Federated Learning via Input-Output Collaborative Distillation

Xuan Gong[1,3 *], Shanglin Li[2 *], Yuxiang Bao[2 *], Barry Yao[1,4], Yawen Huang[5], Ziyan Wu[6], Baochang Zhang[2,7,8 †], Yefeng Zheng[5], David Doermann[1 †]

[1] University at Buffalo, Buffalo, NY, USA
[2] Institute of Artificial Intelligence, Hangzhou Research Institute, Beihang University, Beijing, China
[3] Harvard Medical School, Boston, MA, USA
[4] Virginia Tech, Blacksburg, VA, USA
[5] Jarvis Research Center, Tencent YouTu Lab, Shenzhen, China
[6] United Imaging Intelligence, Burlington, MA, USA
[7] Zhongguancun Laboratory, Beijing, China
[8] Nanchang Institute of Technology, Nanchang, China
xuangong@buffalo.edu, shanglin@buaa.edu.cn, yxbao@buaa.edu.cn, barryyao@vt.edu, yawenhuang@tencent.com, ziyan.wu@uii-ai.com, bczhang@buaa.edu.cn, yefengzheng@tencent.com, doermann@buffalo.edu

Abstract

Federated learning (FL) is a machine learning paradigm in which distributed local nodes collaboratively train a central model without sharing individually held private data. Existing FL methods either iteratively share local model parameters or deploy co-distillation. However, the former is highly susceptible to private data leakage, and the latter design relies on the prerequisites of task-relevant real data. Instead, we propose a data-free FL framework based on local-to-central collaborative distillation with direct input and output space exploitation. Our design eliminates any requirement of recursive local parameter exchange or auxiliary task-relevant data to transfer knowledge, thereby giving direct privacy control to local users. In particular, to cope with the inherent data heterogeneity across locals, our technique learns to distill input on which each local model produces consensual yet unique results to represent each expertise. Our proposed FL framework achieves notable privacy-utility trade-offs with extensive experiments on image classification and segmentation tasks under various real-world heterogeneous federated learning settings on both natural and medical images. Code is available at https://github.com/lsl001006/FedIOD.

Introduction

The recent success of deep learning in various applications can be attributed to data-driven algorithms typically trained in a centralized fashion, *i.e.*, computational units and data samples residing on the same server. Real-world scenarios, however, tend to disperse this wealth of data throughout separate locations and governed by diverse entities. Due to privacy regulations and communication limitations, collecting all data in one location for centralized training is often impractical, especially true for mobile vision and medical applications.

[*] These authors contributed equally.
[†] Corresponding.

Accordingly, federated learning (FL) does not necessarily need all data samples to be centralized; instead, it relies on model fusion/distillation techniques to train one centralized model in a decentralized fashion. Privacy is a critical consideration, and it is vital to prevent private data leakage. Another challenge is data heterogeneity among locals, as distributed data centers tend to collect data in different settings.

Most federated learning methods are based on the recursive exchange of local model parameters during the training process (McMahan et al. 2017; Li et al. 2018; Karimireddy et al. 2020). Each local node uploads its model parameters after a particular time of local training. The central server aggregates the parameters of the local model with different schemes (Wang et al. 2020; Li et al. 2020; Hsu, Qi, and Brown 2020) and then distributes the aggregated parameters. Each local node receives the latest parameters to update its local model accordingly and continues with the next round of local training. However, naively employing such iterative parameter exchange suffers from known weaknesses: (1) All participating models must have exactly homogeneous architectures. (2) Iteratively sharing the model parameters opens all internal states of the model to white-box inference attacks, resulting in significant privacy leakage (Chang et al. 2019). Recent works (Zhu, Liu, and Han 2019; Geiping et al. 2020) obtain private training data from publicly shared model gradients.

Distillation-based methods are proposed to train the central model with aggregated locally-computed logits (Li and Wang 2019; Lin et al. 2020; Gong et al. 2022a), eliminating the requirement of identical network architectures. However, to transfer knowledge, additional public data are commonly assumed to be accessible and sampled from the same underlying distribution as the privately held local data. This assumption can be strong in practice and unavoidably exposes private data to stealthy attacks. Although (Zhu, Hong, and Zhou 2021; Zhang, Wu, and Yuan 2022; Zhang et al. 2022) takes a step further to eliminate the requirement of real data for distillation, iterative model parameter exchange is still

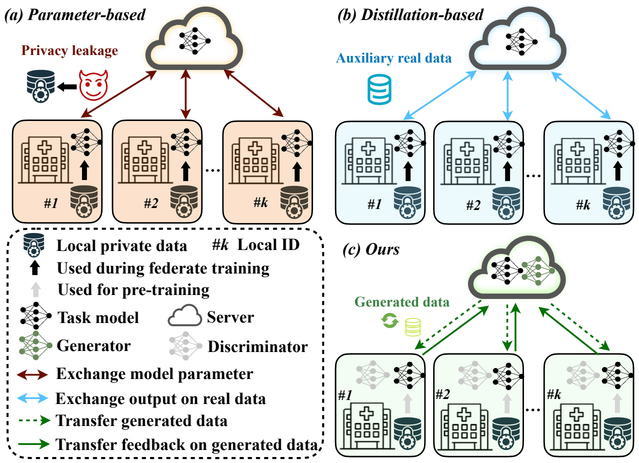

Figure 1: (a) Parameter-based methods recursively exchange model parameters between each local and server-side (McMahan et al. 2017; Li et al. 2018; Karimireddy et al. 2020), which is highly vulnerable to a security attack (Zhu, Liu, and Han 2019). (b) Distillation-based methods utilize auxiliary task-dependent real data to conduct co-distillation between each local and the central server (Li and Wang 2019; Gong et al. 2022a). (c) Our FL method conducts one-way distillation from locals to the server with generated data, eliminating the prerequisite of additional data required by typical distillation, and the security weaknesses of white-box attacks caused by recursive parameter exchange.

essential in these frameworks where knowledge transfer is only an auxiliary module for fine-tuning. As noted above, such parameter exchange is limited by identical model architecture and, more importantly, highly susceptible to privacy leakage. These methods require such recursive parameter exchange primarily because they mainly focus on the output distillation, leaving the input space under-explored.

