Cross-Silo Prototypical Calibration for Federated Learning with Non-IID Data

Zhuang Qi
Shandong University
z_qi@mail.sdu.edu.cn

Lei Meng*
[1]Shandong University
[2]Shandong Research Institute of Industrial Technology
lmeng@sdu.edu.cn

Zitan Chen
Shandong University
chenzt@mail.sdu.edu.cn

Han Hu
Beijing Institute of Technology
hhu@bit.edu.cn

Hui Lin
National Engineering Research Centerfor Risk Perception and Prevention
linhui@whu.edu.cn

Xiangxu Meng
Shandong University
mxx@sdu.edu.cn

ABSTRACT

Federated Learning aims to learn a global model on the server side that generalizes to all clients in a privacy-preserving manner, by leveraging the local models from different clients. Existing solutions focus on either regularizing the objective functions among clients or improving the aggregation mechanism for the improved model generalization capability. However, their performance is typically limited by the dataset biases, such as the heterogeneous data distributions and the missing classes. To address this issue, this paper presents a cross-silo prototypical calibration method (FedCSPC), which takes additional prototype information from the clients to learn a unified feature space on the server side. Specifically, FedCSPC first employs the Data Prototypical Modeling (DPM) module to learn data patterns via clustering to aid calibration. Subsequently, the cross-silo prototypical calibration (CSPC) module develops an augmented contrastive learning method to improve the robustness of the calibration, which can effectively project cross-source features into a consistent space while maintaining clear decision boundaries. Moreover, the CSPC module's ease of implementation and plug-and-play characteristics make it even more remarkable. Experiments were conducted on four datasets in terms of performance comparison, ablation study, in-depth analysis and case study, and the results verified that FedCSPC is capable of learning the consistent features across different data sources of the same class under the guidance of calibrated model, which leads to better performance than the state-of-the-art methods. The source codes have been released at https://github.com/qizhuang-qz/FedCSPC.

CCS CONCEPTS

• **Computing methodologies → Distributed algorithms**.

*Corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612481

KEYWORDS

federated learning, data heterogeneity, prototypical calibration

ACM Reference Format:
Zhuang Qi, Lei Meng, Zitan Chen, Han Hu, Hui Lin, and Xiangxu Meng. 2023. Cross-Silo Prototypical Calibration for Federated Learning with Non-IID Data. In *Proceedings of the 31st International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3581783.3612481

1 INTRODUCTION

Federated learning has gained significant attention for addressing data silos in scenarios where data sources are dispersed and difficult to share. It enables multiple parties to collaboratively train a model without sharing their data and aims to aggregate the local models obtained from the parties to generate a global model with generalization capability [26, 32, 37]. However, the vulnerability of the federated model when confronted with heterogeneous data distribution patterns across clients has been highlighted in recent research [6, 38, 44]. This is mainly due to the bias in optimization objectives among the data sources, which makes it difficult to aggregate multiple ill-posed learners into an excellent model.

To mitigate the challenge of heterogeneous data distribution, three main approaches have been developed: data sharing, mitigating the local drift on the client side and optimizing the aggregation scheme on the server. The first method involves the use of public or synthetic datasets to create balanced data distributions, which can be beneficial in guiding clients to build unbiased models [9, 17]. The second approach typically utilizes global information as a regularizer to guide the learning process of each client, with the purpose of promoting model output consistency among clients [7, 11, 16, 18, 19, 56]. And these methods can also be divided into three subcategories: parameter-based [7, 19], feature-based [18, 56], and prediction-based [11, 16]. The third method considers that directly averaging parameters of local models will lead to a performance decline. They either design novel strategies to enhance the aggregation phase (such as FedMA [46], FedNova [47].) or retrain the global classifier using virtual representations (CCVR [27]). However, the heterogeneity of data distribution across sources results in inconsistent feature spaces, which leads to difficulties in training a model to fit data from all clients.

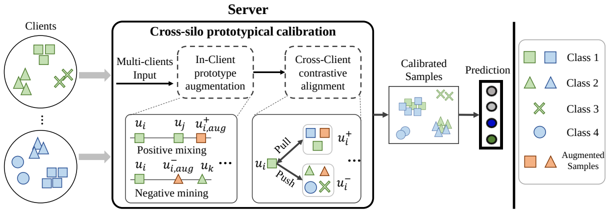


Figure 1: Motivation of FedCSPC. It calibrates the representation space of heterogeneous clients on the server side, which improves the generalization capability of the global model.

To address this problem, this paper presents a novel Cross-Silo Prototypical Calibration method, termed FedCSPC. As illustrated in Figure 1, compared with conventional federated learning method, the proposed FedCSPC performs prototypical calibration, which can map representations from different feature spaces to a unified space while maintaining clear decision boundaries. Specifically, FedCSPC has two main modules: the Data Prototypical Modeling (DPM) module and the Cross-Silo Prototypical Calibration (CSPC) module. To promote the alignment of features across different spaces, the DPM module employs clustering to model the data patterns and provides prototypical information to the server to assist with model calibration. Subsequently, to enhance the robustness of calibration, FedCSPC develops an augmented contrastive learning method in the CSPC module, which increases sample diversity by positive mixing and hard negative mining, and implements contrastive learning to achieve effective alignment of cross-source features. Meanwhile, the calibrated prototypes form a knowledge base in a unified space and generate knowledge-based class predictions to reduce errors. Notably, the CSPC module is a highly adaptable tool that easily integrates into various algorithms. As observed, FedCSPC is capable of alleviating the feature gap between data sources, thus significantly improving the generalization ability.

Experiments are conducted on four datasets in terms of performance comparison, ablation study of the key components of FedCSPC, in-depth analysis and case study for the effectiveness of cross-source calibration and error analysis of FedCSPC. The results verify that FedCSPC can calibrate heterogeneous representations from different sources into a unified space via CSPC, which can mitigate the negative impact of data heterogeneity. Moreover, the error analysis reveals the potential sources of error in FedCSPC and provides insights for future improvements.

To summarize, this paper includes three main contributions:

- A novel cross-silo prototypical calibration method is proposed to alleviate the problem of data distribution heterogeneity among different clients. To the best of our knowledge, this is the first method that can map heterogeneous features from different sources to a unified space.
- The proposed CSPC module is an orthogonal improvement to client-based methods. Its plug-and-play design makes it easy to integrate into existing infrastructure, and it enhances the generalization without altering core components.
- This study reveals the fact that the inconsistent feature spaces across clients pose a challenge for the federated model to fit all clients effectively. And we have verified that FedCSPC can effectively solve this problem.

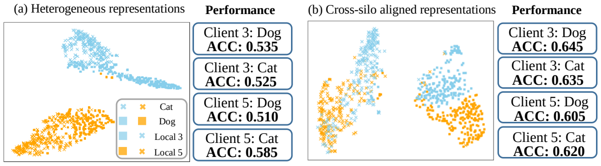


Figure 2: Feature distributions learned by FedAvg and FedCSPC. FedCSPC effectively generalizes to client-side samples by learning to calibrate client-side prototypes.

2 RELATED WORK

To address the issue of data heterogeneity, existing methods typically follow three main approaches: the first approach aims to alleviate the difference between local and global objectives during the local training phase, the second method focuses on optimizing the model aggregation scheme on the central server, and the third approach stems from the data sharing.

