Rethinking the Representation in Federated Unsupervised Learning with Non-IID Data

Xinting Liao[1], Weiming Liu[1], Chaochao Chen[1]*, Pengyang Zhou[1], Fengyuan Yu[1], Huabin Zhu[1], Binhui Yao[1, 2], Tao Wang[2], Xiaolin Zheng[1], Yanchao Tan[3]
[1]Zhejiang University, [2]Midea Group, [3]Fuzhou University
{xintingliao, 21831010, zjuccc, zhoupy, fengyuanyu, zhb2000, xlzheng}@zju.edu.cn,
tony.yao@midea.com, tao.wang.seu@gmail.com, yctan@fzu.edu.cn

Abstract

*Federated learning achieves effective performance in modeling decentralized data. In practice, client data are not well-labeled, which makes it potential for federated unsupervised learning (`FUSL`) with non-IID data. However, the performance of existing `FUSL` methods suffers from insufficient representations, i.e., (1) representation collapse entanglement among local and global models, and (2) inconsistent representation spaces among local models. The former indicates that representation collapse in local model will subsequently impact the global model and other local models. The latter means that clients model data representation with inconsistent parameters due to the deficiency of supervision signals. In this work, we propose `FedU`^{2} which enhances generating uniform and unified representation in `FUSL` with non-IID data. Specifically, `FedU`^{2} consists of flexible uniform regularizer (**FUR**) and efficient unified aggregator (**EUA**). **FUR** in each client avoids representation collapse via dispersing samples uniformly, and **EUA** in server promotes unified representation by constraining consistent client model updating. To extensively validate the performance of `FedU`^{2}, we conduct both cross-device and cross-silo evaluation experiments on two benchmark datasets, i.e., CIFAR10 and CIFAR100.*

1. Introduction

To meet the demands of privacy regulation, federated learning (FL) [31] is boosting to model decentralized data in both academia and industry. This is because FL enables the collaboration of clients with decentralized data, aiming to develop a high-performing global model without the need for data transfer. However, conventional FL work mostly assumes that client data is well-labeled, which is less practical in real-world applications. In this work, we consider the problem of federated unsupervised learning (`FUSL`) with non-IID data [14, 43], i.e., modeling unified representation among imbalanced, unlabeled, and decentralized data.

Utilizing existing centralized unsupervised methods cannot adapt to `FUSL` which has non-IID data [44]. To mitigate it, one of the popular categories is to train self-supervised learning models, e.g., BYOL [8], SimCLR [3], and Simsiam [5], in clients, and aggregate models via accounting extremely divergent model [44, 45], knowledge distillation [9], and combining with clustering [30]. However, two coupling challenges of `FUSL`, i.e., ***CH1****: Mitigating representation collapse entanglement*, and ***CH2****: Obtaining unified representation spaces*, are not well considered.

The first challenge is that representation collapse [13] in the client subsequently exacerbates the representation of global and other local models. Motivated by regularizing Frobenius norm of representation in centralized self-supervised models [15, 21], FedDecorr [36] tackles representation collapse with the global supervision signals in federated supervised learning. But directly applying these methods to `FUSL` has three aspects of limitations. Firstly, it relies on large data batch size [30] to capture reliable distribution statistics, e.g., representation variance. Besides, regularizing the norm of high-dimensional representations inevitably causes inactivated neurons and suppresses meaningful features [18]. Moreover, clients cannot eliminate representation collapse entanglement by decorrelating representations for `FUSL` problem, once clients represent data in different representation spaces.

The second challenge refers to optimizing inconsistent client model parameters toward discrepant parameter spaces, bringing less unified representations among local models. Most of the existing `FUSL` methods aggregate participating models with the ratio of samples, i.e., FedAvg [31]. This not only fails to tackle the client shift from global optimum to local optimum, but also brings sub-optimal results [22, 33]. To mitigate this, `FUSL` meth-

*Chaochao Chen is the corresponding author.

ods maintain consistency by (1) abandoning extremely divergent clients by threshold [44, 45], (2) obtaining global supervised signal via clustering client sub-clusters [6, 30], and (3) scaling angular divergence among client models in a layer-wise way [33]. These methods either forget to adjust clients updated with inconsistent directions, or break down the performance coherence among different layers of the whole model, failing to capture unified representations.

To fill this gap, we propose a framework, i.e., $\texttt{FedU}^2$, to enhance **U**niform and **U**nified representation in $\texttt{FUSL}$ with non-IID data. To tackle **CH1**, we initially devise a **flexible uniform regularizer** (**FUR**) to prevent the sample representation collapse with no regard to data distribution and client discrepancies. In each client, **FUR** minimizes unbalanced optimal transport divergence between client data and uniform random samples, i.e., samples from the same spherical Gaussian distribution among clients. Thus it not only flexibly disperses local data representations toward ideal uniform distribution, but also avoids the representation collapse entanglement among clients without leaking privacy. To mitigate **CH 2**, we propose **efficient unified aggregator** (**EUA**) to aggregate a global model that maintains model consistency among global optimization and different local optimizations. Specifically, **EUA** formulates model aggregation as a multiple-objective optimization based on the model deviation change rates of clients. **EUA** reduces computation by searching exact solutions in the dual formulation with alternating direction methods of multipliers. Compared with conventional aggregation methods, we equivalently maintain consistent model updating based on client model deviation change, enhancing unified representations.

Summarily, we aim to enhance the representation in $\texttt{FUSL}$ by mitigating representation collapse and unifying representation generalization. (1) We enhance uniform representation by approaching data samples to spherical Gaussian distribution, which mitigates representation collapse and its subsequent entangled impacts. (2) We enhance unified representation by constraining the consistent updating of different client models. (3) To reach the above goals, we propose $\texttt{FedU}^2$ with **FUR** and **EUA**, which is agnostic and orthogonal for the backbone of self-supervised models. (4) In our empirical studies, we conduct experiments on two benchmark datasets and two evaluation settings, which extensively validate the performance of $\texttt{FedU}^2$.

2. Related Work

2.1. Federated Unsupervised Learning

To enhance $\texttt{FUSL}$ with non-IID data [24], there are two categories of efforts, i.e., (1) generating global supervised signals, and (2) enhancing unified representation. The former targets at generating global supervised signals via local-global clustering [6], and sharing data representation among clients [41, 43]. But these methods either suffer from randomness in obtaining global supervision[30], or take the risk of leaking privacy [44]. The latter enhances unified representation by adapting existing unsupervised representation methods, and tackling non-IID modeling with divergence-aware model aggregation [17, 30, 33, 44, 45]. Both FedU [44] and FedEMA [45] enhance the awareness of heterogeneity in federated self-supervised learning by divergence-aware predictor update rule, and adaptive global knowledge interpolation, respectively. However, this kind of work overlooks representation collapse in non-IID clients. Orchestra [30] utilizes local-global clustering derived from K-Fed [6] to guide self-supervised learning. This brings additional cost for clustering and is fragile to random initialization. Moreover, FedX [9] devises local relational loss to distill the invariance of data samples, and global relational loss to maintain client inconsistencies. Recently, L-DAWA [33] corrects the $\texttt{FUSL}$ optimization trajectory by measuring and scaling angular divergence among client models in a layer-wise way. However, it is hard to guarantee that the newly aggregated global model is still compatible and performant as a consistent model. Differently, the proposed $\texttt{FedU}^2$ enhances uniform and unified representation without the prior knowledge of unsupervised models, data distribution, and federated settings.

