Prevention over Correction: Learning Aligned Representations in One-shot Federated Learning

YongHoon Kang and Jee-Hyong Lee *

Sungkyunkwan University, Suwon, South Korea
{artfjqm2, john}@skku.edu

Abstract. One-shot Federated Learning (OFL) reduces communication costs by requiring only a single round of model exchange. However, when clients train independently on heterogeneous data, their learned representations develop different means and variances, creating distributional mismatch that prevents effective server-side aggregation. Existing methods attempt post-calibration at the server, but struggle to reconcile representations that have already diverged into incompatible statistical spaces. We propose **FACE** (**F**ederated **A**ligned-**C**onsistent **E**nsemble) to prevent distributional divergence during client training rather than correcting it afterward. FACE uses a shared fixed reference matrix to guide representation means toward consistent directions, applies a variance stabilizer to maintain consistent feature variance, and performs feature-fusion ensemble inference. Extensive experiments on CIFAR-10/100 and Tiny-ImageNet demonstrate that FACE achieves state-of-the-art performance, with particularly strong improvements under extreme data heterogeneity. Comprehensive ablation studies validate the necessity of both components for achieving distributional alignment.

Keywords: Federated Learning, One-shot Federated Learning, Extreme Heterogeneity, Representation Alignment

1 Introduction

Federated Learning (FL) has emerged as a practical solution to privacy concerns and data protection regulations in modern machine learning. In FL, each client keeps its data local, trains a local model, and sends only model parameters to a central server. The server aggregates these local models and broadcasts the updated global model back to clients. By iteratively repeating this process, FL can build high-performance models without directly accessing sensitive data [6, 11, 14, 17, 18, 21, 28, 35]. However, such multi-round interactions incur substantial communication overhead and latency, which can be impractical in resource-constrained or time-sensitive settings [2, 20, 29].

To reduce communication costs, One-shot Federated Learning (OFL) compresses multi-round FL into a single round of model exchange [4, 8, 16, 25, 30]. In OFL, each client trains extensively on its private data and uploads the trained

*Corresponding author.

model to the server only once. While this greatly improves communication efficiency, it introduces a fundamental challenge. OFL provides no intermediate synchronization to guide clients toward a common solution. As a result, clients optimize independently on heterogeneous data, and their learned representations diverge.

This divergence manifests as a mismatch in the feature distributions learned by different clients. Because each client trains on a different local distribution, their representations can develop different statistical characteristics. In particular, feature means may drift toward client-specific directions, and features may collapse into client-specific low-dimensional subspaces, reducing feature variance in the representation space. Existing OFL methods attempt to address this mismatch through server-side post-calibration, including ensemble distillation, selective aggregation, and weighted averaging. However, as data heterogeneity increases, the gap between client representations often becomes too large for these post-calibration methods to reconcile. Post-calibration cannot fully restore alignment that was never established during training.

We therefore take a different paradigm, **Prevention over Correction**. Instead of correcting divergence after training, we aim to prevent it during client training. Specifically, we impose constraints on the local optimization process so that clients learn representations with more consistent feature distributions. With this prevention strategy, independently trained client models arrive at the server with more compatible representations, enabling effective feature-level fusion.

To address this challenge, we propose **FACE** (**F**ederated **A**ligned-**C**onsistent **E**nsemble), a framework that enforces distributional consistency during client training through two complementary mechanisms. First, we introduce Directional Mean Alignment (DMA), which aligns class-conditional feature means from different clients with shared reference directions, thereby reducing mean drift under heterogeneous training. To complement this, we apply a Variance Stabilizer (VS) based on a uniformity regularization objective that discourages representation collapse and preserves sufficient feature variance. DMA and VS jointly regulate feature means and variances during local training. As a result, independently trained client models become more compatible in their representation, enabling effective feature-level fusion at the server without additional post-calibration.

Extensive experiments on CIFAR-10/100 and Tiny-ImageNet show that FACE achieves state-of-the-art performance, particularly under extreme data heterogeneity. Through comprehensive analysis, we further demonstrate the effectiveness of each component.

2 Related Work

2.1 Federated Learning

Federated Learning (FL) enables multiple clients to collaboratively train a model without sharing raw data. In the standard multi-round setting, clients alternate

between local optimization and server-side aggregation, repeatedly synchronizing model states to implicitly reduce client drift. Aggregation follows two main strategies. Parameter-averaging methods merge client weights into a single global model, often with additional corrections to counter client drift [11, 17, 21, 28]. Ensemble-based methods instead keep client models separate and combine their outputs or representations at the server [1, 5, 9, 10, 19].

A broad line of work improves robustness under heterogeneous data by modifying optimization or constraining model components. Prototype-based methods communicate class prototypes to regularize local representation learning [26]. Other approaches decouple representation learning from classifier adaptation, updating the encoder globally while adapting the classifier locally [22]. Inspired by Neural Collapse, fixed classifier designs use a shared simplex ETF structure to mitigate classifier bias and promote more consistent representations across clients [18].

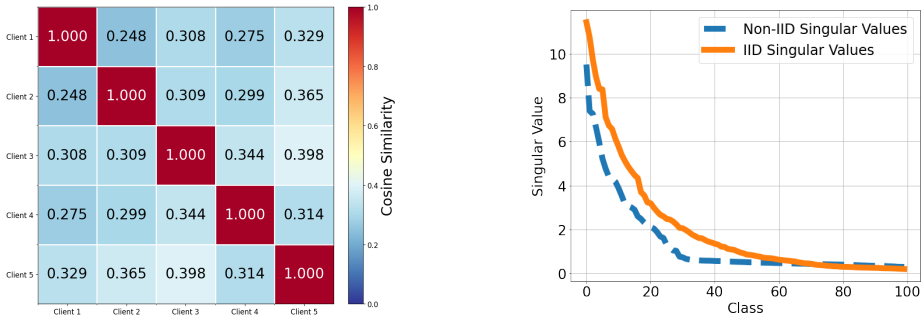
Another related direction studies representation degeneration under heterogeneity. Prior work analyzes dimensional collapse, where representations concentrate in low-dimensional subspaces, and proposes regularization that encourages decorrelated features during local training [23]. More recently, uniformity- and variance-oriented regularization has been explored to stabilize representation dispersion under heterogeneous training [24]. While effective in multi-round FL, these methods typically benefit from repeated synchronization, which is absent in one-shot FL.

