2023 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-6891-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICME55011.2023.00117

# Cross-Training with Prototypical Distillation for improving the generalization of Federated Learning

Tianhan Liu[1], Zhuang Qi[1], Zitan Chen[1], Xiangxu Meng[1] and Lei Meng*[1,2]
[1] School of Software, Shandong University, Jinan, China
[2] Shandong Research Institute of Industrial Technology, Jinan, China
Email: {th_liu, z_qi,chenzt}@mail.sdu.edu.cn, {mxx, lmeng}@sdu.edu.cn

***Abstract*—Cross-training has become a promising strategy to handle data heterogeneity problem in federated learning, which re-train a local model across different clients to improve its generalization capability in a privacy-preserving manner. Its main idea is to make the local models to fit the data of all clients. However, the heterogeneity between data sources may lead the local models to quickly forget the knowledge learned in several rounds of cross-training. To address the problem, this paper presents a novel prototype guided cross training mechanism, termed PGCT, to regularize the change of class-level data representations across clients. It includes two main modules, where the prototype guided representation learning module employs client-aware prototypes of data patterns learned by clustering to guide the learning of consistency representation across feature spaces. This maintains the similar decision boundary across different clients. The prototype-based feature augmentation module uses prototypes as soft attention regularizers to further aggregate rich information to enhance the discrimination of historical features. Experiments were conducted on four datasets in terms of performance comparison, ablation study and case study, and the results verified that PGCT can learn discriminative features with different classes under the guidance of prototypes, which leads to better performance than the state-of-the-art methods.**

***Index Terms*—Federated learning, Cross Training, Prototype Guided, Knowledge Forgetting**

## I. Introduction

Federated learning is an emerging distributed learning paradigm and it enables multiple parties to build a shared model working for them [1]–[3]. Existing federated learning methods allow model-level interaction between client and server without sharing local data [4], [5]. This enables federated learning can effectively avoid the risk of privacy disclosure. However, recent studies have revealed the vulnerability of the federated learning model when exposed to the non-independent and identical distribution scenarios [6]–[8]. This is mainly due to the bias between the local and the global optimization objective, and it is difficult to aggregate multiple biased learners into a high-performance global model.

Existing methods for mitigating data heterogeneity issues can be roughly divided into two categories: regulating the local process under global constraints [5], [9], [10] and improving the generality of the local model [11]–[13]. The former approaches use the global output as the knowledge to guide multiple clients learn a unified objective. Conventional algorithms along this line of research include feature-based, parameter-based, and prediction-based constraints, such as FedDC [10] and MOON [5] align the output of the local and global model in the parameter and feature space respectively. The latter methods focus on the expansion of knowledge, which enable local models to fit distinct data distributions. For example, FedExg [12] and FedMe [13] apply the cross-training strategy to re-train a local model across different clients to learn comprehensive knowledge. As observed, utilizing cross-training strategy can automatically adjust the optimization objective of local models. However, the inconsistency of data distribution among clients also brings knowledge forgetting problems, which leads to limited performance gains.

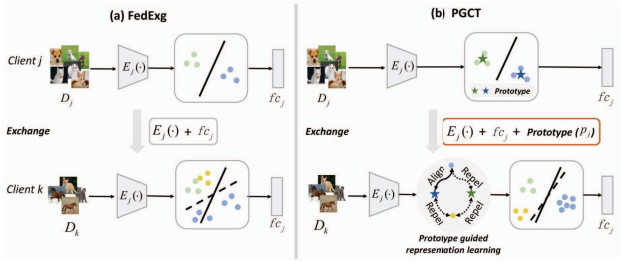


Fig. 1. Motivation of PGCT. (a) The corss-training procedures of FedExg may lead the local models to have a significant change in decision boundaries; (b) PGCT is able to regularized the learned data representations to be similar to those learned from previous clients when performing cross-training.

To address this problem, this paper presents a novel **p**rototype **g**uided **c**ross **t**raining mechanism, termed PGCT. As illustrated in Figure 1, compared with conventional method with cross-training, the proposed PGCT implements prototypical knowledge distillation to learn consistent representation across clients and maintain similar decision boundary. Specifically, PGCT has two main modules: the prototype-guided representation learning (PGRL) module and the prototype-based feature augmentation (PFA) module. Considering that private data cannot be shared, PGCT memorizes class-aware prototypes to replace class-level representations. The PGRL module utilizes these prototypes to guide the learning of consistency representation to maintain the discrimination of corresponding features. The PFA module focuses on stabilizing the decision phase, and it uses representative prototypes as soft-attention regularizers to refine and augment image features, which leverages linear combination in feature-level to fuses information from intra-class representations across clients.

