

Fair Federated Learning under Domain Skew with Local Consistency and Domain Diversity

Yuhang Chen^{1*} Wenke Huang^{1*} Mang Ye^{1,2†}

¹National Engineering Research Center for Multimedia Software,
School of Computer Science, Wuhan University, Wuhan, China

²Taikang Center for Life and Medical Sciences, Wuhan University, Wuhan, China

{yhchen0, wenkehuang, yemang}@whu.edu.cn

<https://github.com/yuhangchen0/FedHEAL>

Abstract

Federated learning (FL) has emerged as a new paradigm for privacy-preserving collaborative training. Under domain skew, the current FL approaches are biased and face two fairness problems. 1) *Parameter Update Conflict*: data disparity among clients leads to varying parameter importance and inconsistent update directions. These two disparities cause important parameters to potentially be overwhelmed by unimportant ones of dominant updates. It consequently results in significant performance decreases for lower-performing clients. 2) *Model Aggregation Bias*: existing FL approaches introduce unfair weight allocation and neglect domain diversity. It leads to biased model convergence objective and distinct performance among domains. We discover a pronounced directional update consistency in Federated Learning and propose a novel framework to tackle above issues. First, leveraging the discovered characteristic, we selectively discard unimportant parameter updates to prevent updates from clients with lower performance overwhelmed by unimportant parameters, resulting in fairer generalization performance. Second, we propose a fair aggregation objective to prevent global model bias towards some domains, ensuring that the global model continuously aligns with an unbiased model. The proposed method is generic and can be combined with other existing FL methods to enhance fairness. Comprehensive experiments on Digits and Office-Caltech demonstrate the high fairness and performance of our method.

1. Introduction

Federated learning (FL) aims to collaboratively train a high-performance model while maintaining data privacy [28, 36].

*Equal contributions. †Corresponding author.

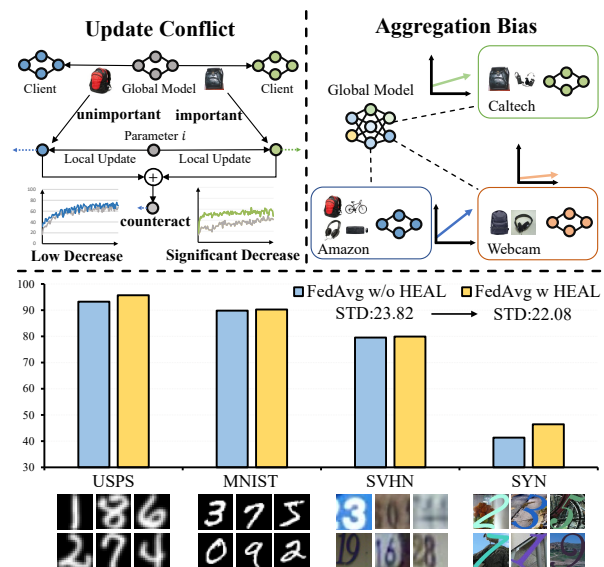


Figure 1. **Problem illustration** of Federated Learning under domain skew. Conventional FL methods (□) exhibit potential performance disparities due to Parameter Update Conflicts and Model Aggregation Bias. The former indicates that **varying parameter importance and inconsistent update directions** lead to an unfair decline in aggregated performance. The latter suggests biased convergence objective, resulting in **performance disparities**. Our method (□) achieves more equitable performance across different domains while enhancing overall performance.

The foundational method, FedAvg [36], allows numerous participants to send their models to the server instead of data. Then, the server aggregates these models into a global model and sends it back for further training. Notably, a significant challenge in FL is data heterogeneity [20, 22, 28, 29, 55], which means that client data appears in a Non-IID (non-independently and identically distributed) manner [32, 50, 54, 56, 59]. One particular heterogeneity type, domain skew [12, 16, 31, 55], refers to the client data

being sampled from **various domains**, leading to different feature distributions for each client.

Under domain skew, the local data are sampled from multiple domains, resulting in a significant disparity in distributed data. This disparity introduces challenges of federated convergent inconsistency. Meanwhile, federated learning aims to achieve a lower overall loss [36]. These two factors lead to FL being biased toward domains with easier convergence, further resulting in the neglect of other domains. This bias leads to distinct performance among domains. However, if some clients feel undervalued, their motivation to participate in the federation will diminish, leading to a narrow scope of knowledge in the federation, hindering its growth and contradicting original intent of FL. This issue gives rise to a pivotal challenge in FL: **Performance Fairness** [17, 45], which aims to ensure the uniform performance across different clients without neglecting clients with inferior performance. The fairness issue is highlighted in Fig. 1, where preliminary methods might overfit some domains, leading to poor performance in other domains. We argue that two primary reasons underlie this fairness issue: **I Parameter Update Conflict**: *The inconsistent parameter update directions and varying parameter importance lead to conflicts between important and unimportant parameters, degrading the performance of clients with poorer results.* Due to domain skew, there can be inconsistencies in parameter update directions among clients. Furthermore, some parameters of the neural network are more important to specific data [24, 46, 53], meaning that changes in these parameters have a larger impact on performance. Domain skew results in varying parameter importance. So important updates from poor-performing client may be potentially overwhelmed by unimportant aspects of others. But the latter can not signally boost performance. So it finally leads to performance disparity. **II Model Aggregation Bias**: *The general weighting distribution method is biased and neglects domain diversity, resulting in unfair convergence objective and performance disparity.* In conventional FL methods [29, 36], the strategy of weighting proportional to sample quantity [36] hinders the global model from adequately learning from domains with few samples. Alternatively, equal weighting overly emphasizes clients with fewer samples. Both strategies introduce biases and amplifying performance diversity. This bias disregards the data diversity among different domains.

