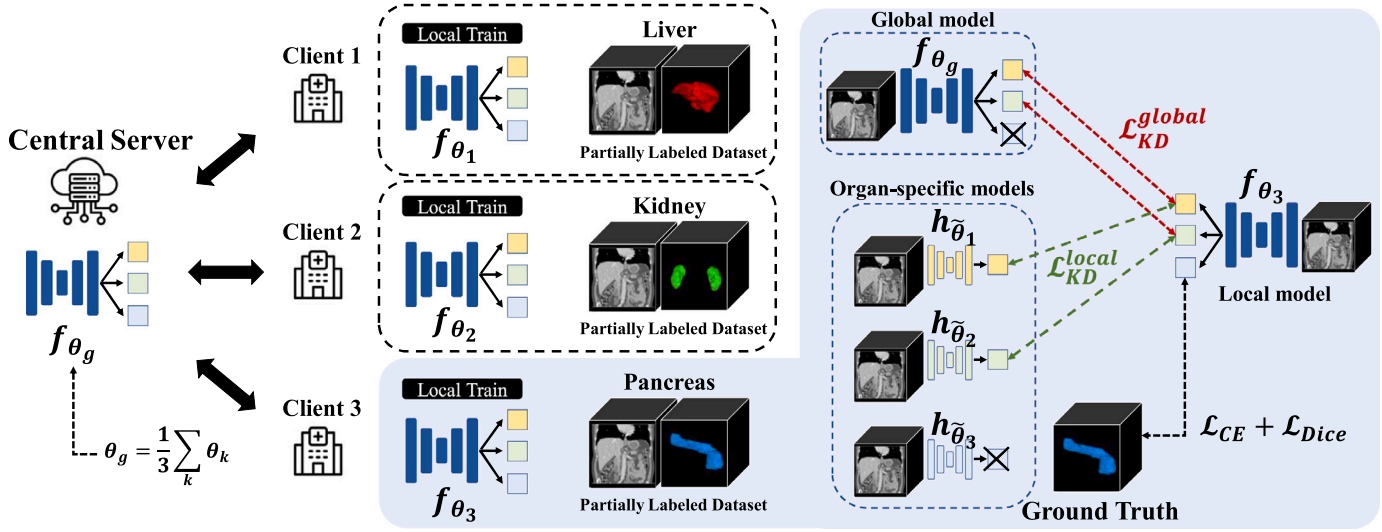

Fig. 1. Proposed federated learning approach for multi-class segmentation from partially labeled datasets. For simplicity, each dataset in this figure only contains segmentation of one organ. f_{θ_g}, f_{θ_k} and $h_{\tilde{\theta}_l}$ denote a global-, local-, and pre-trained organ-specific model, respectively. $h_{\tilde{\theta}_l}$ is pre-trained in each client as illustrated in Fig. 2. During local training, the local model is updated by minimizing local and global KD loss functions, whereas θ_g and $\tilde{\theta}_l$ are frozen.

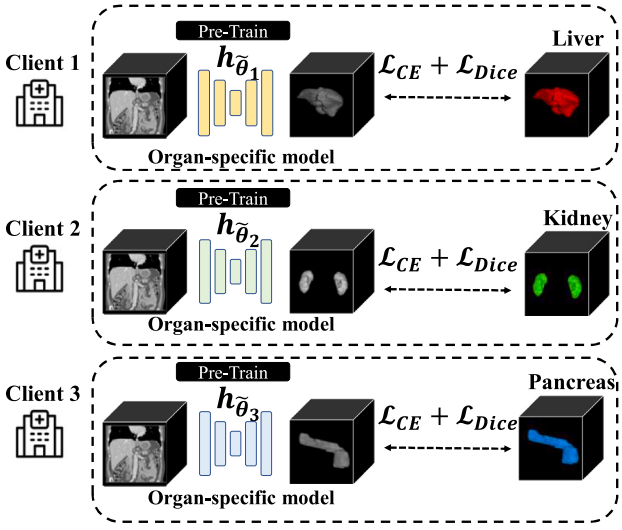

Fig. 2. Organ-specific models are pre-trained in each client and their parameters of the models are shared with other clients before FL starts.

segmentations $Y_i^k \in \mathbb{R}^{Z\times H\times W}$ (Z is the dimension in axial direction, H in coronal, and W in sagittal). Additionally, the datasets may differ with respect to the segmented target objects (i.e., organs or tumors). Let $L = \{1, 2, \ldots, N_C\}$ represent the complete set of classes, where N_C is the total number of classes. Then, each voxel in Y_i^k belongs to a subset $L^k \subseteq L$. We train a single segmentation network f parameterized by θ, which predicts a segmentation $f_\theta(X_i^k, l)$ for the lth target organ in the input image. We optimize θ to minimize loss functions $\mathcal{L}(\cdot,\cdot)$ across partially labeled datasets as follows:

$$\theta = \underset{\theta}{\operatorname{argmin}} \sum_{k=1}^{N_D} \sum_{i=1}^{N_I^k} \sum_{l\in L^k} \mathcal{L}(f_\theta(X_i^k, l), Y_i^{kl}). \tag{1}$$

where Y_i^{kl} denotes binary segmentation label of the lth target organ.

3.2. Federated averaging

Fig. 1 illustrates our proposed approach based on the FedAvg (McMahan et al., 2017) framework for aggregating local models from multiple clients through parameter averaging. We initialize the global model f_{θ_g} with parameters θ_g and transmit it to the clients. Next, each client trains its own model f_{θ_k} using D^k. The model f_{θ_k} is optimized to minimize the loss function between its predictions $f_{\theta_k}(X_i^k, k)$ and the corresponding ground truth Y_i^k for kth organ as:

$$\underset{\theta_k}{\operatorname{argmin}} \sum_{i=1}^{N_I^k} \sum_{l\in L^k} \mathcal{L}(f_{\theta_k}(X_i^k, l), Y_i^{kl}). \tag{2}$$

Existing segmentation methods for partially labeled datasets often use a combination of cross-entropy loss and dice-loss (Zhang et al., 2021b) between the predicted mask and ground truth as:

$$\mathcal{L}(f_{\theta_k}(X_i^k, l), Y_i^{kl}) = \mathcal{L}_{CE,i} + \mathcal{L}_{Dice,i}, \tag{3}$$

where

$$\mathcal{L}_{CE,i} = -\frac{1}{N_V} \sum_{j=1}^{N_V} Y_{ij}^{kl} \log(f_{\theta_k}(X_{ij}^k, l)), \tag{4}$$

$$\mathcal{L}_{Dice,i} = 1 - \frac{2\sum_{j=1}^{N_V} f_{\theta_k}(X_{ij}^k, l) Y_{ij}^{kl}}{\sum_{j=1}^{N_V} f_{\theta_k}(X_{ij}^k, l) + \sum_{j=1}^{N_V} Y_{ij}^{kl}}. \tag{5}$$

j is a voxel index and $N_V = Z\times H\times W$ is the number of voxels. $\mathcal{L}_{CE}$ is the most popularly used voxel-level classification loss. $\mathcal{L}_{Dice}$ measures the similarity of predicted mask and ground truth and is beneficial for small object segmentation. After local training, the global model is updated by averaging the client's parameters as:

$$\theta_g = \frac{1}{N_D} \sum_{k=1}^{N_D} \theta_k. \tag{6}$$

Even if the original FedAvg rebalances the clients' parameters using the number of data points in each client, we do not weigh the client's parameters since the label distributions across clients can be heterogeneous in our problem setting. FedAvg repeats this communication process until convergence. The final segmentation is obtained using f_{θ_g} after final aggregation.

3.3. Global knowledge distillation for federated learning

However, the global model in FedAvg often results in sub-optimal parameters because the client's model may forget the knowledge for segmenting other organs during local training. Specifically, when the

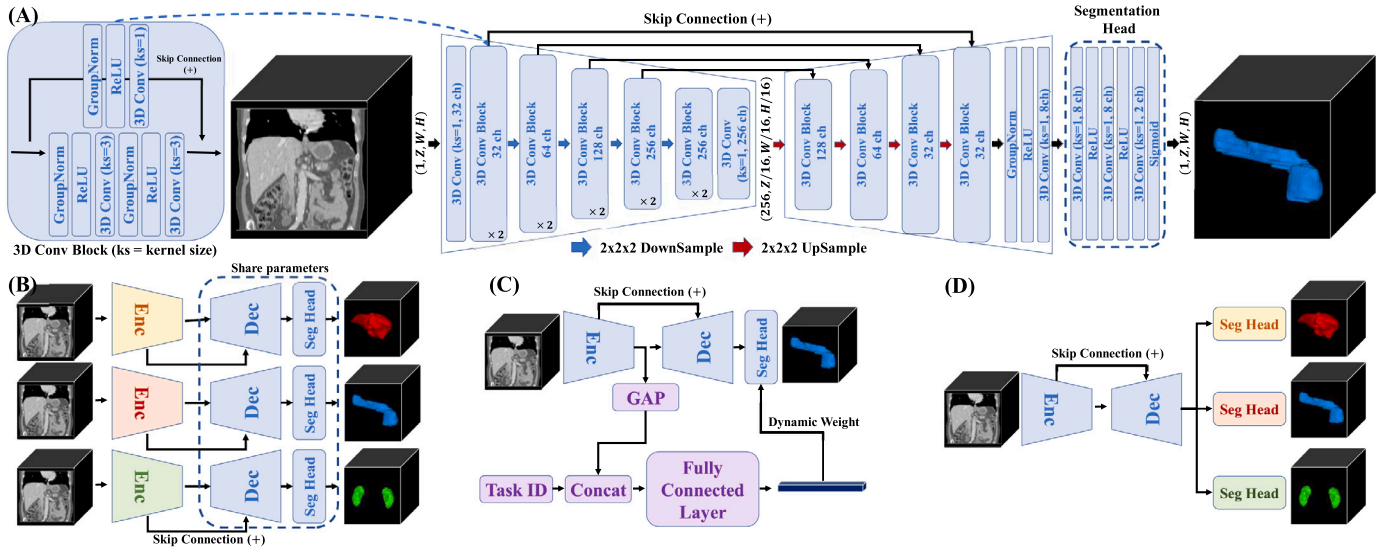

Fig. 3. Different architectures for segmenting partially labeled datasets. (A) 3D U-Net (B) separate encoder (Xu et al., 2023), (C) conditioned network (Zhang et al., 2021b), and (D) our multi-head baseline. Details of the encoder and decoder in (B), (C) are same with (A).

