Rethinking Federated Learning with Domain Shift: A Prototype View

Wenke Huang[1], Mang Ye[1,2*], Zekun Shi[1], He Li[1], Bo Du[1,2*]
[1]National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence,
Hubei Key Laboratory of Multimedia and Network Communication Engineering,
School of Computer Science, Wuhan University, Wuhan, China.
[2] Hubei Luojia Laboratory, Wuhan, China.
https://github.com/WenkeHuang/RethinkFL

Abstract

Federated learning shows a bright promise as a privacy-preserving collaborative learning technique. However, prevalent solutions mainly focus on all private data sampled from the same domain. An important challenge is that when distributed data are derived from diverse domains. The private model presents degenerative performance on other domains (with domain shift). Therefore, we expect that the global model optimized after the federated learning process stably provides generalizability performance on multiple domains. In this paper, we propose Federated Prototypes Learning (FPL) for federated learning under domain shift. The core idea is to construct cluster prototypes and unbiased prototypes, providing fruitful domain knowledge and a fair convergent target. On the one hand, we pull the sample embedding closer to cluster prototypes belonging to the same semantics than cluster prototypes from distinct classes. On the other hand, we introduce consistency regularization to align the local instance with the respective unbiased prototype. Empirical results on Digits and Office Caltech tasks demonstrate the effectiveness of the proposed solution and the efficiency of crucial modules.

1. Introduction

Federated learning is a privacy-preserving paradigm [47, 83], which reaches collaborative learning without leaking privacy. The cornerstone solution, FedAvg [47], aggregates parameters from participants and then distributes the global model (averaged parameters) back for further training, which aims to learn a high-quality model without centralizing private data. However, an inherent challenge in federated learning is data heterogeneity [26, 39, 69, 87]. Specifically, the private data is collected from distinct sources with diverse preferences and presents non-iid (independently and identically distributed) distribution [87]. Each participant optimizes toward the local empirical risk minimum, which is inconsistent with the global direction. Therefore, the averaged global model unavoidably faces a slow convergence speed [40] and achieves limited performance improvement.

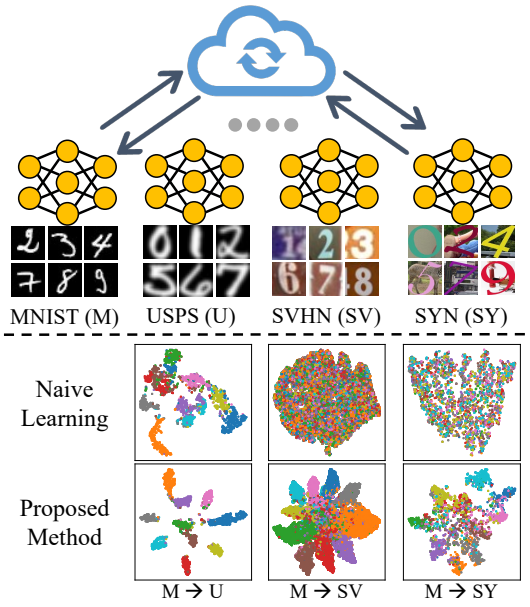


Figure 1. **Illustration of heterogeneous federated learning**. The feature visualization on inter domains (→ represents testing on target domain *i.e.*, M → SV means that local dataset is from MNIST and test model on SVHN). The **top** row indicates that local training results in domain shift. The **bottom** row shows that our method acquires generalizable performance on different domains.

A mainstream of subsequent efforts delves into introducing a variety of global signals to regulate private model [13, 28, 38, 40, 51, 66, 70]. These methods focus on label skew, where distributed data are from the **same domain**, and simulate data heterogeneity via imbalanced sampling, *e.g.*, Dirichlet strategy [32] to generate different label distributions. Nonetheless, another noticeable data het-

*Corresponding Author: Mang Ye, Bo Du

erogeneous property in federated learning is **domain shift** [21,22,41,55,57]. In particular, private data is derived from **various domains**, leading to distinct feature distributions. In this scenario, we argue that naive learning on private data brings poor generalizable ability in Fig. 1. Specifically, the private model fails to provide discrimination on other domains because it overfits local domain distribution. The aforementioned methods mainly regulate the private model via global knowledge (*i.e.*, the average signals from participants). Therefore, these algorithms share a common weakness: *the global information is insufficient to describe diverse domain knowledge*, which is magnified under the domain shift and thus hinders the improvement of generalizability. An intuitive solution is to preserve multiple models for distilling respective domain knowledge. However, it incurs a high cost of both communication and computation.

Taking into account both the effectiveness and efficiency, we rethink the prototype [11, 36, 67, 82, 91], which is the mean value of features with identical semantics. It represents class-wise characteristics and is vector type [90]. Given the enormous participant scale in federated learning, it is not efficient and feasible to maintain all prototypes. However, directly averaging all prototypes to get global prototypes would arise the same impediment as global models because averaging operation weakens the domain diversity. Besides, global prototypes probably yield biased to the dominant domain due to the unknown of private domains proportion, which results in disadvantageous performance on minority domains. Driven by these two issues, on the one hand, we find representative prototypes by clustering all prototypes. Therefore, each class is abstracted by a set of diverse prototypes, capturing rich domain variance. On the other hand, we generate unbiased prototypes based on cluster prototypes to construct fair and stable global signals, which avoid optimizing toward the underlying primary domain and thus ensure stability on different domains. Compared with original feature vectors, cluster and unbiased prototypes are privacy-friendly because it experiences twice and third times averaging operation [70]. Hence, it is less feasible to disentangle each raw representation and subsequently reconstruct private data. We analyze the superiority of cluster prototypes and unbiased prototypes in Sec. 3.2.

In this paper, we propose **F**ederated **P**rototype **L**earning (FPL), which consists of two components. **First**, in order to improve the generalizability on the premise of discriminability. We introduce Cluster Prototypes Contrastive Learning (CPCL), which leverages cluster prototypes to construct contrastive learning [7,19,79,84,85]. CPCL adaptively enforces the query embedding to be more similar to cluster prototypes from the same class than other prototypes with different semantics. In particular, such an objective encourages instance feature to be close to representative prototypes in the same semantic and separates it away from other class prototypes, which incorporates diverse domain knowledge and maintains a clear decision boundary. **Second**, we utilize unbiased prototypes to provide a fair and stable convergence point and propose Unbiased Prototypes Consistent Regularization (UPCR). Specifically, we average cluster prototypes to acquire unbiased prototypes. The local instance is required to minimize the feature-level distance with the corresponding unbiased prototype. Therefore, the local model would not be biased toward dominant domains and exhibits stable performance on inferior domains. We conjecture that these two components together make FPL a competitive method for federated learning with domain shift. The main contributions are summarized below.

