

# FEDFA: FEDERATED LEARNING WITH FEATURE ALIGNMENT FOR HETEROGENEOUS DATA

Anonymous authors

Paper under double-blind review

## ABSTRACT

Federated learning allows multiple clients to collaboratively train a model without exchanging their data, thus preserving data privacy. Unfortunately, it suffers significant performance degradation under heterogeneous data at clients. Common solutions involve designing specific regularizers for local-model training or developing aggregation schemes for global-model aggregation. Nevertheless, we found that these methods fail to achieve the desired performance due to neglecting the importance of feature mapping consistency across client models. We first observe and analyze that, with heterogeneous data, a *vicious cycle* exists between classifier divergence and feature mapping inconsistency across clients, thereby shifting the aggregated global model from the expected optima. We then propose a simple yet effective framework named *Federated learning with Feature Alignment* (FedFA) to tackle the data heterogeneity problem from a novel perspective of shared feature space. A key insight of FedFA is introducing feature anchors to align the feature mappings and calibrate the classifier updates across clients during their local updates, such that client models are updated in a shared feature space. We prove that this modification brings a property of consistent classifier updates if features are class-discriminative. Extensive experiments show that FedFA significantly outperforms the state-of-the-art federated learning algorithms on various image classification datasets under both label and feature distribution skews.

## 1 INTRODUCTION

With massive data located at edge clients of large-scale networks like the internet of things networks, federated learning (McMahan et al., 2017) enables clients to jointly train a machine learning model without collecting client data into a centralized server, thus preserving data privacy. However, the private data are typically heterogeneous across clients, resulting in slower convergence (Li et al., 2020; Karimireddy et al., 2020) and degraded generalization performance (Zhao et al., 2018; Li et al., 2021a). This is because data heterogeneity makes the local objectives inconsistent with the global objective and thus the converged model deviates from the expected optima (Wang et al., 2020b).

Common works tackle the data heterogeneity problem by improving federated optimization on the client or server side. They either add a regularizer in the local objective to control local training on the client side, such as (Li et al., 2020; Acar et al., 2021), or improve aggregation schemes to control the converged optima on the server side, such as (Wang et al., 2020a;b). Nevertheless, recent works (Li et al., 2021a; Chen & Chao, 2022; Luo et al., 2021) found that these methods did not show clear advantages over the canonical method FedAvg (McMahan et al., 2017) on classification tasks.

**Data heterogeneity causes feature mapping inconsistency during local updates.** To unravel the underlying reasons for the ineffectiveness of existing methods, we first observe that data heterogeneity (including heterogeneous label and feature distributions across clients) can induce feature mapping inconsistency across client models. We analyze the existence of a *vicious cycle* between feature inconsistency and classifier divergence across clients, as shown in Figure 1(a). Specifically, inconsistent features diverge the classifier updates, and then the diverged classifiers force feature extractors to map more inconsistent features, thus diverging client updates.

**A new federated learning framework with feature alignment.** We propose a simple yet effective framework called *Federated learning with Feature Alignment* (FedFA) for classification tasks to address the skewed label and feature distributions across clients. FedFA introduces the feature anchors

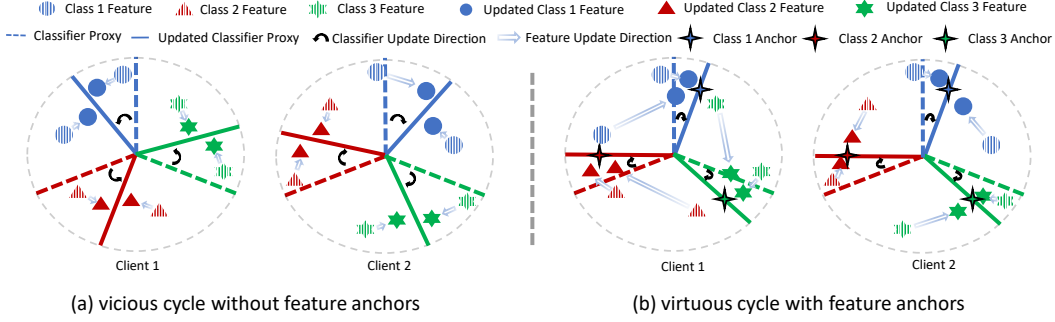


Figure 1: A toy example with three inconsistent class features between two clients to show the rationale of FedFA. Figures 1(a) and 1(b) illustrate the relationship between feature and classifier updates without and with feature anchors, respectively. Since clients keep minimizing the angle between features and their corresponding classifier proxies during local training, the *vicious cycle* in Figure 1(a) describes that the inconsistent feature (to be verified in Figure 2) makes the classifier proxy updates of client 1 contrary to that of client 2, such that feature extractors of client 1 and client 2 to map more inconsistent features to decrease the classification error in local training. The *virtuous cycle* in Figure 1(b) means that feature anchors align class features and calibrate the updates of classifiers between client 1 and client 2, which in turn promotes the consistency between clients’ feature mappings (to be verified in Figure 3).

to regularize where clients should extract the features in a shared feature space and calibrate the updates of classifiers locally. We show theoretically and empirically that FedFA has a property of similar classifier updates under consistent feature maps, which brings a *virtuous cycle* between classifier and feature updates as shown in Figure 1(b), contrary to the above *vicious cycle*. Meanwhile, our experiments show that FedFA significantly outperforms the existing methods under label distribution skew, feature distribution skew, and their combined skew. To the best of our knowledge, we are the first to study the combined label and feature distribution skews.

**Contributions.** The main contributions of this work are summarized as follows:

- We illustrate that data heterogeneity can cause feature mapping inconsistency across client models and can induce a *vicious cycle* between classifier update divergence and feature inconsistency.
- We propose a novel framework FedFA to overcome skewed label and feature distributions, which aligns the features of different clients and calibrates the classifier updates with feature anchors to update all client models in a consistent feature space.
- We analyze that FedFA can bring a *virtuous cycle* between feature consistency and classifier update harmony. Moreover, our experiments show the significant advantage of FedFA over the state-of-the-art algorithms under various data heterogeneity settings.

## 2 RELATED WORK

Due to the limited space, we mainly introduce the methods close to ours (i.e., federated optimization-based methods). Please see Appendix B and E.3 for more detailed discussion.

**Tackle data heterogeneity on the client side.** Many works add a well-designed regularizer to control local updates so that local models will not be far away from the global model to avoid local models converging to their local minima instead of global minima. For example, FedProx (Li et al., 2020) uses the Euclidean distance between the local and global models as a regularizer. FedDyn (Acar et al., 2021) modifies the local objective with a dynamic regularizer consisting of a linear term based on the first-order condition and the Euclidean-distance term, such that the local minima are consistent with the global minima. MOON (Li et al., 2021b) introduces a model-contrastive regularizer to maximize (minimize) the agreement of the features extracted by the local model and that by the global model (the local model of the previous round). In place of the model-contrastive term, FedProc (Mu

et al., 2021) adds a prototype-contrastive term to regularize the features within each class with class prototypes (Snell et al., 2017). Besides, instead of implicit correction by regularization, other works reduce the bias explicitly in local updates by controlling variates, like SCAFFOLD (Karimireddy et al., 2020) that introduces a control variate to correct the client updates to trace the global update. Nevertheless, similar to (Li et al., 2021a; Chen & Chao, 2022; Luo et al., 2021), our experiments in Section 5 show that these works fail to consistently outperform FedAvg in image classification tasks, which motivates us to analyze the relationship between classifier updates and feature mappings.

**Tackle data heterogeneity on the server side.** Many works have also developed alternative aggregation schemes to control the converged optima on the server side. For instance, the authors of (Wang et al., 2020b) find an objective inconsistency problem induced by a different number of local updates because of unbalanced client data, and propose FedNova to eliminate the inconsistency by normalizing the local updates before averaging. Beyond layer-weighted averaging, some works like FedMA (Wang et al., 2020a) and Fed<sup>2</sup> (Yu et al., 2021) introduce neuron-wise averaging because there may exist neuron mismatching from permutation invariance of neural networks. These ideas can complement our work that only performs modification on the client side.

Our work aims at the typical federated learning setting as (McMahan et al., 2017) and tries to improve the local optimization by feature alignment. MOON and FedProc are most similar to ours since they also work from the perspective of feature maps. Different from our method, MOON and FedProc do not involve the feature distribution skew and neglect the *vicious cycle* between classifier update divergence and feature mapping inconsistency, thus impairing their performance.

### 3 INCONSISTENT FEATURE MAPPINGS BETWEEN CLIENTS

#### 3.1 PROBLEM SETUP

Federated learning (McMahan et al., 2017) trains a global model parameterized by vector  $\mathbf{w}$  by collaborating a total of  $N$  clients with a server to solve the following optimization problem:

$$\min_{\mathbf{w} \in \mathbb{R}^d} \mathcal{L}(\mathbf{w}) := \mathbb{E}_i[\mathcal{L}_i(\mathbf{w})] = \sum_i^N \frac{n_i}{n} \mathcal{L}_i(\mathbf{w}) \quad (1)$$

where  $n = \sum_i n_i$  represents the total sample size with  $n_i$  being the sample size of the  $i$ -th client, and  $\mathcal{L}_i(\mathbf{w}) := \mathbb{E}_{\xi \in \mathcal{D}_i}[l_i(\mathbf{w}; \xi)]$  is the local objective function in local dataset  $\mathcal{D}_i$  of the  $i$ -th client. However,  $\mathcal{D}_i$  may differ (i.e., data heterogeneity) between clients such that federated training would not be comparable to centralized training with the global dataset  $\mathcal{D} = \cup_i^N \mathcal{D}_i$  (Zhao et al., 2018).

Suppose that the global dataset  $\mathcal{D}$  consists of  $C$  classes indexed by  $[C]$  for classification tasks. Let  $(\mathbf{x}, y) \in \mathcal{D}$  and  $\mathcal{D} \subseteq \mathcal{X} \times [C]$  where  $(\mathbf{x}, y)$  denotes a sample  $\mathbf{x}$  in the input-feature space  $\mathcal{X}$  with the corresponding label  $y$  in the label space  $[C]$ . We represent  $[C_i]$  ( $\cup_{i=1}^N [C_i] = [C]$ ) as a subset of  $[C]$  and  $\mathcal{D}_{i,c} = \{(\mathbf{x}, c) \in \mathcal{D}_i; c \in [C_i]\}$  as the subset of  $\mathcal{D}_i$  with the label  $c$  at the  $i$ -th client. Moreover, we decompose the classification model parameterized by  $\mathbf{w} = \{\boldsymbol{\theta}, \phi\}$  into a *feature extractor* (i.e., other layers except the last layer of the model denoted by  $f_{\boldsymbol{\theta}} : \mathcal{X} \rightarrow \mathcal{H}$ ) and a *linear classifier* (i.e., the last layer of the model denoted by  $f_{\phi} : \mathcal{H} \rightarrow \mathbb{R}^{[C]}$ ). Specifically, the feature extractor maps a sample  $\mathbf{x}$  into a feature vector  $\mathbf{h} = f_{\boldsymbol{\theta}}(\mathbf{x})$  in the feature space  $\mathcal{H}$ , and then the classifier, given the feature  $\mathbf{h}$ , generates a probability distribution  $f_{\phi}(\mathbf{h})$  as the prediction for  $\mathbf{x}$ .

#### 3.2 MOTIVATION: FEATURE MAPPING INCONSISTENCY DUE TO DATA HETEROGENEITY

This work considers both label and feature distribution skews with concrete definitions provided in Appendix A. Briefly, label (feature) distribution skew denotes the label (input feature) distribution variation across clients with the same conditional distribution given the label (input feature).

**Experimental validation.** We first perform validation experiments by FedAvg with ten clients to train a CNN model with two convolutional neural layers to show the feature mapping inconsistency under both label and feature distribution skews. Herein, we choose FMNIST (Xiao et al., 2017) (a mixed-digit dataset consisting of five common digit datasets from (Li et al., 2021c)) as the training set for label (feature) distribution skew. We separately sample a subset from their corresponding test sets to visualize the feature mappings of the local models based on t-SNE visualization (Van der

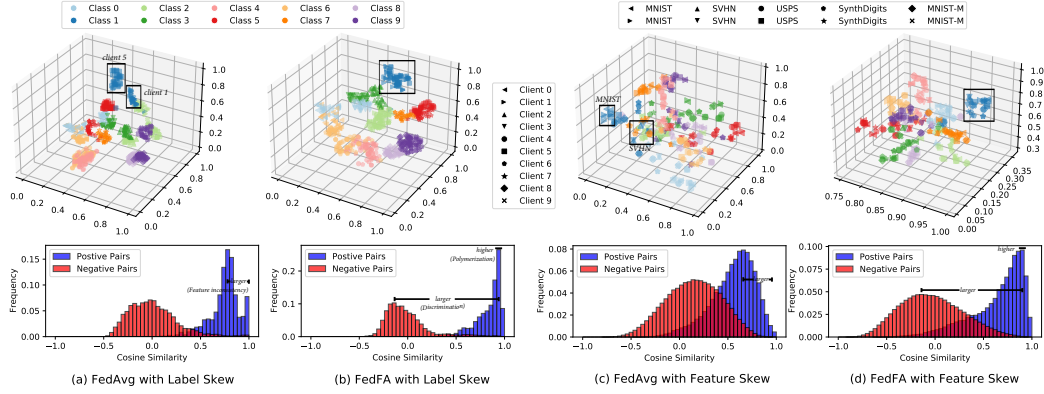


Figure 2: The t-SNE visualization and the histogram of feature cosine similarity on FedAvg and FedFA (Our) under label and feature distribution skews. Taking class 1 (i.e., dark blue) as the example, given the same samples of class 1 into client local models, Figure 2(a) shows that client 1 (i.e., right triangle) and client 5 (i.e., square) extract inconsistent features (in the black box) under label distribution skew, and Figure 2(c) presents the feature mappings of MNIST (i.e., left and right triangles) deviate from that of SVHN (i.e., up and down triangles) under feature distribution skew. In contrast, the inconsistency is significantly mitigated in Figures 2(b) and 2(d). The histograms of FedAvg show a lower similarity for *positive pairs* and a smaller gap between *positive pairs* and *negative pairs* than that of FedFA, meaning more inconsistent features across clients in FedAvg.

Maaten & Hinton, 2008) and min-max normalization, where Appendix E.2 describes the specific experiment settings. We also plot the histogram of feature cosine similarity according to the *positive (negative) pairs*, i.e., features with (without) the same label. Note that we visualize features according to the classes (digit dataset) owned by a client under label (feature) distribution skew. We direct readers to Appendix F.1 for feature visualizations of baselines under the same setting.

