Federated Learning with Domain Shift Eraser

Zheng Wang[1,2,3]   Zihui Wang[5]   Zheng Wang [4]   Xiaoliang Fan [1,2]   Cheng Wang [1,2*]

[1]Fujian Key Laboratory of Sensing and Computing for Smart Cities, Xiamen University, China
[2]Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, School of Informatics, Xiamen University, China
[3] Shanghai Innovation Institution
[4] School of Informatic, Xiamen University, China
[5] Peng Cheng Laboratory, Shenzhen, China

Abstract

Federated learning (FL) is emerging as a promising technique for collaborative learning without local data leaving their devices. However, clients' data originating from diverse domains may degrade model performance due to domain shifts, preventing the model from learning consistent representation space. In this paper, we propose a novel FL framework, Federated Domain Shift Eraser (FDSE), to improve model performance by differently erasing each client's domain skew and enhancing their consensus. First, we formulate the model forward passing as an iterative deskewing process that extracts and then deskews features alternatively. This is efficiently achieved by decomposing each original layer in the neural network into a Domain-agnostic Feature Extractor (DFE) and a Domain-specific Skew Eraser (DSE). Then, a regularization term is applied to promise the effectiveness of feature deskewing by pulling local statistics of DSE's outputs close to the globally consistent ones. Finally, DFE modules are fairly aggregated and broadcast to all the clients to maximize their consensus, and DSE modules are personalized for each client via similarity-aware aggregation to erase their domain skew differently. Comprehensive experiments were conducted on three datasets to confirm the advantages of our method in terms of accuracy, efficiency, and generalizability.

1. Introduction

Federated learning has emerged as a promising paradigm for training machine learning models on distributed data while preserving privacy [24]. However, a notable challenge arises when clients originate from diverse domains, where domain shifts across clients may result in performance degradation of the aggregated model [4, 11]. For
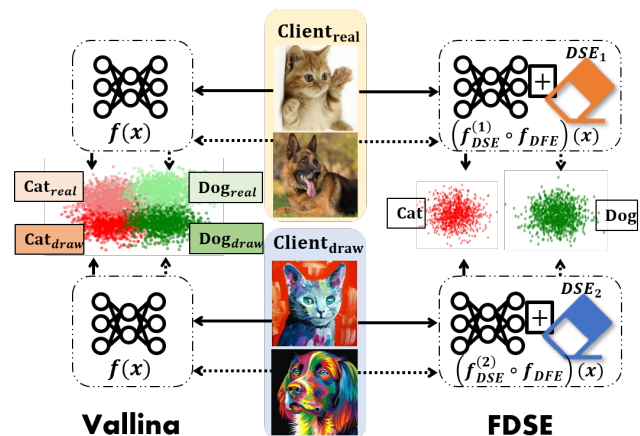
*Corresponding Author

Figure 1. **Illustration of the characteristic of the proposed FDSE v.s. Vallina FL [24]**. FDSE decomposes the model to respectively erase domain skew for each client by $f_{DSE}^{(i)}$ and extract generalizable features for all the clients by f_{DFE}. This promotes consistency in representation space across domains since knowledge is fine-grainedly decoupled for learning and aggregation.

example, hospitals located in different regions may collect data from diverse patient populations using various machines and protocols, leading to distinct domains in their feature distributions [20]. The misalignment in feature spaces can hinder client consensus on the representation space (e.g., Figure 1 left), finally reducing model performance in two ways. First, the model is compelled to additionally learn generalizable representations for vastly different samples across domains, competitively preventing it from focusing on the task objective. Second, the training process may be dominated by a single domain when clients' model updates exhibit conflicts and substantial variations in magnitude [5, 37], which causes the model to overfit on the prevailing domain while sacrificing its utility on others [10].

Recent works addressing the FL domain shift problem can be categorized into two main groups. The first group

(i.e., consensus-based methods) [5, 36, 43] enhances the consensus among clients at various levels to improve the model generalizability across domains. In contrast, another group (i.e., personalization-based methods) [20, 21, 32] mitigates the need for clients' consensus by personalizing their models to fit their local data distributions. Since the two groups exhibit complementary strengths in modeling different types of knowledge (i.e., general and personal knowledge), it's natural to consider a hybrid solution that synergizes their advantages for further improvement. However, model personalization may severely hinder clients' consensus due to over-focusing on clients' local benefits, thereby resulting in invalid integration. This motivates us to rethink the FL domain shift from the hybrid perspective

How to personalize models to enhance clients' consensus?

In this work, we address this issue by **using personalized parameters to erase domain-specific skew for each client while enhancing clients' consensus on other parameters** based on two key observations. First, visual learning tasks often mitigate samples' domain-specific skew differently. For example, in point cloud segmentation, high-density and low-density point clouds from channel-varying lidars are respectively downsampled and upsampled to achieve their alignment in feature space [39]. These inherently opposing processes of feature alignment necessitate personalized modeling. Second, a unified treatment is usually applied to different samples after their domain-specific skew is mitigated [11, 13, 39], revealing the importance of achieving post-deskewing consensus in a domain-agnostic way. Following the above paradigm, we develop a novel FL framework, **F**ederated **D**omain **S**hift **E**raser (**FDSE**), aiming to differently erase clients' domain-specific skew and then enhance their consensus in a decoupled manner. First, we formulate the model forward passing as an iterative deskewing process that extracts and then deskews features alternatively. This is efficiently achieved by decomposing each original layer in the neural network into a Domain-agnostic Feature Extractor (DFE) and a Domain-specific Skew Eraser (DSE). Then, a regularization term is applied to enhance the effectiveness of feature deskewing by pulling local statistics of each DSE's output close to the globally consistent ones. Finally, DFE modules are fairly aggregated and broadcast to all the clients to maximize their consensus, while DSE modules are respectively aggregated for each client based on their similarity to erase heterogeneous domain skew collaboratively. Compared to the Vallina FL in Figure 1 (left), our FDSE benefits from the design that emphasizes fine-grained personalization and consensus maximization on decoupled objectives, promoting both the consistency in the representation space and the efficiency of aggregating different types of knowledge. Our contributions are summarized as follows:

- We rethink the domain shift problem in FL from a novel hybrid perspective that integrates the advantages of two distinct sets of existing methods, i.e., consensus-based and personalization-based methods.
- We develop a novel framework, FDSE, to improve cross-domain FL by differently erasing clients' domain skew while enhancing their consensus in a decoupled manner.
- We conduct comprehensive experiments on three datasets (i.e., Office-Caltech10, PACS, and DomainNet) to confirm FDSE's superiority in terms of accuracy, efficiency, and generalizability against the state-of-the-art methods.

