Mitigating Global Knowledge Forgetting via Adaptive Decoupled Knowledge Distillation

Hung-Chin Jang
Dept. of Computer Science
National Chengchi University
Taipei, Taiwan, R.O.C.
jang@cs.nccu.edu.tw

Ping-Hsien Chou
Department of Computer Science,
National Chengchi University
Taipei, Taiwan, R.O.C.
112971005@nccu.edu.tw

***Abstract*—Federated learning (FL) enables collaborative model training without exposing raw data; however, conventional aggregation schemes suffer from global knowledge forgetting when client data are non-independent and identically distributed (non-IID). This paper introduces FedADKD, an FL framework that performs Adaptive Decoupled Knowledge Distillation. FedADKD splits distillation into true-class (TCKD) and non-true-class (NCKD) components and adjusts the TCKD weight on each client based on quantified data heterogeneity, while keeping the NCKD weight constant to propagate shared dark knowledge. Clients with highly skewed data receive a lower TCKD weight, preserving local characteristics; balanced clients employ a higher weight, reinforcing inter-client consensus. The adaptive scheme incurs negligible communication overhead, as it transmits only a few scalar values per round. Comprehensive experiments on CIFAR-10 and CIFAR-100 with diverse non-IID partitions demonstrate that FedADKD consistently surpasses FedAvg and FedNTD in global accuracy and markedly lowers forgetting rates. Ablation studies confirm the independent contribution of adaptive TCKD weighting, validating its role in reconciling local adaptation with global integration. FedADKD therefore offers an efficient, privacy-preserving solution for heterogeneous FL deployments.**

Keywords—federated learning, knowledge distillation, non-IID, knowledge forgetting, decoupled knowledge distillation

I. INTRODUCTION

The exponential proliferation of Internet of Things (IoT) devices, mobile platforms, and smart-city applications has fundamentally altered the data storage landscape. Rather than residing in centralized repositories, data are increasingly generated and stored across distributed edge devices. While this decentralization mitigates single-point failure risks, it introduces significant challenges regarding data privacy and communication overhead. To address these concerns, Federated Learning (FL) [1] has emerged as a privacy-preserving paradigm that enables clients to collaboratively train a global model by exchanging model updates rather than raw data.

The standard FL algorithm, FedAvg [1], aggregates client parameters via weighted averaging and has proven effective in many general scenarios. However, in real-world deployments—such as smart healthcare, where hospital data vary by patient demographics and equipment—data are often Non-Independent and Identically Distributed (Non-IID). This statistical heterogeneity causes local updates to diverge, a phenomenon known as client drift [12], which leads to unstable convergence and catastrophic forgetting of global knowledge during aggregation [2]. To mitigate the adverse effects of non-IID data, Knowledge Distillation (KD) [6] has been adapted for FL. Unlike parameter averaging, KD operates at the output level, making it more robust to structural model discrepancies. Recent approaches, such as FedNTD [2], utilize KD to preserve global knowledge by distilling "Not-True Class" (non-target) information. While FedNTD effectively recovers some forgotten knowledge, it is theoretically incomplete; it neglects "True Class" (target) knowledge, which is essential for task-specific discrimination. The concept of Decoupled Knowledge Distillation (DKD) [7] posits that both True-Class KD (TCKD) and Non-True-Class KD (NCKD) are requisite for optimal transfer.

However, directly applying DKD to Federated Learning presents a critical unresolved challenge: adaptability. Existing distillation frameworks [10], [11], [18] typically employ fixed weighting coefficients (e.g., α, β) for all clients. This static approach fails to account for the severe heterogeneity in FL, where some clients possess highly biased data while others hold more balanced distributions. Although Adaptive Self-Distillation (ASD) [19] adjusts weights based on label statistics, it has not yet been integrated with DKD's decoupled formulation. Consequently, current state-of-the-art methods suffer from uneven contributions from distillation, leading to suboptimal performance in highly heterogeneous environments. To address these limitations, we propose FedADKD (Federated Adaptive Decoupled Knowledge Distillation). This novel framework integrates the decoupled distillation mechanism into FL while introducing a heterogeneity-aware adaptive weighting scheme. Specifically, FedADKD fixes the NCKD weight to ensure consistent dark knowledge transfer but adaptively assigns the TCKD weight based on each client's data divergence. Clients with high divergence are assigned lower TCKD weights to preserve unique local characteristics, while clients with balanced data receive higher TCKD weights to align closely with the global model.