In this paper, we propose a new federated learning framework (FedIOD) that conducts a collaborative knowledge distillation in both the input and output space (as Figure 1). It is purely based on data-free distillation without any prerequisite of auxiliary real data or locally trained model parameters. Besides, we adopt differential privacy protection on the gradients used to train the generator (Torkzadehmahani, Kairouz, and Paten 2019; Chen, Orekondy, and Fritz 2020). This, by design, gives explicit privacy control to each local node. Unlike the previous data-free federated distillation counterparts (Zhu, Hong, and Zhou 2021; Zhang, Wu, and Yuan 2022; Zhang et al. 2022), which employ both bidirectional distillation and iterative model parameter exchange, our framework makes another difference by conducting one-way distillation from thoroughly trained local models to the central model. These fully trained teacher models immediately enable us to explore the input space and learn the most efficient samples for knowledge distillation. Our critical insight is that each local's unique expertise under the heterogeneous FL setting can be further exploited. Therefore, we implement the input distillation according to the corresponding local products (*c.f.*, Figure 2). This involves learning the transferred input to enable local nodes to reach a consensus on its semantic clarity while simultaneously generating diverse predictions with each task model. The former ensures the fundamental viability of the input data for transferring knowledge. At the same time, the latter allows the input data to leverage the unique aspects of each local node under heterogeneous federated learning scenarios. Such feedback from local nodes enables us to deploy per-input importance weight for output ensemble distillation. We demonstrate the effectiveness of our proposed method on natural and medical images through comprehensive experiments on image classification and segmentation tasks under various real-world federated learning scenarios, including the most challenging cross-domain cross-site settings. Our key contributions can be summarized as follows.

- We propose a federated learning framework with collaborative distillation in both the input and output space. It eliminates any requirement on model parameter exchange, model structure identity, prior knowledge of the local task, or auxiliary real data.
- To cope with the inherent heterogeneity of decentralized clients in federated learning, we introduce an ensemble distillation scheme that learns transferred input with explicit exploitation of each local's consensual and unique expertise.
- We conduct extensive experiments with natural and medical images on classification and segmentation tasks, demonstrating state-of-the-art privacy-utility trade-offs compared to the prior art.

Related Work

Knowledge Distillation

Hinton *et al.* (Hinton, Vinyals, and Dean 2015) first proposed the concept of knowledge distillation *i.e.*, using a cumbersome network as a teacher to generate soft labels to supervise the training of a compact student network. Although most of the following works transfer knowledge with one teacher, some techniques focus on multiple teachers and propose a variety of aggregation schemes, *e.g.*, gate learning in the supervised setting (Asif, Tang, and Harrer 2019; Xiang, Ding, and Han 2020), and relative sample similarity for unsupervised scenarios (Wu et al. 2019). Recent progress in data-free knowledge transfer (Fang et al. 2019; Chen et al. 2019) focuses on an adversarial training scheme to generate hard-to-learn and hard-to-mimic samples. Similarly, DeepInversion (Yin et al. 2020) utilizes backpropagated gradients to generate transfer samples that cause disagreements between the teacher and the student. (Nayak et al. 2019) crafts a transfer set by modeling and fitting data distributions in output similarities.

Distillation-based Federated Learning

Beyond the parameter based FL (McMahan et al. 2017; Hsu, Qi, and Brown 2019; Li et al. 2018), early FL works like (Jeong et al. 2018) employ parameter and model output exchanges. Although the following works (Li and Wang 2019;

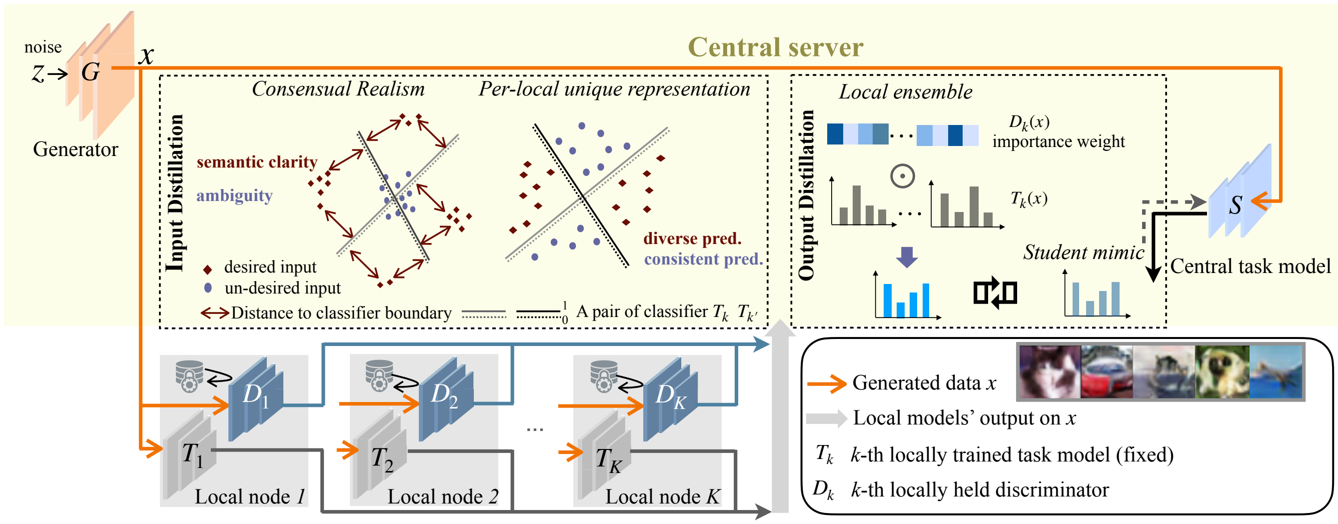

Figure 2: The overall pipeline of the proposed FedIOD. We conduct distillation in input and output spaces to transfer knowledge from the locally trained task model T_k and the auxiliary discriminator D_k to the central task model S. *Input distillation* optimizes central generator G to generate transferred input on which local models (1) achieve consensus on its semantic clarity, (2) and simultaneously produce diverse predictions. The latter is to exploit each local's unique expertise under the heterogeneous FL setting. *Output distillation* leverages per-input importance for output ensemble knowledge transfer.

