FedCD: Towards Consolidated Distillation for Heterogeneous Federated Learning

Yichen Li[1,2], Hang Su[1], Huifa Li[2], Haolin Yang[2], Xinlin Zhuang[2], Haochen Xue[2], Haozhao Wang[1], Imran Razzak[2*]

[1]School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China
[2]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates
{ycli0204, hz_wang}@hust.edu.cn

Abstract

Knowledge Distillation (KD) serves as an effective approach to addressing heterogeneity issues in Federated Learning (FL), leveraging additional datasets to align local and global models better. There are two primary distillation paradigms: feature-based distillation, which utilizes intermediate-layer features of the network, and logit-based distillation, which employs the final layer's logit outputs. However, existing studies often select distillation methods based on intuitive and empirical evidence when facing different heterogeneous settings, neglecting the intrinsic relationship between distillation paradigms and heterogeneity. This oversight may result in suboptimal federated knowledge distillation performance under heterogeneous conditions. In this paper, we propose the Consolidated Distillation for Heterogeneous Federated Learning - **FedCD** that balances knowledge representations from both feature-based and logit-based distillation to enhance performance. Specifically, to address the misalignment between knowledge conveyed by features and logits, we aggregate features from different layers via cross-layer attention to preserve semantic knowledge, followed by distribution modeling using Gaussian Mixture Models. This process strengthens knowledge distillation by constraining the transformation of different network layers' features under a consolidated distribution, thereby mitigating impacts from both data and model heterogeneity. Extensive experiments demonstrate that FedCD outperforms state-of-the-art methods by over 10.72% and validate the effectiveness of our approach.

Introduction

Federated Learning (FL) is a distributed machine learning paradigm enabling multiple devices to collaboratively train models without sharing raw data (McMahan et al. 2017; Li et al. 2025e; Wang et al. 2022). By aggregating model updates on a central server, FL preserves data privacy and reduces communication costs compared to centralized approaches (Wang et al. 2023b; Li et al. 2025g; Liao et al. 2024, 2025). Recently, FL has gained traction in privacy-sensitive domains such as healthcare (Dong et al. 2024; Pfitzner, Steckhan, and Arnrich 2021), IoT (Khan et al. 2021; Nguyen et al. 2021), and mobile apps (Kang et al. 2020; Lim et al. 2020).

*Corresponding author.

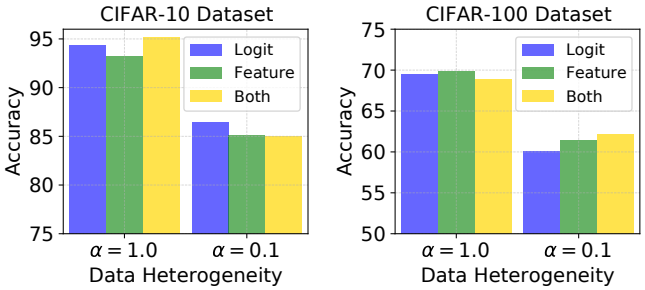

Figure 1: The bar chart illustration of ensemble distillation performance on two datasets. We test three distillation modes: a) logit, b) feature, and c) logit+feature (directly).

However, these aggregation-based FL methods often experience significant drops in model effectiveness when client data follows a heterogeneous setting, a prevalent challenge in federated settings (Li et al. 2020a, 2024a, 2025d; Meng et al. 2024; Qi et al. 2023). This occurs because local model parameters are optimized in conflicting directions across devices, resulting in excessive variance during aggregation.

To address this limitation, federated distillation introduces a method that transfers knowledge from multiple local models to the global model by aggregating their predictions on proxy datasets, a strategy gaining growing interest (Li and Wang 2019). (Lin et al. 2020) utilized a shared public dataset as distillation samples to generate soft logit predictions from local models, then updated the global model by averaging these predictions. Building on this, subsequent works by (Zhu, Hong, and Zhou 2021) and (Wang et al. 2023a) enhanced distillation efficiency by replacing the public dataset with synthetic data generated through generative models. Different from the works mentioned above, (Li et al. 2025f) pioneers in feature-based distillation with orthogonal projection to handle heterogeneous models in FL.

Although these methods significantly enhance the performance of FL approaches in heterogeneous settings, their selection of distillation methods is intuitive and lacks interpretability, potentially leading to suboptimal model optimization. There are generally two distillation modes: Logit-based and Feature-based (Romero et al. 2015). Most existing studies employ logit-based distillation because logit outputs are model-agnostic and possess higher information density. However, (Heo et al. 2019; Huang et al. 2024) also finds that

features often contain richer information, which can better improve distillation effectiveness, particularly on complex data distributions. How to adaptively select a more suitable distillation approach based on heterogeneity settings in FL remains an important and unexplored topic.

To tackle this challenge, we propose investigating a consolidated distillation method to further enhance federated distillation performance while adaptively and interpretably accommodating different heterogeneous FL settings. A simple and direct approach might involve using both logit and feature outputs to compute distillation loss. However, since features and logits differ in information density, determining the weighting of the two loss components in the objective function is challenging and highly dependent on hyperparameter choices. Fig.1 illustrates a simple experiment comparing the performance of different distillation approaches. Empirically, it demonstrates that the relative effectiveness of logit-based versus feature-based distillation is often unpredictable, and naively combining both distillation losses may even yield inferior performance compared to using a single distillation loss.

To explore this idea, we propose a consolidated distillation method for heterogeneous FL that constrains feature representations across different network layers and jointly utilizes them with logit outputs for federated distillation. Specifically, we first perform adaptive feature fusion from layers. We then model their distributions using Gaussian Mixture Models (GMMs) (Wu et al. 2023) to constrain feature knowledge. Finally, we compute feature distribution loss via Wasserstein distance and optimize it jointly with traditional logit-based distillation loss for federated optimization. We conduct extensive experiments on three widely used datasets comparing against eight baselines. The results demonstrate that FedCD effectively enhances federated distillation performance and exhibits greater robustness under heterogeneous settings. Our main contributions are summarized as follows:

- We are the first to explore the selection of distillation modes in heterogeneous FL. It is meaningful to figure out the impact of different distillation modes on federated distillation performance and their underlying selection mechanisms.
- We then propose a novel approach named FedCD that enhances distillation performance by constraining feature representations across network layers while simultaneously leveraging both feature-based and logit-based distillation modes.
- Extensive experiments have been done on three datasets and various settings to evaluate the effectiveness of FedCD. Experimental results demonstrate that the proposed method can outperform other state-of-the-art baselines by up to 10.72% in terms of test accuracy.