2.2 One-shot Federated Learning

One-shot Federated Learning (OFL) compresses multi-round FL into a single round of model exchange, where each client trains extensively on local data and uploads a model only once [4, 8, 16, 25, 30]. The lack of intermediate synchronization makes client representations more likely to diverge, leading to mismatched feature distributions.

Most OFL methods address this mismatch after local training at the server. Data-free and synthetic-data-based distillation methods aggregate client knowledge through distillation pipelines, but the resulting supervision can become unreliable when client representations are misaligned [3, 33, 34]. Other approaches improve one-shot aggregation using specialized model-combination rules such as layer-wise aggregation [20], or enhance local representation quality with additional training objectives and fusion mechanisms [27].

FAFI is most closely related to our work because it introduces both client-side mechanisms and server-side fusion. FAFI learns client-specific prototypes and performs informativeness-aware fusion at the server [31]. In contrast, our approach imposes a shared reference direction and variance regularization during local training using fixed reference directions and uniformity-based regularization. This design directly constrains feature means and variances to become more consistent across clients, enabling effective feature-level fusion without additional server-side post-calibration.



(a) Mean divergence **(b)** Variance divergence

Fig. 1: Feature divergence in one-shot federated learning under data heterogeneity. **(a)** Cosine similarity between client feature means shows consistently low values, indicating that clients learn representations in incompatible spaces. This reflects mean divergence caused by heterogeneous local data. **(b)** Singular values of classifier weights decay faster under non-IID training than IID training, showing that decision boundaries collapse toward subspaces biased to local data. These observations motivate our approach of preventing divergence during client training rather than correcting it at the server.

3 Problem Analysis

3.1 Preliminaries

We consider a one-shot federated learning (OFL) setting with N clients. Each client $i \in \{1, \dots, N\}$ has a private dataset $\mathcal{S}_i = \{(x_j, y_j)\}$ sampled from its local distribution $\mathcal{D}_i$. Starting from a common initialization θ_0, each client independently trains

$$\theta_i^\star = \arg\min_\theta \mathcal{L}_i(\theta) = \arg\min_\theta \mathbb{E}_{(x,y)\sim\mathcal{D}_i}\big[\ell(f(x;\theta), y)\big], \tag{1}$$

where $f(x;\theta) = h(g(x;\theta))$ is decomposed into a feature extractor $g : \mathcal{X} \to \mathbb{R}^d$ and a classifier $h : \mathbb{R}^d \to \mathbb{R}^C$. Unlike multi-round federated learning, OFL allows only a single model upload from each client to the server.

To analyze cross-client representation compatibility, we introduce a shared reference distribution $\mathcal{D}_{\text{ref}}$ with class prior $\pi_{\text{ref}}(c)$ and class-conditional marginal $\mathcal{D}_{\text{ref}}^c$, used solely for analysis. We write $g_i(x) := g(x;\theta_i^\star)$ and $\tilde{\mathbf{z}}_i(x) := g_i(x)/\|g_i(x)\|_2$, and define

$$\tilde{\mu}_i = \mathbb{E}_{x\sim\mathcal{D}_{\text{ref}}}[\tilde{\mathbf{z}}_i(x)], \qquad \tilde{\Sigma}_i = \mathrm{Cov}_{x\sim\mathcal{D}_{\text{ref}}}(\tilde{\mathbf{z}}_i(x)), \qquad \tilde{\mu}_i^c = \mathbb{E}_{x\sim\mathcal{D}_{\text{ref}}^c}[\tilde{\mathbf{z}}_i(x)]. \tag{2}$$

3.2 Empirical Observation: Feature Divergence in One-shot FL

Independent local training under heterogeneous data leads to incompatible representations across clients. Fig. 1a reports the pairwise cosine similarity $\mathrm{sim}(i,j) = \tilde{\mu}_i^\top \tilde{\mu}_j / (\|\tilde{\mu}_i\|_2 \|\tilde{\mu}_j\|_2)$ between client-wise normalized feature means. The off-diagonal

values are consistently low, ranging from 0.25 to 0.40, indicating that clients develop mutually inconsistent feature directions despite sharing the same initialization. This directional inconsistency directly undermines feature-level aggregation: when client representations occupy geometrically inconsistent subspaces, their average retains less class-discriminative information than any individual client representation.

Fig. 1b provides a complementary perspective from the classifier side. Under non-IID training, the singular values of the classifier weight matrix decay markedly faster than in the IID case, reflecting a reduction in the effective rank of the learned decision boundary. This is consistent with each client's classifier collapsing onto a low-dimensional subspace biased toward its local class distribution. We treat this as corroborating evidence for representation collapse rather than a direct estimate of feature covariance.

Taken together, these observations indicate that feature divergence in OFL manifests along two axes: directional inconsistency of feature means across clients, and rank collapse of individual client representations. This motivates addressing both axes during local training, rather than relying on post-hoc correction at the server.

3.3 Problem Setting

In OFL, each client optimizes solely on its local objective, so heterogeneous data induces divergent optimization trajectories from the shared initialization θ_0. We characterize how this parameter divergence propagates to feature statistics under the following standard assumptions on the feature extractor:

$$\|g(x;\theta)\|_2 \le B_g, \qquad \|g(x;\theta) - g(x;\theta')\|_2 \le L_g\|\theta - \theta'\|_2. \tag{3}$$

Under these conditions, parameter divergence transfers directly to feature-moment divergence:

$$\|\mu_i - \mu_j\|_2 \le L_g\|\theta_i^\star - \theta_j^\star\|_2, \qquad \|\Sigma_i - \Sigma_j\|_F \le 4B_gL_g\|\theta_i^\star - \theta_j^\star\|_2. \tag{4}$$

Because server-side inference operates on normalized features, the quantity of primary concern is the pairwise disagreement among $\{\tilde{\mu}_i\}_{i=1}^N$. When clients fuse normalized features as $\mathbf{z}_{\text{avg}}(x) = \sum_{i=1}^N w_i\,\tilde{\mathbf{z}}_i(x)$, the squared norm of the fused representation satisfies

$$\|\mathbf{z}_{\text{avg}}(x)\|_2^2 = \sum_{i=1}^N w_i^2 + 2\sum_{i<j} w_iw_j\,\tilde{\mathbf{z}}_i(x)^\top\tilde{\mathbf{z}}_j(x), \tag{5}$$

showing that low cross-client cosine similarity directly reduces the coherence of the fused representation. Existing OFL methods address this mismatch after local training through post-hoc calibration or distillation; however, once representations have drifted to incompatible directions or collapsed to client-specific subspaces, prediction-level correction cannot recover lost discriminative structure. This motivates a client-side prevention strategy that acts during local training rather than after.