2.2. Representation Collapse

Representation collapse [15, 21] means representation vectors are highly correlated and simply span a lower-dimensional subspace, which is widely studied in metric learning [34], i.e., self-supervised learning [15], and supervised federated learning [36]. In federated supervised learning, FedDecorr [36] finds the dimensional collapse entanglement among server and client models, and decorrelates representations via regularizing the Frobenius norm of batch samples. However, FedDecorr relies on large batch size and deactivates lots of neuron parameters, degrading performance once the scale of clients increases [18]. To avoid representation collapse entanglement in $\texttt{FUSL}$, the proposed **FUR** in $\texttt{FedU}^2$ will regularize data representations to a uniform distribution that is the same among clients. In this way, decorrelating representation is not affected by the data sampling. Meanwhile, the data is uniformly dispersed into the same random distribution space, avoiding intriguing collapse impacts of client collaboration.

3. Method

3.1. Federated Unsupervised Learning Formulation

We introduce the $\texttt{FUSL}$ problem formulation and related assumptions in the following. Empirically, we assume a dataset decentralizes among K clients, i.e., $\mathcal{D} = \cup_{k\in[K]}\mathcal{D}_k$. Data distributions of different clients, i.e., $\mathcal{D}_k = \{\boldsymbol{x}_{k,i}\}_{i=1}^{N_k}$,

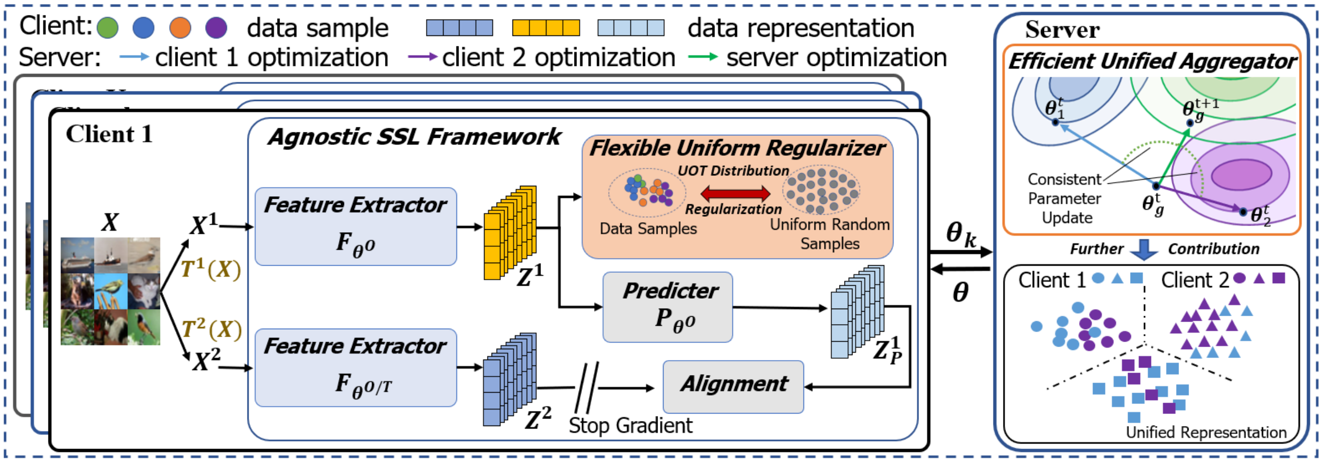


Figure 1. Framework of FedU^2. For clients with agnostic self-supervised framework, **FUR** expands non-IID data uniformly to avoid representation collapse for FUSL. **EUA** in server maintains a balanced aggregation for all client models, bringing unified representations.

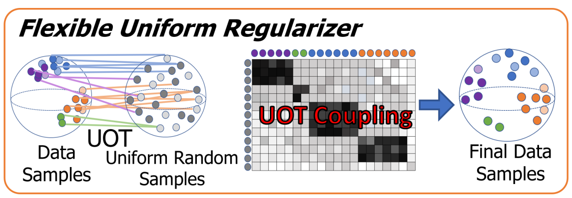


Figure 2. Example of **FUR**. Firstly, data representations collapse in part of the spherical space. Then **FUR** flexibly maps data towards spherical Gaussian distribution with unbalanced optimal transport (UOT), dispersing data uniformly.

are *unlabeled and non-IID* in practice. FUSL can be formulated as a global objective that seeks a collaborative aggregation among clients, i.e.,

$$\mathrm{argmin}_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta};\boldsymbol{p}) = \Sigma_{k=1}^{K} p_k \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_k}[\mathcal{L}_k(\boldsymbol{\theta};\boldsymbol{x})], \quad (1)$$

where $\mathcal{L}_k(\cdot)$ is the unsupervised model loss at client k, $\boldsymbol{p} = [p_1, \ldots, p_K]$, and p_k represents its weight ratio. The common aggregation approach is to assign the ratio of sample amount in client k as the weight ratio, e.g., FedAvg [31]. Nevertheless, the client with a large amount of data will dominate in aggregating, deteriorating the optimization of other clients with inconsistent local optimums [33]. Due to privacy constraints, directly aligning client local optimums with representations is forbidden [44, 45]. Therefore, it is necessary to account for a multi-objective optimal combination, i.e., restraining the consistency between global and local model parameters.

3.2. FedU^2 Overview

To address FUSL with non-IID data, i.e., Eq. (1), we propose FedU^2, whose framework overview is depicted in Fig. 1. There are one server and K clients in FedU^2, which share the same self-supervised model, e.g., Simsiam [5], SimCLR [3], and BYOL [8]. Additionally, FedU^2 contains extra *flexible uniform regularizer* (**FUR**) module, which mitigates representation collapse without requiring prior knowledge for FUSL. We introduce the unsupervised local modeling at each client k, and then illustrate the communication between clients and server. For a batch of image data $\boldsymbol{X}$ at client k, we augment them with two transformations, i.e., $\boldsymbol{X}^v = T^v(\boldsymbol{X})$ for $T^v \sim \mathcal{T}$ with $\mathcal{T}$ denoting transformation set and $v \in \{1, 2\}$. And feature extractor represents two views of augmented data with d-dimensional l_2-normalized representations, i.e., $\boldsymbol{Z}^v = \mathcal{F}_{\boldsymbol{\theta}_k}(\boldsymbol{X}^v)$. Then for each view of sample representations, we maximize its representation space via approximating uniform Gaussian distribution in **FUR**. Meanwhile, we align two views of feature representations by predictor module (if available) and alignment module.

In one communication round, every participating client k uploads its model parameters $\boldsymbol{\theta}_k$ to server. Next, server with *efficient unified aggregator* (**EUA**) module, first formulates the client model aggregation as a multi-objective optimization based on different client model deviation change rates, and searches for a balanced model combination. Server further restrains client parameters in a consistent space, which not only enhances the consistency between global optimum and local optimums, but also captures unified representation for data of the same class but different clients. This communication between server and clients iterates until the performance of FedU^2 converges.

