Improving Global Generalization and Local Personalization for Federated Learning

Lei Meng, *Member, IEEE*, Zhuang Qi, Lei Wu, Xiaoyu Du, Zhaochuan Li, Lizhen Cui, *Senior Member, IEEE*, and Xiangxu Meng

***Abstract*— Federated learning aims to facilitate collaborative training among multiple clients with data heterogeneity in a privacy-preserving manner, which either generates the generalized model or develops personalized models. However, existing methods typically struggle to balance both directions, as optimizing one often leads to failure in another. To address the problem, this article presents a method named personalized federated learning via cross silo prototypical calibration (pFedCSPC) to enhance the consistency of knowledge of clients by calibrating features from heterogeneous spaces, which contributes to enhancing the collaboration effectiveness between clients. Specifically, pFedCSPC employs an adaptive aggregation method to offer personalized initial models to each client, enabling rapid adaptation to personalized tasks. Subsequently, pFedCSPC learns class representation patterns on clients by clustering, averages the representations within each cluster to form local prototypes, and aggregates them on the server to generate global prototypes. Meanwhile, pFedCSPC leverages global prototypes as knowledge to guide the learning of local representation, which is beneficial for mitigating the data imbalanced problem and preventing overfitting. Moreover, pFedCSPC has designed a cross-silo prototypical calibration (CSPC) module, which utilizes contrastive learning techniques to map heterogeneous features from different sources into a unified space. This can enhance the generalization ability of the global model. Experiments were conducted on four datasets in terms of performance comparison, ablation study, in-depth analysis, and case study, and the results verified that pFedCSPC achieves improvements in both global generalization and local personalization performance via calibrating cross-source features and strengthening collaboration effectiveness, respectively.**

***Index Terms*— Data heterogeneity, federated learning (FL), generalization, personalization, prototypical calibration.**

I. Introduction

FEDERATED learning (FL), as a solution to address the problem of data silos, enables multiple clients with data heterogeneity to collaboratively train models in a privacy-preserving manner [1], [2], [3]. FL has two optimization objectives: global generalization [4], [5], [6], [7] and local personalization [8], [9], [10], [11]. Recently, various methods have been proposed to optimize different objectives. However, the simultaneous consideration of both directions has been rarely explored in the existing research. This leads to difficulties in FL systems as they struggle to meet the needs of both servers and clients simultaneously.

To enhance the generalization of the federated model, three methods have been proposed: data sharing, mitigating the local drift on the client side, and optimizing the aggregation scheme on the server. The first method aims to use synthetic or public datasets to create balanced data distributions to build unbiased local models [12], [13]. The second method commonly leverages global knowledge to regularize the learning of local, aiming to enhance model output consistency across client, where global knowledge typically represents the output of a global model or the average of outputs from multiple participants [14], [15], [16]. The third method addresses the performance decline issue caused by directly averaging parameters of local models. It either introduces new strategies to improve the aggregation effectiveness [17], [18] or performs fine-tuning to boost the knowledge transfer [19]. Despite the overall performance improvement, the issue of inconsistency in the cross-source feature space needs to be addressed. On the other hand, personalized FL methods can also be roughly divided into three categories. The first method typically performs model fine-tuning on local data to fit the personalized objective [20], [21]. The second method aims to compute personalized aggregation weights for each client to generate personalized models for specific clients [22], [23], [24], [25], [26]. The third method focuses on decoupling the network architecture by splitting it into personalized layers and global layer, which can improve the flexibility of model [27], [28], [29]. As observed, optimizing for a single objective has yielded impressive results. However, there is a lack of new solutions to balance both directions.

To address this problem, this article presents a method, termed personalized federated learning via cross silo prototypical calibration (pFedCSPC), which regularizes the learning of consistent knowledge among clients to improve the collaboration effectiveness between clients. As illustrated in Fig. 1, compared with conventional methods, pFedCSPC considers both optimizing global generalization and local personalization simultaneously. Specifically, for the global optimization, pFedCSPC utilizes clustering to learn prototypical representations of local data on the client side, which precisely modeling the

Manuscript received 31 May 2023; revised 23 January 2024 and 11 May 2024; accepted 6 June 2024. Date of publication 19 July 2024; date of current version 8 January 2025. This work was supported in part by the National Key Research and Development Program of China under Grant 2021YFC3300203, in part by the TaiShan Scholars Program under Grant tsqn202211289, in part by the National Natural Science Foundation of China under Grant 62172226, and in part by the 2021 Jiangsu Shuangchuang (Mass Innovation and Entrepreneurship) Talent Program under Grant JSSCBS20210200. *(Corresponding author: Zhuang Qi.)*

Lei Meng is with the School of Software, Shandong University, Jinan 250101, China, and also with Shandong Research Institute of Industrial Technology, Jinan 250098, China (e-mail: lmeng@sdu.edu.cn).

Zhuang Qi, Lei Wu, Lizhen Cui, and Xiangxu Meng are with the School of Software, Shandong University, Jinan 250101, China (e-mail: z_qi@mail.sdu.edu.cn; i_lily@sdu.edu.cn; clz@sdu.edu.cn; mxx@sdu.edu.cn).

Xiaoyu Du is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: duxy.me@gmail.com).

Zhaochuan Li is with Inspur, Jinan 250101, China (e-mail: lizhaoch@inspur.com).

Digital Object Identifier 10.1109/TNNLS.2024.3417452

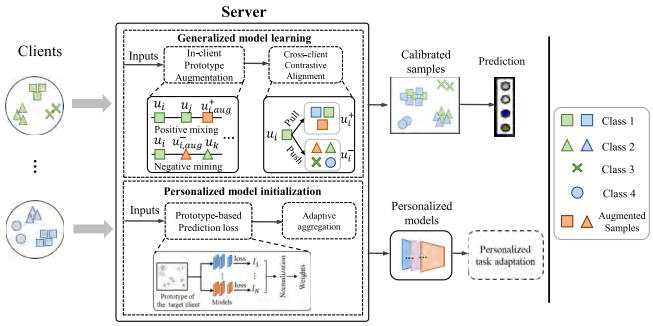

Fig. 1. pFedCSPC calibrates the representation space of heterogeneous clients on the server side, which improves the generalization capability of the global model. Moreover, pFedCSPC adopts adaptive PMI, which allows for rapid adaptation to personalized tasks.

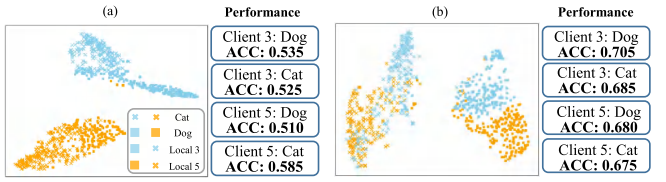

Fig. 2. Feature distributions learned by FedAvg and pFedCSPC. pFedCSPC effectively generalizes to client-side samples by learning to calibrate client-side prototypes. (a) Heterogeneous representations. (b) Cross-silo aligned representations.

distribution of local representations. Furthermore, pFedCSPC utilizes the cross-silo prototypical calibration (CSPC) module on the server to align prototypes from different spaces, which serves as a key component in addressing feature heterogeneity. Notably, to enhance the robustness of the calibration, the CSPC module employs a prototype augmentation (PA) method to increase sample diversity and leverages contrastive learning techniques to achieve prototypical calibration between clients, which aids in strengthening the generalization ability of the global model. For the personalized optimization, pFedCSPC introduces a personalized model initialization (PMI) method based on prototype prediction loss, which provides clients with models that are better suited for personalized tasks. Subsequently, pFedCSPC utilizes global knowledge to guide local representation learning, which effectively addressing the challenges of imbalanced data and overfitting. Importantly, the CSPC method exhibits high adaptability, seamlessly integrating into diverse algorithms. It has been observed that pFedCSPC effectively reduces the feature gap between data sources and improve the collaboration efficiency among clients.

Experiments are conducted in terms of performance comparison, ablation studies of the main components of pFedCSPC, in-depth analysis, and case study for the effectiveness of cross-silo feature alignment and PMI. The results verify that pFedCSPC is capable of mapping heterogeneous representations into a unified space, which leads to improved generalization of the global model. Moreover, it can be observed that regularizing clients to learn consistent knowledge enhances collaboration among them.

To summarize, this article includes three main contributions.

1) This article proposes an FL framework (pFedCSPC) to improve the global generalization and local personalization. It leverages consistent knowledge to facilitate collaborative modeling, which is beneficial for both optimization objectives.
2) This article proposes a plug-and-play module, named CSPC module, that can be easily integrated into existing infrastructure, enhancing versatility without altering core components. It is an orthogonal improvement to client-based methods.
3) This study reveals two findings: the heterogeneity of features between clients poses a challenge for global optimization, and the inconsistency of knowledge hinders the collaboration of clients. Furthermore, pFedCSPC employs CSPC to align features from different spaces and utilizes global knowledge to guide clients in learning similar representations.

II. Related Work

The optimization objectives of existing FL methods can be divided into two types: global optimization objective and personalized optimization objective.

A. FL for Global Optimization

Global optimization methods, which aim to learn a generalized model to fit the data from all clients, can be roughly classified into the following three categories.

1) Methods That Mitigate Client Drift: In the local training phase, common strategies involve leveraging global information as knowledge to guide local updates. Traditional approaches within this line of research include methods based on weights [14], [30], features [31], [32], [33], [34], [35], [36], and predictions [15], [37]. Weight-based methods employ proximal terms to reduce disparities between local and global models or use drift factors to correct parameter deviations. Feature-based approaches aim to penalize inconsistencies by emphasizing feature contrast. They usually involve aligning the local and global outputs in a latent space or utilizing prototypes to constrain clients' learning of similar representations. However, it has been observed that these methods are not able to fully address the issue of feature heterogeneity, leading to limited performance gains (see Fig. 2). Prediction-based methods typically rely on public datasets, and they integrate local soft-label predictions on the auxiliary dataset rather than model parameters or gradients, which is beneficial for reducing communication costs and achieving knowledge distillation. Federated learning framework of bias-eliminating augmentation learning (FedBEAL) utilizes a bias-eliminating augmenter and generated bias-conflicting samples to perform debiasing local updates on each client [38]. Additionally, federated learning framework with a simplex equiangular tight frame (FedETF) utilizes a fixed simplex classifier to guide clients in learning consistent representation spaces [39].