According to Figures 2(a) and 2(c), the feature mapping inconsistency across client models exists under both label and feature distribution skews, such as class 1 (i.e., dark blue), class 5 (i.e., dark red) and class 9 (i.e., dark purple), which is significantly alleviated by FedFA proposed in the next section in Figures 2(b) and 2(d). Moreover, the histograms show that both label and feature distribution skews induce a low similarity for *positive pairs* in FedAvg, indicating inconsistent feature space across clients. The histograms of FedAvg also present a low frequency of *positive pairs* and a small gap between *positive pairs* and *negative pairs*, meaning inconsistent polymerization (i.e., a sizeable intra-class feature distance) and discrimination (i.e., a small inter-class feature distance) across client models. Therefore, these results demonstrate that both label and feature distribution skews would cause feature mapping inconsistency across clients, thus inducing divergence of client updates.

**Theoretical demonstration.** To further analyze the influence of data heterogeneity, we follow (Movshovitz-Attias et al., 2017) to represent the classifier parameters  $\phi_i$  of the  $i$ -th client as the weight vectors  $\{\phi_{i,c}\}_{c=1}^C$ , where  $\phi_{i,c}$  is named the  $c$ -th proxy of the  $c$ -th class samples of the  $i$ -th client. For simplicity, we set all the bias vectors of the classifier as zero vectors and use the cross-entropy loss as the supervised loss. The classifier forward process can be represented as:

$$\mathcal{L}_{\text{sup}_i}(\phi_i) := \mathbb{E}_{(\mathbf{x}_j, y_j) \in \mathcal{D}_i} [l_{\text{sup}_i}(\phi_i; (\mathbf{x}_j, y_j))] = -\frac{1}{n_i} \sum_{j=1}^{n_i} \sum_{c=1}^C \mathcal{I}\{y_j = c\} \log \frac{\exp(\phi_{i,c}^T \mathbf{h}_{i,y_j})}{\sum_{q=1}^C \exp(\phi_{i,q}^T \mathbf{h}_{i,y_j})} \quad (2)$$

where  $\mathcal{I}(\cdot)$  is the indication function, and  $\mathbf{h}_{i,y_j} = f_{\theta_i}(\mathbf{x}_j)$  is the feature mapping of a sample  $(\mathbf{x}_j, y_j)$ .

Firstly, for the  $c$ -th proxy  $\phi_{i,c}$  of the  $i$ -th client, the *positive features* and *negative features* denote the features from the  $c$ -th class and other classes, respectively. Let  $p_{j,c} = \frac{\exp(\phi_{i,c}^T \mathbf{h}_{i,y_j})}{\sum_{q=1}^C \exp(\phi_{i,q}^T \mathbf{h}_{i,y_j})}$  and compute the update of  $\phi_{i,c}$  with the learning rate  $\eta$  as:

$$\Delta\phi_{i,c} = -\frac{\eta}{n_i} \frac{\partial \mathcal{L}_{\text{sup}_i}}{\partial \phi_{i,c}} = \underbrace{\frac{\eta}{n_i} \sum_{j=1, y_j=c}^{n_i} (1 - p_{j,c}^{(i)}) \mathbf{h}_{i,y_j}}_{\text{update by positive features}} - \underbrace{\frac{\eta}{n_i} \sum_{j=1, y_j \neq c}^{n_i} p_{j,c}^{(i)} \mathbf{h}_{i,y_j}}_{\text{update by negative features}}. \quad (3)$$

For label distribution skew, we assume that client  $i$  and client  $j$  hold the same samples of the  $c$ -th class if  $c \in [C_i] \cap [C_j]$  to ablate the impact of feature distribution skew (i.e.,  $\mathbf{h}_{i,c} = \mathbf{h}_{j,c}$ ). Then we have: (i) for class  $c \in \{[C] \setminus [C_i]\}$ , there is no update of  $\phi_{i,c}$  at client  $i$  by *positive features* since the client does not hold any sample with class  $c$  (i.e.,  $\mathbf{h}_{i,c} = 0$ ). That is, the updates of the  $c$ -th proxy (i.e.,  $\Delta\phi_{i,c}$ ) only depend on *negative features* such that clients that hold different  $\{[C] \setminus [C_i]\}$  update the  $c$ -th proxy inconsistently. (ii) for  $c \in [C_i]$ , due to  $\mathbf{h}_{i,c} = \mathbf{h}_{j,c}$ , the inconsistent  $\Delta\phi_{i,c}$  originates from the different updates by *negative features*. This is because the updates depend on the non- $c$ -th classes in  $[C_i]$ , and clients with the skew of  $[C_i]$  hold different  $\{[C_i] \setminus c\}$ . For feature distribution skew, since the global model initiates client models, the skewed input features of samples induce the client models to map inconsistent features (i.e.,  $\mathbf{h}_{i,c} \neq \mathbf{h}_{j,c}$ ) at the beginning of each round, resulting in different  $\Delta\phi_{i,c}$ . Thus, label and feature distribution skews diverge the classifier updates across clients, which explains the classifier divergence observed in (Luo et al., 2021).

Next, we compute the update of  $\mathbf{h}_{i,y_j}$ , which can be represented as:

$$\Delta\mathbf{h}_{i,y_j} = -\frac{\eta}{n_i} \frac{\partial \mathcal{L}_{\text{sup}_i}}{\partial \mathbf{h}_{i,y_j}} = \underbrace{\frac{\eta}{n_i} (1 - p_{j,y_j}^{(i)}) \phi_{i,y_j}}_{\text{update by positive proxy}} - \underbrace{\frac{\eta}{n_i} \sum_{c=1, c \neq y_j}^C p_{j,c}^{(i)} \phi_{i,c}}_{\text{update by negative proxies}}. \quad (4)$$

Similar to the above analysis of classifier updates, the feature updates are inconsistent across clients because of diverged *positive proxies* and *negative proxies*, which demonstrates feature mapping inconsistency across clients and has been also observed by (Li & Zhan, 2021; Tang et al., 2022).

Finally, by combining (3) and (4), we conclude the existence of a *vicious cycle*: to minimize the supervised loss (2) (i.e., the angle between features and their corresponding proxies) at each round, inconsistent feature mappings make the update of classifiers diverge, and these different classifiers in turn force feature extractors to map to more inconsistent features. Hence, the *vicious cycle* diverges the client updates and deviates the global optima from the expected optima. Based on our experimental and analytical demonstrations, we believe that feature mapping inconsistency is one of the primary reasons for the client update divergence in federated classification tasks.

#### 4 FEDERATED LEARNING WITH FEATURE ALIGNMENT (FedFA)

We propose FedFA to tackle the problem of inconsistent feature mappings by feature alignment, which trains client models in a shared feature space with feature anchors.

**Feature anchor loss.** With a total of  $C$  classes in the whole dataset, the server initiates  $C$  feature anchor vectors  $\{\mathbf{a}_c\}_{c=1}^C \in \mathcal{H} \times [C]$  indexed by  $c \in [C]$  before training, which are to align the each-class feature mappings of feature extractors  $f_{\theta}(\mathbf{x})$  by the following loss:

$$\mathcal{L}_{\text{fa}}(\theta_i) := \mathbb{E}_{(\mathbf{x}_j, y_j) \in \mathcal{D}_i} [l_{\text{fa}}(\theta_i; \mathbf{a}_c, (\mathbf{x}_j, y_j), c = y_j)] = \frac{1}{2n_i} \sum_{j=1, c=y_j}^{n_i} \|\mathbf{h}_{i,c} - \mathbf{a}_c\|^2 \quad (5)$$

where  $\mathbf{h}_{i,c}$  denotes the feature vector extracted by  $f_{\theta_i}(\mathbf{x})$  for a sample  $(\mathbf{x}, c)$ . The feature anchor loss measures the average distance between features and their corresponding feature anchors. Minimizing (5) can reduce the intra-class feature distance for a given client, as well as across all clients, as shown in Figure 1 (b), which helps train client models in a shared feature space.

**Local objective function.** In the local optimization, FedFA adds the feature anchor loss to a standard supervised loss (e.g., the cross-entropy loss as shown in (2) represented as  $\mathcal{L}_{\text{sup}}$ ). At the  $t$ -th round, the server sends the current global model  $\mathbf{w}^{(t-1)}$  and feature anchors  $\{\mathbf{a}_c^{(t-1)}\}_{c=1}^C$  to a set  $\mathcal{S}$  of active clients, and then each client  $i \in \mathcal{S}$  locally updates  $\mathbf{w}^{(t-1)}$  to  $\mathbf{w}_i^{(t)}$  with the following local optimization problem:

$$\min_{\mathbf{w}_i \in \mathbb{R}^d} \mathcal{L}_i(\mathbf{w}_i) := \mathcal{L}_{\text{sup}_i}(\mathbf{w}_i) + \mathcal{L}_{\text{fa}}(\theta_i) := \mathbb{E}_{(\mathbf{x}, y) \in \mathcal{D}_i} [\mathcal{L}_{\text{sup}_i}(\mathbf{w}_i) + \mu l_{\text{fa}}(\theta_i)] \quad (6)$$

where  $\theta_i \in \mathbf{w}_i = \{\theta_i, \phi_i\}$  and  $\mu$  is a hyper parameter to balance  $l_{\text{sup}_i}$  and  $l_{\text{fa}_i}$ .

**Local classifier calibration with feature anchors.** Besides aligning the per-class feature, feature anchors are also used to calibrate the updates of classifier proxies. At each local epoch of the  $t$ -th round, the active client  $i \in \mathcal{S}$  calibrates its own classifier after one mini-batch update. The feature anchors  $\{\mathbf{a}_c^{(t-1)}\}_{c=1}^C$  perform one mini-batch input of the classifier  $f_{\phi_i}$  and their corresponding classes as the label set  $\{C\}$ , which can be described as:

$$\mathcal{L}_{\text{cal}_i}(\phi_i) := \mathbb{E}_{(\mathbf{a}_c, c) \in \{\mathbf{a}_c\}_{c=1}^C} [l_{\text{cal}_i}(\phi_i; (\mathbf{a}_c, c))] = -\frac{1}{C} \sum_{c=1}^C \log \frac{\exp(\phi_{i,c}^T \mathbf{a}_c)}{\sum_{q=1}^C \exp(\phi_{i,q}^T \mathbf{a}_c)}. \quad (7)$$

According to (3), the classifier calibration loss  $l_{\text{cal}_i}$  can correct the classifier divergence across clients by reducing the angle between the  $c$ -th class proxy and feature anchor as shown in Figure 1 (b), thus facilitating to update client models in a consistent feature space.

**Feature anchors updating during federated training.** If feature anchors  $\{\mathbf{a}_c\}_{c=1}^C$  are updated locally based on the gradient of the local loss (5) like (Wen et al., 2016), the updates of  $\mathbf{a}_c$  would be inconsistent under heterogeneous data, which contradicts the target of feature alignment across client models. Therefore, to keep  $\mathbf{a}_c$  consistent in local training, client  $i$  accumulates the  $c$ -th class features by the  $c$ -th class momentum  $\mathbf{m}_{c,i}$  but without updating  $\mathbf{a}_c$ , and the server aggregates all  $\mathbf{m}_{c,i}$  of active clients as the next-round feature anchors according to the class when performing federated aggregation, which can be represented at the  $t$ -th round as:

$$\mathbf{a}_c^{(t)} = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \mathbf{a}_{c,i}^{(t)} = \frac{1}{K|\mathcal{S}|} \sum_{i \in \mathcal{S}} \sum_{k=1}^K \mathbf{m}_{c,i}^{(t,k)} = \frac{1}{K|\mathcal{S}|} \sum_{i \in \mathcal{S}} \sum_{k=1}^K \lambda \bar{\mathbf{h}}_{c,i}^{(t,k-1)} + (1-\lambda) \bar{\mathbf{h}}_{c,i}^{(t,k)}. \quad (8)$$

Here  $k \in [0, K]$  denotes the index of the local epoch in local optimization,  $\bar{\mathbf{h}}_{c,i}^{(t,k)}$  denotes the averaging features of the  $c$ -th class in the  $k$ -th local epoch (i.e.,  $\bar{\mathbf{h}}_{c,i}^{(t,k)} = \frac{1}{|\mathcal{D}_{i,c}|} \sum_{(\mathbf{x}, c) \in \mathcal{D}_{i,c}} f_{\theta_i}(\mathbf{x})$  and  $\bar{\mathbf{h}}_{c,i}^{(t,0)} = \mathbf{a}_c^{(t-1)}$ ). However, it is inefficient to compute  $\bar{\mathbf{h}}_{c,i}^{(t,k)}$  with the entire training dataset at each epoch. To solve this problem, the client performs moving average to estimate  $\bar{\mathbf{h}}_{c,i}^{(t,k)}$  based on the  $c$ -th class samples of one mini batch  $\mathcal{B}_i$  when performing prediction. Moreover,  $\lambda$  denotes a hyper parameter that controls the weight of  $\bar{\mathbf{h}}_{c,i}^{(t,k)}$  in the  $k$ -th local epoch. Adapting the value of  $\lambda$  can reduce the variance of model predictions at different epochs by controlling the information of each epoch  $\bar{\mathbf{h}}_{c,i}^{(t,k)}$  into the accumulated momentum  $\mathbf{m}_{c,i}^{(t,k)}$ . If  $\lambda = 0$ ,  $\mathbf{m}_{c,i}^{(t,k)}$  only depends on  $\bar{\mathbf{h}}_{c,i}^{(t,k)}$  of the  $k$ -th local epoch.

**Analysis of FedFA.** We assume that there exist feature extractors so that the features are class-discriminative to show the property of FedFA on the mitigation of classifier divergence:

**Assumption 1** (*Discriminative features*) There exist a constant  $\delta$ ,  $C$  sub hyperspaces  $\{\mathcal{H}_c\}_{c=1}^C$  in the feature space  $\mathcal{H}$ , and  $C$  feature anchors  $\{\mathbf{a}_c\}_{c=1}^C$  such that for all  $c$ -th class features  $\mathbf{h}_c$ ,  $\mathcal{H}_c = \{\mathbf{h}_c \mid \|\mathbf{h}_c - \mathbf{a}_c\|^2 < \delta^2\}$ , where for  $i \neq j$ ,  $\mathcal{H}_i \cap \mathcal{H}_j = \emptyset$ .

