Generalizable Heterogeneity-aware Federated Feature and Basic-matrix Consistency Learning

Xuan Lai, Luying Zhong, Tianying Lu, Junjie Zhang, Zhiqin Huang, and Zheyi Chen*

College of Computer and Data Science, Fuzhou University, Fuzhou 350116, China
241016004@fzu.edu.cn, {luyingzhongfzu, tianyinglufzu}@163.com, 241010005@fzu.edu.cn, zhiqinhuangfzu@163.com, z.chen@fzu.edu.cn

Abstract

As an emerging distributed learning paradigm, Federated Learning (FL) facilitates collaborative training among multiple clients without sharing raw data. However, the classic FL still faces significant challenges due to feature/model heterogeneity and catastrophic forgetting, which seriously hinder knowledge transfer and cause the forgetting of previous knowledge. To address these important challenges, we propose *FBCL*, a novel generalizable heterogeneity-aware Federated features and Basic-matrix Consistency Learning to balance intra-domain discriminability and inter-domain generalization. For feature/model heterogeneity, we align the similarity of feature distribution and construct the high-dimensional basic matrix with irrelevant unlabeled data, thereby overcoming communication barriers and learning generalizable representations while maintaining strict privacy preservation. For catastrophic forgetting during local updating, we introduce constraints in high-dimensional features to retain inter-domain knowledge and then extract accurate knowledge by distilling old models to preserve worthy historical information. Using real-world unlabeled public datasets, extensive experiments validate the superiority of the proposed *FBCL*, which outperforms the state-of-the-art methods on different scenarios of image classification.

Introduction

In recent years, the rapid advancement in Artificial Intelligence (AI) has significantly influenced various aspects of social life, such as healthcare, finance, and smart homes (Feng et al. 2023), while accompanied by serious issues of data privacy and security. To effectively address these issues, Federated Learning (FL) (McMahan et al. 2017) has emerged as a novel decentralized training paradigm, garnering considerable attention for its excellent ability to enable model collaborative training while safeguarding data privacy and security. The clients can only share their model parameters rather than the raw data, thereby mitigating the risk of data leakage. FL not only offers substantial advantages in privacy protection but also presents innovative solutions for collaborative efforts across many parties, such as healthcare and finance. For example, multiple medical institutions can collaboratively train a strong diagnostic model without compromising patient privacy, thereby improving diagnostic accuracy and medical plans while adhering to privacy regulations and ethical standards (Liu et al. 2022).

Although FL offers considerable benefits in protecting data privacy and security, it still encounters numerous challenges in practical applications. First and foremost, the data from different clients commonly exhibits huge heterogeneity, which implies that the data of different clients might possess distinct underlying distributions, and thus each client will optimize towards its local empirical optimum during the training process. Under such a situation, slow convergence, low accuracy, and even model collapse may happen in classic FL algorithms (e.g., FedAvg (McMahan et al. 2017)). Moreover, FL necessitates model updates on multiple devices with varying hardware capabilities and network conditions. Besides, clients typically aim to protect their proprietary information and data privacy, and they are often reluctant to share the details of model design, thereby further increasing the complexity and instability of the training process. Existing studies fall into two categories: 1) Model sharing (Diao, Ding, and Tarokh 2020; Ma et al. 2022; Tan et al. 2023) and 2) Knowledge transfer (Li, Li, and Varshney 2021; Chen et al. 2020; Itahara et al. 2021). As for the studies of model sharing, they usually require designing unique network architectures, incurring substantial communication costs. As for the studies of knowledge transfer, they commonly rely on data quality and the completeness of information transfer, which may inadvertently cause interference with other clients in some cases. It is noted that the latest research trend has predominantly focused on personalized FL (Tan et al. 2022a; Shen, Zhou, and Yu 2022; Zhang et al. 2023) with rare consideration of model generalization, which may detract from the broader applicability and robustness of FL across diverse scenarios.

In FL, the central server and clients conduct model updates via mutual communication. Due to the constraints on bandwidth and privacy, the number of communication rounds is often restricted. Thus, the model should flexibly adapt to various tasks within limited communication rounds. When making model transitions between different data distributions, model parameters need to be fine-tuned with constrained capacities. This may compromise their adaptability to old tasks, resulting in catastrophic forgetting (Tan et al. 2022a). Existing studies focus on parameter isolation (Bon-

*Corresponding author.

ato et al. 2024) and replay mechanism (Li et al. 2024), which significantly increase model complexity and storage resources. Therefore, it remains an open challenge to find efficient ways to mitigate the forgetting of old data distributions while learning new ones.

To address the above important challenges in FL, we propose *FBCL*, a novel Federated features and Basic-matrix Consistency Learning. *FBCL* leverages the accessible unlabeled public data to perform instance-level feature alignment, where client models map same-label data to consistent regions in a shared feature space, reducing inter-domain discrepancies. Further, we introduce feature basis-matrix consistency that enforces orthogonality between categorical features, establishing distinct feature-to-basis vector mappings to prevent interference while enhancing discriminability and generalization. Meanwhile, to mitigate catastrophic forgetting during local updates, we implement dual-consistency constraints: (i) Feature-space stabilization through $L2$-norm regularization limits representation drift; (ii) Distillation of non-target probabilities from prior models preserves global knowledge. Label supervision is maintained throughout to ensure intra-domain discriminability, collectively enabling balanced adaptation to new tasks without sacrificing prior knowledge. The main contributions of this work are summarized as follows.

