DoBlock: Blocking Malicious Association Propagation for Backdoor-Robust Federated Learning under Domain Skew

Zhou Tan[1], De Li[2], Yirui Huang[1], Duanshu Fang[2], Jia-Li Yin[1*], Xiaolei Liu[3], Songze Li[4], Shouling Ji[5]

[1] College of Computer and Data Science, Fuzhou University, Fuzhou, China
[2] School of Computer Science and Engineering, Guangxi Normal University, Guilin, China
[3] National Interdisciplinary Research Center of Engineering Physics, Mianyang, China
[4] School of Cyber Science and Engineering, Southeast University, Nanjing, China
[5] College of Computer Science and Technology, Zhejiang University, Hangzhou, China
tanzhou_obius@163.com, lide@stu.gxnu.edu.cn, hyr642591116@gmail.com, fds19918@outlook.com
jlyin@fzu.edu.cn, luxaole@gmail.com, songzeli@seu.edu.cn, sji@zju.edu.cn

Abstract

Federated Learning (FL) enables privacy-preserving distributed training but remains vulnerable to backdoor attacks. Attackers can embed malicious trigger-label associations into the global model by participating in the aggregation process. Existing defense methods typically defend against backdoor attacks by detecting and filtering malicious updates that deviate from benign ones. However, we find that these defenses fail under domain skew, where differing feature distributions across clients increase update heterogeneity, making it harder to distinguish malicious updates from benign ones. To address this challenge, we propose **DoBlock**, a novel defense that utilizes an aggregatable domain infuser incapable of embedding malicious associations, through federated training to facilitate cross-domain knowledge sharing. Moreover, DoBlock prevents malicious association propagation by isolating local models from aggregation, as local models remain client-specific and rely solely on local data for training. Experiments on five domain skew datasets (Digits, PACS, VLCS, Office-Caltech10, and DomainNet) show that DoBlock maintains attack success rates below 2.5%, while achieving the highest main task accuracy, demonstrating superior robustness without sacrificing benign performance.

Introduction

Federated Learning (FL) has emerged as a transformative approach in distributed machine learning, enabling collaborative model training across diverse clients, such as smartphones, hospitals, or IoT devices, without requiring the centralization of sensitive data (McMahan et al. 2017; Tan et al. 2023; Li et al. 2024, 2025; Lin, Tan, and Liu 2025). By keeping data localized, FL addresses privacy concerns while harnessing the computational power and data diversity of edge devices, making it a promising solution for domains like healthcare (e.g., personalized diagnostics) (Qian et al. 2025), financial modeling (e.g., fraud detection) (Abdul Salam et al. 2024), and smart infrastructure (e.g., edge computing) (Wu et al. 2024). Despite its advantages, the distributed nature of FL introduces unique security challenges, with backdoor attacks standing out as a critical threat (Gu et al. 2019; Bagdasaryan et al. 2020; Tan et al. 2025a,b). These attacks involve attackers injecting hidden triggers into the global model through their local updates, causing the model to misclassify specific inputs while maintaining normal behaviour on benign data. This stealthy manipulation undermines trust in FL systems.

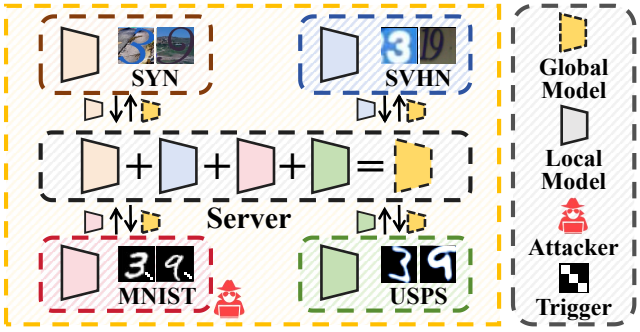
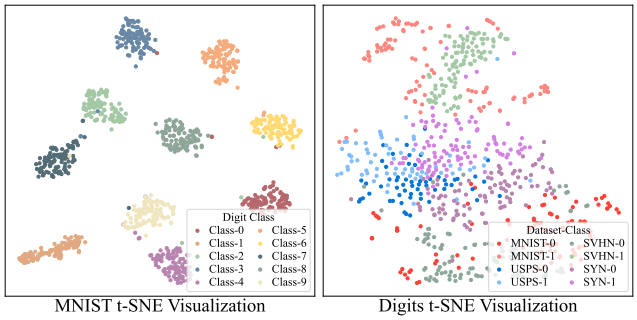


Figure 1: **Problem Illustration** of backdoor attacks in FL under domain skew. **i)** Four clients process data from the **Digits** dataset's distinct domains (MNIST, USPS, SYN, SVHN), exhibiting pronounced feature divergence (e.g., stylistic variations in digits). **ii)** Attackers inject **common triggers** (e.g., pixel patterns) that leverage this domain divergence to camouflage malicious updates.

To mitigate backdoor attacks, defense mechanisms in FL have been widely studied, such as distance-based defense (Blanchard et al. 2017; Huang et al. 2023a), statistical distribution defense of client updates (Pillutla, Kakade, and Harchaoui 2022; Xu, Zhang, and Hu 2025b), and model refinement defense (Xie et al. 2021; Huang et al. 2023b). These methods typically rely on a strong assumption: client data is homogeneous, originating from a similar domain or sharing stylistic consistency. Under this assumption, malicious updates can be identified as outliers, or their impact can be minimized through aggregation methods. However, real-world FL scenarios frequently exhibit domain

*Corresponding Author.



