FedX: Unsupervised Federated Learning with Cross Knowledge Distillation

Sungwon Han$^{\star 1,2[0000-0002-1129-760X]}$, Sungwon Park$^{\star 1,2[0000-0002-6369-8130]}$,
Fangzhao Wu$^{3[0000-0001-9138-1272]}$, Sundong Kim$^{2[0000-0001-9687-2409]}$,
Chuhan Wu$^{4[0000-0001-5730-8792]}$, Xing Xie$^{3[0000-0002-8608-8482]}$, and
Meeyoung Cha$^{2,1[0000-0003-4085-9648]}$

[1] School of Computing, KAIST
{lion4151, psw0416}@kaist.ac.kr
[2] Data Science Group, Institute for Basic Science
{sundong, mcha}@ibs.re.kr
[3] Microsoft Research Asia
wufangzhao@gmail.com, xingx@microsoft.com
[4] Tsinghua University
wuchuhan15@gmail.com

Abstract. This paper presents FedX, an unsupervised federated learning framework. Our model learns unbiased representation from decentralized and heterogeneous local data. It employs a two-sided knowledge distillation with contrastive learning as a core component, allowing the federated system to function without requiring clients to share any data features. Furthermore, its adaptable architecture can be used as an add-on module for existing unsupervised algorithms in federated settings. Experiments show that our model improves performance significantly (1.58–5.52pp) on five unsupervised algorithms.

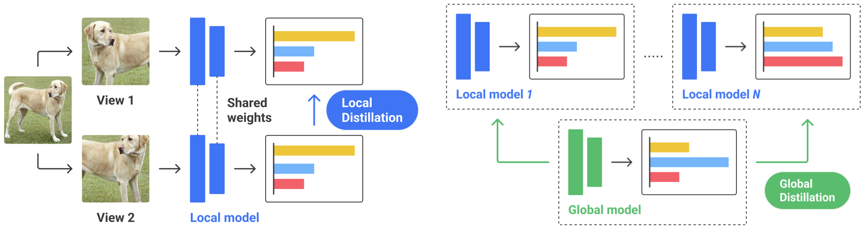
Keywords: Unsupervised representation learning, self-supervised learning, federated learning, knowledge distillation, data privacy

1 Introduction

Most deep learning techniques assume unlimited access to data during training. However, this assumption does not hold in modern distributed systems, where data is stored at client nodes for privacy reasons [28,34]. For example, personal data stored on mobile devices cannot be shared with central servers, nor can patient records in hospital networks. *Federated learning* is a new branch of collaborative technique to build a shared data model while securing data privacy; it is a method to run machine learning by involving multiple decentralized edge devices without exchanging locally bounded data [2,36].

In federated systems, supervised methods have been used for a variety of downstream tasks such as object detection [22], image segmentation [31], and

[⋆] Equal contribution to this work.



(a) Local knowledge distillation

(b) Global knowledge distillation

Fig. 1: Illustration of two knowledge flows in FedX: (a) local knowledge distillation progressively learns augmentation-invariant features, and (b) global knowledge distillation regularizes local models from bias.

person re-identification [45]. The main challenge here is the data's decentralized and heterogeneous nature (i.e., non-IID setting), which obscures the global data distribution. To address this issue, several methods have been proposed, including knowledge distillation [45], control variates [13], and contrastive learning [19]. These methods necessitate that local clients have high-quality data labels.

Nowadays, the need for *unsupervised* federated learning is increasing to handle practical scenarios that lack data labels. This is the new frontier in federated learning. There have been a few new ideas; for instance, Zhang *et al.* proposed FedCA, a model that uses local data features and external datasets to alleviate inconsistency in the representation space [42]. Wu *et al.* proposed FCL, which exchanges encrypted local data features for privacy and introduces a neighborhood matching approach to cluster the decentralized data across clients [38]. However, these approaches allow data sharing among local clients and raise privacy concerns.

We present FedX, a new advancement in unsupervised learning on federated systems that learns semantic representation from local data and refines the central server's knowledge via *knowledge distillation*. Unlike previous approaches, this model is privacy-preserving and does not rely on external datasets. The model introduces two novel considerations to the standard FedAvg [23] framework: *local knowledge distillation* to train the network progressively based on local data and *global knowledge distillation* to regularize data bias due to the non-IID setting. This two-sided knowledge flow distinguishes our model.

Local knowledge distillation (Fig. 1a) maximizes the embedding similarity between two different views of the same data instance while minimizing that of other instances—this process is defined by the *contrastive loss*. We designed an additional loss that relaxes the contrastive loss via soft labeling. Soft labels are computed as similarities between an anchor and randomly selected instances, called *relationship vectors*. We minimize the distance between relationship vectors of two different views in order to transfer structural knowledge and achieve fast training speed—this process is modulated by the *relational loss*.

Global knowledge distillation (Fig. 1b) treats the sample representation passed by the global model as an alternative view that should be placed near the embedding of the local model. This process is also modulated by contrastive loss and relational loss. Concurrent optimization allows the model to learn semantic information while eliminating data bias through regularization. These objectives do not require additional communication rounds or costly computation. Moreover, they do not share sensitive local data or use external datasets.

1. We propose an unsupervised federated learning algorithm, FedX, that learns data representations via a unique two-sided knowledge distillation at local and global levels.
2. Two-sided knowledge distillation helps discover meaningful representation from local data while eliminating bias by using global knowledge.
3. FedX can be applied to extant algorithms to enhance performance by 1.58–5.52pp in top-1 accuracy and further enhance training speed.
4. Unlike other unsupervised federated learning approaches, FedX preserves privacy between clients and does not share data directly. It is also lightweight and does not require complex communication for sending data features.
5. FedX is open-sourced at https://github.com/Sungwon-Han/FEDX.