- We focus on heterogeneous federated learning with domain shift and identify that the inherent limitation of existing methods is that global regularization signal is insufficient to depict diverse domain knowledge and biased toward major domain among participants.
- We propose a simple yet effective strategy to learn a well generalizable global model in federated learning with domain shift. Inspired by the success of prototype learning, we introduce cluster prototypes to provide rich domain knowledge and further construct unbiased prototypes based on the average of cluster prototypes to further offer fair and stable objective signal.
- We conduct extensive experiments on Digits [23, 33, 52, 61] and Office Caltech [16] tasks. Accompanied with a set of ablative studies, promising results validate the efficacy of FPL and the indispensability of each module.

2. Related Work

2.1. Data Heterogeneous Federated Learning

Federated learning is proposed to handle privacy concerns in the distributed learning environment. A pioneering federated method, FedAvg [47] trains a global model by aggregating local model parameters. However, its performance is impeded due to decentralized data, which poses non-i.i.d distribution (called data heterogeneity). Based on FedAvg, existing methods of tackling data heterogeneity problem mainly leverage global penalty term. FedProx [40], FedCurv [66], pFedME [68], and FedDyn [1] calculate global parameter stiffness to control discrepancies. Besides, MOON [38], FedUFO [86], FedProto [70], and FedProc [51] maximize feature-level agreement of local model and global model. Moreover, SCAFFOLD [28] and FedDC [13] leverage global gradient calibration to control local drift. The major limitation of these methods is that they focus on single domain performance under label skew scenario and overlook the problem of domain shift, leading to an unsatisfying generalizable performance on multiple domains. Closely related methods such as FedBN [41], ADCOL [37], FCCL [22] focus on personalized models rather than a shared global model. Besides, the latter two

methods require additional discriminator and public data, which incurs a heavy burden for either the participant or the server side. In this paper, we introduce cluster prototypes (diverse domain knowledge) and unbiased prototypes (consistent optimization direction) to learn a generalizable and stable global model during the federated learning process.

2.2. Clustering Federated Learning

Clustered federated learning involves grouping clients with similar data distributions into clusters, such that each client is uniquely associated with a particular data distribution and contributes to the training of a model tailored to that distribution [64, 80]. Existing methods can mainly leverage four types of clustering signals: model parameters [5, 43], gradient information [10, 64], training loss [15, 46] and exogenous information [2, 42]. However, we leverage the clustering strategy to select representative prototypes in order to address the federated learning with domain shift.

2.3. Prototype Learning

Prototype refers to the mean vector of the instances belonging to the identical class [67]. Due to its exemplar-driven nature and simpler inductive bias, it has boosted great potential in a variety of tasks. For example, in supervised classification tasks, it labels testing images via calculating its distance with prototypes of each class, which is considered to be more robust and stabler [81] in handling few-shot [48, 67, 75, 81], zero-shot [25]. Moreover, it also has been a surge of interest in semantic segmentation task [35, 77, 90], unsupervised learning [18, 36, 79, 84, 85] and so on. As for federated learning, prototypes can provide diverse abstract knowledge while adhering to privacy protocols. There exist a few works incorporating prototypes to handle data heterogeneous federated learning. PGFL [49] leverages prototypes to construct weight attention parameter aggregation. FedProc [51] and FedProto [70] aim to reach feature-wise alignment with global prototypes. CCVR [44] generates virtual feature based on approximated Gaussian Mixture Model. FedPCL [71] focuses on personalized federated learning and utilizes prototypes to learn personalized models. However, these methods focus on single-domain performance. In domain shift, it is vital to consider generalization on diverse domains. Our work sheds light on leveraging cluster and unbiased prototypes to achieve this goal in federated learning.

2.4. Contrastive Learning

Contrastive learning has recently become a promising direction in the self-supervised learning field, achieving competitive performance as supervised learning. The classic methods [7, 19, 53, 79, 84, 85] mainly construct a positive pair and a negative pair for each instance and leverage InfoNCE [54] to contrast positiveness against negativeness. A major branch of subsequent research focuses on elaborating the selection of the informative positive pairs [3, 12, 30, 31, 34, 56, 59, 65] and negative pairs [8, 14, 24, 27, 50, 60, 72, 88]. Another line explicitly investigates the semantic structure and introduces unsupervised clustering methods to construct fruitful prototypes as representative embeddings for groups of semantically similar instances [6, 18, 36, 73, 89]. Differently, in this work, Cluster Prototypes Contrastive Learning (CPCL) is designed for providing generalization ability in federated learning with domain shift. We leverage the unsupervised clustering algorithm to select representative prototypes for each class and then seek to attract each instance to cluster prototypes in the same semantics while pushing away other cluster prototypes from different classes, which brings both generalizable and discriminative ability.

3. Methodology

3.1. Preliminaries

Following the typical federated learning [40, 47], there are M participants (indexed by m) with respective private data, $D_m = \{x_i, y_i\}_{i=1}^{N_m}$, where N_m denotes the local data scale. Under heterogeneous federated learning, the conditional feature distribution $P(x|y)$ varies across participants even if $P(y)$ is consistent, resulting in **domain shift**:

- **Domain shift**: $P_m(x|y) \neq P_n(x|y)$ ($P_m(y) = P_n(y)$). There exists domain shift among private data. Specifically, for the same label space, distinctive feature distribution exists among different participants.

Besides, participants agree on sharing a model with the same architecture. We regard the model with two modules: feature extractor and unified classifier. The feature extractor $f : \mathcal{X} \to \mathcal{Z}$, encodes sample x into a compact d dimensional feature vector $z = f(x) \in \mathbb{R}^d$ in the feature space $\mathcal{Z}$. A unified classifier $g : \mathcal{Z} \to \mathbb{R}^{|I|}$, maps feature z into logits output $l = g(z)$, where I means the classification categories. The optimization direction is to learn a generalizable global model to present favorable performance on multiple domains, through the federated learning process.

3.2. Prototypes Meet Federated Learning

Motivation. Each prototype $c^k \in R^d$ is calculated by the mean vector of the features belonging to same class:

$$c^k = \frac{1}{|S^k|} \sum_{(x_i, y_i) \in S^k} f(x_i), \tag{1}$$

where S^k means the set of samples annotated with class k. Prototypes are typical for respective semantic information. Besides, it carries the specific domain style information because the prototypes are not consistent on different domains. Therefore, it motivates us to leverage prototypes from different domains to learn a generalizable model without leaking privacy information. We further define the k^{th}