MNIST t-SNE Visualization Digits t-SNE Visualization

Figure 2: t-SNE visualization of domain skew: MNIST (distinct class clusters) and Digits (skew-induced chaotic overlap). Please see Appendix A[1] for implementation details.

skew, where samples of the same class exhibit different features across different clients due to variations in data collection methods, devices, and other factors, as illustrated in Figure 1. Unlike label distribution skew, which causes inter-client class imbalance, domain skew directly disrupts feature-level consistency within shared classes, degrading model decision boundaries and performance.

This skew compromises the efficacy of the global model and introduces inconsistencies in client updates. As a result, it alters the core representations learned by the global model. To examine this, we analyse feature embeddings using t-SNE on the homogeneous MNIST dataset and the heterogeneous Digits dataset. As illustrated in Figure 2, MNIST features form ten distinct clusters corresponding to digit classes, whereas Digits features, even when limited to two classes, overlap chaotically without class-specific clustering. This natural heterogeneity causes distance-based and statistical distribution defenses to misclassify benign updates as outliers. Simultaneously, it allows poisoned updates to blend in with the heterogeneous update distribution, thereby obscuring the distinction between malicious and benign updates. Moreover, attackers can exploit this skew by crafting triggers that mimic legitimate variations (e.g., device noise or stylistic variations). This allows them to conceal malicious updates within the natural diversity of client data, thereby evading detection by traditional defenses. Consequently, identifying attackers becomes challenging, which inevitably degrades defense performance, as demonstrated in Tables 2, 3, and 4.

These observations reveal a critical gap in FL security: existing defenses are inadequately prepared to counter backdoor attacks under domain skew scenarios. Motivated by this limitation, we propose DoBlock, a novel defense framework designed to block backdoor attacks that exploit domain skew. We identify that the backdoor attack's success in FL depends on establishing malicious associations in global models. Attackers achieve this through combining trigger injection and label manipulation, propagating poisoned updates through the aggregation. To address this challenge, DoBlock introduces an aggregatable domain infuser incapable of embedding malicious associations, through federated training to facilitate cross-domain knowledge sharing. Crucially, only this domain infuser is shared and aggregated across clients, while the client's local model, responsible for mapping features to labels, remains private and excluded from aggregation. This separation blocks the propagation of malicious associations to benign clients, thereby strengthening FL against backdoor attacks while maintaining collaborative efficacy. The main contributions are summarized as:

- **Vulnerability Discovery:** We reveal FL's backdoor vulnerability, which malicious trigger-label associations propagate through model aggregation. Through rigorous ablation studies, we demonstrate that disrupting either attack component (trigger injection or label manipulation) effectively defends against backdoors.
- **Novel Defense Framework:** We propose DoBlock. To the best of our knowledge, it is the first backdoor-robust FL specifically designed for domain skew. DoBlock isolates local models from aggregation to prevent malicious association propagation while enabling cross-domain knowledge sharing through its domain infuser.
- **Comprehensive Evaluation:** We conduct extensive experiments across multiple datasets, validating the efficacy of DoBlock against various backdoor attacks. Compared to existing state-of-the-art defense methods, DoBlock exhibits superior robustness.

Related Work

Domain Skew in FL. Research addressing domain skew in FL has focused on generalization. FedHEAL (Chen, Huang, and Ye 2024) selectively discards unimportant updates and employs a fair aggregation objective to ensure unbiased global model convergence and equitable performance across domains. ELCFS (Liu et al. 2021) and CCST (Wang, Guo, and Tang 2024) exchange frequency-domain amplitudes or style statistics across clients, incurring privacy leaks and high communication overhead. FedKA (Sun, Chong, and Ochiai 2023), a server-side pseudo-label voting that amplifies computational costs.FedDG-GA (Zhang et al. 2023) introduces variance reduction for fairness but complicates aggregation with dynamic weight calibration. To mitigate these limitations, gPerXAN (Le et al. 2024) introduces a lightweight architectural solution using personalized eXplicitly assembled normalization and a guiding regularizer, filtering domain-specific features. However, domain skew intrinsically elevates backdoor vulnerability by amplifying client update heterogeneity. Existing generalization-oriented methods fail to mitigate this security threat as they lack mechanisms to detect malicious updates.

Backdoor Attacks in FL. FL is vulnerable to backdoor attacks (Bhagoji et al. 2019; Shen et al. 2025). Pixel-level attacks deploy client-specific triggers on local data (Gu et al. 2019; Xie et al. 2019; Barni, Kallas, and Tondi 2019; Liu et al. 2024), while model replacement attacks (Bagdasaryan et al. 2020; Kumar, Mohan, and Cenkeramaddi 2024) hijack global models by scales malicious updates. In contrast, Sybil attacks compromise the system via collusive clients (Fung, Yoon, and Beschastnikh 2018). Neurotoxin (Zhang et al.

[1]The appendix at https://github.com/obius-coder/DoBlock.

2022) enhances stealth and effect by selectively updating benign gradients' least active parameters. Further concealment is achieved through clean-label attacks (Turner, Tsipras, and Madry 2018; Huynh et al. 2024), which embed triggers in target-class samples without altering labels. Our goal is to defend against all these attacks.