1945-788X/23/$31.00 ©2023 IEEE
DOI 10.1109/ICME55011.2023.00117

Experiments are conducted on four commonly used datasets in terms of performance comparison, ablation study of the key components of PGCT, and case study for the effectiveness of representation learning. The results verify that prototypes can be used as effective knowledge to guide representation learning and prototype-guided cross-training can expand the learnable knowledge of the local model to alleviate classification bias. To summarize, this paper includes two main contributions:

- A model-agnostic cross-training mechanism, termed PGCT, is proposed to alleviate the knowledge forgetting problem. To the best of our knowledge, this is the first method that uses data prototypes to guide local models to learn consistent representation across clients in federated learning.
- We found that the knowledge forgetting mainly comes from two aspects, namely, the inconsistency in representation distributions and the loss of discrimination to historical features. And we also verifies the effectiveness of PGCT in solving these problems.

## II. Related Work

### A. Federated learning with cross-training

To solve the data heterogeneity problem in federated learning, there are different training strategies: 1) local training + global aggregation and 2) local training + random exchange + cross training + global aggregation. The former methods typically align local and global optimization objectives, such as FedProx [4] and FedUFO [14] aim to align the output of the local and global model in parameter and feature space, respectively. The latter approaches utilize cross-training mechanism to retrain local models across different clients. This enables local models to be trained on more data to learn comprehensive knowledge, such as FedExg [12] and FedMe [13]. Notably, cross-training is orthogonal to the former methods and it can be combined with these techniques in the local training phase. Therefore, cross-training is a promising strategy to improve the generalization capability of local models.

### B. Knowledge distillation in federated learning

Knowledge distillation is also widely used to handle data heterogeneity problem in federated learning [15]–[18]. Existing methods typically rely on a proxy dataset, and they aggregate the local predictions of proxy dataset rather than model parameter or gradient [15], [16]. However, it is observed that the correlation of proxy data and local data determines the validity of the decision. Inspired by prototype learning, many studies have shown that global prototypes can serve as effective knowledge to guide the update of local models [19]–[21]. Prototypes are derived from the average features of all classes, it is easy to implement and involves no privacy breaches, but may lose some representative information by averaging.

## III. Cross-Training with Prototypical Distillation

### A. Overall framework

The Prototype Guided Cross-Training mechanism (PGCT) in federated learning, as depicted in Figure 2, has three main phases, including local training, cross training, and global aggregation. In phase 1, PGCT can use any federated learning algorithm to optimize the model and it memorizes a class-aware prototype for each class. Then, PGCT introduces a random shuffling for local models and prototypes, then anonymously broadcasts them in sever, which enlarges the trainable dataset for local models without privacy leakage. In phase 2, **each client obtains the model and prototypes learned from phase 1 in another client** for re-training. It has two key modules, the Prototype Guided Representation Learning module and the Prototype-based Feature Augmentation module. PGCT obtains all local models from phase 2 and aggregates them to generate a global model in phase 3.

### B. Prototype-Guided Representation Learning

The Prototype-Guided Representation Learning (PGRL) module retrains a model under the guidance of class-aware prototypes to fit distinct data distribution. It performs a generalized version of contrastive learning to align the distribution of prototypes and image features. This enables a local model to learn invariant representations across clients.

To realize cross-client prototypical distillation, we generate a class-aware prototype by clustering for each class [22]–[24]. It can learn visual patterns of classes and gather similar features of the same classes into a cluster. For example, to calculate prototypes on client $j$, the procedure can be formulated as:

$$C_1^m, C_2^m, ..., C_n^m = clustering(\mathcal{F}_j(x)|x \in D_j^m) \quad (1)$$

where $C_i^m$ denotes i-th cluster of class $m$, $D_j^m$ denote the data of class $m$. $\mathcal{F}_j(\cdot)$ is a feature extractor. And the prototype $p_j^m$ of class $m$ in client$j$ is learned by weighting, defined by

$$p_j^m = \sum_{i=1}^{n} \frac{|C_i^m|}{|D_j^m|} \mathbf{mean}(f^m | f^m \in C_i^m) \quad (2)$$