2. Related Works

2.1. Data Heterogeneity in FL

FL often faces challenges due to data heterogeneity, complicating the training process and degrading overall model accuracy [24, 41]. Previous efforts have been devoted to handling heterogeneous data by maintaining consistency between local objective and global objective [1, 15, 17, 18] or smoothening the clients' loss landscape during model training [6, 26, 30, 33]. Another series of works instead focuses on enhancing models' local performance under data heterogeneity via model personalization techniques, such as partial parameter sharing [12, 21, 32], meta-learning[8, 19], latent representation space alignment[27, 36], hypernetwork[31], and knowledge decoupling[3, 40]. Nevertheless, most of these methods are naturally designed for and verified in label skew scenarios [14], overlooking the domain shift problem and resulting in suboptimal performance in cross-domain scenarios [5].

2.2. Cross-domain FL

Cross-domain FL refers to FL with clients whose datasets come from different domains [5, 11, 13, 20, 22, 29, 32, 43]. The domain shift across clients' datasets can hinder the FL model training process, leading to model performance degradation [5]. Previous methods addressing this issue can be mainly grouped into two sets. Methods in the first set focus on maximizing different clients' consensus across domains at different levels. For example, [13] exchanges a few local data information across clients to normalize their samples in the frequency space at the input level. [43] increases the visibility of all domains' samples to each client during local training by feature augmentation in the intermediate layers of the model, leading to improved global consensus at the feature level. [11, 36] maintain consistent class prototypes before the output layers of the model to increase clients' consensus in the representation space. [5] instead enhances clients' consensus directly at the model parameter level by minimizing the conflicts of model updates of important parameters. Another set of methods uses model personalization techniques to help suit the global model to their

local domain skew. [20] keeps batch normalization layers to be locally maintained to make the model more adaptable to their personal domains. [32] adaptively loads partial parameters instead of the entire model for each local training. Despite the rapid development of the two sets of methods respectively, the way to integrate the advantages of the two sets still remains unexplored, leading to a potentially large space for further performance improvement.

To this end, we rethink the domain shift problem in FL from a hybrid view of existing methods (e.g., consensus enhancement and personalization). Different from previous studies, we fine-grained identify what should be personalized (e.g., domain-specific skew erasing) and on which the consensus should be maximized (the treatment to deskewed features), thus successfully bridging the two distinct insights behind existing approaches.

3. Problem Formulation

Personalized FL. In personalized FL, there exist N clients equipped with their private data $\{\mathcal{D}_i | i \in [N]\}$, and the goal is to obtain a series of model $\boldsymbol{\Theta} = \{\theta_1, \theta_2, ..., \theta_N\}$ that minimize

$$\min_{\boldsymbol{\Theta}} = \sum_{i \in [N]} \frac{|\mathcal{D}_i|}{|\mathcal{D}|} F_i, F_i = \mathbb{E}_{(x,y) \sim \mathcal{D}_i} [\ell(\theta_i; x, y)] \quad (1)$$

where $|\mathcal{D}| = \sum_{i=1}^{N} \mathcal{D}_i$ and $\ell(\cdot)$ is the loss function. In this paper, we consider one popular scheme of personalized FL where some parameters are globally shared among all the clients and the others are privately kept [2, 12]. Let $\theta_i = [\theta_u; \theta_{v,i}]$, we denote θ_u the globally shared parameters and $\theta_{v,i}$ the personalized ones.

Cross-domain FL Cross-domain FL [5, 43] refers to FL with clients that exhibit domain shift among their local datasets. The domain shift means that different clients share the consistent label space $\mathbb{P}(Y)$ but differ in conditional feature distribution $\mathbb{P}(X|Y)$ given Y (e.g., $\mathbb{P}_i(X|Y = y) \neq \mathbb{P}_j(X|Y = y), \forall i \neq j, i, j \in [N])$.

4. Methodology

In this section, we present the main procedure and implementation of FDSE. We begin by illustrating the layer decomposition approach, which allows us to decouple the personalized domain skew erasing from the common task objective by assigning them to different model parameters in Sec. 4.1. Next, we introduce the consistency regularization used during clients' local training to enhance the effectiveness of feature deskewing in Sec. 4.2. Finally, we discuss the model aggregation strategies respectively for common and personalized parameters in Sec. 4.3.

4.1. Layer Decomposition

As aforementioned in Sec.1, we personalize partial model parameters for each client to erase their domain-specific skew. Existing deskewing methods usually align cross-domain features at a certain layer of the neural network. For example, [13, 39] align samples' attributions in the feature space before feeding them into the model, and [21, 36] mitigate the domain divergence in the representation space at the output layer. However, premature deskewing (e.g., the input layer) may lead to insufficient skew elimination while the too-delayed one (e.g., the output layer) can hinder domain-agnostic knowledge extraction. Therefore, instead of only deskewing features at a certain layer, we view the model forward passing as an iterative deskewing process that extracts and then deskews features alternatively. This is efficiently achieved by decomposing each original layer in the neural network into a Domain-agnostic Feature Extractor (DFE) and a Domain-specific Skew Eraser (DSE), as is shown in Figure 2. Concretely, given a convolution layer f with filter size k, input channel number S, and output channel number T, we follow [9] to decompose it into two sub-convolution modules $f_{\text{DSE}} \circ f_{\text{DFE}}$ where each submodule consists of a convolution, a batch normalization layer, and an activation function. The DFE module shares parameters with the original layer f (e.g., kernel size, stride, and padding) except for the number of output channels $T_{\text{DFE}} = \lceil T/G \rceil$ by

$$\mathbf{X}_{\text{DFE}} = f_{\text{DFE}} * \mathbf{X} \quad (2)$$

where $\mathbf{X}_{\text{DFE}} \in \mathbb{R}^{T_{\text{DFE}} \times h \times w}$, h, w are respectively the height and the weight of the original output $f(\mathbf{X})$ and G is the architecture parameter. Then, the DFE module maps each channel in the DFE module's output into G channels via cheap linear operations by