**Property 1** (*Similar classifier updates*). Let Assumption 1 holds. For  $\|\mathbf{a}_c\| > \sqrt{2}\delta$  and the inner product  $\mathbf{h}_c \cdot \mathbf{h}_q \leq 0$  where  $c \neq q$ ,  $\mathbf{h}_c \in \mathcal{H}_c$  and  $\mathbf{h}_q \in \mathcal{H}_q$ , we have:

$$\cos(\Delta\phi_{i,c}, \Delta\phi_{v,c}) > 0. \quad (9)$$

We direct readers to Appendix C for a detailed proof, where the key idea is that the dot product of inter-class features is negative. Therefore, FedFA can achieve similar updates of the  $t$ -th proxy between clients  $i$  and  $v$  even without classifier calibration to regulate classifier updates.

**Benefits of feature anchors.** As shown in Figure 1, FedFA can effectively prevent the *vicious cycle* mentioned in section 3 and achieve a *virtuous cycle*: both feature alignment and classifier calibration can boost consistent classifier updates according to (7) and Property 1, which promotes feature mapping consistency across clients. Feature alignment can also regularize the updates of feature extractors to significantly alleviate feature mapping inconsistency, which is verified in Figure 2. Moreover, the histograms of Figures 2(c) and 2(d) show that FedFA improves the feature polymerization for *positive pairs* and discrimination for *negative pairs* (i.e., FedFA can reduce intra-class feature distance and increase inter-class feature distance across clients to help classification).

## 5 EXPERIMENTS

### 5.1 EXPERIMENTAL SETUP

We briefly describe our experimental setup, including the baselines, datasets, models and federated setup. The experiment and hyper-parameter setup of all methods are described in Appendix E.

**Datasets.** This work aims at image classification tasks under label and feature distribution skews, and it uses federated benchmark datasets as (McMahan et al., 2017; Yurochkin et al., 2019; Li et al., 2021a), including EMNIST (Cohen et al., 2017), FMNIST (Xiao et al., 2017), CIFAR-10, CIFAR-100 (Krizhevsky et al., 2009), and Mixed Digits dataset (Li et al., 2021c). Specifically, for label distribution skew, we consider two settings: (i) Same size of local dataset: following (McMahan et al., 2017), we split data samples based on class to clients (e.g.,  $\#C = 2$  denotes that each client holds two class samples); (ii) Different sizes of local dataset: following (Yurochkin et al., 2019), we set  $\alpha$  of Dirichlet distribution  $Dir(\alpha)$  as 0.1 and 0.5 to generate distribution  $p_{i,c}$  by which the  $c$ -th class samples are splitted to client  $i$ . For feature distribution skew, we consider two settings: (i) Real-world feature skew: we sample a subset with 10 classes of a real-world dataset EMNIST (Cohen et al., 2017) with natural feature skew; (ii) Artificial feature skew: we use a mixed-digit dataset from (Li et al., 2021c)<sup>1</sup> consisting of MNIST (LeCun et al., 1998), SVHN (Netzer et al., 2011), USPS (Hull, 1994), SynthDigits and MNIST-M (Ganin et al., 2015).

**Baselines and models.** We compare FedFA with the canonical federated learning algorithms including FedAvg (McMahan et al., 2017), FedProx (Li et al., 2020) and the state-of-the-art methods including FedDyn (Acar et al., 2021), MOON (Li et al., 2021b) and FedProc (Mu et al., 2021). For a fair comparison, our models follow what is reported in the baselines. Following (Acar et al., 2021), we use a CNN model with two convolution layers for EMNIST, FMNIST, and CIFAR-10. We utilize the ResNet-18 (He et al., 2016) with a linear projector from (Li et al., 2021b) for CIFAR-100 and a CNN model with five batch normalization (BN) layers from (Li et al., 2021c) for Mixed Digits.

**Federated simulation setup.** In total, 100 clients attend federated training, and 10 clients participate in each communication round. We use the SGD optimizer with a 0.01 learning rate and 0.001 weight decay for all experiments except for the CIFAR-100 experiment with 0.9 momentum additionally. The local batch size is 64, the local epochs number is 5, and the targeted communication round is 200. According to Property 1, we initiate the pairwise-orthogonal feature anchors  $\mathbf{a}_c$  by sampling column vector from an identity matrix whose dimension is the same as that of features.

### 5.2 EXPERIMENT RESULTS

**Performance under label distribution skew.** Table 1 shows that FedFA provides significant gains in different label-skew settings regardless of the dataset. Compared with  $\alpha = 0.5$ , both  $\#C = 2$  and  $\alpha = 0.1$  indicate more severe label distribution skew, but clients under  $\#C = 2$  have the same sample number while the ones with  $\alpha = 0.1$  do not. Firstly, we find that the performance of all methods decreases as the degree of data heterogeneity increases. Nevertheless, the decline of FedFA is much smaller than that of other methods. For example, when  $\alpha$  changes from 0.5 to 0.1, the top-1 accuracy of all the baselines goes down by about 13% on FMNIST and CIFAR-10, which is twice as large as FedFA. Secondly, under the same label skew, FedFA achieves larger gains over other methods when label distribution skew becomes more severe, up to 18.06% (i.e., Moon: 34.89% and FedFA 52.95% under  $\alpha = 0.1$  in CIFAR-10). Thirdly, to explore more difficult tasks, we test on CIFAR-100 with ResNet18, and our method still achieves the best performance (i.e., about 3% accuracy advance).

**Performance under feature distribution skew.** According to Table 2, our method obtains higher accuracy than all baselines on EMNIST and Mixed Digits. Specifically, the accuracy of FedFA in EMNIST reaches 99.28%, which is 0.77% higher than the best baseline (i.e., MOON 98.51%). Moreover, we split each digit dataset of Mixed Digits into 20 subsets, one for each client, with the same sample number and label distribution (e.g., a skewed feature distribution exists between the clients with a subset of SVHN and the ones with a subset of MNIST). All the methods with federated BN achieve about 4% gains over the versions with local BN. Compared with the best baseline (i.e., Feddyn: 79.53% under local BN and 83.95% under federated BN) on Mixed Digits, our method achieves performance gains of 6.64% and 6.91%, respectively.

<sup>1</sup>Data source: <https://github.com/med-air/FedBN>

Table 1: The top-1 accuracy of FedFA and all the baselines under label distribution skew on the test datasets. We run three trials and report the mean and standard derivation. For FedAvg and FedFA, we also report their top-1 accuracy without label skew.

Method (lr = 0.01)	Label Distribution Skew								
	#C = 2	FMNIST $\alpha = 0.1$		CIFAR-10 $\alpha = 0.1$			CIFAR-100 $\alpha = 0.1$		
FedAvg w/o skew		85.90(0.14)		59.66(0.05)			25.37(0.28)		
FedFA w/o skew		<b>89.67(0.16)</b>		<b>64.95(0.53)</b>			<b>33.94(0.44)</b>		
FedAvg	74.60(1.42)	69.81(3.00)	82.80(0.65)	36.07(3.02)	35.20(3.72)	48.66(3.00)	22.62(0.84)	21.79(0.79)	26.52(1.09)
FedProx	74.63(1.30)	69.59(2.99)	82.92(0.38)	36.63(2.64)	35.21(3.78)	48.43(2.27)	22.27(0.90)	22.30(0.47)	26.03(0.73)
FedDyn	74.77(1.76)	70.09(2.24)	83.95(0.29)	36.11(3.35)	36.00(3.78)	50.46(2.33)	13.28(2.19)	1.00(0.00)	1.00(0.00)
MOON	74.25(1.59)	68.52(2.26)	82.72(0.42)	35.90(3.17)	34.89(3.18)	48.74(2.45)	22.03(1.00)	22.04(0.62)	26.69(1.03)
FedProc	74.96(1.94)	69.80(3.26)	82.94(0.34)	36.57(3.61)	35.02(4.53)	48.99(2.85)	23.00(0.35)	22.32(0.63)	26.38(0.52)
FedFA (Our)	<b>84.08(1.22)</b>	<b>83.42(1.14)</b>	<b>88.40(0.12)</b>	<b>52.64(1.46)</b>	<b>52.95(2.01)</b>	<b>60.40(0.38)</b>	<b>26.68(1.18)</b>	<b>24.05(2.32)</b>	<b>29.16(1.03)</b>

Table 2: The top-1 accuracy of FedFA and the other baselines under label & feature distribution skews on the test dataset. Note that we report the average top-1 accuracy on five benchmark digit datasets in Mixed Digit for all methods. Federated BN means that the server aggregates the BN layer, while local BN denotes that the BN layer is not aggregated at the server and is stored by clients locally.

Method (lr = 0.01)	Feature Distribution Skew			Label & Feature Distribution Skew					
	EMNIST	Mixed Digits		Mixed Digits with Local BN			Mixed Digits with Federated BN		
		Local BN	Federated BN	#C = 2	$\alpha = 0.1$	$\alpha = 0.5$	#C = 2	$\alpha = 0.1$	$\alpha = 0.5$
FedAvg w/o skew	-	81.67(0.21)	82.33(0.59)	81.67(0.21)			82.33(0.59)		
FedFA w/o skew	-	<b>88.75(0.35)</b>	<b>88.92(0.40)</b>	<b>88.75(0.35)</b>			<b>88.92(0.40)</b>		
FedAvg	98.50(0.04)	78.83(2.51)	82.67(2.60)	54.48(4.92)	60.61(1.52)	75.18(1.11)	54.93(5.84)	62.57(2.14)	78.31(1.75)
FedProx	98.44(0.06)	78.71(2.44)	82.60(2.69)	53.83(5.11)	60.23(1.53)	75.06(1.12)	54.46(6.40)	62.33(2.07)	78.25(1.73)
FedDyn	97.63(0.19)	79.53(1.72)	83.67(2.55)	50.16(6.36)	61.13(1.38)	75.96(1.24)	49.66(7.30)	63.03(1.70)	79.53(1.71)
MOON	98.51(0.06)	78.18(2.43)	81.53(3.28)	55.81(4.72)	61.31(1.45)	75.25(1.25)	55.46(5.51)	62.06(2.05)	78.25(1.83)
FedProc	98.28(0.04)	78.10(2.45)	83.06(2.74)	<b>57.98(3.06)</b>	62.07(1.16)	75.26(0.91)	59.87(3.23)	65.37(2.97)	79.62(1.06)
FedFA (Our)	<b>99.28(0.33)</b>	<b>86.17(1.94)</b>	<b>90.86(1.92)</b>	51.81(1.63)	<b>62.36(0.52)</b>	<b>79.23(0.38)</b>	<b>83.73(2.76)</b>	<b>85.26(0.74)</b>	<b>89.87(0.48)</b>

**Performance under combined label and feature distribution skews.** We combine label skew and feature skew to explore the impact of data heterogeneity further. Namely, we not only split each dataset in Mixed Digits into 20 subsets, one for each client, but also set the different label distributions for each client (i.e., clients are subject to at least one of label distribution skew and feature distribution skew). The results in Table 2 show that all the methods are more susceptible under this setting than that of feature distribution skew. For example, the most significant performance drop reaches 34.01% (i.e., FedDyn with Federated BN drops from 83.67% to 49.66% under  $\#C = 2$ ). Nevertheless, FedFA significantly mitigates this performance degradation with a mild decrease from 90.86% to 83.73%. FedFA with federated BN also substantially alleviates its degradation with local BN and outperforms baselines by at least about 10%, especially keeping 23.86% advantage in  $\#C = 2$ , which indicates the compatibility between FedFA and Federated BN to mitigate data heterogeneity.

**Performance without label or feature distribution skew.** We compare our method with FedAvg under more homogeneous data and take the same learning rate of this case as that of data heterogeneity for comparison, where the results are reported in Table 1 and Table 2. The results demonstrate that FedFA still brings a significant advance in the presence of data homogeneity. For example, FedFA is 8.64% more accurate than FedAvg on CIFAR-100. Incredibly, FedFA under mild data heterogeneity (e.g.,  $\alpha = 0.5$  in FMNIST or Mixed Digits with Federated BN) even obtains higher accuracy than FedAvg without any label or feature skew (e.g., FedFA: 88.40% vs. FedAvg: 85.90% in FMNIST). This illustrates the importance of guaranteeing the consistency of feature mappings.

**Local BN vs. federated BN.** For local BN (Li et al., 2021c), clients would combine their own BN layers with the global model to perform inference on the digit dataset they hold (e.g., client 1 with MNIST would test on the test set of MNIST). For federated BN, we directly test on test sets of all digit datasets based on the global model. Thanks to feature alignment, federated BN sharing the statistics of features helps our method FedFA gain a tremendous advantage over baselines under both label and feature skews, up to 34.07% (i.e., FedDyn: 49.66% vs. FedFA: 83.73% in  $\#C = 2$  setting). Furthermore, Table 2 shows that all baselines with federated BN bring consistent accuracy benefits as the label distributions become less skewed across clients (e.g., increasing by about 3% over local BN under  $\alpha = 0.5$  or without label distribution skew). The results reveal that federated



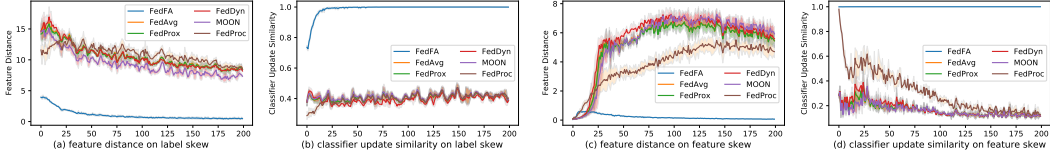


Figure 3: Feature distance (local models given the same inputs) and classifier update similarity for different rounds (x-axis) under label skew ( $\#C = 2$ ) in FMNIST and feature skew in Mixed Digit.

Table 3: The top-1 accuracy of FedFA with federated BN in different ablation settings.

Method (lr = 0.01)	Label Skew (FMNIST $\#C = 2$ )	Feature Skew (Mixed Digits)	Label & Feature Skew (Mixed Digits $\#C = 2$ )
FedFA w/o anchor updating	81.89(1.87)	88.69(0.75)	76.81(1.78)
FedFA w/o classifier calibration	78.07(2.23)	79.25(1.25)	61.36(4.00)
FedFA w/o orthogonal anchor initialization	77.90(2.37)	90.76(0.54)	83.34(1.42)
FedFA (Our)	<b>84.08(1.22)</b>	<b>90.86(1.92)</b>	<b>83.73(2.76)</b>

BN can positively affect federated training in some cases and deserve to be explored more, although there exist some shortcomings as observed in (Li et al., 2021c; Hsieh et al., 2020).