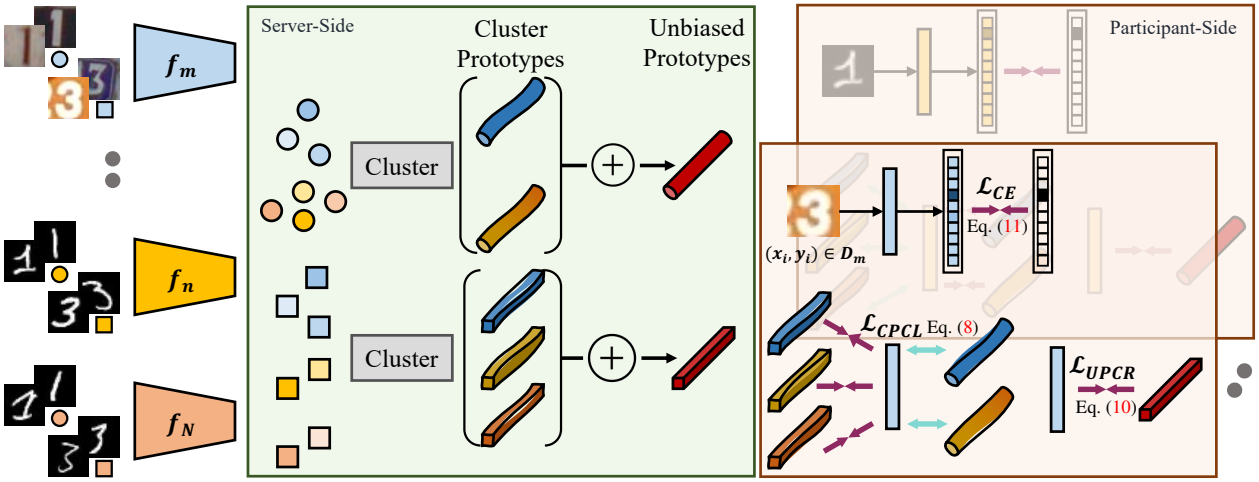


Figure 2. **Architecture illustration** of Federated Prototypes Learning (FPL). Participants upload local prototypes to server. Based on these prototypes, we introduce cluster prototypes ( ) to construct Cluster Prototypes Contrastive Learning (CPCL in Sec. 3.3.1), bringing diverse domain information. Besides, we acquire unbiased prototypes ( ) and propose Unbiased Prototypes Consistent Regularization (UPCR in Sec. 3.3.2) to provide a stable consistency signal. Best viewed in color. Zoom in for details.

class prototype from the m^{th} participant as:

$$c_m^k = \frac{1}{|S_m^k|} \sum_{(x_i,y_i)\in S_m^k} f_m(x_i)$$
$$\mathcal{O}_m = [c_m^1, \dots, c_m^k, \dots, c_m^{|I|}] \in \mathbb{R}^{|I|\times d}, \tag{2}$$

where $S_m^k = \{x_i, y_i | y_i = k\}_{i=1}^{N_m^k} \subset D_m$ represents the private dataset D_m of the k^{th} class for the m^{th} participant.

Global Prototypes. Considering that number of participants is large-scale in federated learning, the straightforward solution to leverage prototypes is the global prototypes ($\mathcal{G}$) via directly averaging operation akin to the global model. Hence, the global prototypes are formulated into:

$$\mathcal{G}^k = \frac{1}{N}\sum_{m=1}^{N} c_m^k \in \mathbb{R}^d$$
$$\mathcal{G} = [\mathcal{G}^1, \dots, \mathcal{G}^k, \dots, \mathcal{G}^{|I|}]. \tag{3}$$

However, global prototypes mainly suffer from two notable problems. ❶ Global prototype unavoidably faces the same dilemma as the global model. In detail, it depicts each class signal by only one prototype, bearing no domain variation under heterogeneous federated learning with domain shift. ❷ Moreover, due to the unknown of participants data distribution in federated learning, global prototypes would be biased toward the dominant domain distribution, leading to a skewed optimization objective during federated process.

Cluster Prototypes. Inspired by such limitations, we first propose cluster prototypes. Compared with global prototypes, we select representative prototypes rather than single one via unsupervised clustering method, FINCH [63]. Compared with well-known clustering techniques such as Kmeans [4, 45] and HAC [78], FINCH is parameter-free and thus suitable for federated learning with uncertain participants scale. Specifically, FINCH views that the nearest neighbor of each sample is a sufficient support for grouping. It implicitly picks characteristic prototypes because prototypes from different domains are less likely to be the first neighbor. Therefore, prototypes from different domains probably fail to merge together, while prototypes from similar domains fall into the same group, conversely. Specifically, we leverage cosine similarity to evaluate the distance between any two prototypes and view the prototype with minimum distance as its 'neighbor', sorted into the same set. We define the k^{th} class prototype adjacency matrix as:

$$A^k(m,n) = \begin{cases} 1, & \text{if } n = v_m^k \text{ or } m = v_n^k \text{ or } v_m^k = v_n^k; \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

where v_m^k denotes the first neighbor (largest cosine similarity) of the class k prototype from the m^{th} participant, c_m^k. Then, we select several representative prototypes in the embedding space based on the clustering results via Eq. (4). Thus, the cluster prototypes ($\mathcal{P}$) are denoted as:

$$\mathcal{P}^k = \{c_m^k\}_{m=1}^N \xrightarrow{\text{Cluster}} \{c_m^k\}_{m=1}^J \in \mathbb{R}^{J\times d}$$
$$\mathcal{P} = \{\mathcal{P}^1, \dots, \mathcal{P}^k, \dots, \mathcal{P}^{|I|}\}. \tag{5}$$

We cluster N prototypes into J representatives of class k, which effectively addresses the aforementioned problem ❶.

Unbiased Prototypes. Nevertheless, the scale of cluster prototypes is variant because the unsupervised clustering methods generate them after each communication, which can not ensure a stable and fair convergent point. Thus, we further average cluster prototypes to get a consistent signal: unbiased prototypes ($\mathcal{U}$), which is calculated as follows:

$$\mathcal{U}^k = \frac{1}{J}\sum_{c^k\in\mathcal{P}^k} c^k \in \mathbb{R}^d$$
$$\mathcal{U} = [\mathcal{U}^1, \dots, \mathcal{U}^k, \dots, \mathcal{U}^{|I|}]. \tag{6}$$

Note that compared with global prototypes ($\mathcal{G}$), unbiased prototypes ($\mathcal{U}$) largely avoid being biased toward dominant domains in heterogeneous federated learning and provide a stable optimization target. Thus, we hypothesize that unbi-

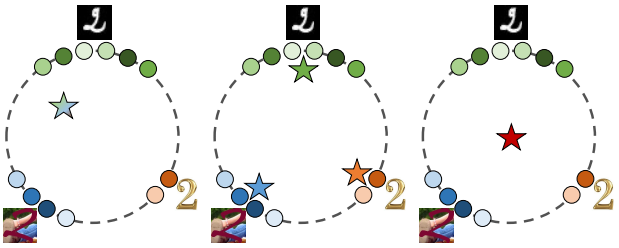


Figure 3. **Illustration of different prototypes**. Global prototype (☆) fails to describe diverse domains information and is biased toward the underlying dominant domain. Cluster prototype (★ ★ ★) and unbiased prototype (★) carry multiple domain knowledge and stable optimization signal. See details in Sec. 3.2.

ased prototypes could depict the considerably fair convergent point and further leverage them to conduct consistent regularization, which productively handles the problem ❷.

