Model Heterogeneous Federated Learning with Personalized Feature Distillation

Yixiang Wang
School of Mathematics and Statistics
South-Central Minzu University
Wuhan, China
yxwang@mail.scuec.edu.cn

Boyi Liu*
Department of Data Science
City University of Hong Kong
Hong Kong, China
boy.liu@my.cityu.edu.hk

***Abstract*—Model-Heterogeneous Federated Learning (MHFL) enables collaborative training across clients with diverse backbones while keeping private data local, yet it renders parameter aggregation infeasible and complicates knowledge fusion across incompatible feature spaces. Public-data-assisted distillation provides a model-agnostic alternative, but logit alignment can be unreliable under feature mismatch, and naive feature matching may over-regularize heterogeneous clients. We propose PerFed, a feature-centric MHFL framework that performs personalized feature distillation without aggregating client backbone parameters. PerFed aligns intermediate features into a shared space and constructs client-specific distillation targets via mask-guided composition, selectively inheriting globally transferable cues while retaining complementary client-specific patterns. Empirical results across standard MHFL benchmarks show that PerFed consistently outperforms strong state-of-the-art baselines.**

***Index Terms*—Federated Learning, Model Heterogeneity, Knowledge Distillation**

I. Introduction

Model-Heterogeneous Federated Learning (MHFL) considers federated training with heterogeneous client backbones, driven by diverse computation and memory budgets. This heterogeneity leads to incompatible model architectures and feature spaces. Unlike conventional homogeneous settings where parameter aggregation (*e.g.* FedAvg [1]) can be directly applied, MHFL must accommodate mismatched parameter spaces and representations. Consequently, parameter-wise aggregation becomes infeasible (Fig. 1), and a central challenge is how to effectively fuse transferable knowledge across incompatible feature spaces.

A common remedy is public-data-assisted knowledge distillation [2], [3], where clients exchange or aggregate soft predictions (logits) on a shared public dataset to enable model-agnostic collaboration without parameter aggregation. This paradigm is appealing because logits provide an architecture-independent label-space interface for cross-model comparison. However, under heterogeneous backbones, logit-level alignment becomes unreliable: feature mismatch leads to miscalibrated and low-agreement predictions on the same public samples, yielding inconsistent supervision across rounds [4]. As a result, the distilled targets may provide weak or conflicting guidance, limiting transferability and robustness.

*Corresponding author

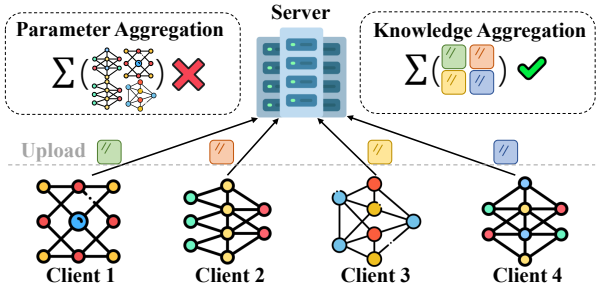


Fig. 1. Model-heterogeneous federated learning overview.

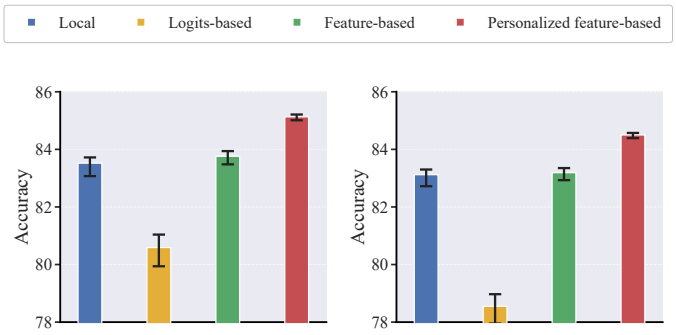
As shown in Fig. 2, logit-level distillation can even underperform local training in MHFL. Aggregating inconsistent logits produces noisy soft targets that may pull models away from local optima, and this effect intensifies with increasing data heterogeneity.