Chang et al. 2019; Li, He, and Song 2021) are purely based on the output of the local model for knowledge transfer, the selection of transfer data is highly dependent on prior knowledge of private data (*i.e.*, they are under similar data distributions). Some recently proposed methods (Lin et al. 2020; Gong et al. 2022a) provide some relaxation on transfer data. However, it is still necessary to carefully select the transfer data according to prior knowledge of the local task and private data. While (Zhu, Hong, and Zhou 2021; Zhang, Wu, and Yuan 2022; Zhang et al. 2022) transfer knowledge without any requirement of real data, all of them need high communication bandwidth due to the iterative exchange of models over hundreds of rounds, leading to high susceptibility to stealth attacks and, hence, privacy concerns.

Approach

Problem Statement

Without loss of generality, we describe our method for the classification task in detail. Suppose that there are K local nodes in a federated learning scenario, each privately holding a labeled dataset $\{\mathcal{X}'_k, \mathcal{Y}'_k\}$, consisting of the input image space $\mathcal{X}' \in \mathbb{R}^{H\times W\times 3}$, and the label space $\mathcal{Y}' \in \{1, \ldots, C\}$, where C is the total number of classes.

The proposed FedIOD includes two stages. First, with each private data $\{\mathcal{X}'_k, \mathcal{Y}'_k\}$ we train the local model T_k to complete. Note that the proposed FedIOD is agnostic to any neural network architecture. Hence, each local node can have its specialized architecture suited to the particular distribution of its local data. In the second stage, each locally trained model, T_k, will be frozen and only used as a teacher model in a one-way distillation paradigm. In contrast to (Gong et al. 2022b; Li, He, and Song 2021) using carefully deliberated real data to transfer knowledge, we exploit ensemble knowledge in the input space $\mathcal{X}$ with a generator G mapping from random noise $\mathcal{W}$ to the input space $\mathcal{X}$. Taking such generated samples $x \sim \mathcal{X}$ as input, local models T_k and the central task model S on the server constitute a student-teacher knowledge transfer problem, with the teacher here being a group of local teachers. Let $\hat{\boldsymbol{z}} = S(x)$ and $\boldsymbol{z}_k = T_k(x)$ be the output logits of the central model and the k-th respectively ($\hat{\boldsymbol{z}}, \boldsymbol{z}_k \in \mathbb{R}^C$), the corresponding probability can be acquired with the following activation function:

$$p_\tau(\boldsymbol{z}) = \left[\frac{e^{z^1/\tau}}{\sum_c e^{z^c/\tau}}, \ldots, \frac{e^{z^C/\tau}}{\sum_c e^{z^c/\tau}}\right], \quad (1)$$

where τ is a temperature parameter set to 1 by default. We abbreviate $p_\tau(\boldsymbol{z}_k)$ and $p_\tau(\hat{\boldsymbol{z}})$ as $\boldsymbol{q}_k = T_k(x;\tau)$ and $\hat{\boldsymbol{q}} = S(x;\tau)$, respectively.

Input Ensemble Distillation

To efficiently exploit the knowledge from local expertise, exploring the input space for the best fit of the global distribution is vital. We expect the optimal input to achieve (1) realism as a consensus achieved by all local nodes and (2) diversity to represent each local's unique knowledge under the heterogeneous federated learning scenarios.

Consensual realism learning. Given the locally trained model T_k as teachers and the central model S as a student, we learn a generative model G from randomly sampled noise w to pseudo-data x, which will be the input for knowledge transfer. To guarantee the realism and practicality of x, we employ an additional discriminator D_k residing at each local node to boost the generative model G training. G is trained to approximate the global data distribution by fooling each local D_k. Following the typical training paradigm of GAN (Goodfellow et al. 2020; Radford, Metz, and Chintala 2015), we train G and D_k in a classical adversarial manner:

$$\max_G \min_{D_k} L^k_{\text{gan}}(G, D_k) = \max_G \min_{D_k} \mathop{\mathbb{E}}_{x'_k \in \mathcal{X}'_k}[\pi_k D_k(x'_k)] + \mathop{\mathbb{E}}_{w \in \mathcal{W}}[1 - D_k(G(w))], \tag{2}$$

where $\pi_k = \frac{|\mathcal{X}'_k|}{\sum_{k'=1}^{K} |\mathcal{X}'_{k'}|}$ is individual local weight and $|\mathcal{X}'_k|$ indicates data size. In addition to this high-level realism, we expect x to be realistic semantically, *i.e.*, with semantic clarity according to the output of each locally trained model. Here, we assume that the input that confuses local models to produce ambiguous results will be less efficient in transferring knowledge. Hence, we expect each local model to produce confident predictions that the input x tends to belong to one particular category. To force such semantic clarity, we maximize the confidence that x belongs to one class. For each local node k, taking $\boldsymbol{q}_k$ as its corresponding probability, we minimize the Shannon entropy $H(\boldsymbol{q}) = -\sum_c \boldsymbol{q}^c \log \boldsymbol{q}^c$, which can be reformulated as:

$$\begin{aligned} \min_G L_{\text{conf}}(G) &= \min_G \mathop{\mathbb{E}}_{x \in \mathcal{X}}[\sum_k \pi_k H(T_k(x; \tau))] \\ &= \min_G \mathop{\mathbb{E}}_{w \in \mathcal{W}}[\sum_k \pi_k H(T_k(G(w); \tau))]. \end{aligned} \tag{3}$$

Per-local unique representation. The supervisions above ensure the realism of x, which are agreed upon by all local nodes. However, it can hardly transfer heterogeneous knowledge across local nodes. Our insight is that each local's expertise must be inconsistent, given the data heterogeneity in a federated learning scenario. Hence, the input must be diverse to generalize and transfer each local's unique knowledge. To this point, we aim to generate x, which will tolerate local diversity, w.r.t., input data on which local models produce divergent results. Specifically, we use Jensen-Shannon divergence to measure the dissimilarity of local probability outputs:

$$\text{JSD}(\boldsymbol{q}_1, \ldots, \boldsymbol{q}_K) = H(\bar{\boldsymbol{q}}) - \sum_{k=1}^{K} \pi_k H(\boldsymbol{q}_k), \tag{4}$$

where $\bar{\boldsymbol{q}} = \sum_{k=1}^{K} \pi_k \boldsymbol{q}_k$ is the weighted ensemble of all locals. We maximize such dissimilarity to encourage the level of local diversity, w.r.t., unique local knowledge which has been exploited:

$$\min_G L_{\text{unique}}(G) = \min_G \mathop{\mathbb{E}}_{w \in \mathcal{W}}[-\text{JSD}(T_1(G(w); \tau), \ldots, T_K(G(w); \tau))]. \tag{5}$$

