Distilling A Universal Expert from Clustered Federated Learning

Zeqi Leng[1,2], **Chunxu Zhang**[1,2*], **Guodong Long**[3], **Riting Xia**[4], **Bo Yang**[1,2*]
[1]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, China
[2]College of Computer Science and Technology, Jilin University, China
[3]Australian Artificial Intelligence Institute, FEIT, University of Technology Sydney
[4]College of Computer Science, Inner Mongolia University, Hohhot, China
lengzq22@mails.jlu.edu.cn, zhangchunxu@jlu.edu.cn, guodong.long@uts.edu.au
xiart@imu.edu.cn, ybo@jlu.edu.cn

Abstract

Clustered Federated Learning (CFL) addresses the challenges posed by non-IID data by training multiple group- or cluster-specific expert models. However, existing methods often overlook the shared information across clusters, which represents the generalizable knowledge valuable to all participants in the Federated Learning (FL) system. To overcome this limitation, this paper introduces a novel FL framework that distills a universal expert model from the knowledge of multiple clusters. This universal expert captures globally shared information across all clients and is subsequently distributed to each client as the initialization for the next round of model training. The proposed FL framework operates in three iterative steps: (1) local model training at each client, (2) cluster-specific model aggregation, and (3) universal expert distillation. This three-step learning paradigm ensures the preservation of fine-grained non-IID characteristics while effectively incorporating shared knowledge across clusters. Compared to traditional gradient-based aggregation methods, the distillation-based model aggregation introduces greater flexibility in handling model heterogeneity and reduces conflicts among cluster-specific experts. Extensive experimental results demonstrate the superior performance of the proposed method across various scenarios, highlighting its potential to advance the state of CFL by balancing personalized and shared knowledge more effectively.

1 Introduction

Federated learning (FL) has emerged as a promising paradigm for privacy-preserving distributed training, enabling collaborative model development without exposing sensitive personal data [Nguyen *et al.*, 2021; Beltrán *et al.*, 2023; Zhang *et al.*, 2023]. The primary goal of FL is to train a high-quality consensus model through joint optimization. However, deploying FL across distributed clients is nontrivial due to the statistical heterogeneity of local data (i.e., non-IID). Standard averaging-based aggregation method [McMahan *et al.*, 2017] tends to result in biased local updates, degrading the performance of the consensus model and slowing convergence. Prior works [Karimireddy *et al.*, 2020; Li *et al.*, 2020] have introduced correction mechanisms that align local and global update directions. While effective in reducing update bias, these approaches compromise local model personalization, a key factor in ensuring high performance under data heterogeneity.

Clustered Federated learning (CFL) effectively compensates for local personalization by training multiple cluster-level consensus models for heterogeneous devices [Tu *et al.*, 2025; Huang *et al.*, 2024; Ma *et al.*, 2023]. Unlike traditional FL, CFL inherently preserves individual data characteristics through its clustering mechanism. Yet even state-of-the-art CFL approaches still face challenges in various practical settings. A key problem lies in the limited information flow across clusters: small clusters are prone to overfitting, and large clusters typically converge to local optima. As a result, CFL consensus models tend to overfit local personalization, limiting their generalization capability.

In fact, neither traditional FL nor CFL methods effectively balance personalized knowledge with global consensus. An ideal solution aims to build a consensus model that preserves client-specific features. Yet, this is a significantly challenge, as naive aggregation of personalized knowledge tends to induce cluster-level model bias. These observations prompt us to explore a new question:

Can the consensus model be successfully conducted without undermining personalization?

Transferring knowledge across heterogeneous clusters offers a potential solution. Recent advances in knowledge distillation (KD) have shown promise in facilitating such knowledge migration [Wang and Yoon, 2021]. However, a major limitation of traditional KD methods is their reliance on auxiliary datasets. The choice of auxiliary data significantly impacts model performance and introduces additional computational overhead [Wang *et al.*, 2024]. These limitations motivate us to adopt a data-free knowledge distillation (DFKD) approach [Lin *et al.*, 2020a], in which a generator is trained through model inversion to synthesize pseudo-data on the fly, enabling knowledge transfer from the multiple teacher networks to the student network.

To this end, we face several key challenges: **C1**. How to

*Corresponding authors.

design a training method with multiple teacher networks such that heterogeneous knowledge remains non-conflicting? **C2**. How to effectively extract and ensemble cross-cluster knowledge, given that dynamic clustering induces random inter-cluster distribution shifts? **C3**. How to mitigate privacy risks in clustering, as the direct exposure of similarity values in existing CFL methods increases the risk of re-identification.

To simultaneously address all the aforementioned challenges, we propose a novel FL framework, DisUE (Distilled Universal Expert)), which enhances the consensus model while preserving local personalization. DisUE follows a three-stage learning paradigm: local model training, iterative clustering, and universal expert distillation. In particular, to tackle **C1**, we construct a universal expert model from CFL and propose a category cluster-level knowledge migration mechanism to enable single expert-to-student distillation. To tackle **C2**, we design two adaptive components at the category cluster-level to handle dynamic inter-cluster heterogeneity. By leveraging category statistics, these components guide both model inversion and ensemble. To tackle **C3**, we introduce a lightweight similarity encryption protocol that prevents direct exposure of similarity values.

In a nutshell, our main contributions are as follows:

- **Novel Framework for FL**: We propose DisUE, a distillation-enhanced framework from CFL, which can maintain both intra- and inter-group knowledge in FL.
- **Adaptive Category Cluster-Distillation Mechanism**: We devise an adaptive Group Label Sampler (*GLS*) and Group Weighting Factors (*GWF*) to process heterogeneous knowledge across groups. It incorporates group-based distillation to reduce distillation aggregation overhead and globally shares a model to improve the performance of minority groups.
- **Flexible plugin.** As a flexible algorithm that is orthogonal to existing CFL optimizers (IFCA, CFL, CFL-GP, PACFL). Our DisUE can easily enhance existing CFL methods, demonstrating the generality and compatibility of our inter-group aggregation mechanism.
- **Superior performance.** Extensive experiments on three real-world datasets against diverse advancing baselines prove the consistently state-of-the-art performance of our proposed DisUE.