A natural remedy is to distill intermediate features instead of aligning logits, as feature-level supervision provides richer guidance (*e.g.* FitNets [5] and Attention Transfer [6]). In MHFL, this is typically achieved via projection modules that map heterogeneous features into a shared space [4]. However, direct feature matching remains challenging: dense point-wise constraints may over-regularize client-specific patterns and propagate non-transferable activations, especially under imperfect alignment.

Feature-based distillation provides limited gains over local training under moderate heterogeneity and becomes nearly ineffective under extreme settings (see Fig. 2). This indicates that naive feature matching yields weak or unstable benefits when cross-model alignment is imperfect. To address this limitation, we propose PerFed, a personalized feature distillation framework for MHFL. It constructs client-specific targets and structured distillation regions, enabling more reliable knowledge transfer while preserving client-specific patterns under severe heterogeneity. Code and data are available at https://github.com/The-ijc/PerFed.

We summarize the main contributions of PerFed as below:

- Through systematic experiments, we observe that logit-level distillation provides no consistent positive guidance in model-heterogeneous settings. In contrast, feature-



(a) Moderate Data Heterogeneity (b) Extreme Data Heterogeneity

Fig. 2. Comparison of Different Strategies Across Data Heterogeneity

level distillation may yield only limited gains, and its improvement diminishes as data heterogeneity increases, approaching zero under severe heterogeneity.

- We propose PerFed, which adopts personalized features for effective knowledge transfer. We select globally consensual regions via mask-based filtering and fuse them with client-specific local masks, yielding distillation knowledge that maintains global consistency while preserving local diversity. This design provides more robust and reliable supervisory feature for client training.
- We conduct extensive evaluations across multiple datasets and experimental settings. In model-heterogeneous settings, PerFed achieves consistent improvements over state-of-the-art baselines.

II. Related Work

Model-Heterogeneous Federated Learning (MHFL) addresses the challenge of model heterogeneity, where FedAvg [1] fails due to mismatched model architectures. MHFL introduces proxy mechanisms to enable knowledge transfer across heterogeneous models. Existing MHFL works can be divided into: *(i) Public dataset-based MHFL*, which aligns the model outputs towards a public dataset [2]–[4], [7], [8]. *(ii) Data-free MHFL*, which relies on proxies like auxiliary models or data generators [9]–[11] instead of public dataset. This paper focuses on MHFL with public dataset, following representative public-data-based distillation frameworks such as FedMD [2], FedDF [3]. However, these methods predominantly rely on response-based distillation and do not explicitly exploit intermediate features for knowledge transfer. In contrast, a recent [4] investigates feature-level distillation, but still assumes that client models are partially aggregatable across heterogeneous architectures. To the best of our knowledge, our work is the first to study feature-level knowledge transfer in a fully heterogeneous MHFL setting, where client models share no compatible backbone parameters, making client-backbone parameter aggregation infeasible.

Knowledge distillation transfers knowledge from a teacher to a student by aligning the student with teacher-provided signals [12]. This domain can be broadly divided into three main areas: *(i) Response-based distillation* matches teacher–student output distributions via KL divergence [13], [14], but offers coarse supervision with limited intermediate guidance [5]. *(ii) Relation-based distillation* differs from response-based distillation that aligns outputs of specific layers [15], [16]. and *(iii) Feature-based distillation* [4], [17]. Early work such as FitNets [5] distills hint features to guide student learning, while Attention Transfer [6] aligns attention maps to convey where the model focuses. Beyond direct feature matching, FSP [18] captures cross-layer feature relations for more structured supervision. It offers richer intermediate-layer supervision, enabling the transfer of fine-grained feature-level knowledge, and can be readily extended with projection/transform modules to reconcile mismatched teacher-student feature dimensions. We are one of the first works that adopt feature-based KD in MHFL.