model is updated only using the ground truth for the target organ disregarding predictions for other organs, the predictions for other organs also can be affected, resulting in unintended changes. To mitigate this issue, we define a KD loss that measures cross-entropy between predictions made by the global model f_{θ_g} and the local model f_{θ_k} for unlabeled organs in X_i^k:

$$\mathcal{L}_{KD,i}^{global} = -\frac{1}{N_V}\frac{1}{N_C - n(L^k)}\sum_{l\notin L^k}\sum_{j=1}^{N_V} f_{\theta_g}(X_{ij}^k, l)\log(f_{\theta_k}(X_{ij}^k, l)). \tag{7}$$

The model is updated using a combination of the KD loss, original cross-entropy, and dice loss functions.

$$\mathcal{L}(f_{\theta_k}(X_i^k, k), Y_i^{kl}) = \mathcal{L}_{CE,i} + \mathcal{L}_{Dice,i} + \mathcal{L}_{KD,i} \tag{8}$$

Even if each client has annotations for a single target organ, this loss function can prevent the client from catastrophic forgetting by preserving predictions for other organs. Even though clients cannot share their data, representation from other datasets can be utilized by averaging the clients' parameters.

3.4. Local knowledge distillation for federated learning

The KD loss can be applied in a different way if each client can train an organ-specific segmentation model and share it with the other clients as illustrated in Fig. 2. Before federated learning, each client trains its organ-specific segmentation model $h_{\tilde{\theta}_k}$ parameterized by $\tilde{\theta}_k$ using the local dataset D^k and transmits the model parameters to the other clients. Once transmitted, the local training of client k utilizes the predictions from all organ-specific segmentation models of the other clients:

$$\mathcal{L}_{KD,i}^{local} = -\frac{1}{N_V}\frac{1}{N_C - n(L^k)}\sum_{l\notin L^k}\sum_{j=1}^{N_V} h_{\tilde{\theta}_l}(X_{ij}^k)\log(f_{\theta_k}(X_{ij}^k, l)). \tag{9}$$

However, this loss is computationally expensive, which negatively impacts training speed and increases in severity with the number of clients. Even if one can pre-compute the outputs of $h_{\tilde{\theta}_l}$ and use them to reduce the computation burden, saving all predictions requires a large amount of storage memory which is much larger than the original dataset. This issue will be more serious as we have more classes to segment. Instead, we randomly sample an index $l \notin L^k$ corresponding to organ-specific model $h_{\tilde{\theta}_l}$ to distill its knowledge. Despite updating the model using only one organ-specific model at a time, we can leverage all organ-specific models during numerous updates of local training. Then, the loss is simplified as:

$$\mathcal{L}_{KD,i}^{local} = -\frac{1}{N_V}\sum_{j=1}^{N_V} h_{\tilde{\theta}_l}(X_{ij}^k)\log(f_{\theta_k}(X_{ij}^k, l)). \tag{10}$$

It can significantly reduce the training time compared to using Eq. (9). We empirically found that the accuracy of our model is not sensitive to this sampling (see Table 11).

Compared to $\mathcal{L}_{KD}^{global}$ that can be more widely applied in FL regardless of the number of data samples in each client and without additional memory consumption, $\mathcal{L}_{KD}^{local}$ requires an additional training process for organ-specific models on each client and increased memory consumption to use multiple organ-specific models in FL. Furthermore, accuracy may be limited if the number of data samples in each client is small, leading to a less reliable organ-specific model. On the other hand, $\mathcal{L}_{KD}^{local}$ can benefit from distilling expert knowledge from pre-trained models. Therefore, $\mathcal{L}_{KD}^{local}$ will be particularly effective when each client can train accurate organ-specific models with a lot of local data. We could observe this tendency in our experiments.

3.5. Baseline model architecture

Since our KD losses rely on predictions for multiple organs, we need to carefully design a model architecture that can quickly obtain predictions for all organs while achieving high accuracy. In Fig. 3, we describe existing model architectures and our multi-head U-Net model. For semantic segmentation task, U-Net (Ronneberger et al., 2015) style encoder–decoder architecture have been popular. We can train multiple U-Nets (A) as a naive approach using each dataset. However, it is inefficient from the perspective of the number of parameters, inference time, and segmentation accuracy, as it cannot utilize representations for different organs. Therefore, we can share parts of the model like (B) and (C). MENU-Net (Xu et al., 2023) (B) shares an encoder for the different organs. DoDNet (Zhang et al., 2021b) (C) shares an encoder and decoder for different organs, and the weights in the segmentation head are dynamically predicted by the task controller. Sharing model benefits from learning various representations of data for different organs.

However, other methods like Multiple Nets (A), Sep. Enc (B), and other conditioned networks (Dmitriev and Kaufman, 2019; Wu et al., 2022) must repeat the feed-forward process to obtain a segmentation of another organ. As a result, a longer inference time slows down the

Table 1
Summary of partially labeled medical image segmentation datasets. O and X indicate the presence or absence of annotations for organs and tumors, respectively. # denotes the number.

Client	Task	Annotations		Spacing	Slice thickness	Slice size	# of slice	# of images	
		Organ	Tumor					Train	Test
0	Liver (Bilic et al., 2023)	O	O	0.56~1.00	0.69~5.00	512	74~987	83	25
1	Kidney (Heller et al., 2019)	O	O	0.43~1.04	0.50~5.00	512, 796	29~1059	158	42
2	Hepatic vessel (Simpson et al., 2019)	O	O	0.57~0.98	0.80~8.00	512	24~251	242	61
3	Pancreas (Simpson et al., 2019)	O	O	0.54~0.98	0.63~7.50	512	37~751	224	57
4	Colon (Simpson et al., 2019)	X	O	0.54~0.98	1.25~7.50	512	36~729	100	26
5	Lung (Simpson et al., 2019)	X	O	0.60~0.98	0.63~2.50	512	112~636	50	13
6	Spleen (Simpson et al., 2019)	O	X	0.61~0.98	1.50~8.00	512	31~168	32	9
BTCV	Liver, Kidney, Pancreas, Spleen (Landman et al., 2015)	O	X	0.59~0.98	2.50~5.00	512	85~198	0	30

overall training process with the KD loss. Moreover, the amount of computation involved is comparable to that of Multiple Nets. This limitation becomes more significant when dealing with partially labeled datasets, making them unsuitable for training with KD loss. Even though DoDNet (C) is a current state-of-the-art method in the centralized setting, performance can depend on the task-specific controller in the federated learning. The baseline architecture of our approach consists of a multi-head U-Nets (D). Employing a multi-head segmenter does not lead to a significant increase in the number of parameters, as each head consists of three convolution layers (1st layer: $(8+1)8$ parameters, 2nd : $(8+1)8$, 3rd : $(8+1)2$, total 162). Moreover, this approach is efficient for multi-organ prediction. Compared to other condition-based methods such as TGNet (Wu et al., 2022) and Cond_Enc (Dmitriev and Kaufman, 2019), multi-head model (D) requires a single feed-forward process and can obtain predictions for various targets by applying different shallow headers on the extracted feature map.

3.6. Implementation details

Each client updates the parameters for a fixed number of iterations instead of epochs since each client has a different number of data samples. We repeat the model aggregation for 1000 communication rounds while updating the local models for 80 iterations. Feature maps of size $[32, 64, 128, 256]$ are sequentially extracted during the encoding process, with their spatial dimensions being gradually reduced. Following this, the feature maps are decoded sequentially with skip connections and upsampling operations. Parameters of all architectures were initialized using Kaiming initialization (He et al., 2015). We used stochastic gradient descent (SGD) optimizer with a 1e−2 initial learning rate, 0.99 momentum, and batch size of 2 per iteration on an NVIDIA RTX A5000 GPU workstation. The learning rate is polynomially decayed following $lr = lr_0(1-e/e_{max})^{0.9}$, where e is communication round in the federated setting. Due to the large dimension of the entire CT volume, we crop a patch with [64,128,128] size to be used as input. To focus more on the foreground regions, 80% of the patches are randomly cropped near the region of interest. To augment the training data, we randomly scale the input patch with a ratio between 0.7~1.4, and mirror it with a 50% probability on each dimension. During inference, we obtain predictions for all patches using a sliding window approach and aggregated them to make an entire prediction map of CT. The same patch size is also used during testing.

4. Experiments

4.1. Experimental setting

We evaluated the proposed model using 8 publicly available abdominal CT segmentation datasets (see Table 1), which have annotations of liver, kidney, pancreas, hepatic vessel, colon, lung, and spleen. 7 datasets were used for training the model in a federated setting. The Beyond the Cranial Vault (BTCV) dataset (Landman et al., 2015) was used as an external test to check the transferability of the model. These datasets were previously utilized in Zhang et al. (2021b) to evaluate DoDNet in a centralized setting. The dataset are summarized in Table 1 (e.g., the number of data samples, target task, spacing, and dimension). We used all samples of the 8 data sets for training or testing with the exception of 25 samples from the liver dataset (Bilic et al., 2023), which missed spatial resolution information. Specifically, we used a total of 889 CT volumes for training and 233 volumes for evaluation following the setting of DoDNet. The BTCV dataset comprises 30 abdominal CT images with labels for 13 organs of which the liver, kidney, pancreas, and spleen were selected for evaluation as they are included in the training datasets. As each dataset contains images with various resolutions and dimensions, we resampled every CT scan to $3.0 \times 1.5 \times 1.5 \times$ mm^3 voxel size and normalized intensities in $[-325, +325]$ to $[-1, 1]$.