$$\mathbf{X}_{\text{DSE},i} = f_{\text{DSE},i} * \text{ReLU}\left(\text{BN}_{\text{DSE}}\left(\mathbf{X}_{\text{DFE},i}\right)\right), i \in [T_{\text{DFE}}]$$
$$\mathbf{X}_{\text{out}} = \text{ReLU}\left(\text{BN}_{\text{DFE}}\left(\text{Concat}(\{\mathbf{X}_{\text{DSE},i}\})\right)\right) \quad (3)$$

where $\mathbf{X}_{\text{DFE},i} \in \mathbb{R}^{1 \times h \times w}, \mathbf{X}_{\text{DSE},i} \in \mathbb{R}^{G \times h \times w}, \mathbf{X}_{\text{out}} \in \mathbb{R}^{T \times h \times w}$, and each $f_{\text{DSE},i}$ augments each channel $\mathbf{X}_{\text{DFE},i}$ to G channels. Although there have been plenty of works that decompose convolution layers for efficiency [25, 34, 35], this layer separation has the natural advantage for our objective decoupling purpose. On one hand, the core components of the output features are independently extracted by the DFE module, each of which is then cheaply expanded to several similar variants by the DSE module. This paradigm enables the DFE module to learn the primary knowledge (e.g., the core components of feature maps), making it suitable for modeling domain-agnostic information. On the other hand, the simplicity of the DSE module avoids over-transforming features during deskewing processes, e.g., $94\times$ smaller than the DFE module in a convolution with 64 channels and kernel size 5. We further enhance
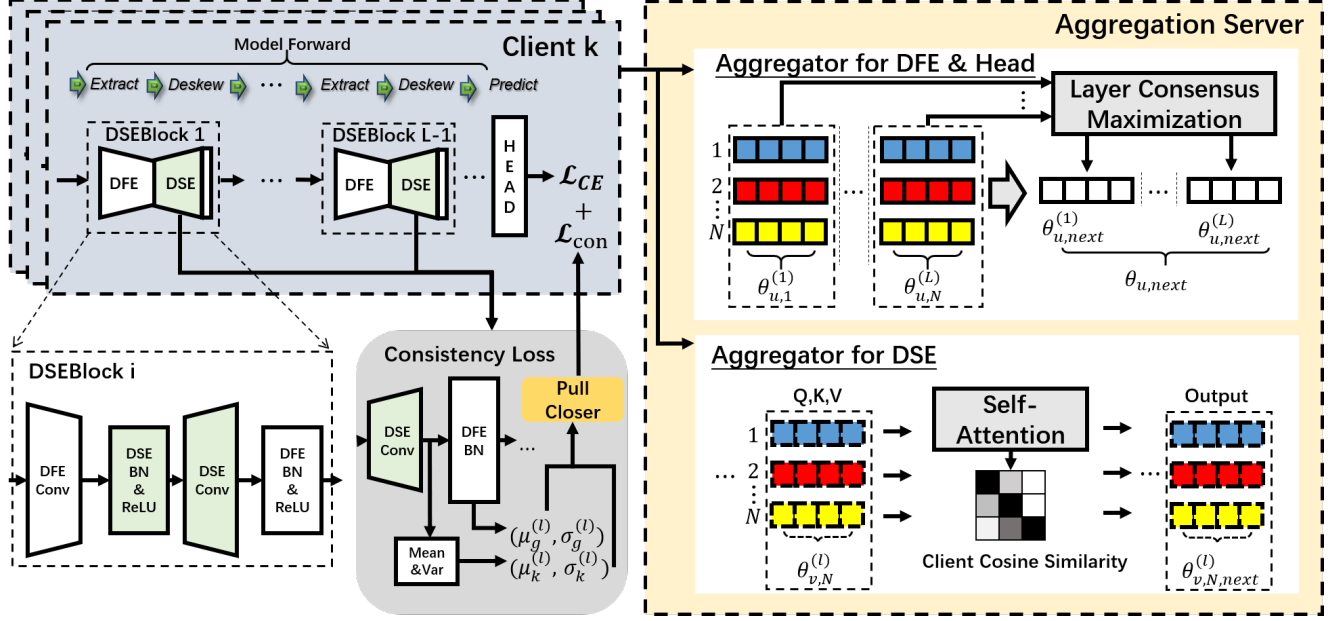
Figure 2. The overview of the FDSE framework.

the knowledge decoupling for the two modules via regularization term in Sec.4.2 and the parameter sharing strategy in Sec.4.3.

4.2. Consistency Regularization

To make the DSE module focus on erasing domain-specific skew for each client, we consider regularizing the DSE module's output features during clients' local training. Supposing that the lth DSE module $f_{\text{DSE}}^{(k,l)}$ in the model has ideally erased the domain skew for the kth client's local data, then the BN layer of the next DFE module $\text{BN}_{\text{DFE}}^{(k,l)}$ cannot infer the samples' domain via the distribution of the seen data when there is no label skew. Based on this insight, we manually pull the statistics of DSE's output to be close to the global statistics of the corresponding DFE module, thus enhancing the deskewing characteristic of the DSE module from the statistical view. Specifically, given the bth batch feature $\mathbf{X}_{k,b}^{(l)}, |\mathbf{X}_{k,b}^{(l)}| = B$ fed to the lth layer of the model, we first compute its statistics (i.e., mean and variance) by

$$\mu_{k,b}^{(l)} = \frac{1}{B} \sum_{i=1}^{B} \mathbf{X}_{k,b,i}^{(l)}, \sigma_{k,b}^{(l)2} = \frac{1}{B} \sum_{i=1}^{B} \left(\mathbf{X}_{k,b,i}^{(l)} - \mu_{k,b}^{(l)} \right)^2 \tag{4}$$

Then, we estimate the local running statistics by exponential averaging with the momentum coefficient γ of BN_{DFE} during local training

$$\hat{\mu}_{k,b}^{(l)} = \gamma \hat{\mu}_{k,b-1}^{(l)} + (1 - \gamma) \mu_{k,b}^{(l)}$$
$$\hat{\sigma}_{k,b}^{(l)2} = \gamma \hat{\sigma}_{k,b-1}^{(l)2} + (1 - \gamma) \sigma_{k,b}^{(l)2} \tag{5}$$