2.1 Models mitigating client drift

Common strategies in the local training phase involve utilizing global information as knowledge to regularize local updates. Conventional approaches along the line of research include weights-based [7, 19, 42], feature-based [18, 56], and prediction-based [11, 16] constraints. Weights-based methods either design proximal terms to constrain the consistency of the local and global models or use a drift factor to track the gap between the global and local models in the parameter space. Feature-based methods focus on feature contrast to penalize inconsistency. They typically align local and global output in latent space or use prototypes to restrict clients from learning similar representations. Nevertheless, it has been observed that there exists feature maps inconsistency in these works, which leads a limited performance (See Figure 2). Prediction-based approaches usually rely on an auxiliary dataset, and they integrate local soft-label predictions on the auxiliary dataset rather than model parameters or gradients, which reduces communication costs and achieves knowledge distillation.

2.2 Models optimizing aggregation scheme

To improve the performance of the federated model, many studies focus on optimizing the aggregation mechanism on the server side. For instance, FedMA uses a Bayesian non-parametric method to match neurons rather than naively averaging [46], FedAvgM applies the momentum rule to update the global model, which can improve robustness to heterogeneous distributed data [13], and FedNova eliminates inconsistencies by normalizing local updates before averaging them [47]. In addition, re-training or fine-tuning schemes are also applied to mitigate the model shift after aggregation, such as FedFTG uses an auxiliary generator to generate pseudo data for retraining, which can model the input space of local models [57]. CCVR [27] and CReFF [40] illustrate that the heterogeneity of the classifier is the main reason for the performance degradation of models trained on non-IID data. Therefore, they retrain the classifier by using the virtual feature generated by the gaussian mixture model and the federated feature with a consistency gradient to the real data, respectively.

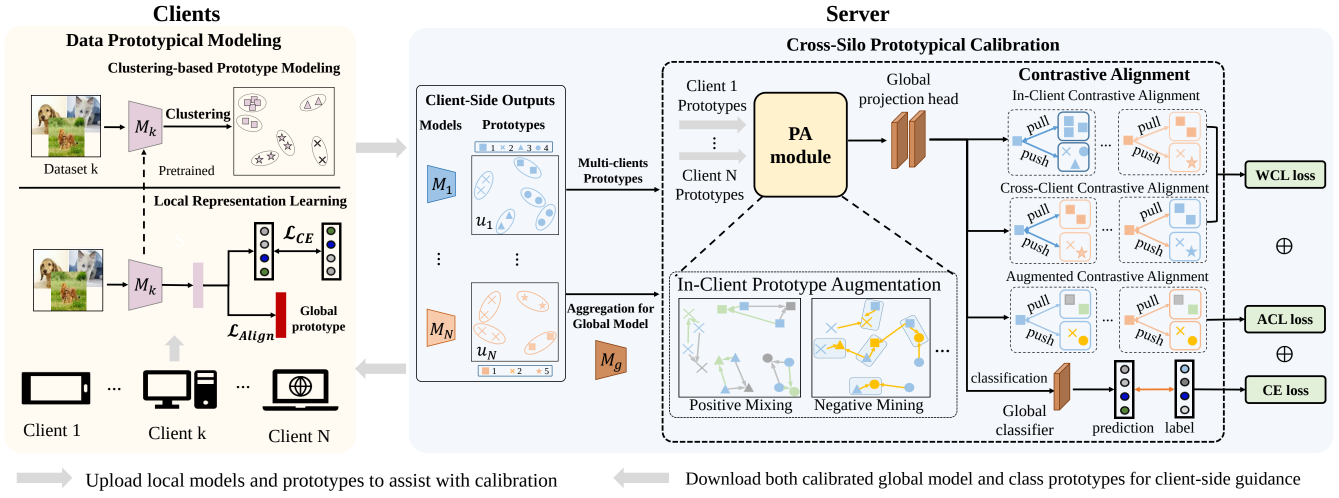

Clients
Data Prototypical Modeling
Clustering-based Prototype Modeling
Clustering
M_k
Dataset k
Pretrained
Local Representation Learning
$\mathcal{L}_{CE}$
$\mathcal{L}_{Align}$
Global prototype
Client 1
Client k
Client N
Server
Cross-Silo Prototypical Calibration
Client-Side Outputs
Models
Prototypes
M_1
u_1
M_N
u_N
Multi-clients Prototypes
Aggregation for Global Model
M_g
Client 1 Prototypes
Client N Prototypes
PA module
Global projection head
Contrastive Alignment
In-Client Contrastive Alignment
Cross-Client Contrastive Alignment
Augmented Contrastive Alignment
pull
push
WCL loss
ACL loss
CE loss
In-Client Prototype Augmentation
Positive Mixing
Negative Mining
classification
Global classifier
prediction
label
Upload local models and prototypes to assist with calibration
Download both calibrated global model and class prototypes for client-side guidance

Figure 3: Illustration of the framework of FedCSPC. FedCSPC offers flexibility in the choice of algorithms for optimizing the local model on the client-side, including FedAvg and MOON, among others. 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. Meanwhile, it uses prototype augmentation to improve the robustness of the calibration.

2.3 Models Trained with Auxiliary Data

Due to the heterogeneity of the data distribution across different sources, the local models trained on the client side may have insufficient generalization ability for certain patterns or samples from absent classes. Therefore, existing studies propose ideas for sharing data. They typically share public datasets [17], synthesized datasets [12], and truncated versions of private data [9]. However, these approaches may violate privacy preservation rules since they expose the raw data to other parties.

3 PROBLEM FORMULATION

In federated learning, there are N clients $C = \{C_1, C_2, ..., C_N\}$ and a sever S. The client C_k holds 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 $w_k = w_k^E \oplus w_k^H \oplus w_k^F$, where E_k is an image encoder with parameters w_k^E, H_k denotes projection head with parameters w_k^H and F_k is a classifier with parameters w_k^F. The goal of federated learning methods is to jointly train a global model with the assistance of a server S without leaking privacy and minimize the following problem:

$$\min L(w) = \min \sum_{C_k \in C} p_k L_k(w; D_k), \tag{1}$$

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

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

The process is repeated for T rounds and the resulting M_g with the parameter w_g represents the final aggregated model.

In contrast, the proposed FedCSPC introduces a **Cross-Silo Prototypical Calibration** (CSPC) module on the server, which aims to relearn the global projection head $H_g \mapsto \hat{H}_g$ the classifier $F_g \mapsto \hat{F}_g$ to align representations from different feature spaces, i.e. $\hat{H}_g(E_i(x_i)) \approx \hat{H}_g(E_j(x_j))$, where x_i and x_j are the samples with the same label in the client C_i and C_j, respectively. FedCSPC first generates class-aware prototypes in all clients, $\mathcal{U} = \{\mathcal{U}_k | k \in C\}$ and $\mathcal{U}_k = \{u_k^i | i \in \mathcal{Y}_k\}$ for each class on the client, and sends them to the server. Subsequently, the CSPC module learns the mapping $\hat{H}_g(\cdot)$ based on these prototypes and the corresponding augmented samples $\mathcal{U}_{aug}$ to gather together cross-source features shared the same label. Finally, calibrated prototypes form a knowledge base to produce knowledge-based prediction $Pred_k$. The final prediction $Pred_{final}$ is achieved by $Pred_k \oplus Pred_{net} \mapsto Pred_{final}$, where $Pred_{net}$ is the prediction of network.