Output Ensemble Distillation

Model distillation techniques typically optimize the student model by minimizing the KL divergence between the student model output $\hat{\boldsymbol{q}}$ and the teacher model output $\bar{\boldsymbol{q}}$ to bridge their performance gap:

$$\text{KL}(\bar{\boldsymbol{q}} || \hat{\boldsymbol{q}}) = H(\bar{\boldsymbol{q}}, \hat{\boldsymbol{q}}) - H(\hat{\boldsymbol{q}}), \tag{6}$$

where $H(\bar{\boldsymbol{q}}, \hat{\boldsymbol{q}}) = -\sum_c \bar{\boldsymbol{q}}^c \log \hat{\boldsymbol{q}}^c$. Hinton *et al.* (Hinton, Vinyals, and Dean 2015) has shown that minimizing Eq. 6 with a high τ (Eq. 1) is equivalent to minimizing the ℓ_2 error between the logits of teacher and student, thereby relating cross-entropy minimization to fitting logits. For multiple teachers, the conventional ensemble takes an average of all teachers' output probability as $\bar{\boldsymbol{q}}$.

However, under the FL scenario, it is not optimal to deploy such a local ensemble under the heterogeneous data distribution. This is mainly due to its inability to cope with the general setting when locally held data are not independent and identically distributed, *e.g.*, they do not share precisely the same set of target classes. Let $P_{\mathcal{X}'_k, \mathcal{Y}'_k}$ be the data distribution of the image and label over the k-th local data, and $P_{\mathcal{X}', \mathcal{Y}'}$ be the global data distribution. Thus, we approximate the importance ratio of local prediction based on its training data distribution:

$$\begin{aligned} \frac{P_{\mathcal{X}'_k, \mathcal{Y}'_k}(y|x)}{P_{\mathcal{X}', \mathcal{Y}'}(y|x)} &= \frac{P_{\mathcal{Y}'_k}(y) P_{\mathcal{X}'_k, \mathcal{Y}'_k}(x|y) P_{\mathcal{X}'}(x)}{P_{\mathcal{Y}'}(y) P_{\mathcal{X}', \mathcal{Y}'}(x|y) P_{\mathcal{X}'_k}(x)} \\ &\approx \frac{P_{\mathcal{Y}'_k}(y)}{P_{\mathcal{Y}'}(y)} \cdot \frac{P_{\mathcal{X}'}(x)}{P_{\mathcal{X}'_k}(x)} \approx \frac{P_{\mathcal{Y}'_k}(y)}{P_{\mathcal{Y}'}(y)} \cdot \frac{P_{\mathcal{X}}(x)}{P_{\mathcal{X}'_k}(x)}, \end{aligned} \tag{7}$$

where we assume $P_{\mathcal{X}'_k, \mathcal{Y}'_k}(x|y) \approx P_{\mathcal{X}', \mathcal{Y}'}(x|y)$ as the local heterogeneity of this term is minor and ignorable compared to the heterogeneity in the image distribution $P_{\mathcal{X}'}(x)$ and the label distribution $P_{\mathcal{Y}'_k}(y)$. And the global image distribution $\mathcal{X}'$ is approximated with the generated input domain $\mathcal{X} \approx \mathcal{X}'$.

To consider this aspect, we introduce the weight of importance per class per input π^c_k for each local node k to reflect the data distribution with which its model was initially trained. Taking x as input, we have the following.

$$\hat{\pi}^c_k(x) = \frac{\mathbb{E}_{y'_k \in \mathcal{Y}'_k} |y'_k = c|}{\mathbb{E}_{k \in \{1, \cdots, K\}, y'_k \in \mathcal{Y}'_k} |y'_k = c|} \cdot \frac{D_k(x)}{\mathbb{E}_{x'_k \in \mathcal{X}'_k} D_k(x'_k)}, \tag{8}$$

where the first term corresponds to $\frac{P_{\mathcal{Y}'_k}(y)}{P_{\mathcal{Y}'}(y)}$ and can be acquired by statistics of local labels, *i.e.*, the number of samples from class c used to train the model at the local node k. The second term corresponds to $\frac{P_{\mathcal{X}}(x)}{P_{\mathcal{X}'_k}(x)}$ which can be approximated by the local discriminator's output on pseudo image x and locally held image x'_k. We then normalize the importance weight between locals for each c: $\pi^c_k(x) = \hat{\pi}^c_k(x) / \sum_{k'=1}^{K} \hat{\pi}^c_{k'}(x)$.

Following the ℓ_2 observation above of Hinton *et al.* (Hinton, Vinyals, and Dean 2015), we consider the case of $\tau \to \infty$ when deploying KL-divergence. Hence, it can be written as the ℓ_2 error between central model logits $\hat{\boldsymbol{z}}$ and local aggregated $\bar{\boldsymbol{z}}$. Let $\boldsymbol{\pi}_k(x) = [\pi^1_k(x), \cdots, \pi^C_k(x)] \in [0, 1]^C$ be the per-sample weight, and $\odot$ is Hadamard product, the local ensemble expertise is indicated as follows:

$$A(\boldsymbol{z}_1, \cdots, \boldsymbol{z}_K, x) = \sum_{k=1}^{K} \boldsymbol{\pi}_k(x) \odot \boldsymbol{z}_k, \tag{9}$$