2 Related Work

2.1 Clustered Federated Learning

CFL addresses statistical heterogeneity by grouping clients into clusters with homogeneous data distributions [Long *et al.*, 2023; Ma *et al.*, 2022]. Existing CFL methods [Ghosh *et al.*, 2020; Sattler *et al.*, 2020; Duan *et al.*, 2021] primarily focus on client partitioning to derive accurate personalized expert models. Prior work has explored clustering under various assumptions, such as temporal domain shifts between training and test data [Fan *et al.*, 2024], joint label and feature skewness [Guo *et al.*, 2024], and multiple types of data distribution shifts [Ruan and Joe-Wong, 2022]. Other approaches optimize aggregation schedules [Kim *et al.*, 2024]. However, these methods restrict inter-cluster communication, causing expert models to overfit limited client data and suffer in generalizability. We bridge this gap by designing an expert model based on CFL, and improving its generalization via cluster-level DFKD transfer.

2.2 Data-free Knowledge Distillation

DFKD removes the need for auxiliary datasets in KD by transferring knowledge through model output exchange [Yoo *et al.*, 2019]. Current approaches fall into two categories based on source networks: single-teacher DFKD [Shin and Choi, 2024; Lin *et al.*, 2020b] and multi-teacher DFKD [Hao *et al.*, 2021; Ye *et al.*, 2020]. While most DFKD methods operate at the client level, they require more training rounds than aggregation methods [McMahan *et al.*, 2017; Karimireddy *et al.*, 2020] due to the larger scale of pseudo-data generated relative to original samples. To address this limitation, we introduce a cluster-level DFKD approach that lowers communication costs via group knowledge fusion.

3 Problem Statement

CFL Pipeline. CFL follows an iterative three-stage framework [Liu *et al.*, 2024]:

- *Local training (L-phase).* Each client performs local training and transmits model updates to the server. Let there be N clients, where the i-th client holds a private dataset $D_i = \{(x_i^{(l)}, y_i^{(l)})\}_{l=1}^{n_i}$ with n_i samples. Client i updates its local parameters ω_i by minimizing the client-level empirical risk:

$$\min_{\omega_i} \frac{1}{n_i} \sum_{l=1}^{n_i} f\left(\omega_i; x_i^{(l)}, y_i^{(l)}\right) \tag{1}$$

  where $f(\cdot)$ denotes the per-sample loss. All clients then send their parameters $\{\omega_i\}_{i=1}^{N}$ to the server.
- *Clustering (C-phase).* The server partitions clients into K disjoint clusters $\mathcal{C} = \{C_1, \ldots, C_K\}$ using similarity metrics (e.g., cosine distance).
- *Group model aggregation (G-phase).* CFL maintains K group models for N clients. Within cluster C_k, m clients with identically distributed data collaboratively train a group model ω_{C_k}. We formulate the cluster-level optimization objective as:

$$\min_{\omega_{C_k}} \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}_i(\omega_i; D_i) \tag{2}$$

  Where $\mathcal{L}$ denotes the loss function for intra-group clients. The server subsequently distributes the updated cluster models $\{\omega_{C_k}\}_{k=1}^{K}$ to their respective clusters.

Emerging Requirements from CFL. The core limitation arises from strict inter-cluster communication constraints. As prior analyses demonstrate, overcoming this challenge requires global information sharing – a critical need that drives our architectural redesign of the *Group model aggregation*:

- *G-phase*: Directly applying existing methods to inter-cluster knowledge transfer encounters two fundamental barriers: (1) **Distribution Discrepancy**. The inherent heterogeneity across clusters in CFL leads to systematic

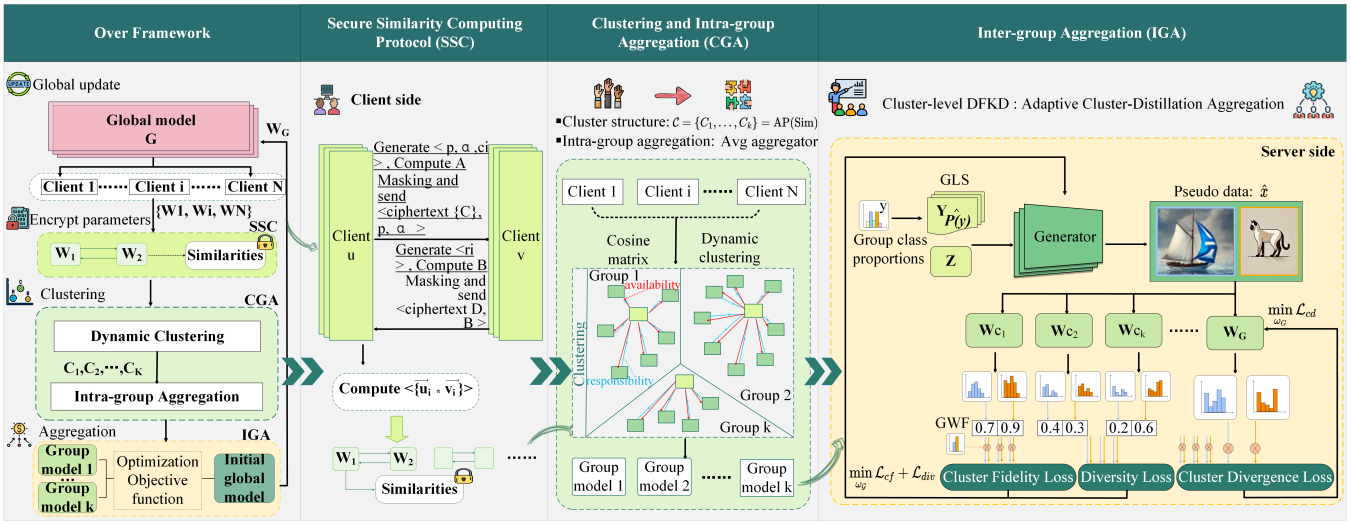


Figure 1: Overview of DisUE workflow. (1) **Local Training**: Clients first train local models using private data. (2) **Secure Similarity Computing**: Clients employ the Secure Similarity Computing protocol to encrypt model parameters. (3) **Clustering and Intra-group Aggregation**: The server partition clients into clusters. Each cluster performs FedAvg averaging within its group. (4) **Inter-group Aggregation**: The framework distills a universal expert model while preserving data privacy.

bias when using traditional aggregation methods (e.g., FedAvg). (2) **Adaptation rigidity**. Static knowledge transfer mechanisms fail to respond to dynamic distribution shifts between clusters. These challenges demand a new CFL paradigm that supports adaptive transfer and dynamically adjusts to distributional changes induced by shifting cluster structures.