Related Work

Federated Learning. Non-IID data poses a primary challenge for model parameter aggregation methods due to significant diversity among local model parameters across clients (Li et al. 2024b; Huang et al. 2025; Li et al. 2025b,h,c,a). This heterogeneity arises as local models optimize toward divergent objectives. To address this, numerous studies aim to reduce parameter discrepancy for effective aggregation. (Li et al. 2020b) introduced regularization terms in local objectives to constrain deviation from the global model. (Karimireddy et al. 2020) mitigated update heterogeneity through local gradient variance reduction. (Diao, Ding, and Tarokh 2020) introduces a heterogeneous local model training accommodating varying computational complexities across clients. Extending this concept, (Wang et al. 2023c) incorporates dual flexibility in both model depth and width, employing skip connections to bypass specific layers alongside structured pruning for adaptive width control. Our work employs ensemble distillation on local models, offering an orthogonal approach to these techniques.

Knowledge Distillation. It leverages knowledge from pretrained teacher models to supervise compact student models, thereby facilitating deployment in resource-constrained environments (Hinton, Vinyals, and Dean 2015). This domain fundamentally encompasses two principal methodologies: logits distillation (Niu et al. 2022; Kim et al. 2021) and feature distillation (Tian, Krishnan, and Isola 2019; Miles, Rodriguez, and Mikolajczyk 2021). Logits distillation, primarily oriented toward classification tasks, incorporates an objective to minimize predictive discrepancies between student and teacher models. This approach originated with KL divergence optimization (Hinton, Vinyals, and Dean 2015), subsequently extended through techniques including spherical normalization (Guo et al. 2020a), label decoupling (Zhao et al. 2022), and probability reweighting (Niu et al. 2022). Feature distillation demonstrates versatility across diverse tasks (Chen et al. 2021) and multi-modal applications (Sanh 2019). Within this paradigm, manually designed flow of solution procedure matrices capture inter-layer feature relationships in residual architectures (Yim et al. 2017). We explore the selection of these two distillation modes in FL and improve the model performance.

Federated Distillation. It involves extracting knowledge from multiple client models and transferring it to the global model (Wu and Gong 2021; Guo et al. 2020b; Bistritz, Mann, and Bambos 2020). (Lin et al. 2020) introduces server-executed knowledge distillation that utilizes an unlabeled proxy dataset to transfer knowledge from local models to a global model. Further developing this paradigm, (Chen and Chao 2021) linearly aggregates multiple local models using Bayesian posterior-derived weights to create combined models, which are subsequently distilled into a single global model. To overcome the limitation of dependency on unlabeled auxiliary datasets, (Zhu, Hong, and Zhou 2021; Zhang et al. 2022; Wang et al. 2023a) propose replacing proxy datasets with generative model-synthesized data, thereby enabling distillation without authentic data requirements. We analyze the challenges inherent in deploying existing methods within heterogeneous FL settings and advocate for developing robust and better distillation methods for performance improvement.

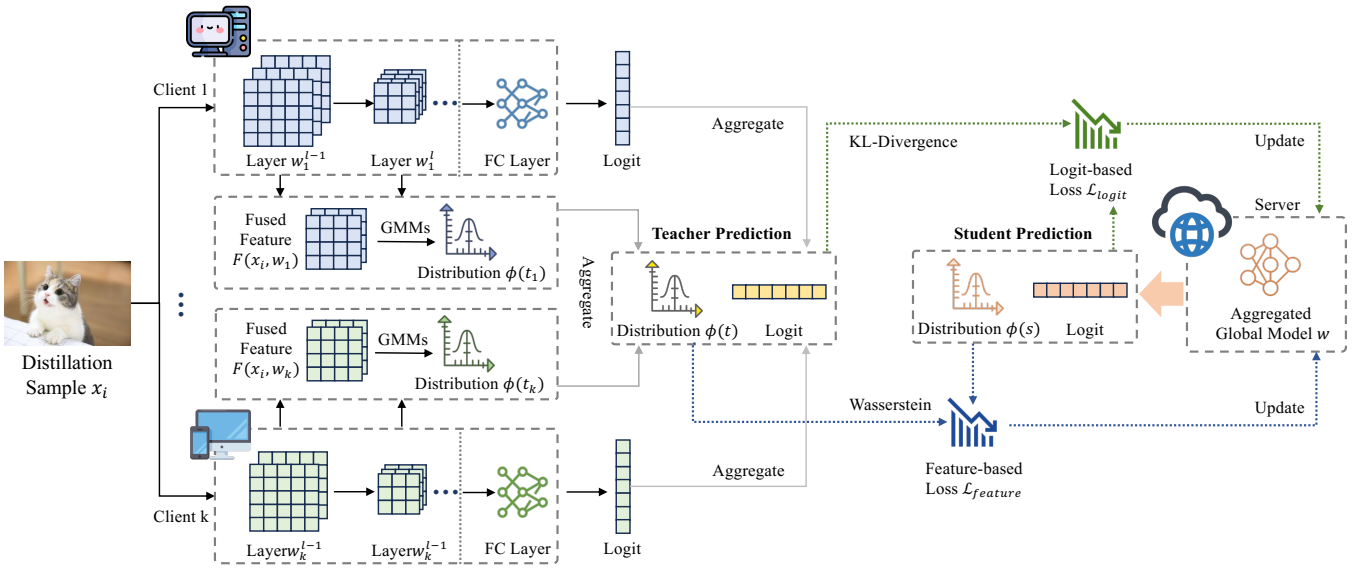

Figure 2: The framework of FedCD. Following global model aggregation in traditional FL, each participating local model and the aggregated global model perform inference on distillation samples. FedCD first employs adaptive feature fusion to consolidate features drawn from different layers into integrated representations, which are subsequently modeled via GMMs to capture statistical distributions. Then, soft predictions are directly extracted from the fully-connected layer output. The aggregated predictions from all participating local models constitute the teacher outputs, while the global model's predictions serve as student outputs. Finally, joint optimization of both feature-based and logit-based distillation losses updates the global model. The server will broadcast the updated global model to the participating clients in the next training round.