To address these issues, we present a novel solution, **Federated Parameter-Harmonized and Aggregation-Equalized Learning** (FedHEAL). For problem **I**, we observe notable consistency in parameters updating during the local training. Specifically, due to the unique domain knowledge, some parameters are consistently pushed to the same direction (*i.e.*, increment and decrement) during local training across consecutive rounds, as detailed in

Sec. 3.3.1. Parameters with strong consistency occur because the global model fails to adapt to certain domains, leading to repeated adjustments in the same direction. Motivated by this, we argue that parameters with strong consistency are more crucial for specific domain. To mitigate parameter update conflict, we aim to prevent unimportant parameters from nullifying the crucial updates of domains with poor performance. By including only essential parameters, we mitigate important updates from being drowned out by less important ones, thus promoting Performance Fairness across multiple domains and clients.

For problem **II**, we argue that more diverse domains result in larger model changes during local training. To account for the data diversity, we propose an optimization objective that minimizes the variance of distances between the global model and all the local models. By reducing the variance, the global model maintains uniform distances to all the local models, preventing bias towards any clients. However, the extensive parameters in neural networks and the large number of clients in FL make it computationally intensive. Thus, we propose a simplified method approximately aligned with this fairness objective. Details of our simplified approach are elaborated in Sec. 3.3.2.

In this paper, FedHEAL consists of two components. **First**, by discarding unimportant parameters, we mitigate conflict among parameter updates. **Second**, we present a fair aggregation objective and a simplified implementation to prevent global model bias towards some domains. FedHEAL ultimately achieves Performance Fairness under domain skew. Since our approach only focuses on aggregation, it can be easily integrated with most existing FL methods. Our main contributions are summarized as follows:

- We identify the parameter update consistency in FL and introduce a partial parameter update method to update only parameters significant to the local domain, enhancing fair performance across domains.
- We propose a new fair federated aggregation objective and a practical approach to consider the domain diversity to improve Performance Fairness.
- We conduct comprehensive experiments on the Digits [18, 25, 40, 43] and Office-Caltech [9] datasets, providing evidence of effectiveness of our method through ablation studies and integrations with existing methods.

2. Related Work

2.1. Federated Learning with Data Heterogeneity

Federated learning aims to collaboratively train models without centralizing data to protect privacy. The pioneering work, FedAvg [36], trains a global model by aggregating participants' local model parameter. However, FedAvg is primarily designed for homogeneous data, and its performance degrades under data heterogeneity. Numerous meth-

Notation	Description	Notation	Description
m	Client Index	p_m	Client Weight
i	Parameter Index	$q_{m,i}$	Parameter Weight
M	Client Volume	$l_{m,i}$	Increment Proportion
G	Parameter Volume	$c_{m,i}$	PUC of Parameter
N_m	Sample Size	Δw_m^t	Model Update
\mathcal{W}^t	Global Model	$\Delta w_{m,i}^t$	Parameter Update
w_m^t	Client Model	d_m	Model Distance
τ	Importance Threshold	β	Update Momentum

Table 1. **Notation table** of this paper.

ods based on FedAvg have emerged to address data heterogeneity [7, 8, 13, 14, 34, 44]. FedProx [29], SCAFFOLD [22] and FedDyn [1] constrain local updates by adding penalty terms. FedProto [48] and MOON [26] enhance the alignment between client-side training at the feature level. However, these methods overlook the issue of domain skew, leading to diminished performance in multi-domain scenarios. Some methods have now been developed to address domain skew, such as FedBN [31] and FCCL [12, 15]. However, these methods focus on personalized models rather than shared models, the latter of which requires additional public datasets. FPL [16] focuses on addressing domain skew but requires each client to upload high-level feature information, contradicting the privacy-preserving nature of FL. FedGA [58] and FedDG [33] focus on the problem of unseen domain generalization, but the former requires an additional validation set, and the latter involves transmitting data information among multiple clients, posing privacy leakage concerns. In this paper, FedHEAL does not require any additional datasets or the transmission of additional signal information. We solely focus on the most fundamental transmitted information in FL: the model updates themselves, to extract the necessary information for enhancing Performance Fairness in multi-domain scenarios.

2.2. Fair Federated Learning

The fairness in FL is currently of widespread interest [4, 11, 41]. The mainstream categorizations of federated fairness fall into three classes [19, 45]: Performance Fairness [19, 27, 30, 38], Individual/Group Fairness [5, 6, 57], and Collaborative Fairness [19, 35, 42, 52, 60]. Performance fairness ensures that all participants experience similar and equitable performance improvements. Individual/Group Fairness aims to minimize model bias towards specific attributes (*e.g.*, gender). Collaborative Fairness ensures that participants are rewarded in proportion to contributions. This paper primarily addresses the issue of Performance Fairness. AFL [38] utilizes a min-max optimization to boost the performance of the worst-performing clients. q-Fedavg [27] recalibrates the aggregate loss by assigning higher weights to devices with higher losses to enhance performance fairness. FedFV [49] uses the cosine similarity to detect and eliminate gradient conflicts to achieve Performance Fairness. But they are tailored for label skew

Non-IID data and do not consider domain skew. Ditto [30] enhances Performance Fairness by incorporating a penalty term but employs a personalized model instead of a shared model. FedCE [19] addresses both Performance Fairness and Collaborative Fairness but necessitates an additional validation set, a requirement that is challenging to meet given the scarcity of client data in FL. Our method is tailored for Performance Fairness under domain skew and consider both the domain diversity. It can be easily integrated with existing methods to enhance their fairness.

3. Methodology

3.1. Preliminaries

Federated Learning. Following typical Federated Learning setup [29, 36, 37], we consider there are M clients (indexed by m). Each client holds private data $D_m = \{x_i, y_i\}_{i=1}^{N_m}$, where N_m represents the data size of client m . The optimization objective of FL is to minimize global loss:

$$\min_w F(w) = \sum_{m=1}^M p_m f_m(w), \quad (1)$$

where $f_m(w) = \frac{1}{N_m} \sum_{i=1}^{N_m} L(x_i, y_i; w)$, p_m is the weight of client. $L(x_i, y_i; w)$ is the loss of data (x_i, y_i) with model parameters w . Each client updates its model locally, and the server then aggregates model updates from all clients.