4 Methodology

In this section, we introduce DisUE, a cross-cluster federated learning framework that overcomes the limitations discussed in Section Introduction (Figure 1). The framework operates through three core components: *(1) Clustering and Intra-group Aggregation (CGA)*: Partitions clients using cosine similarity-based clustering and extracts group-specific knowledge through intra-group model aggregation. *(2) Inter-group Aggregation (IGA)*: Distills heterogeneous cluster expertise into a global model via data-free adversarial learning. *(3) Secure Similarity Computing Protocol (SSC)*: Encrypts client parameters using PCSC, ensuring security against parameter leakage during similarity computation.

4.1 Clustering and Intra-group Aggregation

Principles. This stage achieves client dynamic clustering through angular similarity measurement, hyperparameter-free operation, and non-IID robustness. We quantify pairwise client relationships using cosine similarity between gradient updates. Unlike density-based methods [McInnes *et al.*, 2017; Ester *et al.*, 1996] that require predefined neighborhood thresholds, our design eliminates hyperparameters through message passing. Additionally, these algorithms exhibit unstable cluster structure under data heterogeneity. Our goal is to adopt a dynamic clustering mechanism that maintains robustness in extreme non-IID scenarios while enabling seamless integration with FL frameworks—critical for realizing a universal expert model. These requirements drive our adoption of Affinity Propagation (AP) [Frey and Dueck, 2007], which automatically identifies exemplar clients through iterative responsibility/availability updates.

Clustering. We compute pairwise client similarities using cosine distance between parameter vectors ω_u and ω_v:

$$\mathbf{Sim}(u, v) = \frac{\omega_u \cdot \omega_v}{\|\omega_u\| \, \|\omega_v\|} \tag{3}$$

where $\|\omega_u\|$ denotes the length of the parameter vector ω_u. Let $\mathbf{Sim}$ be the cosine similarity matrix. The AP clustering obtains K disjoint groups as follows:

$$\mathcal{C} = \{C_1, \ldots, C_K\} = \mathrm{AP}(\mathbf{Sim}) \tag{4}$$

Intra-group Aggregation. For each cluster $C_j \in \mathcal{C}$, we aggregate member models through Eq. (2), denotes as $\{\omega_{C_1}, \ldots, \omega_{C_K}\}$. Yielding specialist models $\{\omega_{C_j}\}_{j=1}^{K}$ that capture cluster-specific knowledge patterns.

4.2 Inter-group Aggregation

Principles. This stage addresses two co-existing challenges: adaptive cluster-wise label distribution shifts and cross-cluster distillation aggregation. Some aggregation methods based on subgroups are constrained by inter-group node symmetry or cluster topology symmetry [Jahani-Nezhad *et al.*, 2022; So *et al.*, 2021], while global averaging suppresses cluster-specific knowledge. Our solution breaks this stalemate through adaptive category group-level DFKD aggregation. Specifically, (1) Dynamically adjusting cluster contributions through adaptive ***GLS*** and ***GWF***, and (2) establishing cross-heterogeneous cluster knowledge transfer via prediction logit alignment between cluster experts and the global model. We fuse ***GLS*** and ***GWF*** into the cluster-level DFKD aggregation through a distillation-driven ensemble, creating a flexible mechanism for cross-cluster knowledge transfer.

Handing Cluster-level Distribution Imbalance. Label distribution imbalance constitutes a primary manifestation of distributional discrepancies across heterogeneous clusters. The dynamic cluster assignment mechanism in CFL induces stochastic cluster-level label distribution variations across training rounds. Within cluster-level DFKD methods, this bias manifests through two critical pathways: (1) leveraging cluster-specific distribution characteristics to optimize generator training, and (2) aggregating cluster models according to their category-specific contributions. Inspired by [Zhang *et al.*, 2022] and the underlying principle of counters, we proposed ***GLS*** and ***GWF***.

The GLS samples representative labels guided by intra-cluster category statistics:

$$\hat{p}(y) \propto \sum_{k=1}^{K} \sum_{i=1}^{m} \mathbb{E}_{\left(x_i^{(l)}, y_i^{(l)}\right) \sim \mathcal{D}_i} \left[\mathbf{I}_{y_i=y}\right] = \sum_{k=1}^{K} n_y^k \quad (5)$$

where $\mathbf{I}(\cdot)$ denotes the indicator function, and n_y^k represents the cardinality of category y samples in cluster k. The GW dynamically determines category importance weights per cluster:

$$\alpha_y^k = n_y^k / \sum_{k=1}^{K} \sum_{i \in C_k} n_y^i \quad (6)$$

This adptaive weighting scheme ensures equitable integration of category-specific knowledge across clusters. These adaptive components regulate cluster-level knowledge transfer through accumulated category distribution statistics, enabling effective handling of non-IID data scenarios.

Cluster-level DFKD Aggregation Mechanism. Our framework aims to facilitate cross-cluster knowledge transfer through data-free distillation, enabling the global model ω_G to assimilate heterogeneous cluster characteristics. We preserve inter-group knowledge foundations by maintaining cluster-averaged feature representations through aggregated cluster models serving as student networks. These cluster models subsequently function as teacher networks to supervise student model distillation.

The teacher-student knowledge transfer occurs through data-free distillation where teacher models guide student models to capture their specific data distributions. However, client-side generator deployment imposes prohibitive computational overhead on edge devices. To overcome this limitation, we design a generator $\mathcal{G}$ at the server that synthesizes pseudo-samples $\hat{x}$ conditioned on category labels y:

$$\hat{x} = \mathcal{G}(\theta_{\mathcal{G}}; z, y) \quad (7)$$

where $z \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$ represents random Gaussian noise vectors, and $\theta_{\mathcal{G}}$ denotes the generator's parameters.