As shown in Figure 2, after random exchange, the PGRL module in the client $j$ obtains the local model $E_i$ and class-aware prototypes $p_i$ from client $i$. To prevent the overlapping of representations between different classes in the latent space happening, PGCT learns unified features by maximizing agreement between the sample and the corresponding prototype. Inspired by contrastive learning in representation learning, we define the prototype-based contrastive loss similar to NT-Xent loss [25]:

$$\mathcal{L}_{PCL} = -\log \frac{\exp(sim(f_x, p^+)/\tau)}{\exp(sim(f_x, p^+)/\tau) + \sum \exp(sim(f_x, p^-)/\tau)} \quad (3)$$

where $p^+$ and $p^-$ denote the prototype with the same and different labels as feature $f_x$ respectively. $sim(\cdot)$ is cosine similarity function, $\tau$ is a temperature parameter.

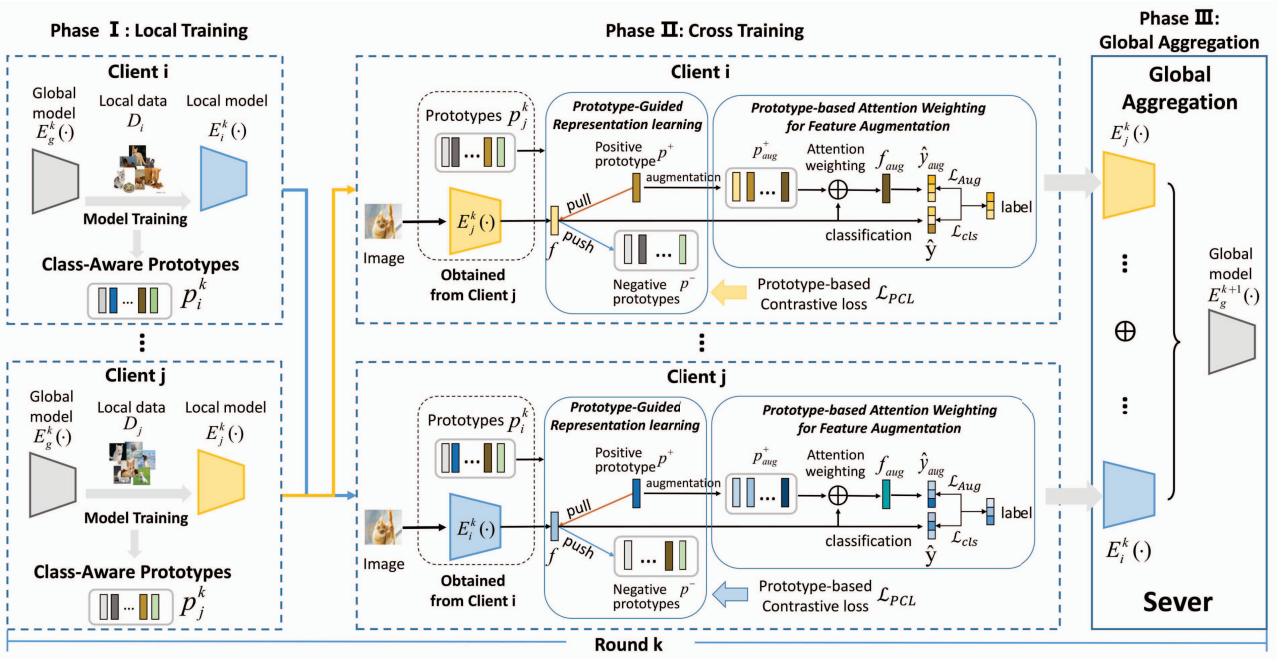


Fig. 2. Illustration of the framework of PGCT. PGCT can use arbitrary algorithms (such as FedAvg and MOON) to optimize local model in phase 1, and generate a class-aware prototype for each class. After random exchange, PGCT learns unified representations across clients to maintain the similar decision boundary and generates augmented feature via prototype-based attention weighting to enhance classification in phase 2. Finally, PGCT aggregates all local models in phase 3.