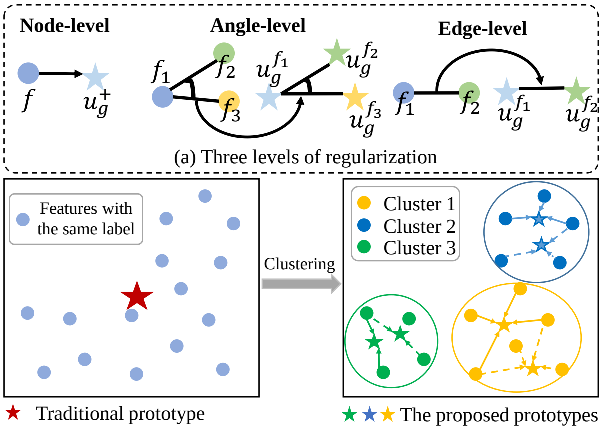
4 APPROACH

4.1 Overall framework

FedCSPC introduces a cross-silo prototypical calibration method to enhance the generalization capability of the global model. Figure 3 illustrates the main framework of FedCSPC. It first designs a novel Data Prototypical Modeling (DPM) module, which is used to model the representation distribution on the clients and provide prototypical information to the server. Afterward, the Cross-Source Prototypical Calibration (CSPC) module obtains prototypical representations from all clients and learns the mapping from the dispersed space to a unified space based on these prototypes to eliminate feature heterogeneity in the heterogeneous space. This enables the global model to generalize to all clients. Meanwhile, the calibrated prototypes form a knowledge base to aid decision-making.

4.2 Data Prototypical Modeling (DPM) module

The Data Prototypical Modeling (DPM) module aims to provide the prototypical information regarding representations to the server, which aids in model calibration. It has two main process: Strengthening local representation learning to alleviate calibration pressure and Modeling prototypical representations for data via clustering.



(a) Three levels of regularization

(b) Clustering-based prototype learning

Figure 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.

4.2.1 ***Local Representation Learning*** *(LRL).* It is understandable that calibration becoming more difficult as the heterogeneity of features among clients increases. To alleviate this problem, the DPM module employs global prototypes $\mathcal{U}_g$ to guide all clients learn similar representations as much as possible within fewer training epochs. It uses three levels of constraints to regularize representation learning, including node, angle, and edge levels, as shown in Figure 4(a). For the point level regularization, we define an prototype-based contrastive loss to optimize the distance between the local representation and the corresponding global prototype,

$$\mathcal{L}_N = -\log \frac{\exp(f \cdot u_g^+/\tau_l)}{\exp(f \cdot u_g^+/\tau_l) + \sum \exp(f \cdot u_g^-/\tau_l)} \tag{3}$$

where f denotes a local representation, u_g^+, 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 the section 4.3. τ_l is a temperature parameter. For the angle level, given three representations f_1, f_2, f_3 with different labels, the corresponding prototypes are $u_g^{f_1}$, $u_g^{f_2}$, $u_g^{f_3}$, and the angle-based alignment loss is defined as

$$\mathcal{L}_A = \left\| (\cos \angle(f1, f2, f3), \cos \angle(u_g^{f_1}, u_g^{f_2}, u_g^{f_3})) \right\|_1 \tag{4}$$

where $\cos \angle(f1, f2, f3) = \left\langle \frac{f_1 - f_2}{\|f_1 - f_2\|_2}, \frac{f_3 - f_2}{\|f_3 - f_2\|_2} \right\rangle$, $\langle \cdot \rangle$ denotes the inner product. For the edge level, it requires the distance between the samples to be consistent with the corresponding prototypes,

$$\mathcal{L}_E = \ell(\|f_1 - f_2\|_2 - \left\| u_g^{f1} - u_g^{f2} \right\|_2) \tag{5}$$

where $\ell(\cdot)$ is the L_2 norm.

4.2.2 ***Clustering-based Prototype Modeling*** *(CPM).* After representation learning, to capture the homogeneity and diversity of representations in each class, the DPM module expands the K-means clustering approach to investigate patterns in representation distributions. Specifically, the procedure of mining different patterns of class j can be formulated as:

$$c_j^1, c_j^2, ..., c_j^k = \textbf{K-means}(E(x), k), x \in D_j \tag{6}$$

where k is the number of clusters. c_j^n denotes the n-th cluster of class j, $E(\cdot)$ is an image encoder, D_j denotes data of class j.

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 Figure 4(b). Compared with using a single prototype to represent the entire class distribution [26, 35], the DPM module is capable of increasing the diversity of prototypes and providing sufficient and effective information for calibration. The calculation of the prototype can be formulated as:

$$u_j^{i,t} = \textbf{mean}\{f \,|f \in \textbf{sampling}(c_j^i, r)\} \tag{7}$$

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

4.3 Cross-Silo Prototypical Calibration (CSPC) module

The CSPC module obtains all local models and local prototype set $\{\mathcal{M}, \mathcal{U}\} = \{(M_k, \mathcal{U}_k) | k = 1, ..., N\}$ from clients. It has been found that regularization of client representation learning cannot completely eliminate heterogeneity. Therefore, the CSPC module is designed to align these prototypical features from heterogeneous spaces, i.e. $\hat{H}_g(\mathcal{U}_i) \approx \hat{H}_g(\mathcal{U}_j)$. Another challenge of learning the generalization mapping $\hat{H}_g(\cdot)$ to align features of heterogeneous spaces is the insufficient amount of data. Therefore, the FedCSPC develops an augmented contrastive learning method in the CSPC module. It has two main process: In-client prototype augmentation for information supplementation and Cross-client contrastive alignment for mitigating heterogeneity.

4.3.1 ***In-Client Prototype Augmentation*** *(PA).* To augment the local prototype set, two strategies are used to generate new sample features, i.e. positive mixing and negative mining. Specifically, we use extrapolation between prototypes of the same class and interpolation between prototypes of different classes to generate positive samples and mine hard negative samples, respectively, i.e.,

$$u_i^+ = (u_j - u_i) \times \lambda_u + u_j, \quad u_i^- = (u_k - u_i) \times \lambda_u + u_i \tag{8}$$

where u_i and u_j share the same label, whereas u_k has a distinct label. λ_u is a constant coefficient. Notably, intra-class extrapolation preserves the core features while increasing diversity. Meanwhile, inter-class interpolation injects positive information into negative samples, making it more difficult for the model to distinguish the decision boundary, which is advantageous for improving the generalization ability of the model.