3.3. FUR for Mitigating Representation Collapse

Representation collapse is a long-standing issue due to its intriguing phenomenon, e.g., constant collapse and partial/full dimensional collapse [8, 15]. In federated learning, representation collapse not only degrades the performance of local clients, but also intricately affects the representation of global and local models [36]. Besides, lacking labels, clients represent data samples to the space around local optimums, where decorrelating sample representations of limited client data suppresses capturing useful features [18].

Without ground truth labels, self-supervised learning not only keeps the invariance of the same sample with different augmentations, but also expands the uniformity of different representations to avoid representation collapse [40]. Given a batch of B representations, i.e., $\boldsymbol{Z}_B^1 = \{\boldsymbol{z}_i^1\}_{i\in[B]}$ and $\boldsymbol{Z}_B^2 = \{\boldsymbol{z}_i^2\}_{i\in[B]}$, we train the self-supervised model by minimizing the total objective as below:

$$\mathcal{L} = \mathbb{E}_{T^1,T^2\sim\mathcal{T}}\ell_a\left(\boldsymbol{Z}_B^1, \boldsymbol{Z}_B^2\right) + \lambda_U\left(\ell_u\left(\boldsymbol{Z}_B^1\right) + \ell_u\left(\boldsymbol{Z}_B^2\right)\right), \tag{2}$$

where ℓ_a and ℓ_u are alignment term and uniformity term, respectively. $\lambda_U > 0$ is a hyperparameter that balances the two terms. The alignment term keeps data samples of the same class to be clustered, while others are separable, i.e., $\ell_a\left(\boldsymbol{Z}_B^1, \boldsymbol{Z}_B^2\right) := \frac{1}{B}\sum_{i\in[B]}\left\|\boldsymbol{z}_i^1 - \boldsymbol{z}_i^2\right\|_2^2$.

The crucial of mitigating representation collapse is to enhance the representation uniformity [40]. To relieve reliance on prior knowledge of client data, we regularize local sample representations to a random distribution with high entropy. Specifically, we select samples following the spherical Gaussian distribution, i.e., $\boldsymbol{s} \sim \mathcal{N}(0,1), s.t., \|\boldsymbol{s}\| = 1$, as the prior. In this way, mitigating representation collapse in FUSL not only avoids leaking privacy, but also disentangles the collapse impacts among clients. Then **FUR** regularizes the divergence between the data representations $\boldsymbol{Z}_B^v$ and a set of random samples following spherical Gaussian distribution $\boldsymbol{S}_B = \{\boldsymbol{s}_i\}_{i\in[B]}$:

$$\ell_u\left(\boldsymbol{Z}_B^v\right) := \mathrm{Div}(\boldsymbol{Z}_B^v, \boldsymbol{S}_B). \tag{3}$$

Thus the uniform term ℓ_u disperses uniformly to avoid representation collapse without repulsing in instance-based contrastive learning.

Since the client data is non-IID, it will break the class separation when strictively constraining sample representation to approach random instances [39]. A more flexible method is to match data samples and random Gaussian samples with arbitrary or proportional masses, i.e., leaving the sampling coupling with lower uncertainties unmatched. And unbalanced Optimal Transport (UOT) [1, 35] is one of the effective resolutions. UOT computes the transport mapping [26, 28, 29] between two sample masses of different distributions under the soft marginal constraints, e.g., l_2-normalization between the predicted margin and the ground truth margin. Given marginal constraints $\boldsymbol{a}$ and $\boldsymbol{b}$ for data and Gaussian distribution respectively, we formulate a UOT problem that searches a coupling matrix $\boldsymbol{\pi}$ with minimal distribution divergence:

$$\min_{\pi_{i,j}\geq 0} \ell_u(\boldsymbol{Z}^v) = \mathrm{vec}(\boldsymbol{C})^\top \mathrm{vec}(\boldsymbol{\pi}) + \frac{\tau_a}{2}\left\|\boldsymbol{\Phi}_\mathrm{r}\,\mathrm{vec}(\boldsymbol{\pi}) - \boldsymbol{a}\right\|_2^2 + \frac{\tau_b}{2}\left\|\boldsymbol{\Phi}_\mathrm{c}\,\mathrm{vec}(\boldsymbol{\pi}) - \boldsymbol{b}\right\|_2^2, \tag{4}$$

where cost matrix $\boldsymbol{C}_{ij} = \|\boldsymbol{Z}_i^v - \boldsymbol{S}_j\|^2$, and $\boldsymbol{\Phi}_\mathrm{r} = \boldsymbol{I}_N \otimes \mathbb{1}_N^\top$ ($\boldsymbol{\Phi}_\mathrm{c} = \mathbb{1}_M^\top \otimes \boldsymbol{I}_M$) are indicators for row-wise (column-wise) Kronecker multiplication with $\boldsymbol{I}$ denoting identity matrix. **Optimization.** Denoting $\tau_a\boldsymbol{\Phi}_\mathrm{p}^\top\boldsymbol{\Phi}_\mathrm{r} + \tau_b\boldsymbol{\Phi}_c^\top\boldsymbol{\Phi}_c = \boldsymbol{Q}$ and $\mathrm{vec}(\boldsymbol{C}) - \tau_a\boldsymbol{\Phi}_\mathrm{r}^\top\boldsymbol{a} - \tau_b\boldsymbol{\Phi}_c^\top\boldsymbol{b} = \boldsymbol{w}$, we rewrite Eq. (4) as a positive definite quadratic form:

$$\min_{\pi_{i,j}\geq 0} \ell_u(\boldsymbol{Z}^v) = \frac{1}{2}\mathrm{vec}(\boldsymbol{\pi})^\top\boldsymbol{Q}\,\mathrm{vec}(\boldsymbol{\pi}) + \boldsymbol{w}^\top\mathrm{vec}(\boldsymbol{\pi}) + \boldsymbol{\Omega}, \tag{5}$$

where constant $\boldsymbol{\Omega} = \frac{1}{2}(\tau_a\boldsymbol{a}^\top\boldsymbol{a} + \tau_b\boldsymbol{b}^\top\boldsymbol{b})$. Next we optimize $\boldsymbol{\pi}$ via steepest gradient descent as bellow:

$$\mathrm{vec}\left(\boldsymbol{\pi}^{(\mathrm{new})}\right) = \max\left(0, \mathrm{vec}\left(\boldsymbol{\pi}^{(\mathrm{old})}\right) - \eta^*\frac{\partial\ell_u(\boldsymbol{Z}^v)}{\partial\,\mathrm{vec}\left(\boldsymbol{\pi}^{(\mathrm{old})}\right)}\right), \tag{6}$$

with $\eta^* = \frac{\left(\boldsymbol{Q}_\mathrm{vec}\left(\boldsymbol{\pi}^{(\mathrm{old})}\right)+\boldsymbol{w}\right)^\top\left(\boldsymbol{Q}_\mathrm{vec}\left(\boldsymbol{\pi}^{(\mathrm{old})}\right)+\boldsymbol{w}\right)}{\left(\boldsymbol{Q}_\mathrm{vec}\left(\boldsymbol{\pi}^{(\mathrm{old})}\right)+\boldsymbol{w}\right)^\top\boldsymbol{Q}\left(\boldsymbol{Q}\,\mathrm{vec}\left(\boldsymbol{\pi}^{(\mathrm{old})}\right)+\boldsymbol{w}\right)}$. Finally, we obtain the uniform UOT divergence by taking the optimal $\boldsymbol{\pi}^*$ back to Eq. (4). **FUR** minimizes UOT divergence to regularize data samples approaching the spherical Gaussian distribution. Note that the spherical Gaussian distribution maximizes its entropy and distributes its samples uniformly. The mapped data representations enjoy the above nice properties of spherical Gaussian distribution and further mitigate the representation collapse entanglement.