2) Methods That Optimize Aggregation Scheme: Recently, many studies aimed to optimize the aggregation scheme on the server side. For example, federated learning with matched averaging (FedMA) leverages a Bayesian nonparametric approach to match neurons; instead of simply averaging them [17], FedAvgM incorporates the momentum method in updating the global model to enhance its resilience to heterogeneous distributed data [40]. FedNova resolves inconsistencies

by normalizing local updates before their aggregation [18], and elastic aggregation adaptively interpolates client models based on parameter sensitivity, which is measured by evaluating the overall change in the prediction function output when each parameter varies [41]. Moreover, additional strategies such as retraining or fine-tuning are employed to address potential model shifts following the aggregation process, such as FedFTG incorporates an auxiliary generator that produces synthetic data for retraining, which can model the input space of local models [42]. CCVR [19] and CReFF [43] demonstrate that the variability in classifiers is the primary factor leading to decreased performance in models trained on non-independent and identically distributed (IID) data. Therefore, they both retrain the classifier using virtual and federated features, respectively. FedSoup utilizes selective interpolation of model parameters to balance local and global performances. Gradient correction method (GRACE) utilizes feature-aligned regularization to correct overfitting in personalized gradients. Additionally, it employs consistency-enhanced reweighted aggregation to calibrate gradients for improved generalization.

3) Methods That Train Model With Auxiliary Data: Due to the heterogeneity of data sources, local models trained on the client side may have limited generalization ability for samples from missing classes. Therefore, existing studies propose ideas for sharing data. Common practices involve sharing public datasets [13], generating synthesized datasets [44], [45], and using truncated versions of private data [12]. However, these approaches may compromise privacy preservation rules as they expose the raw data to other parties.

B. FL for Personalized Optimization

Personalized FL aims to utilize collaborative computing capabilities to train personalized models for each participating party based on their specific needs, which can be categorized into the following three methods.

1) Methods That Fine-Tune the Generalized Model: In this line of work, they typically enhance the generalization of the global model and then perform model fine-tuning on local data [20], [21]. These methods can be further divided into two subclasses: data-based and model-based approaches. The former typically utilizes data augmentation to alleviate statistical heterogeneity among different data sources [45], [46], [47] or adaptively samples client subsets to accelerate convergence [48], [49]. However, while data augmentation can help create balanced datasets, it requires sharing some data samples, which can potentially lead to privacy leakage. The latter utilizes the global model to regularize the local learning process in order to enhance the personalization capability of the local models [50], [51], [54], such as personalize locally, generalize universally (PLGU) employs generalized information to regulate the personalization capability of local models, which prevents clients from falling into suboptimal performance [54].

2) Methods That Aggregate Local Models With Personalized Weights: Recent studies have focused on generating personalized aggregation weights to create client-specific models [22], [23], [24], [25], [26]. For instance, FedAMP introduces a federated attention message-passing mechanism to foster increased collaboration among similar clients [22]. FedPHP utilizes rule-based moving averages and predefined hyperparameters to aggregate the global model and the corresponding local model [23]. FedFomo employs a local validation set to estimate the optimal weights for each client's personalized model [24]. Adaptive personalized cross-silo federated learning (APPLE) conducts local aggregation in every training batch, going beyond local initialization alone [25].

3) Methods That Learn Personalized Layers: Some studies focus on decoupling the network architecture by splitting it into personalized and global layers [27], [28], [29], [55], [56]. For instance, FedPer designates the feature extraction component as the global layer while treating the classifier as the personalized layer [28]. In contrast, LG-Fed aggregates the classifiers of all clients on the server while maintaining personalized feature extractors [55]. Cyclic distillation-guided channel decoupling federated learning framework (CD^2-pFed) assigns an adaptive ratio of learnable personalized weights to each layer of the model and keeps the personalized parameters locally [56].

III. Problem Formulation

In FL system, there are N clients denoted as $\mathcal{C} = \{C_1, C_2, \ldots, C_N\}$ and a server S. The client C_k owns a local dataset $D_k = \{(\mathcal{X}_k, \mathcal{Y}_k)\}$ and a local model $M_k = E_k \odot H_k \odot F_k$ with parameters $w_k = w_k^E \odot w_k^H \odot w_k^F$, where E_k is an image encoder with parameters w_k^E, H_k denotes the projection head with parameters w_k^H, and F_k is a classifier with parameters w_k^F. The aim of global optimization is to jointly train a generalized model under the supervision of the server S while maintaining privacy and minimizing the following objective:

$$w_g^* = \arg\min_w \sum_{C_k \in \mathcal{C}} p_k L_k(w; D_k) \tag{1}$$

where $L_k(w) = \mathbb{E}_{(x,y)\sim\mathcal{D}_k}[\ell_k(w; (x, y))]$ is the objective loss of C_k, and $p_k = (|D_k|/D)$ is the corresponding weight, where $D = \sum_{C_k \in \mathcal{C}} |D_k|$. After local training, clients $C_k \in \mathcal{C}$ upload the local parameters w_k to sever, and the server aggregates these parameters by

$$w_g = \sum_{C_k \in \mathcal{C}} p_k w_k. \tag{2}$$

For personalized optimization, the corresponding objective can be expressed as

$$\left\{w_1^*, \ldots, w_N^*\right\} = \arg\min_w \sum_{k=1}^{N} L_k(w_k; D_k) \tag{3}$$

where $w = \{w_k\}_{k=1}^N$ contains all personal model parameters.

In contrast, to improve the generalization of the global model, the proposed pFedCSPC introduces a CSPC module on the server, which aims to relearn the global projection head $H_g \mapsto \hat{H}_g$ and the classifier $F_g \mapsto \hat{F}_g$ to align representations from different feature spaces, i.e., $\hat{H}_g(E_{k_1}(x_{k_1})) \approx \hat{H}_g(E_{k_2}(x_{k_2}))$, where x_{k_1} and x_{k_2} are the samples with the same label in clients C_{k_1} and C_{k_2}, respectively. pFedCSPC first generates class-aware prototypes in all clients, i.e., $\mathcal{U} = \{\mathcal{U}_k | k \in \mathcal{C}\}$ and $\mathcal{U}_k = \{u_k^i | i \in \mathcal{Y}_k\}$ for each class i, and sends them to the server. Subsequently, the CSPC module learns the

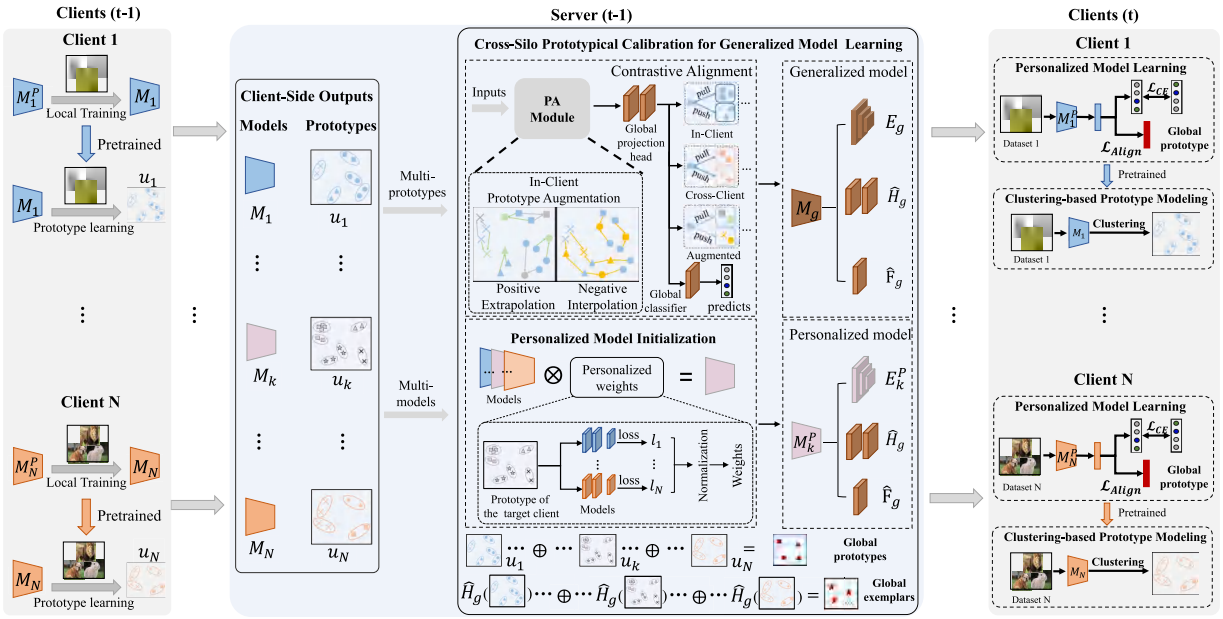

Fig. 3. Illustration of the framework of pFedCSPC. It uses prototypes obtained from the clients to retrain the global projection head $H_g(\cdot)$ and global classifier $F_g(\cdot)$ on the server to align features from different spaces at training round $t-1$. Meanwhile, it uses PA to improve the robustness of the calibration. Moreover, pFedCSPC employs adaptive PMI and utilizes the constraint of global knowledge to regularize local training at training round t, which can enhance the collaboration between clients.

mapping $\hat{H}_g(\cdot)$ based on these prototypes and the corresponding augmented samples $\mathcal{U}_{\text{aug}}$ to gather together cross-source features shared the same label. Finally, calibrated prototypes form a knowledge base to produce knowledge-based prediction (KP) $\text{Pred}_{\text{knowledge}}$. The final prediction $\text{Pred}_{\text{final}}$ is achieved by $\text{Pred}_{\text{knowledge}} \oplus \text{Pred}_{\text{net}} \mapsto \text{Pred}_{\text{final}}$, where Pred_{net} is the prediction of network. Moreover, to fit the personalized tasks, pFedCSPC employs the PMI to provide personalized models $\{m_k^p\}_{k=1}^N$ to clients. In addition, pFedCSPC utilizes knowledge distillation to guide local learning and alleviate the problems of data imbalance and overfitting.