The main contributions of this work are summarized as follows. We introduce FedADKD, the first Federated Learning framework to integrate Decoupled Knowledge Distillation with a client-specific adaptive weighting scheme, thereby addressing the rigidity inherent in existing global knowledge fusion methods. By dynamically adjusting TCKD weights, FedADKD significantly mitigates catastrophic forgetting and local bias; extensive experiments on CIFAR-10 and CIFAR-100 demonstrate that our method consistently outperforms FedAvg and FedNTD across diverse Non-IID settings. Furthermore, the proposed solution enhances convergence stability and accuracy without necessitating raw data sharing or introducing additional communication overhead, making it an efficient and practical solution for bandwidth-constrained edge environments.

II. Problem Statement & Analysis

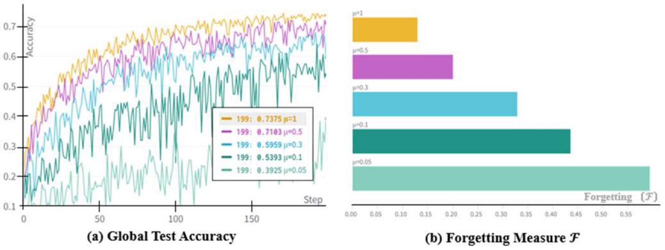
Existing methods face significant challenges under Non-IID conditions: FedAvg suffers from catastrophic forgetting, FedNTD neglects target-class information, and DKD relies on static weights that are unsuited to client diversity. To quantify these limitations, we simulate highly heterogeneous environments using CIFAR-10 [8] partitioned via Latent Dirichlet Allocation (LDA) [9], sharding [1], and mixed configurations.

A. Impact of Non-IID Distributions on Global Knowledge Forgetting

To simulate heterogeneity, we distribute CIFAR-10 across 100 clients using LDA with parameter μ. We train FedAvg for 200 rounds (10 clients per round, five local epochs). Fig. 1(a) demonstrates that lower μ values (higher heterogeneity) significantly degrade convergence accuracy, confirming that divergent client distributions impair global knowledge retention. To quantify this, we adopt the "forgetting measure" [16], defined as the maximum decline in class-wise accuracy:

$$\mathcal{F} = \frac{1}{C}\sum_{c=1}^{C} \max_{t\in\{1,\dots,T-1\}}\left(\mathcal{A}_c^{(t)} - \mathcal{A}_c^{(T)}\right) \tag{1}$$

where $\mathcal{A}_c^{(t)}$ denotes the accuracy for class c after round t and C is the number of classes. As depicted in Fig. 1(b), $\mathcal{F}$ increases sharply as μ decreases, confirming that severe heterogeneity intensifies global knowledge forgetting.



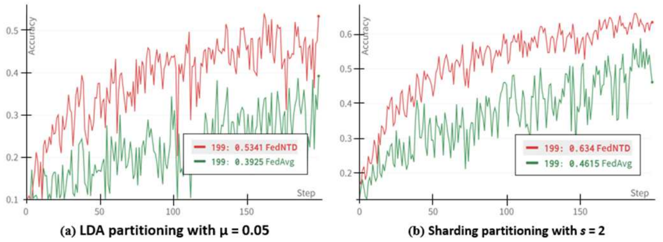
(a) Global Test Accuracy

(b) Forgetting Measure $\mathcal{F}$

Fig. 1. Global model results under different LDA heterogeneity levels (μ): (a) test accuracy; (b) forgetting measure $\mathcal{F}$.

B. TCKD and NCKD in Federated Learning

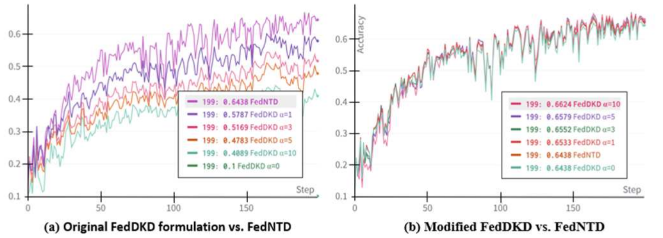
FedNTD mitigates forgetting via NCKD and consistently outperforms FedAvg under high heterogeneity (Fig. 2). To capture target-specific information, we formulate FedDKD, which combines TCKD and NCKD. Fig. 3(a) compares FedDKD with FedNTD under a fixed NCKD weight ($\beta = 1$) and varying TCKD weights ($\alpha \in \{0, 1, 3, 5, 10\}$). Across all values of α, FedDKD consistently underperforms FedNTD in terms of global model accuracy. The original DKD definition of NCKD retains the target-class dimension, leading to residual interference. In contrast, replacing it with the dimension-removed formulation of FedNTD significantly improves performance, as shown in Fig. 3(b), thereby validating the advantage of simultaneously applying TCKD and NCKD in the federated context.



