FedCoSR: Personalized Federated Learning with Contrastive Shareable Representations for Label Heterogeneity in Non-IID Data

Chenghao Huang, Xiaolu Chen, Yanru Zhang, and Hao Wang

***Abstract*—Heterogeneity resulting from label distribution skew and data scarcity can lead to inaccuracy and unfairness in intelligent communication applications that mainly rely on distributed computing. To deal with it, this paper proposes a novel personalized federated learning algorithm, named Federated Contrastive Shareable Representations (FedCoSR), to facilitate knowledge sharing among clients while maintaining data privacy. Specifically, parameters of local models' shallow layers and typical local representations are both considered shareable information for the server and aggregated globally. To address poor performance caused by label distribution skew among clients, contrastive learning is adopted between local and global representations to enrich local knowledge. Additionally, to ensure fairness for clients with scarce data, FedCoSR introduces adaptive local aggregation to coordinate the global model involvement in each client. Our simulations demonstrate FedCoSR's effectiveness in mitigating label heterogeneity by achieving accuracy and fairness improvements over existing methods on datasets with varying degrees of label heterogeneity.**

***Index Terms*—Personalized federated learning, label heterogeneity, contrastive learning, representation learning, intelligent communication**

I. Introduction

A. Background and Motivation

In today's connected world, Intelligent Communication (IC) plays a pivotal role in enabling data-driven applications. From personalized recommendations on smartphones to real-time health monitoring via wearable devices, these systems rely heavily on large volumes of data collected from distributed sources, such as mobile devices or sensors, which can form complex and diverse systems through Internet-of-Things (IoT), data centers, and cloud servers [1]–[5]. Since data-driven solutions based on IC have empowered industries like healthcare, smart homes, and finance to provide tailored services, they also raise privacy concerns, as personal and sensitive data are frequently involved [6], [7].

To address privacy issues, Federated Learning (FL) has emerged as a promising paradigm. Rather than centralizing sensitive data in one location, FL allows distributed clients—such as smartphones, IoT devices, or edge servers—to collaboratively train machine learning models while keeping their data locally stored [8]. This distributed approach helps preserve user privacy while leveraging the computational capabilities of these distributed devices. However, while the classical FL algorithm, FedAvg [8], performs well on Independent and Identically Distributed (IID) data, real-world data generated by distributed clients is often non-IID, also known as statistical heterogeneity [9]. This statistical heterogeneity arises due to varying user behaviors, preferences, and environments across devices, making it challenging for a single global model to generalize effectively across all clients.

In this paper, we focus on two major forms of statistical heterogeneity below.

- **Label distribution skew** refers to the FL situation where the distribution of labels varies significantly across different clients [10]. Due to the imbalance in label distribution, it is challenging to train one global model that generalizes well across all clients. For instance, in smart home devices, geographic location, user interests, and lifestyle differences can result in vastly different user behavior patterns captured by devices.
- **Data scarcity** in a distributed system, also known as data quantity skew among clients, poses an additional challenge on fairness [11], [12]. On one hand, the labels of scarce data are sometimes unique and important, such as rare anomalies in industrial applications or survey results of minority groups. Unfairness may arise from the failure to integrate valuable knowledge contained in scarce data into the global model. On the other hand, scarce data is highly susceptible to causing overfitting, impeding personalization. This usually occurs in scenarios with monopolized clients and minority clients, or newly participating clients with few historical data, resulting in challenges in integrating these data globally.

Consequently, such statistical heterogeneity presents substantial challenges to FL, demanding solutions that balance generalizability, personalization, and fairness. Note that, since we mainly study the labels of scarce data, data scarcity is also regarded as a kind of label skew in this paper. Thus, we collectively refer to the above two types of statistical heterogeneity as label heterogeneity, clarified in Fig. 1.

To deal with label heterogeneity, Personalized Federated Learning (PFL) has emerged as an extension of traditional FL and received attention [9], [12]. PFL tailors personalized models for each client within a collaborative training paradigm to achieve robust performance on heterogeneous datasets.

C. Huang and H. Wang are with the Department of Data Science and AI, Faculty of IT and Monash Energy Institute, Monash University, Melbourne, VIC 3800, Australia (e-mails: {chenghao.huang, hao.wang2}@monash.edu).

X. Chen and Y. Zhang are with the School of Computer Science and Technology, University of Electronic Science and Technology of China (UESTC), Chengdu, and Shenzhen Institute for Advanced Study of UESTC, Shenzhen, China (e-mails: jzzcqebd@gmail.com, yanruzhang@uestc.edu.cn).

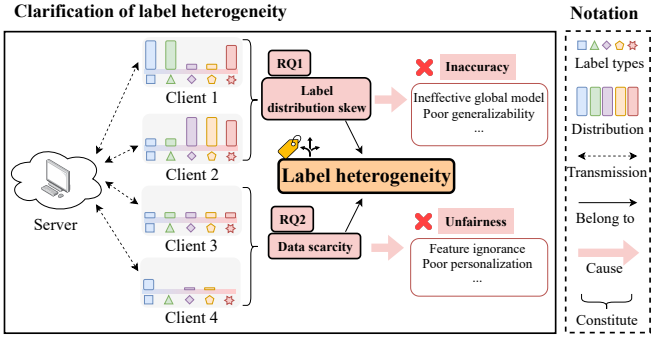

Fig. 1: Clarification of label heterogeneity.

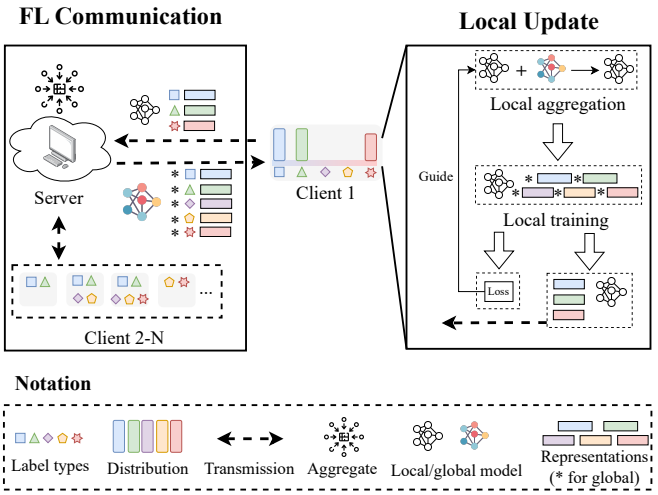

Fig. 2: Brief diagram of FedCoSR, the proposed PFL framework aiming to deal with label heterogeneity.

Notably, it has been proven that it is effective to coordinate the distance between the global model and local models through regularizing the local training objective [13]–[15]. Furthermore, a substantial body of research has focused on splitting the model structure into representation layers, serving as a common feature extractor, and projection layers which address specific tasks [16]–[19]. Despite these advancements, the above methods only improve clients' performance through generic model parameters but ignore more fine-grained characteristics inherent in data, especially label distribution skew. On the other hand, recent studies have explored data quantity skew problems [20], but they have not adequately addressed coexistence of label distribution skew and data scarcity, leaving a research gap.

Thanks to the close correlation between data and representations [21], [22], shareable representations are introduced in FL to improve personalization while preserving privacy [23], and further integrated with the global model to facilitate knowledge integration [24], [25]. However, these methods primarily focus on same-label representation alignment, limiting generalizability, especially with skewed label distributions. In light of this, Contrastive Representation Learning (CRL) which emphasizes deriving knowledge from label-agnostic representations [26], offers a promising perspective. We believe that utilizing shared representations among clients can further contribute to mitigating label heterogeneity.

B. Main Work and Contributions of FedCoSR

This paper introduces a novel approach leveraging CRL [26] on shareable representations among clients, aiming to address the label heterogeneity caused by label distribution skew and data scarcity in distributed ML scenarios. The brief process of the proposed PFL framework is illustrated in Fig. 2.

Globally, the server receives and separately aggregates local model parameters and typical representations of each client. Then, the server sends the global model parameters and the global representations, which provide additional knowledge for performance improvement, to all clients for personalization.

During the local update, each client conducts personalization primarily through local aggregation and local training. For local training, CRL is adopted to foster similarity among representations with the same label and dissimilarity among those with different labels. To both mitigate label distribution skew and enhance knowledge of data-scarce clients, the global representations are used to construct positive and negative sample pairs for the local representations. Moreover, for local aggregation, a loss-wise weighting mechanism is introduced to coordinate personalization among clients with data quantity skew, especially for data-scarce clients. This mechanism ensures that, when the local model's contrastive loss is high, indicating its poor capability of distinguishing samples with different labels, the global model contributes more to drive advancement. Conversely, when the local model's contrastive loss is low, the global model participates less to avoid a compromise on personalization.

The contributions of this work are as follow.

- We consider a challenging FL scenario that simultaneously encompasses both label distribution skew and data scarcity. By effectively tackling these two challenges, we aim to enhance the practicality of our FL algorithm, enabling more robust and equitable performance across diverse and data-limited IC applications in real world.
- We propose Federated Contrastive Shareable Representation (FedCoSR) to guide personalization among clients with heterogeneous labels and data quantity skew, thus enhancing overall performance and fairness among clients. Specifically, through contrastive learning on shareable representations, FedCoSR effectively utilizes local data with label distribution skew to provide additional knowledge for each client model. Moreover, an adaptive local aggregation principle according to contrastive loss is proposed to adjust the proportion of global model participation to ensure fairness, preventing local models with larger datasets from harming the personalized knowledge of those with smaller datasets.
- We provide the theoretical analysis of the effectiveness of the designed local loss function and the communication convergence of the proposed FL algorithm. In our experiments on image classification with varying data heterogeneity, we demonstrate the superiority of FedCoSR compared to other methods addressing the two major forms of heterogeneity.

This paper consists of six sections. Section II reviews the related work of FL and CRL. Section III presents the

formulation of PFL, representation sharing, and the developed FedCoSR algorithm. Section IV analyzes the effectiveness of the designed loss function and the non-convex convergence of the developed algorithm. The experimental results and discussions are provided in Section V. Section VI concludes the whole paper.