4.3.2 ***Cross-Client Contrastive Alignment*** *(CA).* For the raw global model $M_g = E_g \odot H_g \odot F_g$, obtained by Eq. (2), the projection head $H_g(\cdot)$ and $F_g(\cdot)$ need to be calibrated, i.e. $H_g(\cdot) \mapsto \hat{H}_g(\cdot)$, $F_g(\cdot) \mapsto \hat{F}_g(\cdot)$. To enhance the robustness of calibration, augmented samples are used as additional constraints,

$$\mathcal{L}_{ACL}(u_i, u_i^+, u_i^-) = ||H_g(u_i) - H_g(u_i^+)||_2^2 - ||H_g(u_i) - H_g(u_i^-)||_2^2 + \alpha \tag{9}$$

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

$$\mathcal{L}_{WCL}(u_i) = -\frac{1}{P(u_i)} \sum_{u_j \in P(u_i)} \log \frac{\sigma_j \cdot \exp(z_i^T \cdot z_j/\tau_g)}{\sum\limits_{u_k \in I_s} \sigma_k \cdot \exp(z_i^T \cdot z_k/\tau_g)} \qquad (10)$$

where $P(u_i)$ indicates the positive set of u_i. I_s denotes the sample set. $z_i = H_g(u_i)$, 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 u_i and u_j are samples of the same class from different clients or samples of different classes from the same client, $\sigma_j = 1$; otherwise, $\sigma_j = 0.5$.

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

$$\mathcal{L}_{sup}(u_i) = -\sum\nolimits_{j=1}^{N_c} \mathcal{I}(y_i = c)\log(\hat{y}_{i,j}) \qquad (11)$$

where $\mathcal{I}(\cdot)$ denotes the indication function, N_c represent the number of classes. y_i is the label of u_i, $\hat{y}_{i,j}$ is the prediction that u_i belongs to class j.

In addition, FedCSPC 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 knowledge-based prediction. And the final prediction of the test sample x is obtained by fusing the decisions from both the network $Pred_{net}(x)$ and knowledge base $Pred_k(x)$, i.e.,

$$e^i = \frac{1}{N}\sum\nolimits_{j=1}^{N}\frac{1}{n_{repeat}}\sum\nolimits_{t=1}^{n_{repeat}} \hat{H}(u_j^{i,t}) \qquad (12)$$

$$Pred_{final}(x) = (1-\lambda_p) \times Norm(Pred_{net}(x)) + \lambda_p \times Norm(Pred_k(x)) \quad (13)$$

where $Pred_k(x) = [sim(f_x, e_i)\,|i = 1, ..., N_c\,]$ contains the similarity between the sample feature f_x and all exemplars $\{e_i\,|i = 1, ..., N_c\}$, $sim(\cdot)$ and $Norm(\cdot)$ denote the similarity and normalization function, respectively.

Furthermore, to reduce the heterogeneity of features among clients, the CSPC module generates global prototypes $\mathcal{U}_g = \{u_g^i | i = 1, ..., N_c\}$ to regularize the representation learning of all clients,

$$u_g^i = \frac{1}{N}\sum\nolimits_{j=1}^{N}\frac{1}{n_{repeat}}\sum\nolimits_{t=1}^{n_{repeat}} u_j^{i,t} \qquad (14)$$

Finally, the CSPC module sends the calibrated global model $\hat{M}_g = E_g \odot \hat{H}_g \odot \hat{F}_g$ and the global prototypes $\mathcal{U}_g$ to all clients.

4.4 Training Strategies

FedCSPC focuses on calibrating feature space on the server side, which can be combined with multiple client-based methods. Consequently, FedCSPC has the following training strategies.

- **In the client**, the optimization objective varies depending on the base algorithm being used. Moreover, the alignment loss $\mathcal{L}_{align} = \mathcal{L}_N + \mathcal{L}_A + \mathcal{L}_E$ is used to regularize all clients to learn similar representations, which can alleviate the calibration difficulty due to heterogeneity. Therefore, the overall optimization objective for a client is

$$\mathcal{L}_{client} = \mathcal{L}_{base} + \kappa \times \mathcal{L}_{align} \qquad (15)$$

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

Table 1: 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	25250

- **In the server**, FedCSPC aims to align features in heterogeneous spaces to eliminate heterogeneity and obtain clear decision boundaries, and it optimizes the following objective function:

$$\mathcal{L}_{server} = \frac{1}{|I_s|}\sum_{u_i \in I_s} \mathcal{L}_{\mathrm{sup}}(u_i) + \eta[\mathcal{L}_{WCL}(u_i) + \mathcal{L}_{ACL}(u_i, u_i^+, u_i^-)] \quad (16)$$

where η is a weight parameter.

5 EXPERIMENTS

5.1 Experiment Settings

5.1.1 Datasets. To verify the effectiveness of the algorithms, we used four datasets in the experiment, including CIFAR10 [14], CIFAR100 [14] and TinyImageNet [15] which are commonly used in federated learning. And a challenging food classification dataset VireoFood172 [2]. Their statistical information is shown in Table 1. The Dirichlet distribution is used to partition the dataset.

5.1.2 Evaluation Measures. Following previous studies [18, 32], we use the Top-1 Accuracy to evaluate the performance of methods,

$$\text{Accuracy } = (TP + TN)/(P + N) \qquad (17)$$

where P, N, TP and TN are Positives, Negatives, True Positives and True Negatives, respectively.

5.1.3 Hyper-parameter Settings. Following recent studies [18, 35], 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 SGD optimizer with the learning rate $lr = 0.01$ and the weight decay wd is set to 1e-5. For all datasets, the Dirichlet parameter $\beta = 0.5$ and $\beta = 0.1$. In the DPM module, the number of clusters for each class k is selected from $\{2, 3, 4\}$, the sample proportion $r = 0.5$, the number of sampling $n_{repeat} = 5$, and the temperature parameter $\tau_l = 0.5$. In the CSPC module, the constant coefficient λ_u and λ_p are selected from $\{0.1, 0.3, 0.5\}$, the margin parameter $\alpha = 1.0$, the number of augmented samples for each prototype $n_{aug} = 5$, the temperature parameter $\tau_g = 0.5$. 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 hyper-parameters by referring to corresponding papers for fair comparison and optimal performance.

5.2 Performance Comparison

We compare FedCSPC with nine state-of-the-art methods, including FedAvg [32], FedProx [19], MOON [18], CCVR [27], FedDC [7], FedNTD [16], FedASAM [1], FedProc [35] and FedDecorr [41]. And the network architecture used for all methods comprises an image encoder, a projection head and a classifier. For all datasets, we employ a 2-layer MLP as the projection head and the classifier is a 1-layer fully-connected layer. For the CIFAR10 dataset, we

Table 2: Performance comparison between FedCSPC with baselines on CIFAR10, CIFAR100, TinyImagenet, and VireoFood172 datasets. All algorithms were run by three trials, and the mean and standard derivation are reported.