(a) LDA partitioning with μ = 0.05

(b) Sharding partitioning with s = 2

Fig. 2. Global model accuracy of FedAvg vs. FedNTD under different data heterogeneity settings: (a) LDA partitioning with μ = 0.05, (b) Sharding partitioning with s = 2.



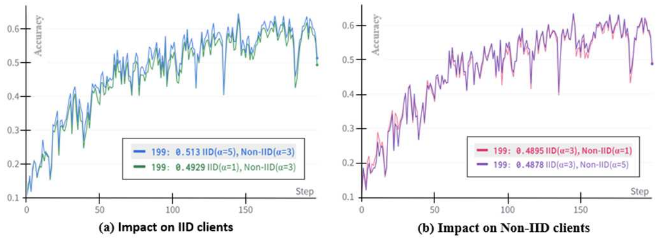
(a) Original FedDKD formulation vs. FedNTD

(b) Modified FedDKD vs. FedNTD

Fig. 3. Global model accuracy of FedDKD vs. FedNTD under LDA partitioning (μ = 0.1) with different TCKD weights (α): (a) original FedDKD formulation; (b) modified FedDKD with dimension-removed NCKD.

C. Effect of TCKD Weight Under Varying Heterogeneity

We hypothesize that the optimal TCKD weight (α) depends on local heterogeneity. While balanced (IID) clients benefit from strong TCKD to enhance discrimination, skewed (Non-IID) clients require weaker TCKD to prevent overfitting. This implies a fixed global α is suboptimal. Experiments in a mixed environment (1:9 IID/Non-IID ratio) confirm this divergence: IID clients improve with higher α (Fig. 4(a)), whereas Non-IID clients perform better with lower α (Fig. 4(b)), validating the necessity of a client-adaptive strategy.



(a) Impact on IID clients

(b) Impact on Non-IID clients

Fig. 4. Effect of TCKD weight (α) on global model accuracy in a 1:9 IID/Non-IID hybrid setup: (a) impact on IID clients; (b) impact on Non-IID clients.

III. Methodology

In the previous section, we demonstrated that jointly distilling target-class and non-target-class knowledge captures multi-level information more completely, and the weight assigned to target-class knowledge must adapt to each client's data heterogeneity. Building on these findings, we propose

FedADKD—Federated Learning with Adaptive Decoupled Knowledge Distillation—which decouples knowledge into TCKD and NCKD and dynamically adjusts the TCKD weight via a heterogeneity metric. The approach aims to suppress global knowledge forgetting while ensuring stable convergence under highly non-IID data.

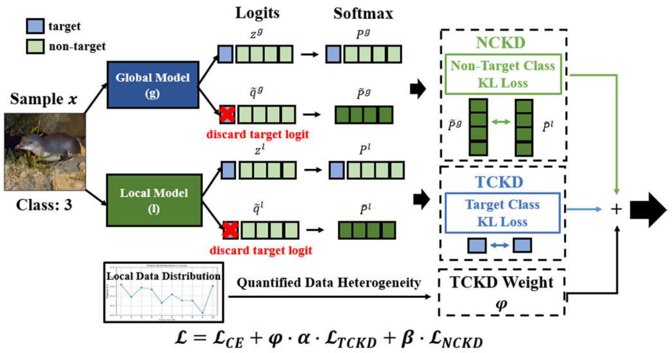


Fig. 5. The working mechanism of adaptive TCKD and NCKD in FedADKD.