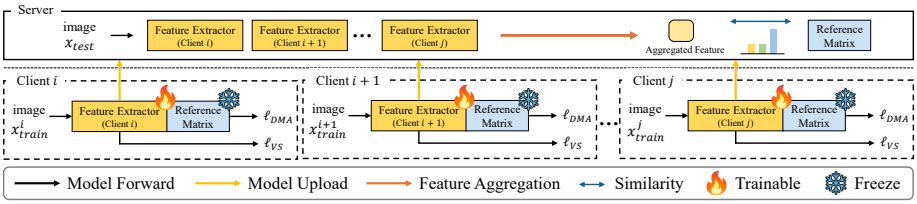


Fig. 2: Overview of FACE. Each client trains with Directional Mean Alignment (DMA) and a Variance Stabilizer (VS), while the server performs feature-level fusion during inference.

4 Proposed Method

4.1 Overview

FACE reduces representation mismatch before server-side fusion. It combines Directional Mean Alignment (DMA), which aligns class-wise feature directions across clients, and Variance Stabilizer (VS), which discourages excessive concentration of normalized features. Each client minimizes

$$\mathcal{L}_{\text{FACE}} = \ell_{\text{DMA}} + \lambda\, \ell_{\text{VS}}, \tag{6}$$

where λ is a hyperparameter. At test time, the server performs feature-level fusion using ℓ_2-normalized features.

4.2 Directional Mean Alignment

A key factor of mismatch in OFL arises when independently trained clients assign geometrically inconsistent directions to the same class. Without a shared reference, each client's feature extractor is free to rotate its class means arbitrarily, so even a well-performing local model may produce features that are irreconcilable with those of its peers upon aggregation.

To impose a common directional reference across all clients, we introduce a fixed, client-shared reference matrix

$$\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_C] \in \mathbb{R}^{d \times C}, \qquad \mathbf{P}^\top \mathbf{P} = \mathbf{I}_C, \tag{7}$$

where the columns $\{\mathbf{p}_c\}$ form an orthonormal basis on the unit hypersphere. This matrix is constructed once before training and never updated, so it introduces no additional learnable parameters and imposes no communication overhead beyond the one-time broadcast of $\mathbf{P}$.

Given input $\mathbf{x}$, client i computes the ℓ_2-normalized feature $\tilde{\mathbf{z}}_i(\mathbf{x}) = g_i(\mathbf{x})/\|g_i(\mathbf{x})\|_2$ and is trained with the cosine-softmax objective

$$\ell_{\text{DMA}} = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_i}\left[-\log \frac{\exp\big(\tilde{\mathbf{z}}_i(\mathbf{x})^\top \mathbf{p}_y\big)}{\sum_{c=1}^{C} \exp\big(\tilde{\mathbf{z}}_i(\mathbf{x})^\top \mathbf{p}_c\big)}\right]. \tag{8}$$

Because both the features and the reference matrix lie on the unit hypersphere, maximizing the cosine similarity $\tilde{\mathbf{z}}_i(\mathbf{x})^\top \mathbf{p}_y$ is equivalent to minimizing the squared Euclidean distance $\|\tilde{\mathbf{z}}_i(\mathbf{x}) - \mathbf{p}_y\|_2^2 = 2(1 - \tilde{\mathbf{z}}_i(\mathbf{x})^\top \mathbf{p}_y)$, so the loss directly pulls each client's class mean toward the shared reference direction.

The effect on mean alignment is made precise by the following proposition. Define the class-wise alignment score of client i on class c as $a_i^c := \mathbb{E}_{x \sim \mathcal{D}_{\text{ref}}^c}[\mathbf{p}_c^\top \tilde{\mathbf{z}}_i(x)]$, which measures how closely the client's normalized class mean tracks the reference matrix.

Proposition 1 (Cross-client class-mean alignment). *For any class $c \in [C]$ and any pair of clients (i, j),*

$$\|\tilde{\mu}_i^c - \tilde{\mu}_j^c\|_2 \;\le\; \sqrt{2(1 - a_i^c)} + \sqrt{2(1 - a_j^c)}. \tag{9}$$

Proposition 1 shows that DMA reduces cross-client directional mismatch in a client-adaptive manner: the bound tightens as each client's alignment score a_i^c approaches one, without requiring a shared margin assumption across clients. Crucially, the bound depends only on quantities each client can optimize independently, which confirms that local training under ℓ_{DMA} is sufficient to control global feature alignment. The proof is provided in the Appendix.

4.3 Variance Stabilizer

DMA aligns feature directions, but it does not directly control how dispersed the normalized features remain on the hypersphere. Without additional regularization, clients may collapse their representations toward the reference directions, reducing intra-class spread and ultimately harming the diversity of information available for aggregation. To preserve sufficient dispersion, we adopt a uniformity-style regularizer:

$$\ell_{\text{VS}} = \mathbb{E}_{\mathbf{x},\mathbf{x}' \sim \mathcal{D}_i}\left[\exp\left(-\frac{\|\tilde{\mathbf{z}}_i(\mathbf{x}) - \tilde{\mathbf{z}}_i(\mathbf{x}')\|_2^2}{2\sigma^2}\right)\right], \tag{10}$$

where σ is a median pairwise distance computed within each minibatch. The loss penalizes pairs of normalized features that are too close on the hypersphere, thereby pushing the feature distribution toward greater spread.