### C. Prototype-based Attention Weighting for Feature Augmentation

The Prototype-based Feature Augmentation (PFA) module aims to exploit the historical features to improve the generalization of local models. A practical idea is to reuse the class-aware prototypes to refine and augment image features. Specifically, it uses Gaussian noise to perform prototype augmentation and regards the augmented prototype as a soft attention regularizer to generate attention weights, defined by

$$\widetilde{p}_{aug}^{n} = p + \mathcal{N}(0,1) \times \varepsilon, \quad n = 1, 2, ... \tag{4}$$

where $\varepsilon$ is a scale parameter. To obtain smoother decision boundaries at feature level, the PFA module realizes feature augmentation by attention weighting to assist classification and the attention weights can be defined by dot product similarity,

$$w_n = softmax(f^T \cdot \widetilde{p}_{aug}^{n}) \tag{5}$$

where the $softmax(\cdot)$ normalizes the scores across all augmented prototypes. And the augmented image features fuse the information of the corresponding prototype and original image features, it can be expressed as

$$\widetilde{f}_{aug} = f + \sum_{j} w_n \cdot \widetilde{p}_{aug}^{n} \tag{6}$$

And we use an augmentation loss to optimize the model,

$$\mathcal{L}_{Aug} = CE(\mathcal{F}(\widetilde{f}_{aug}), \widetilde{y}_{aug}) \tag{7}$$

where $\mathcal{F}(\cdot)$ is a classifier, $\widetilde{y}_{aug}$ denotes the label of augmented features $\widetilde{f}_{aug}$.

### D. Training Strategies of PGCT

PGCT focuses on learning cross-client consistency features under the guidance of prototypes and using augmented features to assist classification. Consequently, the integrated objective of PGCT in cross training is to minimize

$$\mathcal{L}_{total} = \mathbb{E}_{(x,y)\sim D_{local}}[\mathcal{L}_{cls} + \alpha \cdot \mathcal{L}_{PCL} + \lambda \cdot \mathcal{L}_{Aug}] \tag{8}$$

where $\mathcal{L}_{cls}$ is cross entropy loss for image features classification, $\mathcal{L}_{PCL}$ is prototype-based contrastive loss and $\mathcal{L}_{Aug}$ is augmentation loss. $\alpha$ and $\lambda$ are loss weights.

## IV. Experiments

### A. Experiment Settings

*1) Datasets:* We use three benchmarking datasets MNIST [26], CIFAR-10 [27], CIFAR-100 [27] and a medical image dataset PathMNIST [28] that are commonly used in federated learning for experiments. Their statistics are shown in Table II.

*2) Network Architecture:* For a fair comparison, we use the same network architectures for all approaches. And the network includes three modules: an image encoder, a projection head, and a classifier. As in the previous works [5], [10], for all datasets, we use a 2-layer MLP as the projection head, and the classifier is a 1-layer fully-connected network. We use two

TABLE I
PERFORMANCE COMPARISON OF ALGORITHMS. FOR ALL METHODS, WE RUN TWO TRIALS AND REPORT THE MEAN AND STANDARD DERIVATION.

| | | MNIST | | CIFAR10 | | CIFAR100 | | PathMNIST | |
|---|---|---|---|---|---|---|---|---|---|
| | | $\beta = 0.3$ | $\beta = 0.5$ | $\beta = 0.3$ | $\beta = 0.5$ | $\beta = 0.3$ | $\beta = 0.5$ | $\beta = 0.3$ | $\beta = 0.5$ |
| without FL | SOLO | 74.21±2.9 | 77.98±1.2 | 38.65±0.4 | 39.52±0.8 | 22.53±0.4 | 22.87±1.3 | 31.03±1.4 | 38.17±0.6 |
| FL without Cross-Training | Fedavg | 96.14±0.8 | 96.65±0.9 | 65.64±0.8 | 66.36±1.0 | 63.18±1.1 | 63.54±0.7 | 70.71±1.5 | 73.35±0.9 |
| | FedProx | 96.48±0.2 | 97.10±0.3 | 66.03±0.4 | 66.65±0.5 | 63.89±0.5 | 64.57±0.6 | 72.18±0.4 | 74.67±0.2 |
| | SCAFFOLD | 97.11±0.1 | 97.31±0.4 | 66.26±0.2 | 66.82±0.7 | 58.54±0.3 | 58.78±0.6 | 69.47±0.5 | 72.11±0.6 |
| | FedDyn | 97.15±0.3 | 97.29±0.1 | 67.12±0.3 | 67.31±0.2 | 63.89±0.5 | 64.64±0.3 | 71.09±0.6 | 73.27±0.4 |
| | MOON | 96.84±0.7 | 97.36±0.3 | 68.54±0.8 | 69.03±0.4 | 64.92±0.7 | 64.56±0.6 | 73.62±0.7 | 74.49±0.4 |
| | FedDC | 97.24±0.5 | 97.35±0.5 | 68.29±0.6 | 69.10±0.8 | 64.38±0.3 | 64.55±0.6 | 74.21±0.4 | 75.63±0.3 |
| FL with Cross-Training | FedExg | 96.81±0.3 | 96.95±0.8 | 67.34±0.2 | 67.89±0.4 | 63.94±0.2 | 64.58±0.1 | 71.75±0.6 | 73.98±0.9 |
| | FedMe | 96.76±0.4 | 97.05±0.6 | 67.68±0.2 | 68.19±0.3 | 63.54±0.2 | 64.21±0.3 | 72.29±0.4 | 74.41±0.4 |
| | PGCT(Fedavg) | 97.88±0.4 | 98.22±0.3 | 68.89±0.2 | 69.43±0.4 | 65.43±0.5 | 65.74±0.4 | 75.12±0.4 | **77.42±0.5** |
| | PGCT(FedProx) | 98.01±0.2 | **98.31±0.2** | 68.96±0.5 | 69.74±0.3 | 65.14±0.6 | 65.38±0.8 | 75.08±0.3 | 77.32±0.3 |
| | PGCT(MOON) | **98.10±0.2** | 98.14±0.2 | **70.23±0.6** | **71.05±0.4** | **65.44±0.6** | **65.83±0.6** | **75.38±0.7** | 77.36±0.3 |