3.4. EUA for Generalizing Unified Representation

Due to non-IID client data, clients optimize to their local optimums with inconsistent model parameters, causing inconsistent even conflicting model deviations from server to clients. Without the guidance of supervision signals, i.e., data labels, this problem further exacerbates in representing data of the same class but different clients, towards inconsistent spaces. Thus it is vital to constrain the consistency among client models in parameter spaces, which further guarantees unified representations.

In round t, the impact of global aggregation on k-th local optimization can be measured with the model deviation change rate, i.e.,

$$c_k\left(\eta, \boldsymbol{d}^t\right) = \frac{u_k\left(\boldsymbol{\theta}_g^t\right) - u_k\left(\boldsymbol{\theta}_g^{t+1}\right)}{u_k\left(\boldsymbol{\theta}_g^t\right)} \approx \eta_g\nabla\log u_k(\boldsymbol{\theta}_g^t)\boldsymbol{d}^t, \tag{7}$$

where $u_k(\boldsymbol{\theta}_g^t) = \|\boldsymbol{\theta}_g^t - \boldsymbol{\theta}_k^t\|^2$ is the model deviation from server to client k [12, 16], and global model optimization is $\boldsymbol{\theta}_g^{t+1} = \boldsymbol{\theta}_g^t - \eta\boldsymbol{d}^t$ with updating direction $\boldsymbol{d}^t$ and step size η. Overlooking inconsistent model deviations, global aggregated model inevitably gets close to a subset of clients while deviating from others. It corresponds that clients get close to the global model increase the model deviation change rate, and clients away decrease it [12, 32]. Motivated by this, we seek the clients with the worst model deviation change rate, and correct the global optimization with a direction maximizing the overall worst model deviation change rate. This can be formulated as a multi-objective optimization, which benefits for mitigating the inconsistencies and conflicts among clients [12, 42], i.e.,

$$\begin{aligned} &\max_{\boldsymbol{d}^t} \min_{\boldsymbol{p}\in\mathbb{S}^K} \eta_g \sum_{k=1}^{K} p_k \nabla \log u_k(\boldsymbol{\theta}_g^t)\boldsymbol{d}^t, \\ &s.t. \|\boldsymbol{d}^t\|^2 \leq 1, \boldsymbol{p}^\top \mathbf{1} = 1, p_k \geq 0, \end{aligned} \tag{8}$$

where $\boldsymbol{p}$ denotes the weights for different clients.

Optimization. Adding the constraints as Lagrange multipliers, Eq. (8) can be rewritten as:

$$\max_{\boldsymbol{d}^t} \min_{\boldsymbol{p}} J = \eta_g \left\langle \sum_{k=1}^{K} p_k \nabla \log u_k(\boldsymbol{\theta}_g^t), \boldsymbol{d}^t \right\rangle - \frac{\phi}{2}(\|\boldsymbol{d}^t\|^2 - 1). \tag{9}$$

Differentiating Eq. (9) with regarding to $\boldsymbol{d}^t$, we have $\boldsymbol{d}^* = \frac{\eta_g}{\phi}(\nabla \log \boldsymbol{u})^\top \boldsymbol{p}$, where $\boldsymbol{u} = (u_1(\boldsymbol{\theta}_g^t), \ldots, u_K(\boldsymbol{\theta}_g^t))$. Taking back to Eq. (9), we can obtain its strong dual form with dual variable $\boldsymbol{p}$:

$$\boldsymbol{J} = \min_{\boldsymbol{p}} \frac{\eta_g^2}{2\phi}\|(\nabla \log \boldsymbol{u})^\top \boldsymbol{p}\|^2 = \min_{\boldsymbol{p}} \frac{\eta_g^2}{2\phi}\boldsymbol{p}^\top \boldsymbol{G}\boldsymbol{p}, \tag{10}$$

where $\boldsymbol{G} = (\nabla \log \boldsymbol{u})^\top(\nabla \log \boldsymbol{u})$. Then we can rewrite it as an augmented Lagrangian form,

$$\boldsymbol{J} = \min_{\boldsymbol{p}} \frac{\eta_g^2}{2\phi}\boldsymbol{p}^\top \boldsymbol{G}\boldsymbol{p} + \mu(\boldsymbol{p}^\top \mathbf{1} - 1) + \frac{\rho}{2}\|\boldsymbol{p}^\top \mathbf{1} - 1\|^2, \tag{11}$$

where μ denotes the Lagrange multipliers. This can be iteratively solved by alternating direction method of multipliers (ADMM) algorithm [7, 27], i.e., fixing μ to optimize $\boldsymbol{p}$, and vice versa:

$$\begin{cases} \boldsymbol{p} = \max(0, (\frac{\eta_g^2}{\phi}\boldsymbol{G} + \rho\boldsymbol{I})^{-1}(\rho\boldsymbol{I} - \mu\boldsymbol{I})) \\ \mu \leftarrow \mu + \rho(\boldsymbol{p}^\top \mathbf{1} - 1) \end{cases} \tag{12}$$

The ADMM iteration guarantees exact solution in minimal computation complexity, then the global model updates towards $\boldsymbol{d}^*$ with step size η.

Theorem 1 (Optimization consistency of model deviations)**.** *Rethinking the Lagrangian of dual form in Eq. (10),*

$$\boldsymbol{J} = \min_{\boldsymbol{p}} \frac{\eta_g^2}{2\phi}\|(\nabla \log \boldsymbol{u})^\top \boldsymbol{p}\|^2 + \lambda \boldsymbol{p}^\top \mathbf{1}, \tag{13}$$

it holds $\nabla \log(u_i(\boldsymbol{\theta}_g^t)) = \nabla \log(u_j(\boldsymbol{\theta}_g^t)), \forall i \neq j \in [K]$.

Proof. We provide the proof details in Appendix A.1. □

After the convergence of global and local optimization, **EUA** balances the model deviation change rate among all clients, making the global aggregation improves all model equivalently. Therefore, all models optimize towards a consistent parameter spaces, obtaining unified representation.

3.5. Overall Algorithm and Convergence Analysis

We describe the overall algorithm of FedU^2 in Algo. 1. In detail, the server collaborates with clients in steps 1:10. After collecting participating client models in step 8, server uses **EUA** to reach a consistent model updating and obtain unified representations. The client executes self-supervised modeling in steps 11:21, where **FUR** enhances uniform representations to avoid collapse entanglement in step 17.