- We propose a new heterogeneity-aware FL that employs Gaussian distributions to characterize sample features across different channels on irrelevant unlabeled public data. By aligning features across domains, an original cross-domain consistency learning is devised to facilitate communication between heterogeneous clients and learn feature representations with stronger generalization.
- We design a novel feature basis-matrix consistency learning mechanism to establish a well-structured feature space, encouraging features from distinct categories to be distributed along an orthogonal direction, thereby reducing inter-feature interference while improving feature distinguishability and generalization potential.
- We introduce an innovative dual-consistency constraint methodology to preserve the global perspective of old models during local updating. With label supervision, the knowledge of different models is balanced, relieving the issue of catastrophic forgetting and improving both the inter-domain and intra-domain performance.
- Using real-world unlabeled public datasets, extensive experiments validate the superiority of the proposed *FBCL* in intra-domain and inter-domain performance, which outperforms the state-of-the-art methods on different image classification tasks. Moreover, the ablation study verifies the effectiveness of core components in *FBCL*, exhibiting its robustness and utility in diverse scenarios.

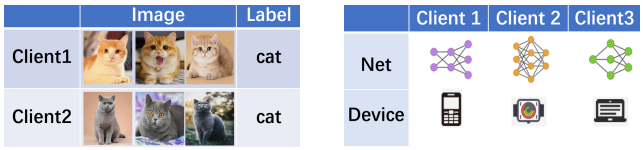
Related Work

Federated Knowledge Distillation

Recently, Federated Knowledge Distillation (FKD) has emerged as an influential advancement in FL, garnering much attention. FedMD (Li and Wang 2019) utilized labeled proxy data and logits for knowledge distillation, enabling knowledge transfer across models. Cronus (Hongyan et al. 2019) aggregated logits to improve robustness against outliers and noise by using a mean estimation on the server side. CFD (Sattler et al. 2021) introduced a quantization mechanism and incremental encoding method to compress logits for reducing communication overheads. FedGEN (Zhu, Hong, and Zhou 2021) executed statistical heterogeneous FL by ensembling client information via a data-free knowledge distillation. FedGKT (He, Annavaram, and Avestimehr 2020) alleviated training burden by transferring knowledge periodically from client-side small models to the server-side large model. FedFTG (Zhang et al. 2022) trained a conditional generator to fit the input space of local models for generating pseudo-data and then transferred local knowledge to the global model by minimizing the Kullback-Leibler (KL) divergence between predicted outputs. Generally, FKD commonly relies on the quality of shared proxy data and extra data generators. However, the transmission of unprocessed logits causes privacy leakage of local data, and the high-dimensional semantic information makes it hard to capture rich local information in feature tilt scenarios, causing confusing directions of gradient updates. Moreover, extra data generators boost the difficulty of parameter tuning and the possibility of model collapses while increasing computation and communication overheads. Different from existing studies, the proposed *FBCL* utilizes unlabeled public data for knowledge transfer, avoiding the dependency on labeled private data while alleviating privacy leakage and distribution mismatch.

Federated Catastrophic Forgetting

The integration of FL with continual learning has led to the development of Federated Continual Learning, aiming to address the problem of catastrophic forgetting in classic FL. FedPer (Xu, Yan, and Huang 2022) decomposed the model into basic and personalized layers, aggregating shallow basic layers to capture generalized knowledge while retaining deep personalized layers to maintain clients' personalized knowledge. FedWeIT (Yoon et al. 2021) decomposed the local model parameters of clients into dense basic parameters and sparse task-adaptive parameters, aiming to achieve more efficient communication. GLFC (Dong et al. 2022) approached from both global and local perspectives, focusing on addressing the problem of catastrophic forgetting by training a global incremental model to support continual learning. FedKNOW (Luopan et al. 2023) integrated the gradients of new and old tasks with the gradients before and after global aggregation. CRSS (Good et al. 2023) replayed samples to enhance the diversity of loss gradients and introduced auxiliary variables to coordinate client selection by making the objectives of clients separable. In general, the existing studies usually require specific configurations for client models or incur high communication costs due to the massive transmission of extra parameters or models. Different from the existing studies, the proposed *FBCL* unifies feature representations under the scenarios of model heterogeneity, mitigating the problem of catastrophic forgetting in distributed environments while maintaining low communi-



(a) Feature heterogeneity. (b) Model heterogeneity.

Figure 1: Examples of feature and model heterogeneity.

cation overheads.

Problem Formulation

In the general FL, there are K participating clients. For the private dataset $D_k = (x_i^k, y_i^k)_{i=1}^{N_k}$ on the k-th client C_k, the total number of samples is denoted as N_k and the predicted output of the sample x_i^k on the local model θ_k is denoted as $\bar{y}_i^k$, where $P_k(X,Y)$ indicates the data distribution. As shown in Figure 1, we specify the problems of feature and model heterogeneity in FL as follows.

- **Feature Heterogeneity.** The data collected by clients often exhibit significant variability due to different scenarios from the perspectives of lighting, background, and resolution, leading to diverse feature distributions across clients. Therefore, the feature heterogeneity can be formulated as $P_k(X) \neq P_l(X), \forall k = l$.
- **Model Heterogeneity.** There is significant diversity in device types and personalized demands across different clients, and clients often adopt distinct local model structures tailored to their specific scenarios. Therefore, the model heterogeneity can be formulated as $shape\{\theta_k\} \neq shape\{\theta_l\}, \forall k = l$.