II. Literature Review

A. Popular FL Frameworks

Traditional FL methods like FedAvg [8], which learn a single global model for all clients, excel with IID data but struggle with non-IID data. FedProx [13] and Ditto [15] mitigate the impact of heterogeneity by regulating the L2 distance between local models and the global model, while pFedMe [14] learns an additional model for each client. To deal with device heterogeneity, PerFedAvg integrates meta learning with FL [27]. However, these methods become less effective with heterogeneous data or numerous clients.

Model splitting has gained traction in FL, and representative works include FedPer [16], LG-FedAvg [17], and FedRep [19], which split model layers for global aggregation. Besides, exploration into client-tailored aggregations, like FedAMP [28] and FedALA [29], aligns with our method. While effective in managing data heterogeneity, these methods only conduct personalization based on model parameters instead of features inherent in raw data. Notably, in MOON [18] and its inspiring variants [30], [31], CL is applied to clients, but only between the local model and global model, neglecting the interactions between clients.

B. Contrastive Representation Learning

Contrastive Representation Learning (CRL), an emergent field in self-supervised learning, effectively captures knowledge from unlabeled data [26], [32]. Its effectiveness hinges on the principle of minimizing the distance between representations with identical labels (positive pairs) and maximizing that between representations with different labels (negative pairs).

Drawing from representation learning [21], FedProto [23] proposes sharing local representations between clients and one server and using them for regularization, while FedGH [24] uploads local representations to train a global projection layer. FedPAC [25] leverages representations to optimize local classifier combination. Both FedPCL and FedProc [33], [34] adopt a prototypical contrastive learning framework, where local prototypes are shared between clients and the server to align local training with global knowledge, mitigating the effects of non-IID data. FedSeg [35] applies pixel-level CRL to address class heterogeneity in semantic segmentation. CreamFL [36] extends CRL to multimodal settings, using inter-modal and intra-modal contrasts to bridge modality and task gaps, showcasing its adaptability to diverse data challenges.

C. Distinguish of Our Work

Though the potential of CRL in FL has been studied in some works, research gaps remain. First of all, due to the self-supervised nature [37], CRL is capable of dealing with data scarcity in FL by extracting minority features from small-dataset clients. Since the distinctiveness of data has yet not been adequately considered in FL communications, CRL built among different label types of all clients is promising to bridge this gap by guiding models to learn distinctive and fine-grained features from small datasets, while maintaining coherence with the global model. Moreover, the reviewed works overlook model-level information, which contains a wealth of knowledge but needs to be properly accessed. To balance the trade-off between the generalization of the global model and the personalization of local models, adaptively deriving parameters from the global model is a reasonable approach. The measure of CRL can serve as a guide, as it dynamically reflects the local model's ability to distinguish between label types, helping determine the extent of assistance from the global model.

Thus, we propose FedCoSR, which shares label-wise knowledge among clients through transmitting typical local representations to the server, and then uses contrastive learning loss to regularize local training and adaptively guide the fusion between global and local models.

III. Our Proposed Federated Contrastive Shareable Representation

A. Problem Statement

We consider N clients, and for the ith client, a local model θ^i is deployed to conduct training on a dataset $\mathcal{D}^i$. For each sample-label pair $(\mathbf{x}^i, \mathbf{y}^i) \sim \mathcal{D}^i$, the local model f_{θ^i}, where $\theta^i : \mathbb{R}^d \to \mathcal{Y}$ is the parameter set, maps $\mathbf{x}^i \in \mathbb{R}^d$ to predict $\hat{\mathbf{y}}^i = f_{\theta^i}(\mathbf{x}^i) \in \mathbb{Y}$ to approximate the true label $\mathbf{y}^i$. All clients have the same objective to improve the performance, in specific, to minimize the empirical risk over local datasets:

$$\mathcal{F} := \mathbb{E}_{(\mathbf{x}^i, \mathbf{y}^i) \sim \mathcal{D}^i} \mathcal{L}(\mathbf{x}^i, \mathbf{y}^i; \theta^i), \tag{1}$$

where $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is the loss function of the specific ML task to penalize the distance between $\mathbf{y}^i$ and $\hat{\mathbf{y}}^i$. The primary goal of the server is to personalize $\{\theta^i\}_{i=1}^N$ for each client to minimize $\mathcal{F}$. Thus, the global objective is to find a set of local model parameters $\Theta^* = \{\theta^{i*}\}_{i=1}^N$ that satisfy

$$\Theta^* = \arg\min_{\Theta^*} \frac{1}{N} \sum_{i=1}^{N} \mathcal{F}^i, \tag{2}$$

where $\mathcal{F}^i := \mathcal{F}(\theta^{i*}, \mathcal{D}^i)$ is the personalized objective of the ith client.

B. Representation Sharing in FL

Heterogeneous data distributed across tasks may share a common representation despite having different labels [21]. Inspired by insights from [19], [23] that representations shared among clients, e.g., shared features across many types of images or across word-prediction tasks, may provide auxiliary information without privacy intrusion, we consider utilizing shared representations to assist local adaptive aggregation and local personalized training.

Briefly, we let $f_{\theta^i} = [f_{\phi^i}; f_{\pi^i}]$, where $f_{\phi^i}(\cdot) \in \mathbb{R}^{d \times k}$ is the representation layers of the ith local model used to generate

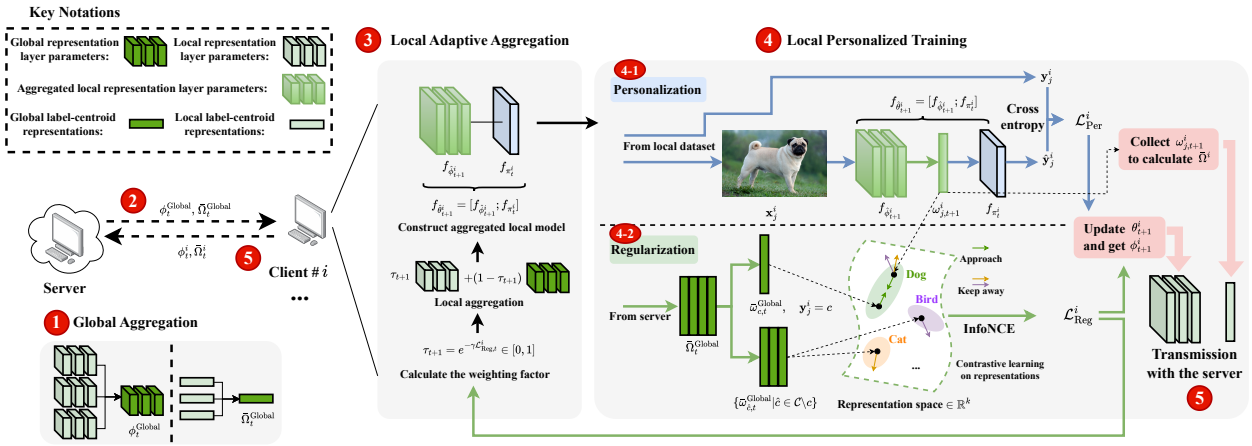

Fig. 3: The entire process of FedCoSR.

representations, and $f_{\pi^i}(\cdot) \in \mathbb{R}^k$ is the projection layers for the output, where d is the data input dimension, and k is the dimension of f_{ϕ^i}'s output, known as a **representation**. Next, the definition of label-centroid representations is provided as follows, which is critical to the holistic framework proposal:

Definition 1 (Label-Centroid Representations)**.** For the ith client, a local label-centroid representation $\bar{\omega}_c^i \in \mathbb{R}^k$ is the mean value of the representations with the label type $c \in \mathcal{C}^i$, where $\mathcal{C}^i$ encompasses all label types existing in $\mathcal{D}^i$. Then, we denote $\mathcal{D}_c^i$ as the subset of $\mathcal{D}^i$ which only contains samples with the label type c, and $\omega^i = f_{\phi^i}(\mathbf{x}^i) \in \mathbb{R}^k$ as the representation of $\mathbf{x}^i$. Formally, the label-centroid representation of the label type c can be calculated as:

$$\bar{\omega}_c^i = \frac{1}{|\mathcal{D}_c^i|} \sum_{\mathbf{x}^i \in \mathcal{D}_c^i} \omega^i, \quad c \in \mathcal{C}^i. \tag{3}$$

Thereby, the local label-centroid representation set of the ith client can be denoted as:

$$\bar{\Omega}^i = \{\bar{\omega}_c^i | c \in \mathcal{C}^i\}. \tag{4}$$

To obtain the final prediction, $f_{\pi^i}(\cdot)$ maps ω^i to the label space. In other words, $f_{\theta^i}(\mathbf{x}^i)$ equals to $f_{\pi^i}(\omega^i)$.

C. Our Developed FedCoSR

At each iteration, the server conducts global aggregation on local models and local representations to generate the global parameters and global representations. Then all clients download the global information for local updates. The main processes of the local update include two parts: local adaptive aggregation by loss-wise weighting and local personalized training by CRL. The overall architecture of FedCoSR is shown in Fig. 3.

1) Global Aggregation: Shallow layers can learn more generic information and they are more suitable for sharing [38]. In FedCoSR, clients only send the parameters of representation layers ϕ to the server for aggregation and keep the parameters of projection layers π local for maintaining the personalization. Also, it is robust for privacy preservation to drop the last one or multiple layers to avoid reverse engineering [39]. Unlike traditional global model aggregation [8], the server in FedCoSR aggregates $\{\phi_t^i\}_{i=1}^N$ at iteration t as follows:

$$\phi_t^{\text{Global}} = \sum_{i=1}^{N} \frac{|\mathcal{D}^i|}{|\mathcal{D}|} \phi_t^i, \tag{5}$$

where $\mathcal{D} = \{\mathcal{D}^i\}_{i=1}^N$ contains all clients' datasets.

Beside the parameter aggregation, all clients upload local label-centroid representations to the server for aggregation as:

$$\bar{\Omega}_t^{\text{Global}} = \Big\{ \sum_{i \in \mathcal{N}_c} \frac{|\mathcal{D}_c^i|}{\sum_i |\mathcal{D}_c^i|} \bar{\omega}_{c,t}^i \Big| c \in \mathcal{C} \Big\} \in \mathbb{R}^{C \times k}, \tag{6}$$

where $\mathcal{N}_c$ denotes the set of clients owning samples with the label type c, and $\mathcal{C} = \bigcup_{i=1}^N \mathcal{C}^i$ encompasses all label types existing in all clients. Rather than averaging, the label volume-wise weighting is adopted for the global label-centroid representation aggregation by considering the size of $\mathcal{D}_c^i$, mitigating the compromise on the global knowledge aggregation caused by local label-centroid representations of scarce samples.