To evaluate our model in the federated setting, we designed an experiment involving multiple clients. Specifically, we created a total of 7 clients, each with a partially labeled dataset as illustrated in Table 1. Additionally, we extended our evaluation by introducing a 21-client scenario, where the data from each client was randomly partitioned into 3 distinct datasets. We implemented DoDNet (Zhang et al., 2021b) (our benchmark for the state-of-the-art for the centralized setting) and Separate Encoder (Sep. Enc) (Xu et al., 2023) in FedAvg as comparison methods. We reproduced other comparison models with the same complexity of architecture(*i.e.* same channel sizes of the encoder and decoder) for a fair comparison. With respect to federated learning methods, we tested FedProx (Li et al., 2020b) and FedScaffold (Karimireddy et al., 2020) with our proposed multi-head baseline as a backbone. We optimized these models using Eq. (3) and set the output channel size at the last layer to 2 for the organ and its tumor to match the annotations of the datasets. With the exception of hepatic vessels, the organs and their tumors may overlap. Therefore, the foreground of the liver, kidney, and pancreas is defined as the region encompassing their respective tumor regions. We closely follow the evaluation process of Zhang et al. (2021b). After normalization with a sigmoid function (as depicted in Fig. 3(A)), a binary segmentation map is generated using a threshold of 0.5. When there are pre-defined numbers of target objects, the largest segment is selected as the final segmentation. For evaluation, dice similarity score (%), Hausdorff distance (voxel), inference time (second), the number of parameters (million), and computational costs (GFLOPs) are used as metrics.

In addition, we evaluated our method in the centralized setting to compare it with existing architectures proposed to address the partially labeled dataset problem. In this case, all datasets in Table 1 are aggregated in the central server, and the model is trained using the loss function in Eq. (3). We reproduce state-of-the-art methods including Multiple Networks, Sep. Enc, Task Adaptive Loss (TAL) (Fang and Yan, 2020), Cond_Enc (Dmitriev and Kaufman, 2019), Cond_Dec (Dmitriev and Kaufman, 2019), and DoDNet (Zhang et al., 2021b) proposed for partially labeled datasets. For TAL, a combination of task adaptive loss is used together with the dice loss. We trained the models for 1000 epochs using the same optimization algorithm.

We also checked whether KD loss is effective in centralized training since the forgetting issue may arise, particularly when the batch size is

Table 2
Accuracy of the proposed model against state-of-the-art methods in the federated setting using 7 clients. The best score is shown in red and the second best in blue.

Method	Avg	Liver		Kidney		Hepatic vessel		Pancreas		Colon	Lung	Spleen
	Dice (%)↑	Organ	Tumor	Organ	Tumor	Organ	Tumor	Organ	Tumor	Tumor	Tumor	Organ
DoDNet$_{FedAvg}$ (Zhang et al., 2021b)	53.20	87.50	53.91	92.47	72.61	42.48	64.20	10.91	6.22	27.88	51.20	75.77
Sep. Enc$_{FedAvg}$ (Xu et al., 2023)	66.82	87.46	54.72	92.80	72.09	58.96	68.44	77.57	45.25	37.88	55.87	84.00
Multihead$_{FedAvg}$ (McMahan et al., 2017)	66.58	93.76	60.66	93.70	71.63	52.61	63.83	75.97	42.06	34.39	50.26	93.50
Multihead$_{FedProx}$ (Li et al., 2020b)	66.62	94.42	58.05	94.56	73.51	55.88	67.86	77.52	39.51	38.08	54.67	78.78
Multihead$_{FedScaffold}$ (Karimireddy et al., 2020)	65.89	92.44	54.34	94.60	74.32	55.24	66.33	73.96	39.14	32.07	57.41	84.94
Ours ($\mathcal{L}_{KD}^{global}$)	69.55	96.21	59.76	94.98	70.48	59.20	72.19	77.71	45.62	46.78	47.38	94.69
Ours ($\mathcal{L}_{KD}^{local}$)	70.80	96.31	62.42	95.01	70.35	59.11	71.97	78.02	48.68	46.25	57.24	93.40
Ours ($\mathcal{L}_{KD}^{global}$+$\mathcal{L}_{KD}^{local}$)	71.00	96.51	62.21	95.46	75.35	59.26	71.03	78.98	49.93	41.75	55.95	94.52
Multihead - centralized setting	72.72	96.48	65.82	95.87	75.37	60.12	74.42	80.77	53.61	44.37	58.66	94.42
Method	Avg	Liver		Kidney		Hepatic vessel		Pancreas		Colon	Lung	Spleen
	HD (voxel)↓	Organ	Tumor	Organ	Tumor	Organ	Tumor	Organ	Tumor	Tumor	Tumor	Organ
DoDNet$_{FedAvg}$ (Zhang et al., 2021b)	58.21	12.52	45.26	8.18	6.01	51.60	69.67	67.03	243.42	51.61	73.19	11.79
Sep. Enc$_{FedAvg}$ (Xu et al., 2023)	21.45	19.83	67.76	7.27	13.81	9.62	25.29	8.03	21.98	29.55	18.00	14.82
Multihead$_{FedAvg}$ (McMahan et al., 2017)	23.84	5.88	34.31	6.87	10.70	14.22	39.05	8.77	46.93	33.93	60.16	1.40
Multihead$_{FedProx}$ (Li et al., 2020b)	22.19	3.93	48.34	5.91	5.00	23.60	43.24	7.85	44.94	29.22	16.02	16.04
Multihead$_{FedScaffold}$ (Karimireddy et al., 2020)	25.36	5.89	72.56	6.12	17.09	26.98	32.09	11.30	44.83	39.65	18.42	4.03
Ours ($\mathcal{L}_{KD}^{global}$)	18.36	2.67	37.58	2.89	24.57	9.47	19.50	7.79	33.68	22.41	40.03	1.40
Ours ($\mathcal{L}_{KD}^{local}$)	15.99	2.91	32.82	4.82	23.60	22.75	9.48	7.24	29.51	25.88	13.35	3.47
Ours ($\mathcal{L}_{KD}^{global}$+$\mathcal{L}_{KD}^{local}$)	16.19	1.96	35.87	4.12	4.89	17.33	9.27	6.16	43.39	22.78	31.10	1.18
Multihead - centralized setting	15.29	2.12	32.64	1.82	17.03	9.49	14.06	5.79	36.45	38.58	9.05	1.14

Table 3
Accuracy of the proposed model against state-of-the-art methods in the federated setting using 3 clients. The best score is shown in red and the second best in blue.

Method	Avg Dice (%)↑	Avg HD (voxel)↓	Liver		Kidney		Pancreas	
			Dice	HD	Dice	HD	Dice	HD
DoDNet$_{FedAvg}$ (Zhang et al., 2021b)	88.53	4.56	94.07	3.38	94.67	3.17	76.84	7.14
Multihead$_{FedAvg}$ (McMahan et al., 2017)	90.37	4.36	96.65	1.84	95.04	4.40	79.43	6.84
Sep. Enc$_{FedAvg}$ (Xu et al., 2023)	89.78	5.46	91.99	9.46	96.27	1.40	81.07	5.52
Ours ($\mathcal{L}_{KD}^{global}$)	91.48	2.67	96.57	1.92	96.16	1.49	81.72	4.58
Ours ($\mathcal{L}_{KD}^{local}$)	91.55	2.82	96.56	2.02	96.28	1.40	81.81	5.05
Ours ($\mathcal{L}_{KD}^{global}$+$\mathcal{L}_{KD}^{local}$)	91.62	2.75	96.44	1.94	96.21	1.93	82.20	4.38

Table 4
Dice similarity scores (%) for organs of Multihead$_{FedAvg}$ and DoDNet$_{FedAvg}$ after 400 rounds of communication. C*i* represents the local model in *i*th client in Table 1 and (organ) is the target organ of the dataset in C*i*. 'Global' indicates the global model's performance.

	Multihead$_{FedAvg}$					DoDNet$_{FedAvg}$				
	Avg	Liver	Kidney	Pancreas	Spleen	Avg	Liver	Kidney	Pancreas	Spleen
C0 (Liver)	39.7	96.3	21.9	40.4	0.0	29.7	96.5	22.3	0.0	0.0
C1 (Kidney)	71.8	44.4	93.9	72.1	76.9	40.2	5.6	95.0	30.0	30.3
C3 (Pancreas)	62.1	7.1	92.1	77.9	71.3	42.4	6.4	84.3	78.7	0.0
C6 (Spleen)	69.7	24.6	85.1	74.4	94.7	38.4	0.9	57.8	0.1	95.0
Global	85.4	89.9	92.9	71.8	87.2	65.2	72.3	91.0	28.5	68.9

restricted due to the high computational cost and significant memory requirements of training a segmentation model on 3D medical images. When the model is updated using a batch of data that only includes labels for a few organs, it may forget the previously learned knowledge for segmenting other organs that are not included in the batch. First, the model is updated to $f_{\theta'}$ using the ground truth for the target organ following Eq. (3). Subsequently, the model generates different predictions for all organs. To address this discrepancy, we minimize the cross-entropy between predictions of f_θ and $f_{\theta'}$ for organs not included in the training batch. When $L_{\bar{B}}$ is a set of organs that are not included in the training batch B and $N_{L_{\bar{B}}}$ is the number of elements in $L_{\bar{B}}$, knowledge distillation loss is defined as:

$$\mathcal{L}_{KD,i} = -\frac{1}{N_V}\frac{1}{N_{L_{\bar{B}}}}\sum_{l\in L_{\bar{B}}}\sum_{j=1}^{N_V} f_\theta(X_{ij}^{k_b}, l)\log(f_{\theta'}(X_{ij}^{k_b}, l)) \tag{11}$$

Table 5
P-values from paired t-tests between Dice scores from our proposed methods and comparison methods. Values below 0.05 are shown in Bold.