Defending Against Backdoor Attacks in FL. To deal with backdoor attacks, existing defense methods can be categorized into: **i)** Distance-based Defense isolates malicious clients by evaluating update discrepancies (Blanchard et al. 2017; Fung, Yoon, and Beschastnikh 2018; Shejwalkar and Houmansadr 2021; Huang et al. 2023a). **ii)** Statistics Distribution Defense employs statistical methods to neutralize attacks (Nguyen et al. 2022; Pillutla, Kakade, and Harchaoui 2022; Xu, Zhang, and Hu 2025a,b). **iii)** Model Refinement Defense enhances resilience through architectural adjustments, including gradient smoothing (Xie et al. 2021), ensemble distillation (Huang, Ye, and Du 2022), network pruning (Huang et al. 2023b). While these methods improve FL's backdoor robustness, they neglect that domain skew conflates benign updates with malicious updates, causing defense failure. This work specifically tackles this challenge.

Methodology

Preliminaries

Following typical FL setup (McMahan et al. 2017; Huang et al. 2024b), we consider a system with M clients (indexed by m), each client m holds a local model $\boldsymbol{w}_m$ and maintains a private dataset $D_m = \{(x_i, y_i)\}_{i=1}^{N_m}$, where N_m represents the data size of the client m. The optimization objective in FL is to minimize a global loss function, defined as a weighted sum of local loss functions across all clients:

$$\min_{\boldsymbol{w}} \frac{1}{M} \sum_{m=1}^{M} \frac{1}{N_m} \sum_{i=1}^{N_m} \mathcal{L}(f(x_i^m; \boldsymbol{w}_m), y_i^m), \tag{1}$$

where $f(\cdot; \boldsymbol{w})$ is the model function, and $\mathcal{L}$ is the loss function (e.g., cross-entropy for classification). The training process involves iterative communication rounds, where clients compute local updates and send them to a central server for aggregation. We define domain skew as the phenomenon where data across clients originates from distinct domains, leading to variations in the conditional feature distribution $P(x|y)$ across clients. This means that the same label may have different features across clients. For example, the same digit "2" may have different handwriting styles (e.g., curved vs. straight-line versions) across clients due to regional writing conventions.

Threat Model

We explore the scenario where there are attackers in the FL system. It is a plausible assumption that attackers intend to inject a backdoor into the benign client's local model by manipulating their local training while obeying the FL protocol.

Attacker Capability. As a client in the FL system, the attacker has access to its data, model, and the gradients computed during the training. For an input pair (x, y), the attacker creates a poisoned version of the data as follows:

$$\hat{x} = (1 - \delta) \odot x + \delta \odot \Phi, \tag{2}$$

Configurations		Digits		PACS	
Trigger Φ	Target $\hat{y}$	$\mathcal{MA}$	$\mathcal{ASR}$	$\mathcal{MA}$	$\mathcal{ASR}$
✔	✖	80.72	1.39	83.55	3.75
✖	✔	76.02	10.23	74.70	22.01
✔	✔	81.99	99.39	85.76	100.00
✖	✖	**82.06**	**0.88**	**86.38**	**2.73**

Table 1: Ablation study on trigger Φ injection and target label $\hat{y}$ manipulation. Please see details in Appendix B.

where δ is a binary mask defining the trigger's location, Φ represents the trigger pattern (e.g., a specific pixel pattern), and $\odot$ denotes element-wise multiplication. The poisoned data $\hat{x}$ is then paired with a maliciously assigned target label $\hat{y}$, and this poisoned set is blended with clean data to train the local model. However, the attacker cannot directly access or alter the data or labels of other clients in the FL system.

Attacker Objective. The attacker aims to achieve a high attack success rate ($\mathcal{ASR}$), ensuring that triggered inputs are consistently misclassified as the target label during inference, while also maintaining a high main task accuracy ($\mathcal{MA}$) to preserve model performance on clean data.

Motivation

To identify why backdoor attacks succeed in FL, we conduct an ablation study on the attacker's poisoning phase across domain skew datasets Digits (Zhou et al. 2020) and PACS (Li et al. 2017) under FedAvg, isolating trigger injection and label manipulation impacts, as shown in Table 1.

i) Trigger Injection Only: When the attacker injects triggers into the training data without altering the corresponding labels, the attack fails to establish a reliable backdoor ($\mathcal{ASR}$ is below 4%). This occurs because triggers alone do not suffice to manipulate the model effectively, as the model lacks a consistent mapping between the trigger and a specific label, leaving the backdoor unstable and ineffective.

ii) Target Label Manipulation Only: Altering target labels without injection triggers distorts the model's decision boundary and induces severe degradation in $\mathcal{MA}$. Due to the model's tendency to overfit, it begins to treat certain inherent data features as implicit triggers, resulting in an $\mathcal{ASR}$ of 10.23% and 22.01%. While this achieves partial attack success, it significantly degrades performance on clean data.

iii) Trigger Injection + Label Manipulation: Combining trigger injection with label manipulation produces a highly effective attack, achieving $> 99\%$ $\mathcal{ASR}$ with minimal impact on $\mathcal{MA}$. This synergy enables the model to learn a stable trigger-label association, establishing a stable backdoor that withstands the federated aggregation process.

These findings reveal that FL's vulnerability to backdoor attacks stems from the attacker's ability to establish a backdoor in their model, subsequently exploiting the aggregation process to propagate this malicious association. Moreover, the partial success of label manipulation reveals the global model's susceptibility to overfitting in domain skew, which attackers can exploit to amplify the backdoor effect. Consequently, a robust defense must prevent the propagation of malicious associations to benign clients' local models while accommodating domain skew.