To see why this controls feature variance, note that for normalized features the expected pairwise squared distance satisfies $\mathbb{E}[\|\tilde{\mathbf{z}} - \tilde{\mathbf{z}}'\|_2^2] = 2 - 2\|\tilde{\mu}_i\|_2^2 = 2\,\mathrm{tr}(\tilde{\Sigma}_i)$, so minimizing ℓ_{VS} is equivalent to maximizing typical pairwise distances, which in turn increases the total variance $\mathrm{tr}(\tilde{\Sigma}_i)$ of the normalized representation. To measure how consistently this dispersion is maintained across clients, we use

$$\Delta_{\text{tr}} = \frac{2}{N(N-1)} \sum_{i<j} \left|\mathrm{tr}(\tilde{\Sigma}_i) - \mathrm{tr}(\tilde{\Sigma}_j)\right|. \tag{11}$$

Proposition 2 (Trace identity and trace-dispersion bound). *For any client i,*

$$\mathrm{tr}(\tilde{\Sigma}_i) = 1 - \|\tilde{\mu}_i\|_2^2. \tag{12}$$

Consequently,

$$\Delta_{\mathrm{tr}} \leq \frac{4}{N(N-1)} \sum_{i<j} \|\tilde{\mu}_i - \tilde{\mu}_j\|_2. \tag{13}$$

Proposition 2 shows that the cross-client dispersion variation Δ_{tr} is governed by the cross-client mean gaps. By the trace identity in Eq. (12), each client's total dispersion is determined by the norm of its mean, so Δ_{tr} is bounded by how much the normalized means differ across clients, as stated in Eq. (13). The better DMA aligns these means, the smaller this bound becomes, which links dispersion consistency directly to directional alignment rather than playing against it. We do not claim that VS alone guarantees a flat covariance spectrum; rather, it serves as a lightweight dispersion regularizer that keeps the per-client variance high, while DMA equalizes it across clients. The proof is provided in the Appendix.

4.4 Joint Control of Feature Means and Variances

DMA and VS are designed as complementary objectives, each targeting a distinct axis of cross-client feature mismatch. DMA controls the directional component by anchoring each client's class means to a shared reference, while VS controls the dispersion component by preventing the normalized features from collapsing toward those directions. They address both mean and variance of the feature distribution, which are the two quantities most directly responsible for aggregation inconsistency.

The class-level alignment guaranteed by DMA propagates to the global level via the decomposition induced by the shared reference prior π_{ref}:

$$\tilde{\mu}_i = \sum_{c=1}^{C} \pi_{\mathrm{ref}}(c) \tilde{\mu}_i^c. \tag{14}$$

Therefore,

$$\|\tilde{\mu}_i - \tilde{\mu}_j\|_2 \leq \sum_{c=1}^{C} \pi_{\mathrm{ref}}(c)\, \|\tilde{\mu}_i^c - \tilde{\mu}_j^c\|_2. \tag{15}$$

Combined with Proposition 1, this shows that reducing class-wise directional mismatch also reduces the global mean mismatch relevant for feature fusion. Substituting Eq. (15) into the bound of Proposition 2 further shows that the same alignment scores control the dispersion mismatch: as each $a_i^c \to 1$, both the global mean gap and Δ_{tr} vanish together. Hence DMA and VS play complementary yet mutually reinforcing roles: DMA improves cross-client directional agreement and thereby equalizes per-client dispersion, while VS preserves the feature dispersion that keeps the fused representation informative.

4.5 Server-side Inference

Because FACE aligns feature statistics during local training, the server requires no post-calibration or re-optimization at inference time. After local training, the server receives N client feature extractors trained under FACE. For a test sample $\mathbf{x}$, client i outputs a normalized feature $\tilde{\mathbf{z}}_i(\mathbf{x}) = g_i(\mathbf{x})/\|g_i(\mathbf{x})\|_2$. The server fuses client features by weighted averaging:

$$\mathbf{z}_{\text{avg}}(\mathbf{x}) = \sum_{i=1}^{N} w_i \, \tilde{\mathbf{z}}_i(\mathbf{x}), \qquad w_i \geq 0, \quad \sum_{i=1}^{N} w_i = 1, \tag{16}$$

where w_i can be uniform or proportional to client data size. The fused feature is then normalized:

$$\tilde{\mathbf{z}}_{\text{fused}}(\mathbf{x}) = \frac{\mathbf{z}_{\text{avg}}(\mathbf{x})}{\|\mathbf{z}_{\text{avg}}(\mathbf{x})\|_2}. \tag{17}$$

The final prediction is obtained by cosine similarity to the shared reference matrix:

$$\hat{y} = \arg\max_{c \in [C]} \tilde{\mathbf{z}}_{\text{fused}}(\mathbf{x})^\top \mathbf{p}_c. \tag{18}$$

This inference requires no additional server-side calibration or optimization.

5 Experiments

5.1 Experimental Setup

Datasets We evaluate our method on three widely used benchmark datasets: CIFAR-10 [12] contains 60,000 32×32 images in 10 classes. CIFAR-100 [12] contains 60,000 32×32 images in 100 classes. Tiny-ImageNet [13] is a subset of ImageNet with 200 classes and 64×64 images. To simulate realistic federated learning scenarios, we create data heterogeneity using a Dirichlet distribution $\text{Dir}(\alpha)$ with concentration parameter α controlling the degree of heterogeneity. We evaluate across $\alpha \in \{0.05, 0.1, 0.3, 0.5\}$ to demonstrate robustness under varying data distribution scenarios.

Baselines We compare FACE against two fundamental federated algorithms. O-FedAvg [21] applies the traditional FedAvg parameter averaging approach in a one-shot setting, serving as the fundamental baseline for federated learning. Ensemble represents a simple voting scheme where predictions are averaged. We also compare five popular and recent one-shot federated learning methods: DENSE [33], Co-Boosting [3], IntactOFL [32], FuseFL [27], and FAFI [31]. FAFI uses self-alignment local training with learnable prototypes and informative feature fusion.

Configuration Following [31], we use ResNet-18 [7] for CIFAR-10/100 and Tiny-ImageNet. All available test data is used to evaluate the final server model. Main experiments are conducted with 5 clients. Results are reported as averages across five random seeds. More implementation details are in the Appendix.

Table 1: Performance of the proposed method across different one-shot federated learning algorithms on three datasets under various Non-IID conditions. The parameter α controls the degree of Non-IID distribution (smaller α indicates more severe Non-IID conditions). Results show mean ± standard deviation of Top-1 Test Accuracy over five random seeds. The best results are in **bold** while the second best results are underlined.