III. Problem Statement

Traditional federated learning *e.g.* FedAvg [1] follows an iterative pipeline of local training at clients and weighted parameter aggregation at server. Formally, given n clients $\{c_1, \ldots, c_n\}$ with local dataset $\{D_1, \ldots, D_n\}$, federated learning aims to train a single global model θ with following objective:

$$\min \sum_{i=1}^{n} p_i F_i(\theta; D_i) \tag{1}$$

where $p_i = \frac{|D_i|}{\sum_{i=1}^{n} |D_i|}$, $|D_i|$ is the size of local dataset D_i, and $F_i(\cdot)$ is the local objective of client i.

However, under model heterogeneity, the server cannot directly aggregate the received model parameters. An alternative approach is to introduce a *public dataset* D_{pub} [2], [8], which enables the server and clients to exchange information and mitigate model discrepancies. By leveraging D_{pub}, model heterogeneity can be addressed by training n individual models $\{\theta_1, \ldots, \theta_n\}$ with the following objective:

$$\min \sum_{i=1}^{n} p_i (F_i(\theta_i; D_i) + \mathcal{R}_i(\theta_i; D_{pub})) \tag{2}$$

where θ_i is local model of client i, $\mathcal{R}_i(\theta_i; D_{pub})$ is a regularization term that incorporates global knowledge from the public dataset into the local model training.

IV. Method

A. PerFed Overview

We propose PerFed (Federated Learning with Personalized Feature Distillation), which bridges heterogeneous model via personalized feature distillation.

Key Idea. This work presents a feature-centric distillation framework for heterogeneous models. Instead of conventional logit-level distillation [2], [3], [7], [8], we distill *intermediate features* to enable finer-grained knowledge transfer.

We instantiate a saliency-consistent distillation objective by aggregating aligned features and imposing sparsity-aware masking to form personalized supervision (see Sec. IV-B). To further mitigate semantic mismatch and prevent unstable feature pulling during alignment, we adopt a dual-projection

Algorithm 1: PerFed

Input: Client set $\mathcal{C}$; Local datasets $\{D_k\}_{k\in\mathcal{C}}$; public dataset D_0 and Top-K settings K_g and $\{K_k\}_{k\in\mathcal{C}}$

Output: Updated client models $\{\theta_k^t\}_{k\in C}$

for *round* $t = 1$ *to* T **do**
- $\mathcal{S}^t \leftarrow \text{Sample}(\mathcal{C})$
- **for** $k \in \mathcal{S}^t$ **do** in parallel
  - Receive global projector ϕ^{t-1}, global mask M_g^{t-1}, and global feature I^{t-1} from server
  - `// Client-side Knowledge Distillation`
  - Fuse personalized feature P_k^t via (7)-(10)
  - Update model parameters on *public data* via (14)
  - `// Client-side Training`
  - Update model parameters via (15)
  - Upload local projector ϕ_k^t, and local feature I_k^t
- `// Server-side Aggregation`
- Aggregate global feature I^t via (3)
- Aggregate global mask M_g^t via (4)
- Aggregate global projector ϕ^t via (13)

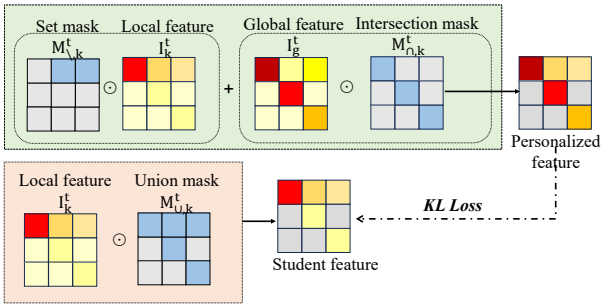


Fig. 3. Personalized feature construction and client-side distillation framework

design that decouples personalization from global alignment (see Sec. IV-C). The specifically tailored knowledge is then distilled to clients by regularizing the features during local training (see Sec. IV-D).

Workflow. PerFed follows an iterative client-server training framework in public dataset-based federated learning [2]–[4], [7], [8]. Each communication round consists of three steps (see Algorithm 1).