Discussion. We further explain the difference of these three kinds of prototypes in Fig. 3. Global prototypes inherently present limited domain knowledge and show skewed feature space toward the potentially dominant domains in heterogeneous federated learning. Cluster and unbiased prototypes complementarily handle these two problems because the former provides fruitful domain knowledge and the latter represents an ideal optimization target, collaboratively ensuring both generalization and stability. Compared with existing methods that leverage the global model to construct regularization term, prototypes are substantially smaller in size than model parameters, which bring less computation cost for participants. Besides, cluster prototypes and unbiased prototypes are privacy-safe because they experience twice and third times averaging operations through unsupervised clustering. Therefore, leveraging these two kinds of prototypes: cluster prototypes and unbiased prototypes is not only a computation-friendly media but also a privacy-preserving solution in heterogeneous federated learning.

3.3. Federated Prototypes Learning

For generalizability and stability in heterogeneous federated learning with domain shift, we leverage cluster prototypes and unbiased prototypes to obtain fruitful domain knowledge and stable consistency signal. The proposed method comprises two key components: Cluster Prototypes Contrastive Learning (CPCL in Sec. 3.3.1) and Unbiased Prototypes Consistent Regularization (UPCR in Sec. 3.3.2).

3.3.1 Cluster Prototypes Contrastive Learning

We deem that a well-generalizable representation should not only be discriminative to provide a clear decisional boundary for different classes but also be as invariant as possible to diverse domain distortions that are applied to this sample. Specifically, for instance $(x_i, y_i) \in D_m$, we feed it into network and acquire its feature vector $z_i = f(x_i)$. Then, we enforce the instance feature to be similar to respective semantic prototypes ($\mathcal{P}^k$) and dissimilar to different semantic prototypes ($\mathcal{N}^k = \mathcal{P} - \mathcal{P}^k$). We define the similarity of the query sample embedding z_i with corresponding cluster prototypes $c \in \mathcal{P}$ as follows:

$$s(z_i, c) = \frac{z_i \cdot c}{||z_i|| \times ||c|| / \tau}, \tag{7}$$

where temperature hyper-parameter, τ controls the concentration strength of representations [76]. Thus, we expect to enlarge the similarity with semantic coincident cluster prototypes than other different cluster prototypes, which aims to maintain a clear class-wise decision boundary. In our work, we introduce Cluster Prototypes Contrastive Learning (CPCL) to contrast cluster prototypes with the same class for each query sample against other remainder of cluster prototypes with different semantics. It is natural to derive the following optimization objective term:

$$\begin{aligned}\mathcal{L}_{CPCL} &= -\log \frac{\sum_{c\in\mathcal{P}^k} \exp(s(z_i, c))}{\sum_{c\in\mathcal{P}^k} \exp(s(z_i, c)) + \sum_{c\in\mathcal{N}^k} \exp(s(z_i, c))} \\ &= \log(1 + \frac{\sum_{c\in\mathcal{N}^k} \exp(s(z_i, c))}{\sum_{c\in\mathcal{P}^k} \exp(s(z_i, c))}),\end{aligned} \tag{8}$$

We give a detailed optimization direction analysis of Eq. (8) and thus reformulate the CPCL loss function as follows:

$$\begin{aligned}&\min \mathcal{L}_{CPCL} \\ &\equiv \log(\frac{\sum_{c\in\mathcal{N}^k} \exp(s(z_i, c))}{\sum_{c\in\mathcal{P}^k} \exp(s(z_i, c))}) \\ &\equiv \underbrace{\log(\sum_{c\in\mathcal{N}^k} \exp(s(z_i, c)))}_{\textbf{Discriminative}} - \underbrace{\log(\sum_{c\in\mathcal{P}^k} \exp(s(z_i, c)))}_{\textbf{Generalizable}}.\end{aligned} \tag{9}$$

Note that minimizing Eq. (8) equally requires pulling embedding vector z_i closely to its assigned positive cluster prototypes ($\mathcal{P}^k$) and pushing z_i far away from others negative prototypes ($\mathcal{N}^k$), which not only aims to be invariant to diverse domain distortions but also enhances the semantic spread-out property, promising both generalizable and discriminative property of feature space and thus acquiring satisfying generalizable performance in federated learning.

3.3.2 Unbiased Prototypes Consistent Regularization

Although cluster prototypes bring diverse domain knowledge for the sake of plasticity under domain shift, the cluster prototypes are dynamically generated at each communication and its scale is changing due to the unsupervised clustering method. Therefore, cluster prototypes could not offer a stable convergence direction at different communication epochs. We assume that unbiased prototypes ($\mathcal{U}$ in Eq. (6)) based on averaged cluster prototypes, could provide a relatively fair and stable optimization point and thus cope with the problem of convergence instability. Thus, in this paper, we purpose Unbiased Prototypes Consistent Regularization

(UPCR) to leverage unbiased prototypes. Specifically, we utilize a consistency regularization term to pull the feature vector z_i closer to the respective unbiased prototype, $\mathcal{U}^k$ as:

$$\mathcal{L}_{UPCR} = \sum_{v=1}^{d} (z_{i,v} - \mathcal{U}_v^k)^2, \tag{10}$$

where v indexes the dimension of feature output. We expect to achieve feature-level alignment between query embedding and the corresponding unbiased prototype. Besides, we construct CrossEntropy [9] loss and use the logits output ($l_i = g(z_i)$) with original annotation signal (y_i) to maintain local domain discriminative ability via:

$$\mathcal{L}_{CE} = -\mathbf{1}_{y_i} \log(\sigma(l_i)), \tag{11}$$

where σ denotes `softmax`. Finally, we carry out the following optimization objective in local updating phase:

$$\mathcal{L} = \mathcal{L}_{CPCL} + \mathcal{L}_{UPCR} + \mathcal{L}_{CE}. \tag{12}$$

The overall federated learning algorithm is shown in Algorithm 1. In each communication epoch, the server distributes the cluster prototypes and unbiased prototypes to participants. In local updating, each participant optimizes on local data, while the objective is defined in Eq. (12).

Algorithm 1: FPL

Input: Communication epochs E, local rounds T, number of participants M, m^{th} participant private data $D_m(x, y)$, private model θ_m

Output: The final global model θ^E

for $e = 1, 2, ..., E$ **do**
- *Participant Side*;
- **for** $m = 1, 2, ..., N$ **in parallel do**
  - $\theta_m^e, \mathcal{O}_m \leftarrow$ **LocalUpdating**$(\theta^e, \mathcal{P}, \mathcal{U})$
- *Server Side*;
- $\theta^{e+1} \leftarrow \sum_{m=1}^{N} \frac{|D_m|}{|D|} \theta_m^e$
- `/* Cluster prototypes */`
- $\mathcal{P}^k = \{c_m^k\}_{m=1}^N \xrightarrow{\text{Cluster}} \{c_m^k\}_{m=1}^J$ via Eq. (5)
- `/* Unbiased prototypes */`
- $\mathcal{U}^k = \frac{1}{J} \sum_{c^k \in \mathcal{P}^k} c^k$ by Eq. (6)

LocalUpdating$(\theta^e, \mathcal{P}, \mathcal{U})$:

$\theta_m^e \leftarrow \theta^e$; `// Distribute global parameter`

for $t = 1, 2, ..., T$ **do**
- **for** $(x_i, y_i) \in D_m$ **do**
  - $z_i = f_m^e(x_i)$
  - $l_i = g_m^e(z_i)$
  - `/* Cluster Prototypes Contrastive Learning */`
  - $\mathcal{L}_{CPCL} \leftarrow (z_i, \mathcal{P})$ in Eq. (8) ; `// Sec. 3.3.1`
  - `/* Unbiased Prototypes Consistent Regularization */`
  - $\mathcal{L}_{UPCR} \leftarrow (z_i, \mathcal{U})$ in Eq. (10) ; `// Sec. 3.3.2`
  - $\mathcal{L}_{CE} \leftarrow (l_i, y_i)$ in Eq. (11)
  - $\mathcal{L} = \mathcal{L}_{CPCL} + \mathcal{L}_{UPCR} + \mathcal{L}_{CE}$
  - $\theta_m^e \leftarrow \theta_m^e - \eta \nabla \mathcal{L}$

$\mathcal{O}_m = \{\}$; `// Initialize local prototypes`

`/* Local prototypes */`

for $k = 1, 2, ..., |I|$ **do**
- $S_m^k = \{x_i, y_i | y_i = k\}^{N_m^k} \subset D_m$
- $c_m^k = \frac{1}{N_m^k} \sum_{(x_i, y_i) \in S_m^k} f_m(x_i)$
- $\mathcal{O}_m = \mathcal{O}_m \cup \{c_m^k\}$ in Eq. (2)

return $\theta_m^e, \mathcal{O}_m$

4. Experiments

4.1. Experimental Setup

Datasets. We evaluate methods on two classification tasks:

- Digits [23, 33, 52, 61] includes four domains: MNIST (M), USPS (U), SVHN (SV) and SYN (SY) with 10 categories (digit number from 0 to 9).
- Office Caltech [16] consists four domains: Caltech (C), Amazon (A), Webcam (W) and DSLR (D), which is formed of ten overlapping classes between Office31 [62] and Caltech-256 [17].

We initialize 20 and 10 participants for Digits and Office Caltech tasks and randomly allocate domains for participants. In detail, the Digits task is MNIST: 3, USPS: 7, SVHN: 6 and SYN: 4. The Office Caltech is Caltech: 3, Amazon: 2, Webcam: 1 and DSLR: 4. For each participant, local data is randomly selected from these domains with different proportions (*i.e.*, 1 % in Digits and 20 % in Office Caltech), based on the difficulty and scale of the tasks. We fix the seed to ensure reproduction of our results.

Model. For these two classification tasks, we conduct experiment with ResNet-10 [20]. The feature vector dimension is 512. Note that all methods use the same network architecture for fair comparison in different tasks.

Counterparts. We compare ours against several sota federated methods focusing on **learning a shared global model**: FedAvg (AISTATS'17 [47]), FedProx (arXiv'18 [40]), MOON (CVPR'21 [38]), FedDyn (ICLR'21 [1]), FedOPT (ICLR'21 [58]), FedProc (arXiv'21 [51]) and FedProto (AAAI'22 [70] with parameter averaging).

Implement Details. To enable a fair comparison, we follow the same setting in [22,38]. We conduct communication epoch for $E = 100$ and local updating round $T = 10$, where all federated learning approaches have little or no accuracy gain with more communications. We use the SGD optimizer with the learning rate $lr = 0.01$ for all approaches. The corresponding weight decay is $1e-5$ and momentum is 0.9. The training batch size is 64. The hyper-parameter setting for FPL presents in the next Sec. 4.2.

Evaluation Metric. Following [40, 47], Top-1 accuracy is adopted for fair evaluation in these two classification tasks. We conduct experiments for three times and utilize the last five communication epochs accuracy as final performance.

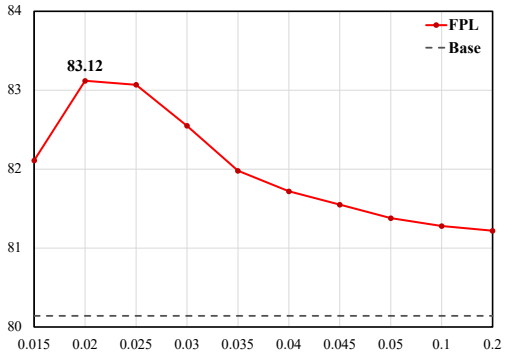
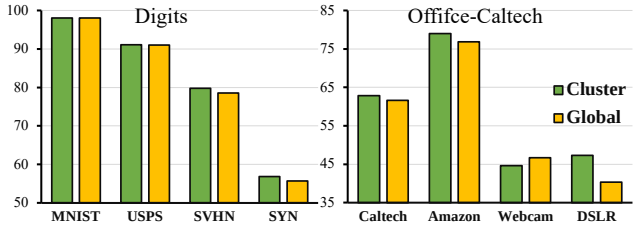


Figure 4. **Analysis of FPL with different temperature (Eq. (7)).** "Base" denotes FedAvg. See details in Sec. 4.2.

CPCL	UPCR	Digits MNIST	USPS	SVHN	SYN	AVG
		98.14	90.85	76.56	55.01	80.14
✓		98.03	91.13	79.76	56.84	81.44
	✓	98.23	93.12	81.18	55.40	81.98
✓	✓	98.31	92.71	80.27	61.20	**83.12**
CPCL	UPCR	Office Caltech Caltech	Amazon	Webcam	DSLR	AVG
		60.15	75.44	45.86	36.00	54.36
✓		61.65	78.16	43.62	45.33	57.19
	✓	64.26	79.54	48.39	44.67	59.21
✓	✓	63.39	79.26	55.86	48.00	**61.63**

Table 1. **Ablation study of key components** of our method in Digits and Office Caltech task. Please see Sec. 4.2 for details.



Prototype	Digits AVG	△	Office Caltech AVG	△
Global $\mathcal{G}$	80.82	-	56.37	-
Cluster $\mathcal{P}$	**81.44**	**+0.62**	**58.45**	**+2.08**

Figure 5. **Comparison of cluster ($\mathcal{P}$ in Eq. (5)) and global prototypes ($\mathcal{G}$ in Eq. (3)) for CPCL (Sec. 3.3.1)** on each domain (Top Row) and overall performance (Bottom Row) in Digits and Office Caltech tasks with $\tau = 0.02$. See details in Sec. 4.2.1.