2) Local Adaptive Aggregation: When completing the global aggregation at iteration t, we start the local update at iteration $t+1$, which begins with the initial step of local adaptive aggregation. This mechanism guides each local model to aggregate a proportion of global model parameters while referring to their own model performance, in order to achieve adaptive personalization. Different from traditional FL overwriting local models by the global model, the ith client aggregates ϕ_t^i with ϕ_t^{Global}. In advance, we introduce the contrastive loss $\mathcal{L}_{\text{Reg},t}^i$ at iteration t, which will be elaborated in Section III-C3. To conduct local adaptive aggregation, a weighting factor τ_{t+1} is determined by exponentially scaling $\mathcal{L}_{\text{Reg},t}^i$ into $[0,1]$. Then, the ith client conducts local aggregation and obtains the aggregated local model $f_{\hat{\theta}^i{}_{t+1}}$ for further local training, formulated as follows:

$$\tau_{t+1} = e^{-\gamma \mathcal{L}_{\text{Reg},t}^i} \in [0,1], \gamma > 0, \tag{7}$$

$$\hat{\phi}_{t+1}^i = \tau_{t+1} \phi_t^i + (1 - \tau_{t+1}) \phi_t^{\text{Global}}, \tag{8}$$

$$f_{\hat{\theta}^i{}_{t+1}} = [f_{\hat{\phi}_{t+1}^i}; f_{\pi_t^i}], \tag{9}$$

where γ is a hyperparameter to control the sensitivity of scaling. When $\mathcal{L}_{\text{Reg},t}^i$ is higher, τ_{t+1} becomes smaller to draw on more parameters from the global model for knowledge

enhancement. On the other hand, when $\mathcal{L}^i_{\text{Reg},t}$ is lower, the local model becomes less dependent on acquiring knowledge from the global model. The exponential design provides a type of nonlinear mapping, guiding the ith client to deal with iterations with high loss by learning from the global model.

The reason for choosing the contrastive loss as the basis for weighting, rather than the supervised loss, is that the contrastive loss directly reflects the model's ability to distinguish among multiple samples with different label types. On the other hand, although the supervised loss can reflect the model's recognition capability for individual samples, it tends to result in severe overfitting in clients with scarce data, leading to low training loss but poor predictive performance. Therefore, using the contrastive loss can objectively reflect the capability of each local model without being affected by data scarcity.

3) Local Personalized Training: The data distributions are significantly different among clients due to heterogeneity. So it is necessary to consider both personalization for adapting patterns of local datasets and regularization for properly acquiring external information.

Personalization: For the ith client, we use $\mathcal{L}$ in Eq. (1) as the personalization objective $\mathcal{L}^i_{\text{Per}}$ based on the local dataset. In the context of classification task, we firstly randomly divide $\mathcal{D}^i$ into batches $\mathcal{D}^i_{b,t+1}$ for training efficiency, each of which has b samples. Then we use the expectation of cross entropy to denote this objective at iteration $t+1$:

$$\begin{aligned}
&\mathcal{L}^i_{\text{Per},t+1} = \mathbb{E}_{\mathcal{D}^i_{b,t+1}\sim\mathcal{D}^i}\mathbb{E}_{(\mathbf{x}^i_j,\mathbf{y}^i_j)\sim\mathcal{D}^i_{b,t+1}}\big[H(\mathbf{y}^i_j,\hat{\mathbf{y}}^i_j)\big],\\
&H(\mathbf{y}^i_j,\hat{\mathbf{y}}^i_j) = -\sum_{c=1}^{C} y^i_{j,c}\log(\hat{y}^i_{j,c}),\\
&\mathbf{y}^i_j = (y^i_{j,1},...,y^i_{j,C}),\quad \hat{\mathbf{y}}^i_j = f_{\hat{\theta}^i_{t+1}}(\mathbf{x}^i_j) = (\hat{y}^i_{j,1},...,\hat{y}^i_{j,C}).
\end{aligned} \tag{10}$$

Through personalization, each client cares about local patterns and can adjust the distance between θ^i and θ^{Global}.

Regularization: To utilize useful external information while maintaining the local personalization, we adopt CRL on representations for regularization.

Inspired by [26], CRL, an ML technique aiming to learn representations by contrasting positive pairs (similar samples) against negative pairs (dissimilar samples), can be utilized between the local representations and the global label-centroid representations to assist personalization. Specifically, at iteration $t+1$, the ith client receives the global label-centroid representations $\bar{\Omega}^{\text{Global}}_t$ from the server. We assume that the labels unseen to $\mathcal{D}^i$ can provide external knowledge for f_{θ^i}. We let b be the batch size, and representations at iteration $t+1$ can be obtained by forwarding $\{\mathbf{x}^i_j\}^b_{j=1}$ towards $f_{\hat{\phi}^i_{t+1}}$:

$$\Omega^i_{t+1} = \{\omega^i_{j,t+1}\}^b_{j=1}. \tag{11}$$

Then, we construct positive pair and negative pairs for the $\omega^i_{j,t+1}$ by $\bar{\omega}^{\text{Global}}_{c,t}$ and $\{\bar{\omega}^{\text{Global}}_{\hat{c},t}|\hat{c}\in\mathcal{C}\backslash c\}$, respectively, where the label type of $\omega^i_{j,t+1}$ is c.

Definition 2 (Positive and Negative Pairs of Representations)**.** Intuitively, for $\omega^i_{j,t+1}$ whose label type is c, we consider the global label-centroid representation $\bar{\omega}^{\text{Global}}_{c,t}$ as the positive sample, which has the same label type. On the other hand, the remaining global label-centroid representations with different label types are considered as negative samples. Thereby, one positive pair $p^+_{j,t+1}$ and $|\mathcal{C}|-1$ negative pairs $\{p^-_{j,t+1}\}$ are constructed as:

$$\begin{aligned}
p^+_{j,t+1} &= (\omega^i_{j,t+1}, \bar{\omega}^{\text{Global}}_{c,t}),\\
\{p^-_{j,t+1}\} &= \{(\omega^i_{j,t+1}, \bar{\omega}^{\text{Global}}_{\hat{c},t})|\hat{c}\sim\mathcal{C}\backslash c\},\\
y^i_j &= c,\quad i\in\{1,...,N\},\quad j\in\{1,...,b\}.
\end{aligned} \tag{12}$$

We adopt InfoNCE [26] as the form of regularization loss function, aiming to maximize the similarity of positive pairs and minimize the similarity of negative pairs. For the ith client, according to Definition. 2, the regularization loss can be expressed as:

$$\begin{aligned}
&\mathcal{L}^i_{\text{Reg},t+1} := \mathbb{E}_{\mathcal{D}^i_{b,t+1}\sim\mathcal{D}^i}\mathbb{E}_{(\mathbf{x}^i_j,\mathbf{y}^i_j)\sim\mathcal{D}^i_{b,t+1}}\\
&\left[-\log\frac{e^{\left[D(p^+_{j,t+1})/\tau_{\text{CL}}\right]}}{e^{\left[D(p^+_{j,t+1})/\tau_{\text{CL}}\right]}+\sum_{\hat{c}\in\mathcal{C}\backslash c}e^{\left[D(p^-_{j,t+1})/\tau_{\text{CL}}\right]}}\right],\\
&D(\omega^i_{j,t+1},\bar{\omega}^{\text{Global}}_{c,t}) = \frac{\omega^i_{j,t+1}\cdot\bar{\omega}^{\text{Global}}_{c,t}}{||\omega^i_{j,t+1}||_2\cdot||\bar{\omega}^{\text{Global}}_{c,t}||_2}\in[-1,1],
\end{aligned} \tag{13}$$

where τ_{CL} is the temperature of CRL which controls the attention on positive samples or negative samples, and $D(\cdot)$ is cosine similarity. Then, the ith client calculates its local label-centroid representations $\bar{\Omega}^i_{t+1}$ referring to Eq. (4).

Note that, the batch size b in Eq. (13) not only affects training effect but also impacts the computational overhead, because a larger b means more negative instances are utilized. For contrastive learning, more negative instances can bring about more knowledge enhancement, thereby improving the training effect but increasing the computational overhead, which needs to be carefully balanced.

Final Objective:

As a result, we obtain the final objective function for locally updating θ^i at iteration t:

$$\mathcal{L}^i_{t+1} := \mathcal{L}^i_{\text{Per},t+1}(\mathcal{D}^i;\hat{\theta}^i_{t+1}) + \alpha\mathcal{L}^i_{\text{Reg},t+1}(\Omega^i_{t+1},\bar{\Omega}^{\text{Global}}_t). \tag{14}$$

The process of local updates can be expressed as follows:

$$\theta^i_{t+1} \leftarrow \hat{\theta}^i_{t+1} - \eta\nabla_{\hat{\theta}^i_{t+1}}\mathcal{L}^i_{t+1}(\mathcal{D}^i;\hat{\theta}^i_{t+1};\Omega^i_{t+1},\bar{\Omega}^{\text{Global}}_t), \tag{15}$$

where α is the hyperparameter for trading off personalization and regularization, and η is the learning rate.

Algorithm 1 presents the entire training process of FedCoSR, including 1) global aggregation for both model parameters and label-centroid representations (line 11-13); 2) local aggregation between each model and the global model (line 17-18); and 3) local training on the aggregated local model (line 20-22). 4) All local information is uploaded to the server (line 24-25). This loop is ended until all the optimal local parameters are found.