The local statistics of the lth layer are finally pulled close to the global ones by

$$\mathcal{L}_{\text{Con}}^{(l)} = \frac{1}{d} \left\| \hat{\mu}_{k,b}^{(l)} - \mu_g^{(l)} \right\|^2 + \left(\frac{\|\hat{\sigma}_{k,b}^{(l)2}\|_1 - \|\sigma_g^{(l)2}\|_1}{d} \right)^2 \tag{6}$$

where $\mu_g^{(l)}, \sigma_g^{(l)}$ are the corresponding layer's global statistics in the received global model and d is the feature dimension. In this way, both the centers and the sizes of the clients' feature spaces are regularized to be consistent. This rule is then applied to each DSE module in a depth-increasing manner, as formulated in the consistency regularization term

$$\mathcal{L}_{\text{Con}} = \sum_{l=1}^{L} w_l \mathcal{L}_{\text{Con}}^{(l)}, \text{w.r.t.,} \ w_l = \frac{\exp(\beta l)}{\sum_{l=1}^{L} \exp(\beta l)} \tag{7}$$

where L is the number of DSE modules and β is the hyperparameter that enables gradual features deskewing across layers. A smaller β can strengthen feature deskewing by emphasizing each layer's statistical consistency. We fix $\beta = 0.001$ in practice and only tune the coefficient λ that scales the regularization term to balance the task objective and the regularization. In addition, this consistency regularization can help adapt the trained model to new unseen domains without any labels, where we fine-tune the DSE modules to minimize the regularization loss as depicted in Sec. 5.4.

4.3. Model Aggregation

We use different strategies to aggregate the two fundamental components of FDSE (e.g., DFE and DSE). For the DFE

modules, we share them among all the clients and aggregate them through a fair layer consensus maximization mechanism. For the DSE modules, we personalized them for each client based on similarity-aware aggregation to erase heterogeneous domain skew collaboratively. As a result, the varying degrees of parameter sharing further facilitate the distinct knowledge acquisition of the DFE and DSE modules in a decoupled manner, as the local knowledge encoded in the DFE modules is much more frequently smoothed over all the clients than the DSE modules. We illustrate the details of the two aggregation strategies below.

4.3.1. Consensus Maximization

We remark that DFE modules are designed to model domain-agnostic knowledge. However, model updates from different clients may largely conflict with each other due to severe domain shifts [5, 10, 37], preventing the model from learning domain-agnostic knowledge by biasing the training process towards the dominant clients. To this end, we propose to maximize client consensus when aggregating fully shared parameters (e.g., DFE modules and the task head) to enhance domain-agnostic knowledge learning. Motivated by the multi-gradient-descent-algorihtm [7, 10] that optimizes the model update to be harmonies with each client's update, we minimize the L2-norm of the aggregated update for each layer to maximize layer-wise consensus. Specifically, given clients' local updates on DSE modules and heads $\{\Delta\theta_{u,k}^{(l)}\}, k \in [N]$, we aggregate them by

$$
\bar{d}_u^{(l)} = \frac{1}{N} \sum_{k=1}^{N} \|\Delta\theta_{u,k}^{(l)}\|_2, \quad \mathbf{d}_{u,k}^{(l)} = \frac{\Delta\theta_{u,k}^{(l)}}{\|\Delta\theta_{u,k}^{(l)}\|_2}
$$
$$
\Delta\theta_u^{(l)} = \bar{d}_u^{(l)} \sum_{k=1}^{N} u_k^{(l)} \mathbf{d}_{u,k}^{(l)}, \quad \mathbf{u}^{(l)} = \mathrm{argmin}_{\mathbf{u}} \| \sum_{k=1}^{N} u_k \mathbf{d}_{u,k}^{(l)} \|_2^2
$$
(8)

where $\Delta\theta_u^{(l)} \cdot \Delta\theta_{u,k}^{(l)} \geq 0, \forall k$ is guaranteed by the above optimization objective [7]. This aggregation scheme enhances the consensus in model updates across clients without sacrificing anyone's benefit. Besides, this process is independently repeated for different layers since simply mitigating conflicts at the model level cannot prevent the final update from favoring parts of clients at some layers, reducing the layer utility for those clients [28].

4.3.2. Similarity-aware Personalization

To enable collaboratively erasing domain skew for clients from similar domains, we aggregate each DSE module's parameters with a self-attention module on the server side based on clients' similarity. Given clients's parameters of the lth DSE module $\{\theta_{v,k}^{(l)}, k \in [N]\}$, we aggregate them by

---

Algorithm 1 Federated Domain Shift Eraser

Input: The global model $\mathcal{M}$, the number of local epochs E, and the learning rate η_t

1: Decompose the initial model into globally shared parameters θ_u^0 and personalized parameters θ_v^0 in Sec. 4.1

2: Initialize clients' personalized parameters $\theta_{v,k}^0 = \theta_v^0, \forall k \in [N]$

3: **for** communication round $t = 0, 1, ..., T-1$ **do**

4:     The server broadcasts the model $\theta_k^t = (\theta_u^t, \theta_{v,k}^t)$ to each client k.

5:     **for** each client $k \in [N]$ **do**

6:         **for** each iteration $i = 0, 1, ..., E-1$ **do**

7:             Compute loss $\mathcal{L}_k = \mathcal{L}_{\mathrm{task}} + \lambda\mathcal{L}_{\mathrm{Con}}$

8:             $\theta_{k,i+1}^t \leftarrow \theta_{k,i}^t - \eta_t \nabla\mathcal{L}_k(\theta_{k,i}^t)$

9:         **end for**

10:         Client k send the model parameters $\theta_k^{t+1} = \theta_{k,E}^t$ to the server.

11:     **end for**

12:     The server respectively aggregates the received globally shared model parameters θ_u^{t+1} by Eq. (8) and personalized model parameters $\left[\theta_{v,k}^{t+1}\right]^\top$ by Eq. (9)

13: **end for**

$$
\mathbf{Q}_l = \mathbf{K}_l = \left[\frac{\mathbf{V}_{lk}}{\|\mathbf{V}_{lk}\|_2}\right]^\top \Big|_{k=1 \text{ to } N}, \mathbf{V}_l = \left[\theta_{v,1}^{(l)}, \cdots, \theta_{v,N}^{(l)}\right]^\top
$$
$$
\left[\theta_{v,1,\mathrm{next}}^{(l)}, \cdots, \theta_{v,N,\mathrm{next}}^{(l)}\right]^\top = \mathrm{softmax}\left(\frac{\mathbf{Q}_l\mathbf{K}_l^\top}{\tau}\right)\mathbf{V}_l \quad (9)
$$

where τ is the temperature parameter controlling the degree of personalization of DSE modules (e.g., smaller τ corresponds to a higher personalization degree and stronger skew elimination). We also perform the personalized aggregation independently for each layer's DSE module, since the optimal personalization degree may vary across layers [23]. For the non-trainable statistical parameters, we did not aggregate them for $\mathrm{BN_{DSE}}$ and directly average them for $\mathrm{BN_{DFE}}$. The pseudo-code of FDSE is summarized in Algorithm 1.