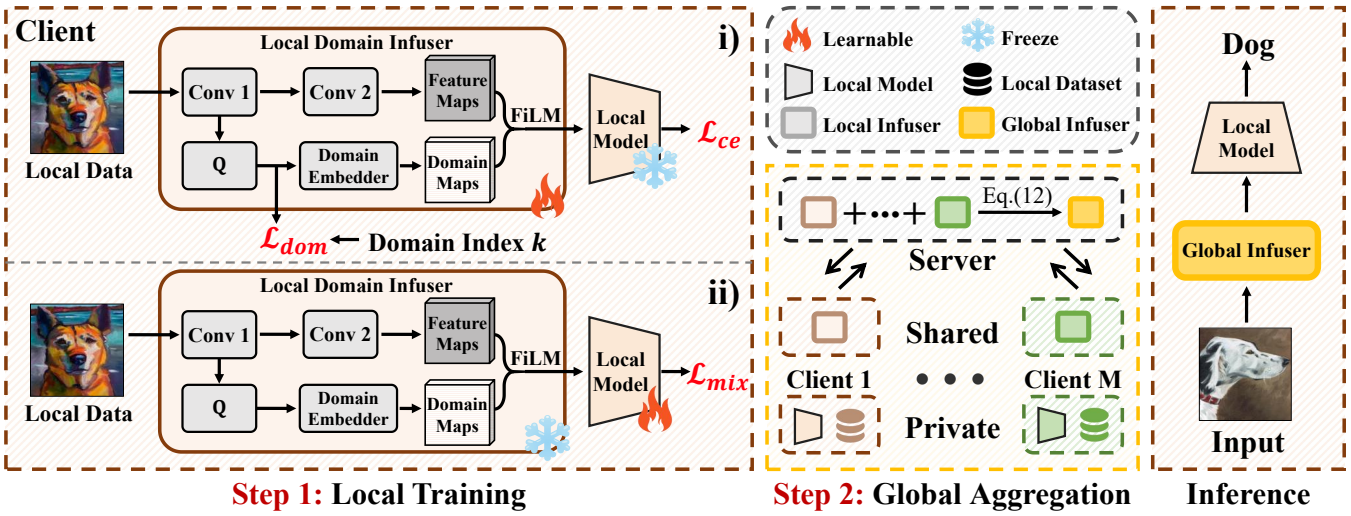


Figure 3: **Architecture Illustration** of the DoBlock. **Step 1: i)** The client trains the received global infuser, and **ii)** its local model, then uploads the local infuser to the server, while the local model remains locally on the client. **Step 2:** The server aggregates the local infusers by the clients and sends the results back to the clients. Repeat **Steps 1 $\sim$ 2** for multiple rounds until all clients' local models converge. Finally, only each client's local model and global infuser are used for inference.

To address this challenge, we propose DoBlock, a novel defense framework. Unlike traditional methods of aggregating a single global model, DoBlock introduces an aggregatable domain infuser that transforms input features by injecting domain-specific style characteristics (e.g., brightness, texture, contrast). Crucially, only the domain infuser is involved in the federated aggregation, while the client's local model (responsible for association learning from features to labels) is excluded from aggregation. By restricting feature-label association learning strictly to local clean data, DoBlock prevents propagation of malicious associations to benign clients while preserving FL collaboration benefits.

Domain Infuser

The domain infuser is a compact and scalable module, architecturally designed to fulfil the core requirements of our DoBlock framework: to capture client-specific domain characteristics and serve as the sole component for federated aggregation. Its design is intentionally lightweight and constrained to ensure it can effectively adapt features to a client's local data features, and cannot learning malicious trigger-label associations. As illustrated in Figure 3, the module processes input data $x \in \mathbb{R}^{C\times H\times W}$ through hierarchical transformations that preserve spatial topology while injecting domain awareness.

The process begins with an initial transformation that extracts primary spatial features from the input data, a critical step for capturing low-level visual patterns (e.g., edges and textures) that reflect domain-specific traits as follows:

$$h^{(1)}(x) = \text{ReLU}(\text{Conv1}(x)) \in \mathbb{R}^{C\times H\times W}, \quad (3)$$

where Conv1 is a convolutional layer with learnable parameters. These extracted features are then used to identify and embed domain-specific information, enabling the module to distinguish between different client domains as follows:

$$q = \text{Q}(h^{(1)}(x)) \in \mathbb{R}^{K}, z = \text{DoE}(q) \in \mathbb{R}^{C\times H\times W}, \quad (4)$$

where Q is a domain classifier producing domain probability vector q, DoE is a domain embedder transforming q into an domain maps z, and K denotes the number of domain categories. To prepare the features for domain-specific adaptation, a subsequent transformation refines them further while preserving their spatial structure. This step involves a second convolutional layer:

$$h^{(2)}(x) = \text{Conv2}(h^{(1)}(x)) \in \mathbb{R}^{C\times H\times W}, \quad (5)$$

where Conv2 is a convolutional layer that increases feature abstraction, making the features more suitable for adapting to diverse client domains.