IV. Theoretical Analysis

Before conducting experiments, we provide a theoretical analysis of the developed FedCoSR algorithm. Firstly, we focus on the effectiveness of the local loss function incorporating cross entropy and InfoNCE, as the local training is

Algorithm 1 FedCoSR Framework

Input: The server and N clients; $\{\mathcal{D}^i\}_{i=1}^N$: datasets of N clients; θ_0^{Global}: the initial global model; η: the learning rate; α: the trade-off factor between personalization and regularization.
Output: $\theta^{1*}, ..., \theta^{N*}$: Optimal local model parameters.
Initialization:
The server sends θ_0^{Global} to all clients to initialize $\{\hat{\theta}_0^i\}_{i=1}^N$ by overwriting, rather than local aggregation in Eq. (8).
Clients train $\{\hat{\theta}_0^i\}_{i=1}^N$ by Eq. (15) in parallel, where $\alpha = 0$. Then clients get $\{\theta_1^i\}_{i=1}^N$.
Clients collect $\{\bar{\Omega}_1^i\}_{i=1}^N$ by Eq. (3) and Eq. (4).
Clients upload $\{\phi_1^i\}_{i=1}^N$ and $\{\bar{\Omega}_1^i\}_{i=1}^N$ to the server.
FL Communication:
for iteration $t = 1, ..., T$ **do**
    Server: ▶ ① **Global aggregation**
    The server aggregates $\{\phi_t^i\}_{i=1}^N$ to get ϕ_t^{Global} by Eq. (5).
    The server aggregates $\{\bar{\Omega}_t^i\}_{i=1}^N$ to get $\bar{\Omega}_t^{\text{Global}}$ by Eq. (6).
    The server sends ϕ_t^{Global} and $\bar{\Omega}_t^{\text{Global}}$ to all clients. ▶ ②
    Clients:
    for the ith client in parallel **do** ▶ **Local Update**
        Local Aggregation: ▶ ③
        Calculate τ_{t+1} based on $\mathcal{L}_{\text{Reg},t}^i$ by Eq. (7).
        Obtain local aggregated model $f_{\hat{\theta}_{t+1}^i}$ by Eq. (9).
        Local Training: ▶ ④
        Construct positive pairs and negative pairs by Eq. (12).
        Calculate $\mathcal{L}_{t+1}^i$ by Eq. (14) for both **personalization** and **regularization**.
        Train $\hat{\theta}_{t+1}^i$ by Eq. (15) and clip it into $[0, 1]$ for normalization to get $f_{\theta_{t+1}^i} = [f_{\phi_{t+1}^i}; f_{\pi_{t+1}^i}]$.
        Upload: ▶ ⑤
        Collect $\bar{\Omega}_{t+1}^i$ by Eq. (3) and Eq. (4).
        Upload ϕ_{t+1}^i and $\bar{\Omega}_{t+1}^i$ to the server.
    end for
end for
return $\theta^{1*}, ..., \theta^{N*}$.

the core part of the personalization of FedCoSR. We explain that FedCoSR minimizes InfoNCE to maximize the mutual information between the local representations and global label-centroid representations, thus enhancing the distinguishing capability of each local model. Secondly, both global model aggregation and local model aggregation linearly change the model parameters, which may cause deviations in the loss expectation during each iteration. If the deviation remains unbounded, the convergence of FedCoSR may not be guaranteed. Therefore, we characterize the overall communication between the server and the clients to establish an upper bound on the loss expectation deviation, thereby ensuring the convergence of FedCoSR. Note that due to the non-convexity of the local loss induced by InfoNCE, we focus on non-convex settings. Moreover, we also present the relationship between the convergence rate and key hyperparameters.

A. Effectiveness of Local Loss Function

Referring to Eq. (14), the local loss function is a linear combination of $\mathcal{L}_{\text{Per}}$ and $\mathcal{L}_{\text{Reg}}$. Since $\mathcal{L}_{\text{Per}}$ is in the form of cross entropy, its convexity and good convergence are known [40]. But the non-convexity of InfoNCE $\mathcal{L}_{\text{Reg}}$ may lead to sub-optimum of the training objective for each local model. Thus, we focus on analyzing $\mathcal{L}_{\text{Reg}}$ to explain the demonstrate of the local loss function.

To quantify the enhancement brought by the InfoNCE loss $\mathcal{L}_{\text{Reg}}$, we introduce the manifestation of mutual information in contrastive learning as follows.

Definition 3 (Mutual Information in Contrastive Learning)**.** To construct contrastive learning task, mutual information $I(\cdot)$ is introduced between anchor samples X and the similar samples X^+ which is also known as positive samples:

$$I(X^+; X) = \sum_{x^+ \in X^+, x \in X} p(x^+, x) \log \left[\frac{p(x^+|x)}{p(x^+)} \right], \quad (16)$$

where $p(\cdot)$ is the notation of probability.

Based on Definition 3, we present the following theorem to illustrate the effectiveness of the local loss function design.

Theorem 1 (InfoNCE Minimization in FedCoSR)**.** *For each data batch of the ith client, minimizing InfoNCE equals to maximizing the mutual information between each anchor representation in this batch and its positive representation, and meanwhile minimizing the mutual information between it and its negative representations, formulated as follows, where b is the batch size:*

$$\mathcal{L}_{\text{Reg},t+1}^i \geq -\frac{1}{b} \sum_{j=1}^{b} I(\bar{\omega}_{c,t}^{\text{Global}}, \omega_j^i) + \log(C). \quad (17)$$

Intuitively, we have $\bar{I}(\cdot) \geq \log(C) - \mathcal{L}_{\text{Reg}}$, where $\bar{I}(\cdot)$ is the mean value of the mutual information. When C becomes larger, the lower bound of similar representations increases, improving the performance of the ith local model. Theorem 1 indicates that optimizing the regularization term $\mathcal{L}_{\text{Reg}}^i$ can enlarge the information acquisition for the ith client, showing the effectiveness of combining $\mathcal{L}_{\text{Per}}$ and $\mathcal{L}_{\text{Reg}}$. Due to the page limitation, the proof of Theorem 1 is omitted from the current paper.

B. Convergence of FedCoSR

The convergence conditions for the ith client is explained in this part. For ease of discussion, we denote the total number of local training epochs as R, which is set to 1 in this paper. Specifically, we define $tR + r$ as the rth epoch at iteration t, tR as the end of iteration t (end of the local training), $tR + 0$ as the beginning of iteration t (local aggregation), and $tR + \frac{1}{2}$ as the utilization of global label-centroid representations at iteration t (after local aggregation).

Before showing the convergence of FL communication, we make three assumptions which are widely used in literature: Assumption 1: Lipschitz Smoothness ensuring consistency in gradient changes [13], [23], [27], [28], Assumption 2: Unbiased Gradient and Bounded Variance providing stable gradient estimates [13], [23], [27], and Assumption 3: Bounded Variance of Representation Layers guaranteeing an acceptable variance between the global model and each local model.

Assumption 1 (Lipschitz Smoothness)**.** The ith local loss function is L_1-Lipschitz smooth, leading to that the gradient

of the local loss function is L_1-Lipschitz continuous, where $L_1 > 0$, $\forall r_1, r_2 \in \{0, \frac{1}{2}, 1, ..., R\}$, and $(\mathbf{x}^i, \mathbf{y}^i) \in \mathcal{D}^i$:

$$||\nabla \mathcal{L}^i_{tR+r_1}(\mathbf{x}^i, \mathbf{y}^i; \hat{\theta}^i_{tR+r_2}) - \nabla \mathcal{L}^i_{tR+r_2}(\mathbf{x}^i, \mathbf{y}^i; \hat{\theta}^i_{tR+r_2})||_2 \leq L_1 ||\hat{\theta}^i_{tR+r_1} - \hat{\theta}^i_{tR+r_2}||_2. \tag{18}$$

Assumption 2 (Unbiased Gradient and Bounded Variance)**.** The stochastic gradient $\nabla \mathcal{L}^i_{tR}(\hat{\theta}^i_{tR}; \mathcal{D}^i_{b,tR})$ is an unbiased estimator of the local gradient, and the variance of $\nabla \mathcal{L}^i_{tR}(\hat{\theta}^i_{tR}; \mathcal{D}^i_{b,tR})$ is bounded by σ:

$$\text{Var}[\nabla \mathcal{L}^i_{tR}(\hat{\theta}^i_{tR}; \mathcal{D}^i_{b,tR})] \leq \sigma^2. \tag{19}$$

Assumption 3 (Bounded Variance of Representation Layers)**.** The variance between f^i_{tR} and f^{Global}_{tR} is bounded, whose parameter bound is:

$$\mathbb{E}[||f^i_{tR} - f^{\text{Global}}_{tR}||^2_2] \leq \varepsilon^2. \tag{20}$$

Based on the above assumptions, the deviation in loss expectation for each iteration is bounded, as shown in Theorem 2, serving as the foundation of our algorithm's convergence.

Theorem 2 (One-Iteration Deviation)**.** *For the ith client, between the iteration t and the iteration $t+1$, we have:*

$$\mathbb{E}[\mathcal{L}^i_{(t+1)R+\frac{1}{2}}] \leq \mathbb{E}[\mathcal{L}^i_{tR+\frac{1}{2}}] - \big(\eta - \frac{L_1\eta^2}{2}\big) \sum_{r=\frac{1}{2}}^{R} ||\nabla \mathcal{L}^i_{tR+r}||^2_2 + \frac{RL_1\eta^2\sigma^2}{2} + \frac{L_1(\varepsilon^2+\varepsilon)}{2} + \frac{2\alpha}{\tau_{CL}}. \tag{21}$$

Theorem 2 indicates that, the deviation in the loss expectation for the ith client is bounded from the iteration t to the iteration $t+1$, which leads to the following corollary for the convergence of FedCoSR in non-convex settings.

Corollary 1 (Non-Convex FedCoSR Convergence)**.** *The loss function of the ith client monotonously decreases between the iteration t and the iteration $t+1$, when the learning rate at iteration r' satisfies:*

$$\eta_{r'} < \frac{\mathbb{S} + \sqrt{\big[\mathbb{S}\big]^2 - \frac{(L_1\mathbb{S}+RL_1\sigma^2)(L_1\varepsilon^2\tau_{CL}+L_1\varepsilon\tau_{CL}+4\alpha)}{\tau_{CL}}}}{L_1\mathbb{S} + RL_1\sigma^2}, \tag{22}$$

where $r' = \frac{1}{2}, 1, ..., R$, and briefly, $\mathbb{S} = \sum_{r=\frac{1}{2}}^{r'} ||\nabla \mathcal{L}^i_{tR+r}||^2_2$.

Corollary 1 indicates that as long as the learning rate is small enough at iteration r', FedCoSR will converge. Furthermore, the convergence rate of FedCoSR can be also obtained as follows.