TABLE II
STATISTICS OF THE DATASETS USED IN THE EXPERIMENTS.

| **Datasets** | **#Classes** | **#Training** | **#Testing** |
|---|---|---|---|
| MNIST | 10 | 60000 | 10000 |
| CIFAR10 | 10 | 50000 | 10000 |
| CIFAR100 | 100 | 50000 | 10000 |
| PathMNIST | 9 | 89996 | 7180 |

TABLE III
ABLATION STUDY OF PGCT ON MNIST AND CIFAR10 DATASETS.

| | MNIST | | CIFAR10 | |
|---|---|---|---|---|
| | $\beta = 0.3$ | $\beta = 0.5$ | $\beta = 0.3$ | $\beta = 0.5$ |
| Base | 96.14±0.8 | 96.65±0.9 | 65.64±0.8 | 66.36±1.0 |
| +Exg | 96.81±0.3 | 96.95±0.8 | 67.34±0.2 | 67.89±0.4 |
| +Exg+PFA | 97.16±0.4 | 97.31±0.3 | 67.43±0.5 | 67.94±0.3 |
| +Exg+PGRL | 97.70±0.2 | 98.06±0.4 | 68.25±0.3 | 68.87±0.1 |
| +Exg+PFA+PGRL | **97.88±0.4** | **98.22±0.3** | **68.89±0.2** | **69.43±0.4** |

fully-connected layers as an encoder for MNIST. For CIFAR10 and PathMNIST, we use a CNN network that has two 5 × 5 convolution layers followed by 2×2 max pooling and two fully connected layers with ReLU activation. For CIFAR100, ResNet-18 without the last fully-connected network is adopted as the encoder.

*3) Hyper-parameter Settings:* For all methods, the local training epoch $E = 10$ in a global round, the client number $N = 10$ with the sample fraction $C = 1.0$, the local optimizer is SGD algorithm, communication round $T = 100$, the shardperuser. For local training, we set the weight decay as 1e-05 and the batchsize as 64, the learning rate is initiated to be 0.01, the Dirichlet parameter $\beta = 0.3$ and $\beta = 0.5$, the temperature parameter $\tau = 0.5$, $\alpha$ and $\lambda$ are selected from $\{0.01, 0.05, 0.1, 0.5, 1.0\}$, the scale parameter $\varepsilon$ is selected from $\{0.1, 0.01\}$. And the settings of other hyper-parameters refer to the corresponding paper.

### B. Performance Comparison

We compare PGCT with existing methods in three categories: 1) local training without federated learning, SOLO; 2) federated learning (FL) methods without cross-training, including FedAvg [1], FedProx [4], SCAFFOLD [6], FedDyn [29], MOON [5] and FedDC [10]; 3) FL methods with cross-training, including FedExg [12] and FedMe (The simplified version adopts the idea of mutual learning) [13]. The following results can be obtained from Table I.