**Classifier similarity and feature distance.** We input the same samples into all local models to compute the mean feature distance and classifier update similarity at the end of each round, as shown in Figure 3. On the one hand, the lower the similarity of classifier updates, the more inconsistent the feature mapping between clients will be according to Figure 3(c) and 3(d), indicating the *vicious cycle* in section 3. All the baselines present a much more prominent feature divergence than FedFA, verifying the effectiveness of preventing inconsistent feature mappings by FedFA. On the other hand, for all the baselines, Figure 3 indicates that the *vicious cycle* under label distribution skew is mildly alleviated as the global model is becoming converged, but the *vicious cycle* under feature distribution skew becomes more serious. Nevertheless, FedFA still breaks the *vicious cycle* to obtain a *virtuous cycle* for feature and classifier updates under both skews, verifying the analysis in section 4.

**Ablation study.** As shown in Table 3, we conduct ablation studies on FedFA without anchor updating in (6), FedFA without classifier calibration in (7), and FedFA without orthogonal anchor initialization according to Property 1 to give an intuition of FedFA performance. First, under both label and feature distribution skews, the anchor updating brings consistent performance benefits (i.e., at least around 2% boost) to FedFA since the updated anchors can keep client models more expressive in the shared feature space. Second, classifier calibration plays the most crucial role in FedFA under feature distribution skew because the skew induces a low classifier update similarity observed in Figure 3. Third, the case of orthogonal anchor initialization verifies the importance of class-discriminative anchors in alleviating performance degradation under label distribution skew, but it does not work under feature distribution skew. Overall, FedFA promotes a shared feature space among clients and keeps classifiers consistent in this space to overcome data heterogeneity.

## 6 CONCLUSION

This work focuses on alleviating performance degradation caused by label and feature distribution skews in federated learning. Firstly, we observe and analyze the existence of a *vicious cycle* between feature mapping inconsistency and classifier update divergence under data heterogeneity. Secondly, we propose FedFA to create a shared feature space across clients, assisted by feature anchors, and keep the classifier consistent in this space, thus bringing a *virtuous cycle* between feature and classifier updates. Finally, FedFA significantly outperforms baselines on various image classification tasks.

We will further explore whether feature mapping inconsistency exists across clients under other tasks besides classification, and if so, we will extend FedFA to resolve the data heterogeneous issue in these settings in the future work. Furthermore, the performance advantage of FedFA tends to decrease for deeper models, which is probably because only aligning the last feature maps does not effectively align the shallow feature maps. Thus, it is promising to align the features of the shallow layers.

## REFERENCES

- Durmus Alp Emre Acar, Yue Zhao, Ramon Matas, Matthew Mattina, Paul Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=B7v4QMR6Z9w>.
- Maruan Al-Shedivat, Jennifer Gillenwater, Eric Xing, and Afshin Rostamizadeh. Federated learning via posterior averaging: A new perspective and practical algorithms. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=GFsU8a0sGB>.
- Hong-You Chen and Wei-Lun Chao. On bridging generic and personalized federated learning for image classification. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=I1hQbx10Kxn>.
- Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 international joint conference on neural networks*, pp. 2921–2926. IEEE, 2017.
- Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning: A meta-learning approach. *arXiv preprint arXiv:2002.07948*, 2020.
- Ganin, Yaroslav, Lempitsky, and Victor. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pp. 1180–1189. PMLR, 2015.
- Chaoyang He, Alay Dilipbhai Shah, Zhenheng Tang, Di Fan, Adarshan Naiynar Sivashunmugam, Keerti Bhogaraju, Mita Shimpi, Li Shen, Xiaowen Chu, Mahdi Soltanolkotabi, and Salman Avestimehr. Fedcv: a federated learning framework for diverse computer vision tasks. *arXiv preprint arXiv:2111.11066*, 2021.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- Kevin Hsieh, Amar Phanishayee, Onur Mutlu, and Phillip Gibbons. The non-iid data quagmire of decentralized machine learning. In *International Conference on Machine Learning*, pp. 4387–4398. PMLR, 2020.
- Jonathan J. Hull. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5):550–554, 1994.
- Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pp. 448–456. PMLR, 2015.
- Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.
- 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*, pp. 5132–5143. PMLR, 2020.
- Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-iid data silos: An experimental study. *arXiv preprint arXiv:2102.02079*, 2021a.

- Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10713–10722, 2021b.
- 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.
- Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. In *International Conference on Learning Representations*, 2021c. URL <https://openreview.net/pdf?id=6YEQUn0QICG>.
- Xin-Chun Li and De-Chuan Zhan. Fedrs: Federated learning with restricted softmax for label distribution non-iid data. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 995–1005, 2021.
- Zijian Li, Jiawei Shao, Yuyi Mao, Jessie Hui Wang, and Jun Zhang. Federated learning with gan-based data synthesis for non-iid clients. *arXiv preprint arXiv:2206.05507*, 2022.
- Mi Luo, Fei Chen, Dapeng Hu, Yifan Zhang, Jian Liang, and Jiashi Feng. No fear of heterogeneity: Classifier calibration for federated learning with non-iid data. *Advances in Neural Information Processing Systems*, 34:5972–5984, 2021.
- Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.
- Yair Movshovitz-Attias, Alexander Toshev, Thomas K Leung, Sergey Ioffe, and Saurabh Singh. No fuss distance metric learning using proxies. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 360–368, 2017.
- Xutong Mu, Yulong Shen, Ke Cheng, Xueli Geng, Jiaxuan Fu, Tao Zhang, and Zhiwei Zhang. Fed-proc: Prototypical contrastive federated learning on non-iid data. *arXiv preprint arXiv:2109.12273*, 2021.
- Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.
- Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and Hugh Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=LkFG31B13U5>.
- Jiawei Shao, Yuchang Sun, Songze Li, and Jun Zhang. Dres-fl: Dropout-resilient secure federated learning for non-iid clients via secret data sharing. *Advances in Neural Information Processing Systems*, 2022.
- Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, 30, 2017.
- Yuchang Sun, Jiawei Shao, Songze Li, Yuyi Mao, and Jun Zhang. Stochastic coded federated learning with convergence and privacy guarantees. *arXiv preprint arXiv:2201.10092*, 2022.
- Canh T Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems*, 33:21394–21405, 2020.
- Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang. Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2022a.

- Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. Fedproto: Federated prototype learning across heterogeneous clients. In *AAAI Conference on Artificial Intelligence*, volume 1, pp. 3, 2022b.
- Zhenheng Tang, Yonggang Zhang, Shaohuai Shi, Xin He, Bo Han, and Xiaowen Chu. Virtual homogeneity learning: Defending against data heterogeneity in federated learning. In *International Conference on Machine Learning*, pp. 21111–21132, 2022.
- Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.
- Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. In *International Conference on Learning Representations*, 2020a. URL <https://openreview.net/forum?id=BkluqlSFDS>.
- Jianguo Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in Neural Information Processing Systems*, 33:7611–7623, 2020b.
- Jianguo Wang, Zachary Charles, Zheng Xu, Gauri Joshi, H Brendan McMahan, Maruan Al-Shedivat, Galen Andrew, Salman Avestimehr, Katharine Daly, Deepesh Data, et al. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*, 2021.
- Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pp. 499–515. Springer, 2016.
- Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.
- Fuxun Yu, Weishan Zhang, Zhuwei Qin, Zirui Xu, Di Wang, Chenchen Liu, Zhi Tian, and Xiang Chen. Fed2: Feature-aligned federated learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 2066–2074, 2021.
- Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks. In *International Conference on Machine Learning*, pp. 7252–7261. PMLR, 2019.
- Dun Zeng, Siqi Liang, Xiangjing Hu, and Zenglin Xu. Fedlab: A flexible federated learning framework. *arXiv preprint arXiv:2107.11621*, 2021.
- Jie Zhang, Zhiqi Li, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, and Chao Wu. Federated learning with label distribution skew via logits calibration. In *International Conference on Machine Learning*, pp. 26311–26329. PMLR, 2022.
- Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

## APPENDIX

## A TERMINOLOGIES

**Global model vs. local model.** Let us first clarify the concepts of “global” vs. “local” models: in each communication round, local models denote the ones updated by the clients after local training, and the global model denotes the model obtained by aggregating all local models at the server. Moreover, client models denote the models being trained during local training.

**Vicious cycle vs. virtuous cycle.** As shown in Figure 1 (a), the *vicious cycle* represents the phenomenon that inconsistent feature mappings of local models diverge the classifier updates, such that the diverged classifiers of different clients induce feature extractors to map to more inconsistent features across clients. As shown in Figure 1 (b), the *virtuous cycle* represents the phenomenon that consistent feature mappings of client local models make the classifier updates similar, such that the updated classifiers make feature extractors of clients map to more consistent features across clients.

**Positive pair vs. negative pair.** A *positive pair* denotes a pair of **samples** with the same label (i.e., the samples belong to the same class). A *negative pair* denotes a pair of **samples** with different labels (i.e., the samples do not belong to the same class).

**Positive feature vs. negative feature.** For the  $c$ -th proxy  $\phi_c$ , the *positive features* denote the **features** of the  $c$ -th class, and the *negative features* denote the **features** of other classes except for the  $c$ -th class.

**Positive proxy vs. negative proxy.** For the feature of the  $c$ -th class, the *positive proxy* denotes the  $c$ -th **proxy**  $\phi_c$ , and the *negative proxies* denote other **proxies** except for the  $c$ -th proxy  $\phi_c$ .

**Label distribution skew vs. feature distribution skew.** Feature and label distribution skews are two representative data heterogeneity (Kairouz et al., 2021), both covered in this work. Suppose that the  $i$ -th client data distribution follows  $P_i(x, y) = P_i(x|y)P_i(y) = P_i(y|x)P_i(x)$  where  $x$  and  $y$  denote the feature and label respectively. Following (Li et al., 2021a), the definition of feature distribution skew and label distribution skew can be given as:

- **Label distribution skew (prior probability):** The label marginal distribution  $P_i(y)$  varies across clients while  $P_i(x|y) = P_j(x|y)$  for all clients  $i$  and  $j$ .
- **Feature distribution skew (covariate shift):** The input feature marginal distribution  $P_i(x)$  varies across clients while  $P_i(y|x) = P_j(y|x)$  for all clients  $i$  and  $j$ .
- **Label & feature distribution skew:** At least one of the label distribution skew and feature distribution skew happens across clients. This means clients  $i$  and  $j$  still suffer from label marginal distribution skew  $P_i(y)$  even if sharing the same  $P_i(x)$ , or clients  $i$  and  $j$  still suffer from input feature marginal distribution skew  $P_i(x)$  even if sharing the same  $P_i(y)$ .
- **Local data distributions without skew:** Herein, for FMNIST, CIFAR-10, and CIFAR-100, we split them evenly into client-side local datasets based on an identical label distribution. For Mixed Digits, we first mix all the digit datasets as a global dataset and evenly split it into client-side local datasets based on an identical label distribution. Note that this case can not guarantee that local distributions share the same global distribution across clients. Still, it means local distributions are more homogeneous than the cases of label or feature distribution skew.

**Center loss vs. feature anchor loss.** Our feature anchor loss borrows the idea of the center loss (Wen et al., 2016), but the purposes of these two losses are different. Besides with a different updating method, the center loss aims to decrease the feature distance for intra-class samples in centralized training, rather than keeping feature mapping consistent across clients. Meanwhile, the feature anchors are not utilized to calibrate the classifier in (Wen et al., 2016), but our work does it to prevent the divergence of classifiers across clients in federated training. Therefore, we use the feature anchor loss to distinguish the center loss proposed by the research community on face recognition.

## B RELATED WORKS

Federated learning is a fast-developing area, and we mainly introduce the methods close to ours (i.e., federated optimization-based methods) and briefly introduce other methods. Comprehensive field studies have appeared in (Kairouz et al., 2021; Wang et al., 2021; Tan et al., 2022a).

**Tackle data heterogeneity on the client side.** To avoid local models converging to their local minima instead of global minima, many works add a well-designed regularization term to penalize local models to make them not far away from the global model. For example, FedProx (Li et al., 2020) uses the Euclidean distance between local models and the global model as the regularization loss. FedDyn (Acar et al., 2021) modifies the local objective with a dynamic regularizer consisting of a linear term based on the first order condition and the Euclidean-distance term, such that the local minima are consistent with the global stationary point. MOON (Li et al., 2021b) utilizes the feature similarity between previous local models and the global model as model-contrastive regularization to correct the local training of each client. In place of the model-contrastive term in MOON, FedProc (Mu et al., 2021) introduces a prototype-contrastive term to regularize the features within each class with class prototypes (Snell et al., 2017). Besides, instead of implicit correction by regularization, a number of works reduce the bias explicitly in local updates by controlling variates or posterior sampling. Borrowing the variance-reduce technique in standard convex optimization, SCAFFOLD (Karimireddy et al., 2020) presents a control variate to correct the client updates, so they are much closer to the global update. Another way to reduce the bias is to run Markov Carlo, instead of stochastic gradient descent, which produces approximate local posterior samples like FedPA (Al-Shedivat et al., 2021). Similar to (Li et al., 2021a; Chen & Chao, 2022; Luo et al., 2021; He et al., 2021), our experiments in Section 5 show that these works may not provide stable better performance gains over FedAvg (McMahan et al., 2017) in classification tasks, which motivates us to analyze the relationship between classifier updates and feature mappings in local training.

**Tackle data heterogeneity on the server side.** In addition to improving on the client side, many works have developed alternative aggregation schemes on the server side to tackle data heterogeneity. For instance, (Wang et al., 2020b) finds an objective inconsistency problem caused by unbalanced data that induces a different number of local updates and propose FedNova to eliminate the inconsistency by normalizing the local updates before averaging. Besides, (Reddi et al., 2021) adopts adaptive momentum update on the server-side to mitigate oscillation of global model updates when the server activates the clients with a limited subset of labels. Beyond layer-weighted averaging, some works like FedMA (Wang et al., 2020a) and Fed<sup>2</sup> (Yu et al., 2021) introduce neuron-wise averaging because there may exist neuron mismatching from permutation invariance of neural networks in federated learning. These ideas complement our work and can be integrated into our method because our method only adds a regularizer on the client side.

**Tackle data heterogeneity based on feature or classifier.** Instead of considering from the federated-optimization view, some recent works such as (Li et al., 2021b; Mu et al., 2021; Li & Zhan, 2021; Luo et al., 2021; Zhang et al., 2022; Tang et al., 2022) pay more attention to feature space across clients. To improve feature consistency, MOON (Li et al., 2021b) and FedProc (Mu et al., 2021) introduce a feature-based local regularizer mentioned above. Meanwhile, (Tang et al., 2022) generates a shared virtual dataset for all clients before training, and calibrates features by minimizing the feature distribution distance between the virtual dataset and the real dataset. To improve classifier consistency, (Luo et al., 2021) observes that the classifier layer (i.e., the last layer of the model) suffers most from label distribution skew and proposes calibration of the classifier with virtual features after training. Moreover, (Li & Zhan, 2021; Zhang et al., 2022) introduce a restricted loss cross-entropy and a fine-grained calibrated cross-entropy loss, respectively. The key idea of the two methods is to prevent the overfitting of missing classes (Li & Zhan, 2021) and minority classes (Zhang et al., 2022) (i.e., both under the label distribution skew) with an improved cross-entropy loss. However, compared with our method, these methods only consider the label distribution skew setting by improving the performance degeneration based on feature calibration or classifier calibration, and neglect *vicious cycle* between different classifier updates and inconsistent features, which hurts their performance.