Method	Ours		
	$\mathcal{L}_{KD}^{global}$	$\mathcal{L}_{KD}^{local}$	$\mathcal{L}_{KD}^{global}$ + $\mathcal{L}_{KD}^{local}$
DoDNet$_{FedAvg}$	**<0.0001**	**<0.0001**	**<0.0001**
Sep. Enc$_{FedAvg}$	0.4125	**0.0088**	**0.0017**
Multihead$_{FedAvg}$	**<0.0001**	**<0.0001**	**<0.0001**
Multihead$_{FedProx}$	**0.0012**	**<0.0001**	**<0.0001**
Multihead$_{FedScaffold}$	**<0.0001**	**<0.0001**	**<0.0001**

4.2. Results

4.2.1. Comparison in federated setting

In Table 2, we present the Dice and HD scores of our method against the comparison methods in the federated setting. DoDNet with FedAvg (DoDNet$_{FedAvg}$) obtained poor accuracy even if DoDNet is considered state-of-the-art in the centralized setting. The task-specific controller

Table 6
Accuracy of the proposed model compared to state-of-the-art methods in the federated setting using 21 clients. The best score is shown in red and the second best in blue.

Method	Avg	Liver		Kidney		Hepatic vessel		Pancreas		Colon	Lung	Spleen
	Dice (%)↑	Organ	Tumor	Organ	Tumor	Organ	Tumor	Organ	Tumor	Tumor	Tumor	Organ
DoDNet$_{FedAvg}$ (Zhang et al., 2021b)	55.57	90.23	52.49	92.09	70.44	30.15	62.77	33.91	20.40	39.37	51.34	68.10
Sep. Enc$_{FedAvg}$ (Xu et al., 2023)	67.92	87.52	52.82	95.64	76.76	59.58	68.36	78.57	50.96	35.73	56.94	84.20
Multihead$_{FedAvg}$ (McMahan et al., 2017)	63.30	92.95	56.99	93.62	67.94	53.37	61.05	73.26	34.77	32.82	35.55	94.03
Multihead$_{FedProx}$ (Li et al., 2020b)	64.08	91.45	57.16	93.02	72.04	55.57	58.11	69.35	37.98	32.16	44.13	93.91
Multihead$_{FedScaffold}$ (Karimireddy et al., 2020)	63.60	93.18	52.57	92.00	65.52	53.55	58.87	71.75	37.12	31.24	49.93	93.87
Ours ($\mathcal{L}_{KD}^{global}$)	69.83	96.39	60.10	94.88	71.46	59.44	70.44	79.99	50.86	49.45	40.87	94.23
Ours ($\mathcal{L}_{KD}^{local}$)	70.05	96.19	61.31	95.49	70.74	59.42	71.13	79.27	50.17	44.05	48.16	94.67
Ours ($\mathcal{L}_{KD}^{global} + \mathcal{L}_{KD}^{local}$)	70.55	92.08	57.10	95.62	73.98	59.85	70.50	78.76	48.92	49.59	55.28	94.36
Method	Avg	Liver		Kidney		Hepatic vessel		Pancreas		Colon	Lung	Spleen
	HD (voxel)↓	Organ	Tumor	Organ	Tumor	Organ	Tumor	Organ	Tumor	Tumor	Tumor	Organ
DoDNet$_{FedAvg}$ (Zhang et al., 2021b)	46.00	11.79	46.76	7.48	18.04	69.03	50.00	45.93	175.54	36.70	33.77	10.92
Sep. Enc$_{FedAvg}$ (Xu et al., 2023)	19.73	17.22	68.05	2.75	6.43	10.54	33.61	7.44	15.58	34.75	16.05	4.65
Multihead$_{FedAvg}$ (McMahan et al., 2017)	29.57	7.44	36.18	6.97	7.84	13.73	57.61	10.95	83.47	41.42	58.20	1.42
Multihead$_{FedProx}$ (Li et al., 2020b)	32.35	7.75	42.61	10.22	29.35	14.59	60.63	15.97	63.64	39.41	69.66	2.03
Multihead$_{FedScaffold}$ (Karimireddy et al., 2020)	29.02	5.52	54.22	10.03	19.29	12.74	46.42	14.54	76.03	42.05	37.11	1.32
Ours ($\mathcal{L}_{KD}^{global}$)	17.56	2.32	48.11	5.03	7.05	9.31	28.42	5.42	21.22	28.42	35.88	2.00
Ours ($\mathcal{L}_{KD}^{local}$)	17.42	2.33	35.07	4.64	19.25	9.00	23.40	6.39	23.80	31.70	34.76	1.24
Ours ($\mathcal{L}_{KD}^{global} + \mathcal{L}_{KD}^{local}$)	15.72	9.31	56.20	2.73	4.09	8.69	23.57	7.74	17.01	23.68	18.36	1.54

used in DoDNet fails to learn multi-class segmentation by averaging the client's parameters. On the other hand, Sep. Enc$_{FedAvg}$ and our multi-head baseline method (Multihead$_{FedAvg}$) obtained substantially improved accuracy compared to DoDNet, i.e., +13.38% avg Dice and −34.37 voxel avg HD (Multihead$_{FedAvg}$). Fig. 5(A) plots the average Dice scores of comparison methods at intervals of 100 communication rounds. While Multihead$_{FedAvg}$ and DoDNet$_{FedAvg}$ show comparable accuracy in the early training stage (around 400 rounds), DoDNet's improvement becomes marginal and inconsistent after 500 rounds with a large accuracy gap against Multihead. Table 4 lists the local models' organ segmentation accuracy in both architectures after 900 rounds of communications. It appears that while local models maintain high accuracy on their target organs in both Multihead and DoDNet, but they forget the knowledge for segmenting non-target organs. For instance, in the Multihead model at client C0, Dice scores for Kidney, Pancreas, and Spleen are below 50%, even though the global model's average Dice score stands at 85.4%. Such forgetting issue is even more pronounced in DoDNet, with average scores for models C0 to C6 falling below those of Multihead models. As a result, DoDNet's accuracy substantially degrades after aggregation compared to Multihead. Our findings suggest that DoDNet is more prone to catastrophic forgetting. This may be attributed to weights predicted by the task-specific controller being updated more dynamically compared to a simple voxel classifier since divergence of local models may lead to degraded accuracy after aggregation. On the other hand, the global model of Multihead showed consistent improvement as we repeated communications between the clients and server.

Advanced FL methods such as Multihead$_{FedProx}$ (Li et al., 2020b) and Multihead$_{FedScaffold}$ (Karimireddy et al., 2020) did not show meaningful improvement as they are not designed for learning from partially labeled datasets. Multihead$_{FedScaffold}$ showed a marginal decrease in accuracy compared to Multihead$_{FedAvg}$. On the other hand, our proposed method Ours($\mathcal{L}_{KD}^{global} + \mathcal{L}_{KD}^{local}$) achieved higher accuracy than Multihead$_{FedAvg}$ methods, i.e., +4.42% avg Dice and −8.55 voxel avg HD scores, and reduced the accuracy gap between the federated and centralized settings. Compared to organs, segmenting tumors is often more challenging as their shape can vary a lot across subjects. When a client trains the model with the tumor label, KD loss prevents forgetting the knowledge for segmenting other organs, which can improve the overall accuracy of the global model as the parameters of the clients are averaged.

We also evaluated our model trained with different distillation losses. Ours($\mathcal{L}_{KD}^{global}$) and Ours($\mathcal{L}_{KD}^{local}$) were comparable to Ours($\mathcal{L}_{KD}^{global} + \mathcal{L}_{KD}^{local}$). Ours($\mathcal{L}_{KD}^{global} + \mathcal{L}_{KD}^{local}$) achieved the highest avg Dice score, while Ours($\mathcal{L}_{KD}^{local}$) achieved the best avg HD score. Distributions of the Dice scores are illustrated using box plots in Fig. 4. The variance of the organs' scores is lower than tumors as segmenting tumors is more challenging. Nevertheless, our proposed method achieved higher average Dice scores and relatively small variance with fewer outliers in most cases. We also confirmed that the distributional differences between our proposed methods and comparison methods are statistically significant, with p-values below 0.05 in all cases except one in paired t-tests as shown in Table 5.

Compared to Ours($\mathcal{L}_{KD}^{global}$), Ours($\mathcal{L}_{KD}^{local}$) is advantageous when each client has a large amount of data to train an accurate organ-specific model. Nonetheless, the precision of their organ-specific models may be compromised when clients have a restricted amount of data. To confirm this effect, we evaluated our proposed methods under conditions where each client had fewer data samples as shown in Table 6. We randomly divided each partially labeled dataset into three separate datasets, resulting in a total of 21(= 7×3) clients. Notably, Ours($\mathcal{L}_{KD}^{local}$) showed a marginal decrease in accuracy compared to the accuracy with 7 clients (Table 2). These results show that our model is robust to the number of clients. In addition, Ours($\mathcal{L}_{KD}^{global} + \mathcal{L}_{KD}^{local}$) achieved the best scores both in avg Dice and avg HD scores. These results show the advantages of employing both $\mathcal{L}_{KD}^{global}$ and $\mathcal{L}_{KD}^{local}$ when the data is distributed across a larger number of sites.