Methodology

As illustrated in Figure 2, the proposed *FBCL* operates through two integrated phases. During collaborative updating, we leverage public data to perform feature instance alignment and represent features through unified basis matrices via Feature Instance-level Alignment (FIA) and Feature Basic-Matrix consistency (FBM), reducing model redundancy while enhancing cross-domain adaptability. This approach minimizes privacy risks and communication overhead inherent in conventional FL by enabling secure knowledge transfer across clients. Transitioning to local updating, *FBCL* employs Correct Not-target Distillation (CND) with Feature Relationship Regularization (FRR) to balance intra/inter-domain information, specifically mitigating catastrophic forgetting under data heterogeneity by preserving prior cross-domain knowledge while adapting to new local distributions. The key steps are presented in Algorithm 1.

Feature Instance-level Alignment

Inspired by the Auto-Encoder (He et al. 2022), the proper feature distribution contains rich information that can replace the direct transmission of model parameters, thereby reducing communication overheads. We split the local model θ_k into a feature extractor f_k and a classifier ψ_k. Different from prototype-based methods(Tan et al. 2022b), which learn domain-invariant representations leaking class-level statistics, we utilize public data D_p to approximate feature extraction capability across models. To address model heterogeneity where feature extractors output different dimensions, we add a mapping branch g_k (i.e., 1×1 $conv$) for client k, to convert features into the same high-dimensional space. For public data $x \in D_p$, features extracted from different clients should be mapped to similar positions as

$$\min_{h_k} \|h_k(x) - \bar{h}(x)\|_F, \forall x \in D_p, \tag{1}$$

where $h_k(x) = g_k(f_k(x))$, $\bar{h}(x) = \frac{1}{K}\sum_{i=1}^{K} h_i(x)$ is the average of the sample x after being mapped by all clients. $\|\cdot\|_F$ measures the similarity between two feature outputs.

Considering conventional pooling operations discard substantial spatial information, we align feature distributions channel-wise to capture intricate activation patterns across the entire feature map. It is limited to comparing the feature offset of individual channels by directly calculating the Euclidean distance, lacking a mechanism for cross-sample comparison. Meanwhile, deep features of images are often high-dimensional and sparse, making the KL divergence unsuitable when distributions do not overlap. To this end, we adopt the Gaussian distribution to model the feature vectors extracted from the high-dimensional space, defined as

$$N(\boldsymbol{\mu}, \boldsymbol{\delta}) = \frac{1}{\sqrt{2\pi\,|\boldsymbol{\delta}|}} e^{-\frac{1}{2}(\boldsymbol{h}-\boldsymbol{\mu})^T \boldsymbol{\delta}^{-1}(\boldsymbol{h}-\boldsymbol{\mu})}, \tag{2}$$

where $\boldsymbol{\delta} = \frac{1}{m}\sum_{i=1}^{m} (\boldsymbol{h_i} - \boldsymbol{\mu})(\boldsymbol{h_i} - \boldsymbol{\mu})^T$ is the covariance matrix, $\boldsymbol{\mu} = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{h_i}$ is the mean vector, $m = height \times width$ is the size of feature map, and $|\boldsymbol{\delta}|$ is the matrix determinant.

By representing features in the above probabilistic manner, we can better align and compare features from different clients, contributing to more consistent and robust feature representations. Therefore, the loss of FIA is defined as

$$L_{FIA} = \|\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}\|_2^2 + \left\|\boldsymbol{\delta} - \bar{\boldsymbol{\delta}}\right\|_2^2. \tag{3}$$

Feature Basic-matrix Consistency

Relying solely on average features for model updating risks convergence to outliers (e.g., trivial solutions like $g_k = 0$ mapping all inputs to identical positions). To capture diverse invariant features and mitigate feature redundancy, we will maximize the differentiation between features of different samples while minimizing the distance for features of the same sample. Since public data lacks labels, it is infeasible to directly assess the distance between different samples in a high-dimensional space. Instead, we measure the positional relationship of samples through the feature direction. The feature vectors of each sample batch can form a feature matrix. If the feature vectors are orthogonal to each other and constitute a subset of the solutions of the basic matrix, the difference between features is maximized. The difference

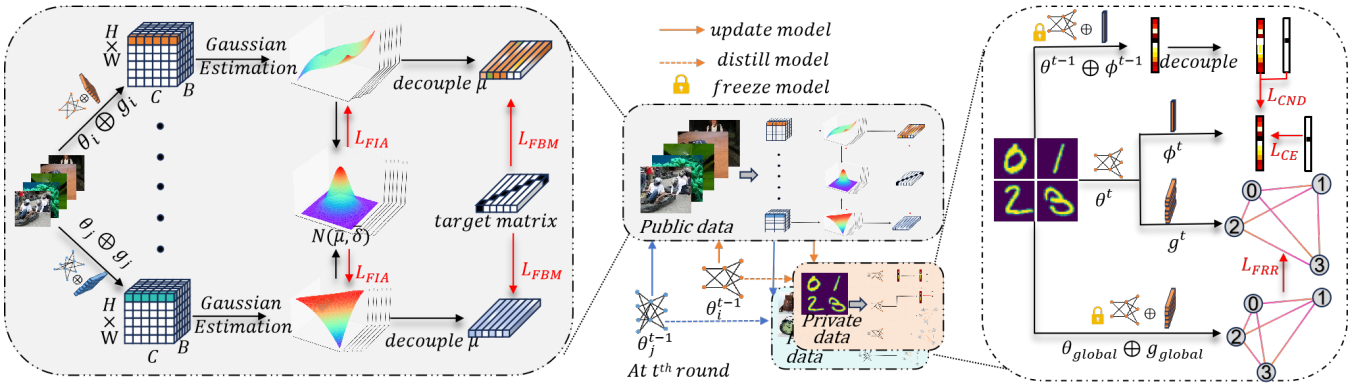


Figure 2: Overview of the proposed *FBCL* for solving the issues of heterogeneity and catastrophic forgetting in FL.