5. Experiments

5.1. Setup

Dataset & Model. We evaluate our method on three popular multi-domain image classification tasks: Office-Caltech10 [32], DomainNet [16], and PACS [42]. We follow [20, 43] to use AlexNet as the backbone for these datasets and allocate a single domain's data to each client respectively for all three datasets.

Method	Domainnet		Office-Caltech10		PACS	
	ALL	AVG	ALL	AVG	ALL	AVG
Local	$57.10_{\pm 0.32}$	$52.96_{\pm 0.33}$	$64.47_{\pm 2.52}$	$62.72_{\pm 7.81}$	$61.29_{\pm 2.47}$	$57.16_{\pm 2.85}$
FedAvg (AISTATS 2017)	$69.17_{\pm 0.46}$	$67.53_{\pm 0.41}$	$82.60_{\pm 3.14}$	$86.26_{\pm 2.54}$	$74.30_{\pm 1.90}$	$72.10_{\pm 1.42}$
LG-FedAvg (NIPSW 2019)	$71.13_{\pm 0.30}$	$67.97_{\pm 0.32}$	$82.69_{\pm 0.53}$	$87.29_{\pm 1.32}$	$79.11_{\pm 0.69}$	$76.72_{\pm 0.54}$
FedProx (MLSys 2020)	$68.81_{\pm 0.71}$	$67.47_{\pm 0.66}$	$82.69_{\pm 1.52}$	$87.36_{\pm 1.87}$	$74.38_{\pm 1.55}$	$72.33_{\pm 1.53}$
Scaffold (ICML 2020)	$70.41_{\pm 0.40}$	$69.06_{\pm 0.43}$	$80.86_{\pm 2.43}$	$85.87_{\pm 1.87}$	$76.30_{\pm 0.93}$	$74.26_{\pm 0.95}$
FedDyn (ICLR 2021)	$70.02_{\pm 0.57}$	$68.86_{\pm 0.54}$	$82.77_{\pm 1.82}$	$87.80_{\pm 1.96}$	$74.92_{\pm 1.29}$	$73.19_{\pm 1.01}$
MOON (CVPR 2021)	$68.35_{\pm 0.32}$	$66.65_{\pm 0.29}$	$80.12_{\pm 1.86}$	$82.48_{\pm 1.71}$	$75.00_{\pm 0.32}$	$72.13_{\pm 0.32}$
Ditto (ICML 2021)	$75.18_{\pm 0.37}$	$72.82_{\pm 0.35}$	$84.12_{\pm 1.32}$	$88.72_{\pm 1.28}$	$82.02_{\pm 1.32}$	$80.03_{\pm 1.37}$
PartialFed (NIPS 2021)	$75.92_{\pm 0.24}$	$73.46_{\pm 0.41}$	$82.70_{\pm 2.12}$	$88.33_{\pm 2.17}$	$81.22_{\pm 0.98}$	$79.18_{\pm 1.09}$
FedBN (ICLR 2021)	$74.75_{\pm 0.24}$	$72.25_{\pm 0.20}$	$83.08_{\pm 1.84}$	$87.01_{\pm 1.30}$	$81.58_{\pm 0.79}$	$79.47_{\pm 0.69}$
FedFA (ICLR 2023)	$69.47_{\pm 0.29}$	$67.60_{\pm 0.29}$	$82.98_{\pm 2.84}$	$86.69_{\pm 3.02}$	$75.44_{\pm 0.71}$	$73.60_{\pm 0.86}$
FedHeal (CVPR 2024)	$69.43_{\pm 0.71}$	$67.96_{\pm 0.65}$	$81.73_{\pm 3.19}$	$86.29_{\pm 2.71}$	$75.46_{\pm 0.84}$	$73.51_{\pm 0.83}$
FDSE (Ours)	$\mathbf{76.77_{\pm 0.41}}$	$\mathbf{74.50_{\pm 0.40}}$	$\mathbf{87.15_{\pm 2.06}}$	$\mathbf{91.58_{\pm 2.01}}$	$\mathbf{83.81_{\pm 1.70}}$	$\mathbf{82.17_{\pm 1.49}}$

Table 1. Comparison of model testing accuracy (%)↑ on DomainNet, Office-Caltech10, and PACS datasets. The optimal results are marked by **bold**. **ALL** refers to the model testing accuracy on all clients' local testing samples and **AVG** refers to the mean of of clients' local testing accuracies. Each result is averaged over 5 trials with different fixed random seeds.

Baselines. We compare our method with four types of baselines: 1) Local training only; 2) Vallina FL with data heterogeneity: FedAvg [24], FedProx [18], Scaffold [15], FedDyn [1], MOON [17] 3) Consensus-based FL: Fed-Heal [5], FedFA [43]; 4) Personalized FL: LG-FedAvg [21], FedBN [20], PartialFed [32].

Hyper-parameters. We tune the learning rate $\eta \in \{0.001, 0.01, 0.05, 0.1, 0.5\}$ by grid search for each method. The batch size is fixed to 50 and the local epochs for Domainnet, Office-Caltech10, and PACS are respectively 5, 1, and 5. We run each trial for 500 communication rounds with the learning rate decay ratio 0.998 per round. All the methods' algorithmic hyper-parameters are respectively tuned to their optimal. More details on the hyper-parameter setting are in the supplementary materials.

5.1.1. Implementation

All our experiments are run on a 64 GB-RAM Ubuntu 22.04.3 server with Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz and 2 NVidia(R) RTX4090 GPUs. All code is implemented in PyTorch 1.12.0 and FLGo 0.3.29 [38].