- **Step 1: Downlink and Client Distillation.** The server broadcasts the aggregated global feature, the global mask, and the updated server-shared projector to the selected clients. Each client then personalizes the global feature to form a client-specific distillation target and performs feature distillation on the public dataset to refine its local model like Fig. 3.
- **Step 2: Local Update and feature Extraction.** Each client performs local training on its private dataset. It then computes intermediate features (*e.g.* feature maps) with updated model on the public dataset.
- **Step 3: Uplink and Server-Side Aggregation.** Clients upload the extracted features (and the required shared-projector updates) to the server. The server aggregates the uploaded features to obtain a global feature and derives a global mask for the subsequent distillation stage. Meanwhile, it updates the server-shared projector by aggregating the received client-side updates.

B. Personalized feature Construction

Principles. feature-level distillation provides richer and more fine-grained supervision than logit matching. However, naively using a globally averaged feature as guidance can be overly coarse and biased toward majority patterns, thereby suppressing client-specific salient responses. To preserve global consistency while retaining personalization, PerFed learns client-wise guidance via mask-adaptive composition: it extracts informative regions with sparse saliency masks and synthesizes a personalized target by selectively inheriting global cues and complementing them with local salient patterns.

Aggregating Global feature and Mask. At round t, the server receives client features produced from heterogeneous backbones. Directly aggregating client-side intermediate activations is unreliable due to architecture-induced semantic mismatch. To ensure comparability, PerFed first maps client features $\hat{F}_k(x)$ into a shared semantic space using a server-shared projector $h(\cdot)$, and then aggregates the aligned maps to obtain a global feature I^t that captures globally consistent saliency patterns:

$$I^t = \frac{1}{|\mathcal{S}^t|} \sum_{k\in\mathcal{S}^t} I_k^t. \tag{3}$$

where $\mathcal{S}^t$ denotes the set of selected clients in round t and I_k^t is the feature generated by client k at round t.

Based on the magnitude of I^t, the server derives a sparse global mask M_g^t that highlights the most informative spatial regions:

$$M_g^t = \textsc{TopKMask}(I^t, K_g). \tag{4}$$

where K_g denotes the number of top-activated spatial positions.

For communication efficiency, the server transmits only the masked global feature I_g^t as downlink guidance:

$$I_g^t = I^t \odot M_g^t. \tag{5}$$

Learning Personalized Guidance via Dual-Mask Combination. Given the broadcast projector $h(\cdot)$ and global guidance I_g^t, client k computes its local feature I_k^t on the same public batch. Following the server-side rule, it constructs a client mask by retaining the Top-K_k, high-response entries:

$$M_k^t = \textsc{TopKMask}(I_k^t, K_k). \tag{6}$$

where K_k denotes the number of top-activated spatial positions.

Prior alignment objectives (e.g., ℓ_1/ℓ_2) enforce point-wise correspondence over all spatial locations, which may over-regularize heterogeneous clients and propagate noisy activations. Instead, PerFed constructs structured distillation regions using mask intersection $M^t_{\cap,k}$ and union $M^t_{\cup,k}$. The intersection captures globally consistent high-response regions, while the union preserves complementary salient patterns. Formally, we define:

$$M^t_{\cap,k} = M^t_g \odot M^t_k, \tag{7}$$

$$M^t_{\cup,k} = \min\big(1,\, M^t_g + M^t_k\big)\,, \tag{8}$$

$$M^t_{\setminus,k} = M^t_{\cup,k} - M^t_{\cap,k}. \tag{9}$$

We then construct the personalized distillation target by inheriting global guidance I^t_g in the intersection region $M^t_{\cap,k}$ and preserving client features in the complementary salient region $M^t_{\setminus,k}$:

$$P^t_k = M^t_{\cap,k} \odot I^t_g + M^t_{\setminus,k} \odot I^t_k. \tag{10}$$

The complement region $(1 - M^t_{\cup,k})$ is excluded from distillation, so no constraint is imposed outside the salient regions.