A. FedADKD Overview

FedADKD combines decoupled distillation and adaptive weighting as shown in Fig. 5. Given a sample x with ground-truth label y:

- The global model g and local model l produce logits, which are converted into full probability vectors (P^g,P^l) and true−class−removed vectors ($\tilde{P}^g$,$\tilde{P}^l$).
- TCKD aligns P^g and P^l on the target class, yielding loss $\mathcal{L}_{TCKD}$; NCKD aligns $\tilde{P}^g$ and $\tilde{P}^l$ on the remaining classes, yielding $\mathcal{L}_{NCKD}$.
- Each client quantifies its data heterogeneity and generates an adaptive TCKD coefficient φ to prevent overly large or small emphasis on the target class.

The composite loss on client i is

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}}(P^l, \mathrm{y}) + \varphi \cdot \alpha \cdot \mathcal{L}_{TCKD}(P^l, P^g) + \beta \cdot \mathcal{L}_{NCKD}(\tilde{P}^l, \tilde{P}^g) \qquad (2)$$

where α and β are global hyperparameters.

B. Decoupled Knowledge Distillation

1) True-Class Knowledge Distillation (TCKD)

TCKD specifically aligns the prediction distributions of the true class between the local and global models. The TCKD loss is defined as the Kullback-Leibler (KL) divergence between the probability distributions on the true class.

$$\mathcal{L}_{TCKD}(P^l, P^g) = \sum_{c=1}^{C} \mathrm{P^g}(c) log \left[\frac{P^g(c)}{P^l(c)}\right],$$

$$\begin{cases} P^g\,(\mathrm{c}) = \dfrac{exp\left(z_y^g\right)}{\sum_j exp\left(z_j^g\right)} \\ P^l\,(\mathrm{c}) = \dfrac{exp\left(z_y^l\right)}{\sum_j exp\left(z_j^l\right)} \end{cases} \qquad (3)$$

z^g and z^l denote the logits produced by the global and local models, respectively. P^g and P^l are the corresponding target-class probability distributions obtained through a soft-max transformation. By designating the ground-truth class as the focal point of distillation, this loss enables the local model to replicate the global model's confidence on that label with higher fidelity, thereby strengthening the retention of target-class information within the federated framework.

2) Non-True-Class Knowledge Distillation (NCKD)

NCKD aligns the probability distributions of non-true classes between global and local models, thereby preserving global generalized knowledge across classes absent or underrepresented in local client data. The NCKD loss is also defined via the KL divergence as follows:

$$\mathcal{L}_{NCKD}(\tilde{P}^l, \tilde{P}^g) = \sum_{c=1, c\neq y}^{C} \tilde{P}^g(c) log \left[\frac{\tilde{P}^g(c)}{\tilde{P}^l(c)}\right],$$

$$\begin{cases} \tilde{P}^g\,(\mathrm{c}) = \dfrac{exp\left(\tilde{q}_k^g\right)}{\sum_k exp\left(\tilde{q}_k^g\right)} \\ \tilde{P}^l\,(\mathrm{c}) = \dfrac{exp\left(\tilde{q}_k^l\right)}{\sum_k exp\left(\tilde{q}_k^l\right)} \end{cases} \qquad (4)$$

$\tilde{q}^g$ and $\tilde{q}^l$ represent the non-target-class logit vectors obtained after masking the target-class entry, whereas $\tilde{P}^l$ and $\tilde{P}^g$ are their soft-max counterparts. Because clients frequently lack sufficient data for certain classes, NCKD allows the global model to retain latent knowledge of these non-local categories and prevents this data from being eroded during successive rounds of local training.

C. Adaptive TCKD Weight

In the previous section, we showed that the contribution of target-class knowledge distillation (TCKD) to both accuracy and knowledge retention is not constant; it must vary with each client's data heterogeneity. To quantify heterogeneity, we adopt the Gini index [17]:

$$G = 1 - \sum_{c=1}^{C} p_c^2, \qquad 0 \leq G \leq 1 - \frac{1}{C} \qquad (5)$$

where p_c is the class c proportion on the client and C is the total number of classes. Lower G values indicate highly skewed (non-IID) data, whereas higher values approach the IID case. To avoid extreme dispersion of G across clients in a single round, we apply a monotonic, nonlinear mapping

$$f(G_i, \delta) = log(1 + \delta \cdot G_i)_{i=1}^{n}, \qquad \delta > 0 \qquad (6)$$

where δ controls the amplification or compression of inter-client differences. After normalization, the adaptive TCKD weight for client i in round t is

$$\varphi_i = \frac{G_i(t)}{\sum_{j=1}^{|S^{(t)}|} G_j(t)} \cdot \left|S^{(t)}\right|, i \in j \qquad (7)$$

with $S^{(t)}$ the set of sampled clients. Hence, clients with highly skewed data automatically receive lower TCKD weight, mitigating over-emphasis on rare local classes. In contrast, clients with near-IID data are assigned a higher weight to exploit their richer class coverage. This adaptive scheme aligns with our earlier conclusion that dynamic adjustment of TCKD is essential for simultaneously accommodating diverse local distributions and global learning objectives.