5.2. Comparison with Baselines

Overall Performance. Table 1 compares our proposed FDSE with several baselines. Notably, FDSE consistently achieves optimal results across all settings, demonstrating a significant advantage in enhancing model performance in cross-domain FL. Additionally, we observe that personal-
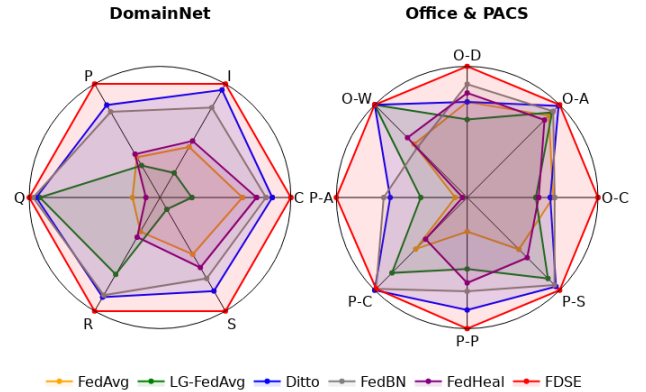

Figure 3. Evaluation results on individual domains (i.e., clients) across the three datasets. Each axis represents the result for a specific domain and is scaled by the axis's maximum value for clarity.

ized methods (e.g., LG-FedAvg, Ditto, and FedBN) outperform almost all non-personalized approaches (e.g., FedAvg, FedHeal, and FedFA). This underscores the importance of incorporating domain-specific knowledge for each domain. Furthermore, enhancing client consensus also contributes to performance improvement, as evidenced by FedHeal and FedFA consistently outperforming FedAvg in nearly all settings. Thus, we attribute the superiority of FDSE to the success in integrating the complementary advantages of both consense-based and personalization-based methods.
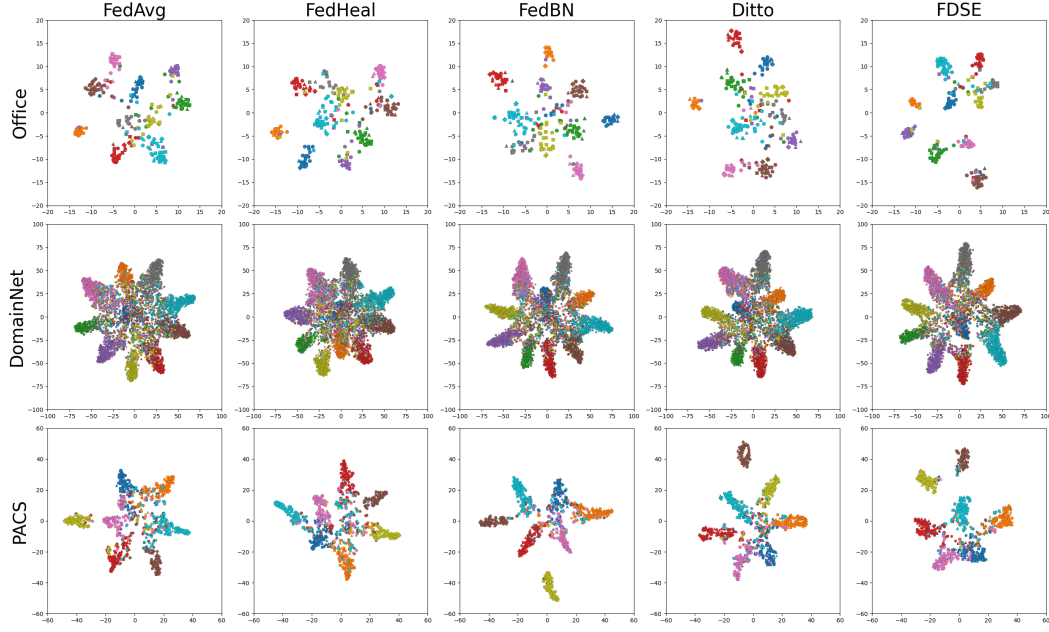
Figure 4. T-SNE visualization for representation space of different methods on Office-Caltech10, DomainNet, and PACS. Each color represents one class of samples and each shape represents one domain.
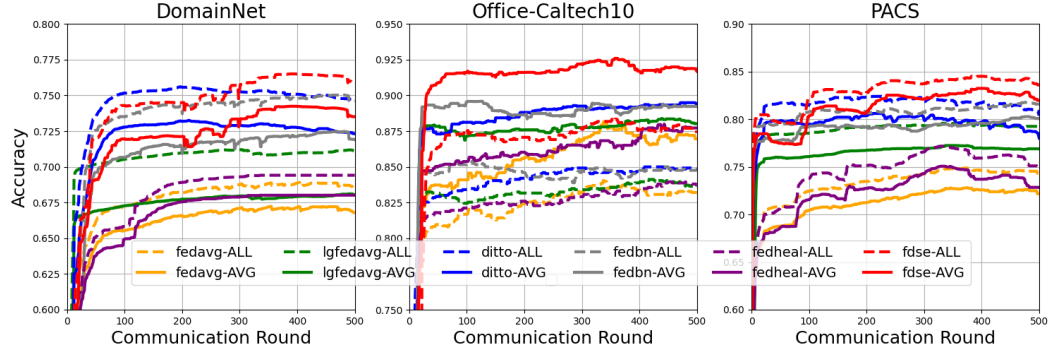

Figure 5. Testing accuracy v.s. communication rounds.

Individual Performance. Figure 3 illustrates the individual performance of clients for different methods, where a larger area under a method's curve indicates better performance. FDSE outperforms all other methods across nearly all clients, confirming its effectiveness across all domains rather than just specific ones.

Convergence. We plot the testing accuracy across communication rounds in Figure 5. While Ditto demonstrates a slightly faster convergence speed than FDSE in the early stages of training on DomainNet and PACS, FDSE ultimately achieves higher performance as training progresses.

5.3. T-SNE Visualization

Figure 4 presents the t-SNE visualization analysis of the representation space (i.e., the feature space before the out-

put layer of the model) for different methods. On one hand, FDSE increases the inter-class distances of samples, with a greater separation between samples of different colors compared to other baselines. On the other hand, FDSE reduces the intra-class distance among samples of the same color, resulting in tighter color clusters than those observed in other baselines. This confirms FDSE's ability to enhance model performance while reducing domain skew.