Convergence Analysis. In the following, we take four mild assumptions [23], and provide the generalization bounds of model divergence and overall convergence error.

Assumption 1. *Let $F_k(\boldsymbol{\theta}_k)$ be the expected model objective for client k, and assume $F_1, \cdots, F_K$ are all L-smooth, i.e., for all $\boldsymbol{\theta}_k$, $F_k(\boldsymbol{\theta}_k) \leq F_k(\boldsymbol{\theta}_k) + (\boldsymbol{\theta}_k - \boldsymbol{\theta}_k)^\top \nabla F_k(\boldsymbol{\theta}_k) + \frac{L}{2}\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k\|_2^2$.*

Assumption 2. *Let $F_1, \cdots, F_N$ are all μ-strongly convex: for all $\boldsymbol{\theta}_k$, $F_k(\boldsymbol{\theta}_k) \geq F_k(\boldsymbol{\theta}_k) + (\boldsymbol{\theta}_k - \boldsymbol{\theta}_k)^\top \nabla F_k(\boldsymbol{\theta}_k) + \frac{\mu}{2}\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k\|_2^2$.*

Assumption 3. *Let ξ_k^t be sampled from the k-th client's local data uniformly at random. The variance of stochastic gradients in each client is bounded: $\mathbb{E}\left\|\nabla F_k\left(\boldsymbol{\theta}_k^t, \xi_k^t\right) - \nabla F_k\left(\boldsymbol{\theta}_k^t\right)\right\|^2 \leq \sigma_k^2$.*

Assumption 4. *The expected squared norm of stochastic gradients is uniformly bounded, i.e., $\mathbb{E}\left\|\nabla F_k\left(\boldsymbol{\theta}_k^t, \xi_k^t\right)\right\|^2 \leq V^2$ for all $k = 1, \cdots, N$ and $t = 1, \cdots, T-1$*

Lemma 1 (Bound of Client Model Divergence)**.** *With assumption 4, η_t is non-increasing and $\eta_t < 2\eta_{t+E}$ (learning rate of t-th round and E-th epoch) for all $t \geq 0$, there exists $t_0 \leq t$, such that $t - t_0 \leq E - 1$ and $\boldsymbol{\theta}_k^{t_0} = \boldsymbol{\theta}^{t_0}$ for all $k \in [N]$. It follows that*

$$\mathbb{E}\left[\sum_{k}^{K} p_k\|\boldsymbol{\theta}^t - \boldsymbol{\theta}_k^t\|^2\right] \leq 4\eta_t^2(E-1)^2V^2. \tag{14}$$

Proof. We provide the proof details in Appendix A.2. □

Theorem 2 (Convergence Error Bound)**.** *Let assumptions 1-4 hold, and L, μ, σ_k, V be defined therein. Let $\kappa = \frac{L}{\mu}, \gamma = \max\{8\kappa, E\}$ and the learning rate $\eta_t = \frac{2}{\mu(\gamma+t)}$. The* FedU^2 *with full client participation satisfies*

$$\mathbb{E}\left[F\left(\overline{\boldsymbol{\theta}}^t\right)\right] - F^* \leq \frac{\kappa}{\gamma+t}\left(\frac{2B}{\mu} + \frac{\mu(\gamma+1)}{2}\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^*\|^2\right),$$

where $B = 4(E-1)^2V^2 + K + 2\Gamma$.

Proof. We provide the proof details in Appendix A.3. □

Algorithm 1 Training procedure of `FedU`^{2}

Input: Batch size B, communication rounds T, number of clients K, local steps E, dataset $\mathcal{D} = \cup_{k\in[K]}\mathcal{D}_k$
Output: Global model $\boldsymbol{\theta}^T$

Server executes():
Initialize $\boldsymbol{\theta}^0$ with random distribution
for $t = 0, 1, ..., T-1$ **do**
  for $k = 1, 2, ..., K$ **in parallel do**
    Send $\boldsymbol{\theta}^t$ to client k
    $\boldsymbol{\theta}_k^{t+1} \leftarrow$ **Client executes**$(k, \boldsymbol{\theta}^t)$
  end for
  EUA optimize Eq. (10) for $\boldsymbol{p}^*$ and update global model $\boldsymbol{\theta}^{t+1}$ with optimal direction $\boldsymbol{d}^t$ in Eq. (9)
end for
return $\boldsymbol{\theta}^T$
Client executes$(k, \boldsymbol{\theta}^t)$**:**
Assign global model to the local model $\boldsymbol{\theta}_k^t \leftarrow \boldsymbol{\theta}^t$
for each local epoch $e = 1, 2, ..., E$ **do**
  for batch of samples $\boldsymbol{X}_{k,B} \in \mathcal{D}_k$ **do**
    Augment samples $\boldsymbol{X}_{k,B}^v = T_{v\in\{1,2\}}^v(\boldsymbol{X}_{k,B})$
    Feature extraction $\boldsymbol{Z}_{k,B}^v \leftarrow \mathcal{F}_{\boldsymbol{\theta}_k^e}(\boldsymbol{X}_{k,B}^v)$
    FUR enhances the uniformity of $\boldsymbol{Z}_{k,B}^v$ by Eq. (4)
    Compute total loss in Eq. (2) and update $\boldsymbol{\theta}_k^e$
  end for
end for
return $\boldsymbol{\theta}_k^E$ to server

4. Experiments

4.1. Experimental Setups

Datasets. We adopt two benchmark datasets, i.e., CIFAR10 and CIFAR100 [20], to evaluate `FedU`^{2}. Both datasets have 50,000 training samples and 10,000 test samples, but differ in the number of classes. Following FedEMA [45], we simulate non-IID data distribution in K clients by assuming class priors follow the Dirichlet distribution parameterized with non-IID degree, i.e., α [11]. The smaller α simulates the more non-IID federated setting. We conduct extensive experiments in both cross-silo ($K = 10$) and cross-device ($K = 100$) settings to validate performance generalization.

Comparison Methods. We compare `FedU`^{2} with three categories of approaches , i.e., (1) combining the existing centralized self-supervised model with FedAvg [31]: **FedSimsiam**, **FedSimCLR**, and **FedBYOL**, (2) the state-of-the-art `FUSL` methods: **FedU** [44], **FedEMA** [45], **FedX** [9], **Orchestra** [30], and **L-DAWA** [33], and (3) adapting existing federated supervised learning models solving representation collapse to `FUSL`: **FedDecorr** [36]. Firstly, we evaluate the representation performance of the above methods with their best-performing models on both cross-silo and cross-device settings on CIFAR10 and CIFAR100 ($\alpha = 0.1$). Secondly, we study the effectiveness of different methods with the same model, i.e., BYOL, for different non-IID degrees [25], i.e., $\alpha = \{0.1, 0.5, 5\}$. We evaluate the pre-trained encoder model via the accuracy of KNN [4], standard linear probing [2], and semi-supervised methods (i.e., fine-tuning 1% and 10% labeled data). All of the above metrics illustrate better performance when the values are higher.