**Other methods.** The data-centric method is one recent new direction, which shares common datasets with all clients like the public dataset in (Zhao et al., 2018). To avoid violating the privacy requirement, some works focus on sharing synthesized data like (Luo et al., 2021; Li et al., 2022; Tang et al., 2022) and coded data (Sun et al., 2022; Shao et al., 2022) with privacy protection to

construct a more homogeneous dataset for federated learning. Moreover, another line of research aims to train a personalized model for each client, rather than a global model (Tan et al., 2022a). Since there is still no standard approach to personalized federated learning, many researchers achieve it by personalized regularization (T Dinh et al., 2020), meta learning (Fallah et al., 2020), prototype learning (Tan et al., 2022b) and personalized layers (Chen & Chao, 2022), etc.

Our work aims at the typical federated learning (McMahan et al., 2017) and tries to improve the local optimization by feature alignment in federated optimization. There are two existing works similar to ours, i.e., MOON (Li et al., 2021b) which introduces a model-contrastive loss to maximize the agreement of the features extracted by the local model and that by the global model, and FedProc (Mu et al., 2021) which proposes a prototype-contrastive loss to correct features by class prototypes. However, compared with our method, although considering the feature mapping inconsistency across local models, MOON and FedProc neglect the *vicious cycle* between different classifiers' updates and inconsistent features, which hurts their performance.

## C PROOF OF PROPERTY 1

**Proof 1 (Property 1)** Let  $A = \sum_{j_i=1, y_{j_i}=c}^{n_i} (1 - p_{j_i,c}^{(i)}) \mathbf{h}_{i,y_{j_i}} - \sum_{j_i=1, y_{j_i} \neq c}^{n_i} p_{j_i,c}^{(i)} \mathbf{h}_{i,y_{j_i}}$ , and  $B = \sum_{j_v=1, y_{j_v}=c}^{n_v} (1 - p_{j_v,c}^{(v)}) \mathbf{h}_{v,y_{j_v}} - \sum_{j_v=1, y_{j_v} \neq c}^{n_v} p_{j_v,c}^{(v)} \mathbf{h}_{v,y_{j_v}}$ .

We compute the similarity of classifier updates between client  $i$  and client  $v$ ,

$$\cos(\Delta\phi_{i,c}, \Delta\phi_{v,c}) = \frac{\langle \Delta\phi_{i,c}, \Delta\phi_{v,c} \rangle}{\|\Delta\phi_{i,c}\| \|\Delta\phi_{v,c}\|} = \frac{\eta^2 \langle A, B \rangle}{n_i n_v \|A\| \|B\|}. \quad (10)$$

Let  $A_1 = \sum_{j_i=1, y_{j_i}=c}^{n_i} (1 - p_{j_i,c}^{(i)}) \mathbf{h}_{i,y_{j_i}}$ ,  $A_2 = \sum_{j_i=1, y_{j_i} \neq c}^{n_i} p_{j_i,c}^{(i)} \mathbf{h}_{i,y_{j_i}}$ ,  $B_1 = \sum_{j_v=1, y_{j_v}=c}^{n_v} (1 - p_{j_v,c}^{(v)}) \mathbf{h}_{v,y_{j_v}}$  and  $B_2 = \sum_{j_v=1, y_{j_v} \neq c}^{n_v} p_{j_v,c}^{(v)} \mathbf{h}_{v,y_{j_v}}$ .

Herein, we represent  $\langle A, B \rangle$  as:

$$\begin{aligned} \langle A, B \rangle &= \langle A_1 - A_2, B_1 - B_2 \rangle \\ &= \langle A_1, B_1 \rangle + \langle A_2, B_2 \rangle - \langle A_1, B_2 \rangle - \langle A_2, B_1 \rangle. \end{aligned} \quad (11)$$

According to Assumption 1, when  $\mathbf{h}_c \in \mathcal{H}_c$  and  $\mathbf{h}_q \in \mathcal{H}_q$  and  $c \neq q$ , the inner product  $\langle \mathbf{h}_c, \mathbf{h}_q \rangle$  is less than or equal to 0. Thus, we have:

$$\begin{aligned} \langle A_1, B_2 \rangle &= \left\langle \sum_{j_i=1, y_{j_i}=c}^{n_i} (1 - p_{j_i,c}^{(i)}) \mathbf{h}_{i,y_{j_i}}, \sum_{j_v=1, y_{j_v} \neq c}^{n_v} p_{j_v,c}^{(v)} \mathbf{h}_{v,y_{j_v}} \right\rangle \\ &= \sum_{j_v=1, y_{j_v} \neq c}^{n_v} \sum_{j_i=1, y_{j_i}=c}^{n_i} (1 - p_{j_i,c}^{(i)}) p_{j_v,c}^{(v)} \langle \mathbf{h}_{i,y_{j_i}}, \mathbf{h}_{v,y_{j_v}} \rangle \\ &\leq 0 \end{aligned} \quad (12)$$

where the inequality holds because  $y_{j_i} = c$  but  $y_{j_v} \neq c$  and thus  $\langle \mathbf{h}_{i,y_{j_i}}, \mathbf{h}_{v,y_{j_v}} \rangle \leq 0$ .

Similarly, we have:

$$\begin{aligned} \langle A_2, B_1 \rangle &= \left\langle \sum_{j_i=1, y_{j_i} \neq c}^{n_i} p_{j_i,c}^{(i)} \mathbf{h}_{i,y_{j_i}}, \sum_{j_v=1, y_{j_v}=c}^{n_v} (1 - p_{j_v,c}^{(v)}) \mathbf{h}_{v,y_{j_v}} \right\rangle \\ &\leq 0. \end{aligned} \quad (13)$$

Also, according to Assumption 1, when  $\|\mathbf{a}_c\| > \sqrt{2}\delta$ , for any  $\mathbf{h}_{i,c} \in \mathcal{H}_c$  and  $\mathbf{h}_{v,c} \in \mathcal{H}_c$ , we obtain that the inner product  $\langle \mathbf{h}_{i,c}, \mathbf{h}_{v,c} \rangle$  is larger than 0 since the arccosine of largest angle between

$\mathbf{h}_{i,c}$  and  $\mathbf{h}_{v,c}$  is  $2\arccos(\|\mathbf{a}_c\|/\sqrt{2}\delta)$  (i.e., the largest angle is smaller than  $\pi/2$ ).

$$\begin{aligned}
\langle A_1, B_1 \rangle &= \langle \sum_{j_i=1, y_{j_i}=c}^{n_i} (1 - p_{j_i,c}^{(i)}) \mathbf{h}_{i,y_{j_i}}, \sum_{j_v=1, y_{j_v}=c}^{n_v} (1 - p_{j_v,c}^{(v)}) \mathbf{h}_{v,y_{j_v}} \rangle \\
&= \sum_{j_v=1, y_{j_v}=c}^{n_v} \sum_{j_i=1, y_{j_i}=c}^{n_i} (1 - p_{j_i,c}^{(i)}) (1 - p_{j_v,c}^{(v)}) \langle \mathbf{h}_{i,y_{j_i}}, \mathbf{h}_{v,y_{j_v}} \rangle \\
&> 0.
\end{aligned} \tag{14}$$

Similarly, we have:

$$\begin{aligned}
\langle A_2, B_2 \rangle &= \langle \sum_{j_i=1, y_{j_i} \neq c}^{n_i} p_{j_i,c}^{(i)} \mathbf{h}_{i,y_{j_i}}, \sum_{j_v=1, y_{j_v} \neq c}^{n_v} p_{j_v,c}^{(v)} \mathbf{h}_{v,y_{j_v}} \rangle \\
&> 0.
\end{aligned} \tag{15}$$

Combining (11) to (15), we obtain that  $\langle A, B \rangle$  is larger than 0, and thus  $\cos(\Delta\phi_{i,c}, \Delta\phi_{v,c}) > 0$ .

## D PSEUDOCODE OF FEDFA

Herein, we get the inspiration from Batch Normalization (Ioffe & Szegedy, 2015) and perform moving average to estimate  $\bar{\mathbf{h}}_{c,i}^{(t,k)}$  based on the  $c$ -th class samples  $\mathcal{B}_{i,c}$  of one mini batch  $\mathcal{B}_i$  when performing prediction in (8), as presented as following:

$$\bar{\mathbf{h}}_{c,i}^{(t,k)} = \frac{1}{2} \sum_{\tau}^B (\bar{\mathbf{h}}_{c,i}^{(t,k_{\tau-1})} + \bar{\mathbf{h}}_{c,i}^{(t,k_{\tau})})$$

where  $\bar{\mathbf{h}}_{c,i}^{(t,k_{\tau})} = \frac{1}{|\mathcal{B}_{i,c}^{(t,k_{\tau})}|} \sum_{(\mathbf{x},c) \in \mathcal{B}_{i,c}^{(t,k_{\tau})}} f_{\theta_i}(\mathbf{x})$ ,  $\bar{\mathbf{h}}_{c,i}^{(t,k_0)} = \bar{\mathbf{h}}_{c,i}^{(t,k-1)}$ , and  $\tau \in [1, B]$  denotes the  $\tau$ -th mini batch of the total number  $B$  at the  $k$ -th epoch.

The pseudocode of FedFA is shown as the following Algorithm 1. Compared with FedAvg, FedFA adds a feature anchor loss and calibrates the classifier locally.

## E DETAILS OF EXPERIMENT SETUP

### E.1 SPECIFIC MODELS

Our validation and test experiments, including label distribution skew, feature distribution skew and label & feature distribution skews, use the models according to Table 4. Herein, to ablate the effect of BN layers, we follow (Hsieh et al., 2020) to replace the BN layer with the GroupNorm layer in all experiments except for the test under Mixed-digit datasets. For a fair comparison, our models follow those reported in the baselines' works. Specifically, following (Acar et al., 2021), we use a CNN model for EMNIST, FMNIST, and CIFAR-10, consisting of two 5x5 convolution layers followed by 2x2 max pooling and two fully-connected layers with ReLU activation. Following (Li et al., 2021b) and (Li et al., 2021c), we utilize the ResNet-18 (He et al., 2016) with a linear projector for CIFAR-100 and a CNN model with three 5x5 convolution layers followed by five batch normalization (BN) layers for the Mixed Digits dataset.

### E.2 VALIDATION EXPERIMENT SETUP

The total number of training samples per client is 1000 in this case. We separately sample a subset from test sets of FMNIST and Mixed Digit to visualize the normalized feature mappings of the local models based on t-SNE visualization (Van der Maaten & Hinton, 2008). In Figure 2, although we input the same Validation samples into all clients' local modes, we only show their features mappings for which clients have the corresponding class (i.e., if client 1 only holds class 1 and class 2 samples,



**Algorithm 1** FedFA (Proposed Framework): Federated Learning with Feature Alignment

---

**Input:** initial model  $\mathbf{w} = \{\theta, \phi\}$ , initial feature anchors  $\{\mathbf{a}_c\}_{c=1}^C$ , learning rate  $\eta$ , local epoch  $K$ , client number  $N$ , class number  $C$

**for** each round  $t = 1, \dots, R$  **do**

  Server samples clients  $\mathcal{S} \subseteq \{1, \dots, N\}$

  Server communicates  $\mathbf{w}^{(t-1)}$  and  $\{\mathbf{a}_c^{(t-1)}\}_{c=1}^C$  to all clients  $i \in \mathcal{S}$

**on client**  $i \in \mathcal{S}$  **in parallel do**

    Initialize the local model  $\mathbf{w}_i \leftarrow \mathbf{w}^{(t-1)}$ , the local feature anchor  $\mathbf{a}_{c,i} \leftarrow \mathbf{a}_c^{(t-1)}$

**for** local epoch  $k = 1, \dots, K$  **do**

      Calculate the local loss  $l_i \leftarrow l_{\text{sup}_i} + \mu l_{\text{fa}_i}$  according to (6)

      Compute mini-batch gradient  $g_i(\mathbf{w}_i) \leftarrow \nabla_{\mathbf{w}_i} l_i$

      Update local model  $\mathbf{w}_i \leftarrow \mathbf{w}_i - \eta g_i(\mathbf{w}_i)$

      Calibrate classifiers with feature anchors according to (7)

      Accumulate dynamic momentum  $\mathbf{m}_i^{(t,k)}$  according to (8)

**end for**

    Update local feature anchors  $\mathbf{a}_{c,i}^{(t)} \leftarrow \sum_{k=1}^K \mathbf{m}_{c,i}^{(t,k)}$

    Communicate  $\mathbf{w}_i^{(t)}$  and  $\{\mathbf{a}_{c,i}^{(t)}\}_{c=1}^C$  back to the server

**end on client**

  Server aggregates the global model  $\mathbf{w}^{(t)} \leftarrow \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \mathbf{w}_i^{(t)}$ , and the feature anchors  $\mathbf{a}_c^{(t)} \leftarrow \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \mathbf{a}_{c,i}^{(t)}$

**end for**

---

Table 4: The specific parameters settings for all the models used in our experiments.