Domain Skew. In heterogeneous federated learning, domain skew among private data occurs when the marginal distribution of labels $P(y)$ is consistent across clients, but the conditional distribution of features given labels $P(x|y)$ varies among different clients [3, 12, 16, 21, 31, 51]:

$$P_m(x|y) \neq P_n(x|y) \quad \text{while} \quad P_m(y) = P_n(y). \quad (2)$$

3.2. Motivation

Observation of Parameter Update Consistency. In heterogeneous federated learning, grasping the characteristic of model updates across clients is crucial. This study introduces a novel observation, termed as **Parameter Update Consistency** (PUC), observed during local training phases. Through a toy experiment involving 4 clients, each sampling from distinct domains and training with a ResNet-10 network, we observe a significant Parameter Update Consistency during local training. As shown in Fig. 2, our findings indicate that a substantial proportion of parameters maintain consistent update directions in consecutive rounds of training. Specifically, most parameters demonstrate significant update consistency. This consistency is noticeable in shorter consecutive rounds (10 rounds) but persists even in the longer term (100 rounds), underscoring its enduring nature. Furthermore, the consistency observed in the last 10 rounds reveals that PUC remains prominent as the global model converges, indicating that **the converged global model has not adapted to specific domains**. These

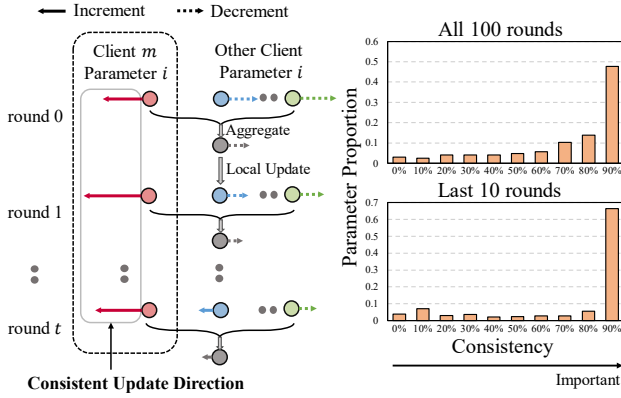


Figure 2. **Illustration of Parameter Update Consistency.** The consistency of parameter updates is displayed over 10 and 100 rounds for a randomly selected layer of the client model update. A significant proportion of parameters maintain a consistent update direction, *i.e.*, almost half of the parameters show the same direction for over 90 of the 100 rounds, indicating a persistent tendency to steer the global model in a fixed direction.

observations suggest a potential global model bias towards other domains, emphasizing the necessity for strategies that mitigate such biases and ensure fairer model aggregation.

Inspired by our observations and insights, to tackle Problem I, we categorize parameters into two distinct classes: important and unimportant. Important parameters are characterized by stable update directions in consecutive rounds, indicating consistent learning behavior within specific domains and inadequacy of the global model in fitting those domains. These parameters are deemed crucial for offering greater contribution in performance improvement. Conversely, unimportant parameters exhibit distinct directional changes across rounds which provide little contribution in performance improvement. We argue that involving unimportant parameters in the updates exacerbates the parameter update conflict, leading to a decline in the performance of some domains and resulting in unfair performance. Therefore, to mitigate this, we aim to minimize the impact of unimportant updates, enhancing the performance of poorly performing domains. During parameter aggregation, we discard unimportant parameters by setting their weights to zero, which prevents them from participating in aggregation and mitigates the parameter update conflict. Then we normalize all weights of a local model parameter. Those with a weight of zero remain at zero but the weights of important parameters increase, amplifying their influence on the global model and improving its adaptation to underperforming domains. The proposed method is detailed in Sec. 3.3.1.

To address Problem II, we employ domain diversity as the guiding metric for allocating weights. Our intuition is that more diverse data will undergo larger parameter changes for the local model to adapt to that domain. The parameter changes exhibit the fitting gap between the global

model and local model, implying the potential for performance improvement. A larger model changes suggests a domain is overlooked and has greater space for performance enhancement, which can also be reflected in the magnitude of model parameter updates. The distance between model iterations can represent the extent of parameter updates. Thus, this issue can be addressed through an optimization objective that minimizes the variance of distances between the global model and each client model. By reducing the variance, the global model maintains a more uniform distance to each client. However, this optimization problem is computationally intensive, so we introduce a simplified approach. We modify the weight allocation strategy with a momentum update [23, 47], assign greater weight to clients that induce larger parameter changes. This method approximately aligns with our objective and draws the new global model closer to neglected clients, thereby reducing the variance. We elaborate the proposed method in Sec. 3.3.2.

3.3. Proposed Method

3.3.1 Federated Parameter-Harmonized Learning

In a Federated Learning system with M clients, the global model in round t is denoted as \mathcal{W}^t . In round t , each client trains their model on private data D_m to obtain local model w_m^t . The local model change is defined as $\Delta w_m^t = w_m^t - \mathcal{W}^t$. The global model of next round, $t + 1$, is then updated by aggregating the model updates of round t :

$$\mathcal{W}^{t+1} = \mathcal{W}^t + \sum_{m=1}^M p_m \Delta w_m^t, \quad (3)$$

where p_m is the aggregation weight of client m , typically proportion to local sample size (*i.e.*, $p_m = \frac{N_m}{\sum_{j=1}^M N_j}$). In neural networks, w_m^t is potentially a large vector of parameters, and Δw_m^t is a vector of parameter changes with the same dimension. For simplicity, we disregard the internal structure of the model and represent Δw_m^t as