2 Related Work

2.1 Unsupervised Representation Learning

There are two common approaches to unsupervised representation learning. One approach is to use generative models like autoencoder [33] and adversarial learning [30] that learn the latent representation by mimicking the actual data distribution. Another method is to use discriminative models with contrastive learning [5,27,40]. Contrastive learning approaches teach a model to pull the representations of the anchor and its positive samples (i.e., different views of the image) in embedding space, while pushing the anchor apart from negative samples (i.e., views from different images) [7,18].

In contrastive learning, SimCLR [3] employs data augmentation to generate positive samples. MoCo [8] introduces a momentum encoder and dynamic queue to handle negative samples efficiently. BYOL [6] reduces memory costs caused by a large number of negative samples. ProtoCL [18] uses prototypes to group semantically similar instances into local clusters via an expectation-maximization framework. However, under distributed and non-IID data settings, as in federated systems, these methods show a decrease in accuracy [42].

2.2 Federated Learning

Federated Averaging (FedAvg) by McMahan *et al.* is a standard framework for supervised federated learning [23]. Several subsequent studies improved the local update or global aggregation processes of FedAvg. For instance, external dataset [43], knowledge distillation [45], control variates [13,20], and contrastive learning [19] can be applied for better local update process. Similarly, global aggregation process can be improved via Bayesian non-parametric approaches [35], momentum updates [11], or normalization methods [37].

Unsupervised federated learning is more difficult to implement because no labels are provided and clients must rely on locally-defined pretext tasks that may be biased. This is a less explored field, with only a few methods proposed. FedCA [42] shares local data features and uses an external dataset to reduce the mismatch in representation space among clients. FCL [38] encrypts the local data features before exchanging them. Because of the explicit data sharing, these methods raise new privacy concerns. We, on the other hand, consider a completely isolated condition that does not permit any local data sharing. FedU [44] is another approach in the field that improves on the global aggregation method. It decides how to update predictors selectively based on the divergence of local and global models. Our model is orthogonal to FedU, and both concepts can be used in tandem.

2.3 Knowledge Distillation

Knowledge distillation aims to effectively train a network (i.e., student) by distilling the knowledge of a pretrained network (i.e., teacher). Knowledge can be defined over the features at the intermediate hidden layers [15,16], logits at the final layer [10], or structural relations among training samples [29,32,24]. Self-knowledge distillation uses the student network itself as a teacher network and progressively uses its knowledge to train the model [12,14]. We leverage this concept to efficiently train the local model while preserving the knowledge of the global model. FedX is the first-of-a-kind approach that uses the knowledge distillation concept for unsupervised federated learning.

3 Model

3.1 Overview

Problem statement. Consider a federated system in which data can only be viewed locally at each client and cannot be shared outside. Our goal is to train a single unsupervised embedding model F_ϕ that maps data points from each client to the embedding space. Let us denote local data and model from client m as $\mathcal{D}^m$ and f_θ^m respectively (i.e., $m \in \{1, ..., M\}$). The main objective for the global model F_ϕ is as follows:

$$\begin{gathered}\arg\min_{\phi} \mathcal{L}(\phi) = \sum_{m=1}^{M} \frac{|\mathcal{D}^m|}{|\mathcal{D}|} \mathcal{L}_m(\phi), \\ \text{where } \mathcal{L}_m(\phi) = \mathbb{E}_{\mathbf{x}\in\mathcal{D}^m}[l_m(\mathbf{x};\phi)]. \end{gathered} \tag{1}$$

$\mathcal{L}_m$ represents the local objective in client m and l_m is the empirical loss objective of $\mathcal{L}_m$ over $\mathcal{D}^m$. For simplicity, we hereafter denote the local model f_θ^m at client m and global model F_ϕ as f^m and F.

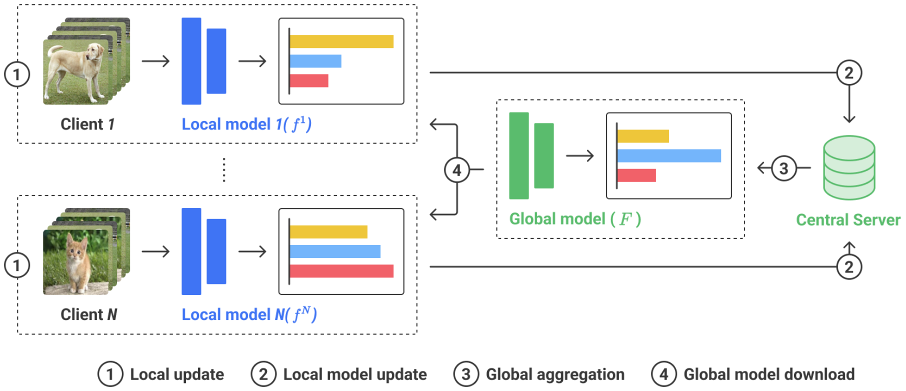


Fig. 2: Illustration of the FedAvg framework [23], which is used as the base structure of many federated systems. FedX modifies the local update process ①.