The cross-cluster distillation method achieves knowledge transfer through prediction discrepancy minimization between ensemble teachers and student network on generated samples, effectively capturing teacher-specific distribution patterns. We formulate the cluster distillation objective as:

$$\min_{\omega_G} \mathcal{L}_{cd} := \mathbb{E}_{y \sim \hat{p}(y)} \mathbb{E}_{z \sim \mathcal{N}(\mathbf{0}, \mathbf{1})} \left[\sum_{k=1}^{K} \alpha_y^k \mathcal{L}_{cd}^k \right] \quad (8)$$

where $\hat{p}(y)$ denotes the cluster-level label distribution from Eq. (5), and α_y^k represents the adaptive cluster weights defined in Eq. (6) . To enable precise cluster knowledge transfer, we leverage the global model baseline while applying cluster-specific regulation through:

$$\mathcal{L}_{cd}^k = KL\left(\sigma\left(\mathbb{C}\left(\omega_{C_k}; \hat{x}\right)\right) \| \sigma\left(\mathbb{C}\left(\omega_G; \hat{x}\right)\right)\right) \quad (9)$$

Here $\mathbb{C}(\cdot)$ denotes the classifier output layer, $\sigma(\cdot)$ the softmax function, and KL the Kullback-Leibler divergence measuring prediction distribution discrepancies.

To maintain semantic coherence in cross-cluster distillation while mitigating computational overhead, we optimize the centralized generator through cluster-conditioned cross-entropy minimization. The objective function for cluster fidelity, denoted as $\mathcal{L}_{cf}$, is given by:

$$\min_{\theta_{\mathcal{G}}} \mathcal{L}_{cf} := \mathbb{E}_{y \sim \hat{p}(y)} \mathbb{E}_{z \sim \mathcal{N}(\mathbf{0}, \mathbf{1})} \left[\sum_{k=1}^{K} \alpha_y^k \mathcal{L}_{cf}^k \right] \quad (10)$$

The per-cluster fidelity loss is computed as:

$$\mathcal{L}_{cf}^k = \sum_{k=1}^{K} \alpha_y^k CE\left(\sigma\left(\mathbb{C}\left(\widetilde{x}; \omega_k\right)\right), y\right) \quad (11)$$

In CFL environments, cluster-specific knowledge exhibits multi-category characteristics. We adopt the diversity-aware regularization term $\mathcal{L}_{div}$ [Zhu *et al.*, 2021] to enhance knowledge transfer through maximizing inter-sample dissimilarity.

$$\mathcal{L}_{div} = e^{\frac{1}{Q*Q} \sum_{i,j \in \{1,\ldots,Q\}} \left(-\|\hat{x}_i - \hat{x}_j\|_2 * \|z_i - z_j\|_2\right)} \quad (12)$$

where Q is the number of pseudo-samples, and z_i denotes the noise of the i-th pseudo-sample.

4.3 Secure Similarity Computation

Principles. Similarity matrix exposure introduces re-identification risks, as adversaries could infer sensitive client attributes through parametric similarity analysis. We address this challenge through a lightweight *Privacy-Preserving Cosine Similarity Computing (PCSC) protocol* protocol for CFL [Lu *et al.*, 2014]. The protocol prevents server access to raw similarity values through encrypted operations.

4.4 Optimization and Algorithm of DisUE

Optimization Objective. Under CFL assumptions, intra-cluster data homogeneity allows cluster-optimal solutions via FedAvg, whereas inter-group label shifts degrade expert model performance. We seek a globally optimal solution ω_G that harmonizes k cluster-specific distributions.

The DisUE framework establishes an adversarial game between the global model ω_G and a generator $\mathcal{G}$ (parameterized by $\theta_{\mathcal{G}}$), governed by a composite loss ($\mathcal{L}_{cd}$, $\mathcal{L}_{cf}$ and $\mathcal{L}_{div}$) that coordinates cross-cluster knowledge distillation through alternating optimization phases: (1) *Maximization Phase* where $\mathcal{G}$ synthesizes hard-classifiable samples from cluster to amplify prediction conflicts, and (2) *Minimization Phase* where ω_G aligns predictions with cluster models using boundary samples. This process progressively sharpens global decision boundaries while absorbing cluster-specific knowledge.

Algorithm 1 Workflow of DisUE

Require: Initial global model $\omega_G^{(0)}$, client set $\mathcal{N}_s$, rounds T, security parameter=sec_params

Ensure: Final global model $\omega_G^{(T)}$, generator parameters $\theta_{\mathcal{G}}^{(T)}$

Initialization:

Broadcast $\omega_G^{(0)}$ to all clients in $\mathcal{N}_s$
for round $t = 1$ **to** T **do**
L-Phase: Local Training
  Client Update
  for all client $i \in \mathcal{N}_s$ **in parallel do**
    Local training: $\omega_i^{(t)} \leftarrow (\omega_G^{(t-1)})$
    $\mathsf{E}(\omega_i^{(t)}) \leftarrow \mathsf{SSC.Encrypt}(\mathsf{sec\_params}, \omega_i^{(t)})$
    Upload $\mathsf{E}(\omega_i^{(t)})$ to server
  end for
  Initialize similarity matrix $\mathbf{Sim}^{(t)} \leftarrow \mathbf{0}_{|\mathcal{N}_s|\times|\mathcal{N}_s|}$
  for all pairs $(u, v) \in \mathcal{N}_s \times \mathcal{N}_s$ where $u < v$ **do**
    Compute encrypted similarity:
    $Sim_{u,v}^{(t)} \leftarrow \mathsf{SSC.Compute}\big(\mathsf{E}(\omega_u^{(t)}), \mathsf{E}(\omega_v^{(t)})\big)$
    Update $\mathbf{Sim}^{(t)}[u, v] \leftarrow Sim_{u,v}^{(t)}$
  end for
C-Phase: Clustering
  $\{\omega_1^{(t)}, \ldots, \omega_K^{(t)}\} \leftarrow \mathsf{CGA\text{-}Clustering}(\mathbf{Sim}^{(t)})$
  Intra-group aggregation:
  $\{\omega_{C_1^{(t)}}, \ldots, \omega_{C_K^{(t)}}\} \leftarrow \mathsf{CGA\text{-}Intra}(\{C_1^{(t)}, \ldots, C_K^{(t)}\})$
G-Phase: Group Model Aggregation
  Global aggregation:
  $\omega_G^{(t)} \leftarrow \mathsf{FedAvg}(\omega_{C_1^{(t)}}, \ldots, \omega_{C_K^{(t)}})$
  Inter-group aggregation:
  $\omega_G^{(t+1)} \leftarrow \mathsf{IGA}\big(\{\omega_{C_1^{(t)}}, \ldots, \omega_{C_K^{(t)}}\}, \omega_G^{(t)}, \theta_{\mathcal{G}}\big)$
  $\mathsf{ServerUpdate}(\omega_G^{(t+1)}, \theta_{\mathcal{G}})$
end for
return $\omega_G^{(T)}, \theta_{\mathcal{G}}^{(T)}$