Layer	Validation Experiment		Test Experiment			
	Label Skew	Feature Skew	Label Skew		Feature Skew	
	FMNIST	Mixed-digit dataset	FMNIST/EMNIST	CIFAR-10	CIFAR-100	Mixed-digit dataset
1	Conv2d(1, 32, 5) ReLU, MaxPool2D(2,2)	Conv2d(3, 64, 5) ReLU, MaxPool2D(2,2)	Conv2d(1, 32, 5) ReLU, MaxPool2D(2,2)	Conv2d(3, 64, 5) ReLU, MaxPool2D(2,2)	Basicbone of Resnet18 with GroupNorm	Conv2d(3, 64, 5, 1, 2) BN(64), ReLU, MaxPool2D(2,2)
2	Conv2d(32, 32, 5) ReLU, MaxPool2D(2,2)	Conv2d(64, 64, 5) ReLU, MaxPool2D(2,2)	Conv2d(32, 32, 5) ReLU, MaxPool2D(2,2)	Conv2d(64, 64, 5) ReLU, MaxPool2D(2,2)	FC(512, 512) ReLU	Conv2d(64, 64, 5, 1, 2) BN(64), ReLU, MaxPool2D(2,2)
3	FC(992, 384) ReLU	FC(1024, 384) ReLU	FC(992, 384) ReLU	FC(1600, 384) ReLU	FC(512, 256)	Conv2d(3, 128, 5, 1, 2) BN(128), ReLU
4	FC(384, 100)	FC(384, 100)	FC(384, 192) ReLU	FC(384, 192) ReLU	FC(256, 100)	FC(6272, 2048) BN(2048), ReLU
5	FC(100, 10)	FC(100, 10)	FC(192, 10)	FC(192, 10)		FC(2048, 512) BN(512), ReLU
6						FC(512, 10)
Source			model from (Acar et al., 2021)	model from (Acar et al., 2021)	model from (Li et al., 2021b)	model from (Li et al., 2021c)

we only offer the feature maps of the client 1 model for these two classes, as it would be unfair to ask the local model of client 1 to map the feature of classes on which it did not learn.). We visualize the feature mappings of client models according to the labels (digit dataset) owned by the corresponding client for label (distribution) distribution skew. The specific setup is described as:

- **Label Distribution Skew:** The experiment has 10 clients where each client has 2 classes with 500 samples per class from FMNIST, and utilizes the SGD optimizer with a 0.01 learning rate and without momentum. The federated setting involves 10 local epoch numbers, 15 communication rounds, and a 100% client sample rate. The top-1 accuracy of global model of all method at the targeted communication round is that FedAvg without skew: 80.32%; FedAvg:52.66%; FedProx:51.43%; FedDyn:51.90%; MOON:45.67%; FedProc:49.87%; FedFA(our): 67.54%.
- **Feature Distribution Skew:** The experiment has 10 clients where each client has 10 classes with 100 samples per class from one of the digit datasets in Mixed Digit (i.e., MNIST, SVHN, USPS, SynthDigits, and MNIST-M), and utilizes the SGD optimizer with a 0.01 learning rate and without momentum. The federated setting involves 10 local epoch numbers, 15 communication rounds, and a 100% client sample rate. The Mean top-1 accuracy of the global models of all methods at the targeted communication round is that FedAvg without

skew: 81.66%; FedAvg:79.56%; FedProx:78.76%; FedDyn:79.60%; MOON: 79.58%; FedProc:79.30%; FedFA(our): 80.44%.

### E.3 TEST EXPERIMENT SETUP

**Baselines.** Federated learning (McMahan et al., 2017) aims to train a global model parameterized by  $\mathbf{w}$  by collaborating a total of  $N$  clients with a central server to solve the following optimization problem:

$$\min_{\mathbf{w} \in \mathbb{R}^d} \mathcal{L}(\mathbf{w}) := \mathbb{E}_i[\mathcal{L}_i(\mathbf{w})] = \sum_i^N \frac{n_i}{n} \mathcal{L}_i(\mathbf{w})$$

where  $n = \sum_i n_i$  represents the total sample size with  $n_i$  being the sample size of the  $i$ -th client, and  $\mathcal{L}_i(\mathbf{w}) := \mathbb{E}_{\xi \in \mathcal{D}_i} [l_i(\mathbf{w}; \xi)]$  is the local objective function in local dataset  $\mathcal{D}_i$  of the  $i$ -th client.

Many methods have been proposed to solve this optimization problem and alleviate the negative impact of data heterogeneity across clients. Herein, from the view of local-optimization methods, we compare FedFA with the common federated learning algorithms, including FedAvg (McMahan et al., 2017), FedProx (Li et al., 2020) and the state-of-the-art methods based on well-designed local regularization including FedDyn (Acar et al., 2021), MOON (Li et al., 2021b) and FedProc (Mu et al., 2021). The specific description of these methods can be denoted as:

- **FedAvg:** As a canonical method to solve (1) proposed by (McMahan et al., 2017), in each communication round, FedAvg firstly selects a subset of clients and initiates client models as  $\mathbf{w}$  and then updates the local models  $\mathbf{w}_i$  by minimizing  $\mathcal{L}_i(\mathbf{w})$ , and finally aggregates the local models  $\mathbf{w}_i$  as the new global model  $\mathbf{w}$  until  $\mathcal{L}(\mathbf{w})$  arrives at a stationary point.
- **FedProx:** FedProx (Li et al., 2020) adds the Euclidean regularization loss between local models and the global model in the local optimization problem, which can be described as:

$$\mathcal{L}_i(\mathbf{w}) = \min_{\mathbf{w}_i} \mathbb{E}_{(\mathbf{x}, y) \in \mathcal{D}_i} [l_i(\mathbf{w}_i; \mathbf{w}^{(t-1)}) + \frac{\mu}{2} \|\mathbf{w}_i - \mathbf{w}^{(t-1)}\|^2]. \quad (16)$$

- **FedDyn:** FedDyn (Acar et al., 2021) modifies the local objective with a dynamic regularization consisting of a linear term based on the first order condition and an above Euclidean-distance term, such that the local minima are consistent with the global stationary point, which can be described as:

$$\mathcal{L}_i(\mathbf{w}) = \min_{\mathbf{w}_i} \mathbb{E}_{(\mathbf{x}, y) \in \mathcal{D}_i} [l_i(\mathbf{w}_i; \mathbf{w}^{(t-1)}) - \langle \nabla \mathcal{L}_i(\mathbf{w}^{(t-1)}), \mathbf{w}_i \rangle + \frac{\mu}{2} \|\mathbf{w}_i - \mathbf{w}^{(t-1)}\|^2]. \quad (17)$$

- **MOON:** MOON (Li et al., 2021b) utilizes the feature similarity of the client model with previous-round local models and with the global model as model-contrastive regularization to correct the local training of each client, which can be described as:

$$\mathcal{L}_i(\mathbf{w}) = \min_{\mathbf{w}_i} \mathbb{E}_{(\mathbf{x}, y) \in \mathcal{D}_i} [l_i(\mathbf{w}_i; \mathbf{w}^{(t-1)}) - \mu \log \frac{\exp(\text{sim}(\mathbf{h}_i, \mathbf{h}_{\text{global}})/\tau)}{\exp(\text{sim}(\mathbf{h}_i, \mathbf{h}_{\text{global}})/\tau) + \exp(\text{sim}(\mathbf{h}_i, \mathbf{h}_{\text{pre}})/\tau)}] \quad (18)$$

where  $\mathbf{h}_i, \mathbf{h}_{\text{global}}, \mathbf{h}_{\text{pre}}$  denote the feature mappings of the local model  $\mathbf{w}_i$ , the global model  $\mathbf{w}$ , and the local model at previous round  $\mathbf{w}_i^{t-1}$  given the same input  $\mathbf{x}$ , respectively;  $\tau$  is the hyperparameter to control the effect of cosine similarity in model-contrastive loss.

- **FedProc:** Instead of the model-contrastive term in MOON, FedProc (Mu et al., 2021) introduces a prototype-contrastive term to regularize the features within each class with class prototypes (Snell et al., 2017), which can be described as:

$$\mathcal{L}_i(\mathbf{w}) = \min_{\mathbf{w}_i} \mathbb{E}_{(\mathbf{x}, y) \in \mathcal{D}_i} [\frac{t}{T} l_i(\mathbf{w}_i; \mathbf{w}^{(t-1)}) + (1 - \frac{t}{T}) \log \frac{\exp(\text{sim}(\mathbf{h}_i, \mathbf{p}_c)/\tau)}{\sum_{c=1}^{c=C} \exp(\text{sim}(\mathbf{h}_i, \mathbf{p}_c)/\tau)}] \quad (19)$$

where  $T$  is the targeted communication round, and  $\mathbf{p}_c$  is the prototype of class  $c$ . In FedProc,  $\mathbf{p}_c$  is updated by the whole local dataset at the end of one communication round (i.e.,  $\mathbf{p}_{c,i}^{(t,k)} = \frac{1}{|\mathcal{D}_{i,c}|} \sum_{(\mathbf{x}, c) \in \mathcal{D}_{i,c}} \mathbf{h}_{i,c}$ ). However, we need to denote that if  $\mathbf{p}_c$  is updated like this, rather than the momentum update as (8), and we found that FedProc would suffer from

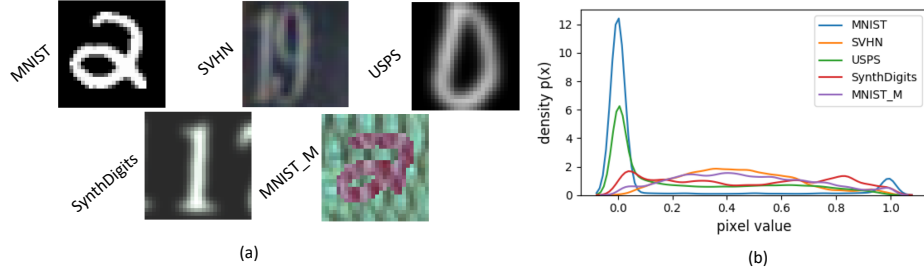


Figure 4: Data visualization. (a) Examples from each dataset (client) in Mixed Digit. (b) feature distributions skew across the datasets (over random 100 samples for each dataset).

the divergence because the update of  $\mathbf{p}_c$  is too drastic in our experiments<sup>2</sup>. Therefore, we improve FedProc with momentum update (8).

**Datasets.** This work aims at image classification tasks under label distribution skew, feature distribution skew and label & feature distribution skew, and uses benchmark datasets with the same data heterogeneity setting as (McMahan et al., 2017; Yurochkin et al., 2019; Li et al., 2021a), including EMNIST (Cohen et al., 2017), FMNIST (Xiao et al., 2017), CIFAR-10, CIFAR-100 (Krizhevsky et al., 2009), and Mixed Digits dataset (Li et al., 2021c). Specifically, for label distribution skew, we consider two settings:

- **Same size of local dataset:** Following (McMahan et al., 2017), we split data samples based on classes to clients (e.g.,  $\#C = 2$  denotes each client holds two class samples), where each client holds 250 samples per class;
- **Different sizes of local dataset:** Following (Yurochkin et al., 2019), we first sample  $p_i$  from Dirichlet distribution  $Dir(\alpha)$  and then assign  $p_{i,c}$  proportion of the samples of class  $c$  to client  $i$ , where we set  $\alpha$  as 0.1 and 0.5 to measure the level of data heterogeneity in our experiments. Moreover, when  $\alpha = 0.1$ , the label distributions across clients are so skewed that the quantity of clients’ local dataset is also skewed. That is, the experiment cases related to  $\alpha = 0.1$  would involve label distribution skew and quantity distribution skew, which denotes the unbalanced data size of the local dataset across clients.

For feature distribution skew, we consider two settings:

- **Real-world feature imbalance:** We use a subset of the real-world dataset with natural feature imbalance, EMNIST (Cohen et al., 2017), including 10 classes and 341873 samples (about 34000 samples per class) totally;
- **Artificial feature imbalance:** We use a mixed-digit dataset from (Li et al., 2021c) consisting of five benchmark digit datasets: MNIST (LeCun et al., 1998), SVHN (Netzer et al., 2011), USPS (Hull, 1994), SynthDigits and MNIST-M (Ganin et al., 2015), including 7430 samples for one digit dataset and 743 sample per class. The data visualization is shown as Figure 4<sup>3</sup>.

Note that for the experiments on Mixed Digits, we report the average top-1 accuracy on five benchmark digit datasets in Table 2 and Table 3, and show the top-1 accuracy on each digit dataset during the training in Figure 13 to Figure 19. For the experiments on other datasets except for Mixed Digits, we test the top-1 accuracy on all datasets based on the global model and report them during the training as shown in Figure 9 to Figure 12.

**Federated Simulation Setup.** All experiments are performed based on PyTorch Paszke et al. (2019) and one node of the High-Performance Computing platform with 4 NVIDIA A30 Tensor Core GPUs with 24GB. We use an existing dataset-split tool FedLab (Zeng et al., 2021) to generate federated local datasets for all clients. There are in total 100 clients, and 10 clients participating in

<sup>2</sup>The codes of FedProc are not open source, and thus our reproduction settings cannot be completely consistent to the original setting, but we fine-tune the hyperparameter of FedProc carefully and report the best results.

<sup>3</sup>Figure comes from (Li et al., 2021c)

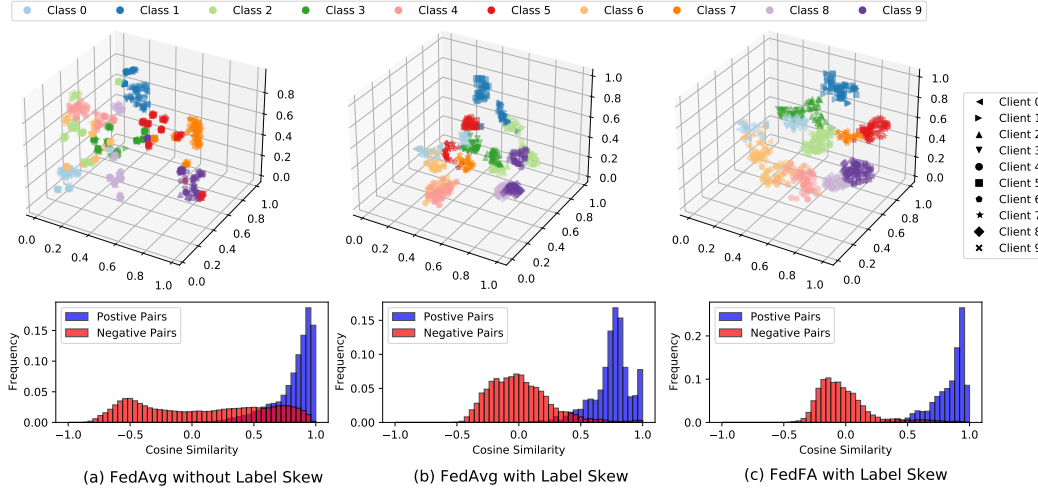


Figure 5: The t-SNE visualization and the histogram of cosine similarity of feature mappings for FedAvg under data homogeneity and for FedAvg and FedFA under label distribution skew with 10 clients.

federated training at each communication round. We use the SGD optimizer with a 0.01 learning rate and 0.001 weight decay for all experiments except for the CIFAR-100 experiment, which uses 0.9 momentum additionally. The local batch size is 64, the number of local epochs is set to 5, and the number of communication rounds is set to 200. Moreover, we carefully select the coefficient of local regularization from  $\{1, 0.1, 0.01\}$  (i.e.,  $\mu/2 = 0.05$  for FedProx and FedDyn,  $\mu = 0.1$  for MOON), set the temperature hyperparameter  $\tau = 0.5$  for MOON and FedProc, and report their best results in our experiments.