- $\text{PGCT}_{\text{FedAvg}}$, $\text{PGCT}_{\text{FedProx}}$ and $\text{PGCT}_{\text{MOON}}$ achieve significant improvements in classification compared to original baselines, which demonstrates the algorithm-agnostic character of the PGCT algorithm.
- PGCT generally achieves better performance than other algorithms. It is reasonable since PGCT is able to enlarge the training set of local models and alleviate the knowledge forgetting problem.
- Federated learning methods with cross-training usually obtain better performance than the corresponding baseline (FedAvg). It verified that cross-training can combine with other algorithms and bring performance gains for them.
- For different distribution parameters $\beta$, the performance of all algorithms increases with the increase of $\beta$. It mainly because a small $\beta$ results in highly skewed local datasets. This proves that it is important to balance the classes of local data.

### C. Ablation Study

This section further studied the effectiveness of different procedures of PGCT. The results are summarized in Table III

- Using solely the model exchange (Exg) may not to lead a significant improvement, since the knowledge forgetting issue. The improvement is still limited even combined with the Prototype-based Feature Augmentation module (PFA).
- Model exchange (Exg) with the assistance of the Prototype Guided Representation Learning (PGRL) outperforms basic model on both datasets with a large margin up to 1.5% and 2.5% which verifies the effectiveness of representation learning .
- PGCT achieves the best performance combining Exg, PGRL and PFA, which demonstrates that consistent representation learning and reinforcement classifiers can alleviate knowledge forgetting.

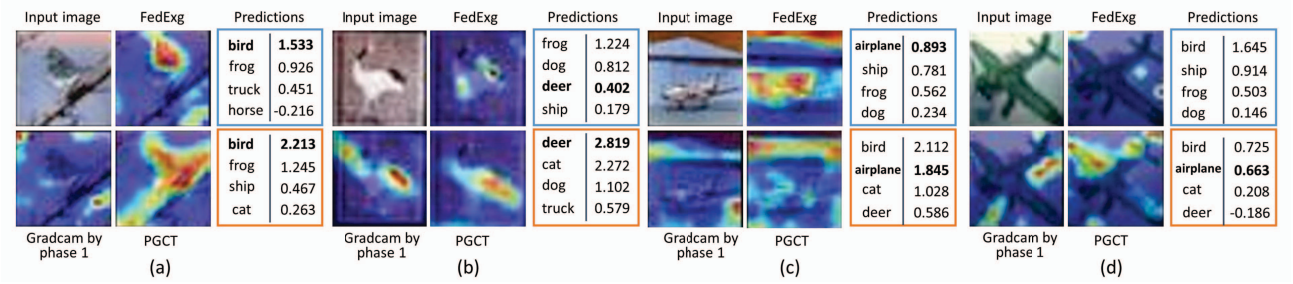


Fig. 3. (a) Cross-training methods can correct errors (b) PGCT achieves knowledge preservation in view of feature attention. (c) PGCT fails due to wrong guidance. (d) PGCT reduces the prediction difference between the ground-truth and top-1.

### D. Case Study

*1) Error Analysis of PGCT.:* This section further analyzes the working mechanism of PGCT from the perspective of feature attention and outputs of the model. And we use GradCAM [30] to generate heatmaps. As observed in Figure 3(a), the model in phase 1 cannot focus on small target, both FedExg and PGCT can learn new knowledge in the cross-training phase to make up for this deficiency and make correct predictions. When the model in phase 1 can focus on the classification objective, PGCT can retain this ability and give a correct prediction, while FedExg fails in classification due to knowledge forgetting, as shown in Figure 3(b). Figure 3(c) shows the case that the model in phase 1 learn a poor attention and gives a unreliable guidance to PGCT, while FedExg can focus on learning new knowledge and attend to the 'airplane' regions and make the correct decisions. Figure 3(d) illustrates a case that both methods make the wrong predictions. However, PGCT better attends to the airplane region and reduces the difference of prediction between 'airplane' and top-1. These observations verify the effectiveness of PGCT for federated classification with cross-training.

*2) Quality analysis of representation learning:* This section further studies the quality of representation learning. As shown in Figure 4, we use TSNE [31] technology to visualize representations in the feature space on the CIFAR10 test dataset. And we randomly selected local models of two clients. Obviously, the local models trained by the FedAvg method learn poor representation distribution, there is an overlap of multiple class features here. This is because local model faces unbalanced data distribution, which leads to poor generalization ability of the local model to the global data. Compared with FedAvg, FedExg and PGCT use cross-training to enlarge the trainable data set of the local model, which enables it to learn comprehensive knowledge. However, FedExg may obtain limited improvement due to knowledge forgetting. Intuitively, PGCT learns more discriminative representation distribution. It verified that Prototype Guided Representation Learning module helps PGCT improves the generalization ability of local models.