The unified optimization objective combines three loss components from Eqs. (8), (10), and (12):

$$\min_{\omega_G} \max_{\theta_{\mathcal{G}}} \; \mathbb{E}_{y\sim\hat{p}(y),\, z\sim\mathcal{N}(\mathbf{0},\mathbf{1})}\big(\mathcal{L}_{cd} + \beta_{cf}\,\mathcal{L}_{cf} + \beta_{div}\,\mathcal{L}_{div}\big) \tag{13}$$

where β_{cf} and β_{div} balance the loss terms. This minimax optimization maximizes cross-cluster knowledge transfer, ultimately producing a universal model with consistent performance across clusters.

Algorithm. Algorithm 1 illustrates the workflow of DisUE. We begin by broadcasting an initial global model $\omega_G^{(0)}$ to all clients. In each communication round t, clients first perform *L-phase* on their private data. Subsequently, each client encrypts its updated parameters using the **SSC** protocol and transmits these encrypted updates to the server. The server obtains a similarity matrix $\mathbf{Sim}^{(t)}$, which it leverages to execute **CGA-Clustering**, partitioning clients into K groups. Within group, parameters are aggregated via **CGA-Intra**, conducing clustering structure $\{\omega_{C_1^t}, \ldots, \omega_{C_K^t}\}$.

Next, the global model ω_G^t is obtained via a federated aggregation (**FedAvg**) of these group models. To enable effective cross-cluster distillation, the **IGA** aggregates distilled knowledge from group models by refining ω_G^t into ω_G^{t+1} without direct access to raw client data. Finally, the server updates the global parameters (and the generator parameters $\theta_{\mathcal{G}}$ before proceeding to the next communication round. After T rounds, DisUE returns the $\omega_G^{(T)}$ and $\theta_{\mathcal{G}}^{(T)}$.

5 Experiments

5.1 Implementation Details

Datasets. We evaluate DisUE on three standard benchmarks: SVHN [Netzer *et al.*, 2011], CIFAR-10 [Krizhevsky *et al.*, 2009], and CIFAR-100 [Krizhevsky *et al.*, 2009]. To simulate federated data heterogeneity, we partition datasets using the Dirichlet distribution $Dir(\epsilon)$ [Hsu *et al.*, 2019], with concentration parameters $\epsilon \in [0.01, 0.06]$ controlling non-IID difficulty levels. Lower ϵ values create more skewed client data distributions, representing challenging FL scenarios.

Baselines. We benchmark DisUE against 11 state-of-the-art FL methods spanning three key categories: (1) CFL methods including IFCA [Ghosh *et al.*, 2020], CFL [Sattler *et al.*, 2020], CFL-GP [Kim *et al.*, 2024], and PACFL [Vahidian *et al.*, 2023]; (2) DFKD approaches FedDF [Lin *et al.*, 2020b] and FedFTG [Zhang *et al.*, 2022]; and (3) conventional FL methods covering FedAvg [McMahan *et al.*, 2017], FedProx [Li *et al.*, 2020], FedDyn [Acar *et al.*, 2021], MOON [Li *et al.*, 2021], and SCAFFOLD [Karimireddy *et al.*, 2020].

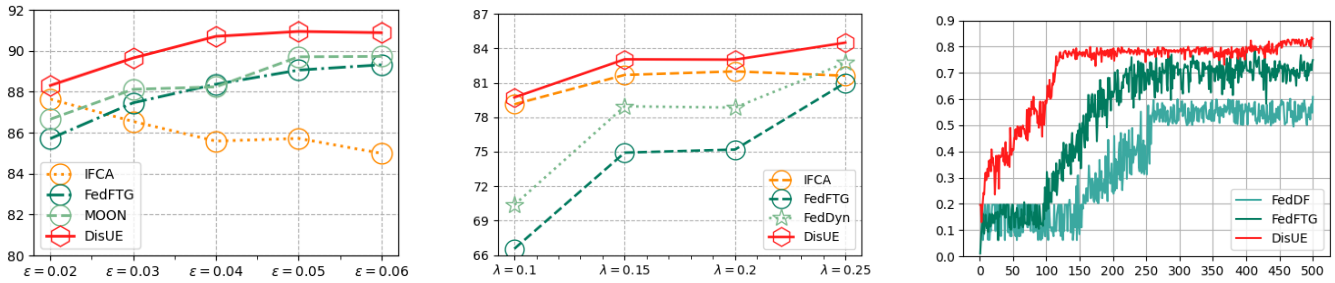
Configurations. For all methods, we set the communication rounds $T = 500$, the number of clients $N = 100$, with an active fraction $Act = 0.15$. For local training, we set the number of local epochs localE = 5, batch size = 50, and the weight decay to 1×10^{-3}. The learning rates for the classifier and generator are initialized to 0.1 and 0.01, respectively. The dimension z is set to 100 for CIFAR-10 and SVHN, and 256 for CIFAR-100. Unless otherwise specified, we adopt $\beta_{\text{cf}} = 1.0$ and $\beta_{\text{div}} = 1.0$. All our experimental results represent the average over five random seeds. For the classifier, we adopt the network architecture from [He *et al.*, 2016]. The generator architecture from [Zhang *et al.*, 2022] is employed for both FedFTG and FedDF.