Theorem 3 (Non-Convex Convergence Rate of FedCoSR)**.** *Given any $\epsilon > 0$, after T iterations, the ith client converges with the rate:*

$$\frac{1}{TR} \sum_{t=1}^{T} \sum_{r=\frac{1}{2}}^{R} \mathbb{E}[||\nabla \mathcal{L}^i_{tR+r}||^2_2] < \epsilon, \tag{23}$$

when

$$T > \frac{2\tau_{CL}(\mathcal{L}^i_{\frac{1}{2}} - \mathcal{L}^{i,*})}{(2\eta - L_1\eta^2)\epsilon\tau_{CL}R - L_1\eta^2\sigma^2\tau_{CL}R - \tau_{CL}L_1(\varepsilon^2 - \varepsilon) - 4\alpha}, \tag{24}$$

$$\eta < \frac{2\epsilon\tau_{CL}R + \sqrt{4\epsilon^2\tau^2_{CL}R^2 - 4L_1\tau_{CL}R(\epsilon+\sigma^2)(\tau_{CL}L_1(\varepsilon^2+\varepsilon) + 4\alpha)}}{2L_1\tau_{CL}R(\epsilon+\sigma^2)}, \tag{25}$$

$$\alpha < \frac{\epsilon^2}{4L_1(\epsilon+\sigma^2)} - \frac{\tau_{CL}L_1}{4}(\varepsilon^2+\varepsilon). \tag{26}$$

Theorem 3 outlines the specific conditions for convergence. To ensure the algorithm converging with a rate, the minimal training iterations T should be determined. Correspondingly, the upper bounds for two hyperparameters, η and α, are also presented. Due to the page limit, the proofs of Theorem 2, Corollary 1, and Theorem 3 are omitted from the paper.

V. Experiments and Discussion

A. Experiment Setup

1) Dataset Description: We consider three popular datasets of image classification for evaluation: CIFAR-10 consists of 10 categories of items, each containing 6,000 images; EMNIST consists of 47 categories of handwritten characters, each containing 2,400 images; and CIFAR-100 consists of 100 categories of items, each containing 600 images. Based on these datasets, we evaluate the performance of our method in tasks with different scales of label classes.

2) Heterogeneity Setting on Datasets: We simulate label distribution skew and data scarcity with two widely adopted settings. The first setting is informed by [29], called practical setting, using the Dirichlet distribution $Dir(\beta)$, where $\beta \in (0, 1]$. We set $\beta = 0.1$ as the default value, since smaller β results in more heterogeneous simulations. The second setting is the pathological setting [8], [41], which samples 2, 10, and 20 label categories from CIFAR-10, EMNIST, and CIFAR-100, respectively. While both settings can lead to differences in label distributions among clients, the difference is: in the practical setting, the Dirichlet distribution controls the proportion of labels assigned to each client, with varying levels of skewness based on the value of β, allowing all clients to potentially receive all labels but in different proportions. In contrast, the pathological setting enforces a hard limit on the number of label categories each client can receive, leading to more extreme label distribution heterogeneity. Thus, the practical setting results in more gradual label distribution shifts, while the pathological setting imposes rigid category constraints.

For evaluation, we ensure that all clients receive datasets of approximately similar sizes, which are about from 8,000 to 10,000 samples. 75% of the local data forms the training dataset, and the remaining 25% is used for testing. Fig. 4 shows the data distribution visualization of the default settings for the three datasets. The corresponding results are presented and analyzed in Section. V-B1 and Section. V-B2.

3) Scarcity Setting on Datasets: To evaluate the performance of the proposed method under data scarcity, we design the following two experiments based on the default heterogeneity settings:

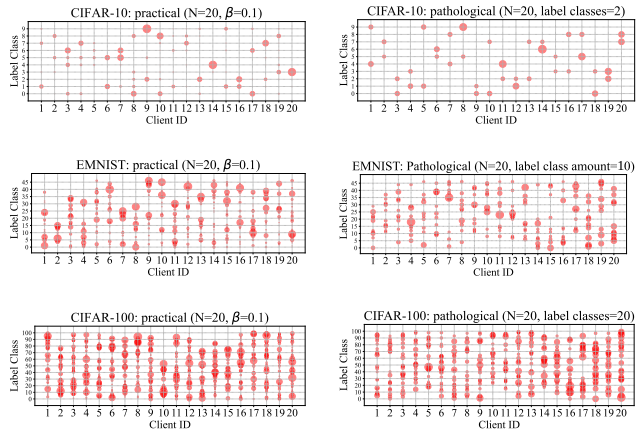

Fig. 4: The data distribution of CIFAR-10, EMNIST, and CIFAR-100 in the default settings.

TABLE I: Taxonomy of compared baselines.

Algorithm	Type	Main technique	RQ1	RQ2
FedAvg [8]	FL	Standard	✗	✗
FedProx [13]	FL	Regularization	✗	✗
MOON [18]	FL	Contrastive learning	✓	✗
pFedMe [14]	PFL	Regularization	✓	✗
Ditto [15]	PFL	Multi-task learning	✗	▲
PerFedAvg [27]	PFL	Meta learning	✓	✗
FedRep [19]	PFL	Model splitting	✓	✗
LG-FedAvg [17]	PFL	Model splitting	✓	✗
FedPer [16]	PFL	Model splitting	✓	▲
FedAMP [28]	PFL	Model collaboration	✓	✗
FedALA [29]	PFL	Local aggregation	✓	✗
FedProto [23]	PFL	Representation sharing	✓	▲
FedPAC [25]	PFL	Representation sharing	✓	✗
FedGH [24]	PFL	Representation sharing	✓	✗
Our proposed	**PFL**	**Representation sharing**	✓	✓

- **Fairness for New Participation:** Under the default CIFAR-10 settings where 20 clients are set, we reduce the data size of the last five clients to 10% of their original amounts while keeping the data distribution unchanged.
- **Robustness against Model Degradation:** Additionally, to explore the robustness of the FL algorithms in a scenario where all clients' data is scarce, we reduce the data size for each client to between 5% and 25%, maintaining the same data distribution.

The corresponding results are presented and analyzed in Section. V-B3 and Section. V-B4.

4) Baselines for Comparison: We compare FedCoSR with individual local training and 14 popular FL methods, categorized in Table I, where ▲ means fairness is discussed but data scarcity is not considered. By default, we adopt the optimal hyperparameters recorded in each work for our comparison.

5) Other Settings: We implement all experiments using PyTorch-1.12 on an Ubuntu 18.04 server with two Intel Xeon Gold 6142M CPUs with 16 cores, 24G memory, and one NVIDIA 3090 GPU. For simplicity, we construct a two-layer Convolutional Neural Network (CNN) followed by two Fully-Connected (FC) layers for all datasets. Each CNN includes a convolution operation, a ReLU activation, and a max-pooling step. The output of the second CNN is flattened and passed through a FC layer with k output features, where k is the dimension of shareable representations. Finally, the k features are mapped by the second FC layer to $|\mathcal{C}|$ label classes, where $\mathcal{C}$ includes all label classes in the FL scenario. In FedCoSR, the representation layers ϕ consists of the two-layer CNN and the first FC layer, and the second FC layer is the projection layer π. Additionally, for reliability, five-time experiments are conducted to calculate the mean and standard deviation, where the mean reflects the overall trend of the results and the standard deviation quantifies the variability.

B. Result Analysis and Discussion on Research Questions

1) RQ1-Label Distribution Skew: Effectiveness: As shown in Table II, FedCoSR achieves the highest test accuracy across all three datasets in both the practical and pathological settings, demonstrating the effectiveness of our method. FedProto achieves the second best performance, just below FedCoSR, demonstrating that FedCoSR's knowledge gained from both representations (data-level) and model parameters (model-level) leads to its superior performance over other methods, particularly with heterogeneous clients.

Another observation is that FedCoSR achieves better performance when the number of label classes increases. As shown in Table II, we can see that FedCoSR's improvement over the second best method becomes more significant with an increase in label classes within two heterogeneous settings: CIFAR-100 (3.54%/4.01%) and CIFAR-10 (1.36%/1.73%). The reason can be that, according to [26], if the labels are fully reliable, the lower bound of the mutual information between positive pairs estimated by InfoNCE will be tighter when the number of negative samples is larger, which usually contributes to model performance improvements. Thus, we can infer that contrastive learning applied to different clients' representation centroids is more effective in tackling heterogeneity, particularly in scenarios with a larger number of label classes. However, we are unsure where the marginal benefits of contrastive learning for FL lie, and further experiments on more diverse datasets may be needed.

Furthermore, we study learning efficiency based on training curves of FedCoSR and other compared methods with relatively high performance in Fig. 5. Specifically, we conduct averaging smooth on original curves whose moving window length is 50. As shown in Fig. 5, FedCoSR achieves the highest accuracy convergence in both practical and pathological settings. Though several methods, such as FedPer and FedRep, achieve higher convergence at the beginning, after a short period of fluctuation, FedCoSR demonstrates an upward trend in accuracy, while others exhibit a prolonged decline in accuracy overtime. This advantage of consistent learning can be attributed to the knowledge enhancement provided by CRL adopted in local personalized training. Although the convergence speed of various algorithms is similar, FedCoSR exhibits more stable trends in both accuracy and loss curves with stability kept by the local adaptive aggregation.

2) RQ1-Label Distribution Skew: Robustness to Varying Heterogeneity: For evaluating FedCoSR's capability of handling varying levels of heterogeneity, we adjust β of the Dirichlet distribution to control practical heterogeneity on CIFAR-10 and change label classes held by each client to adjust the pathological heterogeneity on CIFAR-100, as shown

TABLE II: The accuracy and standard deviations of the three datasets in the practical and pathological heterogeneous setting. We use superscripts $*$, †, and ‡, to emphasize the 1st, 2nd, and 3rd best values in each column, respectively. Green means better and red means worse than the averaged result.