Table 7 summarizes the accuracy of the different methods on the external BTCV dataset using a total of 7 and 21 clients. DoDNet$_{FedAvg}$, Multihead$_{FedAvg}$, Multihead$_{FedProx}$ and Multihead$_{FedScaffold}$ received low accuracy scores. Interestingly, Sep. Enc$_{FedAvg}$ obtained comparatively better accuracy than other comparison methods, although it is limited compared to Ours. The difference in results between Tables 2 and 7 may be attributed to the varying distribution of the datasets. Especially, the overall accuracies of the models for kidney segmentation were notably lower compared to those in Tables 2 and 6 since the CT scans in the training dataset (Heller et al., 2019) are contrast-enhanced, whereas the CT scans in BTCV dataset are not. Despite this gap, Our proposed methods showed substantial improvements over comparison methods, even with a larger number of clients, i.e., 21 clients. When utilizing $\mathcal{L}_{KD}^{global} + \mathcal{L}_{KD}^{local}$, we obtained a better avg Dice score, but a marginal decrease in avg HD score compared to Ours($\mathcal{L}_{KD}^{global}$). However, it is noteworthy that Ours($\mathcal{L}_{KD}^{global} + \mathcal{L}_{KD}^{local}$) trained under 21 clients setting achieved a better Avg HD score than Ours($\mathcal{L}_{KD}^{global}$) when excluding the kidney score, i.e., 12.88 vs. 13.44. Considering a huge gap between the distributions of the datasets, employing a combination of $\mathcal{L}_{KD}^{global}$ and

Table 7
Comparison of the proposed federated learning model against state-of-the-art methods on an external dataset (BTCV). The best score is shown in red and the second best in blue.

Method	Avg		Liver		Kidney		Pancreas		Spleen	
	Dice (%)↑	HD (voxel)↓	Dice	HD	Dice	HD	Dice	HD	Dice	HD
7 Clients			Organ	Organ	Organ	Organ	Organ	Organ	Organ	Organ
DoDNet$_{FedAvg}$ (Zhang et al., 2021b)	30.37	60.14	68.96	37.22	0.01	66.78	17.83	70.29	34.69	66.29
Sep. Enc$_{FedAvg}$ (Xu et al., 2023)	72.41	25.28	93.54	7.72	56.36	42.28	64.13	25.65	75.59	25.45
Multihead$_{FedAvg}$ (McMahan et al., 2017)	49.22	37.75	78.86	18.84	0.01	58.91	53.38	38.28	64.61	34.98
Multihead$_{FedProx}$ (Li et al., 2020b)	53.04	35.86	80.83	22.57	0.00	67.74	65.43	17.51	65.89	35.63
Multihead$_{FedScaffold}$ (Karimireddy et al., 2020)	51.76	36.89	75.67	25.41	0.00	73.41	57.30	30.07	74.08	18.66
Ours ($\mathcal{L}_{KD}^{global}$)	77.53	15.09	93.61	6.00	63.56	22.95	66.40	20.62	86.54	10.80
Ours ($\mathcal{L}_{KD}^{local}$)	76.22	16.17	93.80	7.31	61.09	28.43	64.45	17.88	85.54	11.09
Ours ($\mathcal{L}_{KD}^{global}$+$\mathcal{L}_{KD}^{local}$)	76.66	15.78	93.89	5.69	62.05	25.69	67.23	18.61	83.48	13.15
21 Clients										
Ours ($\mathcal{L}_{KD}^{global}$)	75.97	14.39	92.71	7.83	61.79	18.23	63.65	23.21	85.72	8.29
Ours ($\mathcal{L}_{KD}^{local}$)	75.40	19.93	93.11	8.17	57.01	34.51	64.57	26.19	86.92	10.84
Ours ($\mathcal{L}_{KD}^{global}$+$\mathcal{L}_{KD}^{local}$)	76.60	18.26	93.36	9.08	59.04	34.37	66.59	19.29	87.39	10.28

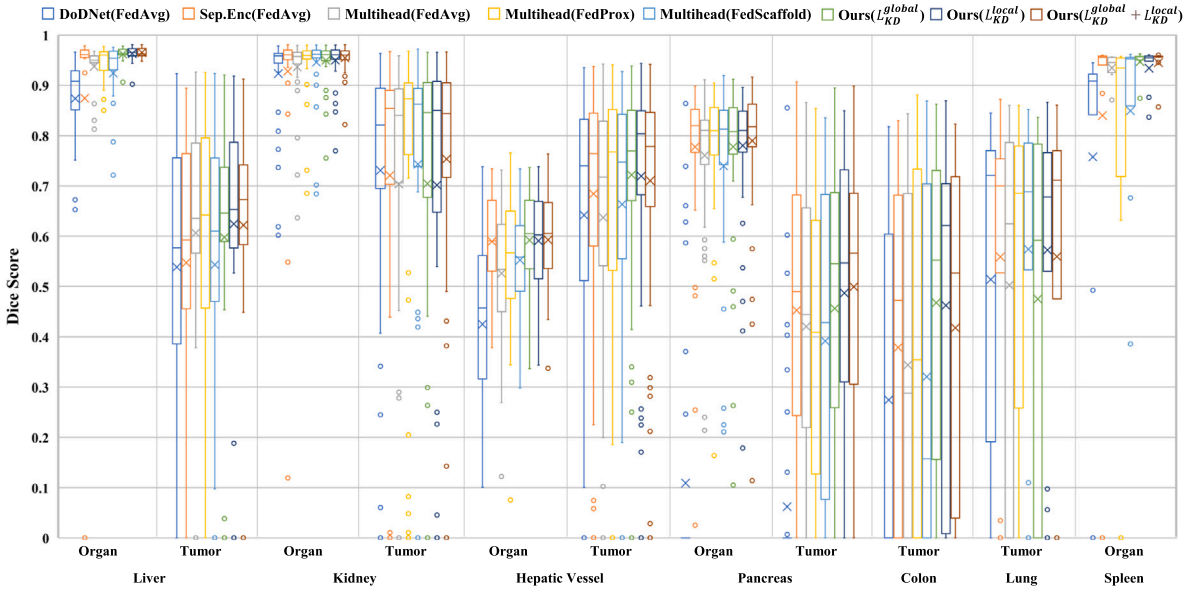

Fig. 4. Box plots of dice similarity scores from different comparison methods. X, -, and o denote average, median and outlier, respectively.

$\mathcal{L}_{KD}^{local}$ still proves beneficial. These results demonstrate the robustness of our proposed methods to different data distributions and the number of clients.

We further investigated the applicability of our proposed method in a simplified setting, federating 3 clients for liver, kidney, and pancreas segmentation. Table 3 lists the accuracy of comparison methods after 1000 rounds of communications. Our findings align with the original experiment: DoDNet$_{FedAvg}$ achieves the lowest accuracy followed by Multihead$_{FedAvg}$ and Sep. Encoder$_{FedAvg}$ yields better. A substantial improvement (i.e., +1.25 avg Dice and −1.61 avg HD score) was achieved by our proposed KD. These consistent improvements across different client numbers underscore the effectiveness of our KD approach in dealing with partially labeled datasets.

In Fig. 8, we visualize the segmentation results of comparison methods trained in the federated setting in 3D. The comparison methods often obtained noisy organ segmentation or inaccurate tumor segmentation results, whereas our proposed method obtained accurate segmentation results in most cases. These results show that our proposed KD-based methods with Multihead architecture effectively regularize the federated learning process. In Fig. 7, we present 2D visualization of multi-organ segmentation by aggregating multiple predictions. However, a potential problem is that the predicted segmentation masks by different heads are not necessarily exclusive to each other, *e.g.,* liver and hepatic vessels. In this case, multiple segmentations can be integrated using anatomical knowledge. For instance, a voxel classified as hepatic vessel and liver should be determined as hepatic vessel because most hepatic vessels exist inside the liver. On the other hand, if a voxel belongs to two different organs exclusive to each other, we select the class with a higher probability. Aggregated segmentation results reveal that noisy predictions from comparison methods as shown in Fig. 8 lead to subpar multi-organ segmentation. Notably, our proposed method demonstrates accurate multi-organ segmentation (*e.g.*, Fig. 7, 2nd row (B), unlike competing methods that misclassify organs like kidney (*e.g.*, Fig. 7, 2nd row (C)–(G)). These results show that our proposed methods are also superior in multi-organ segmentation.

4.2.2. Comparison in centralized setting

In Table 8, we present Dice and HD scores of our method against the reproduced state-of-the-art methods in the centralized setting. TAL and Conditional models such as Cond_Enc and Cond_Dec were of lower accuracy than Multiple Nets. This result implies that we should carefully design the model architecture to improve accuracy by sharing some parts of the model. The accuracy of Multiple Nets is limited in Colon, Lung, and Spleen which have a relatively small number of data samples.

Table 8
Comparison of the proposed model against state-of-the-art methods in the centralized setting.