5.4. Generalizability to Unseen Domains

We evaluate the adaptability of FDSE to unseen domains in Table 2. Each column in Table 2 represents a target client that did not participate in the training process, and we evaluate the model trained on other clients on the target after adaptation. We emphasize that the model adaptation processes for all methods are label-free, as we detailed in the

Method	Office-Caltech10					DomainNet						
	C	A	D	W	AVG	C	I	P	Q	R	S	AVG
FedAvg	51.78	70.52	80.00	65.51	66.95	62.81	30.15	55.53	48.86	59.74	58.92	52.66
FedBN	**60.71**	70.52	80.00	55.17	66.60	62.56	31.14	57.26	53.06	63.17	62.46	54.94
- align	40.17	68.42	86.66	72.41	66.91	60.53	31.57	54.33	50.60	59.85	62.59	53.24
FDSE	57.14	**75.78**	**93.33**	**75.86**	**75.52**	**65.22**	**32.34**	**59.32**	**55.00**	**64.28**	**65.28**	**56.91**

Table 2. Model performance ($\uparrow$) of adapting the trained model to different unseen clients. The optimal results are marked by **bold**.
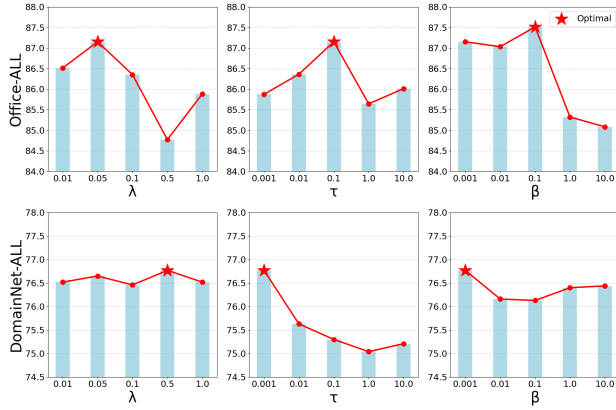

Figure 6. The impact of hyper-parameters on model performance.

Module			DomainNet		Office		PACS	
A	B	C	ALL	AVG	ALL	AVG	ALL	AVG
$\times$	$\times$	$\times$	74.63	72.19	83.94	86.13	82.46	80.72
$\times$	$\checkmark$	$\checkmark$	75.66	73.23	84.44	88.54	82.92	80.97
$\checkmark$	$\times$	$\checkmark$	76.57	74.32	85.23	89.25	83.40	81.73
$\checkmark$	$\checkmark$	$\times$	76.49	74.24	86.43	90.38	83.40	81.78
$\checkmark$	$\checkmark$	$\checkmark$	**76.77**	**74.50**	**87.15**	**91.58**	**83.81**	**82.17**

Table 3. The ablation study of the model performance of FDSE with each submodule to be respectively removed.

Method	$\text{Num}_{\times 10^7}$	Comm.	$\text{FLOPs}_{\times 10^{10}}$	Time_{train}
FedBN	1.30	49.52M	4.41	4.03s
Ditto	1.30	49.52M	4.41	6.59s
FDSE	0.65	24.87M	2.24	4.27s

Table 4. Comparison of communication and communication costs.

tuned while β can be fixed at a low value.

5.6. Ablation Study

We conduct the ablation analysis of the effectiveness of FDSE's modules in Table 3. Module A, B, and C respectively correspond to the consensus-maximization aggregation in Sec.4.3.1, similarity-aware aggregation in Sec.4.3.2, and consistency regularization in Sec.4.2. The results show that the raw architecture of FDSE has achieved comparable results with FedBN. Further, the performance will be degraded after removing each module and the optimal result appears in the full participation of each module, indicating the collaborative effectiveness of these modules.

5.7. Efficiency

Table 4 compares the communication and computation costs. Our FDSE saves more communication efficiency per round (e.g., Comm.) and computation efficiency (e.g., FLOPs) than baseline due to its decomposition-based architecture. Besides, FDSE achieves competitive time costs of local training against FedBN, where the additional training costs of FDSE mainly come from the regularization term.

6. Conclusion

In this work, we rethink the domain shift problem in FL from a hybrid view that integrates the advantages of personalization-based methods and consensus-based methods. We develop a novel framework, FDSE, to differently erase domain skew for each client while maximizing their consensus. Specifically, we efficiently formulate the model forward passing as an iterative deskewing process that extracts and then deskews features alternatively via layer decomposition. Further, we fine-grained design aggregation strategies and the regularization term to enhance the knowledge decoupling, leading to improved consistency in the representation space. We plan to extend this framework to more applications and model architectures in the future.

supplementary materials. **- align** refers to directly evaluating the model trained by FDSE without adaptation. Our FDSE achieves comparable results to the baselines even before model adaptation. After adaptation, FDSE significantly enhances model performance for most clients, confirming FDSE's ability to generalize effectively to unseen domains.

5.5. Impact of Hyper-parameters

Effect of λ. As shown in the first column of Figure 6, slightly increasing λ can improves the accuracy while a too large λ will cause performance reduction, indicating that λ should be carefully tuned to their optimal in practice.

Effect of β and τ. From the last two columns of Figure 6, the effects of both parameters exhibit a similar trend on each dataset as their values change. We attribute this similarity to the varying degrees of domain skew presented in each dataset, since small values of both τ and β strengthen domain skew elimination. The performance is more sensitive to τ than to β, where a small β is preferred while the optimal τ values differ, suggesting that τ should be properly

Acknowledgement

References

[1] Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. *arXiv preprint arXiv:2111.04263*, 2021. 2, 6

[2] Robert J Barro and Xavier Sala-i Martin. Convergence. *Journal of political Economy*, 100(2):223–251, 1992. 3

[3] Hong-You Chen and Wei-Lun Chao. On bridging generic and personalized federated learning for image classification. *arXiv preprint arXiv:2107.00778*, 2021. 2

[4] Yiqiang Chen, Xin Qin, Jindong Wang, Chaohui Yu, and Wen Gao. Fedhealth: A federated transfer learning framework for wearable healthcare. *IEEE Intelligent Systems*, 35 (4):83–93, 2020. 1

[5] Yuhang Chen, Wenke Huang, and Mang Ye. Fair federated learning under domain skew with local consistency and domain diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12077–12086, 2024. 1, 2, 3, 5, 6

[6] Rong Dai, Xun Yang, Yan Sun, Li Shen, Xinmei Tian, Meng Wang, and Yongdong Zhang. Fedgamma: Federated learning with global sharpness-aware minimization. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 2