IV. Approach

A. Overall Framework

pFedCSPC aims to improve the global generalization and the local personalization for the FL system. Fig. 3 illustrates the main framework of pFedCSPC. It first introduces the PMI method to allocate model weights suitable for the corresponding tasks to each client, and then uses global prototypes to guide the learning of local representation to regularize the consistency of local knowledge, which is capable of enhancing collaboration between clients. Furthermore, pFedCSPC performs data prototypical modeling to capture the distribution of representations and provide prototypical knowledge to the server. Finally, pFedCSPC performs CSPC to eliminate feature heterogeneity in the heterogeneous space, which can improve the generalization of the global model.

B. Personalized Model Learning

The personalized model learning module aims to achieve personalized benefits for clients by leveraging collaborative training manner. It has two main processes: PMI and personalized model training.

1) Personalized Model Initialization: PMI aims to customize specific model parameters for personalized tasks, rather than allocating uniform parameters to all users, which helps in rapidly adapting to personalized tasks. An intuitive approach is to assign greater weights to models that benefit specific objectives, and conversely, assign smaller weights to models that do not contribute as significantly. Following this line of thinking, a weight allocation scheme based on prototype prediction loss has been designed to measure the contribution of models to clients:

$$p_k^{k'} = \frac{e^{-\mathcal{L}_{\text{sup}}(\mathcal{U}_k; M_{k'})/\tau_p}}{\sum_{\hat{k}=1}^{n} e^{-\mathcal{L}_{\text{sup}}(\mathcal{U}_k; M_{\hat{k}}/\tau_p)}} \tag{4}$$

where $p_k^{k'}$ denotes the weight of the local model $M_{k'}$ when generating the initial model for client k. τ_p is a temperature parameter. Notably, to better facilitate collaboration between clients, the calibrated projection head $\hat{H}_g$ and classifier $\hat{F}_g$ are combined with personalized image encoder $E_k^p = \sum_{k'=1}^{N} p_k^{k'} \times E_{k'}$ obtained by personalized weighting to fully utilizes global knowledge to supplement the specific needs of each client, i.e., the overall personalized model can be expressed by $M_k^p = E_k^p \odot \hat{H}_g \odot \hat{F}_g$. Note that $\hat{H}_g$ and $\hat{F}_g$ will be provided in Section IV-C2.

2) Personalized Task Adaptation: To mitigate the issues of local data class imbalance and overfitting, pFedCSPC uses global knowledge to guide the local representation learning. It incorporates regularization at the node, angle, and edge levels, as shown in Fig. 4(a). Specifically, we apply a prototype-based contrastive loss for point-level regularization to align the representation with the corresponding global prototype

$$\mathcal{L}_N = -\log \frac{\exp\left(f \cdot u_g^+/\tau_l\right)}{\exp\left(f \cdot u_g^+/\tau_l\right) + \sum \exp\left(f \cdot u_g^-/\tau_l\right)} \tag{5}$$

where f denotes the representation, and u_g^+ and u_g^- are the global prototypes of the same/different class as f, respectively. Note that the method for calculating global prototypes will be provided in Section IV-C2. τ_l is a temperature parameter. For

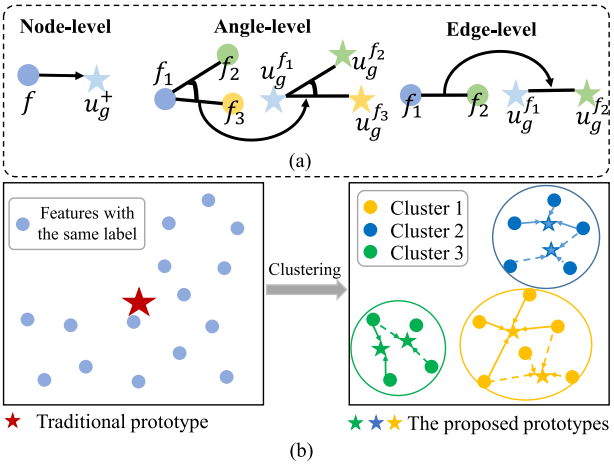

Fig. 4. (a) Three levels of regularization to guide local representation learning. (b) Clustering-based prototype learning method better fits the distribution patterns of samples than traditional methods.

the angle level, given three representations f_1, f_2, and f_3 with different labels, the corresponding prototypes are $u_g^{f_1}$, $u_g^{f_2}$, and $u_g^{f_3}$, and the angle-based alignment loss is defined as

$$\mathcal{L}_A = \left\| \left(\cos\angle(f_1, f_2, f_3), \cos\angle\left(u_g^{f_1}, u_g^{f_2}, u_g^{f_3}\right)\right) \right\|_1 \quad (6)$$

where $\cos\angle(f_1, f_2, f_3) = \langle (f_1 - f_2)/(\|f_1 - f_2\|_2), (f_3 - f_2)/(\|f_3 - f_2\|_2) \rangle$, and $\langle\cdot\rangle$ denotes the inner product. For the edge level, it requires the distance between the samples to be consistent with the corresponding prototypes

$$\mathcal{L}_E = \ell\left(\|f_1 - f_2\|_2 - \left\|u_g^{f1} - u_g^{f2}\right\|_2\right) \quad (7)$$

where $\ell(\cdot)$ is the L_2-norm. Notably, encouraging the learning of consistent knowledge among clients can promote cooperation between them.

C. Generalized Model Learning

The generalized model learning module aims to integrate knowledge from multiple clients to obtain a model that can fit all the data. It has two main processes: clustering-based prototype modeling (CPM) and CSPC.

1) Clustering-Based Prototype Modeling: Inspired by clustering [52], [53], after personalized model training, to assist with model calibration, pFedCSPC utilizes the K-means clustering method to explore patterns in the representation distribution and generate prototypes. The process can be described as follows:

$$c_i^1, c_i^2, \ldots, c_i^j = \textbf{K-means}(E(x), j), \quad x \in D_i \quad (8)$$

where j denotes the number of clusters. c_i^j is the jth cluster of class i, and D_i is the data with label i. $E(\cdot)$ is the encoder.

To better model the distribution of representations, we repeat the process of randomly sampling n_{repeat} times within each cluster to generate multiple class-aware prototypes, as shown in Fig. 4(b). Remarkably, this increases prototype diversity compared to traditional methods [16], [58]. The calculation of the prototype can be formulated as

$$u_i^{\hat{j},t} = \textbf{mean}\left\{ f | f \in \textbf{sampling}\left(c_i^{\hat{j}}, r\right)\right\} \quad (9)$$

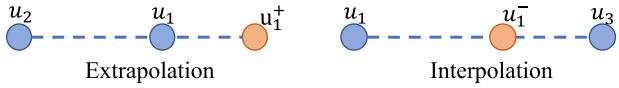

Fig. 5. Illustration of the generation of augmented samples via extrapolation and interpolation.

where $u_i^{\hat{j},t}$ represents the tth local prototype of cluster $c_i^{\hat{j}}$, **mean**$(\cdot)$ is the mean operation, and **sampling**$(c_i^{\hat{j}}, r)$ denotes the randomly select sample features with a proportion of r in cluster $c_i^{\hat{j}}$. Finally, client k sends the local model M_k and local prototype set $\mathcal{U}_k = \{u_i^{\hat{j},t} | i \in \mathcal{Y}_k, \hat{j} = 1, 2, \ldots, j, t = 1, \ldots, n_{\text{repeat}}\}$ as output to the server.

2) CSPC: The server obtains all local models and local prototype set $\{\mathcal{M}, \mathcal{U}\} = \{(M_k, \mathcal{U}_k) | k = 1, \ldots, N\}$ from clients. To align prototypical features obtained from heterogeneous spaces, an intuitive idea is to relearn the projection head $H_g(\cdot)$ and classifier $F_g(\cdot)$. However, there is a challenge hindering the robustness of calibration, which is the insufficient number of prototypes. Therefore, the pFedCSPC develops an augmented contrastive learning method, which contains in-client PA and cross-client contrastive alignment (CA).

a) In-client PA: As shown in Fig. 5, to expand the local prototype set, positive extrapolation and negative interpolation strategies are employed to generate new sample features, which is beneficial for information supplementation, i.e.,

$$u_1^+ = (u_1 - u_2) \times \lambda + u_1, \quad u_1^- = (u_3 - u_1) \times \lambda + u_1 \quad (10)$$

where u_1 and u_2 share the same label, whereas u_3 has a distinct label. λ is a constant coefficient. Notably, intraclass extrapolation keeps the main characteristics while increasing diversity. Meanwhile, interclass interpolation injects positive information into negative samples, which makes it more difficult for the model to distinguish the decision boundary. This is advantageous for improving the generalization of the model.

b) Cross-client CA: For the raw global model $M_g = E_g \odot H_g \odot F_g$, obtained by (2), the projection head $H_g(\cdot)$ and the classifier $F_g(\cdot)$ need to be calibrated, i.e., $H_g(\cdot) \mapsto \hat{H}_g(\cdot)$ and $F_g(\cdot) \mapsto \hat{F}_g(\cdot)$. To enhance the robustness of calibration, augmented samples for each class i are used as additional constraints to regularize the learning of mapping, i.e.,

$$\mathcal{L}_{\text{ACL}}(u_i, u_i^+, u_i^-) = \left|\left|H_g(u_i) - H_g\left(u_i^+\right)\right|\right|_2^2 - \left|\left|H_g(u_i) - H_g\left(u_i^-\right)\right|\right|_2^2 + \alpha \quad (11)$$

where α is the margin parameter. For the real samples of each class i, we maximize the similarity between prototypes of the same class from different sources via weighted contrastive learning

$$\mathcal{L}_{\text{WCL}}(u_i) = -\frac{1}{|P(u_i)|} \sum_{u_i^+ \in P(u_i)} \log \frac{\sigma \cdot \exp\left(z_i^{\text{T}} \cdot z_i^+ / \tau_g\right)}{\sum_{u_s \in I_s} \sigma \cdot \exp\left(z_i^{\text{T}} \cdot z_s / \tau_g\right)} \quad (12)$$

where $P(u_i)$ indicates the positive set of u_i. I_s denotes the sample set. $z_i = H_g(u_i)$, and τ_g is a temperature parameter. σ_j is a weighting factor. Considering that it is more difficult to pull samples from different sources closer and push samples from the same source farther away, we design the following rules: if the two samples being compared are of the same class

but from different clients, or are of different classes but from the same client, $\sigma = 1$; otherwise, $\sigma = 0.5$.