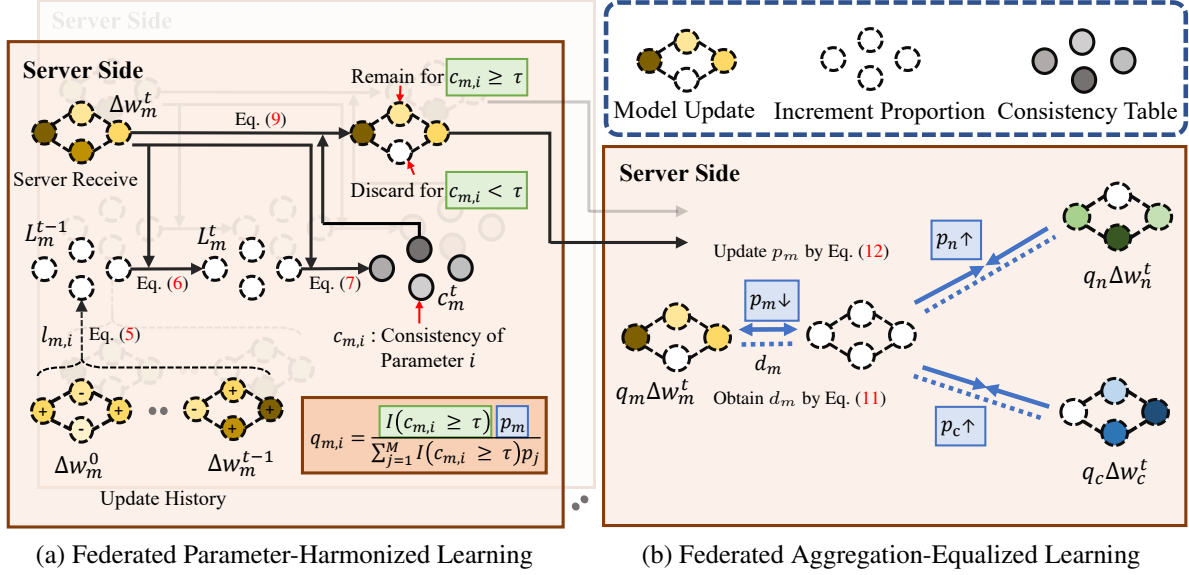
$$\Delta w_m^t = [\Delta w_{m,1}^t, \Delta w_{m,2}^t, \dots, \Delta w_{m,G}^t], \quad (4)$$

where G denotes the total number of parameters, and $\Delta w_{m,i}^t$ represents the change in the i^{th} parameter of the model w_m^t . We aim to compute the PUC for each parameter across all clients. This computation characterizes how significantly the parameters of the global model are pushed towards a consistent direction to fit specific domains. We maintain a list of proportions for parameters that are increasing for each client and then employ dynamic programming [2] to update this list in each round. The list of increment proportions for client m is denoted as

$$L_m = [l_{m,1}, l_{m,2}, \dots, l_{m,G}], \quad (5)$$

$$l_{m,i} = \frac{\sum_{j=0}^{t-1} I(\Delta w_{m,i}^j \geq 0)}{t},$$

where $I(\cdot)$ is the indicator function and $l_{m,i}$ is the proportion of increasing parameter before round t . Applying Dy-



(a) Federated Parameter-Harmonized Learning

(b) Federated Aggregation-Equalized Learning

Figure 3. **Architecture illustration** of FedHEAL. Clients send local model updates to the server. In FPHL (Sec. 3.3.1), the server maintains a consistency table, computes the consistency of current updates with past directions and discards updates with low consistency. Then in FAEL (Sec. 3.3.2), server minimized the variance of distance between global model and client model to mitigate model aggregation bias.

dynamic Programming [2], the $l_{m,i}$ in the round t can be calculated from the previous round's result:

$$l_{m,i} \leftarrow \frac{l_{m,i} * (t-1) + I(\Delta w_{m,i}^t \geq 0)}{t}. \quad (6)$$

Then the PUC of i^{th} parameter can be expressed as

$$c_{m,i} = PUC(w_{m,i}) = \begin{cases} l_{m,i} & \text{if } \Delta w_{m,i}^t \geq 0, \\ 1 - l_{m,i} & \text{otherwise} \end{cases}. \quad (7)$$

To alleviate parameter update conflicts and mitigate the global model from experiencing unfair performance, we select parameters with a strong PUC for retention and discard others. We introduce a hyperparameter τ for this purpose. We categorize the i^{th} parameter as important if $c_{m,i} \geq \tau$ and as unimportant otherwise. Given the model aggregation process in Eq. (3), the aggregation of i^{th} parameter is

$$\mathcal{W}_i^{t+1} = \mathcal{W}_i^t + \sum_{m=1}^M p_m \Delta w_{m,i}^t, \quad (8)$$

where \mathcal{W}_i^t is the i^{th} parameter of global model \mathcal{W}^t . By applying our method, Eq. (8) can be reformulated as

$$\mathcal{W}_i^{t+1} = \mathcal{W}_i^t + \sum_{m=1}^M q_{m,i}^t \Delta w_{m,i}^t, \quad (9)$$

$$q_{m,i}^t = \frac{I(c_{m,i} \geq \tau) p_m}{\sum_{j=1}^M I(c_{j,i} \geq \tau) p_j},$$

where p_m is the aggregation weight of client m and all its parameters, $q_{m,i}^t$ zeros out the weights of insignificant parameters and then further normalizes them to ensure the aggregation weights sum up to 1. Consequently, only important parameters will participate and unimportant updates no longer impact important updates. By normalizing $q_{m,i}^t$, the proportion of important parameter updates is further ampli-

fied, enhancing their contribution during aggregation.