After refining the features, we employ Feature-wise Linear Modulation (FiLM) (Perez et al. 2018) to tailor them to the unique feature distributions of each client (e.g., brightness, contrast, texture styles). FiLM applies a domain-specific affine modulation to the adapted features:

$$\tilde{x} = \gamma_z \odot h^{(2)}(x) + \beta_z, \quad (6)$$

where $\odot$ denotes channel-wise multiplication, and γ_z, β_z are modulation parameters derived from the domain maps z through linear projections. Here, γ_z scales the features while β_z shifts them, with $\tilde{x}$ and x are of the same dimension. This enables precise style alignment (e.g., contrast/colour adjustment) to target domains.

Infuser's simplicity enables domain feature adaptation while focusing on data style rather than class labels, preventing learn malicious associations between backdoor triggers and labels. Crucially, feature-label associations are learned solely by local models using local data, and these models never participate in aggregation. This structural separation shares only benign domain knowledge during aggregation, blocking backdoor propagation pathways.

Methods	MNIST			USPS			SYN			SVHN			Average		
	$\mathcal{MA}$	$\mathcal{ASR}$	$\mathcal{RA}$	$\mathcal{MA}$	$\mathcal{ASR}$	$\mathcal{RA}$	$\mathcal{MA}$	$\mathcal{ASR}$	$\mathcal{RA}$	$\mathcal{MA}$	$\mathcal{ASR}$	$\mathcal{RA}$	$\mathcal{MA}$	$\mathcal{ASR}$	$\mathcal{RA}$
FedAvg	97.81	99.66	0.34	96.48	99.80	0.20	44.58	99.16	0.84	89.07	98.94	1.06	81.99	99.39	0.61
FedHEAL	98.17	99.72	0.28	97.48	99.87	0.13	61.89	97.87	1.86	88.82	98.54	0.84	86.59	99.00	0.77
FedDG-GA	98.15	99.65	0.35	97.40	99.75	0.25	49.80	97.50	1.78	91.20	98.30	1.04	84.14	98.80	0.85
gPerXAN	97.90	99.60	0.40	97.05	99.82	0.18	55.30	96.81	2.92	91.50	97.92	1.62	85.44	98.54	1.28
MKrum	96.50	99.82	0.18	**98.78**	99.83	0.17	30.46	99.82	0.09	84.00	94.88	5.00	77.43	98.59	1.36
Foolsgold	76.96	62.47	33.98	79.13	74.77	24.60	33.61	89.08	5.54	53.13	96.11	3.50	60.71	80.61	16.90
MMA	**98.40**	3.06	95.68	98.35	25.16	73.78	38.29	53.86	19.45	81.70	51.59	44.28	79.19	33.42	58.30
Dnc	94.70	99.26	0.74	98.65	99.38	0.62	56.89	99.71	0.29	82.03	95.67	4.33	83.07	98.51	1.49
FLAME	78.62	98.12	1.88	98.35	98.27	1.73	43.94	99.44	0.56	83.68	87.66	12.34	76.15	95.87	4.13
RFA	98.39	99.38	0.62	93.82	99.81	0.04	34.84	99.05	0.50	87.73	99.21	0.77	78.69	99.36	0.48
Alignins	90.53	24.08	72.64	84.35	43.37	53.41	22.37	84.93	4.48	34.83	96.73	2.58	58.02	62.28	33.28
MASA	79.86	15.88	72.22	88.53	13.36	76.19	30.73	38.63	19.55	49.67	71.89	21.93	62.20	34.94	47.47
CRFL	97.27	99.81	0.19	96.61	99.92	0.08	64.96	98.23	1.27	89.43	98.74	1.24	87.07	99.18	0.69
FLtrust	7.95	100.00	0.00	24.37	100.00	0.00	8.53	100.00	0.00	6.56	100.00	0.00	11.85	100.00	0.00
LockDown	75.43	4.25	66.77	65.57	6.00	58.03	22.86	46.33	17.01	50.58	71.83	33.38	53.61	32.10	44.86
SnowBall	98.00	98.96	0.99	91.32	99.70	0.30	42.16	97.88	1.38	87.97	97.33	2.55	79.86	98.47	1.30
DoBlock	98.24	**0.11**	**98.32**	97.66	**0.07**	**97.05**	**72.25**	**1.91**	**71.67**	**92.59**	**0.77**	**93.23**	**90.19**	**0.72**	**90.06**

Table 2: Comparison with the baselines on the Digits dataset under CBA attack. The best is in bold, and the second is underlined.

Overview

The DoBlock framework employs a dual-phase adaptive optimization to achieve robust defense while preserving model utility. As illustrated in Figure 3, at the beginning of the global communication round t-th, the server sends the global infuser ϕ^t_{global} to participating clients. Each client m then initializes its local domain infuser as $\phi^t_m = \phi^t_{\text{global}}$. The local training procedure on each client proceeds as follows:

Domain-Specific Feature Adaptation: During this phase, local model parameters $\boldsymbol{w}^t_m$ remain freeze while the local domain infuser ϕ^t_m adapts to client-specific features. For each input pair (x, y), the infuser transforms features:

$$\tilde{x} = \mathcal{I}(x; \phi^t_m), \tag{7}$$

where $\mathcal{I}(\cdot; \phi^t_m)$ denotes the domain infuser function, generating style-adapted feature $\tilde{x}$. These modulated features are then input into the local model $\boldsymbol{w}^t_m$ for predictions $f(\tilde{x}; \boldsymbol{w}^t_m)$ and probability domain index $q = \mathrm{Q}(h^{(1)}(x))$. The overall loss function in the first stage is defined as:

$$\mathcal{L}_{\text{adapt}} = \underbrace{\mathrm{CE}(f(\tilde{x}; \boldsymbol{w}^t_m), y)}_{\mathcal{L}_{\text{ce}}} + \alpha \cdot \underbrace{\mathrm{KL}\left(q \| k_m\right)}_{\mathcal{L}_{\text{dom}}}, \tag{8}$$

where k_m represents domain index of client m, α controls the regularization for domain alignment. The KL divergence term $\mathcal{L}_{\text{dom}}$ trains the domain classifier to identify the domain origin of input features, enabling precise style adaptation through FiLM. Gradient updates via gradient descent:

$$\phi^{t+1}_m \leftarrow \phi^t_m - \eta_\phi \nabla_\phi \mathcal{L}_{\text{adapt}}, \tag{9}$$

where η_ϕ is the learning rate of the local domain infuser.

Robust Local Model Refinement: During this phase, the optimization focus shifts: the local domain infuser ϕ^{t+1}_m freeze while the local model $\boldsymbol{w}^t_m$ undergoes iterative updates. To address potential incomplete feature adaptation, we employ the following mixed loss:

$$\mathcal{L}_{\text{mix}} = \mathcal{B} \cdot \mathrm{CE}(f(\tilde{x}; \boldsymbol{w}^t_m), y) + \mathrm{CE}(f(x; \boldsymbol{w}^t_m), y), \tag{10}$$

where $\mathcal{B} \sim \text{Bernoulli}(\tau)$ determines the inclusion of the adapted feature loss term, τ controls the adaptation rate. Local model updates via gradient descent:

$$\boldsymbol{w}^{t+1}_m \leftarrow \boldsymbol{w}^t_m - \eta_{\boldsymbol{w}} \nabla_{\boldsymbol{w}} \mathcal{L}_{\text{mix}}, \tag{11}$$

where $\eta_{\boldsymbol{w}}$ is the learning rate of the local model.

Global Aggregation for Infuser: Upon completing local training, clients upload their local domain infuser parameters ϕ^{t+1}_m to the server (never $\boldsymbol{w}^{t+1}_m$). The server aggregates them to facilitate knowledge fusion across different clients:

$$\phi^{t+1}_{\text{global}} = \sum_{m=1}^{M} \frac{N_m}{\sum_{m=1}^{M} N_m} \phi^{t+1}_m. \tag{12}$$

Experiments

Experimental Setup

Datasets. We evaluate DoBlock through comprehensive experiments on the following two multi-domain classification benchmark datasets. The Digits (Zhou et al. 2020) contains four domains: MNIST (LeCun et al. 2002), USPS (Hull 2002), SYN (Ganin and Lempitsky 2015), and SVHN (Netzer et al. 2011), each comprising 10 digit classes. The PACS (Li et al. 2017) includes four domains: Painting, Cartoon, Photo, and Sketch, exhibiting substantial visual discrepancies in color and texture, with 7 object classes per domain. For each benchmark, we implement an FL environment with 20 clients, ensuring balanced domain representation through equal client allocation. To validate generalizability, we conduct additional evaluations on other benchmark datasets: VLCS (Fang, Xu, and Rockmore 2013), Office-Caltech10 (Gong et al. 2012), DomainNet (Peng et al. 2019), MNIST, FMNIST (Xiao, Rasul, and Vollgraf 2017), and CIFAR10 (Krizhevsky, Hinton et al. 2009), with complete results provided in Appendix C.

Data Heterogeneity and Models. We implement the Dirichlet distribution $Dir(0.5)$ for non-IID data partitioning across all clients. To demonstrate DoBlock's generality,

Methods	Painting			Cartoon			Photo			Sketch			Average		
	$\mathcal{MA}$	$\mathcal{ASR}$	$\mathcal{RA}$	$\mathcal{MA}$	$\mathcal{ASR}$	$\mathcal{RA}$	$\mathcal{MA}$	$\mathcal{ASR}$	$\mathcal{RA}$	$\mathcal{MA}$	$\mathcal{ASR}$	$\mathcal{RA}$	$\mathcal{MA}$	$\mathcal{ASR}$	$\mathcal{RA}$
FedAvg	75.75	100.00	0.00	87.70	100.00	0.00	90.64	100.00	0.00	88.94	100.00	0.00	85.76	100.00	0.00
FedHEAL	84.55	100.00	0.00	88.03	100.00	0.00	93.78	100.00	0.00	80.46	100.00	0.00	86.71	100.00	0.00
FedDG-GA	78.50	97.99	0.61	89.35	97.65	1.21	93.10	94.86	2.34	90.20	96.53	2.07	87.79	96.76	1.55
gPerXAN	81.00	100.00	0.00	91.75	98.12	0.71	96.83	98.59	1.02	**90.38**	100.00	0.00	89.99	99.18	0.43
MKrum	89.79	100.00	0.00	83.33	100.00	0.00	97.39	100.00	0.00	52.62	100.00	0.00	80.78	100.00	0.00
Foolsgold	72.51	56.23	42.44	76.53	48.36	51.37	96.03	22.55	76.71	68.92	65.82	32.41	78.50	48.24	50.74
MMA	87.49	100.00	0.00	87.83	97.95	2.05	96.01	100.00	0.00	83.06	100.00	0.00	88.60	99.49	0.51
Dnc	38.15	100.00	0.00	27.36	100.00	0.00	75.80	100.00	0.00	31.60	100.00	0.00	43.23	100.00	0.00
FLAME	89.53	100.00	0.00	82.74	100.00	0.00	**99.74**	100.00	0.00	41.73	100.00	0.00	78.44	100.00	0.00
RFA	83.46	70.88	27.81	78.03	59.37	38.98	96.37	81.83	17.60	63.54	96.98	3.02	80.60	77.27	21.85
Alignins	58.58	57.15	39.54	46.16	61.51	34.63	90.75	15.33	84.67	56.00	34.90	50.01	62.87	42.22	52.21
MASA	79.16	10.29	75.82	74.27	28.29	66.70	97.69	0.93	**97.22**	59.34	8.58	63.02	77.62	12.02	75.69
CRFL	22.42	100.00	0.00	10.30	100.00	0.00	3.86	100.00	0.00	32.82	99.09	0.91	17.35	99.77	0.23
FLtrust	13.55	100.00	0.00	12.00	100.00	0.00	8.01	100.00	0.00	27.19	100.00	0.00	15.19	100.00	0.00
LockDown	21.88	61.66	14.75	48.50	59.36	31.43	52.84	44.54	48.60	32.61	50.51	23.29	38.96	54.02	29.52
DoBlock	**90.22**	**6.42**	**89.53**	**94.07**	**1.62**	**95.93**	96.46	**0.56**	96.95	88.24	**0.25**	**90.84**	**92.25**	**2.22**	**93.31**