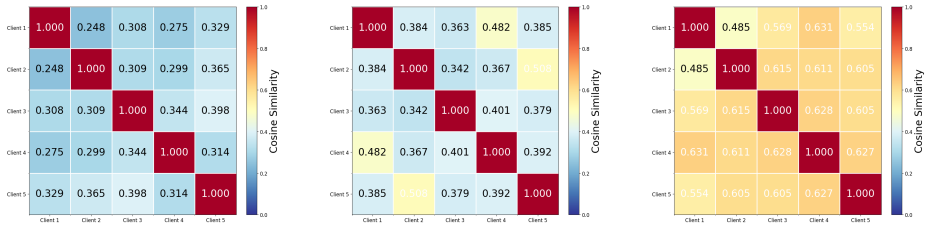
Top-1 Test Accuracy (%)									
Dataset	α	O-FedAvg	Ensemble	DENSE	Co-Boosting	IntactOFL	FuseFL	FAFI	FACE
CIFAR-10	0.5	35.42±0.67	61.61±0.23	62.19±0.12	70.24±2.34	79.93±0.23	**84.34±0.88**	83.88±0.59	83.97±0.36
	0.3	28.07±0.89	62.18±0.34	59.76±0.45	67.21±1.76	70.21±0.60	**81.58±0.91**	77.56±0.43	80.19±0.67
	0.1	17.43±0.51	45.43±0.32	50.26±0.24	58.49±1.24	61.13±0.63	73.79±0.34	68.76±0.72	**73.96±0.51**
	0.05	12.13±2.11	41.36±0.67	38.37±1.08	39.20±0.81	48.22±0.43	54.42±0.41	60.28±0.62	**69.21±0.53**
CIFAR-100	0.5	12.13±0.05	41.61±0.77	38.84±0.39	42.67±1.40	46.78±0.78	49.30±0.32	**54.42±0.41**	53.47±0.45
	0.3	10.67±0.31	38.01±0.67	37.33±0.48	39.30±1.30	41.86±0.60	45.12±0.51	48.76±0.71	**51.18±0.58**
	0.1	6.45±0.71	26.23±0.55	32.03±0.44	27.59±1.35	39.15±0.46	36.86±0.38	41.79±0.47	**45.46±0.39**
	0.05	4.77±0.21	20.46±0.62	18.37±2.43	20.19±1.44	27.99±0.67	29.12±0.23	34.06±0.48	**41.57±0.57**
Tiny-ImageNet	0.5	13.71±0.16	28.50±0.46	32.34±0.32	30.78±2.01	35.09±0.14	34.34±1.81	49.02±0.25	**51.04±0.37**
	0.3	13.61±0.10	17.53±0.31	28.14±0.34	29.24±1.32	30.15±0.12	33.04±1.79	48.28±0.57	**49.32±0.51**
	0.1	8.31±0.21	15.38±0.23	22.25±0.33	21.90±1.20	28.43±0.17	29.28±2.04	39.34±0.35	**42.18±0.42**
	0.05	5.67±0.45	13.28±0.67	18.77±0.67	19.00±1.45	20.45±0.34	22.15±2.11	34.60±0.29	**39.20±0.31**

5.2 Results

Data Heterogeneity As shown in Table 1, FACE outperforms existing one-shot federated learning methods across evaluated datasets and heterogeneity settings, with particularly strong improvements under extreme data heterogeneity. Under severe heterogeneity ($\alpha = 0.05$), FACE achieves 69.21% accuracy on CIFAR-10, 41.57% on CIFAR-100, and 39.20% on Tiny-ImageNet, representing an average improvement of 7% over the best competing method. Even under moderately heterogeneous conditions ($\alpha = 0.1$), FACE maintains an average performance advantage of approximately 3% across all datasets.

These results indicate that FACE effectively addresses the core challenge of one-shot federated learning: distributional mismatch in feature representations. Most existing OFL methods attempt to mitigate this mismatch through server-side post-calibration, relying on mechanisms such as adaptive weighting, selective aggregation, or client filtering. However, these strategies become increasingly ineffective as heterogeneity intensifies. In contrast, FACE reduces representation divergence during client training through Directional Mean Alignment (DMA) and the Variance Stabilizer (VS). DMA encourages class-conditional feature means across clients to align with shared reference directions, while VS maintains sufficient feature dispersion by preventing representation collapse. As a result, the uploaded models exhibit more compatible feature statistics, enabling more effective integration through simple feature-level fusion. The performance gains across multiple datasets, particularly the strong improvements under extreme heterogeneity, support the effectiveness of the proposed prevention-over-correction strategy.

Scalability We assess the scalability of FACE by varying the numbers of clients $N = \{5, 10, 25, 50, 100\}$. We conduct experiments using the CIFAR-10 dataset with $\alpha = 0.5$. As shown in Table 2, FACE consistently achieves strong performance across different numbers of clients.



(a) O-FedAvg **(b)** FAFI **(c)** FACE (Ours)

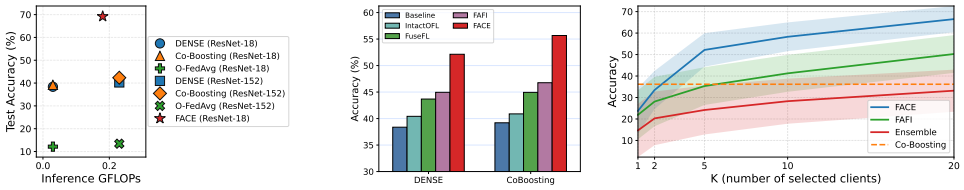
Fig. 3: Feature alignment comparison across different methods. Cosine similarity matrices show: **(a)** O-FedAvg exhibits severe misalignment with low similarities, **(b)** FAFI achieves moderate improvement, and **(c)** FACE demonstrates significantly enhanced alignment with substantially higher similarities.

Existing approaches suffer from significant performance drops as the number of clients increases. This decline happens because when more clients train independently on different data, the distributional mismatch in their learned representations becomes more severe. With more clients, the diversity in representation means and variances increases, making post-calibration approaches less effective since they must reconcile a larger number of misaligned representations. In contrast, FACE performs alignment during client training itself, ensuring that all clients learn distributionally consistent representations regardless of how many clients there are. This prevention strategy enables FACE to maintain stable performance even as the number of clients grows, demonstrating the scalability of our approach.

Table 2: Performance under different numbers of clients, $N = \{5, 10, 25, 50, 100\}$ on CIFAR-10 with $\alpha = 0.5$.