Implemental Details. We conduct image augmentation for SimCLR, BYOL, and SimSiam, following their original papers. We adopt ResNet18 [10] as an encoder module, choose Projector/Predictor architectures like original papers, and optimize each model 5 local epochs per communication round until converging. We set all datasets with batch size as 128 and embedding dimension as 512. To obtain fair comparisons, we conduct every experiment for each method with its best hyper-parameters, and report the average result of 3 repetitions. We choose Adam [19] as the optimizer for each local model, and SGD [37] for updating the global model. We set the uniformity effect $\lambda_U = 0.1$, the soft margin constraints $\tau_a = \tau_b = 0.8$, the coefficient of constraints $\phi = 0.1$, and the coefficient in ADMM $\rho = 1$.

4.2. Experimental Results

Representation Performance Comparison. Firstly, we follow the existing `FUSL` methods [30, 33, 45], to evaluate the performance of pre-trained models learned in Tab. 1. We group the state-of-the-art methods in terms of their best-performing models, i.e., Simsiam, SimCLR, and BYOL. **For the first group**, we can observe that FedDecorr performs the worst, especially on CIFAR100 cross-device task. It indicates that directly avoiding collapse via decorrelating a batch of data representations is unsuitable for `FUSL` with limited data and severely heterogeneous data distribution. **In terms of the second group**, compared with FedX, L-DAWA performs better on CIFAR100, while is less competitive on CIFAR10. We can conclude that: (1) L-DAWA can better control model divergence when clients have inconsistent optimums, and (2) L-DAWA fails to obtain discriminative representations since it takes no action to representation collapse. **On mentioned the third group**, Orchestra captures global supervision signals to guide data representation, whose effectiveness suffers from randomness. **In general**, directly combining existing self-supervised model with FedAvg cannot tackle `FUSL` with non-IID data well. Cross-device simulation on CIFAR100 is so challenging that some existing methods fail dramatically. Moreover, `FedU`^{2} is agnostic to self-supervised model and performs better than existing work, which validates the superiority of enhancing uniform and unified representations.

Effect of Heterogeneity on Generalization. Next, we report the KNN-accuracy of cross-silo methods on CIAFR10 and CIFAR100 in Tab. 2, for validating the performance generalization. We choose the same model, i.e., BYOL,

Table 1. Accuracy (%) of linear probing (LP), fine-tuning (FT) 1%, and 10% labeled data on CIFAR10 and CIFAR100 ($\alpha = 0.1$).

Dataset	CIFAR10						CIFAR100					
Setting $\alpha = 0.1$	Cross-Device (K=100)			Cross-Silo (K=10)			Cross-Device (K=100)			Cross-Silo (K=10)		
Method \Evaluation	LP	FT 1%	FT 10%	LP	FT 1%	FT 10%	LP	FT 1%	FT 10%	LP	FT 1%	FT 10%
FedSimsiam	60.49	44.45	70.46	70.61	57.60	69.88	31.91	12.58	37.33	49.81	21.64	43.08
FedDecorr-Simsiam	43.18	35.15	58.68	74.56	65.21	80.10	17.09	5.36	20.60	47.93	20.53	45.21
FedU^2-Simsiam	**68.50**	**56.43**	**75.33**	**84.92**	**77.11**	**85.21**	**35.59**	**13.08**	**38.22**	**56.55**	**31.42**	**48.75**
FedSimCLR	65.76	51.18	68.33	75.65	62.86	76.15	37.09	11.73	31.97	51.62	19.47	41.60
L-DAWA-SimCLR	65.63	49.66	69.89	75.48	63.4	78.66	41.28	13.52	36.56	51.11	21.07	45.02
FedX-SimCLR	67.33	49.96	70.18	78.29	65.03	79.43	38.11	11.18	33.96	51.67	19.65	42.38
FedU^2-SimCLR	66.49	**51.77**	**70.76**	**82.37**	**69.84**	**82.39**	**41.56**	**14.32**	**36.90**	**56.56**	**26.11**	**47.99**
FedBYOL	61.46	54.36	74.01	83.29	74.04	81.40	28.27	10.43	34.90	48.78	19.79	42.82
FedU-BYOL	60.15	53.53	74.62	82.33	69.24	83.37	28.09	10.46	36.06	58.02	28.38	48.12
FedEMA-BYOL	62.27	54.91	74.76	82.17	71.37	83.78	28.40	10.63	35.62	57.25	30.03	50.33
Orchestra	38.66	41.62	62.97	83.53	78.44	85.40	17.91	6.96	23.40	51.31	26.36	48.85
FedU^2-BYOL	**67.62**	**54.74**	**74.93**	**85.58**	**78.64**	**86.24**	**38.09**	**13.16**	**36.87**	**59.71**	**34.83**	**53.87**

Table 2. KNN accuracy (%) of different α on CIFAR10 and CIFAR100 for cross-silo settings.

Dataset	CIFAR10			CIFAR100		
Method \ α	0.1	0.5	5	0.1	0.5	5
FedBYOL	76.12	77.23	82.71	38.13	43.93	45.30
FedU-BYOL	79.09	79.68	82.75	51.31	51.81	52.05
FedEMA-BYOL	80.32	82.01	82.80	53.18	53.18	53.28
FedDecorr-BYOL	76.76	79.66	81.09	49.87	49.54	52.07
L-DAWA-BYOL	65.40	66.17	82.82	23.09	51.10	52.25
FedX-BYOL	50.94	40.96	41.05	15.83	16.35	16.89
Orchestra	79.25	76.78	76.30	38.52	37.17	34.93
FedU^2-**FUR**-BYOL	81.04	82.18	83.45	53.45	53.86	54.34
FedU^2-**EUA**-BYOL	80.93	82.14	84.01	53.20	53.72	54.61
FedU^2-BYOL	**81.39**	**82.21**	**84.79**	**53.87**	**54.07**	**55.06**

to be comparable among all FUSL methods. We can discover that: (1) Most FUSL methods increase their performance when the non-IID degree α increases, and the performance variances among different methods increase with the decreasing of α. (2) FedEMA-BYOL is not sensitive to the non-IID degrees, while Orchestra behaves on opposite. This states that capturing global supervision signals to guide local representation suffers from clustering randomness. (3) FedU^2 performs the best among all tasks, even in $\alpha = 0.1$, illustrating its performance generalization.

Ablation Studies. In Tab. 2, we also consider two variants of FedU^2: (1) FedU^2 removes **FUR**, i.e., FedU^2-**FUR**, (2) FedU^2 removes **EUA**, i.e., FedU^2-**EUA**, to study the effect of each module. From Tab. 2, we can see that either applying **FUR** or **EUA** can enhance representations, since they have better performance than the existing FUSL methods. Compared with FedU^2, FedU^2-**FUR** and FedU^2-**EUA** drop KNN accuracy slightly, validating the effectiveness of tackle two challenges, i.e., representation collapse entanglement, and generating unified representations. FedU^2-**FUR** performs better than FedU^2-**EUA** when $\alpha = 0.1$, while gets worse when $\alpha = 5$.