Methodology

Problem Formulation

A typical FL framework involves K clients collaboratively training a global model. Each client k exclusively accesses its local private dataset $D_k = (x_k^i, y_k^i)$, where x_k^i denotes the i-th input sample and $y_k^i \in \{1, \ldots, C\}$ represents the corresponding label among C classes. Let $|D_k|$ indicate the cardinality of D_k. The FL objective is to learn a global model w minimizing the global empirical loss over the combined dataset D:

$$\min_w \mathcal{L}(w) = \sum_{k=1}^{K} \frac{|D_k|}{|D|} \mathcal{L}(w_k),$$
$$\text{where } \mathcal{L}(w_k) = \frac{1}{|D_k|} \sum_{i=1}^{|D_k|} \mathcal{L}_{CE}(w_k; x_k^i, y_k^i). \quad (1)$$

Here $\mathcal{L}(w_k)$ is the local loss at client k, and $\mathcal{L}_{CE}$ denotes the cross-entropy loss between predictions and ground-truth labels. For aggregation-based methods, the server obtains the global model through weighted parameter averaging:

$$w := \sum_{k=1}^{K} \frac{|D_k|}{|D|} \cdot w_k. \quad (2)$$

In traditional FL, the aggregated global model will be distributed directly to clients for the next round of local training. In contrast, federated distillation requires an additional distillation step on the server to enhance model performance before distributing the distilled global model to clients. This necessitates a proxy dataset, denoted as P. Each client model and the aggregated global model then perform inference on P to obtain soft predictions $s(\cdot)$. Finally, the predictions from all client models are aggregated as the teacher model's output, while the aggregated global model's predictions serve as the student model's output. The model is then optimized again via distillation:

$$w = \arg\min_w \mathcal{L}_{KD}(w) \quad (3)$$
$$= \frac{1}{|P|} \sum_{x_i \in P} KL\left(\sum_{k=1}^{K} s(w_k; x_i), s(\hat{w}_{t+1}; x_i)\right),$$

where $KL(\cdot)$ is to compute the Kullback-Leibler divergence (KL-divergence).

FedCD: A Consolidated Distillation Framework

The key idea of FedCD is to balance knowledge from logit and feature outputs by constraining feature representations across network layers, thereby enabling simultaneous employment of both feature-based and logit-based distillation losses to ensure distillation performance and robustness in heterogeneous settings. More specifically, we incorporate an adaptive feature fusion module and a GMM-based distribution prediction module. We first replace simple gating mechanisms with cross-layer attention to obtain a feature with multi-scale knowledge. Then, we adopted GMMs assumptions to capture feature representations under complex heterogeneous settings. Finally, aligning the logit prediction

and the distribution drawn from features enables the global model to more effectively assimilate knowledge with effective performance improvements.

Algorithm 1: FedCD

Input : T: communication round; K: client number; D_k: local dataset for the client k; P: proxy dataset; $\mathbb{K}$: convolutional projector; $\mathbf{W}$:projection weight.
Output: w: global model.

Initialize the parameter w;
for $c = 1$ ***to*** T **do** `// communication round`
  Server randomly selects a subset of devices S_t and sends them the global model w;
  for *each selected client* $k \in S_t$ **in parallel do**
    Train the model with local data with (1);
    Send the local model w_k back to the server.
  end
  Adaptive Feature Fusion:
  Obtain the aggregated global model w with (2);
  Compute the cross-layer attention $G_k^{l,m}$ with (4)(5);
  Obtain the fused feature $F(x_i, w_k^l)$ with (6)(7);
  GMMs-based Distribution Prediction:
  Model the GMMs-based distribution $\phi(\cdot)$ with (8-10);
  Calculate two distillation losses $\mathcal{L}_{feature}$ and $\mathcal{L}_{logit}$ with (11)(3);
  Optimize the global model w with total distillation loss $\mathcal{L}_{total}$ with (12).
end

Adaptive Feature Fusion. In traditional FL scenarios, distributed devices are often subject to resource constraints that prevent them from handling substantial computation and communication overhead. Consequently, directly uploading features from different network layers is inadvisable. Furthermore, features from distinct layers carry heterogeneous knowledge, making it challenging to quantify their contributions to the loss function. To address these challenges, we employ cross-layer attention to fuse feature outputs across layers. Assuming client k's local model has L feature extractor layers, we denote its feature outputs for data sample x_i as $f(x_i, w_k^l)$. The attention mechanism employs two critical components derived from layer features:

$$Q_k^l = \mathbb{K}_1\big(f(x_i, w_k^l)\big), \quad K_k^l = \mathbb{K}_2\big(\mathrm{Up}(f(x_i, w_k^m))\big), \quad \forall m > l, \tag{4}$$

where $\mathbb{K}$ denotes the convolutional projector, Q_k^l acts as an information request from layer l, and K_k^l functions as an information index for layers $m > l$ with $\mathrm{Up}(\cdot)$ aligning spatial dimensions. The attention weights are computed as:

$$G_k^{l,m} = \mathrm{Softmax}\left(\frac{Q_k^l [K_k^l]^\top}{\sqrt{d_k}}\right). \tag{5}$$

This formula measures compatibility between current-layer queries and deeper-layer keys. Higher scores indicate stronger semantic relationships. Then, the channel dimensionality of each feature is first increased via an expansion module, then concatenated with the up-sampled output of deeper layers. This combined representation undergoes fusion through a gating mechanism, producing the integrated fused feature $F(x_i, w_k^l)$:

$$\tilde{f}(x_i, w_k^l) = \sum_{m=l+1}^{L} G_k^{l,m} \odot \mathbb{K}_3\big(\mathrm{Up}(f(x_i, w_k^m))\big), \tag{6}$$

$$F(x_i, w_k^l) = \sigma(\mathbf{W}[f(x_i, w_k^l)] \oplus [\tilde{f}(x_i, w_k^l)]), \tag{7}$$

where σ denotes the activation function and $\mathbf{W}$ is the weight matrix. This architecture dynamically prioritizes semantically rich features and maintains training stability in resource-constrained FL devices through learnable and interpretable attention weights.