Methods		CIFAR10		CIFAR100		TinyImagenet		VireoFood172	
		$\beta = 0.1$	$\beta = 0.5$	$\beta = 0.1$	$\beta = 0.5$	$\beta = 0.1$	$\beta = 0.5$	$\beta = 0.1$	$\beta = 0.5$
FL without Calibration	FedAvg (AISTATS'17)	61.18±0.7	66.78±0.4	62.12±0.8	66.54±0.3	42.48±0.8	45.38±0.4	56.78±0.9	59.88±0.8
	FedProx (MLSys'20)	62.85±1.2	67.55±0.6	62.87±0.6	67.13±0.8	42.67±0.6	46.59±0.7	57.52±0.6	60.69±0.9
	MOON (CVPR'21)	63.11±0.8	69.04±0.7	63.45±0.3	67.88±0.4	43.75±0.7	47.31±0.9	58.17±0.5	61.25±1.1
	FedDC (CVPR'22)	63.25±0.7	69.13±0.6	63.76±0.7	67.75±0.6	43.68±0.8	46.81±0.2	58.04±0.4	60.97±0.5
	FedNTD (NeurIPS'22)	62.79±0.9	68.89±0.3	62.97±1.1	67.83±0.2	43.51±1.0	45.79±0.5	57.92±0.7	60.88±0.9
	FedASAM (ECCV'22)	63.16±0.5	68.48±0.6	63.21±0.3	67.71±0.5	43.48±0.6	47.38±0.6	58.12±0.8	61.14±0.2
	Fedproc (FGCS'23)	62.52±1.3	69.18±1.2	63.46±1.3	67.63±0.7	43.75±0.6	47.21±0.4	57.86±0.4	60.46±0.4
	FedDecorr (ICLR'23)	62.38±0.8	68.66±0.8	63.53±0.5	67.79±0.6	43.94±0.3	46.21±0.7	58.01±0.3	61.06±0.7
FL with Calibration	$\text{CCVR}_{\text{FedAvg}}$ (NeurIPS'21)	62.48±0.9	68.56±0.7	63.36±0.7	67.86±0.4	42.48±0.4	46.11±0.4	57.51±0.5	60.87±0.3
	$\text{CCVR}_{\text{MOON}}$ (NeurIPS'21)	63.51±0.6	69.49±0.5	63.89±0.2	67.94±0.3	44.36±0.6	47.89±0.5	58.49±0.7	60.98±0.8
	$\text{CCVR}_{\text{FedASAM}}$ (NeurIPS'21)	63.12±0.8	69.46±0.9	64.18±0.6	68.03±0.5	43.73±0.3	47.94±0.6	58.79±0.5	61.38±0.4
	$\text{FedCSPC}_{\text{FedAvg}}$	64.01±0.7	70.81±0.7	64.19±0.8	68.39±0.4	44.62±0.8	47.89±0.6	59.37±0.4	62.19±0.6
	$\text{FedCSPC}_{\text{MOON}}$	**64.44±0.7**	**71.42±0.4**	64.68±0.3	68.28±0.5	**45.33±0.7**	48.46±0.5	**60.21±0.6**	**62.84±0.5**
	$\text{FedCSPC}_{\text{FedASAM}}$	64.13±0.7	70.65±0.5	**64.81±0.6**	**68.49±0.5**	45.24±0.6	**48.62±0.3**	60.14±0.4	62.61±0.3

employ a convolutional neural network comprising two 5x5 convolutional layers, which are followed by 2x2 max pooling, and two fully connected layers with ReLU function as the image encoder. For other datasets, we use a ResNet18 encoder, excluding its last fully-connected layer. The following can be observed from Table 2.

- **$\text{FedCSPC}_{\text{FedAvg}}$, $\text{FedCSPC}_{\text{MOON}}$, and $\text{FedCSPC}_{\text{FedASAM}}$ have demonstrated substantial improvements in classification compared to their corresponding baselines**, highlighting the model-agnostic character of the FedCSPC.
- **FedCSPC algorithm typically performs better than other algorithms,** which is reasonable because the calibration mechanism of FedCSPC algorithm can effectively alleviate the heterogeneity between features from different sources.
- **Incorporating calibration techniques into the learning process typically yields better results than the baseline method.** This is primarily because the calibration mechanism can assist devices in learning a generalized model from various data sources, such as CCVR and FedCSPC.
- As observed, **the improvement achieved by combining FedCSPC is significant compared to the baseline on CIFAR10, while it is relatively small on other datasets.** This is understandable because the final accuracy depends not only on the degree of bias correction after model calibration but also closely related to the quality of local representation learning.

5.3 Ablation Study

This section further studied the effectiveness of different modules of FedCSPC. We set the sample fraction $C = 0.5$ and $C = 1.0$, the Dirichlet parameter $\beta = 0.5$. The results are summarized in Table 3.

- **Simply combining the traditional prototype generation method (TPG [26]) with the cross-client contrastive alignment (CA) may not bring performance gains**, mainly because a single prototype cannot describe the overall distribution, and an insufficient number of prototypes cannot provide enough information to train a generalizable model.
- **Cross-client contrastive alignment (CA) with the assistance of the clustering-based prototype modeling (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 modeling the representational distribution.

Table 3: Ablation study on the effectiveness of different components of FedCSPC on the CIFAR10 and CIFAR100 datasets.

	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
+LRL+CPM+CA	68.14±0.6	69.51±0.4	67.04±0.1	68.02±0.7
+LRL+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
+LRL+CPM+PA+CA	69.01±0.6	70.42±0.4	67.24±0.1	68.21±0.7
+LRL+CPM+PA+CA+KP	**69.47±0.3**	**70.81±0.4**	**67.36±0.5**	**68.39±0.4**

- In general, **using local representation learning (LRL) and prototype augmentation (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.
- As reported, **knowledge-based prediction (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.

5.4 In-depth Analysis

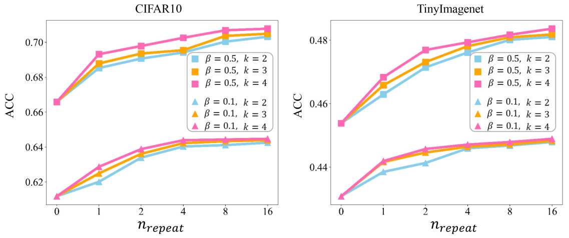


Figure 5: The influence of the number of prototypes (n_{repeat} = 1, 2, 4, 8, 16) on the final performance of FedCSPC on the CIFAR10 and TinyImagenet datasets with different levels of heterogeneity ($\beta = 0.1, 0.5$) and the number of clusters ($k = 2, 3, 4$).

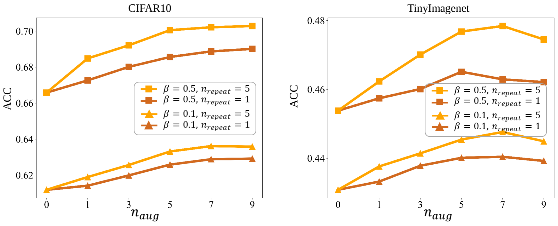


Figure 6: The influence of the number of augmented samples (n_{aug} = 1, 3, 5, 7, 9) on the final performance of FedCSPC on the CIFAR10 and TinyImagenet datasets with different levels of heterogeneity (β = 0.1, 0.5) and the number of prototypes generated per cluster (n_{repeat} = 1, 5).

*5.4.1 **How many prototypes per cluster are enough to assist the server to calibrate a good model?*** The key hyperparameter n_{repeat} is the number of prototypes generated in a cluster. We evaluate its influence by tuning it from $\{0, 1, 2, 4, 8, 16\}$ on CIFAR10 with different heterogeneity $\beta = 0.1$ and $\beta = 0.5$, where $n_{repeat} = 0$ denotes traditional FedAvg. And we tune the number of clusters k from $\{2, 3, 4\}$ for both datasets.

