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

Xinting Liao\textsuperscript{1}, Weiming Liu\textsuperscript{1}, Chaochao Chen\textsuperscript{1,4}, Pengyang Zhou\textsuperscript{1}, Fengyuan Yu\textsuperscript{1}, Huabin Zhu\textsuperscript{1}, Binhui Yao\textsuperscript{1,2}, Tao Wang\textsuperscript{2}, Xiaolin Zheng\textsuperscript{1}, Yanchao Tan\textsuperscript{3}

\textsuperscript{1}Zhejiang University, \textsuperscript{2}Midea Group, \textsuperscript{3}Fuzhou University

\{xintingliao, 21831010, zjuccc, zhoupwy, fengyuanyu, zhb2000, zzheng\}@zju.edu.cn, tony.yao@midea.com, tao.wang.seu@gmail.com, yctan@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 \textit{FedU} which enhances generating uniform and unified representation in FUSL with non-IID data. Specifically, \textit{FedU} 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 \textit{FedU}, 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., \textit{CH1}: Mitigating representation collapse entanglement, and \textit{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-
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., FedU, to enhance Uniform and Unified representation in 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 CH2, 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 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 FedU 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 FedU.

2. Related Work

2.1. Federated Unsupervised Learning

To enhance 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 rules, 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 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 FedU 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 FUSL, the proposed FUR in FedU 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 FUSL problem formulation and related assumptions in the following. Empirically, we assume a dataset decentralizes among $K$ clients, i.e., $D = \bigcup_{k \in [K]} D_k$. Data distributions of different clients, i.e., $D_k = \{x_{k,i}\}_{i=1}^{N_k}$.
are unlabeled and non-IID in practice. FUSL can be formulated as a global objective that seeks a collaborative aggregation among clients, i.e.,

$$\min_{\theta} \mathcal{L}(\theta; p) = \sum_{k=1}^{K} p_k \mathbb{E}_{x \sim \mathcal{D}_k} [\mathcal{L}_k(\theta; x)],$$  \hspace{1cm} (1)$$

where $\mathcal{L}_k(\cdot)$ is the unsupervised model loss at client $k$, $p = [p_1, \ldots, p_K]$ 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 $X$ at client $k$, we augment them with two transformations, i.e., $X^v = T^v(X)$ 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., $Z^v = F_{\theta_k}(X^v)$. 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 $\theta_k$ 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., \( \mathbf{z}_{1,B}^2 = \{ z_i^2 \}_{i \in [B]} \) and \( \mathbf{Z}_{B}^2 = \{ z_i^2 \}_{i \in [B]} \), we train the self-supervised model by minimizing the total objective as below:

\[
\ell = \mathbb{E}_{t \sim \mathcal{T}, r \sim \mathcal{R}} \ell_a (\mathbf{Z}_{B}^0, \mathbf{Z}_{B}^2) + \lambda_U (\ell_u (\mathbf{Z}_{B}^0) + \ell_u (\mathbf{Z}_{B}^2)),
\]

(2)

where \( \ell_a \) and \( \ell_u \) 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.,

\[
\ell_u (\mathbf{Z}_{B}^0, \mathbf{Z}_{B}^2) := \frac{1}{B} \sum_{i \in [B]} \| z_i^1 - z_i^2 \|^2.
\]

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., \( s \sim \mathcal{N}(0, 1), s.t., \| s \| = 1 \), 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 \( \mathbf{Z}_{B}^0 \) and a set of random samples following spherical Gaussian distribution \( \mathbf{S}_{B} = \{ s_i \}_{i \in [B]} \):

\[
\ell_u (\mathbf{Z}_{B}^0) := \text{Div}(\mathbf{Z}_{B}^0, \mathbf{S}_{B}).
\]

(3)

Thus the uniform term \( \ell_u \) 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 strictly 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 \( a \) and \( b \) for data and Gaussian distribution respectively, we formulate a UOT problem that searches a coupling matrix \( \pi \) with minimal distribution divergence:

\[
\min_{\pi \geq 0} \ell_u (\mathbf{Z}^0) = \text{vec}(C^\top) \text{vec}(\pi) + \frac{T_\alpha}{2} \| \mathbf{F}_1 \text{vec}(\pi) - \mathbf{a} \|^2_2 + \frac{T_\alpha}{2} \| \mathbf{F}_2 \text{vec}(\pi) - \mathbf{b} \|^2_2,
\]