D. *FedADKD Algorithm*

FedADKD integrates adaptive TCKD weighting with decoupled knowledge distillation to handle non-IID data while maintaining stable global learning. Algorithm 1 outlines the procedure.

Algorithm 1 A Federated Learning Approach based on Adaptive Decoupled Knowledge Distillation

```
Input: total rounds $T$, local epochs $E$, dataset $D$, sampled clients sets $S^{(t)} \subset S$ in round $t$, learning rate $\gamma$, TCKD Weight $\alpha$, NCKD Weight $\beta$
Initialize: Server initializes $w(0)$ for global server weight and the quantified data heterogeneity $G = 1$ for each node
for each communication round $t = 1, \cdots, T$ do
    Server samples nodes $S^{(t)}$ and calculates the adaptive TCKD weights $\varphi_i$, where $i \in S^{(t)}$ by ( 6 ) , ( 7 )
    Server broadcasts $\varphi_i$ and $\widetilde{\mathrm{w}}^{(t)} \leftarrow w^{(t)}$
    for each node $i \in S^{(t)}$ in parallel do
        Calculate quantified data heterogeneity $G_i$ by ( 5 )
        for Local Steps $e = 1, \cdots, E$ do
            for Batches $b = 1, \cdots, B$ do
                $\widetilde{\mathrm{w}}^{(t)} \leftarrow \widetilde{\mathrm{w}}^{(t)} - \gamma \nabla_{\mathrm{w}} \mathcal{L}(\widetilde{\mathrm{w}}^{(t)}; [D^i]_{\mathrm{j}})$ by ( 2 )
            end for
        end for
    end for
    Upload $G_i$ $and$ $\widetilde{\mathrm{w}}^{(t)}$ to server
    Server Aggregation : $w^{(t+1)} \leftarrow \frac{1}{|S^{(t)}|} i \in S^{(t)} \widetilde{w}_i^{(t)}$
end for
Server output :$w_T$
```

IV. Experiments

This section evaluates FedADKD under non-IID conditions. We outline the experimental setup, compare performance against baselines such as FedAvg and FedNTD, assess global knowledge retention, and conduct ablation studies to validate individual component contributions.

A. *Experimental Design*

1) *Datasets and Training Environment*

Two standard image classification benchmarks are used:

- CIFAR-10: Contains 10 classes, each with 32 × 32 color images, split into 50,000 training samples and 10,000 test samples.
- CIFAR-100: Comprises 100 classes with identical image format and structure to CIFAR-10, posing a more challenging classification task due to increased class diversity.

We simulate a federated environment with 100 clients (10 sampled per round), each training for five local epochs. The global model is a four-layer CNN optimized via Momentum SGD (learning rate 0.01, momentum 0.9, decay 0.99). All experiments are implemented in PyTorch on an NVIDIA RTX 3070 Ti GPU.

2) *Non-IID Data Simulation*

Three common partitioning strategies are adopted to simulate heterogeneous client distributions:

- Sharding: The dataset is label-sorted and split into equal-sized shards, which are randomly assigned so that each client receives s shards. For CIFAR-10 with s=2, each client obtains data from exactly two classes, generating extreme heterogeneity. Increasing s yields more balanced distributions.
- Latent Dirichlet Allocation (LDA): Client class proportions $p_k \sim Dir(\mu)$ are sampled from a Dirichlet distribution. Smaller μ produces higher skewness, whereas larger μ yields more uniform allocations.
- IID/Non-IID hybrid: To reflect mixed real-world settings, clients are partitioned into IID and highly Non-IID groups (the latter using LDA with μ=0.05).

These diverse partitioning schemes enable a systematic evaluation of FedADKD across varying degrees and types of data heterogeneity, thereby demonstrating its generalization capability and practical value in federated learning.