Methods	**Client scales** N				
	5	10	25	50	100
O-FedAvg	35.42	32.09	28.03	28.24	27.14
DENSE	62.19	54.67	49.32	48.67	43.34
Ensemble	61.61	60.44	58.44	52.51	45.72
FuseFL	**84.34**	78.28	62.12	42.18	37.11
IntactOFL	79.93	69.11	64.32	59.45	53.21
FAFI	83.88	81.73	79.21	76.46	71.08
FACE	83.97	**82.61**	**82.39**	**79.34**	**77.18**

Representation Similarity We conduct additional analysis to examine whether FACE effectively aligns feature distributions across clients. Fig. 3 presents cosine similarity matrices of features extracted by models trained on different clients for the same test samples on CIFAR-10 with five clients. O-FedAvg (Fig. 3a) exhibits very low similarity between features produced by different models. This indicates that independently trained models learn incompatible feature spaces when no alignment mechanism is imposed, resulting in substantial divergence in both feature means and variances. FAFI (Fig. 3b) improves similarity by introducing learnable reference directions during training. However, because these directions are optimized independently for each client, their resulting representations still remain partially misaligned. In contrast, FACE (Fig. 3c) produces significantly higher similarity across models. Directional Mean Alignment (DMA) encourages class-conditional feature means to align with shared reference directions,



(a) Budget-matched comparison **(b)** Distillation evaluation **(c)** Random-K selection evaluation

Fig. 5: Inference efficiency analysis under extreme heterogeneity.

while the Variance Stabilizer (VS) prevents representation collapse and maintains consistent feature dispersion. Consequently, models trained under FACE exhibit more compatible feature distributions, enabling effective feature fusion at inference time. These results demonstrate that FACE achieves distributional alignment during local training, directly addressing the representation divergence identified in our problem analysis. Additional experiments on parameter variance are provided in the Appendix.

Training Efficiency Analysis To verify the training efficiency of our method, we conduct additional analysis. Fig. 4 compares our method with existing baselines in terms of accuracy versus GFLOPs. Here, GFLOPs represent the total training cost including both client-side and server-side training. All experiments are conducted under identical configurations on CIFAR-10 $\alpha = 0.05$. As shown in Fig. 4, our method achieves the highest performance while requiring significantly lower training cost. The lower cost compared to O-FedAvg stems from the use of a fixed reference matrix, which does not require classifier training. While Ensemble is cost-efficient, its overall efficiency suffers due to lower performance. Methods requiring server-side training, including IntactOFL, Co-Boosting, and DENSE, incur significant training cost overhead yet yield inferior performance compared to our approach. Although FAFI avoids server-side training, it necessitates higher training cost than standard cross-entropy loss due to client-side contrastive learning mechanisms. These results demonstrate that our method provides the most favorable accuracy-efficiency trade-off among all evaluated approaches, offering significant advantages for OFL.

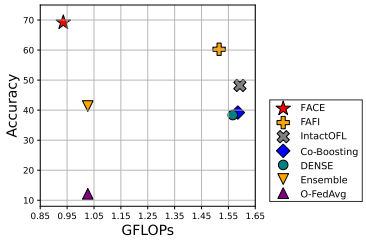


Fig. 4: Accuracy vs. computational cost (GFLOPs) trade-off analysis on CIFAR-10.

Inference Efficiency Analysis To assess the inference efficiency of our method, we conduct additional analysis under the extreme heterogeneity setting with $\alpha = 0.05$. Since FACE retains all N client models at inference, we examine its cost from three angles: whether its gain merely reflects ensemble capacity,

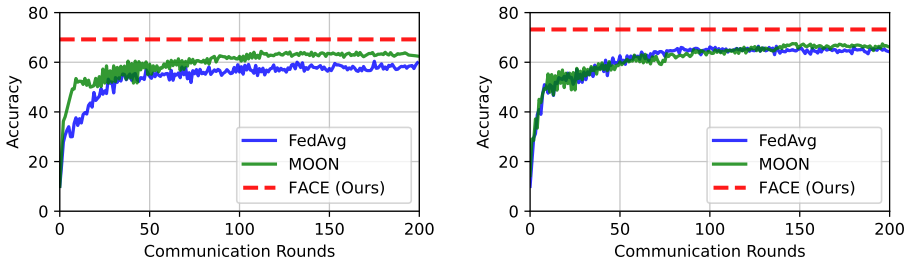
whether it can be compressed into a single model, and how it degrades when only a subset of clients is used.

Compared with methods that aggregate client knowledge into a single global model, FACE incurs higher inference latency and storage when all N client models are retained at the server, as test samples must be processed by multiple client feature extractors. However, this cost profile is not unique to FACE; ensemble-style OFL methods that preserve multiple client models at inference exhibit similar scaling behavior. To disentangle algorithmic gains from model-capacity effects, we compare FACE with single-model baselines under matched inference FLOPs, scaling the baseline model so that its inference cost matches the total inference cost of FACE. As shown in Fig. 5a, FACE consistently and substantially outperforms the budget-matched single model across all settings, confirming that its performance gains arise from genuine representation-level improvements rather than simply from increased model capacity.

We further evaluate whether the gains of FACE can be transferred to a compact single model through distillation. Fig. 5b reports student-model accuracy under the DENSE and Co-Boosting pipelines with Baseline, IntactOFL, FuseFL, FAFI, and FACE as teacher ensembles. FACE consistently yields the highest distilled accuracy across both pipelines by a substantial margin, demonstrating that the distributional alignment enforced during local training produces stronger and more consistent supervisory signals, and that a significant portion of the full ensemble gain can be effectively retained in a more efficient single-model deployment.

Finally, we analyze inference behavior when only K out of N client models are used. For each value of K, we repeatedly sample client subsets uniformly at random and report accuracy with mean and standard deviation. As shown in Fig. 5c, FACE demonstrates markedly superior performance across all values of K, already matching Co-Boosting at $K = 2$ and surpassing it by a clear margin at $K = 5$, while FAFI and Ensemble remain substantially below throughout. Critically, FACE exhibits significantly lower variance across random client subsets compared to all competing methods, indicating that the representation alignment induced by DMA and VS produces individually reliable client models whose contributions remain effective regardless of which subset is selected. Collectively, these results establish that the inference overhead of FACE can be substantially mitigated in practice through distillation or client subset selection, without sacrificing the core performance advantages of the proposed framework.