Algorithm 1: the proposed $FBCL$

Input: Communication rounds T, number of participating clients K, unlabeled public data $D_p(X_p)$, i-th private data $D_i(X_i, Y_i)$, batch size B, hyper-parameters λ, α, β.
Output: θ_i^t of i-th participating client.
for $t = 1$ to T **do**
  # **Global updating**
  for $k = 1$ to K **do**
    Calculate $N(\mu_k^t, \delta_k^t)$ for $x \in D_p$ by Eq.(2)
    Close $N(\mu_k^t, \delta_k^t)$ to other clients by Eq.(3)
    Make $\{\mu_k^t\}_{i=1}^B$ unrelated by Eq.(4)
    Update $\theta_{k,global}^t \leftarrow \theta_i^t - \eta \nabla L_{col}$ by Eq.(5)
  end for
  #**Local updating**
  for $k = 1$ to K **do**
    Distill historical knowledge from Z_k^{t-1} by Eq.(6)
    Close relationship of $\{\mu_k^t\}_{i=1}^B$ to global by Eq.(7)
    Update $\theta_k^{t+1} \leftarrow \theta_{k,global}^t - \eta \nabla L_{loc}$ by Eq.(9)
  end for
end for

degree between the u-th and v-th samples in each batch is denoted as $\frac{|\boldsymbol{\mu}_u^T \boldsymbol{\mu}_v|}{\|\boldsymbol{\mu}_u\| \|\boldsymbol{\mu}_v\|}$, with a value range from 0 (completely unrelated) to 1 (completely consistent).

Since the normal vector of each feature in the high-dimensional space is not unique, when the batch size B is less than the dimension D, a direction should be set to approximate the true direction of feature vectors. For simplicity, we regard the average direction of features as the preset direction. The difference between the mapped feature of the u-th sample and the preset direction is quantified as $\frac{|\boldsymbol{\mu}_u^T \bar{\boldsymbol{\mu}}_u|}{\|\boldsymbol{\mu}_u\| \|\bar{\boldsymbol{\mu}}_u\|}$. Because this optimization direction is consistent with Eq. (3), it is negligible. Therefore, the loss of FBM is defined as

$$L_{FBM} = \frac{1}{B(B-1)} \sum_{u \neq v} \frac{|\boldsymbol{\mu}_u^T \boldsymbol{\mu}_v|}{\|\boldsymbol{\mu}_u\| \|\boldsymbol{\mu}_v\|}. \tag{4}$$

The loss of FBM encourages the orthogonality and diversity of features, enabling *FBCL* to learn comprehensive and robust representations. Furthermore, by combining L_{FIA} and L_{FBM}, the overall loss of collaborative is updated by

$$L_{col} = L_{FIA} + \lambda L_{FBM}. \tag{5}$$

Correct Not-target Distillation

While soft labels from old models provide richer inter-class relational information than hard labels, enhancing knowledge transfer under complex distributions, their inconsistency and potential misalignment with true labels risk propagating errors during distillation. To address this issue, we propose a new CND that calculates the probability distribution by only using the "non-target class" information during the training process, thereby minimizing the negative impact of erroneous labels. Then, we set weight parameters to the prediction probability of corresponding labels. Therefore, the loss of CND for the sample (x, y) is defined as

$$L_{CND} = \gamma \sum_{i \neq y} KL(P_{old}(i|x), P_{cur}(i|x)), \tag{6}$$

where γ is the confidence parameter for historical knowledge. If the output of the old model corresponds exactly to the label y, $\gamma = P_{old}(y|x)$; Otherwise, $\gamma = 0$. $P_{old}(y|x)$ is the soft label distribution of the old model for y, $P_{cur}(i|x)$ is the soft label distribution of the current model for i, and KL is the KL divergence.

Feature Relationship Regularization

Relying solely on local supervised learning risks overfitting to current data distributions, degrading inter-domain performance. To prevent feature representation drift, we apply the corresponding global model $\theta_{i,global}^{(t-1)}$ as a teacher to guide

local updating. Since the global model acts as a feature extractor and does not output specific logits, classic logit-based knowledge distillation strategies cannot be applied directly. To realize effective knowledge transfer between the global and local models, we constrain the changes in local models via a feature relationship matrix ϕ (Detailed design is given in Appendix A), which is defined as

$$L_{FRR} = l(\phi(h(\theta_{i,global}^{(t-1)}, X)), \phi(h(\theta_i^{(t)}, X))), \quad (7)$$

This consistency constraint ensures that the distribution of feature representations in the high-dimensional space remains invariant across both source and target domains, thereby circumventing the over-reliance on data from a particular domain. It enhances the model generalization across diverse domains while alleviating the impact of catastrophic forgetting, thus enabling more efficient knowledge transfer and sharing among clients.