Method	Avg	Liver		Kidney		Hepatic vessel		Pancreas		Colon	Lung	Spleen
	Dice (%)↑	Organ	Tumor	Organ	Tumor	Organ	Tumor	Organ	Tumor	Tumor	Tumor	Organ
Multiple Net	69.70	96.16	62.60	95.67	78.18	61.04	69.72	81.61	52.81	21.74	53.74	93.38
Sep. Enc (Xu et al., 2023)	69.77	91.59	57.78	94.79	77.46	59.97	73.06	81.71	52.50	40.46	45.78	92.32
Cond_Enc (Dmitriev and Kaufman, 2019)	67.19	93.56	53.98	94.24	76.09	50.89	66.55	77.49	49.57	33.11	54.03	89.58
Cond_Dec (Dmitriev and Kaufman, 2019)	54.21	27.58	29.77	92.57	74.75	54.95	73.76	71.08	46.31	38.99	32.13	54.46
TAL (Fang and Yan, 2020)	69.20	95.85	37.66	94.99	72.99	59.68	70.67	81.02	50.18	47.55	55.77	94.88
DoDNet (Zhang et al., 2021b)	72.92	96.35	61.66	96.29	79.05	60.34	71.89	82.05	56.20	49.19	55.92	93.17
Multihead	72.72	96.48	65.82	95.87	75.37	60.12	74.42	80.77	53.61	44.37	58.66	94.42
Ours	73.79	96.58	64.23	96.01	79.29	60.48	72.53	79.86	54.39	51.86	62.23	94.28
Method	Avg	Liver		Kidney		Hepatic vessel		Pancreas		Colon	Lung	Spleen
	HD (voxel)↓	Organ	Tumor	Organ	Tumor	Organ	Tumor	Organ	Tumor	Tumor	Tumor	Organ
Multiple Net	15.00	3.66	45.23	5.93	9.72	10.37	25.59	6.55	14.05	59.47	21.83	1.92
Sep. Enc (Xu et al., 2023)	16.00	2.99	33.92	2.74	12.96	9.87	27.85	7.54	27.59	19.84	10.26	4.46
Cond_Enc (Dmitriev and Kaufman, 2019)	22.00	5.28	37.07	4.22	4.37	12.93	38.87	6.24	47.39	49.54	31.21	4.85
Cond_Dec (Dmitriev and Kaufman, 2019)	46.37	36.67	163.36	9.61	22.04	12.29	15.38	15.03	36.87	27.75	81.35	89.72
TAL (Fang and Yan, 2020)	17.96	5.92	79.14	3.38	6.92	9.79	23.43	6.23	23.94	27.78	10.01	1.05
DoDNet (Zhang et al., 2021b)	16.44	2.13	50.14	1.53	11.99	10.25	29.95	4.88	17.13	29.59	21.42	1.83
Multihead	15.29	2.12	32.64	1.82	17.03	9.49	14.06	5.79	36.45	38.58	9.05	1.14
Ours	13.00	1.99	32.92	1.74	11.96	8.87	26.85	6.54	26.59	16.84	7.26	1.46

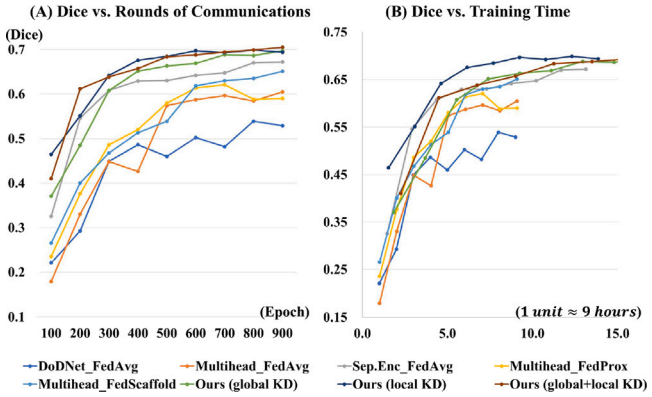

Fig. 5. (A) Average dice similarity score *vs.* rounds of communication. (B) Average dice similarity score *vs.* training time. The unit of training time in (B) is the time taken for 100 rounds in DoDNet$_{FedAvg}$, *i.e.* about 9 h.

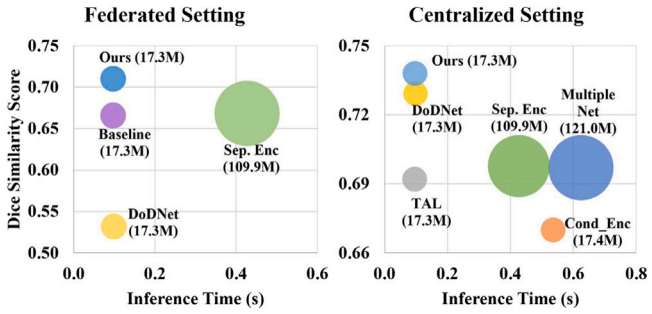

Fig. 6. Dice similarity score vs. inference time for comparison methods. The size of the circle is proportional to the number of parameters also shown in parentheses. 'M' denotes million. We measured inference time to obtain predictions for all possible tasks, i.e., seven tasks in our experiments, from a single 3D patch.

It may be attributed to the fact that Multiple Nets are trained using only limited data without utilizing data from other tasks. On the other hand, Sep. Enc and DoDNet showed improved accuracy compared to Multiple Nets. The accuracy of the multi-head baseline method was on par with DoDNet. Task-specific controller used in DoDNet was not critical for improving accuracy. On the other hand, our proposed method showed significantly improved accuracy compared to the baseline method and achieved the best Dice and HD scores among all comparison methods. This result shows that the knowledge distillation loss is essential in learning from partially labeled datasets.

Table 9
Speed and computational cost of comparison methods. We measured model updating time for a single iteration as train time, and inference time to obtain predictions for all possible tasks as test time. Multihead$_{FedProx}$ and Multhead$_{FedScaffold}$ have almost same speed and cost with Multihead$_{FedAvg}$.

Method	Time (s)		Cost (GFLOPs)	
	Train	Test	Train	Test
DoDNet$_{FedAvg}$	0.280	0.098	458.2	458.4
Sep. Enc$_{FedAvg}$	0.409	0.427	458.5	3219.1
Multihead$_{FedAvg}$	0.283	0.097	458.5	461.0
Ours($\mathcal{L}_{KD}^{global}$)	0.518	0.097	919.5	461.0
Ours($\mathcal{L}_{KD}^{local}$)	0.431	0.097	917.1	461.0
Ours($\mathcal{L}_{KD}^{global} + \mathcal{L}_{KD}^{local}$)	0.631	0.097	1378.1	461.0

4.2.3. Speed, computational cost and the number of parameters

Table 9 presents the speed and computational costs for training and testing various comparison methods in the federated setting. DoDNet$_{FedAvg}$ and Multihead$_{FedAvg}$ exhibited the highest speed and lowest computational cost, while Sep. Enc$_{FedAvg}$ had the slowest speed and incurs the highest computation. Introducing KD losses increased training time and computation, with 461.0 GFLOPs for $\mathcal{L}_{KD}^{global}$ and 458.5 GFLOPs for $\mathcal{L}_{KD}^{local}$ when predicting unlabeled organ segmentation. Nevertheless, once trained, the model has fast inference with low computation. Fig. 5(A) and (B) depict the average dice similarity score w.r.t. rounds of communication and training time, respectively. (A) shows that our model's improvement per communication round of our models is larger than other methods. In addition, our method can achieve the best accuracy within same training time as shown in (B).

In Fig. 6, we show the dice similarity score with the inference time and the number of parameters of comparison methods in the federated and centralized settings. Multiple Nets and Cond_Enc require the longest time for inference in the centralized setting because they need to repeat the entire neural network feed-forward process to get a prediction for each task, and their accuracy is limited. Since Sep. Enc

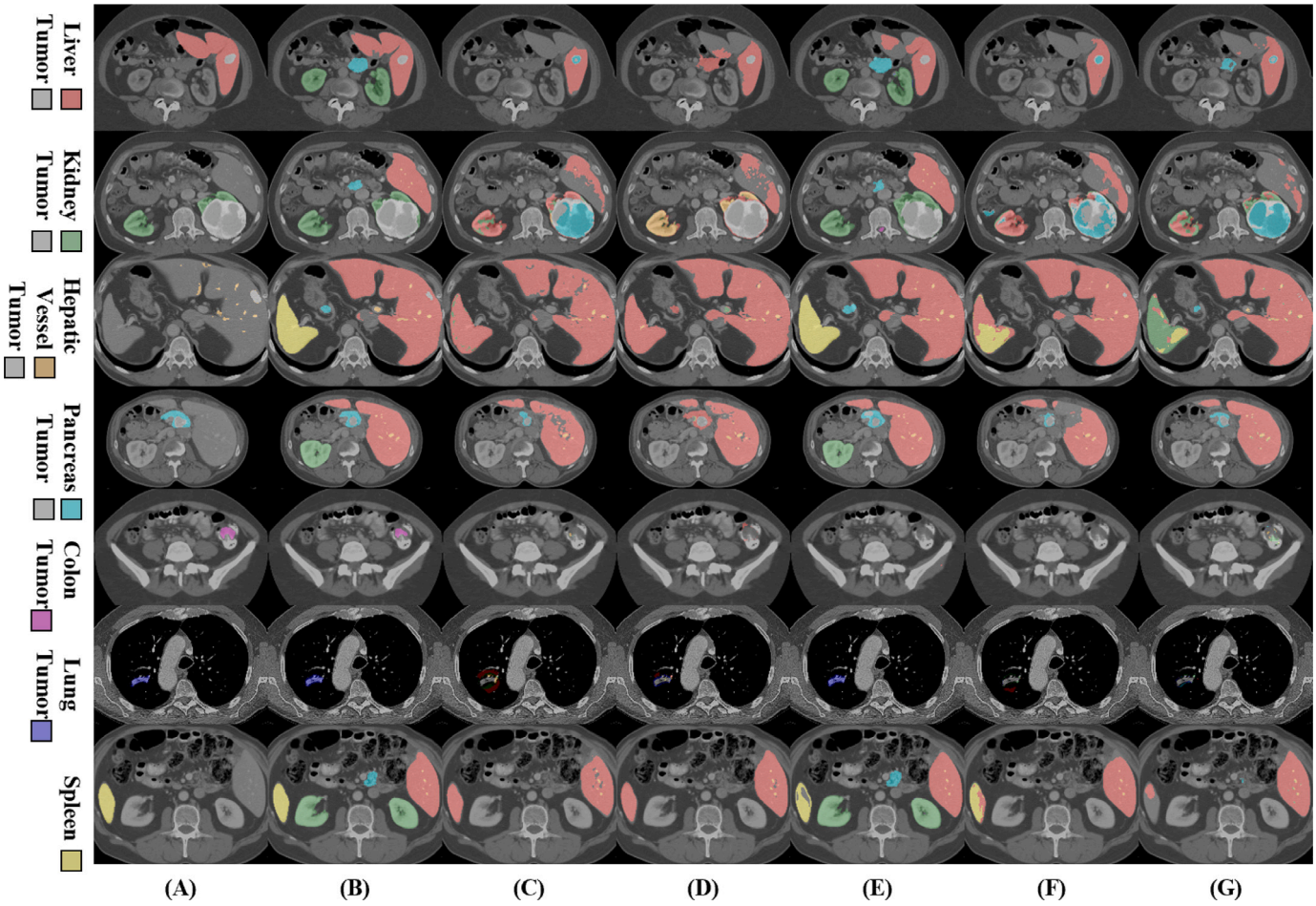

Fig. 7. 2D visualization of multi-organ segmentation with the input image and ground truth. Each column shows (A) Ground truth, (B) Ours ($\mathcal{L}_{KD}^{global} + \mathcal{L}_{KD}^{local}$), (C) Multihead$_{FedAvg}$, (D) DoDNet$_{FedAvg}$, (E) Sep. Enc$_{FedAvg}$, (F) Multihead$_{FedScaffold}$, (G) Multihead$_{FedProx}$. ROI of images is cropped for better visibility.