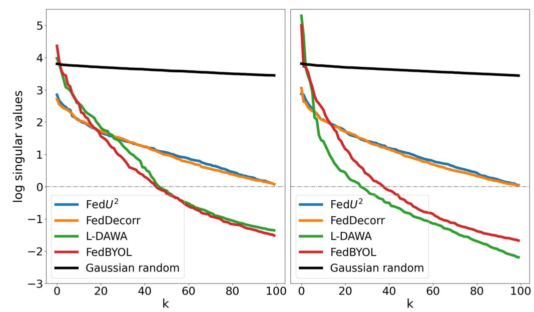
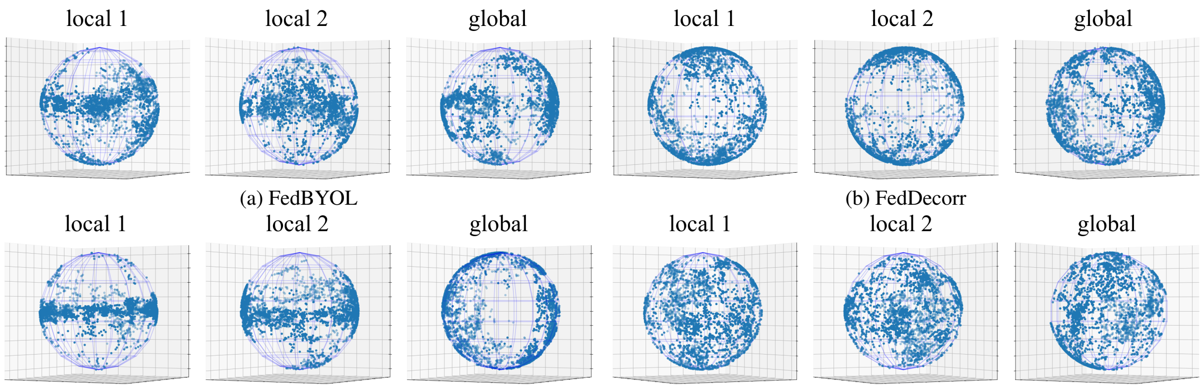


Figure 3. Top k log singular values of the covariance matrix of global model (left) and local model (right) representations.

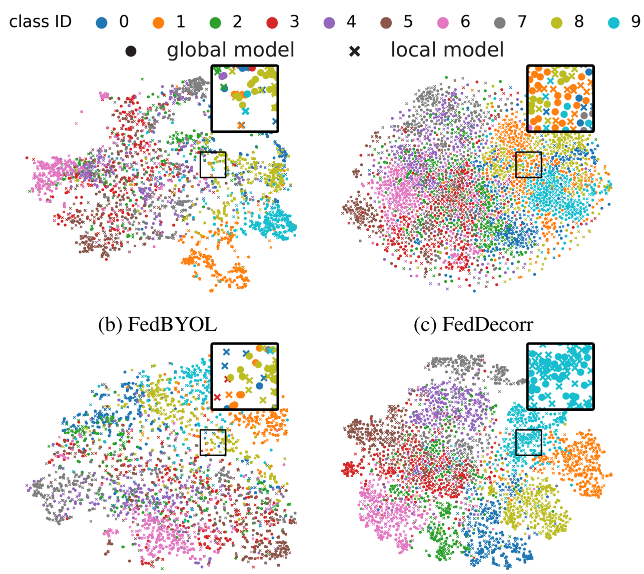
4.3. Representation Visualization

Analysis of Representation Collapse Entanglement. To study the representation collapse entanglement caused by non-IID data, we capture the representation covariance matrices of N test data points on CIFAR10 from both the global and local BYOL models of FUSL methods, i.e., FedDecorr, L-DAWA, FedBYOL, and FedU^2. And we utilize the singular value decomposition on each of the representation covariance matrices, and visualize the top-100 singular values in Fig. 3. Both L-DAWA and FedBYOL suffer from severe representation collapse, because they have less singular values beyond 0 than FedDecorr and FedU^2. The representation collapses in global model and local model are consistent, proving that collapse impacts are entangled intricately. Compared with the singular values decomposed from the covariance matrix of Gaussian random samples, the singular values of FedDecorr and FedU^{2}are not similar. Because a fully uniform distribution breaks down the alignment effect and deteriorates clustering. Furthermore, in Fig. 4, we visualize the representation collapse on 3-D spherical space, where the existing FUSL methods leave evident blank space and suffer from collapse entanglements.



(a) FedBYOL (b) FedDecorr (c) L-DAWA (d) FedU^2

Figure 4. The representations collapse issue on the sphere using BYOL model (on CIFAR10 $\alpha = 0.1$ Cross-silo). The more blank representation space means the more severe collapse issue is.

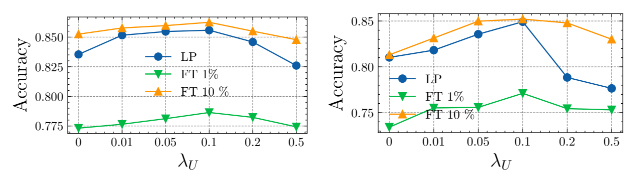


(b) FedBYOL (c) FedDecorr (d) L-DAWA (e) FedU^2

Figure 5. The distributions of data representations using global and local BYOL model (on CIFAR10 $\alpha = 0.1$ Cross-silo).

Analysis of Unified Representation. We also use t-SNE [38] to picture the 2-D representation of both global (circle) and local (cross) BYOL models in Fig. 5. There are three interesting conclusions: Firstly, FedU^2 has clearer cluster boundary than FedDecorr, validating that directly decomposing the Frobenius norm of representations deteriorates generalization. Secondly, compared with FeBYOL, the global and local representations of L-DAWA are looser, implying its ineffectiveness in controlling conflicting model deviations. Lastly, with the effect of **EUA**, FedU^2 achieves tighter distribution consistency between global and local representations, as well as more clear cluster bound.

Hyper-parameters sensitivity. We consider the sensitivity



(a) BYOL (b) Simsiam

Figure 6. The effect of λ_U (on CIFAR10 $\alpha = 0.1$ Cross-silo).

of highly relevant hyper-parameters, i.e., the effect of uniformity term $\lambda_U = \{0, 0.01, 0.05, 0.1, 0.2, 0.5\}$ on Cifar10 Cross-silo ($\alpha = 0.1$), in Fig. 6. We set $\lambda_U = 0.1$ in experiments since it reaches the highest performance. And we leave the number of clients $K = \{5, 10, 20, 50, 100\}$ and the local epochs $E = \{5, 10, 20, 50\}$ in Appendix B.

5. Conclusion

In this work, we propose a FUSL framework, i.e., FedU^2, to enhance **U**niform and **U**nified representation. FedU^2 consists of flexible uniform regularizer (**FUR**) and efficient unified aggregator (**EUA**). **FUR** encourages data representations to uniformly distribute in a spherical Gaussian space, mitigating representation collapse and its subsequent entangled impacts. **EUA** further constrains the consistent optimization improvements among different client models, which is good for unified representation. In our empirical studies, we set both cross-silo and cross-device settings, and conduct experiments on CIFAR10 and CIFAR100 datasets, which extensively validate the superiority of FedU^2.

Acknowledgements. This work was supported by National Key R&D Program of China (2022YFB4501500, 2022YFB4501504), and the National Natural Science Foundation of China (No.72192823).