Intra-domain Discrimination

In addition to ensuring the similarity between local and global features, we expect the classification features of the local model to be sufficiently discriminative to distinguish different classes in the feature space. To this end, we retain the label loss, which is defined as

$$L_{CE} = -\sum_{c=1}^{C} y_i log(p_i). \quad (8)$$

Therefore, the overall loss of local is updated by

$$L_{loc} = L_{CE} + \alpha L_{CND} + \beta L_{FRR}. \quad (9)$$

Global Generalization Bound

Considering the target domain to be the set of all client data (i.e., $D_t = \cup_{k=1}^{K} D_k$), there exists an upper bound on the generalization error of the global model.

Theorem 1. (Upper Bound on Global Generalization Error) When the global classifier ψ^* and feature transformers $\{g_k\}$ satisfy the Lipschitz continuity and the distribution discrepancy between the public dataset D_p and the target domain D_t is measured by $d_H(D_p, D_t)$, the generalization error of the global model is bounded by

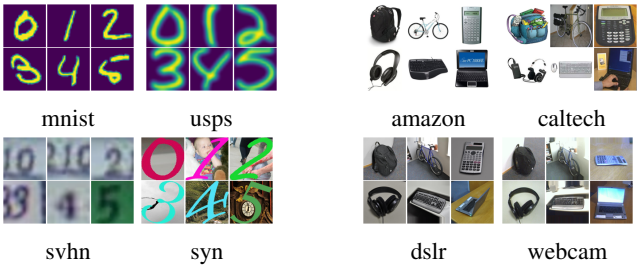
$$\begin{aligned} \epsilon_g(h^*, \psi^*) \leq \frac{1}{K}\sum_{k=1}^{K} \epsilon_k(f_k, \psi_k) + \rho L_\psi L_g R_g \sqrt{\Delta z_g} \\ + \lambda \cdot d_H(D_p, D_t) + O\left(\sqrt{\frac{\log K}{nK}}\right), \end{aligned} \quad (10)$$

where Δz_g is the feature alignment loss and n is the minimum size of samples per client. The upper bound indicates that the global performance is jointly determined by the client-specific error, alignment error, domain discrepancy, and statistical term. Detailed proof is given in Appendix.

Performance Evaluation

Experiment Setup

Datasets and Backbones. We evaluate the proposed *FBCL* on two real-world datasets from Digits (LeCun et al. 1998; Netzer et al. 2011; Roy et al. 2018; Hull 1994) and Office (Gong et al. 2012). The Digits dataset encompasses four domains: mnist, usps, svhn, and syn, where each contains 10 classes of digits 0-9. We randomly select 1% of the samples per class in each domain to form the training set while keeping the testing set unchanged. For the Office dataset, we randomly select 30% of the samples per class in each domain as the testing set while randomly selecting 50% of the remaining samples as the training set. Based on the above setup, each client is assigned data from one domain exclusively, without overlapping with data from other domains. As illustrated in Figure 3, the datasets from different domains exhibit significant feature heterogeneity. For the public dataset, we adopt the datasets of CIFAR and ImageNet, where 5,000 images are randomly selected. To verify the robustness and generalizability of *FBCL*, we consider various mainstream model structures as backbones, including ResNet (He et al. 2016), EfficientNet (Tan and Le 2019), and MobileNet (Howard 2017). Table 1 summarizes the settings of model structures for different datasets.



(a) Digits. (b) Office-Caltech.

Figure 3: Feature heterogeneity on Digits and Office-Caltech.

Dataset	Structure	Dataset	Structure
mnist	Resnet10	amazon	Mobilenetv2
usps	Resnet12	caltech	Resnet12
svhn	efficientnet	dslr	Resnet10
syn	Mobilenetv2	webcam	Resnet12

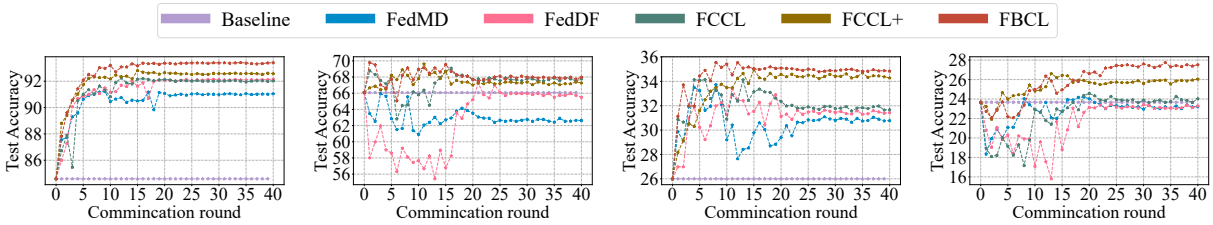
Table 1: Settings of model structures for different datasets

Comparison Methods. We compare the proposed *FBCL* with the following state-of-the-art methods: 1) FedMD (Li and Wang 2019) relies on relevant public data for knowledge distillation; 2) FedDF (Lin et al. 2020) integrates knowledge distillation with unlabeled data for model fusion; 3) FCCL (Huang, Ye, and Du 2022) constructs a cross-correlation matrix to align unlabeled data and extract cross-domain knowledge; and 4) FCCL+ (Huang et al. 2023) adds instance-similarity distribution alignment and extracts cross-domain information through posterior class relations. For model homogeneity, we further compare *FBCL* with FedFSA (Qi et al. 2025), which learns a unified feature space for clients to bridge inconsistency.