Meanwhile, to enhance the classification capability, the cross-entropy loss is used to further optimize the classifier $F_g(\cdot) \mapsto \hat{F}_g(\cdot)$

$$\mathcal{L}_{\text{sup}}(u) = -\sum_{j=1}^{N_c} \mathcal{I}(y = i) \log\left(\hat{y}_i\right) \tag{13}$$

where $\mathcal{I}(\cdot)$ denotes the indication function, and N_c represents the number of classes. y is the label of sample u, and $\hat{y}_i$ is the prediction that u belongs to class i.

In addition, pFedCSPC is unique in that it generates an exemplar e_i for each class in the unified space, which serves as a knowledge base to form a KP. For predictions based on knowledge $\text{Pred}_{\text{knowledge}}(x)$ and network $\text{Pred}_{\text{net}}(x)$ that may be in different ranges, both of them are normalized before fusion, which enables them to be mapped to the same interval. And the final prediction is generated by taking a weighted average of the two predictions, i.e.,

$$e^i = \frac{1}{N}\sum_{k=1}^{N} \frac{1}{n_{\text{repeat}}} \sum_{t=1}^{n_{\text{repeat}}} \hat{H}\left(u_k^{i,t}\right) \tag{14}$$

$$\begin{aligned}\text{Pred}_{\text{final}}(x) = (1-\lambda) \times \text{Norm}(\text{Pred}_{\text{net}}(x)) \\ + \lambda \times \text{Norm}\left(\text{Pred}_{\text{knowledge}}(x)\right)\end{aligned} \tag{15}$$

where $\text{Pred}_{\text{knowledge}}(x) = [\text{sim}(f_x, e_i)|i = 1, \ldots, N_c]$, $\text{sim}(f_x, e_i)$ denotes the similarity between the sample feature f_x and all exemplars $\{e_i|i = 1, \ldots, N_c\}$, $\text{Norm}(\cdot)$ denotes the normalization function, and λ is a weight parameter.

Finally, to encourage the learning of consistent knowledge between clients, pFedCSPC generates global prototypes $\mathcal{U}_g = \{u_g^i|i = 1, \ldots, N_c\}$ and sends them to all clients

$$u_g^i = \frac{1}{N}\sum_{k=1}^{N} \frac{1}{n_{\text{repeat}}} \sum_{t=1}^{n_{\text{repeat}}} u_k^{i,t}. \tag{16}$$

D. Training Strategies

For global optimization, pFedCSPC focuses on calibrating feature space on the server side, which can be combined with multiple client-based methods. Consequently, pFedCSPC has the following training strategy.

1) In the server, pFedCSPC aims to align heterogeneous features to eliminate heterogeneity and obtain clear decision boundaries, and it optimizes the following objective:

$$\begin{aligned}\mathcal{L}_{\text{server}} = \frac{1}{|I_s|}\sum_{u \in I_s} \mathcal{L}_{\text{sup}}(u) \\ + \eta\left[\mathcal{L}_{\text{WCL}}(u) + \mathcal{L}_{\text{ACL}}\left(u, u^+, u^-\right)\right]\end{aligned} \tag{17}$$

where η is a weight parameter.
For personalized optimization, pFedCSPC focuses on learning local knowledge without overfitting. It follows the strategy to achieve its goals.

2) In the client, the optimization objective varies depending on the base algorithm being used. Moreover, the alignment loss $\mathcal{L}_{\text{align}} = \mathcal{L}_N + \mathcal{L}_A + \mathcal{L}_E$ is used to regularize all clients to learn similar representations. Therefore, the overall optimization objective for a client is

$$\mathcal{L}_{\text{client}} = \mathcal{L}_{\text{base}} + \kappa \times \mathcal{L}_{\text{align}} \tag{18}$$

where κ is a weight parameter, and the base algorithm could be FedAvg, FedASAM, and so on.

TABLE I
STATISTICS OF CIFAR10, CIFAR100, TINYIMAGENET, AND VIREOFOOD172 DATASETS USED IN THE EXPERIMENT

Datasets	#Class	#Training	#Testing
CIFAR10	10	50000	10000
CIFAR100	100	50000	10000
TinyImagenet	200	100000	10000
VireoFood172	172	68175	33154

V. EXPERIMENTS

A. Experimental Settings

1) Datasets: To evaluate the performance of the algorithms, four datasets are used in the experiment, including CIFAR10 [59], [74], [75], CIFAR100 [59], TinyImagenet [60], and a challenging food classification dataset VireoFood172 [61], [70], [77]. Table I presents their statistical information. Following recent studies [16], [19], [31], the Dirichlet distribution is used to partition the training dataset. For personalized optimization, the local testing set and its corresponding training set exhibit an identical distribution to assess personalized performance.

2) Evaluation Measures: Following [1] and [31], for the global optimization, we use the top-1 accuracy (Acc) to evaluate the performance of methods, i.e.,

$$\text{ACC}_{\text{global}} = (\text{TP} + \text{TN})/(P + N) \tag{19}$$

where P, N, TP, and TN are positives, negatives, true positives, and true negatives, respectively. For personalized optimization, we evaluate both Acc and macro-F1 (MF1) score to mitigate the negative impact of overfitting. The indicator (I_{local}) is defined as

$$I_{\text{local}} = \frac{2 * \text{Acc} * \text{MF1}}{\text{Acc} + \text{MF1}}. \tag{20}$$

Moreover, we also calculated the harmonic mean (HM) of global generalization and local personalization performance, i.e.,

$$\text{HM} = \frac{2 \times \text{ACC}_{\text{global}} \times I_{\text{local}}}{\text{ACC}_{\text{global}} + I_{\text{local}}}. \tag{21}$$

3) Hyperparameter Settings: Following [16] and [31], for all methods, we set the number of clients $N = 10$ with the sample fraction $C = 1.0$, the number of local training epochs $E = 10$, the batch size $B = 64$, the communication round $T = 100$ for CIFAR10 and CIFAR100 datasets, $T = 50$ for TinyImagenet and VireoFood172 datasets, and the stochastic gradient descent (SGD) optimizer with the learning rate $\text{lr} = 0.01$ and the weight decay $\text{wd} = 1e-05$ is used in the local training. For all datasets, the Dirichlet parameter $\beta = 0.5$. The number of clusters for each class k is selected from $\{2, 3, 4\}$, the sample proportion $r = 0.5$, and the number of sampling $n_{\text{repeat}} = 5$. The constant coefficient λ is randomly

TABLE II

PERFORMANCE COMPARISON BETWEEN PFEDCSPC WITH BASELINES ON FOUR DATASETS. ALL ALGORITHMS WERE RUN BY THREE TRIALS, AND THE MEAN AND STANDARD DERIVATION ARE REPORTED

Methods	CIFAR10			CIFAR100			TinyImagenet			VireoFood172		
	Global	Local	HM	Global	Local	HM	Global	Local	HM	Global	Local	HM
FedAvg	66.78±0.4	64.13±0.7	65.43±0.5	66.54±0.3	63.15±0.6	64.80±0.4	45.38±0.4	41.32±1.0	43.25±0.6	59.88±0.8	54.21±0.6	56.89±0.7
MOON	69.04±0.7	63.49±0.5	66.14±0.6	67.88±0.4	64.79±0.7	66.29±0.6	47.31±0.9	42.48±0.7	44.76±0.8	61.25±1.1	54.65±0.6	57.76±0.8
FedDC	69.13±0.6	63.23±0.7	66.05±0.6	67.75±0.6	64.51±0.6	66.09±0.6	46.81±0.2	39.63±0.5	42.92±0.3	60.97±0.5	53.63±0.8	57.06±0.6
FedNTD	68.89±0.3	61.75±0.7	65.12±0.4	67.83±0.2	62.59±0.7	65.10±0.5	45.79±0.5	40.18±0.3	42.80±0.4	60.88±0.9	51.19±0.8	55.62±0.8
FedASAM	68.48±0.6	64.63±1.0	66.50±0.8	67.71±0.5	63.37±0.8	65.47±0.7	47.38±0.6	42.78±0.5	44.96±0.5	61.14±0.2	52.87±0.7	56.71±0.3
Fedproc	69.18±1.2	65.57±1.1	67.32±1.1	67.63±0.7	64.49±0.5	66.02±0.6	47.21±0.4	42.98±0.6	44.99±0.5	60.46±0.4	54.72±0.5	57.45±0.4
FedDecorr	68.66±0.8	65.81±0.6	67.20±0.7	67.79±0.6	65.21±0.9	66.47±0.8	46.21±0.7	41.83±0.5	43.91±0.6	61.06±0.7	53.65±0.9	57.11±0.8
CCVR	68.56±0.7	63.54±0.8	65.95±1.1	67.86±0.4	62.35±0.6	64.99±0.5	46.11±0.4	39.76±0.3	42.70±0.3	60.87±0.3	51.98±0.6	56.07±0.4
FedSoup	68.74±0.6	66.37±0.9	67.53±0.6	67.36±0.7	**65.51±0.6**	66.42±0.6	47.81±0.4	45.48±0.3	45.08±0.3	61.36±0.3	56.82±0.4	58.46±0.4
GRACE	69.21±0.7	67.86±0.5	68.52±0.6	67.14±0.7	66.22±0.7	65.90±0.5	47.36±1.1	**45.58±0.7**	46.45±0.8	60.61±0.3	57.13±0.8	58.81±0.5
FedETF	**70.84±0.8**	68.43±0.7	69.61±0.8	**68.67±0.7**	65.32±0.4	**66.95±0.5**	47.51±0.5	44.32±0.6	45.85±0.5	61.27±0.3	57.56±0.5	59.36±0.4
pFedCSPC	70.81±0.7	**69.81±0.5**	**70.31±0.6**	68.39±0.4	65.28±0.6	66.80±0.5	**47.87±0.6**	45.27±0.4	**46.53±0.5**	**62.19±0.6**	**57.89±0.6**	**59.96±0.6**

selected from $\{0.1, 0.3, 0.5\}$, the margin parameter $\alpha = 1.0$, the number of augmented samples for each prototype $n_{\text{aug}} = 5$, and the temperature parameters $\tau_l, \tau_g = 0.5$. For training strategies, both weight parameters κ and η are adjusted from $\{0.01, 0.05, 0.1, 0.5\}$. For other compared methods, we tuned their hyperparameters by referring to corresponding articles for fair comparison and optimal performance.