5.2 Performance Comparison

Test Accuracy. Table 1 reports the test accuracy achieved by all methods on the SVHN, CIFAR10, and CIFAR100 datasets. ***(1)*** Among all scenarios, our method consistently outperforms the three baseline algorithm types on both IID and non-IID settings. ***(2)*** Compared to CFL algorithms, DisUE achieves significantly better performance, primarily due to cross-group knowledge transfer, which enables its global model to leverage more comprehensive information. ***(3)*** Compared to the DFKD algorithms, performance is further improved. This is attributed to the group-based distillation protocol having a pre-trained component, clustering, thereby mitigating the non-IID problem in advance. ***(4)*** Experimental observations reveal distinct performance patterns across algorithms under varying data distributions. Global FL algorithm achieves superior accuracy in IID and $Dir(\epsilon = 0.1)$. This phenomenon arises from the relatively balanced category distributions across clients, which induces aligned

Type	Method	SVHN			CIFAR10			CIFAR100		
		IID	0.1	0.01	IID	0.1	0.01	IID	0.1	0.01
CFL	IFCA	91.66	88.63	<u>81.68</u>	78.06	66.25	53.86	41.09	42.34	30.86
	CFL	80.05	69.03	68.16	69.29	55.37	47.60	39.65	38.28	26.33
	CFL-GP	79.95	71.53	69.01	71.09	55.30	48.56	41.22	34.13	26.75
	PACFL	93.09	87.46	80.65	78.30	67.26	<u>55.89</u>	34.37	<u>43.14</u>	<u>30.96</u>
DFKD	FedDF	94.91	89.80	60.85	81.02	71.81	45.79	44.53	39.79	7.08
	FedFTG	94.73	90.89	74.92	81.67	73.19	51.61	42.71	36.61	30.85
FL	FedAvg	94.80	90.86	73.83	81.28	74.49	51.90	44.43	42.08	30.92
	FedProx	94.93	90.52	73.61	81.77	74.96	51.86	44.27	41.42	30.06
	FedDyn	<u>95.10</u>	90.59	78.95	81.95	75.01	51.50	45.44	40.64	16.13
	MOON	95.06	<u>91.62</u>	76.94	81.23	73.96	51.45	42.65	41.23	30.07
	SCAFFOLD	95.08	91.18	69.46	<u>82.02</u>	<u>75.39</u>	48.85	<u>49.55</u>	25.32	17.84
DisUE		**95.99**	**93.25**	**83.04**	**83.25**	**76.22**	**58.83**	**51.80**	**44.41**	**31.87**

Table 1: Comparative Performance Analysis of Federated Learning Methods. Best results in **bold**, second best <u>underlined</u>.



(a) Test accuracy on dir($\epsilon = 0.01$) (b) Test accuracy on act($\cdot$), dir$\epsilon = 0.01$ (c) Test accuracy on SVHN dir($\epsilon = 0.01$)

Figure 2: (a) Test accuracy versus data heterogeneity measured by $\epsilon(\cdot)$. (b) Test accuracy versus the fraction C of active clients per round ($\epsilon = 0.01$). (c) Test accuracy over 500 communication rounds (learning curve) with $\epsilon = 0.01$. All experiments are conducted on SVHN.

parameter update directions that undermine clustering efficacy. Conversely, under $Dir(\epsilon = 0.01)$ non-IID conditions, CFL demonstrates better performance over global FL.

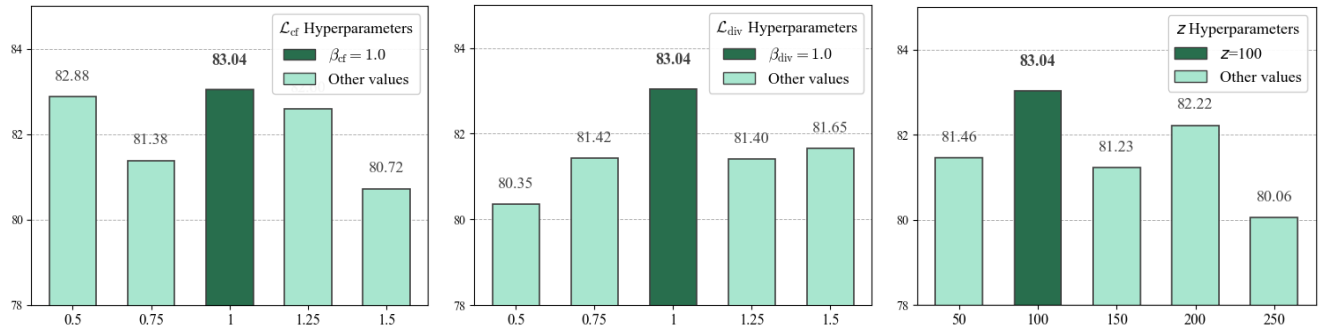
Communication Overhead. We assessed the communication rounds needed for convergence by DFKD SOTA methods on SVHN with on-IID degrees set to *0.01* in Figure. 2c. ***(1)*** Our method achieves accelerated convergence rates through cluster-level adversarial distillation. ***(2)*** During the first 100 iterations, our methods demonstrate greater performance gains compared to other DFKD approaches, evidencing that integration clustering effectively mitigates data heterogeneity issues and prevents early-stage local noise.

Data Heterogeneity. We evaluate the robustness of algorithms under varying degrees of non-IID data by measuring the test accuracy of all methods with non-IID degrees set to (0.02, 0.03, 0.04, 0.05, 0.06). Figure 2a illustrates the test accuracy as a function of varying $Dir(\cdot)$, respectively showing the best-performing algorithm for each type of method. ***(1)***DisUE outperforms baseline methods across diverse configurations, demonstrating the effectiveness of CFL's clustering mechanism in (i) preserving heterogeneous data characteristics through clustering and (ii) maximizing knowledge utility via inter-group distillation. ***(2)*** Global FL algorithms exhibit performance improvements with increasing Dirichlet coefficient ϵ. ***(3)*** CFL performance degrades progressively with higher values due to accumulating clustering errors.

Client Participation Scale. We evaluated all approaches across varying client scales to observe how the impact of each algorithm responds to an increasing number of participants. Thus, we set *act*= (0.1, 0.15, 0.2, and 0.25) per training round. ***(1)*** As shown in Figure 2b, DisUE demonstrates superior performance. ***(2)*** Increasing $act(\cdot)$ improves the accuracy of all methods, as more local information becomes available per round. ***(3)*** Compared with FL and DFKD, CFL and DisUE demonstrate a more stable increase in accuracy. This indicates that, under a fixed number of clients, the CFL algorithm should avoid selecting a high client participation rate when adapting to a Dirichlet distribution.