3.3.2 Federated Aggregation-Equalized Learning

We first define the distance between the global model and the local model of client m in round t as $d_m = \|U - w_m^t\|_2^2$, where U is the new global model and $\|\cdot\|_2^2$ is the square of the euclidean distance. To minimize the variance of distances between the global model and each client model, we introduce the following optimization objective:

$$U_t^* = \arg \min_U \text{Var}(\{d_m\}_{m=1}^M) \\ = \arg \min_U \text{Var}(\{\|U - w_m^t\|_2^2\}_{m=1}^M) \quad (10)$$

$$\text{s.t. } U = \sum_{m=1}^M p_m w_m^t, \sum_{m=1}^M p_m = 1, \text{ and } \forall m, p_m \geq 0,$$

where U_t^* represents the unbiased global model. The time complexity of this optimization is $\mathcal{O}(qMG)$, where q is the number of iterations needed for convergence. However, computational resources are often limited in FL. Thus, we propose a simplified approach that reduces the time complexity to $\mathcal{O}(MG)$, requiring only a single distance calculation for each client. Specifically, in round t , we measure the distance between the trained client model and global model \mathcal{W}^t as $d_m = \|\Delta w_m^t\|_2^2$. Notably, if we combine it with FPHL, d_m can be rewritten as

$$d_m = \left\| \sum_{i=1}^G I(c_{m,i} \geq \tau) \cdot \Delta w_{m,i}^t \right\|_2^2, \quad (11)$$

implying that we only compute the important parameter change distance which better reflects the alterations made for specific domain. We then apply a momentum update

strategy [23, 47] to update the weight for each client:

$$\begin{aligned}\Delta p_m^t &= (1 - \beta)\Delta p_m^{t-1} + \beta \frac{d_m}{\sum_{j=1}^M d_j}, \\ p_m^t &= p_m^{t-1} + \Delta p_m^t, \quad p_m^t = \frac{p_m^t}{\sum_{j=1}^M p_j^t}.\end{aligned}\quad (12)$$

where β is a hyper-parameter, with larger values indicating a more pronounced influence of the distance on p_m . When $\beta = 0$, the method degenerates to the FedAvg. At $\beta = 1$, the weights are assigned purely based on the distance set of current round. The simplified method strives to minimize the variance, aligning with the unbiased global model.

3.4. Discussion and Limitation

The key notations are summarized in Tab. 1 and the pseudo-code of FedHEAL is presented in Algorithm 1.

Comparison with Analogous Methods. q-FFL [27] and FedCE [19] increase weights for poor-performing clients based on single loss or accuracy metrics. However, increasing the weights of these clients does not guarantee significant performance improvement. FAEL adjusts weights based on the domain diversity, which implies fitting gap between the global model and local models. It is a better way is to infer the effectiveness of weight modification based on the potential space for performance improvement. Similar work FedFV [49] alleviates model gradient conflicts. but it modifies gradients based on cosine similarity and gradient projection of the model, still allows less significant gradients to influence crucial ones. In contrast, FPHL, grounded in the observed PUC characteristics, selectively discards unimportant updates to safeguard the important ones. It demonstrates targeted conflict resolution, leading to better fairness and higher overall performance.

Discussion on FPHL. FPHL selectively discards unimportant parameters to reduce update conflicts, meaning it is particularly effective in large-scale FL systems where such conflicts are more pronounced. Similar with other methods aimed at Performance Fairness [27, 38], FPHL can increase the influence of clients with poorer performance, which may sometimes reduce the relative weight and performance of other clients. Yet, by discarding unimportant parameters uniformly, FPHL can diminish the adverse impact of both poorly performing client updates and better-performing clients to some extent. Consequently, while it significantly boosts the performance of the former, it may also help to prevent performance decline in the latter. So it achieves higher average accuracy across domains.

Limitation. Our method leverages the parameter update consistency and the fitting gap between the global model and local models to guide parameter aggregation and client aggregation weights. However, our method’s performance is sensitive to the selected hyperparameters. When a hyperparameter is not selected properly, our method may become unstable. Additionally, our method is designed for scenar-

Algorithm 1: FedHEAL

Input: Communication rounds T , local epochs \mathcal{K} , number of participants M , m^{th} participant private data D_m , private model w_m

Output: The final global model w^T

Server: initialize the global model w^0 and

$$L_m^0 = [0, 0, \dots, 0]_G$$

for $t = 0, 1, 2, \dots, T - 1$ **do**

Client:

for $m = 1, 2, \dots, M$ **in parallel do**

$$\quad w_m^t \leftarrow \mathcal{W}^t$$

for $k = 1, 2, \dots, \mathcal{K}$ **do**

$$\quad \quad w_m^t \leftarrow w_m^t - \eta \nabla \mathbf{CE}(w_m^t, D_m)$$

$$\Delta w_m^t \leftarrow w_m^t - \mathcal{W}^t$$

Server:

$$q^t, L^t \leftarrow \mathbf{FedHEAL}(L^{t-1})$$

for $i = 1, 2, \dots, G$ **do**

$$\quad \mathcal{W}_i^{t+1} = \mathcal{W}_i^t + \sum_{m=1}^M q_{m,i}^t \Delta w_{m,i}^t$$

FedHEAL(L^{t-1}):

for $m = 1, 2, \dots, M$ **do**

for $i = 1, 2, \dots, G$ **do**

$$\quad \quad l_{m,i} \leftarrow (l_{m,i}, \Delta w_{m,i}^t) \text{ in Eq. (6)}$$

$$\quad \quad c_{m,i} \leftarrow (l_{m,i}, \Delta w_{m,i}^t) \text{ in Eq. (7)}$$

$$\quad d_m \leftarrow (c_{m,i}, \Delta w_{m,i}^t) \text{ in Eq. (11)}$$

for $m = 1, 2, \dots, M$ **do**

$$p_m^t, \Delta p_m^t \leftarrow (p_m^{t-1}, \Delta p_m^{t-1}, \Delta w_{m,i}^t, D^t, \beta) \text{ in Eq. (12)}$$

for $i = 1, 2, \dots, G$ **do**

$$\quad \quad q_{m,i}^t \leftarrow (p_m^t, c_{m,i}^t) \text{ in Eq. (9)}$$

 return q^t, L^t

ios where all clients share the same network architecture, so it may fail in cases where clients have different architectures and parameter update consistency cannot be assumed.