shares the decoder, it showed a shorter inference time, but accuracy is still limited. The models sharing the encoder and decoder of U-Net such as DoDNet, TAL, and Ours obtained the shortest inference time. However, the accuracy of DoDNet was substantially degraded in the federated setting. On the other hand, our proposed method achieved the best Dice score in both federated and centralized settings. Even though 0.62 s inference time of Multiple Nets for a single 3D patch seems already short, obtaining a prediction for an entire 3D volume will take much longer as we use a sliding window strategy for inference. In our setting, about 50 patches were extracted from one data sample. The number of repetitions will be greater with higher resolution.

Our proposed method's learning speed correlates with the number of models sampled for KD in Eq. (9). We assess Ours($\mathcal{L}_{KD}^{local}$) accuracy with varying sampling numbers in Table 11. The consistent accuracy across all cases indicates insensitivity to the sampled models, allowing us to enhance training speed and reduce computational costs for repeatedly predicting unlabeled organ segmentations.

4.2.4. Ablation study on different communication frequencies

Communication frequency is an important factor in practical federated learning since repeatedly sending and receiving parameters of the model induce a network bottleneck. In Table 10, the segmentation accuracy of Multihead$_{FedAvg}$ and Ours($\mathcal{L}_{KD}^{global}$) is shown with different communication frequencies. The number of iterations denotes the number of model updates during the local training step in each client. We set the total number of updates as 120,000 and gradually doubled the communication frequency. Our proposed method is successfully trained on various communication (round × iteration) combinations and outperforms Multihead$_{FedAvg}$ consistently. Though we obtained better accuracy in most cases with more frequent communication, accuracy saturates with more than 1000 rounds of communication.

Table 10
Accuracy of $FedAvg$ and Ours ($\mathcal{L}_{KD}^{global}$) with different communication frequencies.

Communication frequency	$FedAvg$		Ours ($\mathcal{L}_{KD}^{global}$)	
Round × Iteration	Dice (%)↑	HD (voxel)↓	Dice (%)↑	HD (voxel)↓
250 × 480	56.27	40.39	67.27	22.03
500 × 240	57.40	40.37	69.33	18.71
1000 × 120	67.14	20.79	70.73	15.63
2000 × 60	67.05	20.93	70.84	14.27
4000 × 30	67.57	19.56	70.22	16.03

Table 11
Accuracy of Ours ($\mathcal{L}_{KD}^{local}$) with different numbers of organ-specific models $h_{\hat{\theta}_i}$ sampled for KD in Eq. (9). Eqs. (9) and (10) sample 7 and 1 models, respectively.

# of sampled local models	Avg	
	Dice (%)↑	HD (voxel)↓
1	70.80	15.99
3	70.02	14.15
5	70.50	16.77
7	70.69	13.22

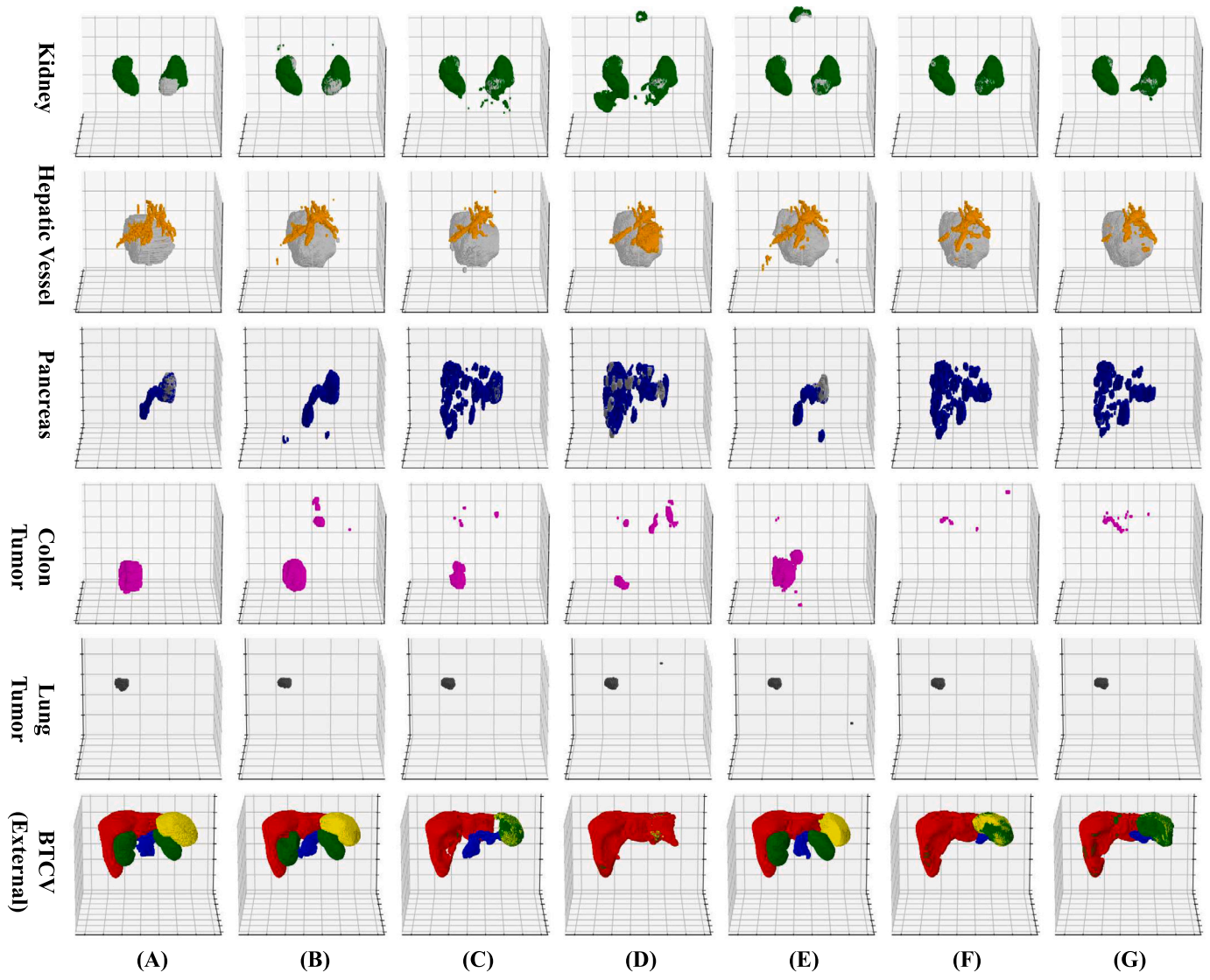

Fig. 8. 3D Visualization of model predictions with the input image and ground truth. Each column shows (A) ground truth, (B) Ours ($\mathcal{L}_{KD}^{global}$), (C) Multihead$_{FedAvg}$, (D) DoDNet$_{FedAvg}$, (E) Sep. Enc$_{FedAvg}$, (F) Multihead$_{FedScaffold}$, (G) Multihead$_{FedProx}$. In Kidney, Hepatic Vessel and Pancreas, tumors are shown as gray. This figure contains different examples with Fig. 7.

5. Conclusion

In this paper, we proposed a federated learning method for multi-class segmentation from partially labeled datasets, which leverages knowledge distillation. Our approach introduced global and local knowledge distillation losses, resulting in improved accuracy by utilizing the knowledge of the global model and pre-trained organ-specific segmentation models. In addition, a multi-head U-Net architecture was designed to have a short inference time with a relatively small number of parameters by sharing most parts of the encoder and decoder. Extensive experiments on internal and external datasets verified the effectiveness and robustness of our proposed method over the state-of-the-art methods. Although our model improves performance, it requires substantial communication between the central server and clients which can be a bottleneck of training depending on the network infrastructure of the nodes. In future work, we will design a segmentation model able to learn within a few communication rounds.

CRediT authorship contribution statement

Soopil Kim: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Data curation, Conceptualization. **Heejung Park:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology. **Myeongkyun Kang:** Writing – review & editing, Writing – original draft. **Kyong Hwan Jin:** Writing – review & editing, Writing – original draft, Supervision. **Ehsan Adeli:** Writing – review & editing, Writing – original draft, Supervision. **Kilian M. Pohl:** Writing – review & editing, Writing – original draft, Supervision, Funding acquisition. **Sang Hyun Park:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Sang Hyun Park reports financial support was provided by Korea Ministry of Science and ICT, and Daegu Metropolitan City.

Data availability

We used public data and will release the code on a github repository.

Acknowledgments

This work was supported by funding from the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) (No. 2019R1C1C1008727 and 2022K2A9A1A01097840) and the DGIST R&D program of the Ministry of Science and ICT of KOREA (22-KUJoint-02) and the Digital Innovation Hub project supervised by the Daegu Digital Innovation Promotion Agency (DIP) grant funded by the Korea government (MSIT and Daegu Metropolitan City) in 2023 (DBSD1-01) and the Stanford University Human-Centered Artificial Intelligence (HAI) Google Cloud Credits Award.