B. Performance Comparison

For global optimization, we compare pFedCSPC with ten state-of-the-art methods, including FedAvg [1], MOON [31], CCVR [19], FedDC [30], FedNTD [15], FedASAM [62], FedProc [16], FedDecorr [63], FedSoup [64], GRACE [65], and FedETF [39]. To validate the effectiveness of the proposed personalized strategies, we also conducted performance comparisons with specialized personalized FL methods, including FedAvg [1], FedAvg-FT (FedAvg with local fine-tuning), LG-Fed [55], FedPer++ [29], Ditto [66], FedFomo [24], and FedALA [26]. The network architecture used for all methods comprises an image encoder, a projection head, and a classifier. For all datasets, we employ a two-layer multilayer perceptron (MLP) as the projection head and the classifier is a fully connected layer. For the CIFAR10 dataset, we employ a convolutional neural network comprising two 5 $\times$ 5 convolutional layers, which are followed by 2 $\times$ 2 max pooling, and two fully connected layers with rectified linear unit (ReLU) function as the image encoder. For other datasets, we use a ResNet18, excluding its last fully connected layer. The following can be observed from Tables II and III.

1) pFedCSPC achieves the best results in most metrics, which can balance global generalization and local personalization. This is reasonable as its prototypical calibration effectively mitigates the challenges of feature heterogeneity and improves the generalization. Moreover, PMI aids in optimizing personalized tasks.
2) Incorporating calibration techniques into the learning process typically yields better global generalization than the baseline method. This is because the calibration mechanism can assist devices in learning a generalized model from various data sources, such as CCVR and pFedCSPC.
3) Due to the issue of data heterogeneity, focusing on optimizing a global model may not yield the most suitable model for personalized tasks. This is because not all client-side knowledge is useful for other personalized tasks.
4) As observed, the improvement achieved by combining pFedCSPC is significant compared to the baseline on CIFAR10, while it is relatively small on other datasets. This is understandable because the final Acc depends not only on the degree of bias correction after model calibration but also closely related to the quality of local representation learning.
5) pFedCSPC consistently outperforms other methods in personalized tasks. This may be attributed to two main factors: the adaptive weighting scheme and the stronger collaborative training approach.
6) With increasing difficulty of classification tasks, FedAvg with fine-tuning demonstrates stronger advantages. This is main because fine-tuning can alleviate the issue of underfitting and improve performance.
7) Fine-tuning or retraining additional models tends to outperform network decoupling methods. This is because the globally aggregated model may underfit personalized tasks, while retraining focuses on capturing local knowledge, and global constraints prevent overfitting.

TABLE III

PERSONALIZED OPTIMIZATION PERFORMANCE COMPARISON BETWEEN PFEDCSPC AND BASELINES ON FOUR DATASETS. ALL ALGORITHMS WERE RUN FOR THREE TRIALS, AND THE MEAN AND STANDARD DEVIATION ARE REPORTED

Methods	CIFAR10	CIFAR100	TinyImagenet	VireoFood172
FedAvg	63.17±1.3	61.03±0.8	38.07±0.9	49.08±1.0
FedAvg-FT	64.13±0.7	63.15±0.6	41.32±1.0	54.21±0.6
LG-Fed	58.92±0.3	60.38±0.5	40.76±0.5	53.25±0.8
FedPer++	62.82±0.8	62.17±1.1	41.89±0.6	55.23±1.1
Ditto	66.28±0.9	64.26±0.5	42.62±0.8	54.14±0.7
FedFomo	67.78±1.2	64.69±0.9	43.22±0.7	55.34±0.7
FedALA	67.14±0.4	63.53±0.7	43.74±0.9	56.46±0.5
pFedCSPC	**69.81±0.5**	**65.28±0.6**	**45.27±0.4**	**57.89±0.6**

C. Ablation Study

This section further investigates the effectiveness of different modules in both global and personalized optimizations. The results are summarized in Tables IV and V, respectively.

1) Simply combining the traditional prototype generation method (TPG [58]) with the cross-client CA may not bring performance gains, mainly because a single prototype cannot fit the overall distribution, and an insufficient

TABLE IV

ABLATION STUDY ON THE EFFECTIVENESS OF DIFFERENT COMPONENTS OF PFEDCSPC IN GLOBAL OPTIMIZATION FOR CIFAR10 AND CIFAR100 DATASETS WITH THE SAMPLE FRACTION $C = 0.5$ AND $C = 1.0$ AND THE DIRICHLET PARAMETER $\beta = 0.5$

	CIFAR10		CIFAR100	
	C=0.5	C=1.0	C=0.5	C=1.0
Base	65.71±0.6	66.78±0.4	65.01±0.7	66.54±0.3
+TPG+CA	63.49±0.7	64.32±0.5	62.37±0.5	64.68±0.5
+CPM+CA	67.66±0.2	68.89±0.3	66.32±0.4	67.35±0.3
+PTA+CPM+CA	68.14±0.6	69.51±0.4	67.04±0.1	68.02±0.7
+PTA+CPM+CA+KP	68.69±0.6	69.94±0.4	67.18±0.4	68.14±0.3
+TPG+PA+CA	64.84±0.3	66.19±0.6	61.79±0.6	65.74±0.2
+CPM+PA+CA	68.64±0.4	69.66±0.2	66.77±0.3	67.77±0.4
+PTA+CPM+PA+CA	69.01±0.6	70.42±0.4	67.24±0.1	68.21±0.7
+PTA+CPM+PA+CA+KP	**69.47±0.3**	**70.81±0.4**	**67.36±0.5**	**68.39±0.4**

TABLE V

ABLATION STUDY ON THE EFFECTIVENESS OF DIFFERENT COMPONENTS OF PFEDCSPC IN PERSONALIZED OPTIMIZATION FOR CIFAR10 AND CIFAR100 DATASETS WITH THE SAMPLE FRACTION $C = 1.0$ AND THE DIRICHLET PARAMETER $\beta = 0.5$

	CIFAR10	**CIFAR100**
FedAvg	63.17±1.3	59.03±1.1
+FT$_{CE}$	64.31±0.7	60.32±0.8
+PMI+FT$_{CE}$	65.12±0.5	61.43±0.4
+PMI+FT$_{CE+Align}$	**66.74±0.4**	**63.03±0.8**
FedCSPC	65.63±1.1	62.18±1.0
+FT$_{CE}$	66.98±0.4	63.47±0.4
+PMI+FT$_{CE}$	68.02±0.8	64.13±0.7
+PMI+FT$_{CE+Align}$	**69.81±0.5**	**65.28±0.6**

number of prototypes cannot provide enough information to train a generalized model.

2) Cross-client CA with the assistance of the CPM outperforms the base on both datasets with a large margin of up to 1.95%, 2.11%, 1.31%, and 0.81%, which verifies the effectiveness of the modeling of data patterns.
3) In general, using personalized task adaptation (PTA) and PA can further yield superior performance, as they improve the quality of client-side representation learning and increase sample diversity, respectively, which enhances the robustness of calibration.
4) As reported, KP demonstrates greater efficacy on the CIFAR10 dataset compared to the CIFAR100 dataset. This is mainly because it is easier to learn reliable classification boundaries in the representation space of CIFAR10 compared to CIFAR100.

1) *Simple fine-tuning* (FT$_{\text{CE}}$) can bring performance gains to personalized tasks. This is because it focuses on learning personalized knowledge while avoiding the interference of irrelevant information.
2) Improving the generalization ability of the global model is beneficial for accomplishing personalized tasks (see line FedCSPC and FedAvg, where FedCSPC denotes the global model without personalization). This demonstrates that collaborative training with local adaptation is a viable approach.
3) PMI method can provide benefits by initializing client-specific model that is more conducive to personalized task.
4) Prototype-guided feature learning (Align) further enhances the performance of the model on personalized tasks. This is mainly because it can guide the personalized model to learn clear class boundaries, which can alleviate the negative impact of imbalanced data distribution and avoid overfitting.

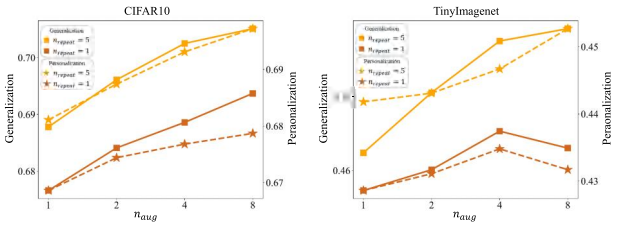

Fig. 6. Influence of the number of prototypes ($n_{\text{repeat}} = 1, 5$) on the final performance of pFedCSPC on the CIFAR10 and TinyImagenet datasets with the heterogeneity $\beta = 0.5$ and the number of augmented samples ($n_{\text{aug}} = 1, 2, 4, 8$).