**FedFA setup.** We set the anchor momentum coefficient  $\lambda = 0.5$  in (8) and local loss coefficient  $\mu = 0.1$  in (6) like our baselines, and according to Property 1, we initiate the pairwise orthogonal feature anchors  $\mathbf{a}_c$  by sampling column vector from an identity matrix whose dimension is the same as the size of the feature mappings. Other settings of FedFA are the same as baselines in all experiments, such as the same random seed (seed: 2021, 2022, 2023) and the same training and test dataset.

## F ADDITIONAL EXPERIMENT RESULTS

### F.1 ADDITIONAL VALIDATION EXPERIMENT RESULTS

#### F.1.1 FEATURE VISUALIZATION AND SIMILARITY HISTOGRAM FOR ALL METHODS UNDER LABEL DISTRIBUTION SKEW

Figures 5 and 6 show the t-SNE visualization and the histogram of cosine similarity of feature mappings for label distribution skew for all methods. We observe that all baselines under label skew exist feature mapping inconsistency across clients. Still, our method FedFA alleviates it significantly, such as class 1 (i.e., dark blue), class 5 (i.e., dark red) and class 9 (i.e., dark purple) in Figures 5 and 6. Besides, similar to the analysis of Figure 2, the histograms also show that label distribution skew could induce the lower similarity for *positive pairs*, which means feature inconsistency. Moreover, there exists a low frequency of *positive pairs* and a small gap between *positive pairs* and *negative pairs*, which indicates inconsistent polymerization and discrimination (i.e., sizeable intra-class feature distance and small inter-class feature distance) across clients in classification tasks. These results of label distribution skew reveal that all client models are trained in inconsistent feature spaces by our baselines, which hurts their performance.

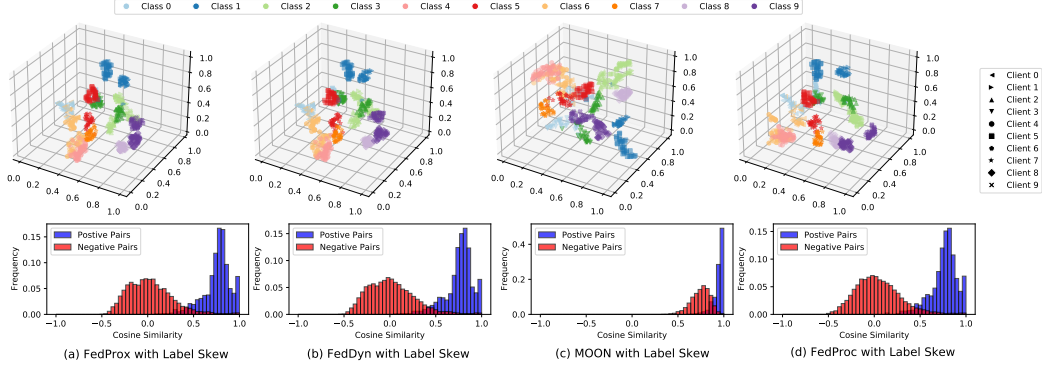


Figure 6: The t-SNE visualization and the histogram of cosine similarity of feature mappings for all baselines under label distribution skew with 10 clients.

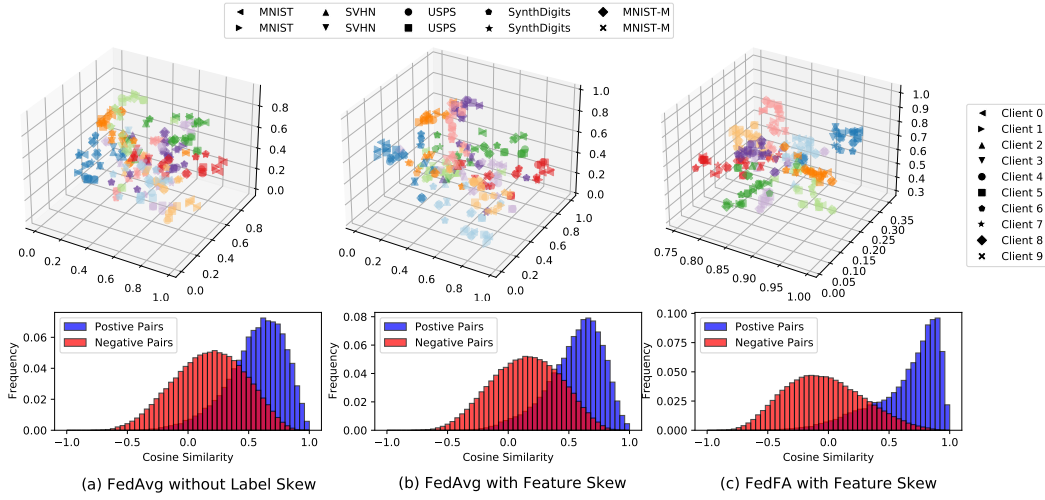


Figure 7: The t-SNE visualization and the histogram of cosine similarity of feature mappings for FedAvg under data homogeneity and for FedAvg and FedFA under feature distribution skew with 10 clients.

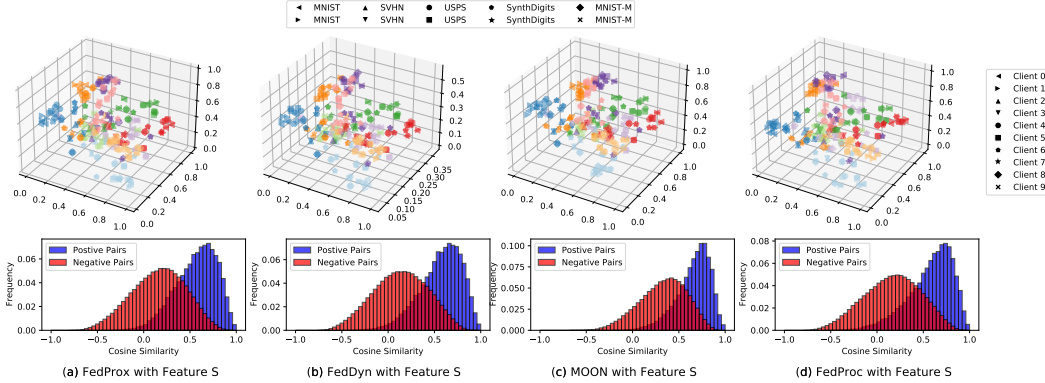


Figure 8: The t-SNE visualization and the histogram of cosine similarity of feature mappings for all baselines under feature distribution skew with 10 clients.

### F.1.2 FEATURE VISUALIZATION AND SIMILARITY HISTOGRAM FOR ALL METHODS UNDER FEATURE DISTRIBUTION SKEW

Similar to label distribution skew, Figures 7 and 8 show the t-SNE visualization and the histogram of cosine similarity of feature mappings for feature distribution skew for all methods. We also observe that all baselines under feature skew still suffer from feature mapping inconsistency across clients, but our method does not. Moreover, without the feature alignment, all baselines present the weak feature polymerization and feature discrimination of clients' local models, which would make the classifier updates divergent as denoted in (3). Therefore, these results of feature distribution skew reveal that all client models are trained in inconsistent feature spaces by our baselines, which hurts their performance.

## F.2 ADDITIONAL TEST EXPERIMENT RESULTS

### F.2.1 COMMUNICATION EFFECTIVENESS

To compare the communication effectiveness, we perform the experiments under label and feature skew to test the communication round number when the global model trained by all methods can reach the target accuracy. As shown in Tables 5 and 6, FedFA realizes better communication efficiency than our baselines with data heterogeneity or not on all datasets except for CIFAR-100. Meanwhile, the training accuracy shown in the following figures also presents similar results.

However, FedFA achieves better generalization than baselines on CIFAR-100 with ResNet but takes more communication rounds to converge. Moreover, as the number of model layers increases, the performance advantages of FedFA diminish. This is probably because the feature anchor loss only regularizes the feature maps of the penultimate layer and is less effective on the shallow layers, which we will explore in future work.

Table 5: The top-1 accuracy (round number when the accuracy reaches the target accuracy) the of all methods under feature distribution skew on the test dataset.

Method (lr = 0.01)	Label Distribution Skew								
	FMNIST $\alpha = 0.1$			CIFAR-10 $\alpha = 0.1$			CIFAR-100 $\alpha = 0.1$		
	#C = 2	$\alpha = 0.1$	$\alpha = 0.5$	#C = 2	$\alpha = 0.1$	$\alpha = 0.5$	#C = 20	$\alpha = 0.1$	$\alpha = 0.5$
Targeted Accuracy	59.682666	55.85066	66.24000	28.85600	28.1600	38.92533	18.09866	17.4346	21.22399
FedAvg w/o skew	62.99(4)	57.31(3)	67.38(6)	29.56(9)	28.34(8)	39.24(23)	18.52(21)	17.69(20)	21.25(28)
FedFA w/o skew	<b>62.62(2)</b>	<b>58.89(1)</b>	<b>67.40(3)</b>	<b>30.06(3)</b>	30.06(3)	40.57(8)	<b>18.17(35)</b>	18.17(35)	<b>21.48(45)</b>
FedAvg	60.22(25)	57.43(34)	66.47(12)	29.64(38)	28.19(43)	39.04(47)	18.34(52)	17.56(73)	21.39(48)
FedProx	59.88(25)	58.55(34)	66.42(13)	29.90(38)	28.62(42)	39.21(42)	18.13(51)	17.76(72)	21.41(52)
FedDyn	62.45(28)	61.99(34)	66.88(12)	30.55(32)	30.19(40)	40.68(36)	0	0	0
Moon	60.54(28)	57.49(34)	66.77(13)	29.09(38)	30.14(48)	39.46(43)	18.35(59)	17.46(81)	21.23(48)
FedProc	61.83(29)	56.44(32)	67.62(22)	30.79(42)	30.10(48)	39.85(49)	18.41(51)	17.52(81)	21.23(52)
FedFA (Our)	<b>62.14(10)</b>	<b>59.46(7)</b>	<b>67.63(5)</b>	<b>30.67(15)</b>	<b>29.94(11)</b>	<b>41.72(14)</b>	<b>18.76(75)</b>	<b>17.70(106)</b>	<b>22.99(87)</b>

Table 6: The top-1 accuracy (round number when the accuracy reaches the target accuracy) of all methods under feature distribution skew on the test dataset.

Method	Feature Distribution Skew			Label & Feature Distribution Skew					
	EMNIST	Mixed Digits Local BN	Mixed Digits Federated BN	Mixed Digits with Local BN #C = 2	$\alpha = 0.1$	$\alpha = 0.5$	Mixed Digits with Federated BN #C = 2	$\alpha = 0.1$	$\alpha = 0.5$
	78.800286	63.07	63.767359	43.58	48.48	60.14	43.94	50.05	62.64
FedAvg w/o skew	80.34(6)	63.44(65)	66.61(72)	44.13(34)	49.57(42)	60.53(61)	44.26(35)	50.63(45)	62.66(64)
FedFA w/o skew	<b>87.85(2)</b>	<b>63.25(45)</b>	<b>66.53(49)</b>	<b>44.74(24)</b>	<b>48.81(28)</b>	<b>60.95(42)</b>	<b>44.61(24)</b>	<b>50.55(30)</b>	<b>63.00(43)</b>
FedAvg	81.12(6)	63.31(72)	66.13(70)	43.75(111)	48.61(84)	60.80(76)	45.87(112)	50.15(85)	63.39(68)
FedProx	81.43(7)	63.91(73)	66.47(72)	44.31(112)	48.54(90)	60.30(76)	45.20(112)	50.49(90)	62.70(68)
FedDyn	79.28(5)	64.42(63)	66.47(63)	43.90(105)	49.82(78)	61.45(64)	44.13(111)	51.24(78)	63.21(63)
Moon	80.99(6)	63.79(74)	67.20(79)	43.63(105)	49.70(84)	60.81(68)	44.13(111)	51.24(94)	63.02(68)
FedProc	82.62(8)	63.21(83)	66.20(79)	44.18(72)	48.82(61)	60.98(60)	44.87(69)	51.43(61)	62.83(58)
FedFA	<b>87.99(2)</b>	<b>64.21(14)</b>	<b>67.80(14)</b>	<b>46.03(24)</b>	<b>48.65(18)</b>	<b>60.89(17)</b>	<b>45.20(20)</b>	<b>50.47(18)</b>	<b>64.85(17)</b>

### F.2.2 TRAINING ACCURACY UNDER LABEL DISTRIBUTION SKEW

The training accuracy under label skew is shown as Figure 9 to Figure 11, which illustrates that the performance of FedFA is better than all baselines on FMNIST, CIFAR-10, and CIFAR-100. FedFA achieves better generalization than baselines on CIFAR-100 with ResNet but takes more communication rounds to converge. This observation is reasonable because regularizing only the penultimate layer by feature anchor loss takes more time to align the feature maps of the shallow layers, which we will explore in future work.

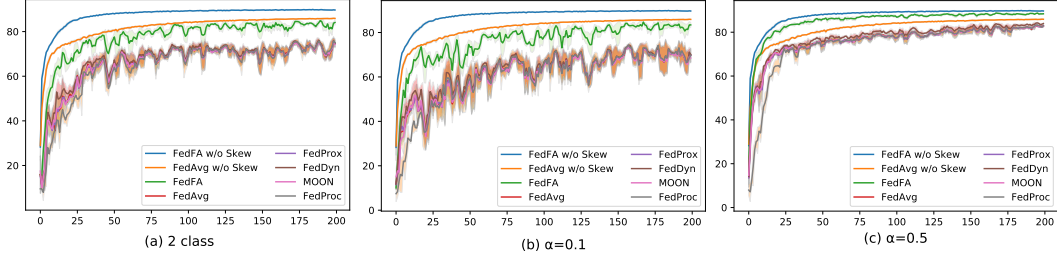


Figure 9: FMNIST with label skew.

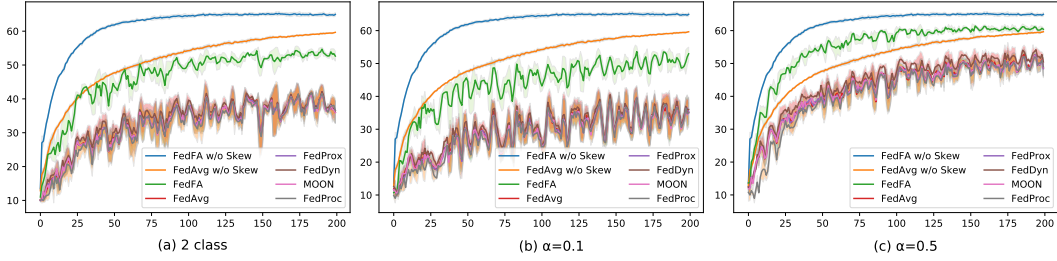


Figure 10: CIFAR-10 with label skew.