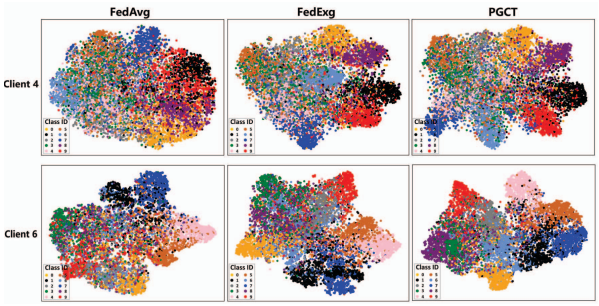


Fig. 4. Visualization of representations in the feature space. PGCT learned discriminative feature distribution under the guidance of the prototype.

## V. Conclusion

This paper presents a novel cross-training mechanism, termed PGCT, to handle the knowledge forgetting problem. PGCT performs prototype-guided representation learning to learn cross-client consistency representation. And PGCT adopts prototype-based feature augmentation for enhancing classification. Experimental results show that PGCT can effectively learn the invariant representations of the same class and discriminative representations of different classes across clients. This improves the generalization ability of local models.

This study can be further explored in two directions. First is to introduce causal inference [32], [33] to improve the effectiveness of prototype learning. Second, expanding the PGCT to more challenging tasks is valuable, such as multimodal learning [34]–[37], domain generalization [38], [39], recommendation task [40]–[44], and long-tail image classification [45].

## VI. Acknowledgments

This work is supported in part by the National Key R&D Program of China (Grant no. 2021YFC3300203), the TaiShan Scholars Program (Grant no. tsqn202211289), the Oversea Innovation Team Project of the "20 Regulations for New Universities" funding program of Jinan (Grant no. 2021GXRC073).

## REFERENCES


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

[2] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM T INTEL SYST TEC*, 10(2):1–19, 2019.

[3] Z. Qi, Y. Wang, Z. Chen, R. Wang, X. Meng, and L. Meng, "Clustering-based curriculum construction for sample-balanced federated learning," in *Artificial Intelligence: Second CAAI International Conference, CICAI 2022, Beijing, China, August 27–28, 2022, Revised Selected Papers, Part III*. Springer, 2022, pp. 155–166.

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

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

[6] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, and et al. Scaffold: Stochastic controlled averaging for federated learning. In *ICML*, pages 5132–5143. PMLR, 2020.

[7] Peter Kairouz, H Brendan McMahan, Brendan Avent, and et al. Bellet, Aurélien. Advances and open problems in federated learning. *FOUND TRENDS MACH LE*, 14(1–2):1–210, 2021.

[8] Dashan Gao, Xin Yao, and Qiang Yang. A survey on heterogeneous federated learning. *arXiv preprint arXiv:2210.04505*, 2022.

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

[10] Liang Gao, Huazhu Fu, Li Li, Yingwen Chen, Ming Xu, and Cheng-Zhong Xu. Feddc: Federated learning with non-iid data via local drift decoupling and correction. In *CVPR*, pages 10112–10121, 2022.

[11] Matias Mendieta, Taojiannan Yang, Pu Wang, Minwoo Lee, Zhengming Ding, and Chen Chen. Local learning matters: Rethinking data heterogeneity in federated learning. In *CVPR*, pages 8397–8406, 2022.

[12] Zhicheng Mao, Wenrui Dai, Chenglin Li, and et al. Fedexg: Federated learning with model exchange. In *ISCAS*, pages 1–5. IEEE, 2020.

[13] Koji Matsuda, Yuya Sasaki, and et al. Fedme: Federated learning via model exchange. In *SDM*, pages 459–467. SIAM, 2022.

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

[15] Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. *NeurIPS*, 33:2351–2363, 2020.

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

[17] Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. Data-free knowledge distillation for heterogeneous federated learning. In *ICML*, pages 12878–12889. PMLR, 2021.

[18] Hyowoon Seo, Jihong Park, Seungeun Oh, and et al. Federated knowledge distillation. *arXiv preprint arXiv:2011.02367*, 2020.

[19] Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, and et al. Fedproto: Federated prototype learning across heterogeneous clients. In *AAAI*, volume 1, page 3, 2022.

[20] Xutong Mu, Yulong Shen, Ke Cheng, Xueli Geng, and et al. Fedproc: Prototypical contrastive federated learning on non-iid data. *arXiv preprint arXiv:2109.12273*, 2021.