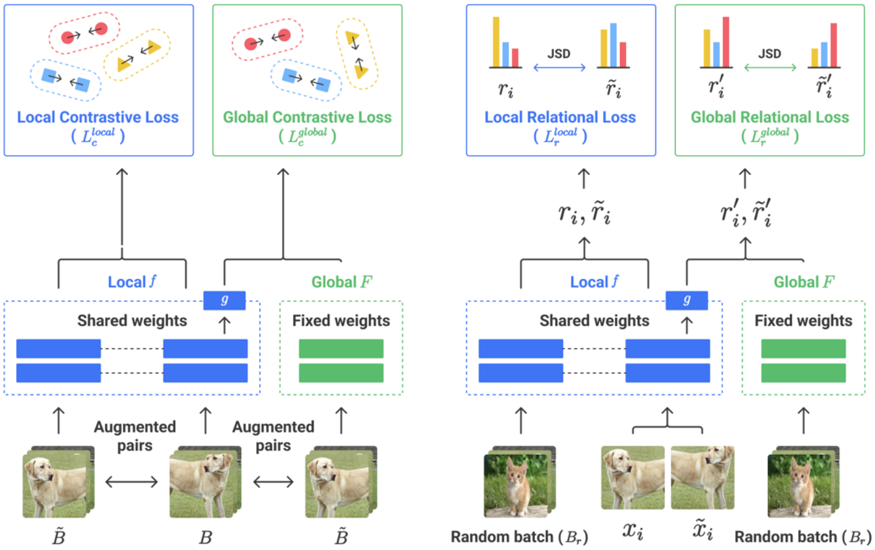
We use FedAvg [23] as the underlying structure, and the data flow is depicted in Fig. 2. Four processes run in each communication round: Process ① on local update is when each local client trains a model f^m with its data $\mathcal{D}^m$ for E local epochs; Process ② on local model upload occurs when clients share the trained model weights with the server; Process ③ on global aggregation occurs when the central server averages the received model weights and generates a shared global model F; Process ④ on global model download is when clients replace their local models with the downloaded global model (i.e., averaged weights). These processes run for R communication rounds.

FedX modifies the Process ① by redesigning loss objectives in order to distill knowledge at both the local and global scales. The following sections introduce the design components of our unsupervised federated learning model.

3.2 Local Knowledge Distillation

The first significant change takes place with local clients, whose goal is to learn meaningful representations from local data. Let us define a data pair; $\mathbf{x}_i$ and $\tilde{\mathbf{x}}_i$ be two augmented views of the same data instance. The *local contrastive loss* $L_{\mathrm{c}}^{\mathrm{local}}$ learns semantic representation by maximizing the agreement between $\mathbf{x}_i$ and $\tilde{\mathbf{x}}_i$ while minimizing the agreement of views from different instances (i.e., negative samples). We showcase the proposed contrastive loss from two of the unsupervised representation learning methods as vanilla baselines.

- SimCLR [3] utilizes a contrastive objective based on the InfoNCE loss [26]. Given a batch $\mathcal{B}$ with size n and its augmented version $\tilde{\mathcal{B}}$, each anchor has a single positive sample and considers all other $(2n-2)$ data points to be negative samples. The following is the definition of this $(2n-1)$-way instance discrimination loss, where τ is the temperature used to control the entropy



(a) Two-sided contrastive loss

(b) Two-sided relational loss

Fig. 3: The overall architecture of FedX, with the local model f^m, the projection head h^m, and the global model F at local client m. Two-sided (a) contrastive loss and (b) relational loss enable the model to learn semantic information from local data while regularizing the bias by distilling knowledge from the global model. FedX modifies the process ① on local update in Fig. 2.

value and sim(·) is the cosine similarity function between two embeddings:

$$L_{\mathrm{c}}^{\mathrm{local}} = -\log \frac{\exp(\mathrm{sim}(\mathbf{z}_i, \tilde{\mathbf{z}}_i)/\tau)}{\sum_{k \in (\mathcal{B} \cup \tilde{\mathcal{B}} - \{i\})} \exp(\mathrm{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)}, \tag{2}$$

$$\text{where } \mathbf{z}_i = f^m(\mathbf{x}_i),\ \tilde{\mathbf{z}}_i = f^m(\tilde{\mathbf{x}}_i). \tag{3}$$

∗ BYOL [6] does not train on negative samples. Instead, an asymmetric architecture is used to prevent the model from learning trivial solutions. The model f^m with a prediction layer g^m is trained to predict a view from the exponential moving average model f^m_{ema}. The loss is defined as follows:

$$L_{\mathrm{c}}^{\mathrm{local}} = \left\| \mathbf{z}_i/\|\mathbf{z}_i\| - \tilde{\mathbf{z}}_i^{\mathrm{ema}}/\|\tilde{\mathbf{z}}_i^{\mathrm{ema}}\| \right\|^2, \tag{4}$$

$$\text{where } \mathbf{z}_i = g^m \circ f^m(\mathbf{x}_i),\ \tilde{\mathbf{z}}_i^{\mathrm{ema}} = f^m_{\mathrm{ema}}(\tilde{\mathbf{x}}_i). \tag{5}$$

We consider another design aspect to help the model learn structural knowledge more effectively. Motivated by the concept of relational knowledge distillation [1,41], structural knowledge represented as relations among samples is extracted from the local model and progressively transferred back to itself. This entails selecting a set of instances at random $\mathcal{B}_r$ and computing the cosine similarity between the embeddings of two different views $\mathbf{x}_i$, $\tilde{\mathbf{x}}_i$ and random instances $\mathcal{B}_r$. We then apply the softmax function to the similarity vector to compute relationship probability distributions $\mathbf{r}_i$ and $\tilde{\mathbf{r}}_i$ (Eq. 6). In vector notation, the superscript j represents the j-th component value of a given vector.

$$\mathbf{r}_i^j = \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k\in\mathcal{B}_r}\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)}, \quad \tilde{\mathbf{r}}_i^j = \frac{\exp(\text{sim}(\tilde{\mathbf{z}}_i, \mathbf{z}_j)/\tau)}{\sum_{k\in\mathcal{B}_r}\exp(\text{sim}(\tilde{\mathbf{z}}_i, \mathbf{z}_k)/\tau)} \tag{6}$$

The above concept, *local relational loss*, is defined as the Jensen-Shannon divergence (JSD) between two relationship probability distributions $\mathbf{r}_i$ and $\tilde{\mathbf{r}}_i$ (Eq. 7). Minimizing the discrepancy between two distributions make the model to learn structural knowledge invariant to data augmentation. In contrastive learning with soft targets, this divergence loss can also be interpreted as relaxing the InfoNCE objective.