B. *Performance Analysis under Varying Data-Heterogeneity*

We evaluate FedADKD against representative baselines—including FedAvg, FedNTD, FedCurv [14], FedProx [3], FedNova [5], SCAFFOLD [4], and MOON [13]—across three data partitioning strategies. The results are summarized in Tables I–III.

1) *Sharding Strategy*

Table I details top-1 accuracy on CIFAR-10 and CIFAR-100 across varying shard sizes (s). Under extreme heterogeneity (CIFAR-10, s=2), FedADKD reaches 65.68%, significantly outperforming FedAvg (46.15%) and FedNTD (63.40%). As shard availability increases (s=5 and s=10), FedADKD consistently maintains the highest accuracy, notably achieving 33.19% on CIFAR-100 with s=5.

TABLE I. Top-1 accuracy (%) on CIFAR-10 and CIFAR-100 with Sharding

Non-IID Partition Strategy : Sharding				
Method	CIFAR-10			CIFAR-100
	$s = 2$	$s = 5$	$s = 10$	$s = 5$
FedAvg	46.15	64.64	72.15	24.70
FedNTD	63.40	73.74	75.50	32.11
FedNTD+ASD	56.22	71.59	75.09	23.77
FedCurv	51.07	61.64	69.63	21.23
FedProx	43.79	60.65	67.49	24.71
FedNova	44.26	62.95	70.39	21.86

SCAFFOLD	46.76	73.18	76.24	32.62
MOON	43.06	64.04	72.37	24.50
FedADKD	**65.68**	**75.48**	**77.74**	**33.19**

2) *LDA Partition*

Table II summarizes the performance under varying levels of data heterogeneity using the LDA partitioning scheme, where the Dirichlet parameter μ controls the degree of class imbalance. When $\mu = 0.05$, FedADKD achieves 56.46% accuracy on CIFAR-10 and 37.62% on CIFAR-100. Across all values of μ considered, FedADKD consistently attains the highest accuracy among the evaluated methods, demonstrating robustness to different levels of data heterogeneity.

TABLE II. TOP-1 ACCURACY (%) ON CIFAR-10 AND CIFAR-100 WITH LDA

Non-IID Partition Strategy : LDA				
Method	CIFAR-10			CIFAR-100
	$\mu = 0.05$	$\mu = 0.1$	$\mu = 0.5$	$\mu = 0.05$
FedAvg	39.25	53.93	71.03	34.50
FedNTD	53.41	64.38	73.42	36.66
FedNTD+ASD	46.57	62.59	72.29	34.10
FedCurv	46.02	49.40	67.79	32.52
FedProx	44.32	47.55	64.08	29.21
FedNova	21.44	31.51	67.24	30.83
SCAFFOLD	10.00	27.31	72.59	37.18
MOON [25]	37.65	52.63	70.87	33.79
FedADKD	**56.46**	**65.98**	**75.41**	**37.62**

3) *IID/Non-IID Hybrid Partition*

Table III presents the performance of FedADKD and baseline methods under mixed data distributions, where clients are divided into IID and highly non-IID groups (LDA with $\mu = 0.05$) at ratios of 1:9, 3:7, and 5:5. Notably, even with only 10% of clients holding IID data (1:9), FedADKD achieves 55.83% accuracy on CIFAR-10 and 40.10% on CIFAR-100, outperforming all compared methods. This result highlights FedADKD's ability to effectively leverage a small fraction of IID data to stabilize training. The performance advantage remains consistent as the proportion of IID clients increases.

TABLE III. TOP-1 ACCURACY (%) ON CIFAR-10 AND CIFAR-100 WITH IID/NON-IID HYBRID

Mix Partition Strategy : IID and Non-IID （LDA $\mu = 0.05$）				
Method	CIFAR-10			CIFAR-100
	1：9	3：7	5：5	1：9
FedAvg	27.86	51.59	53.19	35.35
FedNTD	44.31	60.24	67.54	38.23
FedNTD+ASD	54.51	63.06	71.00	36.52
FedCurv	32.22	51.44	59.46	33.00
FedProx	38.71	55.01	64.16	29.72
FedNova	13.98	37.67	42.25	32.61
SCAFFOLD	10.00	10.00	10.00	38.29
MOON	28.04	52.46	55.64	34.78
FedADKD	**55.83**	**66.44**	**73.11**	**40.10**

C. *Global Knowledge Retention*

This section evaluates the global knowledge retention capability of FedADKD under non-IID federated settings. Two evaluation metrics are employed: the Forgetting Measure [15] and Knowledge Outside of Local Distribution (KOLD) [2]. The results are compared with those of FedAvg and FedNTD.