4. Experiments

4.1. Experiment Details

Datasets. We evaluate our methods on two multi-domain image classification tasks.

- Digits [18, 25, 40, 43] includes four domains: MNIST, USPS, SVHN and SYN, each with 10 categories.
- Office-Caltech [9] includes four domains: Caltech, Amazon, Webcam, and DSLR, each with 10 categories.

We allocate 20 clients for each task and distribute an equal number of clients to each domain. We randomly sample a certain proportion for each client from their datasets, based on task difficulty and task size. Specifically, we sample 1% for Digits and 10% for Office-Caltech. We fix the seed to ensure reproduction of results. The example cases in each domain are presented in Fig. 4.

Model. For both classification tasks, we use ResNet-10 [10] as the shared model architecture for training.

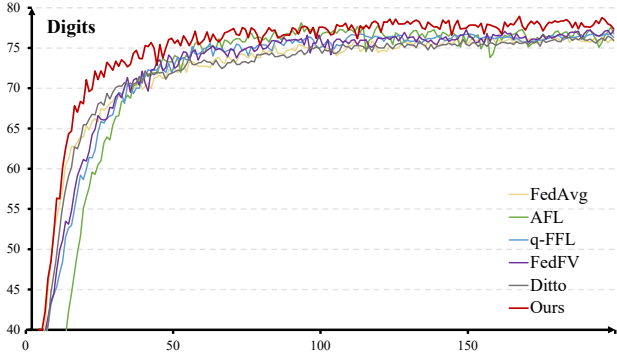


Figure 5. Comparison of convergence of average accuracy with counterparts on Digits. Please see details in Sec. 4.3.

β	0.0	0.2	0.4	0.6	0.8	1.0
AVG \uparrow	76.00	76.88	77.20	76.83	76.96	77.13
STD \downarrow	23.82	22.70	22.64	22.83	22.71	22.70

Table 4. Hyper-parameter study with different β (Eq. (12)) on Digits datasets. See details in Sec. 4.2.

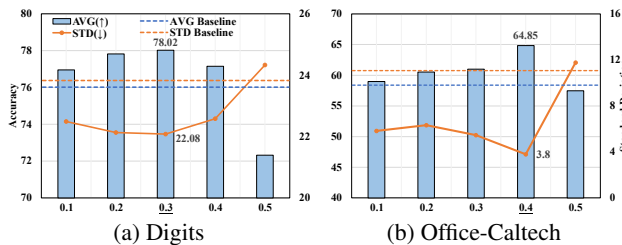


Figure 6. Hyper-parameter study with variant τ (Eq. (5)) when fix $\beta = 0.4$. Please see details in Sec. 4.2.

Ablation Study. To provide a comprehensive analysis of the effectiveness of FPHL and FAEL, we carried out an ablation study on both the Digits and Office-Caltech in Tab. 5. They contribute positively to the performance enhancement and their combination results in optimal performance.

Compatibility Study. To validate the compatibility of FedHEAL, we compared the results of several widely-adopted FL methods, FedAvg [36], FedProx [29], Scaffold [22], MOON [26], FedDyn [1], FedProc [39], FedProto [48] without and with FedHEAL. The results are shown in Tab. 2. They reveal tangible benefits offered by our system, *i.e.*, FedProto [48] with FedHEAL achieves 5.60% reduction in STD and 15.12% increase in AVG on Office-Caltech. We plot the differences in convergence between the benchmark without and with FedHEAL in Fig. 7, demonstrating faster convergence and higher accuracy of FedHEAL.

4.3. Comparison to State-of-the-Arts

The Tab. 2 and Tab. 3 shows the accuracy and standard deviation at the end of communication with SOTAs that address Performance Fairness in FL. The results depict that our method outperforms counterparts in both standard deviation and mean accuracy. This demonstrates that FedHEAL

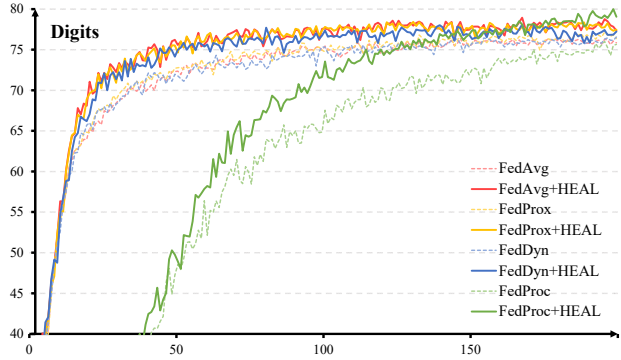


Figure 7. Comparison of convergence of average accuracy with and without the integration of FedHEAL, across selected FL methods. Please see details in Sec. 4.2.

		Digits					
FPHL	FAEL	MNIST	USPS	SVHN	SYN	AVG \uparrow	STD \downarrow
		89.84	93.25	79.54	41.35	76.00	23.82
\checkmark		92.19	95.32	76.32	44.37	77.05	23.32
	\checkmark	90.05	95.16	78.76	44.83	77.20	22.64
\checkmark	\checkmark	90.27	95.69	79.94	46.45	78.09	22.08
		Office-Caltech					
FPHL	FAEL	Amazon	DSLRL	Caltech	Webcam	AVG \uparrow	STD \downarrow
		72.63	56.67	58.57	45.52	58.35	11.13
\checkmark		68.42	66.00	57.95	66.55	64.73	4.64
	\checkmark	66.73	63.33	57.59	53.10	60.19	6.05
\checkmark	\checkmark	67.90	66.00	59.28	66.21	64.85	3.80

Table 5. Ablation study on Digits and Office-Caltech. Please refer to Sec. 4.2 for detailed discussion.

achieves better Performance Fairness and further improves accuracy across multiple domains. We plot the average accuracy at each epoch in Fig. 5, which illustrates the faster convergence of FedHEAL.