C. Feature Projection Modules

Principles. While intermediate-feature distillation offers richer supervision than logits, it becomes unreliable under heterogeneous backbones. Feature mismatch arises not only in tensor shapes (channels/resolutions) but also in semantics, where naive normalization fails to make activations comparable. Thus, robust transfer should decouple shape normalization from semantic alignment. PerFed realizes this via a bi-level projector: $g_k(\cdot)$ adapts client features to a unified tensor space, and $h(\cdot)$ enforces a shared embedding to stabilize aggregation and distillation.

Bi-level Projector Framework. Client k extracts an intermediate feature map $F_k(x)$ from its backbone as the distillation carrier. Compared with logits, intermediate features encode richer and more transferable knowledge for heterogeneous distillation.

- *Personalized projector*. Due to architectural discrepancies, client feature maps often differ in channel dimensions and spatial resolutions. We thus introduce a client-private projector $g_k(\cdot)$ to adapt $F_k(x)$ into a unified tensor space:

$$\hat{F}_k(x) = g_k\big(F_k(x)\big). \tag{11}$$

  The resulting $\hat{F}_k(x)$ provides a shape-compatible feature carrier for subsequent semantic alignment.
- *Server-shared projector*. Following the bi-level design, while $g_k(\cdot)$ resolves shape incompatibility, the resulting features may still lie in client-specific semantic spaces. To ensure semantic comparability for reliable aggregation and mask construction, we introduce a server-shared projector $h(\cdot)$ that maps $\hat{F}_k(x)$ into a unified embedding space:

$$I_k(x) = h\big(\hat{F}_k(x)\big). \tag{12}$$

The shared projector $h(\cdot)$ is collaboratively optimized via standard federated averaging. After local updates on client k, its parameters ϕ^t_k are uploaded and aggregated as:

$$\phi^t = \sum_{k\in\mathcal{S}^t} \frac{|D_k|}{\sum_{j\in\mathcal{S}^t}|D_j|}\,\phi^t_k, \tag{13}$$

where $|D_k|$ denotes the number of training samples on client k, ϕ^t_k is the locally updated parameter set of the server-shared projector on client k at round t, and ϕ^t is the aggregated global projector parameter broadcast to selected clients in the next round. The aggregated ϕ^t is then broadcast in the next round, stabilizing the shared embedding space over communication rounds and enabling consistent global guidance.

D. Local Training and Distillation

Here, $\mathcal{R}_i(\theta_i; D_{\text{pub}})$ is instantiated as our mask-guided feature distillation regularizer:

$$\mathcal{R}_i(\theta_i; D_{\text{pub}}) = \mathbb{E}_{x\sim D_{\text{pub}}}\Big[\mathrm{KL}\Big(\sigma\big(M^t_{\cup,i} \odot P^t_i(x)\big) \;\Big\|\; \sigma\big(M^t_{\cup,i} \odot I^t_i(x)\big)\Big)\Big]. \tag{14}$$

where $P^t_i(x)$ is the personalized target constructed in Eq. (10). We instantiate $F_i(\theta_i; D_i)$ as the standard supervised objective on private data:

$$F_i(\theta_i; D_i) = \mathbb{E}_{(x,y)\sim D_i}\big[\ell_{\text{CE}}(f(x;\theta_i), y)\big]. \tag{15}$$

V. Experiments

A. Experimental Setup

Configurations. We set client number $n = 20$, with a sampling rate of 40%. We set communication rounds $T = 50$, local epochs $E = 5$, batch size $B = 64$. The learning rate η is tuned within $\{0.01, 0.001\}$.

Baselines. We compare PerFed with the following MHFL baselines: FedDF [3], FedProto [19], FedMD [2], LG-FedAvg [20], FedMRL [10], KT-pFL [8].