(4)

where cost matrix \( C_{ij} = \| \mathbf{Z}_i^1 - \mathbf{S}_j \|^2 \) and \( \mathbf{F}_1 = \mathbf{I}_N \otimes \mathbf{I}_N \) (\( \mathbf{F}_c = \mathbf{I}_M \otimes \mathbf{I}_M \) are indicators for row-wise (column-wise) Kronecker multiplication with \( I \) denoting identity matrix.

**Optimization.** Denoting \( \tau_a \mathbf{F}_1^\top \mathbf{F}_1 + \tau_b \mathbf{F}_2^\top \mathbf{F}_2 = \mathbf{Q} \) and vec\((\mathbf{C}) - \tau_a \mathbf{F}_1 \mathbf{a} - \tau_b \mathbf{F}_2 \mathbf{b} = \mathbf{w} \), we rewrite Eq. (4) as a positive definite quadratic form:

\[
\min_{\pi \geq 0} \ell_u (\mathbf{Z}^0) = \frac{1}{2} \text{vec}(\pi)^\top \mathbf{Q} \text{vec}(\pi) + \mathbf{w}^\top \text{vec}(\pi) + \Omega,
\]

(5)

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

\[
\text{vec}\left( \pi^{\text{new}} \right) = \max \left( 0, \text{vec}(\pi^{\text{old}}) - \eta^* \frac{\partial \ell_u (\mathbf{Z}^0)}{\partial \text{vec}(\pi^{\text{old}})} \right),
\]

(6)

with \( \eta^* = \frac{(\mathbf{Q}_{vec}(\pi^{\text{old}}) + \mathbf{w})^\top (\mathbf{Q}_{vec}(\pi^{\text{old}}) + \mathbf{w})}{(\mathbf{Q}_{vec}(\pi^{\text{old}}) + \mathbf{w}) \mathbf{Q}_{vec}(\pi^{\text{old}}) + \mathbf{w})} \). Finally, we obtain the uniform UOT divergence by taking the optimal \( \pi^* \) 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.,

\[
c_k(\eta, d^t) = \frac{u_k(\theta^t)}{u_k(\theta^t)} - \frac{u_k(\theta^{t+1})}{u_k(\theta^t)} \approx \eta_\theta \nabla \log u_k(\theta^t) d^t,
\]

(7)

where \( u_k(\theta^t) = \| \theta^t_k - \theta^t \|^2 \) is the model deviation from server to client \( k \) [12, 16], and global model optimization is \( \theta^{t+1}_g = \theta^t_g - \eta d^t \) with updating direction \( d^t \) and step size \( \eta \). 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.,

$$\max_{d^t} \min_{p \in \mathbb{R}^K} \sum_{k=1}^{K} p_k \nabla \log u_k(\theta^*_d) d^t,$$

s.t. $\|d^t\|^2 \leq 1, p^T 1 = 1, p_k \geq 0,$

where $p$ denotes the weights for different clients.

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

$$J = \min_{p} \frac{\eta}{2} \|\nabla \log u\|^T p \|p\|^2 = \min_{p} \frac{\eta}{2} p^T \mathbf{G} p,$$

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

$$J = \min_{p} \frac{\eta}{2} \|\nabla \log u\|^T p \|p\|^2 + \rho \|p^T 1 - 1\|^2,$$

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

$$p = \max(0, \frac{\eta}{2\rho} \mathbf{G} + \rho \mathbf{I})^{-1} (\rho \mathbf{I} - \rho \mathbf{I})$$

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

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

$$J = \min_{p} \frac{\eta}{2} \|\nabla \log u\|^T p \|p\|^2 + \lambda p^T 1,$$

it holds $\nabla \log(u_i(\theta^*_d)) = \nabla \log(u_j(\theta^*_d)), \forall i \neq j \in [K]$. 