Algorithm 1: FedIOD

Input: Total number of local nodes K, locally held data $\{\mathcal{X}'_k, \mathcal{Y}'_k\}$, local models $\{T_k\}$, central task model S, central generator G, auxiliary local discriminator $\{D_k\}$.
for each local node $k = 1, \cdots, K$ **do**
  Train T_k with $(\mathcal{X}'_k, \mathcal{Y}'_k)$ to complete
end for
for each distillation step **do**
  □ Input distillation
  $w \leftarrow$ randomly sampled from $\mathcal{W}$
  $x \leftarrow G(w)$
  for $k = 1, ..., K$ **do**
    $\boldsymbol{z}_k, \boldsymbol{q}_k \leftarrow T_k(x)$
    $x'_k \leftarrow$ randomly sampled from $\mathcal{X}'_k$
    $L^k_{\text{gan}}(G, D_k) \leftarrow D_k(x'_k), D_k(x)$ ▷ Eq. 2
    Update D_k by descending its stochastic gradient $\nabla_{D_k} L_{\text{gan}}$
  end for
  $L_{\text{conf}}(G), L_{\text{unique}}(G) \leftarrow \{\boldsymbol{q}_k\}$ ▷ Eq. 3,5
  □ Output distillation
  $\hat{\boldsymbol{z}}, \hat{\boldsymbol{q}} \leftarrow S(x)$
  $L_{\text{mimic}}(G, S) \leftarrow \hat{\boldsymbol{z}}, \{\boldsymbol{z}_k\}$ ▷ Eq. 10
  □ Update
  Update G by descending its stochastic gradient $\nabla_G[L_{\text{conf}} + L_{\text{unique}} - L_{\text{mimic}} - \sum_{k=1}^K L^k_{\text{gan}}]$
  Update S by descending its stochastic gradient $\nabla_G L_{\text{mimic}}$
end for

where the central model S is optimized to mimic the local ensemble of expertise, while the generator G is a critic to generate x on which S lags behind local experts. The motivation is that such challenging input will transfer the hard-to-mimic knowledge from local to central. Therefore, we tailor the input data on which the central model produces a result diverged from the local output. Using KL-divergence as a dissimilarity evaluation, we train G and S in an adversarial manner:

$$\max_G \min_S L_{\text{mimic}}(G, S) = \max_G \min_S \mathop{\mathbb{E}}_w |S(G(w)) - A(T_1(G(w)), \cdots, T_K(G(w)))|^2, \tag{10}$$

where $A(\cdot)$ is the aggregation function detailed in Eq. 9. To sum up, the overall loss function can be written as

$$\max_G \min_{D_k} L^k_{\text{gan}}(G, D_k) + \min_G [L_{\text{conf}}(G) + L_{\text{unqiue}}(G)] + \max_G \min_S L_{\text{mimic}}(G, S). \tag{11}$$

And the overall process is explained in Algorithm 1.

Experiments

We provide comprehensive empirical studies with various heterogeneous FL settings on natural image classification and more privacy-sensitive medical tasks, including brain tumor segmentation and histopathological nuclei instance segmentation.

CIFAR-10/100 Classification

We use heterogeneous data splits with Dirichlet distribution following the prior art (Hsu, Qi, and Brown 2019) for distributed local training sets. The value of α in the Dirichlet distribution controls the degree of non-IIDness: $\alpha \to \infty$ indicates an identical local data distribution, and a smaller α indicates a higher non-IIDness. We report average accuracy over three split seeds on the corresponding test set.

We conduct experiments following the typical FL setting (Lin et al. 2020) under K=20 and α=1, 0.1 with ResNet-8. w is randomly sampled with a dimension of 100, and $x = G(w)$ has a size of 32×32. We use a patch discriminator as D_k, of which the output is of size 8×8. The comparison in Table 1 shows that our method achieves superior or competitive results and a much stronger privacy guarantee. Without the requirement of auxiliary data or prior knowledge of the local task, our method outperforms relevant-data-dependent distillation-based and parameter-based counterparts. Moreover, our method demonstrates other benefits, including eliminating prerequisites of identical local model architecture or task-relevant real data.

Magnetic Resonance Image Segmentation

We use the dataset from the 2018 Multimodal Brain Tumor Segmentation Challenge (BraTS 2018) (Menze et al. 2014; Bakas et al. 2018). Each subject was associated with voxel-level annotations of “whole tumor”, “tumor core,” and “enhancing tumor.” Following the experimental protocol of one prior art, (Chang et al. 2020), we deploy 2D segmentation of the whole tumor on T2 images of HGG cases, among which 170 were for training and 40 for testing. The local data split also follows (Chang et al. 2020).

We employ the same network structure of G, D_k, S, and the same data preprocessing as (Chang et al. 2020) for a fair comparison. Following its label condition $\mathcal{W}$, we improve our L_{gan} with additional perceptual loss (Johnson, Alahi, and Fei-Fei 2016). The Dice score, sensitivity (Sens.), specificity (Spec.), and Hausdorff distance (HD95) are used as evaluation metrics, where “HD95” represents 95% quantile of the distances instead of the maximum.

Table 2 compares our method with the prior art of distributed learning (Chang et al. 2020) and the classical parameter-based FedAvg method. Ours performs best segmentation on pixel-level overlap metrics (Dice and Sens.) and shape similarity metrics (HD95).

Histopathological Image Segmentation

In real-world medical applications, the heterogeneity of data distributed among medical entities is not limited to the local size of the data or various subjects. Local data held by different clinical sites can be quite a domain variant, *e.g.*, targeting different organs or collected with different infrastructures, which is relatively underexplored in contemporary FL methods. To this end, we evaluate our method in a cross-organ, cross-site setting where locally held data are from different organs and institutes. We experiment on nuclei instance segmentation task with pathological datasets, including TCGA (Kumar et al. 2017), Cell17 (Vu et al. 2019) and TNBC (Naylor et al. 2018).

	Method	Model-agnostic	Auxiliary Prerequisite	CIFAR-10 $\alpha = 1$	CIFAR-10 $\alpha = 0.1$	CIFAR-100 $\alpha = 1$	CIFAR-100 $\alpha = 0.1$
	Standalone (mean $\pm$ std)	-	-	65.25$\pm$ 5.14	30.92$\pm$ 11.17	27.60$\pm$ 1.58	16.99$\pm$ 2.46
Parameter-based	FedAvg (McMahan et al. 2017)	✗	-	78.57$\pm$ 0.22	68.37$\pm$ 0.50	42.54$\pm$ 0.51	36.72$\pm$ 1.50
	FedProx (Li et al. 2018)	✗	-	76.32$\pm$ 1.95	68.65$\pm$ 0.77	42.94$\pm$ 1.23	35.74$\pm$ 1.00
	FedAvgM (Hsu, Qi, and Brown 2019)	✗	-	77.79$\pm$ 1.22	68.63$\pm$ 0.79	42.83$\pm$ 0.36	36.29$\pm$ 1.98
	FedGEN (Zhu, Hong, and Zhou 2021)	✗	task-relevant data	80.31$\pm$ 0.97	68.13$\pm$ 1.37	45.97$\pm$ 0.23	35.97$\pm$ 0.31
	FedDF (Lin et al. 2020)	✗	task-relevant data	**80.69$\pm$ 0.43**	**71.36$\pm$ 1.07**	**47.43$\pm$ 0.45**	**39.33$\pm$ 0.03**
Distill-based	FedMD (Li and Wang 2019)	✓	task-relevant data	80.37$\pm$ 0.37	69.23$\pm$ 1.31	**45.83$\pm$ 0.58**	38.86$\pm$ 0.78
	FedKD (Gong et al. 2022a)	✓	task-relevant data	80.98$\pm$ 0.11	65.46$\pm$ 3.45	45.55$\pm$ 0.38	40.61$\pm$ 2.54
	FedIOD	✓	None	**82.78$\pm$ 0.18**	**70.08$\pm$ 0.37**	45.36$\pm$ 0.32	**41.88$\pm$ 0.16**