5.3 Compatibility Study

To validate the compatibility of the proposed aggregation mechanism within CFL frameworks, we integrated it into sev-



(a) $\mathcal{L}_{cf}$ (b) $\mathcal{L}_{div}$ (c) noise dimension

Figure 3: Performance of DisUE using different hyperparameters on SVHN, dir($\epsilon = 0.01$)

Methods	SVHN	CIFAR10	CIFAR100
IFCA	81.68	53.86	30.86
IFCA+IGA	82.89	55.31	32.01
Improvement	↑ 1.21%	↑ 1.45%	↑ 1.15%
CFL	68.16	47.60	26.33
CFL+IGA	69.80	49.12	28.71
Improvement	↑ 1.64%	↑ 1.52%	↑ 2.38%
CFL-GP	69.01	48.56	26.75
CFL-GP+IGA	71.40	50.62	28.80
Improvement	↑ 2.39%	↑ 2.06%	↑ 2.05%
PACFL	80.65	55.89	30.96
PACFL+IGA	84.09	58.94	34.27
Improvement	↑ 3.44%	↑ 3.05%	↑ 3.31%

Table 2: Compatibility analysis by integrating our proposed IGA into representative CFL methods with $dir(\epsilon = 0.01)$.

eral existing CFL optimizers and evaluated their accuracy. As shown in Table 2, our cross-cluster distillation module consistently enhances the global performance of each CFL method, with the combination involving PACFL achieving the highest test accuracy. Essentially, the clustering algorithm acts as a fundamental component in CFL. Therefore, these results also demonstrate our method's compatibility with various clustering strategies, as it improves their performance while remaining orthogonal to the underlying clustering approach.

5.4 Ablation Study

To assess the necessity of each component, we test accuracy after removing different key modules and loss functions. In all experiments, we use $Dir(\epsilon = 0.01)$ on the SVHN task. For the modules, we remove *GLS*, *GWF*, and IGA, while for the key loss functions, we eliminate $\mathcal{L}_{cf}$ and $\mathcal{L}_{div}$. In Table 3, after removing the IGA, DisUE reverts to FedAvg, resulting in the most performance degradation. This underscores the critical importance of extracting and migrating cross-cluster information. *GLS* and *GWF* modules cause the model to average samples and integrate identical weights, respectively. Consequently, the global model becomes incapable

	Method	Accuracy (%)
Baseline	DisUE	**83.04**
Module	− *GLS*	82.23
	− *GWF*	82.14
	− IGA	72.76
Loss	$-\mathcal{L}_{cf}$	81.55
	$-\mathcal{L}_{div}$	82.69

Table 3: The impact of each module and loss. The experiments are conducted on SVHN, dir($\epsilon = 0.01$).

of addressing differences in data distributions across groups, leading to further performance declines. $\mathcal{L}_{cf}$ and $\mathcal{L}_{div}$ play key roles in pseudo data generation. Dropping $\mathcal{L}_{cf}$ leads to blurred or less realistic outputs, whereas dropping $\mathcal{L}_{div}$ results in less diverse categories.

5.5 Hyperparameters Sensitivity Analysis

We evaluate the sensitivity of key hyperparameters: β_{cf}, β_{div}, and noise dimension. The β_{cf} and β_{div} weights are assessed at (0.5, 0.75, 1, 1.25, and 1.5), while the noise dimensions are set to (50, 100, 150, 200, and 250), following the settings in [Zhang *et al.*, 2022]. ***(1)*** As illustrated in Figure 3a and Figure 3b, DisUE achieves optimal performance when $\beta_{\text{cf}} = 1$ and $\beta_{\text{div}} = 1$. ***(2)*** An unsuitable β_{cf} may lead to capturing the group model semantic imbalances in the pseudo-data. Similarly, the improper ratio of β_{div} could reduce the diversity of pseudo-data. ***(3)*** As shown in Figure 3c, a noise dimension of 100 produces the best performance, indicating that excessively high noise dimensions should be avoided in this setup.

6 Conclusion

Training a consensus model that effectively balances personalized and global knowledge under statistical heterogeneity remains a challenge in FL. We address this by introducing a novel framework that dynamically clusters clients to capture personalization and leverages a cluster-level DFKD strategy to distill and transfer global knowledge across groups.

Acknowledgements

Zeqi Leng, Chunxu Zhang and Bo Yang are supported by the National Natural Science Foundation of China under-Grant Nos.U22A2098,62172185,62206105, and 62202200; the Major Science and Technology Development Plan of Jilin Province under Grant No. 20240302078GX; the Major Science and Technology Development Plan of Changchun under Grant No.2024WX05. Riting Xia is supported by the Inner Mon-golia University High-level Talent Project under Grant No. 10000-23112101/2861.

References

[Acar *et al.*, 2021] Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. *arXiv preprint arXiv:2111.04263*, 2021.

[Beltrán *et al.*, 2023] Enrique Tomás Martínez Beltrán, Mario Quiles Pérez, Pedro Miguel Sánchez Sánchez, Sergio López Bernal, Gérôme Bovet, Manuel Gil Pérez, Gregorio Martínez Pérez, and Alberto Huertas Celdrán. Decentralized federated learning: Fundamentals, state of the art, frameworks, trends, and challenges. *IEEE Communications Surveys & Tutorials*, 25(4):2983–3013, 2023.

[Duan *et al.*, 2021] Moming Duan, Duo Liu, Xinyuan Ji, Yu Wu, Liang Liang, Xianzhang Chen, Yujuan Tan, and Ao Ren. Flexible clustered federated learning for client-level data distribution shift. *IEEE Transactions on Parallel and Distributed Systems*, 33(11):2661–2674, 2021.

[Ester *et al.*, 1996] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. Density-based spatial clustering of applications with noise. In *Int. Conf. knowledge discovery and data mining*, volume 240, 1996.

[Fan *et al.*, 2024] Jiamin Fan, Kui Wu, Guoming Tang, Yang Zhou, and Shengqiang Huang. Taking advantage of the mistakes: Rethinking clustered federated learning for iot anomaly detection. *IEEE Transactions on Parallel and Distributed Systems*, 35(6):862–876, 2024.