$$L_r^{\text{local}} = \frac{1}{2}\text{KL}(\mathbf{r}_i \| \mathbf{r}_i^{\text{target}}) + \frac{1}{2}\text{KL}(\tilde{\mathbf{r}}_i \| \mathbf{r}_i^{\text{target}}), \text{ where } \mathbf{r}_i^{\text{target}} = \frac{1}{2}(\mathbf{r}_i + \tilde{\mathbf{r}}_i) \tag{7}$$

The total loss term for local knowledge distillation is given in Eq. 8:

$$L_{\text{local-KD}} = L_c^{\text{local}} + L_r^{\text{local}}. \tag{8}$$

3.3 Global Knowledge Distillation

The second major change is to regularize the bias contributed by the inconsistency between local and overall data distribution. The inconsistency addresses the issue of decentralized non-IID settings, where local clients are unaware of global data distribution. Training the local model f^m will be suboptimal in this case because the local update process becomes biased towards local minimizers [42]. Such data inconsistency among local clients can be resolved by distilling knowledge on a global scale.

We consider two kinds of losses: *global contrastive loss* and *global relational loss*. Because the global model simply aggregates model weights at the local clients in FedAvg, we can think of the sample's embedding from the global model as an alternate view of the same data instance. The global contrastive loss maximizes the agreement between the views of the local and global models from the same instance while minimizing that of all other views from different instances.

Each communication round assumes that the central server sends a fixed set of averaged model weights (i.e., global model F) to the client. The batch $\mathcal{B}$ and its augmented version $\tilde{\mathcal{B}}$ are then used to train the local model f^m as in Eq. 9 with the InfoNCE loss. To match the embedding space between the local and global models, we consider an additional prediction layer h^m on top of local models. Similar method has been used in [4,6].

$$L_c^{\text{global}} = -\log\frac{\exp(\text{sim}(\mathbf{z}_i^l, \tilde{\mathbf{z}}_i^g)/\tau)}{\sum\limits_{k\in(\mathcal{B}-\{i\})}\exp(\text{sim}(\mathbf{z}_i^l, \mathbf{z}_k^l)/\tau) + \sum\limits_{k\in(\tilde{\mathcal{B}}-\{i\})}\exp(\text{sim}(\mathbf{z}_i^l, \mathbf{z}_k^g)/\tau)},$$

$$\text{where } \mathbf{z}_i^l = h^m \circ f^m(\mathbf{x}_i),\ \tilde{\mathbf{z}}_i^l = h^m \circ f^m(\tilde{\mathbf{x}}_i),\ \mathbf{z}_i^g = F(\mathbf{x}_i),\ \tilde{\mathbf{z}}_i^g = F(\tilde{\mathbf{x}}_i). \tag{9}$$

We introduce the *global relational loss* on top of the global contrastive loss. This loss is defined in the same way as the local relational loss (Eq. 7), but it includes global model embeddings. It regularizes the model by penalizing any mismatch between two augmented views over the global embedding space after the prediction layer h^m. As a result, the model maintains its local knowledge based on local data while learning augmentation-invariant knowledge using the global contrastive loss.

Given two different views $\mathbf{x}_i$, $\tilde{\mathbf{x}}_i$ and random instances $\mathcal{B}_r$, the relationship probability distributions for global relational loss, $\mathbf{r}'_i$ and $\tilde{\mathbf{r}}'_i$, are defined (Eq. 10). We again adopt the JS divergence between two relationship probability vectors $\mathbf{r}'_i$ and $\tilde{\mathbf{r}}'_i$ as the global relational loss (Eq. 11).

$$\mathbf{r}'^{j}_i = \frac{\exp(\mathrm{sim}(\mathbf{z}^l_i, \mathbf{z}^g_j)/\tau)}{\sum_{k\in\mathcal{B}_r}\exp(\mathrm{sim}(\mathbf{z}^l_i, \mathbf{z}^g_k)/\tau)}, \quad \tilde{\mathbf{r}}'^{j}_i = \frac{\exp(\mathrm{sim}(\tilde{\mathbf{z}}^l_i, \mathbf{z}^g_j)/\tau)}{\sum_{k\in\mathcal{B}_r}\exp(\mathrm{sim}(\tilde{\mathbf{z}}^l_i, \mathbf{z}^g_k)/\tau)} \tag{10}$$

$$L^{\text{global}}_r = \frac{1}{2}\mathrm{KL}(\mathbf{r}'_i \| \mathbf{r}'^{\text{target}}_i) + \frac{1}{2}\mathrm{KL}(\tilde{\mathbf{r}}'_i \| \mathbf{r}'^{\text{target}}_i), \text{ where } \mathbf{r}'^{\text{target}}_i = \frac{1}{2}(\mathbf{r}'_i + \tilde{\mathbf{r}}'_i) \tag{11}$$

The total loss for global knowledge distillation is given in Eq. 12. The overall model then combines losses from knowledge distillation at the local and global levels, as shown in Eq. 13. The detailed algorithm is described in the appendix.

$$L_{\text{global-KD}} = L^{\text{global}}_c + L^{\text{global}}_r \tag{12}$$

$$L_{\text{total-KD}} = L_{\text{local-KD}} + L_{\text{global-KD}} \tag{13}$$

4 Experiment

Using multiple datasets, we compared the performance of our model to other baselines and investigated the role of model components and hyperparameters. We also used embedding analysis to examine how the proposed model achieves the performance gain. Finally, we applied the model in a semi-supervised setting.

4.1 Performance Evaluation

Data settings. Three benchmark datasets are used. CIFAR-10 [17] contains 60,000 images of 32×32 pixels from ten classes that include airplanes, cats, and dogs. SVHN [25] contains 73,257 training images and 26,032 test images with small cropped digits of of 32×32 pixels. F-MNIST [39] contains 70,000 images of 28×28 pixels from ten classes, including dresses, shirts, and sneakers.