4.2. Diagnostic Analysis

For thoroughly analyzing the efficacy of essential modules in our approach, we perform an ablation study on Digits and Office Caltech to investigate: Cluster Prototypes Contrastive Learning (CPCL) and Unbiased Prototypes Consistent Regularization (UPCR). We firstly present the overall performance with different contrastive temperature (τ in Eq. (8)). The Fig. 4 reveals that a smaller temperature benefits training more than higher ones, but extremely low temperatures are harder to train due to numerical instability ($\mathcal{L} = NaN$ in Eq. (12) when $\tau = 0.01$), corroborating relevant observations reported in [7, 29, 74, 76]. Specifically, the accuracy progressively increases as τ enlarges, and the amelioration becomes marginal when $\tau = 0.02$. Hence, we choose $\tau = 0.02$ by default. We further give a quantitative result on these two components in Tab. 1. The first row refers to the FedAvg which directly averages model parameter without extra operation. Three crucial conclusions can be drawn. **First**, CPCL leads to significant performance improvements against the baseline on different tasks. This evidences that CPCL strategy is able to produce generalizable feature space. **Second**, we notice gains by incorporating UPCR into the baseline. This proves the importance of considering consistent regularization. **Third**, combining CPCL and UPCR achieves better performance, which supports our motivation of exploiting joint generalization and stability in heterogeneous federated learning.

4.2.1 Cluster Prototypes Contrastive Learning

To prove the superiority of cluster prototypes ($\mathcal{P}$ in Eq. (5)) in providing generalizable and discriminative ability, we compare them with global prototypes ($\mathcal{G}$ in Eq. (3)) on Office Caltech task under contrastive temperature $\tau = 0.02$ in Fig. 5. The results reveal that leveraging cluster prototypes performs better than global prototypes and thus confirm our motivation of leveraging multiple prototypes to capture diverse domain knowledge. For example, in Office Caltech task, cluster prototypes achieve 2.08% overall performance gain compared with global prototypes.

4.2.2 Unbiased Prototypes Consistent Regularization

Note that both global prototypes ($\mathcal{G}$) and unbiased prototypes ($\mathcal{U}$) are able to offer consistent regularization. We conduct experiments on Digits and Office Caltech in Tab. 3. These results confirm the superiority of utilizing unbiased prototypes to offer consistent signal. As seen, the unbiased prototypes provide a better convergence signal than global prototypes and present the increased performance on different tasks *i.e.*, Digits ($+0.82$) and Office Caltech ($+0.73$).

4.3. Comparison to State-of-the-Arts

The Tab. 2 plots the final accuracy metric by the end of federated learning process with popular sota methods. It clearly depicts that our method performs significantly better than counterparts, which confirms that FPL can acquire well generalizable ability and thus effectively boost performance on different domains. Take the result of Office Caltech as an example, our method outperforms the best counterpart with a gap of 4.59%. We visualize the t-SNE visualization analysis of FPL at different communication epochs in Fig. 7,

Methods	Digits						Office Caltech					
	MNIST	USPS	SVHN	SYN	AVG	△	Caltech	Amazon	Webcam	DSLR	AVG	△
FedAvg [ASTAT17] [47]	98.14	90.85	76.56	55.01	80.14	-	60.15	75.44	45.86	36.00	54.36	-
FedProx [arXiv18] [40]	98.11	90.24	77.01	56.66	80.50	+0.36	60.21	77.44	48.62	37.33	55.90	+1.54
MOON [CVPR21] [38]	97.44	92.15	77.62	38.79	76.50	-3.64	56.19	71.54	41.04	30.22	49.74	-4.62
FedDyn [ICLR21] [1]	98.01	91.00	78.95	54.22	80.54	+0.40	61.64	75.54	48.28	35.56	55.25	+0.89
FedOPT [ICLR21] [58]	96.23	91.80	73.03	57.85	79.72	-0.42	56.31	56.74	63.33	48.89	56.31	+1.95
FedProc [arXiv21] [51]	97.86	88.99	78.90	45.84	77.89	-2.25	58.07	73.65	42.76	30.22	51.17	-3.19
FedProto [AAAI22] [70]	98.30	92.44	80.35	53.58	81.16	+1.02	64.02	79.37	50.17	40.33	58.47	+4.11
FPL	98.31	92.71	80.27	61.20	**83.12**	**+2.98**	63.39	79.26	55.86	48.00	**61.63**	**+7.27**

Table 2. **Comparison with the sota methods** on Digits and Office Caltech tasks. AVG denotes average accuracy calculated on all domains. See details in Sec. 4.3. Best in bold and second with underline. These notes are the same to others.

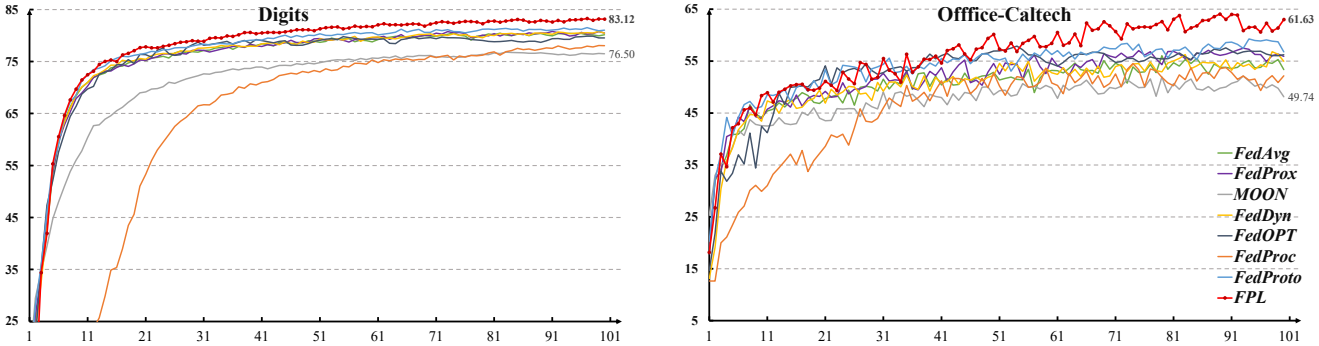


Figure 6. **Comparison of average accuracy on different communication epochs with counterparts** on Digits and Office Caltech tasks. Please see details in Sec. 4.3.

Prototype	Digits					
	MNIST	USPS	SVHN	SYN	AVG	△
$\mathcal{G}$	98.30	92.44	80.35	53.58	81.16	-
$\mathcal{U}$	98.23	93.12	81.18	55.40	**81.98**	**+0.82**
	Office Caltech					
	Caltech	Amazon	Webcam	DSLR	AVG	△
$\mathcal{G}$	64.02	79.37	50.17	40.33	58.47	-
$\mathcal{U}$	64.26	79.54	48.39	44.67	**59.21**	**+0.73**

Table 3. **Comparison of consistent regularization** with global prototypes ($\mathcal{G}$ in Eq. (3)) and unbiased prototypes ($\mathcal{U}$ in Eq. (6)) on Digits and Office Caltech tasks. See details in Sec. 4.2.2.