It can be found from Figure 5 that **generating more class-aware prototypes generally leads to higher accuracy**. Moreover, with an increase in the number of prototypes, the improvement stabilizes gradually. This result is understandable, since many prototypes with high similarity are produced, which provide limited information. An impressive result is that even when only one prototype is learned per cluster, FedCSPC can still achieve an average improvement of 1.2% and 2.2% on CIFAR10 when $\beta = 0.1$ and $\beta = 0.5$ respectively. It is worth noting that the performance of FedCSPC approach the upper limit when $n_{repeat} = 4$. This finding would be conducive to mitigating the costs incurred in communication between clients and server. Additionally, **as the number of clusters generated for each class increases, the final performance gradually increases**, since clustering can effectively capture different patterns in the data, which enables the prototypes to exhibit diversity. In conclusion, although FedCSCP can bring performance gain to the baseline, the number of prototypes should be tuned carefully to achieve higher performance.

*5.4.2 **How does the number of augmented samples generated for each sample affect the final performance?*** This section explores the impact of the number of augmented samples n_{aug} on the final results. We tune the n_{aug} from $\{1, 3, 5, 7, 9\}$. We considered two cases, $n_{repeat} = 1$ and $n_{repeat} = 5$.

In general, **the more augmented samples generated, the greater the performance gain for CIFAR10.** This is because the model can learn good local representations on CIFAR10, which enables the augmented positive samples to effectively increase diversity, and the augmented negative samples can help suppress overfitting, thereby enhancing the robustness of calibration. 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 model generalization capacity and results in a decrease in performance. Therefore, we should carefully select the number of augmented samples based on factors such as data complexity to achieve the best performance.

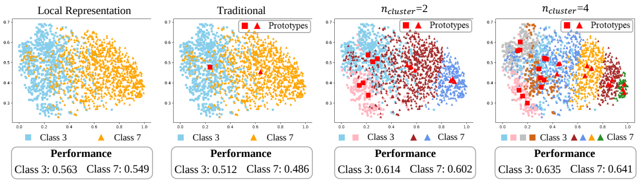
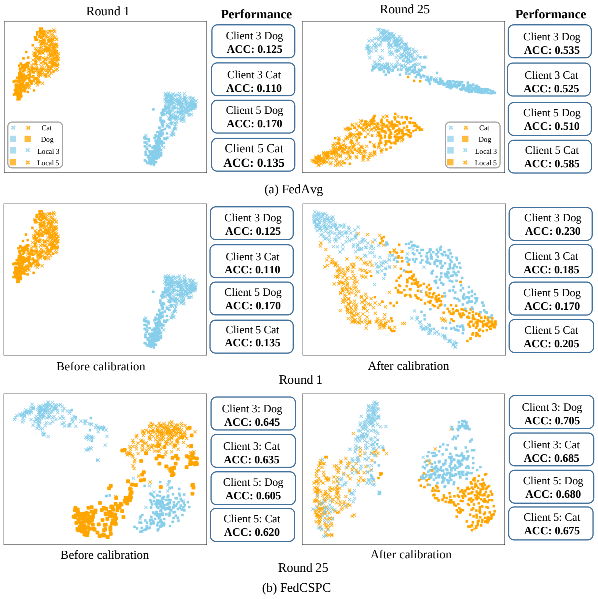


Figure 7: The influence of different clustering results ($n_{cluster}$=2,4 and traditional method) on the representation distribution modeling and global model performance.



(a) FedAvg

(b) FedCSPC

Figure 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) FedCSPC effectively learns the common space of the same class but from different clients, which enables the global model to generalize to different clients.

5.5 Case Study

*5.5.1 **Clustering-based Prototype Generation.*** This section evaluates the influence of prototype generation on the representation distribution modeling and global model performance. We randomly selected the results of a training round and used TSNE [45] to visualize the feature of two classes and their corresponding prototypes for a random user. As shown in Figure 7, **the traditional prototype cannot exhibit the intra-class diversity, while the clustering-based prototype generation method can capture the distribution patterns of representations well.** Meanwhile, we observed that clustering-based prototype generation can capture the overlapping representations between different classes (cornflower blue in the right figure). This significantly increases the diversity within each class and, as hard samples, can improve the robustness of calibration. In addition, **the more clusters generated, the more accurate the modeling of representation distribution will be, which brings more performance gains to the model ($n_{cluster}$ = 4 in the figure).** This is mainly because a larger

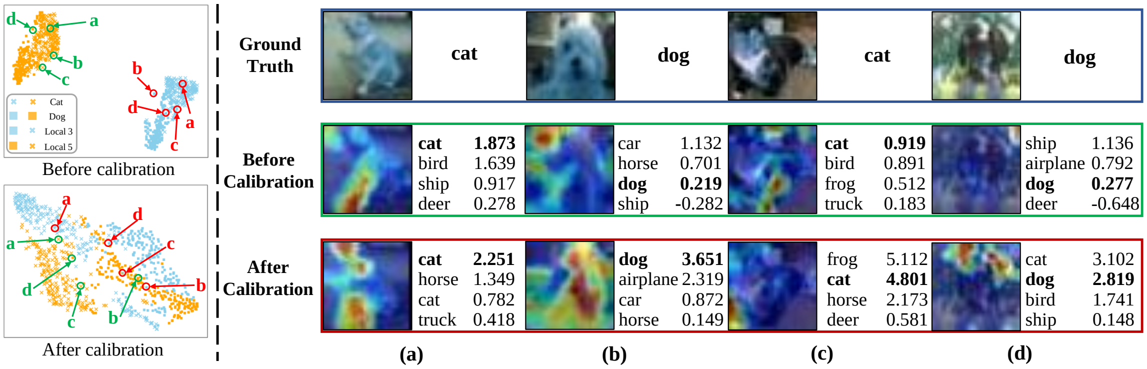


Figure 9: Left: the representation distribution before and after calibration. Right: the error analysis of FedCSPC. (a) FedCSPC employs cross-silo prototypical calibration method to enhance the recognition ability of the correct class (b) FedCSPC can correct the error of prediction and calibrate the feature attention. (c) FedCSPC failed due to poor representation learning. (d) FedCSPC reduces the prediction difference between the ground-truth and top-1.

number of clusters can better capture the details and diversity of sample distribution.

5.5.2 ***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 [45] method was used to visualize the feature distribution of samples before and after calibration. We also output the corresponding classification accuracy of the model before and after calibration. As shown in Figure 8, **FedCSPC 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, **FedCSPC not only corrects the distribution of heterogeneous representations in the current round but also promotes consistency in learning representations among clients.** For instance, in the 25-th round, FedCSPC 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 FedCSPC in federated classification tasks. In addition, we note that FedCSPC was already able to calibrate heterogeneous feature distributions in the first round, but the classification accuracy remained low. This is due to the poor representation learning of local models in the 1-th round, and the limited effective information provided to the server, resulting in unreliable decision boundaries.