5. Conclusion

In this paper, we address Performance Fairness in federated learning with domain skew by tackling parameter update conflicts and model aggregation bias. We discover a property in federated learning which we term *Parameter Update Consistency*. Leveraging this characteristic, we propose a simple yet effective approach. By discarding unimportant parameters, FedHEAL alleviates parameter update conflicts for poor-performing clients. Moreover, considering domain diversity, we reduce the variance of distances between the global model and local models, addressing model aggregation bias. Extensive experiments demonstrate the effectiveness and compatibility of FedHEAL. We believe that this newly discovered property and our work will offer fresh research directions and insights for the community.

Acknowledgement. This work is supported by National Natural Science Foundation of China under Grant (62361166629, 62176188, 62272354).

References

- [1] Durmus Alp Emre Acar, Yue Zhao, Ramon Matas, Matthew Mattina, Paul Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *ICLR*, 2021. 3, 7, 8
- [2] Richard Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, 1957. 4, 5
- [3] Debora Caldarola, Massimiliano Mancini, Fabio Galasso, Marco Ciccone, Emanuele Rodolà, and Barbara Caputo. Cluster-driven graph federated learning over multiple domains. In *CVPRW*, pages 2749–2758, 2021. 3
- [4] Huiqiang Chen, Tianqing Zhu, Tao Zhang, Wanlei Zhou, and Philip S Yu. Privacy and fairness in federated learning: on the perspective of trade-off. *ACM Computing Surveys*, 2023. 3
- [5] Sen Cui, Weishen Pan, Jian Liang, Changshui Zhang, and Fei Wang. Addressing algorithmic disparity and performance inconsistency in federated learning. *NeurIPS*, 34:26091–26102, 2021. 3
- [6] Yahya H Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and A Salman Avestimehr. Fairfed: Enabling group fairness in federated learning. In *AAAI*, pages 7494–7502, 2023. 3
- [7] Xiuwen Fang and Mang Ye. Robust federated learning with noisy and heterogeneous clients. In *CVPR*, pages 10072–10081, 2022. 3
- [8] Xiuwen Fang, Mang Ye, and Xiyuan Yang. Robust heterogeneous federated learning under data corruption. In *ICCV*, pages 5020–5030, 2023. 3
- [9] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, pages 2066–2073. IEEE, 2012. 2, 6, 7
- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6
- [11] Samuel Horvath, Stefanos Laskaridis, Mario Almeida, Ilias Leontiadis, Stylianos Venieris, and Nicholas Lane. Fjord: Fair and accurate federated learning under heterogeneous targets with ordered dropout. *NeurIPS*, 34:12876–12889, 2021. 3
- [12] Wenke Huang, Mang Ye, and Bo Du. Learn from others and be yourself in heterogeneous federated learning. In *CVPR*, pages 10143–10153, 2022. 1, 3
- [13] Wenke Huang, Mang Ye, Bo Du, and Xiang Gao. Few-shot model agnostic federated learning. In *ACM MM*, pages 7309–7316, 2022. 3
- [14] Wenke Huang, Guancheng Wan, Mang Ye, and Bo Du. Federated graph semantic and structural learning. In *IJCAI*, pages 139–143, 2023. 3
- [15] Wenke Huang, Mang Ye, Zekun Shi, and Bo Du. Generalizable heterogeneous federated cross-correlation and instance similarity learning. *IEEE PAMI*, 2023. 3
- [16] Wenke Huang, Mang Ye, Zekun Shi, He Li, and Bo Du. Rethinking federated learning with domain shift: A prototype view. In *CVPR*, pages 16312–16322, 2023. 1, 3
- [17] Wenke Huang, Mang Ye, Zekun Shi, Guancheng Wan, He Li, Bo Du, and Qiang Yang. A federated learning for generalization, robustness, fairness: A survey and benchmark. *arXiv*, 2023. 2
- [18] Jonathan J. Hull. A database for handwritten text recognition research. *IEEE PAMI*, 16(5):550–554, 1994. 2, 6, 7
- [19] Meirui Jiang, Holger R Roth, Wenqi Li, Dong Yang, Can Zhao, Vishwesh Nath, Daguang Xu, Qi Dou, and Ziyue Xu. Fair federated medical image segmentation via client contribution estimation. In *CVPR*, pages 16302–16311, 2023. 3, 6
- [20] Xuefeng Jiang, Sheng Sun, Yuwei Wang, and Min Liu. Towards federated learning against noisy labels via local self-regularization. In *CIKM*, pages 862–873, 2022. 1
- [21] 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. 3
- [22] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *ICML*, pages 5132–5143. PMLR, 2020. 1, 3, 7, 8
- [23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014. 4, 6
- [24] Yann LeCun, John Denker, and Sara Solla. Optimal brain damage. *NeurIPS*, 2, 1989. 2
- [25] 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. 2, 6, 7
- [26] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *CVPR*, pages 10713–10722, 2021. 3, 7, 8
- [27] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. In *ICLR*, 2019. 3, 6, 7
- [28] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3):50–60, 2020. 1
- [29] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *MLSys*, 2:429–450, 2020. 1, 2, 3, 7, 8
- [30] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *ICML*, pages 6357–6368. PMLR, 2021. 3, 7
- [31] Xiaoxiao Li, Meirui JIANG, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. In *ICLR*, 2021. 1, 3
- [32] Zexi Li, Xinyi Shang, Rui He, Tao Lin, and Chao Wu. No fear of classifier biases: Neural collapse inspired federated learning with synthetic and fixed classifier. In *ICCV*, pages 5319–5329, 2023. 1