References

[1] Jean-David Benamou. Numerical resolution of an “unbalanced” mass transport problem. *ESAIM: Mathematical Modelling and Numerical Analysis*, 37(5):851–868, 2003. 4

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 6

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 3

[4] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021. 6

[5] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021. 1, 3

[6] Don Kurian Dennis, Tian Li, and Virginia Smith. Heterogeneity for the win: One-shot federated clustering. In *International Conference on Machine Learning*, pages 2611–2620. PMLR, 2021. 2

[7] Jonathan Eckstein and Wang Yao. Augmented lagrangian and alternating direction methods for convex optimization: A tutorial and some illustrative computational results. *RUTCOR Research Reports*, 32(3):44, 2012. 5

[8] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 1, 3

[9] Sungwon Han, Sungwon Park, Fangzhao Wu, Sundong Kim, Chuhan Wu, Xing Xie, and Meeyoung Cha. Fedx: Unsupervised federated learning with cross knowledge distillation. In *European Conference on Computer Vision*, pages 691–707. Springer, 2022. 1, 2, 6

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[11] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019. 6

[12] Zeou Hu, Kiarash Shaloudegi, Guojun Zhang, and Yaoliang Yu. Federated learning meets multi-objective optimization. *IEEE Transactions on Network Science and Engineering*, 9 (4):2039–2051, 2022. 4, 5

[13] Tianyu Hua, Wenxiao Wang, Zihui Xue, Sucheng Ren, Yue Wang, and Hang Zhao. On feature decorrelation in self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9598–9608, 2021. 1

[14] Yilun Jin, Yang Liu, Kai Chen, and Qiang Yang. Federated learning without full labels: A survey. *arXiv preprint arXiv:2303.14453*, 2023. 1

[15] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations*, 2021. 1, 2, 3

[16] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143. PMLR, 2020. 4

[17] Hansol Kim, Youngjun Kwak, Minyoung Jung, Jinho Shin, Youngsung Kim, and Changick Kim. Protofl: Unsupervised federated learning via prototypical distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6470–6479, 2023. 2

[18] Jaeill Kim, Suhyun Kang, Duhun Hwang, Jungwook Shin, and Wonjong Rhee. Vne: An effective method for improving deep representation by manipulating eigenvalue distribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3799–3810, 2023. 1, 2, 4

[19] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 6

[20] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[21] Alexander C Li, Alexei A Efros, and Deepak Pathak. Understanding collapse in non-contrastive siamese representation learning. In *European Conference on Computer Vision*, pages 490–505. Springer, 2022. 1, 2

[22] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020. 1

[23] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*, 2019. 5

[24] Xinting Liao, Chaochao Chen, Weiming Liu, Pengyang Zhou, Huabin Zhu, Shuheng Shen, Weiqiang Wang, Mengling Hu, Yanchao Tan, and Xiaolin Zheng. Joint local relational augmentation and global nash equilibrium for federated learning with non-iid data. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1536–1545, 2023. 2

[25] Xinting Liao, Weiming Liu, Chaochao Chen, Pengyang Zhou, Huabin Zhu, Yanchao Tan, Jun Wang, and Yue Qi. Hyperfed: Hyperbolic prototypes exploration with consistent aggregation for non-iid data in federated learning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 3957–3965, 2023. 6

[26] Weiming Liu, Jiajie Su, Chaochao Chen, and Xiaolin Zheng. Leveraging distribution alignment via stein path for cross-

domain cold-start recommendation. *Advances in Neural Information Processing Systems*, 34:19223–19234, 2021. 4

[27] Weiming Liu, Xiaolin Zheng, Chaochao Chen, Mengling Hu, Xinting Liao, Fan Wang, Yanchao Tan, Dan Meng, and Jun Wang. Differentially private sparse mapping for privacy-preserving cross domain recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6243–6252, 2023. 5

[28] Weiming Liu, Xiaolin Zheng, Chaochao Chen, Jiajie Su, Xinting Liao, Mengling Hu, and Yanchao Tan. Joint internal multi-interest exploration and external domain alignment for cross domain sequential recommendation. In *Proceedings of the ACM Web Conference 2023*, pages 383–394, 2023. 4

[29] Weiming Liu, Xiaolin Zheng, Jiajie Su, Longfei Zheng, Chaochao Chen, and Mengling Hu. Contrastive proxy kernel stein path alignment for cross-domain cold-start recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 2023. 4

[30] Ekdeep Lubana, Chi Ian Tang, Fahim Kawsar, Robert Dick, and Akhil Mathur. Orchestra: Unsupervised federated learning via globally consistent clustering. In *International Conference on Machine Learning*, pages 14461–14484. PMLR, 2022. 1, 2, 6

[31] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017. 1, 3, 6

[32] Zibin Pan, Shuyi Wang, Chi Li, Haijin Wang, Xiaoying Tang, and Junhua Zhao. Fedmdfg: Federated learning with multi-gradient descent and fair guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9364–9371, 2023. 5

[33] Yasar Abbas Ur Rehman, Yan Gao, Pedro Porto Buarque de Gusmao, Mina Alibeigi, Jiajun Shen, and Nicholas D Lane. L-dawa: Layer-wise divergence aware weight aggregation in federated self-supervised visual representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16464–16473, 2023. 1, 2, 3, 6

[34] Karsten Roth, Timo Milbich, Samarth Sinha, Prateek Gupta, Bjorn Ommer, and Joseph Paul Cohen. Revisiting training strategies and generalization performance in deep metric learning. In *International Conference on Machine Learning*, pages 8242–8252. PMLR, 2020. 2

[35] Thibault Séjourné, Jean Feydy, François-Xavier Vialard, Alain Trouvé, and Gabriel Peyré. Sinkhorn divergences for unbalanced optimal transport. *arXiv preprint arXiv:1910.12958*, 2019. 4

[36] Yujun Shi, Jian Liang, Wenqing Zhang, Vincent Tan, and Song Bai. Towards understanding and mitigating dimensional collapse in heterogeneous federated learning. In *The Eleventh International Conference on Learning Representations*, 2022. 1, 2, 4, 6

[37] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147. PMLR, 2013. 6

[38] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 8

[39] Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504, 2021. 4

[40] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020. 4

[41] Y Wu, Z Wang, D Zeng, M Li, Y Shi, and J Hu. Federated contrastive representation learning with feature fusion and neighborhood matching (2021). In *URL https://openreview.net/forum*. 2

[42] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems*, 33:5824–5836, 2020. 5

[43] Fengda Zhang, Kun Kuang, Long Chen, Zhaoyang You, Tao Shen, Jun Xiao, Yin Zhang, Chao Wu, Fei Wu, Yueting Zhuang, et al. Federated unsupervised representation learning. *Frontiers of Information Technology & Electronic Engineering*, 24(8):1181–1193, 2023. 1, 2

[44] Weiming Zhuang, Xin Gan, Yonggang Wen, Shuai Zhang, and Shuai Yi. Collaborative unsupervised visual representation learning from decentralized data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4912–4921, 2021. 1, 2, 3, 6

[45] Weiming Zhuang, Yonggang Wen, and Shuai Zhang. Divergence-aware federated self-supervised learning. In *International Conference on Learning Representations*, 2022. 1, 2, 3, 6