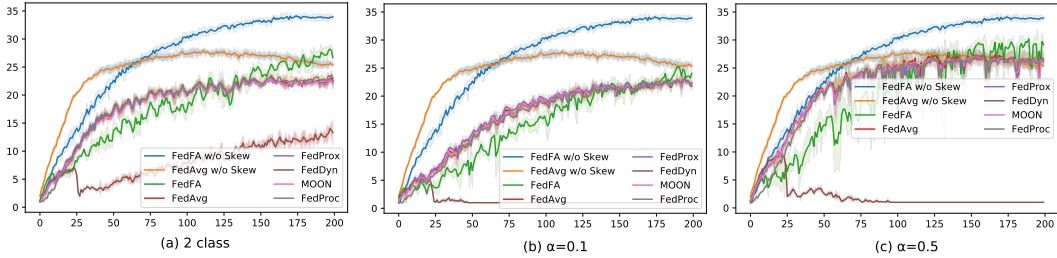


Figure 11: CIFAR-100 with label skew.

### F.2.3 TRAINING ACCURACY UNDER FEATURE DISTRIBUTION SKEW

Figures 12 to 13 show the better performance of FedFA over all baselines under feature distribution skew on EMNIST and Mixed Digit.

### F.2.4 TRAINING ACCURACY UNDER LABEL & FEATURE DISTRIBUTION SKEW

Figures 14 to 20 show the better performance of FedFA over all baselines under label & feature skew on EMNIST and Mixed Digit.



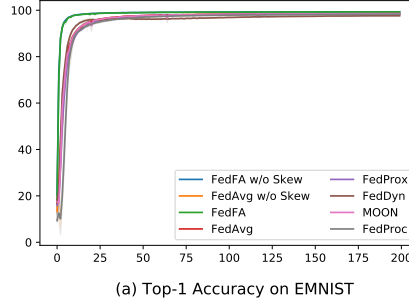


Figure 12: EMNIST with feature skew.

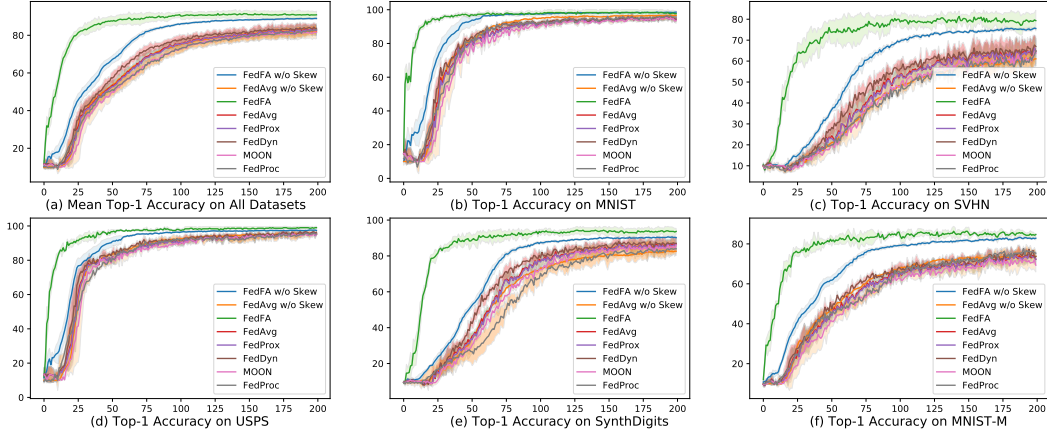
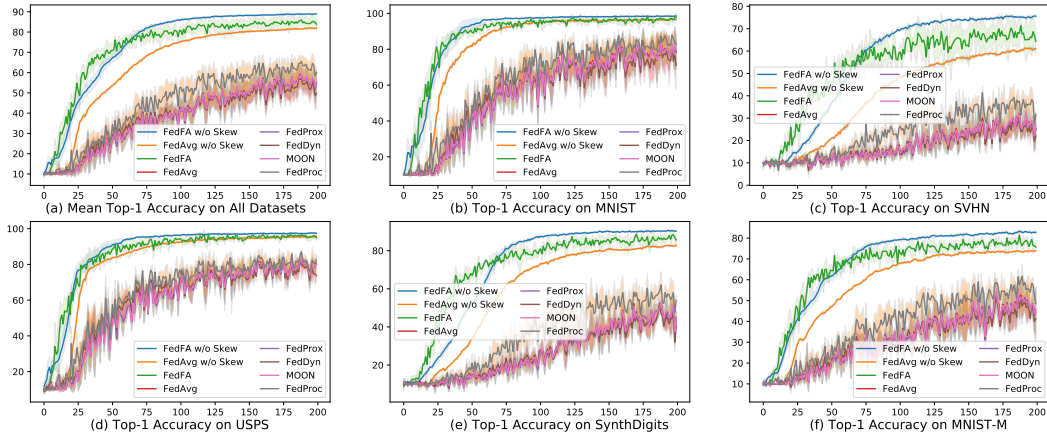


Figure 13: Mixed Digit without label skew.

Figure 14: Mixed Digit with label skew  $\#C = 2$  and federated BN.



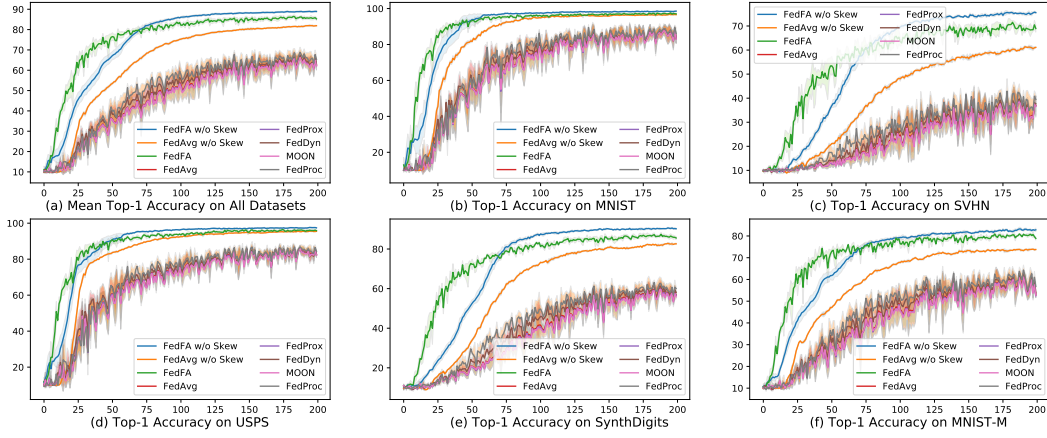
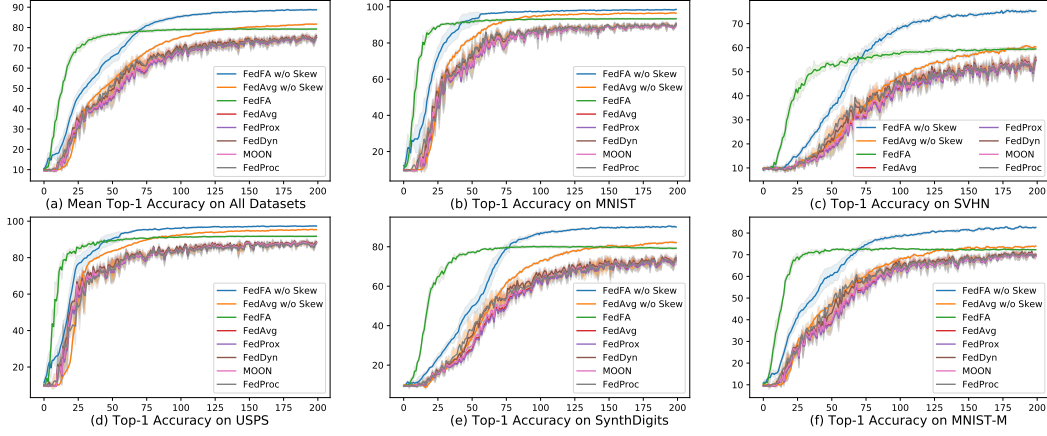
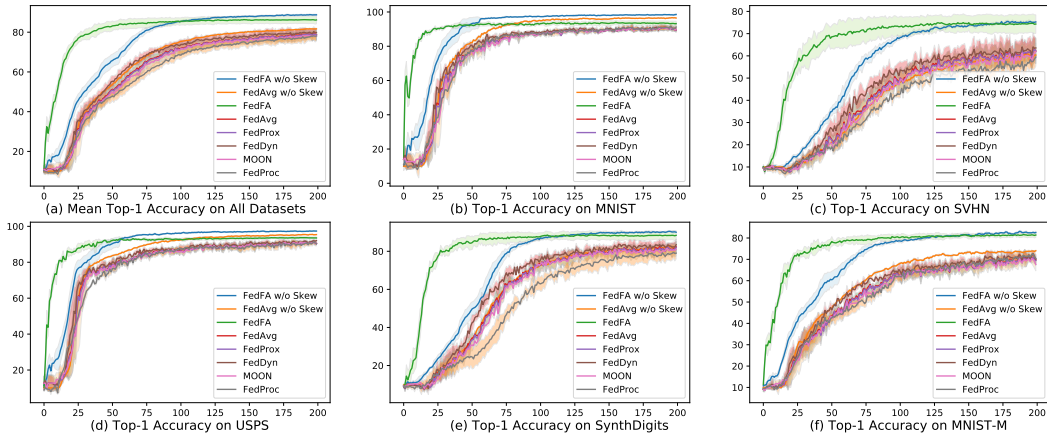
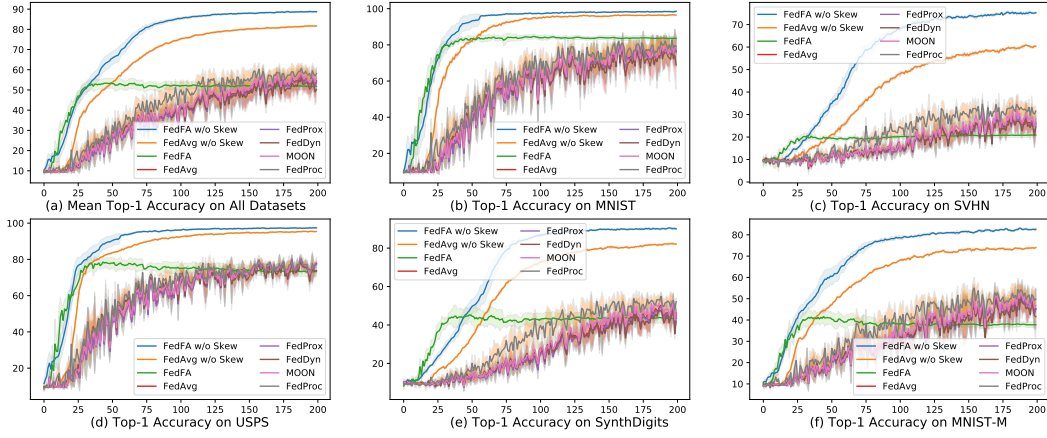
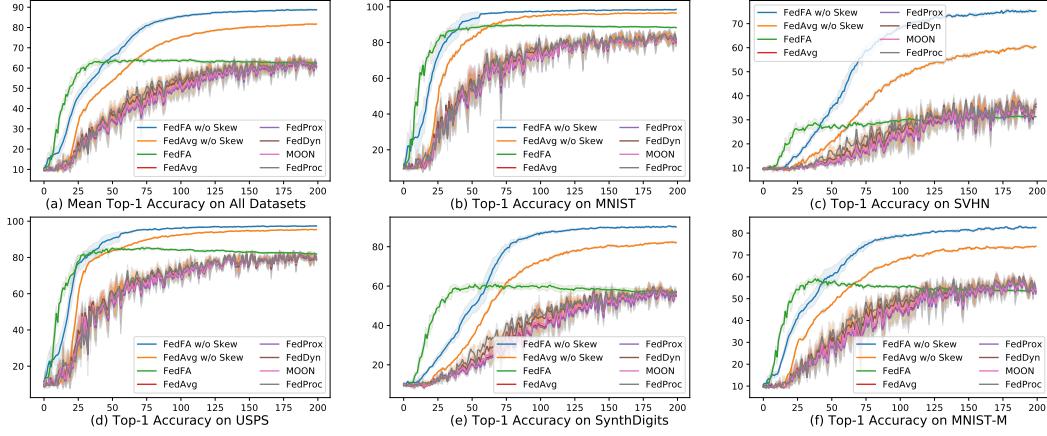
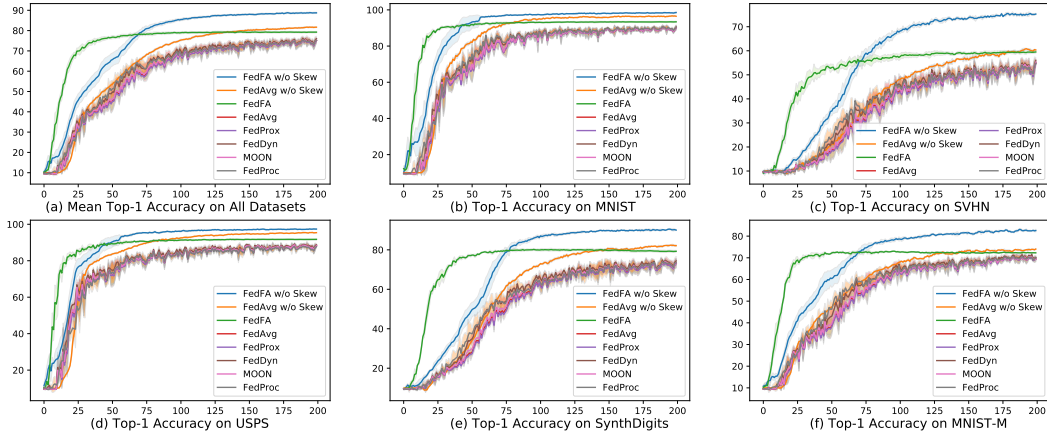
Figure 15: Mixed Digit with label skew  $\alpha = 0.1$  and federated BN.Figure 16: Mixed Digit with label skew  $\alpha = 0.5$  and federated BN.

Figure 17: Mixed Digit without label skew and with local BN.

Figure 18: Mixed Digit with label skew  $\#C = 2$  and with local BN.Figure 19: Mixed Digit with label skew  $\alpha = 0.1$  and with local BN.Figure 20: Mixed Digit with label skew  $\alpha = 0.5$  and with local BN.