Datasets and Models. This study focuses on the image classification task, evaluating PerFed and baselines on CIFAR-10 [21] with public dataset CIFAR-100 [21], CIFAR-100 [21] with public dataset CIFAR-10 [21] and EMNIST [22] with public dataset MNIST [23]. To simulate heterogeneous data distributions, two partitioning schemes are used: (i) the pathological partition [24] with classes per client C=2 for CIFAR-10 and C=10 for CIFAR-100; and (ii) a Dirichlet partition [25] with concentration parameter α=0.1, α=0.5, α=1. Five heterogeneous CNN backbones (CNN-1 to CNN-5) are employed for both datasets. Client-specific convolutional layers serve as personalized projectors, while a server-shared convolutional layer acts as a common projector.

B. Main Result

Study on Comprehensive Benchmarks. Figure 4 and Table I report test accuracy (%) on CIFAR-10, CIFAR-100, and EMNIST under Dirichlet label skew (Dir$(0.1/0.5/1)$) in MHFL, where backbone aggregation is infeasible.

TABLE I
TEST ACCURACY (%) UNDER DIRICHLET LABEL SKEW.

Method	CIFAR10			CIFAR100			EMNIST		
	Dir(0.1)	Dir(0.5)	Dir(1)	Dir(0.1)	Dir(0.5)	Dir(1)	Dir(0.1)	Dir(0.5)	Dir(1)
FedProto	83.97	84.46	84.89	46.15	26.83	20.04	87.70	87.69	87.99
LG-FedAvg	84.05	84.37	84.72	45.73	26.76	20.37	87.27	87.58	87.45
KT-pFL	84.01	84.12	84.29	45.92	26.17	20.13	87.52	87.58	87.67
FedMD	78.55	79.97	80.59	40.73	22.13	17.98	86.94	87.14	87.02
FedDF	12.54	10.64	10.46	4.47	21.93	**22.11**	83.80	83.88	83.90
FedMRL	83.83	83.96	83.85	44.84	26.64	20.43	87.65	87.43	87.89
PerFed	**84.10**	**84.55**	**85.13**	**47.08**	**27.85**	20.77	**87.83**	**87.91**	**88.09**

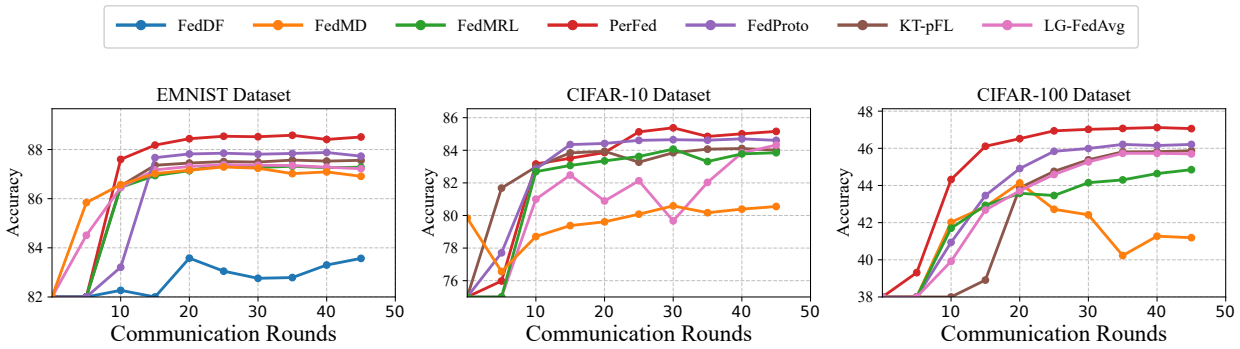


Fig. 4. Comparison of PerFed with different baseline methods across diverse datasets. For clarity, we plot one point every 5 communication rounds.

TABLE II
TEST ACCURACY COMPARISON UNDER PATHOLOGICAL LABEL SKEW IN RECENT EXPERIMENTS.