- [33] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *CVPR*, pages 1013–1023, 2021. 3
- [34] Kangyang Luo, Xiang Li, Yunshi Lan, and Ming Gao. Grandma: A gradient-memory-based accelerated federated learning with alleviated catastrophic forgetting. In *CVPR*, pages 3708–3717, 2023. 3
- [35] Lingjuan Lyu, Xinyi Xu, Qian Wang, and Han Yu. Collaborative fairness in federated learning. *Federated Learning: Privacy and Incentive*, pages 189–204, 2020. 3
- [36] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguerre y Arcas. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, pages 1273–1282, 2017. 1, 2, 3, 7, 8
- [37] Jiayu Miao, Zongxin Yang, Leilei Fan, and Yi Yang. Fedseg: Class-heterogeneous federated learning for semantic segmentation. In *CVPR*, pages 8042–8052, 2023. 3
- [38] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *ICLR*, pages 4615–4625. PMLR, 2019. 3, 6, 7
- [39] Xutong Mu, Yulong Shen, Ke Cheng, Xueli Geng, Jiaxuan Fu, Tao Zhang, and Zhiwei Zhang. Fedproc: Prototypical contrastive federated learning on non-iid data. *arXiv preprint arXiv:2109.12273*, 2021. 7, 8
- [40] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bisaccho, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 2, 6, 7
- [41] Tao Qi, Fangzhao Wu, Chuhan Wu, Lingjuan Lyu, Tong Xu, Hao Liao, Zhongliang Yang, Yongfeng Huang, and Xing Xie. Fairvfl: A fair vertical federated learning framework with contrastive adversarial learning. *NeurIPS*, 35:7852–7865, 2022. 3
- [42] Bhaskar Ray Chaudhury, Linyi Li, Mintong Kang, Bo Li, and Ruta Mehta. Fairness in federated learning via core-stability. *NeurIPS*, 35:5738–5750, 2022. 3
- [43] Prasun Roy, Subhankar Ghosh, Saumik Bhattacharya, and Umapada Pal. Effects of degradations on deep neural network architectures. *arXiv preprint arXiv:1807.10108*, 2018. 2, 6, 7
- [44] Jiangming Shi, Shanshan Zheng, Xiangbo Yin, Yang Lu, Yuan Xie, and Yanyun Qu. Clip-guided federated learning on heterogeneous and long-tailed data. *arXiv preprint arXiv:2312.08648*, 2023. 3
- [45] Yuxin Shi, Han Yu, and Cyril Leung. Towards fairness-aware federated learning. *IEEE TNNLS*, 2023. 2, 3
- [46] Neta Shoham, Tomer Avidor, Aviv Keren, Nadav Israel, Daniel Benditkis, Liron Mor-Yosef, and Itai Zeitak. Overcoming forgetting in federated learning on non-iid data. *arXiv preprint arXiv:1910.07796*, 2019. 2
- [47] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *ICML*, pages 1139–1147. PMLR, 2013. 4, 6
- [48] Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. Fedproto: Federated prototype learning across heterogeneous clients. In *AAAI*, pages 8432–8440, 2022. 3, 7, 8
- [49] Zheng Wang, Xiaoliang Fan, Jianzhong Qi, Chenglu Wen, Cheng Wang, and Rongshan Yu. Federated learning with fair averaging. *IJCAI*, 2021. 3, 6, 7
- [50] Nannan Wu, Li Yu, Xin Yang, Kwang-Ting Cheng, and Zengqiang Yan. Fediic: Towards robust federated learning for class-imbalanced medical image classification. In *MICCAI*, pages 692–702. Springer, 2023. 1
- [51] An Xu, Wenqi Li, Pengfei Guo, Dong Yang, Holger R Roth, Ali Hatamizadeh, Can Zhao, Daguang Xu, Heng Huang, and Ziyue Xu. Closing the generalization gap of cross-silo federated medical image segmentation. In *CVPR*, pages 20866–20875, 2022. 3
- [52] Xinyi Xu, Lingjuan Lyu, Xingjun Ma, Chenglin Miao, Chuan Sheng Foo, and Bryan Kian Hsiang Low. Gradient driven rewards to guarantee fairness in collaborative machine learning. *NeurIPS*, 34:16104–16117, 2021. 3
- [53] Xiyuan Yang, Wenke Huang, and Mang Ye. Dynamic personalized federated learning with adaptive differential privacy. In *NeurIPS*, 2023. 2
- [54] Xiyuan Yang, Wenke Huang, and Mang Ye. Fedas: Bridging inconsistency in personalized federated learning. In *CVPR*, 2024. 1
- [55] Mang Ye, Xiuwen Fang, Bo Du, Pong C Yuen, and Dacheng Tao. Heterogeneous federated learning: State-of-the-art and research challenges. *arXiv preprint arXiv:2307.10616*, 2023. 1
- [56] Mang Ye, Wenke Huang, Zekun Shi, He Li, and Du Bo. Revisiting federated learning with label skew: An overconfidence perspective. *SCIS*, 2024. 1
- [57] Yuchen Zeng, Hongxu Chen, and Kangwook Lee. Improving fairness via federated learning. *arXiv preprint arXiv:2110.15545*, 2021. 3
- [58] Ruipeng Zhang, Qinwei Xu, Jiangchao Yao, Ya Zhang, Qi Tian, and Yanfeng Wang. Federated domain generalization with generalization adjustment. In *CVPR*, pages 3954–3963, 2023. 3
- [59] 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. 1
- [60] Zirui Zhou, Lingyang Chu, Changxin Liu, Lanjun Wang, Jian Pei, and Yong Zhang. Towards fair federated learning. In *ACM SIGKDD*, pages 4100–4101, 2021. 3