D. In-Depth Analysis

1) Analysis of the Impact of Key Parameters on Generalization and Personalization: There are two key parameters that affect the robustness of model calibration, i.e., n_{repeat} and n_{aug}. We tune the parameters $n_{\text{repeat}} = \{1, 5\}$ and $n_{\text{aug}} = \{1, 2, 4, 8\}$. Fig. 6 shows the results of the comparison. In general, generating more class-aware prototypes leads to higher generalization score. Moreover, as the calibration robustness improves, the enhancement of model generalization ability also contributes to the performance gains in personalized fine-tuning. This is primarily because a generalized model is capable of learning more diverse and universal features and patterns, thus providing a stronger initial foundation for personalized tasks, enabling personalized models to adapt more quickly to the specific tasks of individual clients. Additionally, the knowledge transfer from the global model can guide the personalized training process and mitigate the risk of overfitting to local data. Meanwhile, the generalized model contains both relevant and irrelevant information for individual clients simultaneously, which results in a smaller improvement in personalized performance compared to generalization performance. However, due to the higher complexity of the representation space in TinyImagenet, the augmented samples may contain misleading information, which increases with the number of augmented samples. This hinders the improvement of the generalization and results in a decrease in both performances.

2) Visual Representation Analytics and Performance Interpretability: This section provides an in-depth analysis of the impact of the way local training on personalized performance from two perspectives: the quality of local representation learning and the efficiency of cooperation between clients. As shown in Fig. 7, the imbalance in the data distribution has led to poor representation learning in the traditional method. pFedCSPC can maintain clear decision boundaries under the guidance of global prototypes, enabling the model to better adapt to personalized tasks. This is because pFedCSPC explicitly delineates class-specific regions in the feature space, which can alleviate the drawback of insufficient quantity. Moreover, the heterogeneity of features hinders effective collaboration between clients, as illustrated in Fig. 7. This is because the

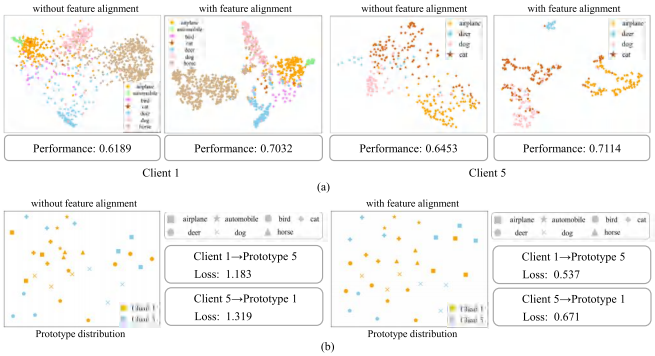

Fig. 7. (a) pFedCSPC learns more discriminative features. (b) pFedCSPC facilitates the learning of consistent representation knowledge between clients, which can promote collaboration.

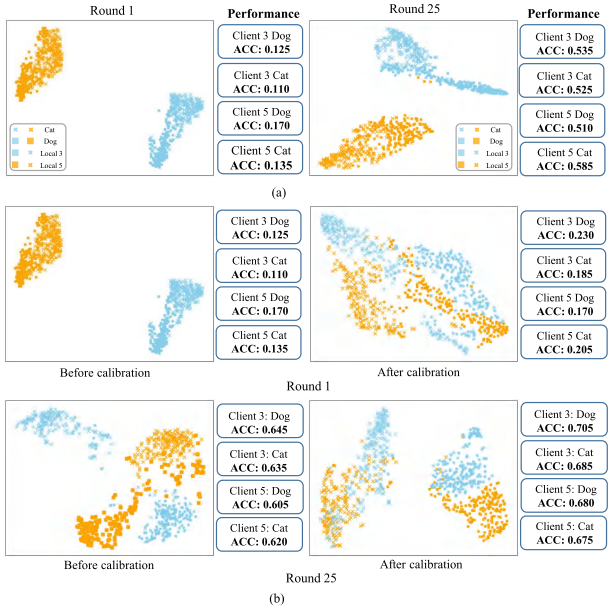

Fig. 8. Illustration of the effectiveness of cross-silo representation alignment. (a) For the same test data, the representation distributions extracted by different local models in the FedAvg method exhibit heterogeneity. (b) pFedCSPC effectively learns the common space of the same class but from different clients, which enables the global model to generalize to all clients.

features learned by clients cannot be correctly classified by other classifiers, which makes it difficult for a single client to leverage the extensive knowledge of others. pFedCSPC enables the establishment of shared knowledge representations among different clients, which allows each client to benefit from others and achieve a higher level of collaboration. For example, Fig. 7 shows the nearly consistent decision boundaries learned by clients. These observations provide tangible evidence for the interpretability and effectiveness of pFedCSPC.

3) Orthogonality of pFedCSPC With Existing FL Methods: Table VI gives the generalization performance of pFedCSPC using FedProx, MOON, and FedASAM. Overall, existing FL methods combined with pFedCSPC have demonstrated substantial improvements in classification compared to their corresponding baselines, highlighting the effectiveness and the orthogonality character of the pFedCSPC. Moreover, MOON and FedASAM have achieved the best performance on the CIFAR10 and TinyImagenet datasets, respectively. Besides, we have observed that all methods exhibited greater

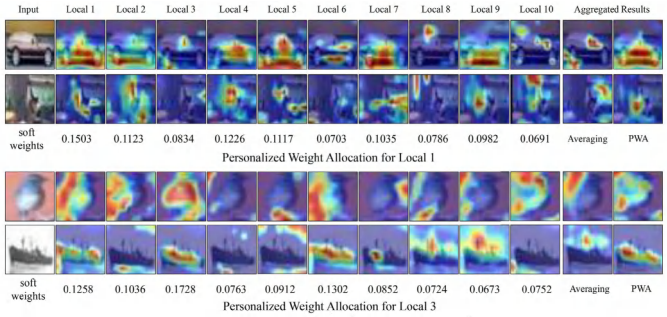

Fig. 9. Illustration of the effectiveness of PMI. FedAvg lacks the ability to assess client importance. pFedCSPC can focus on the clients that have significant contributions and obtain a better initialized model that is more suitable for personalized tasks.

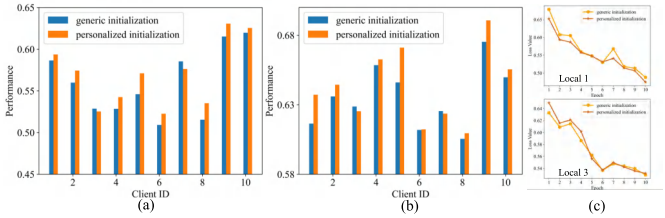

Fig. 10. Quantitative analysis of the performance of PMI. (a) Test performance of the initialized model. (b) Test performance after training different initialized models. (c) Training loss for different initializations.

improvements on the CIFAR10 dataset compared to TinyImagenet. This result is consistent with the conclusion in Table II. This is mainly because the quality of local representation learning is one of the factors that determine the performance of cross- silo calibration.

4) Communication Efficiency of the pFedCSPC: This section evaluates the number of communication rounds required by different methods to achieve the same test Acc on CIFAR10 and TinyImagenet with different levels of heterogeneity ($\beta = 0.1$ and 0.5), aiming for accuracies of 61.0 ($\beta = 0.1$) and 66.0 ($\beta = 0.5$) on CIFAR10 and 42.0 ($\beta = 0.1$) and 45.0 ($\beta = 0.5$) on TinyImagenet, respectively. Table VII shows the comparison results. We can observe that the number of communication rounds is significantly reduced in pFedCSPC. Notably, pFedCSPC achieves the same test Acc with less than one-third of the number of communication rounds required by FedAvg. Compared to other methods, pFedCSPC also demonstrates a significant advantage in communication efficiency.

E. Case Study

1) Cross-Silo Prototypical Calibration: In this section, we randomly selected two local models and two easily confused classes (cat and dog) and extracted 200 samples from the test set for each class. The TSNE [67] method was used to visualize the feature distribution of samples before and after calibration. We also output the corresponding classification Acc of the model before and after calibration. As shown in Fig. 8, pFedCSPC is capable of mapping features from disparate spaces to a unified space and maintaining clear decision boundaries. In contrast, FedAvg cannot eliminate heterogeneity during training. Moreover, pFedCSPC not only corrects the distribution of heterogeneous representations in the current round, but also promotes consistency in learning

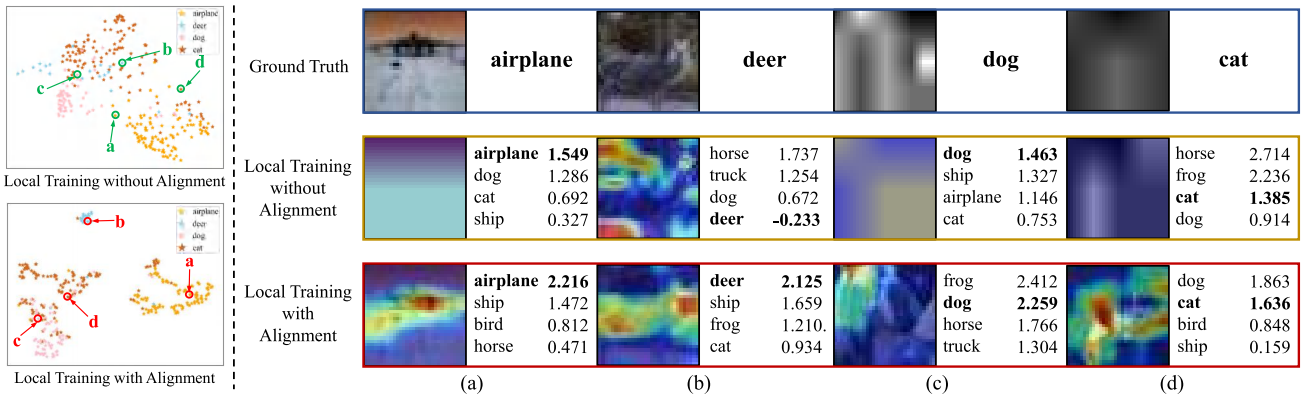

Fig. 11. Comparing the performance of representation learning with and without alignment (left). Error analysis of personalized optimization in pFedCSPC (right). (a) pFedCSPC employs prototype-guided feature alignment to enhance the recognition ability of the correct class. (b) pFedCSPC can correct the error of prediction and improve the performance of feature learning. (c) pFedCSPC failed due to the confusion between the target object and other classes. (d) pFedCSPC reduces the prediction difference between the ground truth and top-1.