Method	Practical heterogeneous ($\beta = 0.1$, $N = 20$)						Pathological heterogeneous ($N = 20$)					
	CIFAR-10		EMNIST		CIFAR-100		CIFAR-10		EMNIST		CIFAR-100	
	Acc. (%)	Std. (%)	Acc. (%)	Std. (%)	Acc. (%)	Std. (%)	Acc. (%)	Std. (%)	Acc. (%)	Std. (%)	Acc. (%)	Std. (%)
Local	69.90 ↓	9.44 ↓	91.80 ↓	4.50 ↓	35.12 ↓	6.25 ↓	61.94 ↓	11.56 ↓	81.21 ↓	7.25 ↓	30.66 ↓	6.33 ↓
FedAvg	61.51 ↓	11.23 ↓	83.59 ↓	10.98 ↓	30.88 ↓	3.46* ↑	51.72 ↓	17.72 ↓	74.19 ↓	13.36 ↓	25.84 ↓	6.32 ↓
FedProx	62.91 ↓	9.75 ↓	85.33 ↓	7.54 ↓	32.45 ↓	3.69† ↑	62.09 ↓	7.01 ↑	84.32 ↓	6.09 ↓	31.23 ↓	4.28‡ ↑
MOON	82.96 ↓	10.19 ↓	93.61 ↓	4.77 ↓	48.04 ↑	5.11 ↑	86.88 ↑	13.49 ↓	95.24 ↑	7.40 ↓	51.31 ↑	5.91 ↓
pFedMe	84.75 ↑	9.55 ↓	95.55 ↑	1.47 ↑	46.20 ↓	6.89 ↓	85.72 ↑	9.63 ↓	95.78 ↑	3.03 ↑	46.88 ↑	4.76 ↑
Ditto	87.53 ↑	8.92 ↓	96.53 ↑	1.17‡ ↑	46.40 ↓	4.81 ↑	89.08 ↑	7.37 ↑	96.48 ↑	2.24 ↑	48.81 ↑	4.51 ↑
PerFedAvg	88.95 ↑	6.98 ↑	94.63 ↑	1.72 ↑	48.80 ↑	5.32 ↓	89.45 ↑	7.69 ↑	93.95 ↑	2.47 ↑	48.03 ↑	4.23† ↑
FedRep	90.66‡ ↑	6.28† ↑	97.00† ↑	1.21 ↑	52.06 ↑	5.15 ↑	91.47 ↑	7.43 ↑	96.59† ↑	2.58 ↑	53.12 ↑	4.90 ↑
LG-FedAvg	88.72 ↑	8.05 ↑	95.90 ↑	1.50 ↑	48.19 ↑	6.25 ↓	91.35 ↑	7.15 ↑	94.21 ↑	1.95* ↑	46.67 ↓	5.84 ↓
FedPer	89.94 ↑	6.51‡ ↑	96.18 ↑	1.61 ↑	53.42‡ ↑	5.23 ↓	91.23 ↑	6.83 ↑	94.82 ↑	2.56 ↑	53.34 ↑	5.25 ↓
FedAMP	89.27 ↑	7.38 ↑	96.32 ↑	1.30 ↑	51.62 ↑	5.88 ↓	91.42 ↑	6.78‡ ↑	95.99 ↑	2.42 ↑	51.27 ↑	5.83 ↓
FedALA	83.33 ↓	9.81 ↓	92.37 ↓	3.21 ↓	40.41 ↓	3.98 ↑	83.88 ↓	13.47 ↓	92.24 ↓	3.87 ↑	45.31 ↓	5.20 ↓
FedProto	89.82 ↑	7.18 ↑	96.82‡ ↑	1.21 ↑	53.47† ↑	6.00 ↓	91.58† ↑	6.52† ↑	96.49‡ ↑	2.25 ↑	53.52‡ ↑	5.09 ↑
FedPAC	90.79† ↑	6.72 ↑	96.66 ↑	1.02* ↑	53.14 ↑	4.27 ↑	91.55‡ ↑	7.52 ↑	95.95 ↑	2.15‡ ↑	53.59† ↑	4.44 ↑
FedGH	88.95 ↑	7.54 ↑	96.01 ↑	1.46 ↑	46.85 ↑	6.27 ↓	91.39 ↑	7.04 ↑	95.66 ↑	2.58 ↑	48.65 ↑	5.44 ↓
FedCoSR	**92.15*** ↑	**5.57*** ↑	**97.98*** ↑	**1.15**† ↑	**57.01*** ↑	**3.76**‡ ↑	**93.31*** ↑	**6.05*** ↑	**97.85*** ↑	**2.01**† ↑	**57.60*** ↑	**4.07*** ↑
Averaged	83.83	8.21	94.13	2.86	46.43	5.15	83.95	9.00	92.59	4.02	46.34	5.15

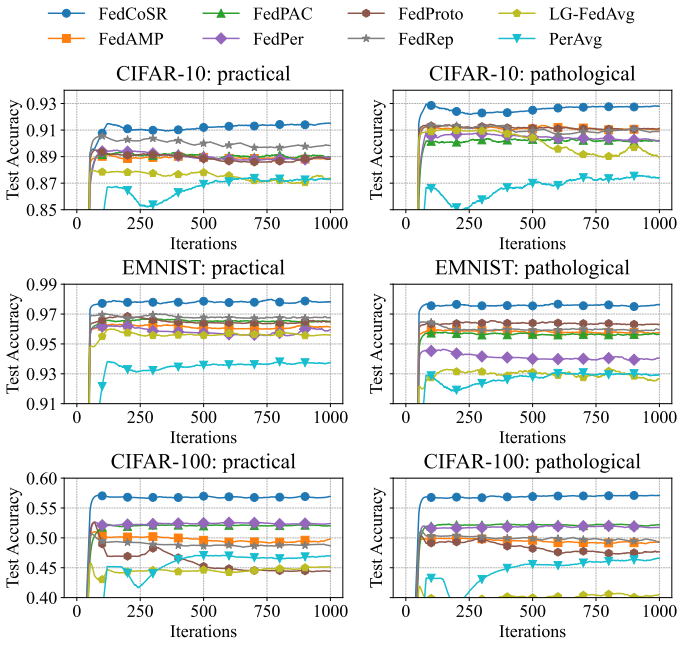

Fig. 5: The smoothed learning curves of well-performing methods in the default settings.

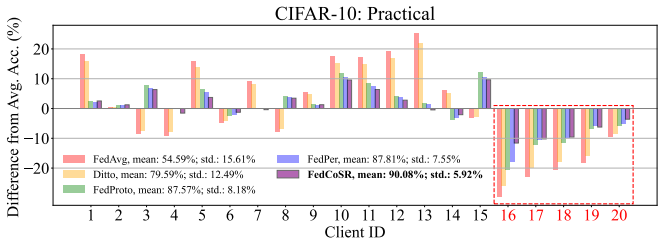

Fig. 6: Performance of each client on CIFAR-10 with practical heterogeneity using different FL methods. The client number in red means its dataset size is much smaller than the others.

in Table III. We find that most PFL methods achieve better performance in more heterogeneous settings, with our FedCoSR consistently ranking in the top three. An interesting trend is that almost all methods degrade in accuracy but achieve improved stability when the heterogeneity becomes moderate. This can be attributed to the reduced performance gap between clients with abundant data and those with scarce data, when the setting becomes less heterogeneous. Specially, FedPAC is not applicable on practical CIFAR-10 with $\beta = 0.01$, whose optimization problem among clients may not have feasible solutions due to the extreme scarcity of label classes. In contrast, FedCoSR shows a stronger robustness to both extreme and moderate label heterogeneity than the other methods.

3) RQ2-Data Scarcity: Fairness Maintenance: In Fig. 6, the performance of FedAvg and the algorithms discussing fairness is assessed under an FL scenario where the data size for the last five clients (Clients 16-20) is significantly reduced to 10% of their original amounts. The goal is to evaluate how well different algorithms handle fairness in terms of model performance for clients with very little data. FedCoSR stands out by maintaining a high level of fairness, even with data scarcity. It achieves a mean accuracy of 90.08% with a standard deviation of 5.92%, the best among the methods. This is due to FedCoSR's ability to enhance shared representation through contrastive learning and utilize local adaptive aggregation, which helps compensate for the reduced data size and allows for better personalization.

FedPer and FedProto show strong results, effectively mitigating heterogeneity through model splitting and personalized aggregation, but still fall short of the fairness achieved by FedCoSR. Ditto demonstrates moderate performance, indicating its regularization strategy struggles in data-scarce environments. Meanwhile, FedAvg performs the worst, highlighting its limitations in managing client heterogeneity and data imbalance due to its sole reliance on model aggregation without personalization.

Additionally, we also analyze the standard deviations in Tables II and III. In all experiments, FedCoSR consistently ranks in the top three for fairness, showing smaller differences among client groups. Traditional FL methods like FedProx and Ditto are often the fairest due to their focus on generalization, while PFL methods like FedRep and PerFedAvg also achieve

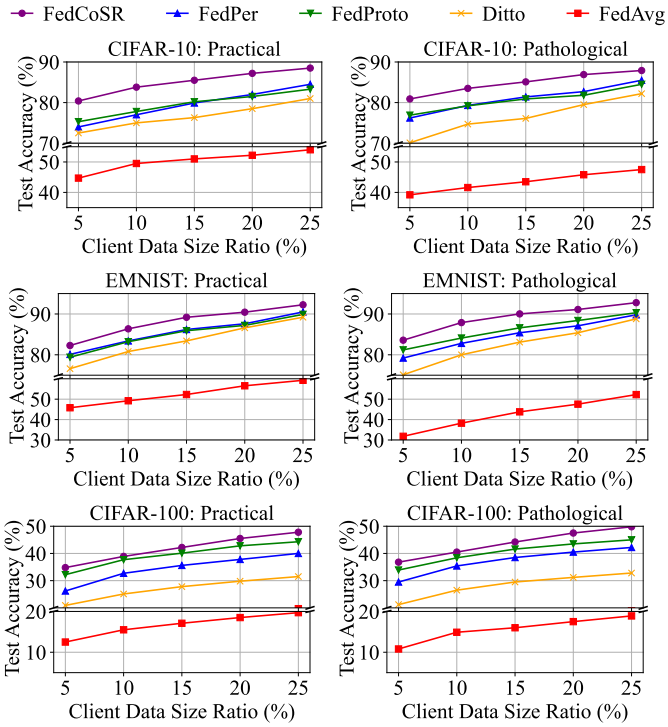

Fig. 7: Robustness to the scenario where the data sizes of all clients are small.

good fairness by balancing global and local updates. FedCoSR strikes a better balance between generalization and personalization, resulting in both high accuracy and fairness, making it effective for clients with scarce data and diverse labels.

Overall, the results demonstrate that FedCoSR provides the most balanced performance across clients, thanks to its privacy-preserving knowledge sharing and local adaptation techniques.