Implementation Details. We implement FBCL using PyTorch on one NVIDIA RTX 3090Ti GPU. The code has

Method	Digits					Office-Caltech				
	mnist	usps	svhn	syn	AVG	amazon	caltech	dslr	webcam	AVG
Baseline	98.67	94.87	82.41	62.40	84.59	52.06	45.18	86.11	80.95	66.08
FedMD	99.13	96.21	86.09	82.75	91.05	57.22	42.54	77.78	73.02	62.64
FedDF	99.08	94.11	85.91	89.50	92.15	62.98	43.93	80.33	74.78	65.51
FCCL	99.08	94.76	86.22	88.05	92.03	67.53	45.61	83.33	74.60	67.77
FCCL+	99.02	95.96	86.06	89.05	92.52	62.37	45.61	83.33	77.78	67.27
FBCL	99.12	95.71	87.22	90.35	**93.10**	65.46	46.93	83.33	76.19	**67.98**

Table 2: Comparison of intra-domain performance on different datasets with CIFAR-100



(a) Intra-Acc in Digits. (b) Intra-Acc in Office-Caltech. (c) Inter-Acc in Digits. (d) Inter-Acc in Office-Caltech.

Figure 4: Comparison of average intra-domain and inter-domain accuracy on Digits and Office-Caltech with CIFAR-100.

Method	Digits					Office-Caltech				
	mnist+	usps+	svhn+	syn+	AVG	amazon+	caltech+	dslr+	webcam+	AVG
Baseline	14.08	15.05	43.95	30.91	26.00	15.35	35.53	20.41	23.39	23.67
FedMD	23.71	10.32	49.07	40.01	30.78	15.12	32.00	23.47	22.20	23.20
FedDF	23.98	12.06	49.60	40.11	31.44	17.95	32.36	17.77	25.08	23.29
FCCL	18.30	15.23	49.21	43.94	31.67	20.80	35.10	20.61	19.56	24.02
FCCL+	15.23	30.38	42.25	49.93	34.45	19.61	42.00	18.36	24.22	26.00
FBCL	17.64	24.57	49.92	47.24	**34.84**	25.15	41.18	20.19	23.54	**27.51**

Table 3: Comparison of inter-domain performance on different datasets with CIFAR-100 (e.g., "mnist+" indicates that "mnist" is private data and the model is tested on all other domains except it)

been included in the supplementary materials. For fairness, we use the same parameter settings for both the proposed *FBCL* and state-of-the-art methods. Specifically, the number of clients is set to 4 for all datasets. The Adam optimizer is used with a learning rate of 0.001. We set the hyper-parameters $\lambda = 0.05$, $\alpha = 3$, and $\beta = 0.1$. The number of unlabeled public data is 5,000 for CIFAR-100. During local updating, the batch size is set to 128 with 40 iterations. During collaborative updating, the batch size is set to 256. For pre-processing, we resize all input images into $32 \times 32 \times 3$ for compatibility without data augmentation and train the baseline optimized on private data only for 100 iterations.

Performance Metrics. The following commonly-used performance metrics are used to evaluate model accuracy.

$$\mathcal{A}_i^{intra} = \frac{\sum(argmax(f_i(X_i^{Test})) == Y_i^{Test})}{|D_i|}, \quad (11)$$

$$\mathcal{A}_i^{inter} = \sum_{i \neq j} \frac{(argmax(f_i(X_j^{Test})) == Y_j^{Test})}{(K-1)|D_j|}, \quad (12)$$

where $|D_j|$ is the number of testing samples on j-th client.

Experiment Results and Analysis

Intra-domain Performance. As illustrated in Table 2, Baseline is optimized solely with local data without the benefit of collaborative learning with other domains, exhibiting the worst performance on the simple digits dataset while obtaining comparable performance on the complex office dataset. This suggests other algorithms may have failed to effectively extract domain-invariant features, indicating that blind joint communication across domains can be counterproductive rather than beneficial. In contrast, the proposed *FBCL* achieves the best or near-best performance across all datasets. As shown in Figure 4, *FBCL* converges with fewer communication rounds and exhibits minimal oscillation thereafter, demonstrating its efficiency and stability.

Inter-domain Performance. To evaluate the effectiveness of different methods in addressing catastrophic forgetting, we analyze their inter-domain performance on different datasets. As depicted in Table 3, Baseline underperforms in all cases, indicating that communication generally has a positive impact. FCCL+ and *FBCL* significantly outperform

Method	amazon	caltech	dslr	webcam	AVG	amazon+	caltech+	dslr+	webcam+	AVG
FedAvg	76.80	54.39	86.11	85.71	75.75	61.20	79.09	53.40	64.67	64.59
FedMD	78.87	52.63	80.56	84.13	74.05	57.62	74.36	59.94	67.41	64.83
FedDF	77.84	55.70	86.11	85.71	76.34	67.04	77.58	45.78	66.93	64.33
FCCL	77.32	50.88	91.67	80.95	75.21	62.00	71.91	54.90	61.98	62.69
FCCL+	83.51	56.14	86.11	80.95	76.68	66.16	76.26	55.47	66.15	66.01
FedPSA	80.41	54.39	86.11	85.71	76.66	66.16	78.51	62.54	67.99	68.80
FBCL	78.35	57.02	86.11	88.89	**77.59**	63.90	76.75	62.67	72.62	**68.99**

Table 4: Comparison with the state-of-the-art methods under model homogeneity on Office-Caltech with CIFAR-100