5.5.3 ***Error Analysis of FedCSPC.*** This section presents a case study based on the TSNE visualization in Section 5.5.2 that delves deeper into the workings of FedCSPC. To this end, GradCAM [39] is employed to generate heatmaps. As depicted in Figure 9(a), both methods achieve accurate predictions for image classes. Meanwhile, FedCSPC employs calibration to attain a more precise focus on image subjects after calibration. When the target object is highly confused with other classes, the model before calibration may fail to capture the main object and make incorrect predictions. FedCSPC relies on the calibration strategy to align cross-client features (see red and green b in the left figure), which corrects prediction errors and calibrates feature attention, as illustrated in Figure 9(b). Figure 9(c) exemplifies a scenario where the model produced accurate classifications before calibration, but suboptimal representation learning hindered subsequent calibration performance, leading to an inaccurate prediction by FedCSPC (see red c in the left figure). Finally, Figure 9(d) shows the case where the model makes incorrect predictions both before and after calibration. Nonetheless, the calibration method improves the feature distribution (see red d in the left figure), which makes the model pay more attention to the dog region, reducing the discrepancy between the "dog" and the top-1 prediction. These observations not only demonstrate the effectiveness of calibration mechanisms in federated classification but also emphasize the importance of local representation learning.

6 CONCLUSION

This paper presents a novel cross-silo prototypical calibration mechanism, termed FedCSPC, to handle the heterogeneity of feature space across clients. FedCSPC first employs the DPM module to mine the pattern of sample features and provide prototypical information for the server. Subsequently, the CSPC module aligns the features in the dispersed space to the unified space, and adopts prototype augmentation to improve the robustness of the alignment. Experimental results show that FedCSPC can not only calibrate the heterogeneous representation distribution of the current round, but also promote clients to learn a consistent representation in subsequent rounds and using this scheme makes FedCSPC outperform existing methods in federated classification.

Despite the significant performance improvements achieved by FedCSPC, there are two directions that could be further explored in future work. First, stronger representation learning techniques that better learning discriminative features in clients can significantly improve performance [20, 24, 36, 48, 49]. Second, it would be worthwhile to extend the FedCSPC to more challenging tasks, such as multimodal learning [3, 8, 10, 23, 25, 33, 34, 50, 52–55], recommendation [28–31] and some generative tasks [4, 5, 21, 22, 43, 49, 51].

ACKNOWLEDGMENTS

This work is supported in part by the National Key R&D Program of China (Grant no. 2021YFC3300203), the National Natural Science Foundation of China (Grant no. 62006141), the Oversea Innovation Team Project of the "20 Regulations for New Universities" funding program of Jinan (Grant no. 2021GXRC073), the Excellent Youth Scholars Program of Shandong Province (Grant no. 2022HWYQ-048), and the TaiShan Scholars Program (Grant no. tsqn202211289).

REFERENCES

[1] Debora Caldarola, Barbara Caputo, and Marco Ciccone. 2022. Improving generalization in federated learning by seeking flat minima. In *ECCV*. Springer, 654–672.

[2] Jingjing Chen and Chong-Wah Ngo. 2016. Deep-based ingredient recognition for cooking recipe retrieval. In *MM*. 32–41.

[3] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. 2021. Dual Encoding for Video Retrieval by Text. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[4] Pei Dong, Lei Wu, Lei Meng, and Xiangxu Meng. 2022. Disentangled Representations and Hierarchical Refinement of Multi-Granularity Features for Text-to-Image Synthesis. In *ICMR*. 268–276.

[5] Pei Dong, Lei Wu, Lei Meng, and Xiangxu Meng. 2022. HR-PrGAN: High-resolution story visualization with progressive generative adversarial networks. *Information Sciences* (2022).

[6] Dashan Gao, Xin Yao, and Qiang Yang. 2022. A Survey on Heterogeneous Federated Learning. *preprint arXiv:2210.04505* (2022).

[7] Liang Gao, Huazhu Fu, Li Li, Yingwen Chen, Ming Xu, and Cheng-Zhong Xu. 2022. FedDC: Federated Learning with Non-IID Data via Local Drift Decoupling and Correction. In *CVPR*. 10112–10121.

[8] Qing-Ling Guan, Yuze Zheng, Lei Meng, Li-Quan Dong, and Qun Hao. 2023. Improving the Generalization of Visual Classification Models Across IoT Cameras via Cross-modal Inference and Fusion. *IEEE Internet of Things Journal* (2023).

[9] Neel Guha, Ameet Talwalkar, and Virginia Smith. 2019. One-shot federated learning. *preprint arXiv:1902.11175* (2019).

[10] Wenya Guo, Ying Zhang, Xiangrui Cai, Lei Meng, Jufeng Yang, and Xiaojie Yuan. 2020. LD-MAN: Layout-driven multimodal attention network for online news sentiment recognition. *IEEE Transactions on Multimedia* 23 (2020), 1785–1798.

[11] Sungwon Han, Sungwon Park, Fangzhao Wu, Sundong Kim, Chuhan Wu, Xing Xie, and Meeyoung Cha. 2022. FedX: Unsupervised Federated Learning with Cross Knowledge Distillation. In *ECCV*. Springer, 691–707.

[12] Weituo Hao, Mostafa El-Khamy, Jungwon Lee, Jianyi Zhang, Kevin J Liang, Changyou Chen, and Lawrence Carin Duke. 2021. Towards fair federated learning with zero-shot data augmentation. In *CVPR*. 3310–3319.

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

[14] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. (2009).

[15] Ya Le and Xuan Yang. 2015. Tiny imagenet visual recognition challenge. *CS 231N* 7, 7 (2015), 3.

[16] Gihun Lee, Minchan Jeong, Yongjin Shin, Sangmin Bae, and Se-Young Yun. 2022. Preservation of the global knowledge by not-true distillation in federated learning. In *NeurIPS*.

[17] Daliang Li and Junpu Wang. 2019. Fedmd: Heterogenous federated learning via model distillation. *preprint arXiv:1910.03581* (2019).

[18] Qinbin Li, Bingsheng He, and Dawn Song. 2021. Model-contrastive federated learning. In *CVPR*. 10713–10722.

[19] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2020. Federated optimization in heterogeneous networks. *MLSys* 2 (2020), 429–450.

[20] Xiangxian Li, Haokai Ma, Lei Meng, and Xiangxu Meng. 2021. Comparative Study of Adversarial Training Methods for Long-tailed Classification. In *ADVM*. 1–7.

[21] Xiang Li, Lei Wu, Xu Chen, Lei Meng, and Xiangxu Meng. 2022. Dse-net: Artistic font image synthesis via disentangled style encoding. In *ICME*. IEEE, 1–6.

[22] Xiang Li, Lei Wu, Changshuo Wang, Lei Meng, and Xiangxu Meng. 2023. Compositional Zero-Shot Artistic Font Synthesis. *IJCAI* (2023).

[23] Xiangxian Li, Yuze Zheng, Haokai Ma, Zhuang Qi, Xiangxu Meng, and Lei Meng. 2023. Cross-modal Learning Using Privileged Information for Long-tailed Image Classification. *CVM* (2023).

[24] Chuang Lin, Sicheng Zhao, Lei Meng, and Tat-Seng Chua. 2020. Multi-source domain adaptation for visual sentiment classification. In *AAAI*, Vol. 34. 2661–2668.

[25] Jinxing Liu, Junjin Xiao, Haokai Ma, Xiangxian Li, Zhuang Qi, Xiangxu Meng, and Lei Meng. 2022. Prompt Learning with Cross-Modal Feature Alignment for Visual Domain Adaptation. In *CICAI*.