**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 $F_k(\theta)$ be the expected model objective for client $k$, and assume $F_1, \ldots, F_K$ are all $L$-smooth, i.e., for all $\theta_k, F_k(\theta_k) \leq F_k(\theta) + \langle \theta_k - \theta \rangle \nabla F_k(\theta) + \frac{L}{2} \|\theta_k - \theta\|^2$.

**Assumption 2.** Let $F_1, \ldots, F_N$ be all $\mu$-strongly convex: for all $\theta_k, F_k(\theta_k) \geq F_k(\theta) + \langle \theta_k - \theta \rangle \nabla F_k(\theta) + \frac{\mu}{2} \|\theta_k - \theta\|^2$.

**Assumption 3.** Let $\xi^t_k$ be sampled from the $k$-th client’s local data uniformly at random. The variance of stochastic gradients in each client is bounded: $E\|\nabla F_k(\theta^*_k, \xi^t_k) - \nabla F_k(\theta^*_k)\|^2 \leq \sigma^2_k$.

**Assumption 4.** The expected squared norm of stochastic gradients is uniformly bounded, i.e., $E\|\nabla F_k(\theta^*_k, \xi^t_k)\|^2 \leq V^2$ for all $k = 1, \ldots, N$ and $t = 1, \ldots, T - 1$.

**Lemma 1** (Bound of Client Model Divergence). With assumption 4, $\eta_t$ is non-increasing and $\eta_t < 2\eta_{t+1}$ (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 $\theta^*_k = \theta^* \in [N]$. It follows that

$$E \left[ \sum_{k=1}^{K} p_k \|\theta^t_k - \theta^*_k\|^2 \right] \leq 4 \eta^2_0 (E - 1)^2 V^2.$$

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

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

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

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

**Proof.** We provide the proof details in Appendix A.3.
4.1. Experimental Setups

4. Experiments

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 than 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 CIFAR10 and CIFAR100 in Tab. 2, for validating the performance generalization. We choose the same model, i.e., BYOL,

### Algorithm 1 Training procedure of FedU$^2$

| **Input:** Batch size $B$, communication rounds $T$, number of clients $K$, local steps $E$, dataset $D = \bigcup_{k \in [K]} D_k$ |
| **Output:** Global model $\theta^T$ |
| 1: **Server executes($\theta^0$):** |
| 2: Initialize $\theta^0$ with random distribution |
| 3: for $t = 0, 1, ..., T - 1$ do |
| 4: for $k = 1, 2, ..., K$ in parallel do |
| 5: Send $\theta^t$ to client k |
| 6: $\theta_k^{t+1} \leftarrow$ Client executes($k$, $\theta^t$) |
| 7: end for |
| 8: EUA optimize Eq. (10) for $p^*$ and update global model $\theta^{t+1}$ with optimal direction $d^t$ in Eq. (9) |
| 9: end for |
| 10: return $\theta^T$ |

11: **Client executes:$k$, $\theta^t$:** |
12: Assign global model to the local model $\theta_k^t \leftarrow \theta^t$ |
13: for each local epoch $e = 1, 2, ..., E$ do |
14: for batch of samples $X_{k,B} \in D_k$ do |
15: Augment samples $X_{k,B}^{\epsilon} = T_{v \in \{1, 2\}}(X_{k,B})$ |
16: Feature extraction $Z_{k,B}^{t} \leftarrow \mathcal{F}_{\theta_k^t}(X_{k,B}^{\epsilon})$ |
17: Compute total loss in Eq. (2) and update $\theta_k^t$ |
18: end for |
19: end for |
20: end for |
21: return $\theta_k^T$ to server |

#### 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 than 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 CIFAR10 and CIFAR100 in Tab. 2, for validating the performance generalization. We choose the same model, i.e., BYOL,

tiveness of tackle two challenges, i.e., representation col-

drop KNN accuracy slightly, validating the effec-

Applying FUR, illustrates its performance generalization.

FedEMA-BYOL is not sensitive to the non-IID degrees, while Orchestra behaves on oppo-

site. This states that capturing global supervision signals to guide local representation suffers from clustering random-

ness. (3) FedU2 performs the best among all tasks, even in

α = 0.1, illustrating its performance generalization.