[21] Umberto Michieli and Mete Ozay. Prototype guided federated learning of visual feature representations. *arXiv preprint arXiv:2105.08982*, 2021.

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

[23] L. Meng, A.-H. Tan, and D. Wunsch, *Adaptive resonance theory in social media data clustering*. Springer, 2019.

[24] L. Meng, A.-H. Tan, and D. Xu, "Semi-supervised heterogeneous fusion for multimedia data co-clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 9, pp. 2293–2306, 2013.

[25] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020.

[26] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

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

[28] Jiancheng Yang, Rui Shi, and Bingbing Ni. Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis. In *ISBI*, pages 191–195. IEEE, 2021.

[29] Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, and et al. Federated learning based on dynamic regularization. *arXiv preprint arXiv:2111.04263*, 2021.

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

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

[32] Y. Wang, X. Li, Z. Qi, J. Li, X. Li, X. Meng, and L. Meng, "Meta-causal feature learning for out-of-distribution generalization," in *European Conference on Computer Vision*. Springer, 2023, pp. 530–545.

[33] 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 *Artificial Intelligence: Second CAAI International Conference, CICAI 2022, Beijing, China, August 27–28, 2022, Revised Selected Papers, Part I*. Springer, 2022, pp. 572–583.

[34] A.-H. Tan, B. Subagdja, D. Wang, and L. Meng, "Self-organizing neural networks for universal learning and multimodal memory encoding," *Neural Networks*, vol. 120, pp. 58–73, 2019.

[35] Q.-L. Guan, Y. Zheng, L. Meng, L.-Q. Dong, and Q. Hao, "Improving the generalization of visual classification models across iot cameras via cross-modal inference and fusion," *IEEE Internet of Things Journal*, 2023.

[36] W. Guo, Y. Zhang, X. Cai, L. Meng, J. Yang, and X. Yuan, "Ld-man: Layout-driven multimodal attention network for online news sentiment recognition," *IEEE Transactions on Multimedia*, vol. 23, pp. 1785–1798, 2020.

[37] L. Meng, A.-H. Tan, C. Leung, L. Nie, T.-S. Chua, and C. Miao, "Online multimodal co-indexing and retrieval of weakly labeled web image collections," in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, 2015, pp. 219–226.

[38] J. Liu, J. Xiao, H. Ma, X. Li, Z. Qi, X. Meng, and L. Meng, "Prompt learning with cross-modal feature alignment for visual domain adaptation," in *Artificial Intelligence: Second CAAI International Conference, CICAI 2022, Beijing, China, August 27–28, 2022, Revised Selected Papers, Part I*. Springer, 2022, pp. 416–428.

[39] C. Lin, S. Zhao, L. Meng, and T.-S. Chua, "Multi-source domain adaptation for visual sentiment classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 03, 2020, pp. 2661–2668.

[40] H. Ma, X. Li, L. Meng, and X. Meng, "Comparative study of adversarial training methods for cold-start recommendation," in *Proceedings of the 1st International Workshop on Adversarial Learning for Multimedia*, 2021, pp. 28–34.

[41] L. Meng, F. Feng, X. He, X. Gao, and T.-S. Chua, "Heterogeneous fusion of semantic and collaborative information for visually-aware food recommendation," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 3460–3468.

[42] H. Ma, R. Xie, L. Meng, X. Chen, X. Zhang, L. Lin, and J. Zhou, "Triple sequence learning for cross-domain recommendation," *arXiv preprint arXiv:2304.05027*, 2023.

[43] H. Wu, X. Chen, X. Li, H. Ma, Y. Zheng, X. Li, X. Meng, and L. Meng, "Vafa: A visually-aware food analysis system for socially-engaged diet management," in *Artificial Intelligence: Second CAAI International Conference, CICAI 2022, Beijing, China, August 27–28, 2022, Revised Selected Papers, Part III*. Springer, 2022, pp. 554–558.

[44] H. Wu, X. Chen, X. Li, H. Ma, Y. Zheng, X. Li, X. Meng, and L. Meng, "A visually-aware food analysis system for diet management," in *2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. IEEE, 2022, pp. 1–1.

[45] X. Li, H. Ma, L. Meng, and X. Meng, "Comparative study of adversarial training methods for long-tailed classification," in *Proceedings of the 1st International Workshop on Adversarial Learning for Multimedia*, 2021, pp. 1–7.