GMMs-based Distribution Prediction. While the fused feature representations are obtained, their knowledge density remains inconsistent with that of the logits. This makes it challenging to balance their respective contributions to the distillation loss and may introduce substantial redundant information, thereby degrading distillation performance that we empirically demonstrate in Fig.1. Inspired by (Wu et al. 2023), GMMs are employed to model local data distributions per client, leveraging their statistical properties to distill essential information from data distributions. Here, we utilize GMMs to model the fused feature representations $F(\cdot)$ generated in the preceding step. This process begins with spatial compression to extract global statistics:

$$\mathbf{h}_k = \mathrm{Flatten}(\mathrm{GlobalAvgPool}(F(x_i, w_k^l)), \tag{8}$$

where $\mathbf{h} \in \mathbb{R}^+$ consolidates spatial information, thus eliminating positional variance while preserving channel-wise semantics. Three parallel fully-connected layers then predict GMM parameters:

$$\pi_k^j = \mathrm{Softmax}(\mathbf{W}_\pi \mathbf{h}_k), \quad \mu_k^j = \mathbf{W}_\mu \mathbf{h}_k, \quad \Sigma_k^j = \mathrm{diag}(e^{\mathbf{W}_\Sigma \mathbf{h}_k}), \tag{9}$$

where j denotes the j-th mixture component, $\mathbf{W}$ is the learnable projection weights, and exponential activation on Σ_k^j can ensure positive-definite covariances. The resulting distribution for student and teacher models is formalized as:

$$\phi(s) = \sum_{j=1}^{J} \pi^j \mathcal{N}(\mu^j, \Sigma^j), \quad \phi(t) = \sum_{k=1}^{K} \sum_{j=1}^{J} \pi_k^j \mathcal{N}(\mu_k^j, \Sigma_k^j). \tag{10}$$

Based on this, the feature-based distillation loss aligns and matches statistical parameters and distributions between the teacher and student models:

$$\mathcal{L}_{feature} = \frac{1}{K \cdot J} \sum_{k=1}^{K} \sum_{j=1}^{J} \left(\|\mu_k^j - \mu^j\|_2 + \|\Sigma_k^j - \Sigma^j\|_F \right) + \lambda \cdot \exp(-\|\phi(s) - \phi(t)\|^2), \tag{11}$$

Categories	Methods	Metrics	CIFAR-10			CIFAR-100			Tiny-ImageNet		
			α=10.0	α=1.0	α=0.1	α=10.0	α=1.0	α=0.1	α=10.0	α=1.0	α=0.1
Classic-FL	FedAvg	Acc	94.63±0.35	92.75±0.21	80.59±1.48	68.24±0.78	65.40±1.65	54.05±2.15	41.43±1.92	37.58±1.64	27.15±3.40
		Round	143.00±4.00	155.50±8.50	132.50±11.00	179.50±9.00	175.00±6.50	185.50±13.50	166.50±4.00	143.50±5.50	177.00 ±10.50
	FedProx	Acc	94.09±0.26	92.14±0.47	81.22±0.53	68.80±0.55	66.13±1.44	54.72±1.06	40.09±1.21	38.56±0.80	28.77±3.08
		Round	137.00±6.50	141.50±7.00	141.50±8.00	168.00±5.00	159.50±3.50	191.00±9.50	144.00±2.50	161.00±7.00	192.00±8.50
	MOON	Acc	94.81±0.16	92.67±0.40	80.75±1.19	67.42±0.19	64.86±1.12	54.31±1.92	38.89±0.75	37.98±0.99	28.95±2.07
		Round	125.50±7.00	144.00±10.00	136.50±12.50	166.50±14.50	170.00±9.00	177.00±2.50	135.50±15.50	174.00±5.50	196.00±9.50
Federated Distillation	FedMD	Acc	82.32±1.65	76.44±1.92	70.74±2.12	50.23±2.66	48.70±2.34	42.83±3.17	30.19±3.05	27.37±3.20	21.25±2.40
		Round	120.50±23.50	133.00±17.50	145.50±12.00	148.00±16.00	145.50±23.00	162.00±12.50	116.50±33.00	152.50±10.50	177.00±7.50
	FedFusion	Acc	95.88±0.07	94.38±0.64	86.46±0.73	70.13±0.25	69.48±0.59	60.11±1.15	44.16±0.68	40.37±1.17	32.58±1.84
		Round	106.00±6.50	119.50±1.50	106.50±10.00	139.50±8.00	115.50±6.50	148.00±7.00	125.00±12.50	139.50±14.00	167.00±7.50
	FedGen	Acc	95.11±0.35	93.02±1.88	84.79±1.51	69.08±2.07	67.58±1.32	57.41±2.78	42.90±1.24	39.49±2.05	33.68±1.79
		Round	131.50±4.50	142.00±12.50	126.00±5.50	146.00±2.00	136.50±10.00	155.00±8.00	147.50±11.50	**130.00±21.50**	159.50±4.50
	DaFKD	Acc	95.66±0.51	94.28±1.01	87.24±0.97	71.37±0.92	68.56±1.19	62.01±1.93	43.67±2.55	41.50±0.24	34.43±2.15
		Round	138.00±7.50	128.50±6.50	130.50±4.00	152.00±7.00	132.50±9.00	171.50±13.00	138.00±5.50	142.00±12.50	142.50±17.50
	FedFD	Acc	94.65±0.13	93.92±1.83	86.21±2.16	70.20±2.07	69.41±1.70	59.37±2.18	44.08±1.52	40.55±1.65	35.11±0.83
		Round	**98.50±8.50**	117.00±7.00	109.50±6.00	132.00±3.00	122.00±1.50	140.00±10.00	111.50±13.50	136.00±16.50	152.00±17.00
	FedCD	Acc	**95.92±0.44**	**95.52±0.79**	**89.18±1.33**	**72.96±0.74**	**70.26±1.04**	**64.77±1.40**	**46.70±2.19**	**43.90±1.55**	**37.31±1.17**
		Round	110.00±4.00	**114.50±6.50**	**105.00±9.00**	**126.00±12.00**	**113.50±6.00**	**139.00±9.50**	**108.00±14.00**	143.50±14.50	**132.50±6.50**

Table 1: Performance comparison of various methods with the test accuracy. Here we report the best test accuracy and the communication rounds to reach the corresponding test accuracy. The best results are **bold**.

where F denotes the Frobenius norm, λ is a hyperparameter (set to 0.1 here), and the maximum mean discrepancy with RBF kernel is used to ensure holistic distribution matching beyond moment alignment.