Method	CIFAR10		CIFAR100	
	Pat2	Pat5	Pat10	Pat20
FedProto	86.51	68.64	58.56	43.32
FedMRL	84.93	67.30	56.00	40.30
LG-FedAvg	86.29	68.03	58.36	42.18
PerFed	87.31	69.64	58.73	43.45

Overall, PerFed achieves consistent improvements across datasets and heterogeneity levels. The gains are more pronounced under challenging settings. Logit-based distillation (FedMD) is less reliable in MHFL. Aligning output distributions often yields weak or unstable supervision when feature mismatch exists across heterogeneous backbones. In contrast, PerFed adopts feature-centric distillation. It provides more transferable guidance and leads to higher accuracy and improved robustness across benchmarks.

Study over Pathological Non-IID. Table II compares PerFed with representative MHFL baselines on CIFAR-10 and CIFAR-100 under pathological label partitioning. CIFAR-10 is evaluated on Pat2 and Pat5, while CIFAR-100 uses Pat10 and Pat20, reflecting increasing label skew.

PerFed achieves the best accuracy across all settings, showing strong robustness under highly fragmented label spaces. The improvements are consistent in both mild (Pat2/Pat10) and severe (Pat5/Pat20) regimes, indicating better preservation of transferable knowledge under strong heterogeneity.

C. Ablation Study

Table III shows that PerFed's gains mainly come from personalized feature construction and client-side distillation. Removing personalized features consistently reduces accuracy on CIFAR-10/100, especially under Dir(0.1), indicating that directly distilling global features can over-regularize heterogeneous clients by enforcing dense alignment and transferring non-transferable activations. In contrast, PerFed builds structured distillation targets by selecting salient regions and fusing global and local cues, thus applying supervision only to comparable representations. Removing distillation causes further degradation, confirming that public-data distillation is important for transferable knowledge injection and stable cross-model transfer. Overall, mask-guided personalized feature transfer better balances global consistency and client specificity than direct global feature distillation.

D. Alignment Degradation under Data Heterogeneity

We measure region consistency using the intersection-over-union (IoU) between the client mask M_k and the global mask M_g:

$$\mathrm{IoU}(M_k, M_g) = \frac{|M_k \cap M_g|}{|M_k \cup M_g|}. \tag{16}$$

Larger IoU indicates stronger agreement between local and global salient regions. As shown in Fig. 5, IoU decreases

TABLE III
ABLATION STUDY OF PERSONALIZED FEATURE CONSTRUCTION AND DISTILLATION.

Method	CIFAR10		CIFAR100	
	Dir(0.1)	Dir(1)	Dir(0.1)	Dir(1)
PerFed	84.10	85.13	47.08	20.77
-w/o personalized feature	83.19	83.76	44.61	19.12
-w/o distillation	83.12	83.52	42.57	17.65

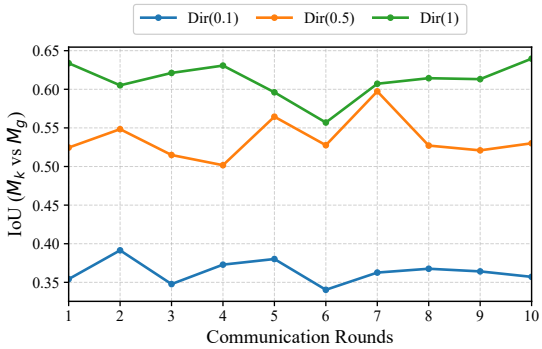


Fig. 5. IoU variations across communication rounds under different degrees of data heterogeneity.

monotonically with increasing heterogeneity, indicating fewer shared salient regions under mismatched label distributions.

Under stronger heterogeneity, clients rely more on client-specific discriminative cues, making global saliency less representative. As a result, the overlap between M_k and M_g shrinks, and supervision based on global regions is more likely applied to non-overlapping activations, introducing noise or conflicting signals.

The declining IoU explains the observations in Fig. 2: generic feature distillation provides limited gains under moderate heterogeneity and becomes ineffective under extreme heterogeneity.