We used the Dirichlet distribution to enforce the non-IID property of local clients. Let $Dir_N(\beta)$ denote the Dirichlet distribution with N clients and β as the concentration parameter. We take a sample $p_{k,j}$ from $Dir_N(\beta)$ and assign class k to client j based on the sampled proportion $p_{k,j}$. With this data allocation strategy, each client will be assigned a few data samples for each class (or even none) to ensure bias. By default, N and β are to 10 and 0.5, respectively, similar to other research [19].

Table 1: Performance improvement with FedX on classification accuracy over three datasets. Both the final round accuracy and the best accuracy show that our model brings substantial improvement for all baseline algorithms.

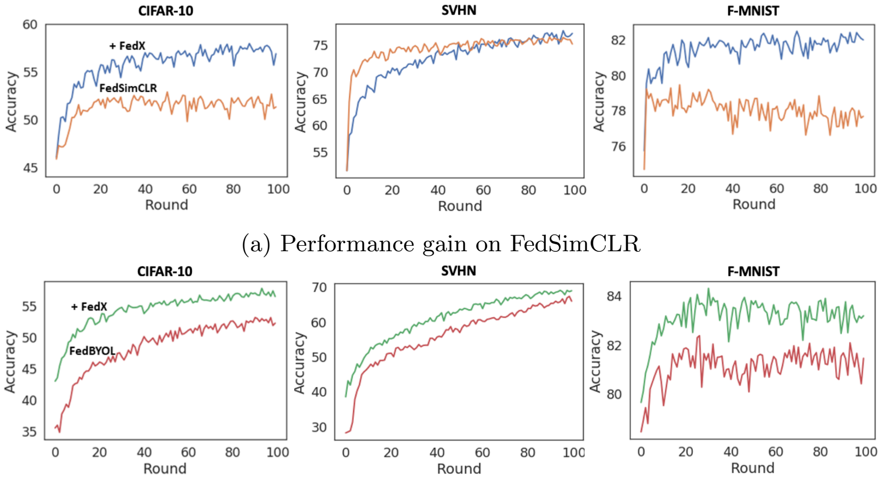
Method	CIFAR-10		SVHN		F-MNIST	
	Last	Best	Last	Best	Last	Best
FedSimCLR	51.31	52.88	75.19	76.50	77.66	79.44
+ FedX	**56.88**	**57.95**	**77.19**	**77.70**	**81.98**	**82.47**
FedMoCo	56.74	57.82	70.69	70.99	82.31	83.58
+ FedX	**58.23**	**59.43**	**73.57**	**73.92**	**83.62**	**84.65**
FedBYOL	52.24	53.14	65.95	67.32	81.45	82.37
+ FedX	**56.49**	**57.79**	**68.94**	**69.05**	**83.18**	**84.30**
FedProtoCL	51.33	52.12	49.85	50.19	81.76	**83.57**
+ FedX	**55.36**	**56.76**	**69.31**	**69.75**	**82.74**	83.34
FedU	50.79	50.79	66.02	66.22	80.59	82.03
+ FedX	**56.15**	**57.26**	**68.13**	**68.39**	**83.73**	**84.12**

Implementation details. The model was trained for 100 communication rounds, with 10 local epochs in each round. The ResNet18 backbone [9] and the SGD optimizer with a learning rate of 0.01 were used. SGD weight decay was set to 1e-5, SGD momentum was set to 0.9, and batch size was set to 128. For all objectives, the temperature τ was set as 0.1. Augmentations included random crop, random horizontal flip, and color jitter. We used four A100 GPUs.

Baselines. We implemented five baselines: (1) FedSimCLR based on SimCLR [3], (2) FedMoCo based on MoCo [8], (3) FedBYOL based on BYOL [6], and (4) FedProtoCL based on ProtoCL [18]. These are unsupervised models that are built on top of FedAvg [23]. The final baseline (5) FedU [44] is built over FedBYOL and downloads a global model by divergence-aware module (see process ④ in Figure 2). For a fair comparison, we applied the same experimental settings on these baselines, including the backbone network, optimizer, augmentation strategy, number of local epochs, and communication rounds. We used the original implementations and hyper-parameter settings for FedU. Unless otherwise specified, we refer to FedSimCLR as the representative baseline in the remainder of this section.

Evaluation. All models were compared using the linear evaluation protocol, which is a method for training a linear classifier on top of representations [42,44]. We freeze the backbone network of each trained model after training. Then, for the next 100 epochs, a new classifier is appended and trained with ground-truth labels. The top-1 classification accuracy over the test set is reported as an evaluation metric.

Results. Table 1 summarizes the performance comparison, where FedX brings meaningful performance improvements over the baseline algorithms. On average,



(a) Performance gain on FedSimCLR

(b) Performance gain on FedBYOL

Fig. 4: Performance comparison between two vanilla baselines (i.e., FedSimCLR and FedBYOL) and FedX-enhanced versions over communication rounds. FedX helps models outperform in all three benchmark datasets and continues to bring advantage with increasing communication rounds.

our model improves CIFAR-10 by 4.29 percent points (pp), SVHN by 5.52pp, and F-MNIST by 1.58pp across all baselines. One exception is F-MNIST, where FedProtoCL by itself has a slightly higher best accuracy. However, adding FedX still contributes to improved final round accuracy, implying that the model has good training stability.

We then examine how quickly the model improves baselines across the various communication rounds. Figure 4 shows the trajectory for two example baselines on FedSimCLR and FedBYOL.[5] These plots confirm that model-enhanced models outperform vanilla baselines; most plots show this benefit early in the communication rounds. We see that local bias can degrade the performance of a baseline model during the early training phase in some cases (see the F-MNIST case in Figure 4a). This is most likely due to the biased contrastive objective caused by locally sampled negatives. In contrast, adding FedX prevents such deterioration and even continues to improve accuracy as communication rounds increase.