1) *Forgetting Measure Results*

Experiments are conducted on CIFAR-10 using mixed IID and non-IID client splits, generated via LDA with $\mu = 0.05$, at ratios of 1:9, 3:7, and 5:5. Table IV reports the forgetting rates across methods, where FedADKD consistently achieves the lowest values. In the most heterogeneous setting (1:9), FedADKD reduces forgetting to 38.90%, outperforming FedNTD by approximately 13% and FedAvg by approximately 33%. This advantage remains as heterogeneity decreases (3:7 and 5:5), indicating that the adaptive decoupled distillation mechanism in FedADKD effectively balances TCKD and NCKD to enhance generalization.

TABLE IV. FORGETTING RATES (%) OF FEDAVG, FEDNTD, AND FEDADKD UNDER MIXED IID/NON-IID SPLITS.

Mix Partition Strategy : IID and Non-IID			
Method	CIFAR-10		
	1：9	3：7	5：5
FedAvg	71.91	46.95	44.29
FedNTD	51.97	35.91	25.78
FedADKD	**38.90**	**28.74**	**17.99**

2) *Knowledge Outside of Local Distribution (KOLD)*

To assess knowledge retention, we evaluate FedADKD on CIFAR-10 using mixed IID/Non-IID splits (1:9, 3:7, 5:5). Fig. 6 compares performance on in-local (seen) versus out-local (unseen) classes. While in-local accuracy remains comparable across methods, FedADKD significantly outperforms FedAvg and FedNTD on out-local classes, particularly in the high-heterogeneity 1:9 setting. This superior cross-client transfer enables FedADKD to achieve the highest global accuracy with lower variance, demonstrating robust resilience to data sparsity.

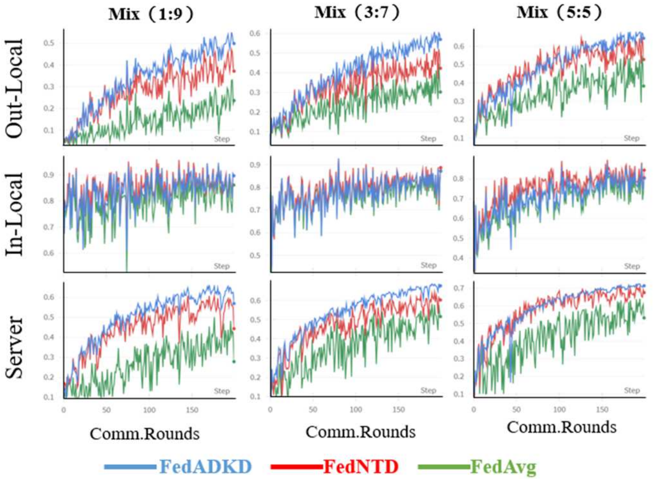
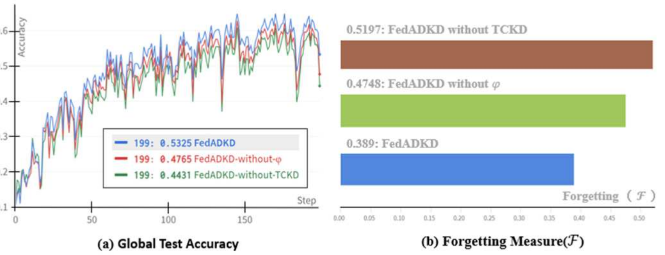


Fig. 6. The comparisons among FedAvg, FedNTD, and FedADKD on CIFAR-10 across mixed IID/Non-IID settings (1:9, 3:7, 5:5). Accuracy is shown for each client's seen classes (in-local), unseen classes (out-local), and the global server test set.

D. Ablation Studies

We conduct an ablation study on CIFAR-10 (1:9 mixed ratio, LDA μ=0.05) to isolate the contributions of the adaptive weight (φ) and the TCKD branch. Fig. 7(a) reveals that removing φ degrades accuracy from 55.83% to 48.54%, while excluding TCKD entirely further drops performance to 44.31% and introduces severe oscillation. Knowledge retention (Fig. 7(b)) mirrors this trend: forgetting increases from 38.90% (Full) to 47.48% (w/o φ) and 51.97% (w/o TCKD). Together, these components yield an 11% accuracy gain and a 13% reduction in forgetting, validating the critical role of adaptive target-class distillation in balancing performance and stability.