TABLE VI

GENERALIZATION PERFORMANCE OF FL METHODS WITH PFEDCSPC ON CIFAR10 AND TINYIMAGENET DATASETS WITH $\beta = 0.5$

Methods	CIFAR10	TinyImagenet
FedProx	67.55±0.6	46.59±0.7
FedProx+pFedCSPC	69.46±0.5	47.68±0.4
MOON	69.04±0.7	47.31±0.9
MOON+pFedCSPC	**71.42±0.4**	48.46±0.5
FedASAM	68.48±0.6	47.38±0.6
FedASAM+pFedCSPC	70.65±0.5	**48.62±0.3**

TABLE VII

EVALUATION OF DIFFERENT METHODS ON CIFAR10 AND TINYIMAGENET WITH DIFFERENT LEVELS OF HETEROGENEITY ($\beta = 0.1$ AND 0.5), IN TERMS OF THE NUMBER OF COMMUNICATION ROUNDS TO REACH TARGET TEST ACC [CIFAR10: ACC = 61.0 ($\beta = 0.1$) AND ACC = 66.0 ($\beta = 0.5$), AND TINYIMAGENET: ACC = 42.0 ($\beta = 0.1$) AND ACC = 45.0 ($\beta = 0.1$)]

Methods	CIFAR10		TinyImagenet	
	acc = 61% (β=0.1)	acc = 66% (β =0.5)	acc = 42% (β =0.1)	acc = 45% (β =0.5)
FedAvg	88	82	89	85
FedProx	77	67	85	71
MOON	54	32	62	48
FedDC	58	34	68	49
FedNTD	76	54	69	42
FedASAM	60	34	72	39
Fedproc	62	36	55	46
FedDecorr	72	41	61	58
CCVR	68	48	74	57
pFedCSPC	**41**	**24**	**44**	**33**

representations among clients. For instance, in the 25th round, pFedCSPC almost eliminates the heterogeneity boundary of the feature distribution before calibration. This improvement in feature alignment may be a factor in the outstanding performance of pFedCSPC in federated classification tasks. In addition, we note that pFedCSPC was already able to calibrate heterogeneous feature distributions in the first round, but the classification Acc remained low. This is due to the poor representation learning of local models in the first round, and the limited effective information provided to the server, resulting in unreliable decision boundaries.

2) Personalized Model Initialization: This section evaluates the impact of model initialization on clients. Specifically, we present the initial performance of different initialization models, test performance after training various initialized models, and the training loss associated with different initializations. And GradCAM [68] is used to demonstrate the model's attention. Figs. 9 and 10 show the results of qualitative and quantitative analyses, respectively. As shown in Fig. 10(a), the initial personalized performance of most personalized models is superior to that of traditional models. As illustrated in Fig. 9, in the personalized initialization tailored to local 1, it is evident that locals 1, 2, 5, and 7 exhibit accurate attention toward the focal objects within the images, and as a result, PMI allocates them with relatively higher weights. Moreover, we observed that PMI may overlook certain beneficial clients. For instance, for the initialization of local 3, it is observed that even though local client 9 accurately focuses on the main subject, it is assigned a relatively smaller weight. Fortunately, PMI is capable of effectively reducing the interference from irrelevant clients, which contributes to enhancing the effectiveness of personalized initialization. Furthermore, personalized initialization models typically achieve better performance after the training, as they attain lower losses during the training process. Significantly, in instances where personalized models exhibit suboptimal performance, the application of localized training serves to mitigate these discrepancies. In conclusion, personalized weight allocation plays a significant role in initialization.

3) Error Analysis of pFedCSPC: This section presents a case study based on the TSNE visualization in Section V-D2 that delves deeper into the workings of pFedCSPC. To this end, GradCAM [68], [81] is employed to generate heatmaps. As shown in Fig. 11(a), both methods achieve accurate predictions for image classes. Meanwhile, pFedCSPC employs prototype-guided feature alignment to attain a more precise focus on image subjects. When the number of samples in the target class is small, the method without feature alignment may fail to capture the main objects and make incorrect predictions. pFedCSPC relies on the guidance of global knowledge to improve the learning of local representations (see red and green b in the left figure), which corrects prediction errors and calibrates feature attention, as illustrated in Fig. 11(b). Fig. 11(c) exemplifies that both methods produce suboptimal representations for low-quality image. This indicates that it is difficult to eliminate the confusion between the target

object and other classes within a few training epochs. Finally, Fig. 11(d) shows the case where both methods make incorrect predictions. Nonetheless, pFedCSPC uses the alignment mechanism to improve the distribution of some representations (see red d in the left figure), which makes the model pay more attention to the image subject and reduces the discrepancy between the "cat" category and the top-1 prediction. These observations demonstrate the effectiveness of global prototype-guided representation learning in personalized federated classification and the importance of fully considering image quality when training.

VI. Conclusion

This article introduces pFedCSPC, an FL approach designed to improve global generalization and local personalization simultaneously. Specifically, pFedCSPC employs adaptive model aggregation to benefit clients. It utilizes global prototypes to guide local representation learning, which can enhance the collaboration between clients. Additionally, pFedCSPC performs data prototypical modeling for aiding in CSPC to address the issue of feature heterogeneity in diverse spaces. The experimental results show that pFedCSPC can not only calibrate the heterogeneous representation distribution, but also promote clients to learn a consistent representation in subsequent rounds, and using this scheme makes pFedCSPC outperform existing methods both in the generalization and personalization tasks.

Despite the performance improvements achieved by pFedCSPC, there are three directions that could be further explored in future work. First, stronger long-tailed representation learning techniques that better learning discriminative features in clients can significantly improve performance [73], [78], [80]. Second, it would be worthwhile to extend the pFedCSPC to more challenging tasks, such as multimodal learning [57], [72] and out-of-distribution generalization [69], [71], [76].

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artif. Intell. Statist.*, 2017, pp. 1273–1282.

[2] L. Lyu, H. Yu, and Q. Yang, "Threats to federated learning: A survey," 2020, *arXiv:2003.02133*.

[3] Z. Liu et al., "Contribution-aware federated learning for smart healthcare," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2022, vol. 36, no. 11, pp. 12396–12404.

[4] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, "Robust and communication-efficient federated learning from non-IID data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 9, pp. 3400–3413, Sep. 2020.

[5] Z. Qi, Y. Wang, Z. Chen, R. Wang, X. Meng, and L. Meng, "Clustering-based curriculum construction for sample-balanced federated learning," in *Proc. 2nd CAAI Int. Conf. Artif. Intell. (CICAI)*, Aug. 2022, pp. 155–166.

[6] Z. Qi, L. Meng, Z. Chen, H. Hu, H. Lin, and X. Meng, "Cross-silo prototypical calibration for federated learning with non-IID data," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 3099–3107.

[7] F. Sattler, T. Korjakow, R. Rischke, and W. Samek, "FedAUX: Leveraging unlabeled auxiliary data in federated learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 9, pp. 5531–5543, Sep. 2023.

[8] X. Liu, Y. Li, Q. Wang, X. Zhang, Y. Shao, and Y. Geng, "Sparse personalized federated learning," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 8, 2023, doi: 10.1109/TNNLS.2023.3250658.

[9] L. Lyu, X. Xu, Q. Wang, and H. Yu, "Collaborative fairness in federated learning," in *Federated Learning: Privacy and Incentive*. Cham, Switzerland: Springer, 2020, pp. 189–204.

[10] Y. Shi, H. Yu, and C. Leung, "Towards fairness-aware federated learning," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Apr. 10, 2023, doi: 10.1109/TNNLS.2023.3263594.

[11] H. Yu et al., "A fairness-aware incentive scheme for federated learning," in *Proc. AAAI/ACM Conf. AI Ethics Soc.*, Feb. 2020, pp. 393–399.

[12] N. Guha, A. Talwalkar, and V. Smith, "One-shot federated learning," 2019, *arXiv:1902.11175*.

[13] D. Li and J. Wang, "FedMD: Heterogenous federated learning via model distillation," 2019, *arXiv:1910.03581*.

[14] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. 3rd Mach. Learn. Syst. Conf.*, 2020, pp. 429–450.

[15] G. Lee, M. Jeong, Y. Shin, S. Bae, and S.-Y. Yun, "Preservation of the global knowledge by not-true distillation in federated learning," in *Proc. NeurIPS*, vol. 35, 2022, pp. 38461–38474.

[16] X. Mu et al., "FedProc: Prototypical contrastive federated learning on non-IID data," *Future Gener. Comput. Syst.*, vol. 143, pp. 93–104, Jun. 2023.

[17] H. Wang, M. Yurochkin, Y. Sun, D. Papailiopoulos, and Y. Khazaeni, "Federated learning with matched averaging," 2020, *arXiv:2002.06440*.

[18] J. Wang, Q. Liu, H. Liang, G. Joshi, and V. H. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," in *Proc. NIPS*, 2020, pp. 7611–7623.

[19] M. Luo, F. Chen, D. Hu, Y. Zhang, J. Liang, and J. Feng, "No fear of heterogeneity: Classifier calibration for federated learning with non-IID data," in *Proc. NIPS*, 2021, pp. 5972–5984.

[20] A. Z. Tan, H. Yu, L. Cui, and Q. Yang, "Towards personalized federated learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 12, pp. 9587–9603, Mar. 2022.

[21] Y. Mansour, M. Mohri, J. Ro, and A. T. Suresh, "Three approaches for personalization with applications to federated learning," 2020, *arXiv:2002.10619*.

[22] Y. Huang et al., "Personalized cross-silo federated learning on non-IID data," in *Proc. AAAI*, vol. 35, no. 9, May 2021, pp. 7865–7873.

[23] X.-C. Li, D.-C. Zhan, Y. Shao, B. Li, and S. Song, "FedPHP: Federated personalization with inherited private models," in *Proc. Eur. Conf. ECML PKDD*, Sep. 2021, pp. 587–602.