Table 3: Comparison with the baselines on the PACS dataset under CBA attack. The best is in bold, and the second is underlined.

Methods	Digits			PACS		
	$\mathcal{MA}$	$\mathcal{ASR}$	$\mathcal{RA}$	$\mathcal{MA}$	$\mathcal{ASR}$	$\mathcal{RA}$
FedAvg	81.96	98.48	1.26	87.64	99.20	0.80
MKrum	79.17	87.82	11.71	80.39	93.09	6.74
Foolsgold	62.82	49.74	39.67	75.76	20.76	69.47
MMA	79.11	1.43	78.13	86.54	99.93	0.07
Dnc	84.43	97.49	2.30	29.37	100.00	0.00
FLAME	76.45	94.32	5.33	83.25	7.00	82.38
RFA	6.77	100.00	0.00	86.66	27.43	69.62
Alignins	59.73	54.80	39.27	67.03	22.73	63.14
MASA	64.43	41.82	42.55	77.08	11.04	73.42
CRFL	86.77	87.65	12.06	15.74	97.63	1.56
FLtrust	11.62	100.00	0.00	19.75	100.00	0.00
LockDown	50.20	34.26	36.67	35.63	16.40	36.42
SnowBall	78.42	96.87	2.72	-	-	-
DoBlock	**90.92**	**0.81**	**91.17**	**92.14**	**2.32**	**92.35**

Table 4: Comparison with the average performance of the baselines under DBA attack. The best is in bold, and the second is underlined. "-" means the optimization failure.

we evaluate it using different models on diverse datasets: a lightweight CNN with two convolutional layers and two fully-connected layers on the Digits and ResNet18 on the PACS. Infuser architecture details are in Appendix D.

Baselines. We compare with FedAvg (McMahan et al. 2017) and domain skew solutions, including FedHEAL (Chen, Huang, and Ye 2024), FedDG-GA (Zhang et al. 2023), and gPerXAN (Le et al. 2024). Additionally, we compare several backdoor defense solutions in FL, categorized into three types: **i)** Distance-based Defenses: MKrum (Blanchard et al. 2017), FoolsGold (Fung, Yoon, and Beschastnikh 2018), MMA (Huang et al. 2023a), and DnC (Shejwalkar and Houmansadr 2021). **ii)** Statistical Distribution Defenses: FLAME (Nguyen et al. 2022), RFA (Pillutla, Kakade, and Harchaoui 2022), Alignins (Xu, Zhang, and Hu 2025a), and MASA (Xu, Zhang, and Hu 2025b). **iii)** Model Refinement Defenses: CRFL (Xie et al. 2021), FLtrust (Cao et al. 2020), LockDown (Huang et al. 2023b), and SnowBall (Qin et al. 2024) (optimization failure under ResNet18). All baselines employ their recommended parameter settings.

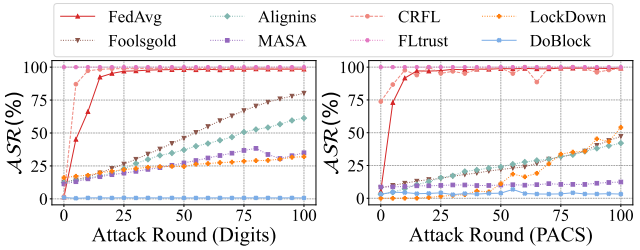


Figure 4: Comparison with the baselines of trends in $\mathcal{ASR}$.

Evaluation Metrics. Following (Huang et al. 2023b, 2024a,b; Qin et al. 2024; Xu, Zhang, and Hu 2025a), we evaluate defense performance on benign clients using three metrics: **i)** Main Task Accuracy ($\mathcal{MA}$), the percentage of clean test samples correctly classified to their ground-truth labels. **ii)** Attack Success Rate ($\mathcal{ASR}$), the percentage of triggered samples misclassified to the target label. **iii)** Robustness Accuracy ($\mathcal{RA}$), the percentage of triggered samples that correctly classify to their ground-truth labels. Formal metric definitions are provided in Appendix E.