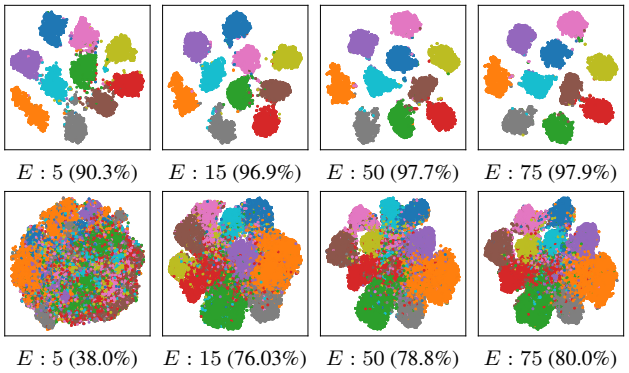


Figure 7. **t-SNE Visualization of FPL at different communication epoch** on randomly participants from MNIST (Top Row) and SVHN (Bottom Row). Please refer to Sec. 4.3 for details.

which depicts that FPL is feasible to learn a generalizable decision boundary. We draw the the average accuracy metric in each communication epoch during training phase in Fig. 6. We observe that FPL presents faster and stabler convergence speed than other methods in these two tasks.

5. Conclusion

In this paper, we explore the generalizability and stability problem under domain shift in heterogeneous federated learning. Our work introduces a simple yet effective federated learning algorithm, Federated Prototypes Learning (FPL). We leverage prototypes (class prototypical representation) to tackle these two problems by enjoying the complementary advantages of cluster prototypes and unbiased prototypes: diverse domain knowledge and stable convergence signal. The effectiveness of FPL has been thoroughly validated with many popular counterparts over various classification tasks. We wish this work to pave the way for future research on heterogeneous federated learning.

Acknowledgement. This work is partially supported by National Natural Science Foundation of China under Grant (62176188, 62225113), the Key Research and Development Program of Hubei Province (2021BAA187), Zhejiang lab (NO.2022NF0AB01), CCF-Huawei Populus Grove Fund (CCF-HuaweiTC2022003), the Special Fund of Hubei Luojia Laboratory (220100015) and the Science and Technology Major Project of Hubei Province (Next-Generation AI Technologies) under Grant (2019AEA170).

References

[1] Durmus Alp Emre Acar, Yue Zhao, Ramon Matas, Matthew Mattina, Paul Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *ICLR*, 2021. 2, 6, 8

[2] Mohammed H Alsharif, Abdullah Alrashoudi, Abdullah Alabdulwahab, Saleh A Alshebeili, and Usman Tariq. Collaborative federated learning for healthcare: Multi-modal covid-19 diagnosis at the edge. *IEEE ITJ*, 2021. 3

[3] Anonymous. Soft neighbors are positive supporters in contrastive visual representation learning. In *ICLR*, 2023. 3

[4] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *ACM-SIAM*, 2006. 4

[5] Christopher Briggs, Zhong Fan, and Peter Andras. Federated learning with hierarchical clustering of local updates to improve training on non-iid data. In *IJCNN*, pages 1–9, 2020. 3

[6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, pages 9912–9924, 2020. 3

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020. 2, 3, 7

[8] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. In *NeurIPS*, pages 8765–8775, 2020. 3

[9] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. *Ann. Oper. Res.*, pages 19–67, 2005. 6

[10] Moming Duan, Duo Liu, Xinyuan Ji, Renping Liu, Liang Liang, Xianzhang Chen, and Yujuan Tan. Fedgroup: Efficient clustered federated learning via decomposed data-driven measure. In *ISPA*, 2021. 3

[11] Richard O Duda, Peter E Hart, et al. *Pattern classification and scene analysis*, volume 3. Wiley New York, 1973. 2

[12] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *ICCV*, pages 9588–9597, 2021. 3

[13] 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*, 2022. 1, 2

[14] Songwei Ge, Shlok Mishra, Chun-Liang Li, Haohan Wang, and David Jacobs. Robust contrastive learning using negative samples with diminished semantics. In *NeurIPS*, 2021. 3

[15] Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. In *NeurIPS*, pages 19586–19597, 2020. 3

[16] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, pages 2066–2073, 2012. 2, 6

[17] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007. 6

[18] Yuanfan Guo, Minghao Xu, Jiawen Li, Bingbing Ni, Xuanyu Zhu, Zhenbang Sun, and Yi Xu. Hcsc: Hierarchical contrastive selective coding. In *CVPR*, pages 9706–9715, 2022. 3

[19] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 2, 3

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6

[21] Kevin Hsieh, Amar Phanishayee, Onur Mutlu, and Phillip Gibbons. The non-iid data quagmire of decentralized machine learning. In *ICML*, pages 4387–4398, 2020.

[22] Wenke Huang, Mang Ye, and Bo Du. Learn from others and be yourself in heterogeneous federated learning. In *CVPR*, 2022. 2, 6

[23] Jonathan J. Hull. A database for handwritten text recognition research. *IEEE PAMI*, pages 550–554, 1994. 2, 6

[24] Tri Huynh, Simon Kornblith, Matthew R Walter, Michael Maire, and Maryam Khademi. Boosting contrastive self-supervised learning with false negative cancellation. In *WACV*, pages 2785–2795, 2022. 3

[25] Saumya Jetley, Bernardino Romera-Paredes, Sadeep Jayasumana, and Philip Torr. Prototypical priors: From improving classification to zero-shot learning. *arXiv preprint arXiv:1512.01192*, 2015. 3

[26] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019. 1

[27] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. In *NeurIPS*, pages 21798–21809, 2020. 3

[28] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for on-device federated learning. In *ICML*, 2020. 1, 2

[29] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NeurIPS*, 2020. 7

[30] Sungnyun Kim, Gihun Lee, Sangmin Bae, and Se-Young Yun. Mixco: Mix-up contrastive learning for visual representation. In *NeurIPS Workshop*, 2020. 3

[31] Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. Mean shift for self-supervised learning. In *CVPR*, pages 10326–10335, 2021. 3

[32] Samuel Kotz, Narayanaswamy Balakrishnan, and Norman L Johnson. *Continuous multivariate distributions, Volume 1: Models and applications*, volume 1. John Wiley & Sons, 2004. 1

[33] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, pages 2278–2324, 1998. 2, 6

[34] Kibok Lee, Yian Zhu, Kihyuk Sohn, Chun-Liang Li, Jinwoo Shin, and Honglak Lee. i-mix: A domain-agnostic strategy for contrastive representation learning. In *ICLR*, 2021. 3

[35] Gen Li, Varun Jampani, Laura Sevilla-Lara, Deqing Sun, Jonghyun Kim, and Joongkyu Kim. Adaptive prototype learning and allocation for few-shot segmentation. In *CVPR*, pages 8334–8343, 2021. 3

[36] Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. In *ICLR*, 2021. 2, 3

[37] Qinbin Li, Bingsheng He, and Dawn Song. Adversarial collaborative learning on non-iid features. *arXiv*, 2021. 2

[38] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *CVPR*, pages 10713–10722, 2021. 1, 2, 6, 8

[39] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE SPM*, pages 50–60, 2020. 1

[40] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 2020. 1, 2, 3, 6, 8

[41] Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fed{bn}: Federated learning on non-{iid} features via local batch normalization. In *ICLR*, 2021. 2