Table 1: Accuracy (%) comparisons on the CIFAR-10 and CIFAR-100 datasets with ResNet-8 and K=20. "Standalone" indicates the performance of local models trained with individual private data. Several popular FL methods are compared with parameter-based and distillation-based FL prior arts.

	Dice(%)↑	Sens.(%)↑	Spec.(%)↑	HD95↓
Standalone	65.03 $\pm$3.31	69.27 $\pm$4.72	99.35 $\pm$0.15	24.61 $\pm$3.62
Centralized	74.85	79.83	99.55	12.85
FedAvg	70.71	67.31	**99.85**	11.88
AsynDGAN	70.43	72.95	99.57	14.94
FedIOD	**75.38**	**79.47**	99.60	**11.76**

Table 2: Comparisons on the BraTS2018 dataset with K=10 under the same net with FedAvg and AsynDGAN. "Centralized": centralized training with all local data.

		Dice(%)↑	Obj-Dice(%)↑	AJI(%)↑	HD95↓
Standalone	breast	77.92	73.47	53.64	12.34
	liver	79.16	75.38	55.63	12.47
	kidney	74.99	69.67	50.99	14.64
	prostate	77.46	73.74	54.40	15.59
	FedAvg	78.12	75.05	55.56	12.96
	AsynDGAN	79.30	72.73	56.08	14.49
	FedIOD	**80.48**	**77.03**	**58.37**	**11.22**

Table 3: Comparisons on the TCGA dataset with four cross-organ local nodes. All methods use the same segmentation net provided by (Chang et al. 2020) for a fair comparison.

We cropped the images into patches of size 256×256 for training and inference. For metrics evaluation, the cropped patches are stitched back into the whole image with the original size. For G, D_k, and S, we use the same model structure provided by (Chang et al. 2020) and the additional perceptual loss (Johnson, Alahi, and Fei-Fei 2016) for L_{gan}. We use object-level Dice (Chen et al. 2016) and Aggregated Jaccard Index (AJI) (Vu et al. 2019) as metrics to evaluate the instance overlap or shape similarities for an individual object. Let $\boldsymbol{y}^i$ be the ground truth mask for the i-th instance of the total n instances, and $\hat{\boldsymbol{y}}^j$ be the predicted mask for the j-th instance of the total $\hat{n}$ instances. $J(\boldsymbol{y}^i) = \text{argmax}_{\hat{\boldsymbol{y}}^j} |\boldsymbol{y}^i \cap \hat{\boldsymbol{y}}^j|/|\boldsymbol{y}^i \cup \hat{\boldsymbol{y}}^j|$ is the predicted instance that maximally overlaps $\boldsymbol{y}^i$, and $J(\hat{\boldsymbol{y}}^j) = \text{argmax}_{\boldsymbol{y}^i} |\boldsymbol{y}^i \cap \hat{\boldsymbol{y}}^j|/|\boldsymbol{y}^i \cup \hat{\boldsymbol{y}}^j|$ denotes the ground-truth instance that maximally overlaps $\hat{\boldsymbol{y}}^j$. For instance, for shape similarity, we use the Aggregated Jaccard Index (AJI):

$$\text{AJI}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{\sum_{i=1}^{n} |\boldsymbol{y}^i \cap J(\boldsymbol{y}^i)|}{\sum_{i=1}^{n} |\boldsymbol{y}^i \cup J(\boldsymbol{y}^i)| + \sum_{j \in \mathcal{J}} |\hat{\boldsymbol{y}}^j|}, \tag{12}$$

where $J(\boldsymbol{y}^i)$ is the predicted instance that has maximum overlap with $\boldsymbol{y}^i$ concerning the Jaccard index (sorted and nonrepeated). $\mathcal{J}$ is the set of predicted instances that have not been assigned to any ground-truth instance.

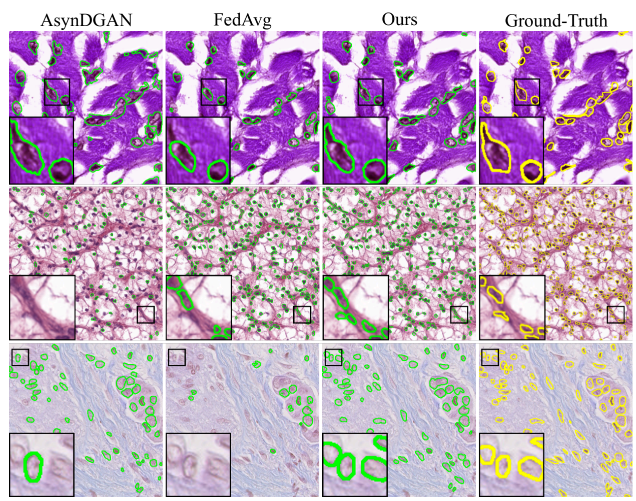

Figure 3: Qualitative comparisons on cross-site cross-organ nuclei segmentation tasks. The three rows visualize instance contours on test images from Cell17, TCGA, and TNBC.