[24] M. Zhang, K. Sapra, S. Fidler, S. Yeung, and J. M. Alvarez, "Personalized federated learning with first order model optimization," 2020, *arXiv:2012.08565*.

[25] J. Luo and S. Wu, "Adapt to adaptation: Learning personalization for cross-silo federated learning," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Jul. 2022, pp. 2166–2173.

[26] J. Zhang et al., "FedALA: Adaptive local aggregation for personalized federated learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, 2023, pp. 11237–11244.

[27] L. Collins, H. Hassani, A. Mokhtari, and S. Shakkottai, "Exploiting shared representations for personalized federated learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jul. 2021, pp. 2089–2099.

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

[29] J. Xu, Y. Yan, and S.-L. Huang, "FedPer++: Toward improved personalized federated learning on heterogeneous and imbalanced data," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2022, pp. 01–08.

[30] L. Gao, H. Fu, L. Li, Y. Chen, M. Xu, and C.-Z. Xu, "FedDC: Federated learning with non-IID data via local drift decoupling and correction," in *Proc.*, Jun. 2022, pp. 10112–10121.

[31] Q. Li, B. He, and D. Song, "Model-contrastive federated learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10713–10722.

[32] L. Zhang, Y. Luo, Y. Bai, B. Du, and L.-Y. Duan, "Federated learning for non-IID data via unified feature learning and optimization objective alignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4400–4408.

[33] T. Zhou, J. Zhang, D. H. K. Tsang, and F. A. Fed, "Federated learning with feature anchors to align features and classifiers for heterogeneous data," *IEEE Trans. Mobile Comput.*, vol. 23, no. 6, pp. 6731–6742, Jun. 2023.

[34] Z. Qi et al., "Cross-training with multi-view knowledge fusion for heterogenous federated learning," 2024, *arXiv:2405.20046*.

[35] Z. Qi, W. He, X. Meng, and L. Meng, "Attentive modeling and distillation for out-of-distribution generalization of federated learning," in *Proc. ICME*, 2024, pp. 648–653.

[36] T. Liu, Z. Qi, Z. Chen, X. Meng, and L. Meng, "Cross-training with prototypical distillation for improving the generalization of federated learning," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2023, pp. 648–653.

[37] S. Han et al., "FedX: Unsupervised federated learning with cross knowledge distillation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 691–707.

[38] Y.-Y. Xu, C.-S. Lin, and Y.-C.-F. Wang, "Bias-eliminating augmentation learning for debiased federated learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 20442–20452.

[39] Z. Li, X. Shang, R. He, T. Lin, and C. Wu, "No fear of classifier biases: Neural collapse inspired federated learning with synthetic and fixed classifier," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 5319–5329.

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

[41] D. Chen, J. Hu, V. J. Tan, X. Wei, and E. Wu, "Elastic aggregation for federated optimization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 12187–12197.

[42] L. Zhang, L. Shen, L. Ding, D. Tao, and L.-Y. Duan, "Fine-tuning global model via data-free knowledge distillation for non-IID federated learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, May 2022, pp. 10174–10183.

[43] X. Shang, Y. Lu, G. Huang, and H. Wang, "Federated learning on heterogeneous and long-tailed data via classifier re-training with federated features," 2022, *arXiv:2204.13399*.

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

[45] W. Hao et al., "Towards fair federated learning with zero-shot data augmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 3310–3319.

[46] H. Wen, Y. Wu, J. Li, and H. Duan, "Communication-efficient federated data augmentation on non-IID data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 3377–3386.

[47] M. Duan, D. Liu, X. Chen, R. Liu, Y. Tan, and L. Liang, "Self-balancing federated learning with global imbalanced data in mobile systems," *IEEE Trans. Parallel Distrib. Syst.*, vol. 32, no. 1, pp. 59–71, Jan. 2020.

[48] H. Wang, Z. Kaplan, D. Niu, and B. Li, "Optimizing federated learning on non-IID data with reinforcement learning," in *Proc. IEEE Conf. Comput. Commun.*, Jul. 2020, pp. 1698–1707.

[49] Y. J. Cho, J. Wang, and G. Joshi, "Towards understanding biased client selection in federated learning," in *Proc. 25th Int. Conf. Artif. Intell. Statist. (AISTATS)*, Mar. 2022, pp. 10351–10375.

[50] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach," in *Proc. NeurIPS Conf.*, vol. 33, Dec. 2020, pp. 3557–3568.

[51] C. T. Dinh, N. H. Tran, and T. D. Nguyen, "Personalized federated learning with Moreau envelopes," in *Proc. NIPS*, Dec. 2020, pp. 21394–21405.

[52] L. Meng, A.-H. Tan, and D. Xu, "Semi-supervised heterogeneous fusion for multimedia data co-clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 9, pp. 2293–2306, Sep. 2014.

[53] L. Meng, A.-H. Tan, and C. Miao, "Salience-aware adaptive resonance theory for large-scale sparse data clustering," *Neural Netw.*, vol. 120, pp. 143–157, Dec. 2019.

[54] Z. Qu, X. Li, X. Han, R. Duan, C. Shen, and L. Chen, "How to prevent the poor performance clients for personalized federated learning?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 12167–12176.

[55] P. Pu Liang et al., "Think locally, act globally: Federated learning with local and global representations," 2020, *arXiv:2001.01523*.

[56] Y. Shen, Y. Zhou, and L. Yu, "CD2-pFed: Cyclic distillation-guided channel decoupling for model personalization in federated learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10031–10040.

[57] L. Meng et al., "Learning using privileged information for food recognition," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 557–565.

[58] Y. Tan et al., "FedProto: Federated prototype learning across heterogeneous clients," in *Proc. 36th AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 8, pp. 8432–8440.

[59] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Univ. Toronto, Toronto, ON, Canada, Tech. Rep. TR-2009, 2009.

[60] Y. Le and X. Yang, "Tiny ImageNet visual recognition challenge," *CS 231N*, vol. 7, no. 7, p. 3, 2015.

[61] J. Chen and C.-W. Ngo, "Deep-based ingredient recognition for cooking recipe retrieval," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 32–41.

[62] D. Caldarola, B. Caputo, and M. Ciccone, "Improving generalization in federated learning by seeking flat minima," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 654–672.

[63] Y. Shi, J. Liang, W. Zhang, V. Y. Tan, and S. Bai, "Towards understanding and mitigating dimensional collapse in heterogeneous federated learning," in *Proc. ICLR*, 2023.

[64] M. Chen, M. Jiang, Q. Dou, Z. Wang, and X. Li, "FedSoup: Improving generalization and personalization in federated learning via selective model interpolation," in *Proc. MICCAI*, 2023, pp. 318–328.

[65] R. Zhang, Z. Fan, Q. Xu, J. Yao, Y. Zhang, and Y. Wang, "Grace: A generalized and personalized federated learning method for medical imaging," in *Proc. MICCAI*, 2023, pp. 14–24.

[66] T. Li, S. Hu, A. Beirami, and V. Smith, "Ditto: Fair and robust federated learning through personalization," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 6357–6368.

[67] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

[68] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[69] Y. Wang et al., "Meta-causal feature learning for out-of-distribution generalization," in *Proc. ECCV*, 2022, pp. 530–545.

[70] J. Liu et al., "Prompt learning with cross-modal feature alignment for visual domain adaptation," in *Proc. 2nd CAAI Int. Conf. Artif. Intell. (CICAI)*, Aug. 2022, pp. 416–428.

[71] Y. Wang, X. Li, H. Ma, Z. Qi, X. Meng, and L. Meng, "Causal inference with sample balancing for out-of-distribution detection in visual classification," in *Proc. 2nd CAAI Int. Conf. Artif. Intell. (CICAI)*, Aug. 2022, pp. 572–583.

[72] Y. Wang, Z. Qi, X. Li, J. Liu, X. Meng, and L. Meng, "Multi-channel attentive weighting of visual frames for multimodal video classification," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jun. 2023, pp. 1–8.

[73] X. Li, Y. Zheng, H. Ma, Z. Qi, X. Meng, and L. Meng, "Cross-modal learning using privileged information for long-tailed image classification," *Comput. Vis. Media*, pp. 1–12, Jun. 2024.

[74] Z. Chen, Z. Qi, X. Li, Y. Wang, L. Meng, and X. Meng, "Class-aware convolution and attentive aggregation for image classification," in *Proc. ACM Multimedia Asia*, Dec. 2023, pp. 1–7.

[75] Z. Chen, Z. Qi, X. Cao, X. Li, X. Meng, and L. Meng, "Class-level structural relation modeling and smoothing for visual representation learning," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 2964–2972.

[76] Y. Wang, X. Li, Y. Liu, X. Cao, X. Meng, and L. Meng, "Causal inference for out-of-distribution recognition via sample balancing," *CAAI Trans. Intell. Technol.*, pp. 1–13, Apr. 2024.

[77] Q. Guan, Y. Zheng, L. Meng, L. Dong, and Q. Hao, "Improving the generalization of visual classification models across IoT cameras via cross-modal inference and fusion," *IEEE Internet Things J.*, vol. 10, no. 18, pp. 15835–15846, Sep. 2023.

[78] X. Li, H. Ma, L. Meng, and X. Meng, "Comparative study of adversarial training methods for long-tailed classification," in *Proc. 1st Int. Workshop Adversarial Learn. Multimedia (ADVM)*, 2021, pp. 1–7.

[79] H. Ma et al., "Triple sequence learning for cross-domain recommendation," *ACM Trans. Inf. Syst.*, vol. 42, no. 4, pp. 1–29, Jul. 2024.

[80] C. Lin, S. Zhao, L. Meng, and T.-S. Chua, "Multi-source domain adaptation for visual sentiment classification," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 3, pp. 2661–2668.

[81] J. Li, H. Ma, X. Li, Z. Qi, L. Meng, and X. Meng, "Unsupervised contrastive masking for visual haze classification," in *Proc. ICMR*, Jun. 2022, pp. 426–434.