FIA	FBM	CND	FRR	mnist	usps	svhn	syn	AVG	mnist+	usps+	svhn+	syn+	AVG
				98.67	94.87	82.41	62.40	84.59	14.08	15.05	43.95	30.91	26.00
√				98.99	96.01	85.97	82.60	90.89	24.43	18.37	45.25	32.59	30.16
√	√			99.10	96.11	85.61	82.30	90.78	26.42	17.44	46.51	36.73	31.77
√	√	√		99.14	95.57	86.74	89.30	92.69	14.14	23.97	50.62	41.78	32.63
√	√		√	99.12	95.42	85.07	90.35	92.49	23.61	17.36	46.77	42.85	32.65
√	√	√	√	99.12	95.71	87.22	90.35	93.10	17.64	24.57	49.92	47.24	34.84

Table 5: Ablation study for core components in *FBCL* on Digits with CIFAR-100

other methods, attributed to their unique designs targeting catastrophic forgetting, demonstrating that preserving global model information substantially enhances cross-domain performance. Notably, except for the Office-31 dataset, *FBCL* slightly outperforms FCCL+, particularly achieving an approximate 1.46% higher performance on the Office-Caltech dataset. The results underscore the good robustness of *FBCL* in maintaining high performance across different domains and alleviating the impact of catastrophic forgetting.

Performance under Model Homogeneity. We compare *FBCL* with other methods under model homogeneity. We set the shared model as ResNet-18 and add the averaging parameters operation between collaborative updating and local updating. The Table 4 presents both inter-domain and intra-domain performance on Office-Caltech with CIFAR-100. It can be observed that merely using logits knowledge for guidance achieves results comparable to FedAvg, while algorithms leveraging feature-level knowledge demonstrate notable improvements. Prototype-based methods like FedFSA perform well in label-skewed settings partitioned from a single dataset, yet slightly underperform in feature-skewed settings with stylistic heterogeneity – the primary focus of this study. Crucially, our FBCL approach achieves competitive results without leaking local feature prototypes.

Ablation Study. We conduct ablation experiments to evaluate the effectiveness of core components in the proposed *FBCL*, including Feature Instance-level Alignment (FIA), Feature Basic-Matrix consistency (FBM), Correct Not-target Distillation (CND), and Feature Relationship Regularization (FRR). Table 5 illustrates the intra-domain and inter-domain performance with these components. Specifically, the first row refers to the model that is independently trained on private data. Both FIA and FBM contribute to better intra-domain and inter-domain performance. This is because FIA can align feature instances, and FBM ensures consistency of the feature basic matrix. CND enhances intra-domain accuracy but leads to decreasing inter-domain performance in some scenarios. This is because the teacher model focuses on retaining the local private data distribution rather than offering guidance based on the global data distribution. Therefore, CND is effective for intra-domain generalization, but it might introduce biases when the model is applied across domains. In contrast, FRR excels in improving inter-domain performance. By constraining the changes in model representations, FRR effectively mitigates domain shifts, leading to better generalization to other domains. When combining CND and FRR, *FBCL* achieves improvement in both intra-domain and inter-domain performance, surpassing the gains only obtained by feature exchange.

Performance under Various Scenarios. We validate the superior intra-domain and inter-domain performance of *FBCL* under various scenarios, including diverse model structures, varying feature dimensions, and different data types and volumes. Detailed analysis is given in Appendix.

Conclusion

In this paper, we propose *FBCL*, a novel generalizable heterogeneity-aware FL framework, to solve the issues of heterogeneity and catastrophic forgetting. In *FBCL*, we first align instance feature distribution and construct the basic matrix to extract client-invariant features on unlabeled public data. Next, we preserve accurate historical knowledge and feature relationship information to balance the inter-domain and intra-domain performance. Extensive experiments on real-world datasets of image classification demonstrate that *FBCL* significantly surpasses the state-of-the-art methods in discriminability and generalization under various scenarios. Moreover, the ablation study validates the effectiveness of the core components designed in *FBCL*.

Acknowledgments

This work was partly supported by the National Natural Science Foundation of China under Grant No. 62202103, the Natural Science Foundation of Fujian Province for Distinguished Young Scholars under Grant No. 2025J010020, the Central Funds Guiding the Local Science and Technology Development under Grant No. 2022L3004, the Fujian Province Technology and Economy Integration Service Platform under Grant No. 2023XRH001, and the Fuzhou-Xiamen-Quanzhou National Independent Innovation Demonstration Zone Collaborative Innovation Platform under Grant No. 2022FX5.

References

Bonato, J.; Pelosin, F.; Sabetta, L.; and Nicolosi, A. 2024. MIND: Multi-Task Incremental Network Distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 11105–11113.

Chen, Y.; Qin, X.; Wang, J.; Yu, C.; and Gao, W. 2020. Fedhealth: A federated transfer learning framework for wearable healthcare. *IEEE Intelligent Systems*, 35(4): 83–93.

Diao, E.; Ding, J.; and Tarokh, V. 2020. Heterofl: Computation and communication efficient federated learning for heterogeneous clients.

Dong, J.; Wang, L.; Fang, Z.; Sun, G.; Xu, S.; Wang, X.; and Zhu, Q. 2022. Federated class-incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10164–10173.

Feng, T.; Bose, D.; Zhang, T.; Hebbar, R.; Ramakrishna, A.; Gupta, R.; Zhang, M.; Avestimehr, S.; and Narayanan, S. 2023. Fedmultimodal: A benchmark for multimodal federated learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4035–4045.