Cross-organ scenario. We first focus on cross-organ settings where each distributed local node holds only the data of one organ. Following (Chang et al. 2020), from the TCGA dataset, we take 16 images of the breast, liver, kidney, and prostate for training and eight images of the same organs for testing. Table 3 shows the experimental results of this cross-organ setting and compares them with the baseline method (Chang et al. 2020) and the classical FedAvg. We can note that our method achieves the best results on semantic segmentation (Dice and Hausdorff) and instance segmentation

	Test Data	Dice(%)↑	Obj-Dice(%)↑	AJI(%)↑	HD95↓	Average Dice(%)↑	Average Obj-Dice(%)↑	Average AJI(%)↑	Average HD95↓
FedAvg	Cell17	68.74	65.82	39.37	24.15	63.42	64.00	37.29	54.51
	TCGA	77.57	72.94	50.03	15.87				
	TNBC	43.95	53.23	22.48	123.51				
AsynDGAN	Cell17	79.82	59.03	34.64	19.27	66.64	61.15	34.21	35.46
	TCGA	52.29	57.12	26.03	47.47				
	TNBC	67.80	67.31	41.96	39.63				
FedIOD	Cell17	86.23	68.03	44.75	7.01	**79.28**	**71.58**	**49.52**	**16.41**
	TCGA	76.59	72.67	53.04	12.69				
	TNBC	75.01	74.03	50.76	29.54				

Table 4: Comparisons of cross-site cross-organ nuclei segmentation tasks with Cell17, TCGA, TNBC as distributed local data. For a fair comparison, all methods use the same U-Net architecture and the same post-processing method.

(object-level Dice and AJI) metrics.

Cross-site cross-organ scenario. We also conduct experiments on more challenging settings with cross-site cross-organ datasets, where locally held data are from different organ nuclei datasets. Taking the training set of Cell17, TCGA, and TNBC as private data distributed over local nodes, we evaluate on the corresponding test sets. Table 4 compares our method with two prior arts (Chang et al. 2020; McMahan et al. 2017) on various segmentation metrics to evaluate semantic/instance level overlap and shape. Our proposed FedIOD outperforms the prior art on all these metrics for overlap and shape evaluation, demonstrating our efficacy in coping with heterogeneous FL scenarios. The qualitative comparisons shown in Figure 3 also demonstrate the superiority of our method over its counterparts.

Privacy budget ε ↓		3.5	6.0	7.7	10.0
FedKD	w/ DP ↑	45.64	56.08	61.80	70.90
	w/o DP ↑	66.79	79.30	80.28	81.55
FedIOD	w/ DP ↑	44.45	58.96	62.14	73.58
	w/o DP ↑	74.31	80.02	82.03	82.69

Table 5: Compare FedIOD and FedKD in terms of accuracy (%) on CIFAR10 (K=20, α=1) under same privacy cost.

Privacy Analysis

Comparison with data-dependent distillation-based FL. The significant difference between ours and typical FL based on distillation is that FedIOD generates data for knowledge distillation, while others rely on auxiliary real data. We adopt the differential privacy (DP) analysis in DP-CGAN (Torkzadehmahani, Kairouz, and Paten 2019) and GS-WGAN (Chen, Orekondy, and Fritz 2020) to measure the privacy cost of the gradients used to train the generator. For a fair comparison, we apply PATE (Papernot et al. 2018) on the local model output and then transfer them to the server to satisfy DP for both FedIOD and our counterpart FedKD (Gong et al. 2022a). Table 5 compares FedIOD with FedKD in terms of accuracy under a series of rigid differential privacy protections (ε <10).

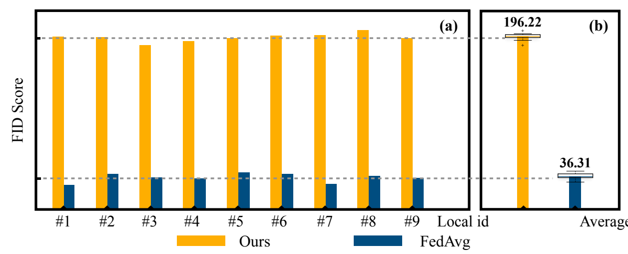

Figure 4: Comparison of FID scores between FedIOD and FedAvg on (a) 9 randomly selected local clients; and (b) average score under CIFAR10 (K=20, α=1) FL setting.

Comparison with parameter-based FL. We use DLG (Zhu, Liu, and Han 2019) as an attacker to recover private data using its iterative shared model parameters for parameter-based FL. We then measure the quality of the recovered data using Fréchet Inception Distance (FID). We assume a larger FID, *i.e.*, a larger distance between the recovered data and private data, indicates a stronger privacy guarantee. For our method, we measure the FID between the synthetic images and the private images. The comparison in Figure 4 shows that our method has a much higher FID, thus far more privacy protected than the FL parameter-sharing method such as FedAvg (McMahan et al. 2017).

Conclusions

In this work, we propose a novel federated learning framework, FedIOD, that protects local data privacy by distilling input and output to transfer knowledge from locals to the central server. To cope with the highly non-i.i.d. data distribution across local nodes, we learn the input on which each local achieves both consensual and unique results to represent individual heterogeneous expertise. We conducted extensive experiments with natural and medical images on classification and segmentation tasks in a variety of real, in-the-wild, heterogeneous FL settings. All demonstrate the efficacy of FedIOD, showing its superior privacy-utility tradeoff to others and significant flexibility in FL scenarios without any prior knowledge or auxiliary real data.

Acknowledgments

This research was supported in part by Zhejiang Provincial Natural Science Foundation of China under Grant No. LD24F020007, Beijing Natural Science Foundation L223024, National Natural Science Foundation of China under Grant 62076016, the National Key Research and Development Program of China (Grant No. 2023YFC3300029) and "One Thousand Plan" projects in Jiangxi Province Jxsg2023102268 and a generous gift from Amazon.

References

Asif, U.; Tang, J.; and Harrer, S. 2019. Ensemble knowledge distillation for learning improved and efficient networks. *arXiv preprint arXiv:1909.08097*.

Bakas, S.; Reyes, M.; Jakab, A.; Bauer, S.; Rempfler, M.; Crimi, A.; Shinohara, R. T.; Berger, C.; Ha, S. M.; Rozycki, M.; et al. 2018. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv preprint arXiv:1811.02629*.

Chang, H.; Shejwalkar, V.; Shokri, R.; and Houmansadr, A. 2019. Cronus: Robust and Heterogeneous Collaborative Learning with Black-Box Knowledge Transfer. *arXiv preprint arXiv:1912.11279*.

Chang, Q.; Qu, H.; Zhang, Y.; Sabuncu, M.; Chen, C.; Zhang, T.; and Metaxas, D. N. 2020. Synthetic learning: Learn from distributed asynchronized discriminator GAN without sharing medical image data. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13856–13866.