Ablation Study We conduct comprehensive ablation studies to examine how each component of FACE contributes to performance. Table 3 shows that both components provide substantial improvements over the baseline. When applied individually, both Directional Mean Alignment (DMA) and the Variance Stabilizer (VS) lead to clear performance gains. DMA reduces mean divergence by encouraging class-conditional feature means across clients to align with shared reference directions, while VS prevents representation collapse and maintains sufficient feature dispersion. Applying DMA alone aligns feature directions but



(a) CIFAR-10 ($\alpha = 0.05$) **(b)** CIFAR-10 ($\alpha = 0.1$)

Fig. 6: Comparison with multi-round FL methods on CIFAR-10. While FedAvg and MOON require 200 communication rounds to gradually approach cross-client alignment, FACE attains it in a single round, consistently surpassing FedAvg and remaining competitive with MOON.

leaves their dispersion unconstrained, so features tend to concentrate around the shared references and part of the achievable gain remains unrealized until VS counteracts this collapse. This suggests that an alignment-only objective, as used in fixed-classifier designs, is insufficient on its own, and that dispersion control is a necessary complement rather than an optional regularizer. In other words, effective distributional alignment requires controlling both the first and second moments of feature representations. Combining DMA and VS yields the best performance, with FACE consistently outperforming its variants across different heterogeneity settings. Beyond these two training-time objectives, Feature-level Ensemble (FLE) contributes only marginally on its own, yet its benefit grows once DMA and VS are in place, indicating that feature fusion becomes effective only after client representations are made compatible, in line with our align-then-fuse design. These results confirm that jointly addressing directional mean alignment and variance stabilization during local training mitigates distributional mismatch and enables more effective feature fusion. Additional hyperparameter sensitivity analyses are provided in the Appendix.

Table 3: Ablations on different components of our method."DMA" for Directional Mean Alignment, "VS" for Variance Stabilizer, and "FLE" for Feature-level Ensemble.

DMA	VS	FLE	CIFAR-10	CIFAR-100
			41.46	20.46
✓			59.39	28.83
	✓		55.76	25.89
✓		✓	61.97	30.14
	✓	✓	56.41	27.05
✓	✓		65.37	34.07
✓	✓	✓	**69.21**	**41.57**

Comparison with Multi-round FL In this section, we compare FACE against two widely-used multi-round algorithms, FedAvg [21] and MOON [15], run over 200 communication rounds with 5 clients performing 1 local epoch per round. As shown in Fig. 6, FACE consistently surpasses multi-round FedAvg and remains competitive with MOON, despite operating within a single communication round. This contrast reflects a difference in mechanism rather than mere efficiency. Multi-round methods mitigate client drift incrementally, relying on

repeated synchronization to pull divergent local models back toward a shared solution round after round. FACE instead eliminates the source of divergence during local training, establishing in a single round the cross-client alignment that multi-round methods must progressively reconstruct. Its strong single-round performance is therefore a direct consequence of Prevention over Correction, not of extended optimization. Crucially, this decouples accuracy from the prolonged interaction that multi-round training presupposes, avoiding the communication overhead, client dropout, and accumulated attack surface inherent to that paradigm. More results are provided in the Appendix.

6 Conclusion

We presented FACE, a one-shot federated learning framework designed to address distributional mismatch in feature representations. We showed that independent training on heterogeneous data causes learned representations to develop inconsistent statistical structures, particularly in their means and variances, which hinders effective model aggregation. To address this, FACE employs two complementary mechanisms: DMA aligns class-conditional feature means toward shared reference directions, while VS preserves sufficient feature dispersion by discouraging representation collapse. By jointly regularizing feature means and variances during local optimization, FACE reduces cross-client distributional mismatch and enables effective feature-level fusion without server-side post-calibration. Extensive experiments on CIFAR-10/100 and Tiny-ImageNet demonstrate that FACE achieves state-of-the-art performance, particularly under severe data heterogeneity where distributional mismatch is most pronounced.

Acknowledgements

This work was supported by Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2019-II190421, 16%; IITP-2026-RS-2024-00437633, 16%; RS-2025-25442569, 16%; RS-2023-00228970, 16%) and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST) (RS-2024-00352717, 16%) and the Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (RS-2025-25410203, 20%).

References

1. Chen, H.Y., Chao, W.L.: Fedbe: Making bayesian model ensemble applicable to federated learning. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=dgtpE6gKjHn
2. Chen, J., Zhu, J., Zheng, Q.: Towards fast and stable federated learning: Confronting heterogeneity via knowledge anchor. Proceedings of the 31st ACM International Conference on Multimedia (2023), https://api.semanticscholar.org/CorpusID:264492141

3. Dai, R., Zhang, Y., Li, A., Liu, T., Yang, X., Han, B.: Enhancing one-shot federated learning through data and ensemble co-boosting. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=tm8s3696Ox
4. Dennis, D.K., Li, T., Smith, V.: Heterogeneity for the win: One-shot federated clustering. In: International Conference on Machine Learning (2021), https://api.semanticscholar.org/CorpusID:232075682
5. Gong, X., Sharma, A., Karanam, S., Wu, Z., Chen, T., Doermann, D., Innanje, A.: Ensemble attention distillation for privacy-preserving federated learning. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15056–15066 (2021). https://doi.org/10.1109/ICCV48922.2021.01480
6. Han, S., Park, S., Wu, F., Kim, S., Wu, C., Xie, X., Cha, M.: Fedx: Unsupervised federated learning with cross knowledge distillation. ArXiv **abs/2207.09158** (2022), https://api.semanticscholar.org/CorpusID:250644323
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
8. Heinbaugh, C.E., Luz-Ricca, E., Shao, H.: Data-free one-shot federated learning under very high statistical heterogeneity. In: International Conference on Learning Representations (2023), https://api.semanticscholar.org/CorpusID:259298646
9. Kang, Y., Kim, H.S., Lee, J.H.: Reducing computational cost in federated ensemble learning via rank-one matrix. 2022 Joint 12th International Conference on Soft Computing and Intelligent Systems and 23rd International Symposium on Advanced Intelligent Systems (SCIS&ISIS) pp. 1–5 (2022), https://api.semanticscholar.org/CorpusID:255418092
10. Kang, Y., Lee, J.H.: Federated domain generalization with source knowledge preservation via discriminative ensembles. Information Sciences **727**, 122804 (2026). https://doi.org/https://doi.org/10.1016/j.ins.2025.122804, https://www.sciencedirect.com/science/article/pii/S0020025525009405
11. Karimireddy, S.P., Kale, S., Mohri, M., Reddi, S., Stich, S., Suresh, A.T.: SCAFFOLD: Stochastic controlled averaging for federated learning. In: Proceedings of the 37th International Conference on Machine Learning. vol. 119, pp. 5132–5143. PMLR (2020), https://proceedings.mlr.press/v119/karimireddy20a.html
12. Krizhevsky, A.: Learning multiple layers of features from tiny images (2009), http://www.cs.toronto.edu/~kriz/cifar.html
13. Le, Y., Yang, X.S.: Tiny imagenet visual recognition challenge (2015), https://api.semanticscholar.org/CorpusID:16664790
14. Li, Q., Diao, Y., Chen, Q., He, B.: Federated learning on non-iid data silos: An experimental study. 2022 IEEE 38th International Conference on Data Engineering (ICDE) pp. 965–978 (2021), https://api.semanticscholar.org/CorpusID:231786564
15. Li, Q., He, B., Song, D.: Model-contrastive federated learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
16. Li, Q., He, B., Song, D.: Practical one-shot federated learning for cross-silo setting. In: Zhou, Z.H. (ed.) Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21. pp. 1484–1490. International Joint Conferences on Artificial Intelligence Organization (8 2021). https://doi.org/10.24963/ijcai.2021/205, https://doi.org/10.24963/ijcai.2021/205, main Track

17. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. In: Dhillon, I., Papailiopoulos, D., Sze, V. (eds.) Proceedings of Machine Learning and Systems. vol. 2, pp. 429–450 (2020)
18. Li, Z., Shang, X., He, R., Lin, T., Wu, C.: No fear of classifier biases: Neural collapse inspired federated learning with synthetic and fixed classifier. 2023 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 5296–5306 (2023), https://api.semanticscholar.org/CorpusID:257623057
19. Lin, T., Kong, L., Stich, S.U., Jaggi, M.: Ensemble distillation for robust model fusion in federated learning. In: Advances in Neural Information Processing Systems. vol. 33, pp. 2351–2363 (2020), https://proceedings.neurips.cc/paper_files/paper/2020/file/18df51b97ccd68128e994804f3eccc87-Paper.pdf
20. Liu, X., Liu, L., Ye, F., Shen, Y., Li, X., Jiang, L., Li, J.: Fedlpa: One-shot federated learning with layer-wise posterior aggregation. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing Systems. vol. 37, pp. 81510–81548. Curran Associates, Inc. (2024). https://doi.org/10.52202/079017-2591, https://proceedings.neurips.cc/paper_files/paper/2024/file/9482f45fdd89aba9130bb04c44f788a9-Paper-Conference.pdf
21. McMahan, B., Moore, E., Ramage, D., Hampson, S., Arcas, B.A.y.: Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. vol. 54, pp. 1273–1282. PMLR (20–22 Apr 2017)
22. Oh, J., Kim, S., Yun, S.Y.: Fedbabu: Toward enhanced representation for federated image classification. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=HuaYQfggn5u
23. Shi, Y., Liang, J., Zhang, W., Tan, V., Bai, S.: Towards understanding and mitigating dimensional collapse in heterogeneous federated learning. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=EXnIyMVTL8s
24. Son, H.M., Kim, M.H., Chung, T.M., Huang, C., Liu, X.: Feduv: Uniformity and variance for heterogeneous federated learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5863–5872 (June 2024)
25. Su, S., Li, B., Xue, X.: One-shot federated learning without server-side training. Neural Networks **164**, 203–215 (2023). https://doi.org/https://doi.org/10.1016/j.neunet.2023.04.035, https://www.sciencedirect.com/science/article/pii/S0893608023002216
26. Tan, Y., Long, G., Liu, L., Zhou, T., Lu, Q., Jiang, J., Zhang, C.: Fedproto: Federated prototype learning across heterogeneous clients. In: AAAI Conference on Artificial Intelligence (2022)
27. Tang, Z., Zhang, Y., Dong, P., ming Cheung, Y., Zhou, A.C., Han, B., Chu, X.: Fusefl: One-shot federated learning through the lens of causality with progressive model fusion. Neural Information Processing Systems **abs/2410.20380** (2024), https://api.semanticscholar.org/CorpusID:273654190
28. Wang, J., Liu, Q., Liang, H., Joshi, G., Poor, H.V.: Tackling the objective inconsistency problem in heterogeneous federated optimization. In: Advances in Neural Information Processing Systems. vol. 33, pp. 7611–7623 (2020), https://proceedings.neurips.cc/paper_files/paper/2020/file/564127c03caab942e503ee6f810f54fd-Paper.pdf

29. Wu, X., Yao, X., Wang, C.L.: Fedscr: Structure-based communication reduction for federated learning. IEEE Transactions on Parallel and Distributed Systems **32**, 1565–1577 (2021), https://api.semanticscholar.org/CorpusID:232043479
30. Yang, M., Su, S., Li, B., Xue, X.: One-shot heterogeneous federated learning with local model-guided diffusion models. In: Forty-second International Conference on Machine Learning (2025), https://openreview.net/forum?id=PqJFVbJAMR
31. Zeng, H., Huang, W., Zhou, T., Wu, X., Wan, G., Chen, Y., Cai, Z.: Does one-shot give the best shot? mitigating model inconsistency in one-shot federated learning. In: Forty-second International Conference on Machine Learning (2025), https://openreview.net/forum?id=2XvF67vbCK
32. Zeng, H., Xu, M., Zhou, T., Wu, X., Kang, J., Cai, Z., Niyato, D.T.: One-shot-but-not-degraded federated learning. Proceedings of the 32nd ACM International Conference on Multimedia (2024), https://api.semanticscholar.org/CorpusID:273645346
33. Zhang, J., Chen, C., Li, B., Lyu, L., Wu, S., Ding, S., Shen, C., Wu, C.: Dense: Data-free one-shot federated learning. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (eds.) Advances in Neural Information Processing Systems (2022), https://openreview.net/forum?id=QFQoxCFYEkA
34. Zhang, J., Liu, S., Wang, X.: One-shot federated learning via synthetic distiller-distillate communication. Neural Information Processing Systems **abs/2412.05186** (2024), https://api.semanticscholar.org/CorpusID:274581292
35. Zhang, L., Luo, Y., Bai, Y., Du, B., yu Duan, L.: Federated learning for non-iid data via unified feature learning and optimization objective alignment. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 4400–4408 (2021), https://api.semanticscholar.org/CorpusID:244426444