4) RQ2-Data Scarcity: Robustness against Model Degradation: We test our FedCoSR with varying local dataset sizes on CIFAR-10 under the two scenarios. As shown in Fig. 7, clients of varying local dataset sizes consistently reap the advantages of engaging in FL, while our method achieves the best performance. Compared with results in Table II, small local datasets indeed cause substantial performance degradation. Despite all methods have decreasing trends when the local dataset size becomes smaller, the degree of decline under FedCoSR is less than that of other methods, demonstrating FedCoSR's strongest robustness to data scarcity. This can be attributed to that the information embedded in the shared representations is effectively enhanced through contrastive learning among clients, compensating for the lack of data volume.

C. Result Analysis and Discussion on Practicality

1) Scalability: In Table III, we study the scalability by increasing the number of clients N to 100. Compared to the results when $N = 20$ in Table II, most methods experience significant degradation when N increases, with about $6-20\%$ degradation in performance. This is attributed to the challenges arising from label scarcity and label distribution skew in pathological setting sbecoming more extreme with increasing N. As such, these methods encounter difficulties in personalization. Overall, model splitting-based methods and representation sharing-based methods demonstrate good scalability. FedCoSR shows only a 6% drop in the practical setting and still achieves the best in both settings, highlighting its strong scalability and the advantage of applying CRL among shared representations.

2) Communication Overhead: We evaluate the communication overhead per client in single iteration. We denote $\varphi(\cdot)$ as the number of parameters. Based on Eq. (9) where θ is concatenated by ϕ and π, most methods incur the same computational overhead as FedAvg, which uploads and downloads only one entire model, denoted as $2\varphi(\theta)$. FedProto only transmits representations, thus in general (we suppose a representation is much smaller than a model), it has the least communication overhead denoted as $2\varphi(\bar{\Omega})$. FedGH uploads representations and downloads projection layers, costing $\varphi(\bar{\Omega})+\varphi(\pi)$, but also takes time for global training on the server. Since FedCoSR uploads and downloads parameters of representation layers and averaged representations of each label, its overhead can be regarded as $2[\varphi(\phi)+\varphi(\bar{\Omega})]$, where the depth of ϕ is $|\theta|-1$. This communication overhead is similar to the major overhead $2\varphi(\theta)$ and thus deemed acceptable.

3) Privacy Concerns: In FL, the risk of data privacy leakage is inevitable due to reverse engineering techniques, such as gradient inversion [42]. However, specific strategies can mitigate these risks without significantly affecting system performance. In our work, we adopt the following approaches to reduce privacy leakage:

- **Partial model parameter sharing:** By uploading only a portion of the local model parameters, we reduce the amount of information available for potential exploitation, avoiding the risk of full data reconstruction.
- **Parameter-only uploads:** We avoid sharing the model structure itself, limiting the server's ability to exploit model-specific details for data inference.
- **Uploading mean values of representations:** Instead of uploading raw embeddings, we only share mean values of representations. This abstracts the data further, reducing the potential for inversion attacks.

In addition to these methods, Differential Privacy (DP) can be employed on model parameters and representations to further reduce privacy risks [43]. DP adds noise to the shared data, but this comes with a trade-off between privacy and model utility, as too much noise may degrade performance.

D. Result Analysis and Discussion on Methodology

1) Ablation Study: To validate the effectiveness of main techniques employed in FedCoSR, we remove them and create three variants ("w/o" is short for "without"): (1) FedCoSR w/o LA: w/o loss-wise local aggregation; (2) FedCoSR w/o Sep: w/o separating the representation layer f and the linear layer g; (3) FedCoSR w/o CRL: w/o CRL loss term in local training. The results are shown in Fig. 8.

The ablation results show that removing LA leads to a notable performance drop, indicating that our loss-wise local aggregation effectively enhances model performance by leveraging model-level information. When Sep is disabled, the performance decline is minimal, but variance increases

TABLE III: The accuracy of changing N of CIFAR-10 for scalability evaluation, β of CIFAR-10 for practical heterogeneity evaluation, and label classes of each client of CIFAR-100 for pathological heterogeneity evaluation. We use superscripts $*$, †, and ‡, to emphasize the 1st, 2nd, and 3rd best values in each column, respectively. Purple + means improvement on previous settings, e.g., $N = 100$ v.s. $N = 20$, and $\beta = 0.01$ v.s. $\beta = 0.1$ in Table II, while blue − means degradation. Green ↑ means better and red ↓ means worse than the averaged result.

Method	Scalability (CIFAR-10)				Heterogeneity (CIFAR-10 practical)				Heterogeneity (CIFAR-100 pathological)			
	Prac. $N = 100$		Path. $N = 100$		$\beta = 0.01$		$\beta = 1$		Classes/Client$= 10$		Classes/Client$= 50$	
	Acc. (%)	Std. (%)	Acc. (%)	Std. (%)	Acc. (%)	Std. (%)	Acc. (%)	Std. (%)	Acc. (%)	Std. (%)	Acc. (%)	Std. (%)
Local	65.72 - ↓	18.47 - ↓	67.44 + ↓	15.90 - ↓	44.25 - ↓	20.77 - ↓	77.57* + ↑	6.82 + ↓	30.13 - ↓	7.12 - ↓	34.89 + ↑	6.80 - ↓
FedAvg	57.86 - ↓	10.34‡ + ↑	59.80 + ↓	13.80 - ↓	30.45 - ↓	22.34 - ↓	71.23 + ↑	4.26 + ↑	20.94 - ↓	6.24 - ↓	31.19 + ↓	2.59 + ↑
FedProx	60.82 - ↓	8.81* - ↑	65.26 - ↑	7.69‡ - ↑	46.13 - ↓	13.93 - ↓	70.31 + ↑	4.11‡ + ↑	28.84 - ↓	4.08‡ - ↑	33.37 + ↑	1.85* + ↑
MOON	80.05 - ↑	16.80 - ↓	74.15 - ↓	15.33 - ↓	95.92 + ↓	12.55 - ↑	66.32 - ↓	4.97 - ↑	57.62 + ↑	4.99 + ↓	23.47 - ↓	4.58 + ↓
pFedMe	80.22 - ↑	13.99 - ↑	76.09 - ↑	8.65 - ↑	98.82 + ↑	8.98‡ + ↑	69.01 - ↓	4.39 + ↑	59.22 + ↑	4.66 + ↑	28.15 - ↓	2.84 + ↑
Ditto	82.68 - ↑	15.78 - ↓	74.85 - ↓	8.48 + ↑	99.05 + ↑	8.60† + ↑	65.03 - ↓	4.86 + ↑	57.21 + ↑	4.59 - ↑	27.70 - ↓	2.90 + ↑
PerFedAvg	84.03 - ↑	12.95 - ↑	79.30 - ↑	8.55 + ↑	99.02 + ↑	8.28* - ↑	73.93 - ↑	4.07† + ↑	63.46 + ↑	4.93 + ↓	36.17† - ↑	2.50‡ + ↑
FedRep	86.08† - ↑	13.23 - ↑	82.31‡ - ↑	7.44† - ↑	99.18 + ↑	9.06 - ↑	71.84 -↑	4.59 + ↑	68.85† + ↑	4.16 + ↑	33.71 - ↑	3.83 + ↓
LG-FedAvg	82.87 - ↑	16.44 - ↓	76.72 - ↑	8.80 + ↑	99.17 + ↑	22.72 - ↓	61.38 -↓	6.04 + ↓	62.45 + ↑	7.53 + ↓	24.52 - ↓	5.17 + ↓
FedPer	84.43‡ - ↑	13.60 - ↑	82.08 - ↑	8.04 - ↑	98.70 + ↑	12.58 - ↑	73.00 - ↑	4.16 + ↑	68.35 + ↑	4.28 + ↑	35.84 - ↑	4.04 + ↓
FedAMP	82.26 - ↑	13.25 - ↑	75.03 - ↓	13.04 - ↓	99.25‡ + ↑	15.82 - ↓	61.58 - ↓	9.65 - ↓	67.51 + ↑	4.80 + ↑	28.89 - ↓	4.34 + ↓
FedALA	79.27 - ↑	15.71 - ↓	70.44 - ↓	15.28 - ↓	97.09 + ↑	10.26 - ↑	72.99 - ↑	4.78 + ↑	30.84 - ↓	4.67 + ↑	30.82 - ↓	4.29 + ↓
FedProto	83.61 - ↑	15.96 - ↓	78.01 - ↑	8.20 - ↑	99.32† + ↑	11.21 - ↑	62.92 - ↓	5.17 + ↓	68.81‡ + ↑	4.56 + ↑	31.48 - ↓	4.64 + ↓
FedPAC	70.80 - ↓	20.04 - ↓	82.33† - ↑	9.59 - ↑	-	-	74.84‡ - ↑	3.44* + ↑	63.30 + ↑	2.93* + ↑	36.12‡ - ↑	3.60 + ↑
FedGH	82.64 - ↑	15.66 - ↓	76.98 - ↑	8.52 - ↑	99.23 + ↑	22.45 - ↓	61.16 - ↓	5.25 + ↓	65.50 + ↑	4.74 - ↑	27.55 - ↓	3.98 + ↓
FedCoSR	**87.57*** - ↑	**12.26**‡ - ↑	**84.81*** - ↑	**7.37*** - ↑	**99.40*** + ↑	**10.78** - ↑	**76.59**† - ↑	**4.17** + ↑	**72.44*** + ↑	**3.56**† + ↑	**42.78*** + ↑	**2.17**† + ↑
Averaged	78.12	14.58	75.28	10.29	86.98	13.35	69.24	5.04	55.24	4.86	31.48	3.75

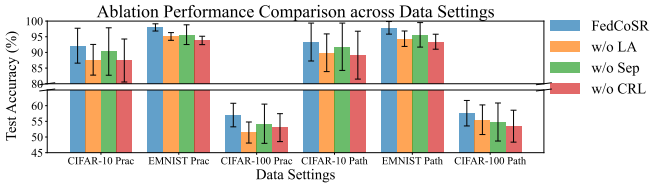

Fig. 8: The accuracy of FedCoSR and its ablated variants under the default settings.