Federated Distillation Loss. Following the feature-based distillation loss, we compute the logit-based distillation loss $\mathcal{L}_{logit}$ by calculating the KL divergence of soft predictions as specified in Eq.(3). The knowledge distillation for the aggregated global model is then implemented by jointly optimizing both losses to enhance performance. The model is optimized as follows:

$$w = \arg\min_{w} \mathcal{L}_{total} = \frac{1}{2}(\mathcal{L}_{feature} + \mathcal{L}_{logit}). \tag{12}$$

Experiments

Setup

Datasets: We conduct our experiments with partitioned datasets over three datasets: **CIFAR-10, CIFAR-100** (Krizhevsky, Hinton et al. 2009), and **Tiny-ImageNet** (Le and Yang 2015). Like (Zhu, Hong, and Zhou 2021; Wang et al. 2023a), we use the Dirichlet distribution Dir(α) on labels to simulate the data heterogeneity. We apply all the training samples to clients and use all the testing samples for performance evaluation.

Baselines: Alongside the fundamental **FedAvg** algorithm (McMahan et al. 2017), we evaluate several comparative methods: **1) Representative FL Model: FedProx** (Li et al. 2020b) introduces a proximal term to the local objective function to regularize client updates; **MOON** (Li, He, and Song 2021) uses the model-contrastive loss to narrow the representation between local and global models to alleviate data heterogeneity; **2) Federated Distillation Model: FedMD** (Li and Wang 2019) maintains the aggregated class scores to assist in regulating the local updating; **FedFusion** (Lin et al. 2020) implements data-based knowledge distillation using unlabeled training samples as a proxy dataset; **FedGen** (Zhu, Hong, and Zhou 2021) employs a data-free distillation framework where synthetic samples generated by the server directly regulate local model updates; **DaFKD** (Wang et al. 2023a) employs a domain discriminator for each client to identify the correlation factor between distillation and local samples, and **FedFD** (Li et al. 2025f) explores the feature-based distillation with both feature alignment and orthogonal projection.

Configurations: Unless otherwise mentioned, we set the local training epochs E = 10, communication rounds T = 200, and the client number K = 20 with an active ratio r = 0.4. For local training, the batch size is 64 and the weight decay is $1e-4$. The learning rate is 0.01 for distillation and 0.001 for training the local model. We employ ResNet-18 (He et al. 2016) as the basic backbone. Correspondingly, we used {CIFAR-100, Tiny-ImageNet, and ImageNet (Deng et al. 2009)} as proxy datasets for {CIFAR-10, CIFAR-100, and Tiny-ImageNet}. For the feature fusion module, we use the 1×1 convolution kernel as the projector and the SiLU function as the activation function. For the distribution prediction module, we set the number of the mixture components $J = 3$. All experiments were run twice, and we take each run's final ten rounds' accuracy and calculate the average value and standard deviation.

Performance Overview

Test Accuracy. Table 1 presents the comparative performance of our proposed FedCD method against baselines across three image datasets. Evaluated under varying levels of data heterogeneity for each dataset, FedCD achieves superior performance in most experimental settings. Among

Categories	Methods	CIFAR-10				CIFAR-100				Tiny-ImageNet			
		acc=80%	Δ	*acc*=85%	Δ	*acc*=55%	Δ	*acc*=60%	Δ	*acc*=30%	Δ	*acc*=35%	Δ
Classic-FL	FedAvg	63.00±3.50	27.3%↑	97.50±7.00	37.3%↑	74.00±2.50	97.3%↑	108.00±8.00	91.2%↑	96.00±10.50	23.9%↑	127.00±8.00	17.6%↑
	FedProx	65.50±2.00	31.3%↑	102.00±1.50	43.7%↑	70.00±3.50	86.7%↑	110.00±6.50	94.7%↑	109.50±12.50	41.3%↑	134.00±9.50	24.1%↑
	MOON	71.00±4.00	43.4%↑	100.00±5.50	40.8%↑	69.50±8.00	85.3%↑	96.50±5.00	70.8%↑	93.00±14.50	20.0%↑	147.50±7.00	36.6%↑
Federated Distillation	FedMD	> 200.00	∞%↑	> 200.00	∞%↑	> 200.00	∞%↑	> 200.00	∞%↑	> 200.00	∞%↑	> 200.00	∞%↑
	FedFusion	**46.50±5.00**	6.1%↓	78.00±4.50	9.9%↑	38.00±9.50	1.3%↑	59.00±3.00	4.4%↑	**73.50±12.00**	5.2%↑	116.00±11.50	7.4%↑
	FedGen	77.00±8.50	55.6%↑	92.00±9.50	29.6%↑	60.00±7.00	60.0%↑	88.50±11.00	56.6%↑	72.50±9.50	6.5%↑	114.50±16.00	5.6%↑
	DaFKD	66.50±4.00	34.3%↑	76.00±8.50	7.0%↑	59.00±3.50	57.3%↑	91.00±8.00	61.1%↑	80.00±7.50	3.2%↑	121.00±9.00	12.0%↑
	FedFD	52.00±2.00	5.1%↑	74.00±3.50	4.2%↑	46.50±6.50	18.6%↑	67.00±7.50	18.6%↑	84.50± 14.00	9.0%↑	126.00±13.00	16.7%↑
	FedCD	49.50±1.50	/	**71.00±6.50**	/	**37.50±8.00**	/	**56.50±9.00**	/	77.50±4.00	/	**108.00±10.50**	/

Table 2: Evaluation of different baselines on three datasets (α=1.0), in terms of the number of communication rounds to reach target test accuracy. Δ represents the percentage of improvement in communication efficiency. The best results are **bold**.