VI. CONCLUSION

We propose PerFed for fully model-heterogeneous federated learning, where client backbones are not aggregatable. PerFed enables personalized feature distillation by aligning heterogeneous features into a shared space and applying mask-guided selective distillation. Extensive experiments demonstrate consistent improvements over strong MHFL baselines.

ACKNOWLEDGMENT

Special thanks go to Mr. Qiyang Liu and Mrs. Tongju Jiang.

REFERENCES

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *AISTATS*, 2017, pp. 1273–1282.

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

[3] T. Lin, L. Kong, S. U. Stich, and M. Jaggi, "Ensemble distillation for robust model fusion in federated learning," *NeurIPS*, vol. 33, pp. 2351–2363, 2020.

[4] Y. Li, X. Wang, W. Xu, H. Wang, Y. Qi, J. Dong, and R. Li, "Feature distillation is the better choice for model-heterogeneous federated learning," *arXiv preprint arXiv:2507.10348*, 2025.

[5] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: hints for thin deep nets; 2014," *arXiv preprint arXiv:1412.6550*, vol. 3, 2014.

[6] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," *arXiv preprint arXiv:1612.03928*, 2016.

[7] X. Gong, A. Sharma, S. Karanam, Z. Wu, T. Chen, D. Doermann, and A. Innanje, "Ensemble attention distillation for privacy-preserving federated learning," in *ICCV*, 2021, pp. 15 076–15 086.

[8] J. Zhang, S. Guo, X. Ma, H. Wang, W. Xu, and F. Wu, "Parameterized knowledge transfer for personalized federated learning," *NeurIPS*, vol. 34, pp. 10 092–10 104, 2021.

[9] J. Zhang, C. Chen, B. Li, L. Lyu, S. Wu, S. Ding, C. Shen, and C. Wu, "Dense: Data-free one-shot federated learning," *NeurIPS*, vol. 35, pp. 21 414–21 428, 2022.

[10] L. Yi, H. Yu, C. Ren, G. Wang, X. Liu, and X. Li, "Federated model heterogeneous matryoshka representation learning," in *NeurIPS*, 2024.

[11] C. Guo, Q. He, X. Tang, Y. Liu, and Y. Jie, "Parameterized data-free knowledge distillation for heterogeneous federated learning," *Knowledge-Based Systems*, vol. 317, p. 113502, 2025.

[12] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *IJCV*, vol. 129, no. 6, pp. 1789–1819, 2021.

[13] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[14] J. Ba and R. Caruana, "Do deep nets really need to be deep?" *NeurIPS*, vol. 27, 2014.

[15] S. H. Lee, D. H. Kim, and B. C. Song, "Self-supervised knowledge distillation using singular value decomposition," in *ECCV*, 2018, pp. 335–350.

[16] C. Zhang and Y. Peng, "Better and faster: knowledge transfer from multiple self-supervised learning tasks via graph distillation for video classification," *arXiv preprint arXiv:1804.10069*, 2018.

[17] B. Heo, J. Kim, S. Yun, H. Park, N. Kwak, and J. Y. Choi, "A comprehensive overhaul of feature distillation," in *ICCV*, 2019, pp. 1921–1930.

[18] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *CVPR*, 2017, pp. 4133–4141.

[19] Y. Tan, G. Long, L. Liu, T. Zhou, Q. Lu, J. Jiang, and C. Zhang, "Fedproto: Federated prototype learning across heterogeneous clients," in *AAAI*, vol. 36, no. 8, 2022, pp. 8432–8440.

[20] P. P. Liang, T. Liu, L. Ziyin, N. B. Allen, R. P. Auerbach, D. Brent, R. Salakhutdinov, and L.-P. Morency, "Think locally, act globally: Federated learning with local and global representations," *arXiv preprint arXiv:2001.01523*, 2020.

[21] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[22] G. Cohen, S. Afshar, J. Tapson, and A. van Schaik, "Emnist: an extension of mnist to handwritten letters," 2017. [Online]. Available: https://arxiv.org/abs/1702.05373

[23] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[24] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," *arXiv preprint arXiv:1806.00582*, 2018.

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