4.2 Component Analyses

Ablation study. FedX used learning objectives at the local and global levels separately, with two types of losses: contrastive loss and relational loss. In this section, we look at ablations by removing each learning objective or loss component and testing the added value of each design choice to overall performance.

[5] Results for other baselines are presented in the Appendix.

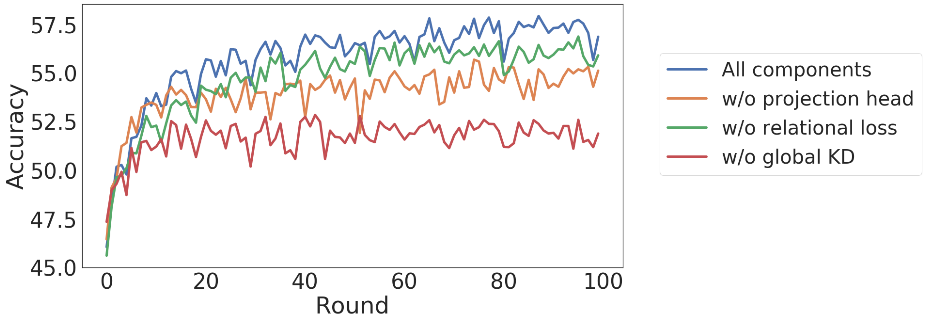


Fig. 5: Performance comparison of ablations over communication rounds for CIFAR-10. Removing any module leads to performance degradation. Ablation on contrastive loss L_c showed the best accuracy of 35.13% and hence excluded.

Table 2: Ablation results with different global-scale regularization methods. The proposed global knowledge distillation performs the best among them.

Method	CIFAR-10		SVHN		F-MNIST	
	Last	Best	Last	Best	Last	Best
$L_{\text{local-KD}}$ only	51.89	52.85	76.64	77.20	79.79	80.42
$L_{\text{local-KD}}$ + SCAFFOLD	52.73	53.20	75.18	75.52	79.45	80.36
$L_{\text{local-KD}}$ + FedProx	52.48	53.34	**77.43**	**77.79**	79.83	80.24
$L_{\text{local-KD}}$ + $L_{\text{global-KD}}$	**56.88**	**57.95**	77.19	77.70	**81.98**	**82.47**

Figure 5 plots the performance comparison of different ablations across the communication round. The complete model has the highest accuracy, implying that removing any component reduces performance. It also confirms the importance of a global knowledge distillation objective.

FedX used global knowledge distillation to convey global model knowledge and regularize the local bias caused by the inconsistency between local and overall data distribution. Several studies in supervised settings have addressed a similar challenge using extra regularization or gradient update processes. We replaced the global knowledge distillation loss ($L_{\text{global-KD}}$ – Eq. 12) with extant strategies, such as FedProx [21] or SCAFFOLD [13] and verified its efficacy. The performance comparison of different ablations across three benchmark datasets is summarized in Table 2. The findings imply that our global knowledge distillation technique is more effective than alternative designs.

Robustness test. The model's robustness is then tested by varying key hyperparameters in different simulation settings. This allows us to test the system in difficult scenarios, such as (a) when each client is only allowed to hold a small amount of data (i.e., data size $|\mathcal{D}|$), (b) when more clients participate in the federated system (i.e., client count N), and (c) when communication with the central server becomes limited and costly (i.e., the number of communication

Table 3: Analysis of accuracy on CIFAR-10 over varying hyper-parameters indicates FedX consistently enhances the baseline performance.

(a) Effect of the data size $|\mathcal{D}|$

Data size	Baseline		Baseline+FedX	
	Last	Best	Last	Best
10%	46.80	47.37	51.03	53.96
25%	48.42	49.79	52.84	54.45
50%	51.17	52.04	54.62	55.85
100%	51.31	52.88	56.88	57.95

(b) Effect of the client count N

Clients #	Baseline		Baseline+FedX	
	Last	Best	Last	Best
5	52.87	53.87	58.55	58.55
10	51.31	52.88	56.88	57.95
15	52.31	53.06	55.12	56.82
20	50.70	52.89	56.56	56.56

(c) Effect of the communication round count R

Communication round	Baseline		Baseline+FedX	
	Last	Best	Last	Best
20	52.01	52.80	56.97	56.97
50	51.95	53.53	57.29	57.29
100	51.31	52.88	56.88	57.95
200	52.79	53.23	57.35	57.58

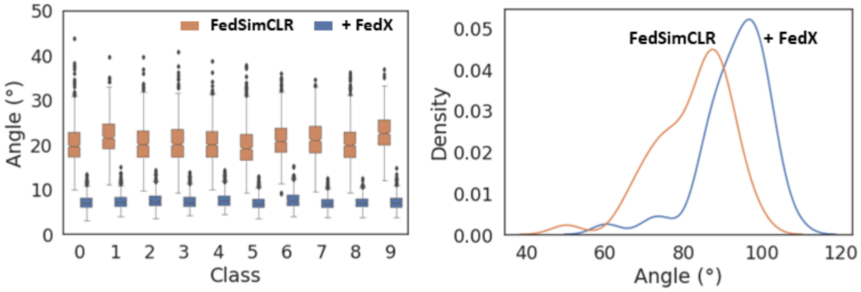
rounds R) [36]. We test how our model performs under these scenarios in Table 3. We note that when varying the communication rounds R, we also changed the number of local epochs E accordingly such that $R \times E = 1000$.

The table summarizes the effect of each hyperparameter for the baseline model (FedSimCLR) and the FedX-enhanced model. We make several observations. First, reducing the data size $|\mathcal{D}|$ degrades performance. The drop, however, is not severe and remains nearly 5pp drop even when clients only hold 10% of the data. Second, increasing the number of clients N will add complexity and degrade performance. However, when N increases from 10 to 20, the drop is only marginal near 1pp. Third, while increasing communication rounds generally provides additional benefits, the gain appears to be marginal after some rounds, as shown in the example. Regardless of these changes, FedX consistently leads to nontrivial improvements over baseline.