References

Acar, D.A.E., Zhao, Y., Navarro, R.M., Mattina, M., Whatmough, P.N., Saligrama, V., 2021. Federated learning based on dynamic regularization. arXiv preprint arXiv:2111.04263.

Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G.E.H., Chartrand, G., et al., 2023. The liver tumor segmentation benchmark (lits). Med. Image Anal. 84, 102680.

Chen, H.-Y., Chao, W.-L., 2020. FedBE: Making bayesian model ensemble applicable to federated learning. arXiv:2009.01974.

Chen, S., Ma, K., Zheng, Y., 2019. Med3D: Transfer learning for 3D medical image analysis. arXiv:1904.00625.

Dmitriev, K., Kaufman, A.E., 2019. Learning multi-class segmentations from single-class datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9501–9511.

Duarte, K., Rawat, Y., Shah, M., 2021. PLM: Partial label masking for imbalanced multi-label classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2739–2748.

Durand, T., Mehrasa, N., Mori, G., 2019. Learning a deep convnet for multi-label classification with partial labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 647–657.

Elskhawy, A., Lisowska, A., Keicher, M., Henry, J., Thomson, P., Navab, N., 2020. Continual class incremental learning for ct thoracic segmentation. In: Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning: Second MICCAI Workshop, DART 2020, and First MICCAI Workshop, DCL 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 2. Springer, pp. 106–116.

Fang, X., Yan, P., 2020. Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction. IEEE Trans. Med. Imaging 39 (11), 3619–3629.

Feng, L., Song, J.H., Kim, J., Jeong, S., Park, J.S., Kim, J., 2019. Robust nucleus detection with partially labeled exemplars. IEEE Access 7, 162169–162178.

Fidon, L., Aertsen, M., Emam, D., Mufti, N., Guffens, F., Deprest, T., Demaerel, P., David, A.L., Melbourne, A., Ourselin, S., et al., 2021. Label-set loss functions for partial supervision: application to fetal brain 3D MRI parcellation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24. Springer, pp. 647–657.

Gibson, E., Giganti, F., Hu, Y., Bonmati, E., Bandula, S., Gurusamy, K., Davidson, B., Pereira, S.P., Clarkson, M.J., Barratt, D.C., 2018. Automatic multi-organ segmentation on abdominal CT with dense V-networks. IEEE Trans. Med. Imaging 37 (8), 1822–1834.

Gou, J., Yu, B., Maybank, S.J., Tao, D., 2021. Knowledge distillation: A survey. Int. J. Comput. Vis. 129 (6), 1789–1819.

He, X., Zemel, R., 2008. Learning hybrid models for image annotation with partially labeled data. Adv. Neural Inf. Process. Syst. 21.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1026–1034.

Heller, N., Sathianathen, N., Kalapara, A., Walczak, E., Moore, K., Kaluzniak, H., Rosenberg, J., Blake, P., Rengel, Z., Oestreich, M., et al., 2019. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. arXiv preprint arXiv:1904.00445.

Hu, P., Wu, F., Peng, J., Liang, P., Kong, D., 2016. Automatic 3D liver segmentation based on deep learning and globally optimized surface evolution. Phys. Med. Biol. 61 (24), 8676.

Kang, M., Kim, S., Jin, K.H., Adeli, E., Pohl, K.M., Park, S.H., 2024. FedNN: Federated learning on concept drift data using weight and adaptive group normalizations. Pattern Recognit. 149, 110230.

Karimireddy, S.P., Kale, S., Mohri, M., Reddi, S., Stich, S., Suresh, A.T., 2020. Scaffold: Stochastic controlled averaging for federated learning. In: International Conference on Machine Learning. PMLR, pp. 5132–5143.

Kim, S., An, S., Chikontwe, P., Park, S.H., 2021. Bidirectional rnn-based few shot learning for 3d medical image segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35, pp. 1808–1816.

Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A., 2015. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In: Proc. MICCAI Multi-Atlas Labeling beyond Cranial Vault—Workshop Challenge. Vol. 5.

Lee, G., Jeong, M., Shin, Y., Bae, S., Yun, S.-Y., 2022. Preservation of the global knowledge by not-true distillation in federated learning. Adv. Neural Inf. Process. Syst. 35, 38461–38474.

Li, L., Fan, Y., Tse, M., Lin, K.-Y., 2020a. A review of applications in federated learning. Comput. Ind. Eng. 149, 106854.

Li, Z., Hoiem, D., 2017. Learning without forgetting. IEEE Trans. Pattern Anal. Mach. Intell. 40 (12), 2935–2947.

Li, W., Milletarì, F., Xu, D., Rieke, N., Hancox, J., Zhu, W., Baust, M., Cheng, Y., Ourselin, S., Cardoso, M.J., et al., 2019. Privacy-preserving federated brain tumour segmentation. In: Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 10. Springer, pp. 133–141.

Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V., 2020b. Federated optimization in heterogeneous networks. Proc. Mach. Learn. Syst. 2, 429–450.

Li, D., Wang, J., 2019. FedMD: Heterogenous federated learning via model distillation. arXiv:1910.03581.

Lu, M.Y., Chen, R.J., Kong, D., Lipkova, J., Singh, R., Williamson, D.F., Chen, T.Y., Mahmood, F., 2022. Federated learning for computational pathology on gigapixel whole slide images. Med. Image Anal. 76, 102298.

Ma, J., Zhang, Y., Gu, S., An, X., Wang, Z., Ge, C., Wang, C., Zhang, F., Wang, Y., Xu, Y., et al., 2022. Fast and low-GPU-memory abdomen CT organ segmentation: the flare challenge. Med. Image Anal. 82, 102616.

McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A., 2017. Communication-efficient learning of deep networks from decentralized data. In: Artificial Intelligence and Statistics. PMLR, pp. 1273–1282.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer, pp. 234–241.

Sheller, M.J., Reina, G.A., Edwards, B., Martin, J., Bakas, S., 2018. Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. In: International MICCAI Brainlesion Workshop. Springer, pp. 92–104.

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., 2021. 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. Springer, pp. 101–110.

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., 2022. Joint multi organ and tumor segmentation from partial labels using federated learning. In: International Workshop on Distributed, Collaborative, and Federated Learning. Springer, pp. 58–67.

Shi, G., Xiao, L., Chen, Y., Zhou, S.K., 2021. Marginal loss and exclusion loss for partially supervised multi-organ segmentation. Med. Image Anal. 70, 101979.

Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., et al., 2019. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063.

Verbeek, J., Triggs, B., 2008. Scene segmentation with conditional random fields learned from partially labeled images. In: Proc. NIPS.

Vu, M.H., et al., 2021. A data-adaptive loss function for incomplete data and incremental learning in semantic image segmentation. IEEE Trans. Med. Imaging 41 (6), 1320–1330.

Wang, P., Shen, C., Roth, H.R., Yang, D., Xu, D., Oda, M., Misawa, K., Chen, P.-T., Liu, K.-L., Liao, W.-C., et al., 2020. Automated pancreas segmentation using multi-institutional collaborative deep learning. In: Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning. Springer, pp. 192–200.

Wicaksana, J., Yan, Z., Zhang, D., Huang, X., Wu, H., Yang, X., Cheng, K.-T., 2022. FedMix: Mixed supervised federated learning for medical image segmentation. IEEE Trans. Med. Imaging.

Wu, H., Pang, S., Sowmya, A., 2022. TGNet: A task-guided network architecture for multi-organ and tumour segmentation from partially labelled datasets. In: 2022 IEEE 19th International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 1–5.

Xia, Y., Yang, D., Li, W., Myronenko, A., Xu, D., Obinata, H., Mori, H., An, P., Harmon, S., Turkbey, E., et al., 2021. Auto-FedAvg: learnable federated averaging for multi-institutional medical image segmentation. arXiv:2104.10195.

Xiao, J.-W., Zhang, C.-B., Feng, J., Liu, X., van de Weijer, J., Cheng, M.-M., 2023. Endpoints weight fusion for class incremental semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7204–7213.

Xu, X., Deng, H.H., Gateno, J., Yan, P., 2023. Federated multi-organ segmentation with inconsistent labels. IEEE Trans. Med. Imaging.

Yan, K., Cai, J., Zheng, Y., Harrison, A.P., Jin, D., Tang, Y., Tang, Y., Huang, L., Xiao, J., Lu, L., 2020. Learning from multiple datasets with heterogeneous and partial labels for universal lesion detection in CT. IEEE Trans. Med. Imaging 40 (10), 2759–2770.

Yang, D., Xu, Z., Li, W., Myronenko, A., Roth, H.R., Harmon, S., Xu, S., Turkbey, B., Turkbey, E., Wang, X., et al., 2021. Federated semi-supervised learning for COVID region segmentation in chest CT using multi-national data from China, Italy, Japan. Med. Image Anal. 70, 101992.

Yu, Q., Yang, D., Roth, H., Bai, Y., Zhang, Y., Yuille, A.L., Xu, D., 2020. C2FNAS: Coarse-to-fine neural architecture search for 3D medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4126–4135.

Zhang, L., Shen, L., Ding, L., Tao, D., Duan, L.-Y., 2022. Fine-tuning global model via data-free knowledge distillation for non-iid federated learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10174–10183.

Zhang, J., Xie, Y., Xia, Y., Shen, C., 2021b. 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.

Zhang, G., Yang, Z., Huo, B., Chai, S., Jiang, S., 2021a. Multiorgan segmentation from partially labeled datasets with conditional nnU-Net. Comput. Biol. Med. 136, 104658.

Zhou, Y., Li, Z., Bai, S., Wang, C., Chen, X., Han, M., Fishman, E., Yuille, A.L., 2019. Prior-aware neural network for partially-supervised multi-organ segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10672–10681.