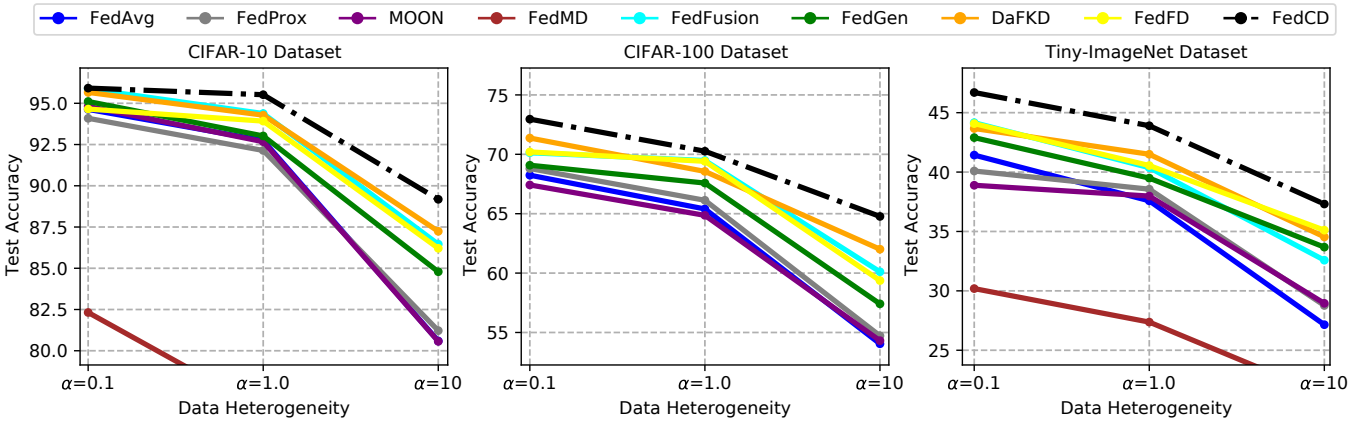

Figure 3: Performance of various baselines w.r.t data heterogeneity α.

these baselines, FedAvg, FedProx, and MOON represent three classic FL methods that optimize global models solely through local training and model aggregation. Unsurprisingly, their performance lags significantly behind federated distillation methods, though they exhibit minor differences among themselves. FedMD performs poorly due to the absence of aggregated parameters. FedGen and DaFKD are data-free federated distillation methods that eliminate the need for proxy datasets by employing generative models to synthesize data. However, their effectiveness exhibits strong dataset dependency: they tend to perform well on simple and homogeneous datasets but suffer from degraded synthetic data quality and consequently reduced distillation efficacy on complex dataset distributions. In contrast, FedCD outperforms all methods by leveraging feature and logit knowledge, achieving performance gains of up to 10.72%.

Communication Efficiency. Table 1 compares the communication efficiency of various methods by measuring the communication rounds required to achieve the best test accuracy, and Table 2 records the minimum communication rounds required to reach the target accuracy. The results demonstrate that federated distillation significantly enhances convergence efficiency, particularly during later stages when achieving superior model performance. On the CIFAR-10 dataset, classic FL methods initially exhibit comparable convergence rates but encounter performance plateaus, necessitating substantially more communication rounds for marginal improvements. Compared to other baselines, FedCD achieves the fastest convergence while maintaining training stability.

Method	CIFAR-10		CIFAR-100	
	α=1.0	α=0.1	α=1.0	α=0.1
FedCD	**95.52**	**89.18**	**70.26**	**64.77**
-w/o logit	93.01.	85.54	67.25	58.97
-w/o feature	94.38	86.46	69.48	60.11
-w/o fusion	94.12	84.99	68.86	62.16
-w/o distribution	92.88	85.33	67.81	58.84

Table 3: Ablation study of FedCD on two datasets.

Ablation Study. As shown in Table 3, we evaluate the effects of each module in our model via ablation studies. -w/o feature and -w/o logit denote the performance of our model without using feature-based loss and logit-based loss. - w/o fusion means we directly align the feature prediction, and - w/o distribution means we align the integrated feature prediction. First, FedCD benefits from both distillation losses, as using either logit or feature knowledge alone degrades model performance. Furthermore, FedCD -

Method	Model Heterogeneity			
	a-d-g	a-f	b-d-f	a-c-d-f-i
FedAvg-hetero	87.24±0.15	88.39±0.23	84.89±0.22	81.74±0.15
FedFusion	84.91±0.28	86.19±0.90	85.93±0.15	84.94±0.08
DaFKD	86.76±0.26	88.04±0.62	86.12±0.20	85.09±0.87
FedFD	88.68±0.66	88.59±0.50	87.19±1.01	86.12±0.73
FedCD	**90.48±0.26**	**89.88±0.24**	**88.98±0.73**	**87.84±0.80**

Table 4: Evaluation of combination of various client-side models levels for CIFAR-10 dataset ($\alpha = 1.0$).

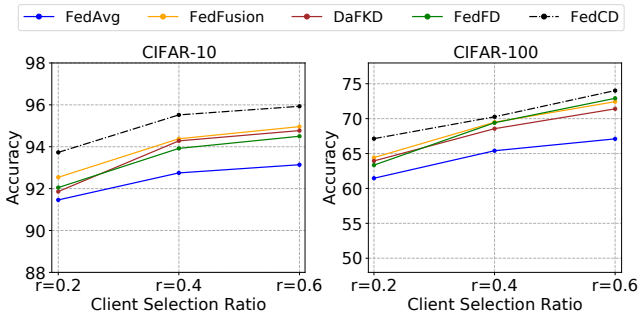

Figure 4: Performance comparison of various methods w.r.t. ratio r between active clients and total clients in each round.

w/o fusion demonstrates that naively summing logit-based and feature-based losses yields unstable performance gains, which is consistent with our findings in Fig.1. Finally, preprocessing feature predictions across network layers via the adaptive feature fusion module enhances sample feature extraction, thereby improving model performance. Experimental results verify that all modules are essential for training a robust global model through federated distillation.

Heterogeneous Setting. Table 1 and Fig.3 illustrate the variation in test accuracy across different levels of data heterogeneity. A consistent trend emerges where all methods demonstrate improved accuracy as data heterogeneity decreases. However, classic FL methods exhibit significant performance degradation under data heterogeneity. Federated distillation inherently mitigates heterogeneous impacts through proxy datasets, demonstrating more substantial performance gains in highly heterogeneous scenarios. Furthermore, we observe that data-free federated distillation methods face additional heterogeneity challenges, likely due to biases introduced during generator aggregation that compromise synthetic data quality. FedCD consistently maintains superior performance across all heterogeneous settings.

Although the above experiments evaluate within the data heterogeneity setting, the federated distillation framework is backbone-agnostic. Following (Diao, Ding, and Tarokh 2020), we establish model heterogeneity for experimentation. As shown in Table 4, we construct ten computational complexity levels $\{a, b, c, \ldots, j\}$ with a hidden channel decay rate of $\alpha = 10\%$. Under high model heterogeneity, the feature-based method FedFD and our proposed method marginally outperform logit-based approaches since logits fail to capture inter-model discrepancies.