4.3 Analysis of the Embedding Space

We next quantitatively examine the embedding space characteristics to see how well FedX distills global knowledge into the local model and encodes data semantic structure. We calculated the angle difference between the normalized embeddings passed by local model f and global model F as a quality metric:

$$\text{Angle}(\mathbf{x}) = \arccos\left(\text{sim}(f(\mathbf{x}), F(\mathbf{x}))\right), \tag{14}$$



(a) Box plot on local vs. global models (b) Histogram on inter-class difference

Fig. 6: Embedding analysis of baseline and FedX-enhanced models on CIFAR-10 comparing the angle difference of the embedded features.

where $\mathbf{x}$ is an instance from the test data $\mathcal{D}_{\text{test}}$ and $\text{sim}(\cdot)$ is the cosine similarity function. It should be noted that a larger angle represents more significant deviance in the embedding distributions of the two models.

Figure 6a visualizes, for each of the ten classes in CIFAR-10, the angle difference between the embedding of each item between the local model and the global model computed by Eq. 14. Compared to the baseline (FedSimCLR), FedX-enhanced model reports a remarkably lower angle difference between the local and global models. This indicates that the local model can learn the refined knowledge of the global model through knowledge distillation.

When it comes to the embedding space of different class items, it is best to have a large gap. Given $\mathcal{D}^c_{\text{test}}$ as a set of instances from class c, we can compute a representative class prototype by averaging embeddings from $\mathcal{D}^c_{\text{test}}$ (Eq. 15). Then, the inter-class angle difference can be defined between any pair of class prototypes (Eq. 16). Figure 6b plots the histogram of the inter-class angle difference of every class pair, showing that FedX-enhanced models have larger angles of 93.15° on average than the baseline model of 82.36°. This demonstrates that our model can better discriminate between different class items.

$$\mathbf{z}_c = \frac{1}{|\mathcal{D}^c_{\text{test}}|} \sum_{\mathbf{x} \in \mathcal{D}^c_{\text{test}}} f(\mathbf{x}) \tag{15}$$

$$\text{Angle}(c_i, c_j) = \arccos(\text{sim}(\mathbf{z}_{c_i}, \mathbf{z}_{c_j})) \tag{16}$$

4.4 Extension to Semi-Supervised Settings

Finally, as a practical extension, consider a scenario in which each client has a small set of partially labeled data. This may be a more natural setting in many real-world federated systems [44]. To convert our model to a semi-supervised setting, we first trained it without supervision and then fine-tuned it with an additional classifier on labeled data for an additional 100 epochs. For fine-tuning, an SGD optimizer with a learning rate of 1e-3 was used.

Table 4: Classification accuracy in a semi-supervised setting on CIFAR-10. FedX enhances the baseline performance even with a small set of labels.

Label Ratio	FedSimCLR		FedMoCo		FedBYOL		FedProtoCL		FedU	
	Vanilla	FedX	Vanilla	FedX	Vanilla	FedX	Vanilla	FedX	Vanilla	FedX
1%	21.37	**23.33**	23.02	**25.18**	18.10	**21.86**	**18.44**	18.17	**21.41**	21.23
5%	30.68	**35.86**	34.24	**37.63**	29.77	**34.48**	19.64	**26.66**	32.19	**35.41**
10%	31.14	**39.40**	38.15	**39.32**	32.23	**37.89**	22.90	**27.54**	34.51	**37.51**

Table 4 shows the performance results on CIFAR-10 in the semi-supervised setting with varying label ratios of 1%, 5%, and 10%. As expected, increasing the labeling ratio from 1% to 5% brings an immediate performance gain. FedX-enhanced models outperform most cases in the semi-supervised setting for multiple baselines. Only minor exceptions can be seen with a 1% labeling rate, where our model performs similarly to the baseline. Our model, on the other hand, benefits more quickly from increasing the label ratio and can learn the data representation from distributed local clients.

5 Conclusion

This work presented the first-of-its-kind unsupervised federated learning approach called FedX. We elaborate the local update process of the common federated learning framework and the model does not share any data directly across local clients. Its unique two-sided knowledge distillation can efficiently handle data bias in a non-IID setting while maintaining privacy. It is straightforward and does not require any complex communication strategy.

The substantial performance gain of FedX shows great potential for many future applications. For example, distributed systems with strict data privacy and security requirements, such as learning patterns of new diseases across hospital data or learning tending content in a distributed IoT network, can benefit from our model. Unsupervised learning is facilitated even when local clients lack data labels and contain heterogeneous data. This versatile and robust trait makes unsupervised learning the new frontier in federated systems. We hope that our technique and implementation details will be useful in tackling difficult problems with decentralized data.

Acknowledgements

We thank Seungeon Lee and Xiting Wang for their insights and discussions on our work. This research was supported by the Institute for Basic Science (IBS-R029-C2, IBS-R029-Y4), Microsoft Research Asia, and Potential Individuals Global Training Program (2021-0-01696) by the Ministry of Science and ICT in Korea.