[Frey and Dueck, 2007] Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007.

[Ghosh *et al.*, 2020] Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. *Advances in Neural Information Processing Systems*, 33:19586–19597, 2020.

[Guo *et al.*, 2024] Yongxin Guo, Xiaoying Tang, and Tao Lin. Fedrc: Tackling diverse distribution shifts challenge in federated learning by robust clustering, 2024.

[Hao *et al.*, 2021] Zhiwei Hao, Yong Luo, Han Hu, Jianping An, and Yonggang Wen. Data-free ensemble knowledge distillation for privacy-conscious multimedia model compression. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1803–1811, 2021.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016.

[Hsu *et al.*, 2019] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

[Huang *et al.*, 2024] Honglan Huang, Wei Shi, Yanghe Feng, Chaoyue Niu, Guangquan Cheng, Jincai Huang, and Zhong Liu. Active client selection for clustered federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 35(11):16424–16438, 2024.

[Jahani-Nezhad *et al.*, 2022] Tayyebeh Jahani-Nezhad, Mohammad Ali Maddah-Ali, Songze Li, and Giuseppe Caire. Swiftagg: Communication-efficient and dropout-resistant secure aggregation for federated learning with worst-case security guarantees. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 103–108. IEEE, 2022.

[Karimireddy *et al.*, 2020] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143. PMLR, 2020.

[Kim *et al.*, 2024] Heasung Kim, Hyeji Kim, and Gustavo De Veciana. Clustered federated learning via gradient-based partitioning. In *Forty-first International Conference on Machine Learning*, 2024.

[Krizhevsky *et al.*, 2009] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[Li *et al.*, 2020] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

[Li *et al.*, 2021] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10713–10722, 2021.

[Lin *et al.*, 2020a] Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2351–2363. Curran Associates, Inc., 2020.

[Lin *et al.*, 2020b] Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. *Advances in neural information processing systems*, 33:2351–2363, 2020.

[Liu *et al.*, 2024] Boyi Liu, Yiming Ma, Zimu Zhou, Yexuan Shi, Shuyuan Li, and Yongxin Tong. Casa: Clustered federated learning with asynchronous clients. In *Proceedings*

of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 1851–1862, 2024.

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

[Lu *et al.*, 2014] Rongxing Lu, Hui Zhu, Ximeng Liu, Joseph K Liu, and Jun Shao. Toward efficient and privacy-preserving computing in big data era. *IEEE Network*, 28(4):46–50, 2014.

[Ma *et al.*, 2022] Jie Ma, Guodong Long, Tianyi Zhou, Jing Jiang, and Chengqi Zhang. On the convergence of clustered federated learning. *arXiv preprint arXiv:2202.06187*, 2022.

[Ma *et al.*, 2023] Jie Ma, Tianyi Zhou, Guodong Long, Jing Jiang, and Chengqi Zhang. Structured federated learning through clustered additive modeling. *Advances in Neural Information Processing Systems*, 36:43097–43107, 2023.

[McInnes *et al.*, 2017] Leland McInnes, John Healy, Steve Astels, et al. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205, 2017.

[McMahan *et al.*, 2017] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[Netzer *et al.*, 2011] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 4. Granada, 2011.

[Nguyen *et al.*, 2021] Dinh C Nguyen, Ming Ding, Pubudu N Pathirana, Aruna Seneviratne, Jun Li, and H Vincent Poor. Federated learning for internet of things: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 23(3):1622–1658, 2021.

[Ruan and Joe-Wong, 2022] Yichen Ruan and Carlee Joe-Wong. Fedsoft: Soft clustered federated learning with proximal local updating, 2022.

[Sattler *et al.*, 2020] Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE transactions on neural networks and learning systems*, 32(8):3710–3722, 2020.

[Shin and Choi, 2024] Hyunjune Shin and Dong-Wan Choi. Teacher as a lenient expert: Teacher-agnostic data-free knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14991–14999, 2024.

[So *et al.*, 2021] Jinhyun So, Başak Güler, and A Salman Avestimehr. Turbo-aggregate: Breaking the quadratic aggregation barrier in secure federated learning. *IEEE Journal on Selected Areas in Information Theory*, 2(1):479–489, 2021.

[Tu *et al.*, 2025] Kaifei Tu, Xuehe Wang, and Xiping Hu. Entrocfl: Entropy-based clustered federated learning with incentive mechanism. *IEEE Internet of Things Journal*, 12(1):986–1001, 2025.

[Vahidian *et al.*, 2023] Saeed Vahidian, Mahdi Morafah, Weijia Wang, Vyacheslav Kungurtsev, Chen Chen, Mubarak Shah, and Bill Lin. Efficient distribution similarity identification in clustered federated learning via principal angles between client data subspaces. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8):10043–10052, Jun. 2023.

[Wang and Yoon, 2021] Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):3048–3068, 2021.

[Wang *et al.*, 2024] Jiaqi Wang, Chenxu Zhao, Lingjuan Lyu, Quanzeng You, Mengdi Huai, and Fenglong Ma. Bridging model heterogeneity in federated learning via uncertainty-based asymmetrical reciprocity learning. *CoRR*, abs/2407.03247, 2024.

[Ye *et al.*, 2020] Jingwen Ye, Yixin Ji, Xinchao Wang, Xin Gao, and Mingli Song. Data-free knowledge amalgamation via group-stack dual-gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12516–12525, 2020.

[Yoo *et al.*, 2019] Jaemin Yoo, Minyong Cho, Taebum Kim, and U Kang. Knowledge extraction with no observable data. *Advances in Neural Information Processing Systems*, 32, 2019.

[Zhang *et al.*, 2022] Lin Zhang, Li Shen, Liang Ding, Dacheng Tao, and Ling-Yu Duan. Fine-tuning global model via data-free knowledge distillation for non-iid federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10174–10183, 2022.

[Zhang *et al.*, 2023] Chunxu Zhang, Guodong Long, Tianyi Zhou, Peng Yan, Zijian Zhang, Chengqi Zhang, and Bo Yang. Dual personalization on federated recommendation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23, 2023.

[Zhu *et al.*, 2021] Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. Data-free knowledge distillation for heterogeneous federated learning. In *International conference on machine learning*, pages 12878–12889. PMLR, 2021.