Ablation Studies. In Tab. 2, we also consider two variants of FedU2: (1) FedU2 removes FUR, i.e., FedU2-FUR, (2) FedU2 removes EUA, i.e., FedU2-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 FedU2, FedU2-FUR and FedU2-EUA drop KNN accuracy slightly, validating the effectiveness of tackle two challenges, i.e., representation collapse entanglement, and generating unified representations. FedU2-FUR performs better than FedU2-EUA when α = 0.1, while gets worse when α = 5.

### Ablation Studies

<table>
<thead>
<tr>
<th>Dataset</th>
<th>CIFAR10</th>
<th>CIFAR100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Setting α</td>
<td>Cross-Device (K=100)</td>
<td>Cross-Silo (K=10)</td>
</tr>
<tr>
<td>α = 0.1</td>
<td>LP</td>
<td>FT</td>
</tr>
<tr>
<td>FedSimSiam</td>
<td>60.49</td>
<td>44.45</td>
</tr>
<tr>
<td>FedDecorr-SimSiam</td>
<td>43.18</td>
<td>35.15</td>
</tr>
<tr>
<td>FedU2-SimSiam</td>
<td>68.50</td>
<td>56.43</td>
</tr>
<tr>
<td>FedSimCLR</td>
<td>65.76</td>
<td>51.18</td>
</tr>
<tr>
<td>L-DAVA-SimCLR</td>
<td>65.63</td>
<td>49.66</td>
</tr>
<tr>
<td>FedX-SimCLR</td>
<td>67.33</td>
<td>49.96</td>
</tr>
<tr>
<td>FedU2-SimCLR</td>
<td>66.49</td>
<td>51.77</td>
</tr>
<tr>
<td>FedBYOL</td>
<td>61.46</td>
<td>54.36</td>
</tr>
<tr>
<td>FedU2-BYOL</td>
<td>60.15</td>
<td>53.53</td>
</tr>
<tr>
<td>FedBYOL</td>
<td>61.46</td>
<td>54.36</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>CIFAR10</th>
<th>CIFAR100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method α</td>
<td>Cross-Device (K=100)</td>
<td>Cross-Silo (K=10)</td>
</tr>
<tr>
<td>α = 0.1</td>
<td>LP</td>
<td>FT</td>
</tr>
<tr>
<td>FedBYOL</td>
<td>76.12</td>
<td>77.23</td>
</tr>
<tr>
<td>FedU2-BYOL</td>
<td>79.09</td>
<td>79.68</td>
</tr>
<tr>
<td>FedEMA-BYOL</td>
<td>80.32</td>
<td>82.01</td>
</tr>
<tr>
<td>FedDecorr-BYOL</td>
<td>76.76</td>
<td>79.66</td>
</tr>
<tr>
<td>L-DAVA-BYOL</td>
<td>65.40</td>
<td>66.17</td>
</tr>
<tr>
<td>FedX-BYOL</td>
<td>50.94</td>
<td>40.96</td>
</tr>
<tr>
<td>Orchestra</td>
<td>79.25</td>
<td>76.78</td>
</tr>
<tr>
<td>FedU2-FUR-BYOL</td>
<td>81.04</td>
<td>82.18</td>
</tr>
<tr>
<td>FedU2-EUA-BYOL</td>
<td>80.93</td>
<td>82.14</td>
</tr>
<tr>
<td>FedU2-BYOL</td>
<td>81.39</td>
<td>82.21</td>
</tr>
</tbody>
</table>

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

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-DAVA, FedBYOL, and FedU2. 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-DAVA and FedBYOL suffer from severe representation collapse, because they have less singular values beyond 0 than FedDecorr and FedU2. 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 FedU2 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.
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 has clearer cluster boundary than FedDecorr, validating that directly decomposing the Frobenius norm of representations deteriorates generalization. Secondly, compared with FedBYOL, 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 achieves tighter distribution consistency between global and local representations, as well as more clear cluster bound.

Hyper-parameters sensitivity. We consider the sensitivity 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, to enhance Uniform and Unified representation. FedU 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.
References


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