References

1. Bhat, P., Arani, E., Zonooz, B.: Distill on the go: Online knowledge distillation in self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2678–2687 (2021)
2. Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konečnỳ, J., Mazzocchi, S., McMahan, B., et al.: Towards federated learning at scale: System design. Proceedings of Machine Learning and Systems **1**, 374–388 (2019)
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proceedings of the International Conference on Machine Learning. pp. 1597–1607. PMLR (2020)
4. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15750–15758 (2021)
5. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. In: International Conference on Learning Representations (2018)
6. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in Neural Information Processing Systems **33**, 21271–21284 (2020)
7. Han, S., Park, S., Park, S., Kim, S., Cha, M.: Mitigating embedding and class assignment mismatch in unsupervised image classification. In: Proceedings of the European Conference on Computer Vision. pp. 768–784. Springer (2020)
8. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2020)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
10. Hinton, G., Vinyals, O., Dean, J., et al.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
11. Hsu, T.M.H., Qi, H., Brown, M.: Measuring the effects of non-identical data distribution for federated visual classification. arXiv preprint arXiv:1909.06335 (2019)
12. Ji, M., Shin, S., Hwang, S., Park, G., Moon, I.C.: Refine myself by teaching myself: Feature refinement via self-knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10664–10673 (2021)
13. Karimireddy, S.P., Kale, S., Mohri, M., Reddi, S., Stich, S., Suresh, A.T.: Scaffold: Stochastic controlled averaging for federated learning. In: Proceedings of the International Conference on Machine Learning. pp. 5132–5143. PMLR (2020)
14. Kim, K., Ji, B., Yoon, D., Hwang, S.: Self-knowledge distillation with progressive refinement of targets. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6567–6576 (2021)
15. Komodakis, N., Zagoruyko, S.: Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In: Proceedings of the International Conference on Learning Representations (2017)
16. Koratana, A., Kang, D., Bailis, P., Zaharia, M.: Lit: Learned intermediate representation training for model compression. In: Proceedings of the International Conference on Machine Learning. pp. 3509–3518. PMLR (2019)

17. Krizhevsky, A.: Learning multiple layers of features from tiny images. Tech. rep., Citeseer (2009)
18. Li, J., Zhou, P., Xiong, C., Hoi, S.: Prototypical contrastive learning of unsupervised representations. In: Proceedings of the International Conference on Learning Representations (2020)
19. Li, Q., He, B., Song, D.: Model-contrastive federated learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10713–10722 (2021)
20. Li, T., Sahu, A.K., Talwalkar, A., Smith, V.: Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine **37**(3), 50–60 (2020)
21. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. Proceedings of Machine Learning and Systems **2**, 429–450 (2020)
22. Liu, Y., Huang, A., Luo, Y., Huang, H., Liu, Y., Chen, Y., Feng, L., Chen, T., Yu, H., Yang, Q.: Fedvision: An online visual object detection platform powered by federated learning. In: Proceedings of the Association for the Advancement of Artificial Intelligence. vol. 34, pp. 13172–13179 (2020)
23. McMahan, B., Moore, E., Ramage, D., Hampson, S., Aguera y Arcas, B.: Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the Artificial Intelligence and Statistics. pp. 1273–1282. PMLR (2017)
24. Mitrovic, J., McWilliams, B., Walker, J.C., Buesing, L.H., Blundell, C.: Representation learning via invariant causal mechanisms. In: International Conference on Learning Representations (2020)
25. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning (2011)
26. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
27. Park, S., Han, S., Kim, S., Kim, D., Park, S., Hong, S., Cha, M.: Improving unsupervised image clustering with robust learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12278–12287 (2021)
28. Park, S., Kim, S., Cha, M.: Knowledge sharing via domain adaptation in customs fraud detection. arXiv preprint arXiv:2201.06759 (2022)
29. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3967–3976 (2019)
30. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: Proceedings of the International Conference on Learning Representations (2016)
31. Sheller, M.J., Reina, G.A., Edwards, B., Martin, J., Bakas, S.: Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. In: Proceedings of the International MICCAI Brainlesion Workshop. pp. 92–104. Springer (2018)
32. Tejankar, A., Koohpayegani, S.A., Pillai, V., Favaro, P., Pirsiavash, H.: Isd: Self-supervised learning by iterative similarity distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9609–9618 (2021)
33. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A.: Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on Machine learning. pp. 1096–1103 (2008)
34. Voigt, P., Von dem Bussche, A.: The EU General Data Protection Regulation (GDPR): A Practical Guide. Springer (2017)

35. Wang, H., Yurochkin, M., Sun, Y., Papailiopoulos, D., Khazaeni, Y.: Federated learning with matched averaging. In: Proceedings of the International Conference on Learning Representations (2020)
36. Wang, J., Charles, Z., Xu, Z., Joshi, G., McMahan, H.B., Al-Shedivat, M., Andrew, G., Avestimehr, S., Daly, K., Data, D., et al.: A field guide to federated optimization. arXiv preprint arXiv:2107.06917 (2021)
37. Wang, J., Liu, Q., Liang, H., Joshi, G., Poor, H.V.: Tackling the objective inconsistency problem in heterogeneous federated optimization. Advances in neural information processing systems **33**, 7611–7623 (2020)
38. Wu, Y., Wang, Z., Zeng, D., Li, M., Shi, Y., Hu, J.: Federated contrastive representation learning with feature fusion and neighborhood matching (2021), `https://openreview.net/forum?id=6LNPEcJAGWe`
39. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747 (2017)
40. Xu, Y.Z., Han, S., Park, S., Cha, M., Li, C.T.: A comprehensive and adversarial approach to self-supervised representation learning. In: 2020 IEEE International Conference on Big Data (Big Data). pp. 709–717. IEEE (2020)
41. Yang, C., An, Z., Cai, L., Xu, Y.: Mutual contrastive learning for visual representation learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 3045–3053 (2022)
42. Zhang, F., Kuang, K., You, Z., Shen, T., Xiao, J., Zhang, Y., Wu, C., Zhuang, Y., Li, X.: Federated unsupervised representation learning. arXiv preprint arXiv:2010.08982 (2020)
43. Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., Chandra, V.: Federated learning with non-iid data. arXiv preprint arXiv:1806.00582 (2018)
44. Zhuang, W., Gan, X., Wen, Y., Zhang, S., Yi, S.: Collaborative unsupervised visual representation learning from decentralized data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4912–4921 (2021)
45. Zhuang, W., Wen, Y., Zhang, X., Gan, X., Yin, D., Zhou, D., Zhang, S., Yi, S.: Performance optimization of federated person re-identification via benchmark analysis. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 955–963 (2020)