Gong, B.; Shi, Y.; Sha, F.; and Grauman, K. 2012. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE conference on computer vision and pattern recognition*, 2066–2073. IEEE.

Good, J.; Majmudar, J.; Dupuy, C.; Wang, J.; Peris, C.; Chung, C.; Zemel, R.; and Gupta, R. 2023. Coordinated replay sample selection for continual federated learning.

He, C.; Annavaram, M.; and Avestimehr, S. 2020. Group knowledge transfer: Federated learning of large cnns at the edge. volume 33, 14068–14080.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hongyan, C.; Virat, S.; Reza, S.; and Amir, H. 2019. Cronus: Robust and heterogeneous collaborative learning with black-box knowledge transfer. *arXiv preprint arXiv:1912.11279*.

Howard, A. G. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Huang, W.; Ye, M.; and Du, B. 2022. Learn from others and be yourself in heterogeneous federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10143–10153.

Huang, W.; Ye, M.; Shi, Z.; and Du, B. 2023. Generalizable heterogeneous federated cross-correlation and instance similarity learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Hull, J. J. 1994. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5): 550–554.

Itahara, S.; Nishio, T.; Koda, Y.; Morikura, M.; and Yamamoto, K. 2021. Distillation-based semi-supervised federated learning for communication-efficient collaborative training with non-iid private data. *IEEE Transactions on Mobile Computing*, 22(1): 191–205.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Li, C.; Li, G.; and Varshney, P. K. 2021. Decentralized federated learning via mutual knowledge transfer. *IEEE Internet of Things Journal*, 9(2): 1136–1147.

Li, D.; and Wang, J. 2019. Fedmd: Heterogenous federated learning via model distillation. *arXiv preprint arXiv:1910.03581*.

Li, Y.; Li, Q.; Wang, H.; Li, R.; Zhong, W.; and Zhang, G. 2024. Towards Efficient Replay in Federated Incremental Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12820–12829.

Lin, T.; Kong, L.; Stich, S. U.; and Jaggi, M. 2020. Ensemble distillation for robust model fusion in federated learning. volume 33, 2351–2363.

Liu, R.; Wu, F.; Wu, C.; Wang, Y.; Lyu, L.; Chen, H.; and Xie, X. 2022. No one left behind: Inclusive federated learning over heterogeneous devices. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3398–3406.

Luopan, Y.; Han, R.; Zhang, Q.; Liu, C. H.; Wang, G.; and Chen, L. Y. 2023. Fedknow: Federated continual learning with signature task knowledge integration at edge. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, 341–354. IEEE.

Ma, X.; Zhang, J.; Guo, S.; and Xu, W. 2022. Layer-wised model aggregation for personalized federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10092–10101.

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

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; Ng, A. Y.; et al. 2011. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep*

learning and unsupervised feature learning, volume 2011, 4. Granada.

Qi, Z.; Meng, L.; Li, Z.; Hu, H.; and Meng, X. 2025. Cross-Silo Feature Space Alignment for Federated Learning on Clients with Imbalanced Data. In Walsh, T.; Shah, J.; and Kolter, Z., eds., *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, 19986–19994. AAAI Press.

Roy, P.; Ghosh, S.; Bhattacharya, S.; and Pal, U. 2018. Effects of degradations on deep neural network architectures. *arXiv preprint arXiv:1807.10108*.

Sattler, F.; Marban, A.; Rischke, R.; and Samek, W. 2021. CFD: Communication-efficient federated distillation via soft-label quantization and delta coding. *IEEE Transactions on Network Science and Engineering*, 9(4): 2025–2038.

Shen, Y.; Zhou, Y.; and Yu, L. 2022. Cd2-pfed: Cyclic distillation-guided channel decoupling for model personalization in federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10041–10050.

Tan, A. Z.; Yu, H.; Cui, L.; and Yang, Q. 2022a. Towards personalized federated learning. *IEEE transactions on neural networks and learning systems*, 34(12): 9587–9603.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, 6105–6114. PMLR.

Tan, Y.; Liu, Y.; Long, G.; Jiang, J.; Lu, Q.; and Zhang, C. 2023. Federated learning on non-iid graphs via structural knowledge sharing. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 9953–9961.

Tan, Y.; Long, G.; Liu, L.; Zhou, T.; Lu, Q.; Jiang, J.; and Zhang, C. 2022b. FedProto: Federated Prototype Learning across Heterogeneous Clients. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, 8432–8440. AAAI Press.

Xu, J.; Yan, Y.; and Huang, S.-L. 2022. FedPer++: toward improved personalized federated learning on heterogeneous and imbalanced data. In *2022 International Joint Conference on Neural Networks (IJCNN)*, 01–08. IEEE.

Yoon, J.; Jeong, W.; Lee, G.; Yang, E.; and Hwang, S. J. 2021. Federated continual learning with weighted inter-client transfer. In *International Conference on Machine Learning*, 12073–12086. PMLR.

Zhang, L.; Shen, L.; Ding, L.; Tao, D.; and Duan, L.-Y. 2022. Fine-tuning global model via data-free knowledge distillation for non-iid federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10174–10183.

Zhang, R.; Fan, Z.; Xu, Q.; Yao, J.; Zhang, Y.; and Wang, Y. 2023. Grace: A generalized and personalized federated learning method for medical imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 14–24. Springer.

Zhu, Z.; Hong, J.; and Zhou, J. 2021. Data-free knowledge distillation for heterogeneous federated learning. In *International conference on machine learning*, 12878–12889. PMLR.