[7] Jean-Antoine Désidéri. *Multiple-gradient descent algorithm (MGDA)*. PhD thesis, INRIA, 2009. 5

[8] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in Neural Information Processing Systems*, 33:3557–3568, 2020. 2

[9] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu. Ghostnet: More features from cheap operations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1580–1589, 2020. 3

[10] Zeou Hu, Kiarash Shaloudegi, Guojun Zhang, and Yaoliang Yu. Federated learning meets multi-objective optimization. *IEEE Transactions on Network Science and Engineering*, 9 (4):2039–2051, 2022. 1, 5

[11] Wenke Huang, Mang Ye, Zekun Shi, He Li, and Bo Du. Rethinking federated learning with domain shift: A prototype view. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16312–16322. IEEE, 2023. 1, 2

[12] Muhammad Akbar Husnoo, Adnan Anwar, Nasser Hosseinzadeh, Shama Naz Islam, Abdun Naser Mahmood, and Robin Doss. Fedrep: towards horizontal federated load forecasting for retail energy providers. In *2022 IEEE PES 14th Asia-Pacific Power and Energy Engineering Conference (APPEEC)*, pages 1–6. IEEE, 2022. 2, 3

[13] Meirui Jiang, Zirui Wang, and Qi Dou. Harmofl: Harmonizing local and global drifts in federated learning on heterogeneous medical images. In *Proceedings of the AAAI Confer-*

[14] 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. 2

[15] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143. PMLR, 2020. 2, 6

[16] Aristotelis Leventidis, Laura Di Rocco, Wolfgang Gatterbauer, Renée J Miller, and Mirek Riedewald. Domainnet: Homograph detection and understanding in data lake disambiguation. *ACM Transactions on Database Systems*, 48(3): 1–40, 2023. 5

[17] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10713–10722, 2021. 2, 6

[18] 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. 2, 6

[19] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, pages 6357–6368. PMLR, 2021. 2

[20] Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. *arXiv preprint arXiv:2102.07623*, 2021. 1, 2, 3, 5, 6

[21] Paul Pu Liang, Terrance Liu, Liu Ziyin, Nicholas B Allen, Randy P Auerbach, David Brent, Ruslan Salakhutdinov, and Louis-Philippe Morency. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523*, 2020. 2, 3, 6

[22] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1013–1023, 2021. 2

[23] Xiaosong Ma, Jie Zhang, Song Guo, and Wenchao Xu. Layer-wised model aggregation for personalized federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10092–10101, 2022. 5

[24] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017. 1, 2, 6

[25] Yiqun Mei, Pengfei Guo, Mo Zhou, and Vishal Patel. Resource-adaptive federated learning with all-in-one neural

ence on Artificial Intelligence, pages 1087–1095, 2022. 2, 3

composition. In *Advances in Neural Information Processing Systems*, pages 4270–4284. Curran Associates, Inc., 2022. 3

[26] Matias Mendieta, Taojiannan Yang, Pu Wang, Minwoo Lee, Zhengming Ding, and Chen Chen. Local learning matters: Rethinking data heterogeneity in federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8397–8406, 2022. 2

[27] Jaehoon Oh, Sangmook Kim, and Se-Young Yun. Fedbabu: Towards enhanced representation for federated image classification. *arXiv preprint arXiv:2106.06042*, 2021. 2

[28] Zibin Pan, Chi Li, Fangchen Yu, Shuyi Wang, Haijin Wang, Xiaoying Tang, and Junhua Zhao. Fedlf: Layer-wise fair federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14527–14535, 2024. 5

[29] Xingchao Peng, Zijun Huang, Yizhe Zhu, and Kate Saenko. Federated adversarial domain adaptation. *arXiv preprint arXiv:1911.02054*, 2019. 2

[30] Zhe Qu, Xingyu Li, Rui Duan, Yao Liu, Bo Tang, and Zhuo Lu. Generalized federated learning via sharpness aware minimization. In *International conference on machine learning*, pages 18250–18280. PMLR, 2022. 2

[31] Aviv Shamsian, Aviv Navon, Ethan Fetaya, and Gal Chechik. Personalized federated learning using hypernetworks. In *International Conference on Machine Learning*, pages 9489–9502. PMLR, 2021. 2

[32] Benyuan Sun, Hongxing Huo, Yi Yang, and Bo Bai. Partialfed: Cross-domain personalized federated learning via partial initialization. *Advances in Neural Information Processing Systems*, 34:23309–23320, 2021. 2, 3, 5, 6

[33] Yan Sun, Li Shen, Tiansheng Huang, Liang Ding, and Dacheng Tao. Fedspeed: Larger local interval, less communication round, and higher generalization accuracy, 2023. 2

[34] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 3

[35] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021. 3

[36] Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. Fedproto: Federated prototype learning across heterogeneous clients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8432–8440, 2022. 2, 3

[37] Zheng Wang, Xiaoliang Fan, Jianzhong Qi, Chenglu Wen, Cheng Wang, and Rongshan Yu. Federated learning with fair averaging. *arXiv preprint arXiv:2104.14937*, 2021. 1, 5

[38] Zheng Wang, Xiaoliang Fan, Zhaopeng Peng, Xueheng Li, Ziqi Yang, Mingkuan Feng, Zhicheng Yang, Xiao Liu, and Cheng Wang. Flgo: A fully customizable federated learning platform. *arXiv preprint arXiv:2306.12079*, 2023. 6

[39] Zhimin Yuan, Wankang Zeng, Yanfei Su, Weiquan Liu, Ming Cheng, Yulan Guo, and Cheng Wang. Density-guided translator boosts synthetic-to-real unsupervised domain adaptive segmentation of 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23303–23312, 2024. 2, 3

[40] Jianqing Zhang, Yang Hua, Hao Wang, Tao Song, Zhengui Xue, Ruhui Ma, Jian Cao, and Haibing Guan. Gpfl: Simultaneously learning global and personalized feature information for personalized federated learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5041–5051, 2023. 2

[41] 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. 2

[42] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13025–13032, 2020. 5

[43] Tailin Zhou, Jun Zhang, and Danny HK Tsang. Fedfa: Federated learning with feature anchors to align features and classifiers for heterogeneous data. *IEEE Transactions on Mobile Computing*, 2023. 2, 3, 5, 6