Method	CIFAR-10		CIFAR-100	
	$K = 50$	$K = 100$	$K = 50$	$K = 100$
FedAvg	75.65±2.01	68.57±1.88	48.03±3.14	39.53±2.50
FedFusion	81.72±1.09	75.08±0.96	56.87±2.11	44.09±1.32
DaFKD	82.10±0.57	77.39±1.25	56.23±1.20	43.11±3.08
FedFD	80.45±0.78	74.47±1.53	57.28±2.30	44.91±3.00
FedCD	**83.96±1.32**	**79.29±0.98**	**59.74±2.35**	**47.02±2.88**

Table 5: The scalability of FedCD and other baselines on two datasets (α=0.01).

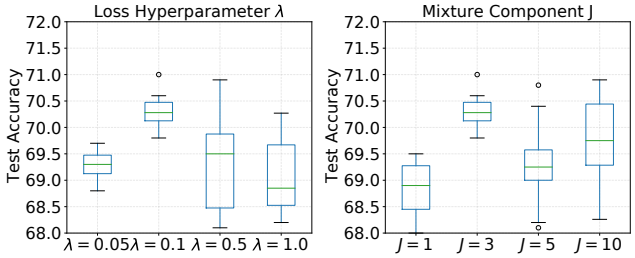

Figure 5: Performance under different configurations (a) hyperparameter λ, (b) mixture component J in the GMMs.

Parameter Sensitivity Analysis. As shown in Table 5, while all methods suffer marked performance degradation in large-scale settings where clients possess limited data samples, FedCD consistently outperforms baselines. This demonstrates robust effectiveness under challenging scalability conditions. Then, we test the client selection ratio r, hyperparameter λ, and the mixture component J. Experimental results demonstrate a positive correlation between increasing r ratios and performance improvement. In contrast, variations in λ and J do not substantially alter model performance, with FedCD consistently maintaining its leading advantage. This shows FedCD is insensitive to hyperparameter choices, demonstrating its parameter-related robustness.

Conclusion

We introduced FedCD, a novel framework to integrate feature-based and logit-based distillation modes within federated learning. FedCD effectively constrains feature representations by adaptively fusing predictions from different network layers and modeling them via GMMs, enabling synergistic optimization of the global model with logit-based distillation loss. Extensive experiments have demonstrated that FedCD can effectively enhance federated distillation.

Acknowledgments

This work is supported by the National Key Research and Development Program of China under grant 2024YFC3307900; the National Natural Science Foundation of China under grants 625B2073, 62376103, 62302184 and 62436003; Major Science and Technology Project of Hubei Province under grants 2025BAB011 and 2024BAA008; and Hubei Science and Technology Talent Service Project under grant 2024DJC078.

References

Bistritz, I.; Mann, A.; and Bambos, N. 2020. Distributed Distillation for On-Device Learning. In *Proceedings of Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems, NeurIPS*.

Chen, H.; and Chao, W. 2021. FedBE: Making Bayesian Model Ensemble Applicable to Federated Learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Chen, P.; Liu, S.; Zhao, H.; and Jia, J. 2021. Distilling knowledge via knowledge review. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5008–5017.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Diao, E.; Ding, J.; and Tarokh, V. 2020. Heterofl: Computation and communication efficient federated learning for heterogeneous clients. *arXiv preprint arXiv:2010.01264*.

Dong, J.; Li, H.; Cong, Y.; Sun, G.; Zhang, Y.; and Van Gool, L. 2024. No One Left Behind: Real-World Federated Class-Incremental Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4): 2054–2070.

Guo, J.; Chen, M.; Hu, Y.; Zhu, C.; He, X.; and Cai, D. 2020a. Reducing the teacher-student gap via spherical knowledge disitllation. *arXiv preprint arXiv:2010.07485*.

Guo, Q.; Wang, X.; Wu, Y.; Yu, Z.; Liang, D.; Hu, X.; and Luo, P. 2020b. Online knowledge distillation via collaborative learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 11020–11029.

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.

Heo, B.; Kim, J.; Yun, S.; Park, H.; Kwak, N.; and Choi, J. Y. 2019. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1921–1930.

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

Huang, S.; Fu, L.; Liao, T.; Deng, B.; Zhang, C.; and Chen, C. 2025. FedBG: Proactively Mitigating Bias in Cross-Domain Graph Federated Learning Using Background Data.

Huang, Y.; Yan, Z.; Shen, C.; Fang, F.; and Zhang, G. 2024. Harmonizing Knowledge Transfer in Neural Network with Unified Distillation. In *European Conference on Computer Vision*, 58–74. Springer.

Kang, J.; Xiong, Z.; Niyato, D.; Zou, Y.; Zhang, Y.; and Guizani, M. 2020. Reliable federated learning for mobile networks. *IEEE Wireless Communications*, 27(2): 72–80.

Karimireddy, S. P.; Kale, S.; Mohri, M.; Reddi, S. J.; Stich, S. U.; and Suresh, A. T. 2020. SCAFFOLD: Stochastic Controlled Averaging for Federated Learning. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, 5132–5143.

Khan, L. U.; Saad, W.; Han, Z.; Hossain, E.; and Hong, C. S. 2021. Federated learning for internet of things: Recent advances, taxonomy, and open challenges. *IEEE Communications Surveys & Tutorials*, 23(3): 1759–1799.

Kim, Y.; Park, J.; Jang, Y.; Ali, M.; Oh, T.-H.; and Bae, S.-H. 2021. Distilling global and local logits with densely connected relations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6290–6300.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Le, Y.; and Yang, X. 2015. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7): 3.

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

Li, L.; Fan, Y.; Tse, M.; and Lin, K.-Y. 2020a. A review of applications in federated learning. *Computers & Industrial Engineering*, 149: 106854.

Li, Q.; He, B.; and Song, D. 2021. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10713–10722.

Li, T.; Sahu, A. K.; Talwalkar, A.; and Smith, V. 2020b. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3): 50–60.

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

Li, Y.; Qin, Q.; Zhu, G.; Xu, W.; Wang, H.; Li, Y.; Zhang, R.; and Li, R. 2025a. A Systematic Survey on Federated Sequential Recommendation. arXiv:2504.05313.