[42] Niklaus Liu, Zhaonan Liang, Junyang Lin, and Yang Liu. Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. *Journal of biomedical informatics*, 99:103291, 2019. 3

[43] Guodong Long, Ming Xie, Tao Shen, Tianyi Zhou, Xianzhi Wang, and Jing Jiang. Multi-center federated learning: clients clustering for better personalization. *WWW*, 26(1):481–500, 2023. 3

[44] Mi Luo, Fei Chen, Dapeng Hu, Yifan Zhang, Jian Liang, and Jiashi Feng. No fear of heterogeneity: Classifier calibration for federated learning with non-iid data. In *NeurIPS*, 2021. 3

[45] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *BSMSP*, pages 281–297, 1967. 4

[46] Yishay Mansour, Mehryar Mohri, Jae Theertha Suresh Ro, and Ananda. Three approaches for personalization with applications to federated learning. *arXiv preprint arXiv:2002.10619*, 2020. 3

[47] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, pages 1273–1282, 2017. 1, 2, 3, 6, 8

[48] Pascal Mettes, Elise van der Pol, and Cees Snoek. Hyperspherical prototype networks. *Advances in neural information processing systems*, 32, 2019. 3

[49] Umberto Michieli and Mete Ozay. Prototype guided federated learning of visual feature representations. *arXiv preprint arXiv:2105.08982*, 2021. 3

[50] Jovana Mitrovic, Brian McWilliams, Jacob Walker, Lars Buesing, and Charles Blundell. Representation learning via invariant causal mechanisms. In *ICLR*, 2021. 3

[51] Xutong Mu, Yulong Shen, Ke Cheng, Xueli Geng, Jiaxuan Fu, Tao Zhang, and Zhiwei Zhang. Fedproc: Prototypical contrastive federated learning on non-iid data. *arXiv preprint arXiv:2109.12273*, 2021. 1, 2, 3, 6, 8

[52] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NeurIPS Workshop*, 2011. 2, 6

[53] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[54] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[55] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE TKDE*, pages 1345–1359, 2009. 2

[56] Xiangyu Peng, Kai Wang, Zheng Zhu, Mang Wang, and Yang You. Crafting better contrastive views for siamese representation learning. In *CVPR*, pages 16031–16040, 2022. 3

[57] Joaquin Quiñonero-Candela, Masashi Sugiyama, Neil D Lawrence, and Anton Schwaighofer. *Dataset Shift in Machine Learning*. Mit Press, 2009. 2

[58] Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and H Brendan McMahan. Adaptive federated optimization. In *ICLR*, 2021. 6, 8

[59] Sucheng Ren, Huiyu Wang, Zhengqi Gao, Shengfeng He, Alan Yuille, Yuyin Zhou, and Cihang Xie. A simple data mixing prior for improving self-supervised learning. In *CVPR*, pages 14595–14604, 2022. 3

[60] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *ICLR*, 2021. 3

[61] Prasun Roy, Subhankar Ghosh, Saumik Bhattacharya, and Umapada Pal. Effects of degradations on deep neural network architectures. *arXiv preprint arXiv:1807.10108*, 2018. 2, 6

[62] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*, pages 213–226, 2010. 6

[63] M. Saquib Sarfraz, Vivek Sharma, and Rainer Stiefelhagen. Efficient parameter-free clustering using first neighbor relations. In *CVPR*, pages 8934–8943, 2019. 4

[64] Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE TNNLS*, 32(8):3710–3722, 2020. 3

[65] Zhiqiang Shen, Zechun Liu, Zhuang Liu, Marios Savvides, Trevor Darrell, and Eric Xing. Un-mix: Rethinking image mixtures for unsupervised visual representation learning. In *AAAI*, pages 2216–2224, 2022. 3

[66] 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. In *NeurIPS Workshop*, 2019. 1, 2

[67] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017. 2, 3

[68] Canh T. Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with moreau envelopes. In *NeurIPS*, pages 21394–21405, 2020. 2

[69] Yue Tan, Yixin Liu, Guodong Long, Jing Jiang, Qinghua Lu, and Chengqi Zhang. Federated learning on non-iid graphs via structural knowledge sharing. In *AAAI*, 2023. 1

[70] Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. Fedproto: Federated prototype learning across heterogeneous clients. In *AAAI*, 2022. 1, 2, 3, 6, 8

[71] Yue Tan, Guodong Long, Jie Ma, Lu Liu, Tianyi Zhou, and Jing Jiang. Federated learning from pre-trained models: A contrastive learning approach. In *NeurIPS*, 2022. 3

[72] Ajinkya Tejankar, Soroush Abbasi Koohpayegani, Vipin Pillai, Paolo Favaro, and Hamed Pirsiavash. Isd: Self-supervised learning by iterative similarity distillation. In *ICCV*, 2021. 3

[73] Yonglong Tian, Olivier J Henaff, and Aäron van den Oord. Divide and contrast: Self-supervised learning from uncurated data. In *ICCV*, pages 10063–10074, 2021. 3

[74] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*, pages 776–794. Springer, 2020. 7

[75] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *ECCV*, 2020. 3

[76] Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *CVPR*, pages 2495–2504, 2021. 5, 7

[77] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *CVPR*, pages 9197–9206, 2019. 3

[78] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *JASA*, pages 236–244, 1963. 4

[79] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, pages 3733–3742, 2018. 2, 3

[80] Chenhao Xu, Youyang Qu, Yong Xiang, and Longxiang Gao. Asynchronous federated learning on heterogeneous devices: A survey. *arXiv preprint arXiv:2109.04269*, 2021. 3

[81] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Attribute prototype network for zero-shot learning. In *NeurIPS*, pages 21969–21980, 2020. 3

[82] Hong-Ming Yang, Xu-Yao Zhang, Fei Yin, and Cheng-Lin Liu. Robust classification with convolutional prototype learning. In *CVPR*, pages 3474–3482, 2018. 2

[83] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM TIST*, pages 1–19, 2019. 1

[84] Mang Ye, Jianbing Shen, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. Augmentation invariant and instance spreading feature for softmax embedding. *IEEE PAMI*, 2020. 2, 3

[85] Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *CVPR*, pages 6210–6219, 2019. 2, 3

[86] 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. 2

[87] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018. 1

[88] Mingkai Zheng, Shan You, Fei Wang, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Ressl: Relational self-supervised learning with weak augmentation. In *NeurIPS*, pages 2543–2555, 2021. 3

[89] Huasong Zhong, Jianlong Wu, Chong Chen, Jianqiang Huang, Minghua Deng, Liqiang Nie, Zhouchen Lin, and Xian-Sheng Hua. Graph contrastive clustering. In *CVPR*, pages 9224–9233, 2021. 3

[90] Tianfei Zhou, Wenguan Wang, Ender Konukoglu, and Luc Van Gool. Rethinking semantic segmentation: A prototype view. In *CVPR*, 2022. 2, 3

[91] Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. Prototype augmentation and self-supervision for incremental learning. In *CVPR*, pages 5871–5880, 2021. 2