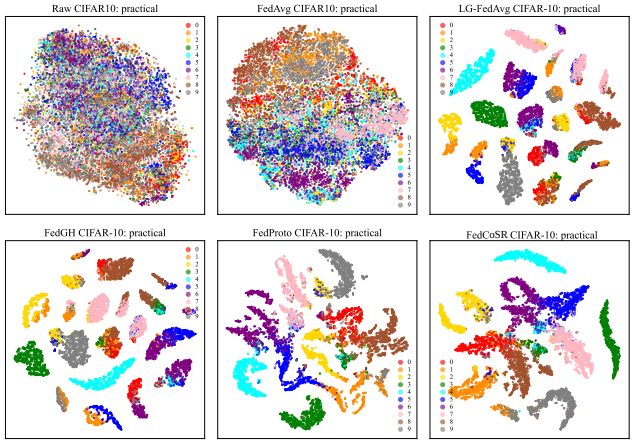

Fig. 9: Practical setting: Visualization of the raw data, representations of FedCoSR and other four baselines through t-SNE.

significantly, highlighting that model separation helps preserve local knowledge without disrupting the global model's convergence, and enhances fairness by maintaining balanced learning. Lastly, removing CRL harms both accuracy and fairness, which can validate that CRL effectively integrates label-wise information across clients into each local model, thereby enhancing overall prediction accuracy in a fair manner.

2) Visualization of Representations: We visualize the CIFAR-10 dataset using t-SNE. Fig. 9 shows that while FedAvg achieves some clustering, PFL demonstrates more distinct clusters based on data patterns. Model splitting-based methods like LG-FedAvg and FedGH mix at least two label types within clusters due to incomplete aggregation, hindering effective fusion of representation and projection layers, making model-splitting suboptimal. In contrast, FedProto and FedCoSR more clearly separate representations into uniform clusters, with FedProto showing some overlap between different labels, especially in the pathological setting. This occurs because FedProto lacks model-level information and focuses only on local personalization, whereas FedCoSR combines model aggregation with shared representation learning, balancing generalization and personalization.

VI. Conclusion and Future Work

This paper presents FedCoSR, a PFL framework that applies contrastive learning to shareable representations to deal with label heterogeneity, including label distribution skew and data scarcity. It enhances local model training by leveraging global representations to form sample pairs, thereby enriching the knowledge of clients, especially those with limited data. The proposed loss-wise weighting model aggregation dynamically balances local and global models, ensuring personalized performance. Experiments demonstrate that FedCoSR outperforms compared methods in various heterogeneous settings, showing its effectiveness and fairness with heterogeneous or scarce data. In future work, we intend to extend the practicality of FedCoSR by studying its potential for addressing other forms of statistical heterogeneity, including feature condition skew, where clients exhibit similar label distributions but distinct sample distributions.

References

[1] G. Fortino, C. Savaglio, G. Spezzano, and M. Zhou, "Internet of things as system of systems: A review of methodologies, frameworks, platforms, and tools," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 1, pp. 223–236, 2021.

[2] W. Y. B. Lim, N. C. Luong, D. T. Hoang, Y. Jiao, Y.-C. Liang, Q. Yang, D. Niyato, and C. Miao, "Federated learning in mobile edge networks: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 2031–2063, 2020.

[3] X. Li, Z. Qu, B. Tang, and Z. Lu, "Fedlga: Toward system-heterogeneity of federated learning via local gradient approximation," *IEEE Transactions on Cybernetics*, vol. 54, no. 1, pp. 401–414, 2024.

[4] C. Zhang, Y. Xie, H. Bai, X. Hu, B. Yu, and Y. Gao, "Federated active semi-supervised learning with communication efficiency," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 53, no. 11, pp. 6744–6756, 2023.

[5] M. Kaheni, M. Lippi, A. Gasparri, and M. Franceschelli, "Selective trimmed average: A resilient federated learning algorithm with deterministic guarantees on the optimality approximation," *IEEE Transactions on Cybernetics*, vol. 54, no. 8, pp. 4402–4415, 2024.

[6] L. Zhang, W. Cui, B. Li, Z. Chen, M. Wu, and T. S. Gee, "Privacy-preserving cross-environment human activity recognition," *IEEE Transactions on Cybernetics*, vol. 53, no. 3, pp. 1765–1775, 2023.

[7] S. Liang, J. Lam, and H. Lin, "Secure estimation with privacy protection," *IEEE Transactions on Cybernetics*, vol. 53, no. 8, pp. 4947–4961, 2023.

[8] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[9] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *Foundations and trends® in machine learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[10] M. Ye, X. Fang, B. Du, P. C. Yuen, and D. Tao, "Heterogeneous federated learning: State-of-the-art and research challenges," *ACM Computing Surveys*, vol. 56, no. 3, pp. 1–44, 2023.

[11] A. Imteaj, U. Thakker, S. Wang, J. Li, and M. H. Amini, "A survey on federated learning for resource-constrained iot devices," *IEEE Internet of Things Journal*, vol. 9, no. 1, pp. 1–24, 2022.

[12] A. Z. Tan, H. Yu, L. Cui, and Q. Yang, "Towards personalized federated learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 12, pp. 9587–9603, 2023.

[13] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proceedings of Machine Learning and Systems*, I. Dhillon, D. Papailiopoulos, and V. Sze, Eds., vol. 2, 2020, pp. 429–450. [Online]. Available: https://proceedings.mlsys.org/paper_files/paper/2020/file/1f5fe83998a09396ebe6477d9475ba0c-Paper.pdf

[14] C. T Dinh, N. Tran, and J. Nguyen, "Personalized federated learning with moreau envelopes," *Advances in neural information processing systems*, vol. 33, pp. 21 394–21 405, 2020.

[15] T. Li, S. Hu, A. Beirami, and V. Smith, "Ditto: Fair and robust federated learning through personalization," in *International conference on machine learning*. PMLR, 2021, pp. 6357–6368.

[16] M. G. Arivazhagan, V. Aggarwal, A. K. Singh, and S. Choudhary, "Federated learning with personalization layers," *arXiv preprint arXiv:1912.00818*, 2019.

[17] P. P. Liang, T. Liu, L. Ziyin, N. B. Allen, R. P. Auerbach, D. Brent, R. Salakhutdinov, and L.-P. Morency, "Think locally, act globally: Federated learning with local and global representations," *arXiv preprint arXiv:2001.01523*, 2020.

[18] Q. Li, B. He, and D. Song, "Model-contrastive federated learning," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 10 708–10 717.

[19] L. Collins, H. Hassani, A. Mokhtari, and S. Shakkottai, "Exploiting shared representations for personalized federated learning," in *International conference on machine learning*. PMLR, 2021, pp. 2089–2099.

[20] S. Wang, X. Fu, K. Ding, C. Chen, H. Chen, and J. Li, "Federated few-shot learning," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, pp. 2374–2385.

[21] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[22] B. Karg and S. Lucia, "Efficient representation and approximation of model predictive control laws via deep learning," *IEEE Transactions on Cybernetics*, vol. 50, no. 9, pp. 3866–3878, 2020.

[23] Y. Tan, G. Long, L. LIU, T. Zhou, Q. Lu, J. Jiang, and C. Zhang, "Fedproto: Federated prototype learning across heterogeneous clients," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 8, pp. 8432–8440, Jun. 2022. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/20819

[24] L. Yi, G. Wang, X. Liu, Z. Shi, and H. Yu, "Fedgh: Heterogeneous federated learning with generalized global header," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 8686–8696.

[25] J. Xu, X. Tong, and S.-L. Huang, "Personalized federated learning with feature alignment and classifier collaboration," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=SXZr8aDKia

[26] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[27] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach," *Advances in neural information processing systems*, vol. 33, pp. 3557–3568, 2020.

[28] Y. Huang, L. Chu, Z. Zhou, L. Wang, J. Liu, J. Pei, and Y. Zhang, "Personalized cross-silo federated learning on non-iid data," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 9, pp. 7865–7873, May 2021. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/16960

[29] J. Zhang, Y. Hua, H. Wang, T. Song, Z. Xue, R. Ma, and H. Guan, "Fedala: Adaptive local aggregation for personalized federated learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 9, pp. 11 237–11 244, Jun. 2023. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/26330

[30] X. Shi, L. Yi, X. Liu, and G. Wang, "Ffedcl: Fair federated learning with contrastive learning," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[31] Y. Zhang, Y. Xu, S. Wei, Y. Wang, Y. Li, and X. Shang, "Doubly contrastive representation learning for federated image recognition," *Pattern Recognition*, vol. 139, p. 109507, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320323002078

[32] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.

[33] Y. Tan, G. Long, J. Ma, L. Liu, T. Zhou, and J. Jiang, "Federated learning from pre-trained models: A contrastive learning approach," *Advances in neural information processing systems*, vol. 35, pp. 19 332–19 344, 2022.

[34] X. Mu, Y. Shen, K. Cheng, X. Geng, J. Fu, T. Zhang, and Z. Zhang, "Fedproc: Prototypical contrastive federated learning on non-iid data," *Future Generation Computer Systems*, vol. 143, pp. 93–104, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X23000262

[35] J. Miao, Z. Yang, L. Fan, and Y. Yang, "Fedseg: Class-heterogeneous federated learning for semantic segmentation," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 8042–8052.

[36] Q. Yu, Y. Liu, Y. Wang, K. Xu, and J. Liu, "Multimodal federated learning via contrastive representation ensemble," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=Hnk1WRMAYqg

[37] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 1, pp. 857–876, 2023.

[38] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" *Advances in neural information processing systems*, vol. 27, 2014.

[39] B. Ghimire and D. B. Rawat, "Recent advances on federated learning for cybersecurity and cybersecurity for federated learning for internet of things," *IEEE Internet of Things Journal*, vol. 9, no. 11, pp. 8229–8249, 2022.

[40] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Annals of operations research*, vol. 134, pp. 19–67, 2005.

[41] A. Shamsian, A. Navon, E. Fetaya, and G. Chechik, "Personalized federated learning using hypernetworks," in *International Conference on Machine Learning*. PMLR, 2021, pp. 9489–9502.

[42] J. Jeon, K. Lee, S. Oh, J. Ok *et al.*, "Gradient inversion with generative image prior," *Advances in neural information processing systems*, vol. 34, pp. 29 898–29 908, 2021.

[43] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. S. Quek, and H. Vincent Poor, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3454–3469, 2020.