(a) Global Test Accuracy

(b) Forgetting Measure(𝓕)

Fig. 7. Performance of FedADKD and its ablated variants under mixed IID/Non-IID data distribution (1:9 ratio) on CIFAR-10: (a) Global Test Accuracy; (b) Forgetting Measure (𝓕).

V. CONCLUSION

This paper presented FedADKD, a federated distillation framework that mitigates global knowledge forgetting in non-independent and identically distributed (non-IID) settings. Building upon the decoupled knowledge distillation paradigm, FedADKD introduces a client-adaptive mechanism that dynamically balances target-class and non-target-class knowledge distillation, with modulation guided by each client's data heterogeneity. Extensive experiments on CIFAR-10 and CIFAR-100 under various partitioning schemes demonstrate that FedADKD consistently outperforms existing baselines in both accuracy and knowledge retention. Importantly, these improvements are achieved without requiring synthetic data, public datasets, or additional communication overhead, highlighting the practicality and scalability of FedADKD for real-world federated learning applications.

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 Proc. 20th Int. Conf. Artificial Intelligence and Statistics (AISTATS), Fort Lauderdale, FL, USA, Apr. 2017, pp. 1273–1282.

[2] G. Lee, M. Jeong, Y. Shin, S. Bae, and S. Y. Yun, "Preservation of the global knowledge by not-true distillation in federated learning," in Advances in Neural Information Processing Systems, vol. 35, 2022, pp. 38461–38474.

[3] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in Proc. Machine Learning and Systems, vol. 2, 2020, pp. 429–450.

[4] S. P. Karimireddy et al., "SCAFFOLD: Stochastic controlled averaging for federated learning," in Proc. 37th Int. Conf. Machine Learning (ICML), 2020, pp. 5132–5143.

[5] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 7611–7623.

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

[7] B. Zhao, Q. Cui, R. Song, Y. Qiu, and J. Liang, "Decoupled knowledge distillation," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11953–11962.

[8] A. Krizhevsky, V. Nair, and G. Hinton, "The CIFAR-10 and CIFAR-100 datasets," [Online]. Available: https://www.cs.toronto.edu/kriz/cifar.html, 2009.

[9] M. Luo et al., "No fear of heterogeneity: Classifier calibration for federated learning with non-IID data," in Advances in Neural Information Processing Systems, vol. 34, 2021, pp. 5972–5984.

[10] C. Wu, F. Wu, L. Lyu, Y. Huang, and X. Xie, "Communication-efficient federated learning via knowledge distillation," Nat. Commun., vol. 13, art. no. 2032, 2022.

[11] Z. Zhou et al., "A decentralized federated learning based on node selection and knowledge distillation," Mathematics, vol. 11, no. 14, art. no. 3162, 2023.

[12] G. I. Parisi et al., "Continual lifelong learning with neural networks: A review," Neural Netw., vol. 113, pp. 54–71, May 2019.

[13] Q. Li, B. He, and D. Song, "Model-contrastive federated learning," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10713–10722.

[14] N. Shoham et al., "Overcoming forgetting in federated learning on non-IID data," arXiv preprint arXiv:1910.07796, 2019.

[15] A. Chaudhry, P. K. Dokania, T. Ajanthan, and P. H. S. Torr, "Riemannian walk for incremental learning: Understanding forgetting and intransigence," in Proc. Eur. Conf. Computer Vision (ECCV), 2018, pp. 532–547.

[16] Z. Wang et al., "Learning to prompt for continual learning," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2022, pp. 139–149.

[17] W. Y. Loh, "Classification and regression trees," Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 1, no. 1, pp. 14–23, Jan. 2011.

[18] L. Su, D. Wang, and J. Zhu, "DKD-pFed: A novel framework for personalized federated learning via decoupling knowledge distillation and feature decorrelation," Expert Syst. Appl., vol. 259, art. no. 125336, 2025.

[19] M. Yashwanth, G. K. Nayak, A. Singh, Y. Simmhan, and A. Chakraborty, "Adaptive self-distillation for minimizing client drift in heterogeneous federated learning," arXiv preprint arXiv:2305.19600, 2023.