Chen, D.; Orekondy, T.; and Fritz, M. 2020. Gs-wgan: A gradient-sanitized approach for learning differentially private generators. *Advances in Neural Information Processing Systems*, 33: 12673–12684.

Chen, H.; Qi, X.; Yu, L.; and Heng, P.-A. 2016. DCAN: deep contour-aware networks for accurate gland segmentation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2487–2496.

Chen, H.; Wang, Y.; Xu, C.; Yang, Z.; Liu, C.; Shi, B.; Xu, C.; Xu, C.; and Tian, Q. 2019. Data-free learning of student networks. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, 3514–3522.

Fang, G.; Song, J.; Shen, C.; Wang, X.; Chen, D.; and Song, M. 2019. Data-free adversarial distillation. *arXiv preprint arXiv:1912.11006*.

Geiping, J.; Bauermeister, H.; Dröge, H.; and Moeller, M. 2020. Inverting Gradients–How easy is it to break privacy in federated learning? *arXiv:2003.14053*.

Gong, X.; Sharma, A.; Karanam, S.; Wu, Z.; Chen, T.; Doermann, D.; and Innanje, A. 2022a. Preserving Privacy in Federated Learning with Ensemble Cross-Domain Knowledge Distillation. In *Association for the Advancement of Artificial Intelligence*.

Gong, X.; Song, L.; Vedula, R.; Sharma, A.; Zheng, M.; Planche, B.; Innanje, A.; Chen, T.; Yuan, J.; Doermann, D.; and Ziyan, W. 2022b. Federated Learning with Privacy-Preserving Ensemble Attention Distillation. *IEEE Transactions on Medical Imaging*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hsu, T.-M. H.; Qi, H.; and Brown, M. 2019. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*.

Hsu, T.-M. H.; Qi, H.; and Brown, M. 2020. Federated Visual Classification with Real-World Data Distribution. In *Proceedings of European Conference on Computer Vision*.

Jeong, E.; Oh, S.; Kim, H.; Park, J.; Bennis, M.; and Kim, S.-L. 2018. Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data. *arXiv preprint arXiv:1811.11479*.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of European conference on computer vision*, 694–711. Springer.

Karimireddy, S. P.; Kale, S.; Mohri, M.; Reddi, S. J.; Stich, S. U.; and Suresh, A. T. 2020. Scaffold: Stochastic controlled averaging for on-device federated learning. In *Proceedings of International Conference on Machine Learning*.

Kumar, N.; Verma, R.; Sharma, S.; Bhargava, S.; Vahadane, A.; and Sethi, A. 2017. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE Transactions on Medical Imaging*, 36(7): 1550–1560.

Li, D.; and Wang, J. 2019. Fedmd: Heterogenous federated learning via model distillation. *arXiv preprint arXiv:1910.03581*.

Li, Q.; He, B.; and Song, D. 2021. Practical one-shot federated learning for cross-silo setting. In *Proceedings of International Joint Conference on Artificial Intelligence*.

Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2018. Federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*.

Li, T.; Sanjabi, M.; Beirami, A.; and Smith, V. 2020. Fair resource allocation in federated learning. In *Proceedings of International Conference on Learning Representations*.

Lin, T.; Kong, L.; Stich, S. U.; and Jaggi, M. 2020. Ensemble Distillation for Robust Model Fusion in Federated Learning. In *Proceedings of Conference on Neural Information Processing Systems*.

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

Menze, B. H.; Jakab, A.; Bauer, S.; Kalpathy-Cramer, J.; Farahani, K.; Kirby, J.; Burren, Y.; Porz, N.; Slotboom, J.; Wiest, R.; et al. 2014. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10): 1993–2024.

Nayak, G. K.; Mopuri, K. R.; Shaj, V.; Radhakrishnan, V. B.; and Chakraborty, A. 2019. Zero-shot knowledge distillation in deep networks. In *Proceedings of International Conference on Machine Learning*, 4743–4751. PMLR.

Naylor, P.; Laé, M.; Reyal, F.; and Walter, T. 2018. Segmentation of nuclei in histopathology images by deep regression of the distance map. *IEEE Transactions on Medical Imaging*, 38(2): 448–459.

Papernot, N.; Song, S.; Mironov, I.; Raghunathan, A.; Talwar, K.; and Erlingsson, Ú. 2018. Scalable private learning with pate. *arXiv preprint arXiv:1802.08908*.

Radford, A.; Metz, L.; and Chintala, S. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Torkzadehmahani, R.; Kairouz, P.; and Paten, B. 2019. Dp-cgan: Differentially private synthetic data and label generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 0–0.

Vu, Q. D.; Graham, S.; Kurc, T.; To, M. N. N.; Shaban, M.; Qaiser, T.; Koohbanani, N. A.; Khurram, S. A.; Kalpathy-Cramer, J.; Zhao, T.; et al. 2019. Methods for segmentation and classification of digital microscopy tissue images. *Frontiers in bioengineering and biotechnology*, 53.

Wang, H.; Yurochkin, M.; Sun, Y.; Papailiopoulos, D.; and Khazaeni, Y. 2020. Federated learning with matched averaging. In *Proceedings of International Conference on Learning Representations*.

Wu, A.; Zheng, W.-S.; Guo, X.; and Lai, J.-H. 2019. Distilled person re-identification: Towards a more scalable system. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1187–1196.

Xiang, L.; Ding, G.; and Han, J. 2020. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In *Proceedings of European Conference on Computer Vision*, 247–263. Springer.

Yin, H.; Molchanov, P.; Alvarez, J. M.; Li, Z.; Mallya, A.; Hoiem, D.; Jha, N. K.; and Kautz, J. 2020. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8715–8724.

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

Zhang, L.; Wu, D.; and Yuan, X. 2022. FedZKT: Zero-Shot Knowledge Transfer towards Resource-Constrained Federated Learning with Heterogeneous On-Device Models. In *2022 IEEE 42nd International Conference on Distributed Computing Systems*, 928–938. IEEE.

Zhu, L.; Liu, Z.; and Han, S. 2019. Deep leakage from gradients. In *Proceedings of Conference on Neural Information Processing Systems*, 14774–14784.

Zhu, Z.; Hong, J.; and Zhou, J. 2021. Data-free knowledge distillation for heterogeneous federated learning. In *Proceedings of International Conference on Machine Learning*, 12878–12889. PMLR.