[26] Tianhan Liu, Zhuang Qi, Zitan Chen, Xiangxu Meng, and Lei Meng. 2023. Cross-Training with Prototypical Distillation for improving the generalization of Federated Learning. *ICME* (2023).

[27] Mi Luo, Fei Chen, Dapeng Hu, Yifan Zhang, Jian Liang, and Jiashi Feng. 2021. No fear of heterogeneity: Classifier calibration for federated learning with non-iid data. *NeurIPS* 34 (2021), 5972–5984.

[28] Haokai Ma, Xiangxian Li, Lei Meng, and Xiangxu Meng. 2021. Comparative study of adversarial training methods for cold-start recommendation. In *ADVM*.

[29] Haokai Ma, Zhuang Qi, Xinxin Dong, Xiangxian Li, Yuze Zheng, and Xiangxu Mengand Lei Meng. 2023. Cross-Modal Content Inference and Feature Enrichment for Cold-Start Recommendation. *IJCNN* (2023).

[30] Haokai Ma, Ruobing Xie, Lei Meng, Xin Chen, Xu Zhang, Leyu Lin, and Jie Zhou. 2023. Exploring False Hard Negative Sample in Cross-Domain Recommendation. In *Recsys*.

[31] Haokai Ma, Ruobing Xie, Lei Meng, Xin Chen, Xu Zhang, Leyu Lin, and Jie Zhou. 2023. Triple Sequence Learning for Cross-domain Recommendation. *preprint arXiv:2304.05027* (2023).

[32] Brendan McMahan, Eider Moore, Daniel Ramage, and et al. 2017. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*. PMLR, 1273–1282.

[33] Lei Meng, Long Chen, Xun Yang, Dacheng Tao, Hanwang Zhang, Chunyan Miao, and Tat-Seng Chua. 2019. Learning using privileged information for food recognition. In *MM*. 557–565.

[34] Lei Meng, Fuli Feng, Xiangnan He, Xiaoyan Gao, and Tat-Seng Chua. 2020. Heterogeneous fusion of semantic and collaborative information for visually-aware food recommendation. In *MM*. 3460–3468.

[35] Xutong Mu, Yulong Shen, Ke Cheng, Xueli Geng, Jiaxuan Fu, Tao Zhang, and Zhiwei Zhang. 2023. Fedproc: Prototypical contrastive federated learning on non-iid data. *Future Generation Computer Systems* 143 (2023), 93–104.

[36] Zhuang Qi and Xiaming Chen. 2022. A novel density-based outlier detection method using key attributes. *Intelligent Data Analysis* 26, 6 (2022), 1431–1449.

[37] Zhuang Qi, Yuqing Wang, Zitan Chen, Ran Wang, Xiangxu Meng, and Lei Meng. 2022. Clustering-based Curriculum Construction for Sample-Balanced Federated Learning. In *CICAI*. Springer, 155–166.

[38] Liangqiong Qu, Yuyin Zhou, Paul Pu Liang, Yingda Xia, Feifei Wang, Ehsan Adeli, Li Fei-Fei, and Daniel Rubin. 2022. Rethinking architecture design for tackling data heterogeneity in federated learning. In *CVPR*. 10061–10071.

[39] Ramprasaath R Selvaraju, Michael Cogswell, and et al. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*. 618–626.

[40] Xinyi Shang, Yang Lu, Gang Huang, and Hanzi Wang. 2022. Federated learning on heterogeneous and long-tailed data via classifier re-training with federated features. *preprint arXiv:2204.13399* (2022).

[41] Yujun Shi, Jian Liang, Wenqing Zhang, Vincent YF Tan, and Song Bai. 2023. Towards Understanding and Mitigating Dimensional Collapse in Heterogeneous Federated Learning. In *ICLR*.

[42] Neta Shoham, Tomer Avidor, Aviv Keren, Nadav Israel, Daniel Benditkis, Liron Mor-Yosef, and Itai Zeitak. 2019. Overcoming forgetting in federated learning on non-iid data. *preprint arXiv:1910.07796* (2019).

[43] Weilin Sun, Xiangxian Li, Manyi Li, Yuqing Wang, Yuze Zheng, Xiangxu Meng, and Lei Meng. 2022. Sequential Fusion of Multi-view Video Frames for 3D Scene Generation. In *CICAI*. Springer, 597–608.

[44] Zhenheng Tang, Yonggang Zhang, Shaohuai Shi, Xin He, Bo Han, and Xiaowen Chu. 2022. Virtual homogeneity learning: Defending against data heterogeneity in federated learning. In *ICML*. PMLR, 21111–21132.

[45] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *J MACH LEARN RES* 9, 11 (2008).

[46] Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. 2020. Federated learning with matched averaging. *preprint arXiv:2002.06440* (2020).

[47] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. 2020. Tackling the objective inconsistency problem in heterogeneous federated optimization. *NeurIPS* 33 (2020), 7611–7623.

[48] Yuqing Wang, Xiangxian Li, Haokai Ma, Zhuang Qi, Xiangxu Meng, and Lei Meng. 2022. Causal Inference with Sample Balancing for Out-of-Distribution Detection in Visual Classification. In *CICAI*. Springer, 572–583.

[49] Yuqing Wang, Xiangxian Li, Zhuang Qi, Jingyu Li, Xuelong Li, Xiangxu Meng, and Lei Meng. 2022. Meta-causal feature learning for out-of-distribution generalization. In *ECCVW*. Springer, 530–545.

[50] Yuqing Wang, Zhuang Qi, Xiangxian Li, Jinxing Liu, Xiangxu Meng, and Lei Meng. 2023. Multi-channel Attentive Weighting of Visual Frames for Multimodal Video Classification. *IJCNN* (2023).

[51] Lei Wu, Xi Chen, Lei Meng, and Xiangxu Meng. 2020. Multitask adversarial learning for Chinese font style transfer. In *IJCNN*. IEEE, 1–8.

[52] Y Xun, W Meng, Z Luming, and Dacheng Tao. 2016. Empirical risk minimization for metric learning using privileged information. In *IJCAI*.

[53] Xun Yang, Jianfeng Dong, Yixin Cao, Xun Wang, Meng Wang, and Tat-Seng Chua. 2020. Tree-Augmented Cross-Modal Encoding for Complex-Query Video Retrieval. *SIGIR* (2020).

[54] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. 2021. Deconfounded Video Moment Retrieval with Causal Intervention. *SIGIR* (2021).

[55] Xun Yang, Meng Wang, and Dacheng Tao. 2018. Person Re-Identification With Metric Learning Using Privileged Information. *IEEE Transactions on Image Processing* 27, 2 (2018), 791–805. https://doi.org/10.1109/TIP.2017.2765836

[56] Lin Zhang, Yong Luo, Yan Bai, Bo Du, and Ling-Yu Duan. 2021. Federated learning for non-iid data via unified feature learning and optimization objective alignment. In *ICCV*. 4420–4428.

[57] Lin Zhang, Li Shen, Liang Ding, Dacheng Tao, and Ling-Yu Duan. 2022. Fine-tuning global model via data-free knowledge distillation for non-iid federated learning. In *CVPR*. 10174–10183.