Li, Y.; Shan, Y.; LIU, Y.; Wang, H.; Wang, C.; wangshi.ww; Wang, Y.; and Li, R. 2025b. Efficient Knowledge Transfer in Federated Recommendation for Joint Venture Ecosystem. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Li, Y.; Shan, Y.; Liu, Y.; Wang, H.; Wang, W.; Wang, Y.; and Li, R. 2025c. Personalized Federated Recommendation for Cold-Start Users via Adaptive Knowledge Fusion. In *Proceedings of the ACM on Web Conference 2025*, 2700–2709.

Li, Y.; Wang, H.; Qi, Y.; Liu, W.; and Li, R. 2025d. Re-fed+: A better replay strategy for federated incremental learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Li, Y.; Wang, H.; Xu, W.; Xiao, T.; Liu, H.; Tu, M.; Wang, Y.; Yang, X.; Zhang, R.; Yu, S.; et al. 2025e. Unleashing the power of continual learning on non-centralized devices: A survey. *IEEE Communications Surveys & Tutorials*.

Li, Y.; Wang, X.; Xu, W.; Wang, H.; Qi, Y.; Dong, J.; and Li, R. 2025f. Feature Distillation is the Better Choice for Model-Heterogeneous Federated Learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Li, Y.; Wang, Y.; Dong, J.; Wang, H.; Qi, Y.; Zhang, R.; and Li, R. 2025g. Resource-Constrained Federated Continual Learning: What Does Matter? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Li, Y.; Wang, Y.; Wang, H.; Qi, Y.; Xiao, T.; and Li, R. 2025h. FedSSI: Rehearsal-Free Continual Federated Learning with Synergistic Synaptic Intelligence. In *Forty-second International Conference on Machine Learning*.

Li, Y.; Xu, W.; Qi, Y.; Wang, H.; Li, R.; and Guo, S. 2024b. SR-FDIL: Synergistic Replay for Federated Domain-Incremental Learning. *IEEE Transactions on Parallel and Distributed Systems*, 35(11): 1879–1890.

Liao, T.; Fu, L.; Chen, J.; Wang, Z.; Zheng, Z.; and Chen, C. 2024. A swiss army knife for heterogeneous federated learning: Flexible coupling via trace norm. *Advances in Neural Information Processing Systems*, 37: 139886–139911.

Liao, T.; Xu, Z.; Hu, Q.; Dai, H.-N.; Huang, H.; Zheng, Z.; and Chen, C. 2025. FedBRB: A Solution to the Small-to-Large Scenario in Device-Heterogeneity Federated Learning. *IEEE Transactions on Mobile Computing*.

Lim, W. Y. B.; Luong, N. C.; Hoang, D. T.; Jiao, Y.; Liang, Y.-C.; Yang, Q.; Niyato, D.; and Miao, C. 2020. Federated learning in mobile edge networks: A comprehensive survey. *IEEE communications surveys & tutorials*, 22(3): 2031–2063.

Lin, T.; Kong, L.; Stich, S. U.; and Jaggi, M. 2020. Ensemble distillation for robust model fusion in federated learning. *Advances in Neural Information Processing Systems*, 33: 2351–2363.

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.

Meng, L.; Qi, Z.; Wu, L.; Du, X.; Li, Z.; Cui, L.; and Meng, X. 2024. Improving global generalization and local personalization for federated learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Miles, R.; Rodriguez, A. L.; and Mikolajczyk, K. 2021. Information theoretic representation distillation. *arXiv preprint arXiv:2112.00459*.

Nguyen, D. C.; Ding, M.; Pathirana, P. N.; Seneviratne, A.; Li, J.; and Poor, H. V. 2021. Federated learning for internet of things: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 23(3): 1622–1658.

Niu, Y.; Chen, L.; Zhou, C.; and Zhang, H. 2022. Respecting transfer gap in knowledge distillation. *Advances in Neural Information Processing Systems*, 35: 21933–21947.

Pfitzner, B.; Steckhan, N.; and Arnrich, B. 2021. Federated learning in a medical context: a systematic literature review. *ACM Transactions on Internet Technology (TOIT)*, 21(2): 1–31.

Qi, Z.; Meng, L.; Chen, Z.; Hu, H.; Lin, H.; and Meng, X. 2023. Cross-silo prototypical calibration for federated learning with non-iid data. In *Proceedings of the 31st ACM international conference on multimedia*, 3099–3107.

Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; Gatta, C.; and Bengio, Y. 2015. FitNets: Hints for Thin Deep Nets. arXiv:1412.6550.

Sanh, V. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Tian, Y.; Krishnan, D.; and Isola, P. 2019. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*.

Wang, C.; Chen, X.; Wang, J.; and Wang, H. 2022. ATPFL: Automatic Trajectory Prediction Model Design under Federated Learning Framework. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, New Orleans, LA, USA,June 18-24,June 18-24*, 6553–6562.

Wang, H.; Li, Y.; Xu, W.; Li, R.; Zhan, Y.; and Zeng, Z. 2023a. DaFKD: Domain-aware Federated Knowledge Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20412–20421.

Wang, H.; Xu, H.; Li, Y.; Xu, Y.; Li, R.; and Zhang, T. 2023b. FedCDA: Federated Learning with Cross-rounds Divergence-aware Aggregation. In *The Twelfth International Conference on Learning Representations*.

Wang, K.; He, Q.; Chen, F.; Chen, C.; Huang, F.; Jin, H.; and Yang, Y. 2023c. Flexifed: Personalized federated learning for edge clients with heterogeneous model architectures. In *Proceedings of the ACM Web Conference 2023*, 2979–2990.

Wu, G.; and Gong, S. 2021. Peer Collaborative Learning for Online Knowledge Distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence, AAAI*.

Wu, Y.; Zhang, S.; Yu, W.; Liu, Y.; Gu, Q.; Zhou, D.; Chen, H.; and Cheng, W. 2023. Personalized federated learning under mixture of distributions. In *International Conference on Machine Learning*, 37860–37879. PMLR.

Yim, J.; Joo, D.; Bae, J.; and Kim, J. 2017. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4133–4141.

Zhang, L.; Shen, L.; Ding, L.; Tao, D.; and Duan, L. 2022. Fine-tuning Global Model via Data-Free Knowledge Distillation for Non-IID Federated Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 10164–10173.

Zhao, B.; Cui, Q.; Song, R.; Qiu, Y.; and Liang, J. 2022. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 11953–11962.

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.