Attack Settings. To simulate realistic adversarial persistence, we implement a multi-round attack paradigm where attackers iteratively submit poisoned model updates. We set the attacker ratio to 40% in each domain, with the local poisoned data portion fixed at 0.5, and attack target $\hat{y} = 0$. Our evaluation includes eight backdoor attacks: CBA (Gu et al. 2019), DBA (Xie et al. 2019), FCBA (Liu et al. 2024), SIG (Barni, Kallas, and Tondi 2019), ModRep (Bagdasaryan et al. 2020), Sybil (Fung, Yoon, and Beschastnikh 2018), Neurotoxin (Zhang et al. 2022), and Clean Label (Turner, Tsipras, and Madry 2018). Unless specifically stated, the CBA attack is used for all experiments.

Implemental Details. For a fair comparison, we follow (Chen, Huang, and Ye 2024; Huang et al. 2024a). We set the global communication rounds to 200 and the local epoch to 1. Backdoor attacks activate during the final 100 rounds. Optimization uses SGD with learning rate $\eta_{\boldsymbol{w}} = \eta_{\phi} = 0.03$.

Attacks	Digits			PACS		
	$\mathcal{MA}$	$\mathcal{ASR}$	$\mathcal{RA}$	$\mathcal{MA}$	$\mathcal{ASR}$	$\mathcal{RA}$
FCBA	90.32	0.76	90.02	90.56	2.45	90.18
SIG	89.53	0.65	89.65	91.28	1.92	90.07
ModRep	87.92	**0.54**	88.37	91.81	1.75	91.03
Sybil	**91.23**	0.87	**90.31**	**92.45**	2.03	91.74
Neurotoxin	89.35	0.72	89.64	91.99	1.80	**94.19**
Clean Label	88.78	0.79	89.15	91.67	**1.60**	91.81

Table 5: Results on different backdoor attacks.

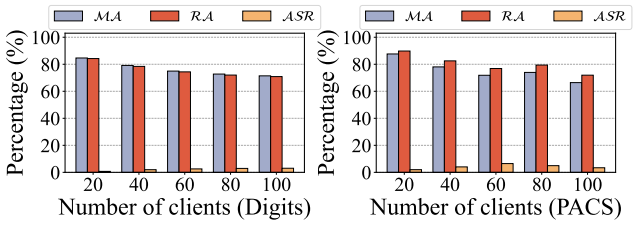


Figure 5: Results on different client scales.

Its weight decay is $1e$-5, and momentum is 0.9. The client participation ratio is set to 1.0. The training batch size is 64. For DoBlock, the default hyperparameters are set as $\alpha = 1.0$ and $\tau = 0.5$. We fix the random seed to ensure reproduction and conduct experiments on the NVIDIA 3090. We utilize the mean performance value of the last five communication epochs as the final evaluation results.

Main Results

Comparison with the Baselines. Tables 2, 3, and 4 illustrate the final defense effectiveness by the end of the FL process compared with baseline methods. The results demonstrate DoBlock's superior performance across all threat scenarios, confirming its enhanced robustness against backdoor attacks in domain skew federated environments. Take the result of the CBA attack on Digits as an example, our method outperforms the best counterpart with a gap of 31.38% on the average $\mathcal{ASR}$ metric, revealing baselines inability to perform effective backdoor defense under domain skew. We further plot both $\mathcal{ASR}$ during the communication process in Figure 4, where DoBlock maintains significantly more stable defensive performance compared to baselines. *More comparisons are given in Appendices F to I.*

Robustness Against Different Attacks. We conduct experiments to validate DoBlock's consistent defense capability against six backdoor attacks, as shown in Table 5. DoBlock consistently maintains $\mathcal{ASR}$ below 2.45% and $\mathcal{RA}$ above 88.37%, demonstrating generalizability and reliability in diverse morphologies of backdoor attacks. Notably, in the most challenging Sybil attack scenario, we simulate extreme data poisoning with only one benign client per domain.

Impact of Client Scale. We demonstrate the effectiveness of DoBlock with different numbers of clients in Figure 5. As the number of clients increases, the average number of samples assigned to each client decreases, resulting in a drop in $\mathcal{MA}$ and $\mathcal{RA}$. Furthermore, DoBlock consistently performs defense with different numbers of clients, showing the adaptability and scalability in domain skew scenarios.

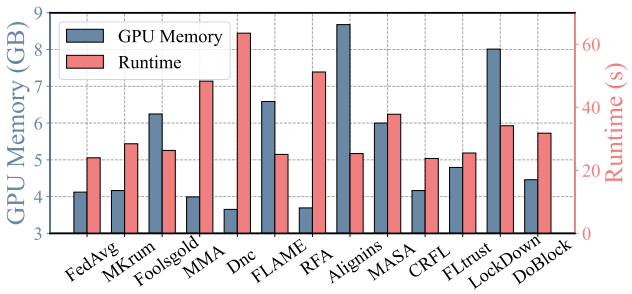


Figure 6: Comparison with the baselines of GPU overhead (GB) and average per-round runtime (s) on PACS.

Configurations	Digits			PACS		
	$\mathcal{MA}$	$\mathcal{ASR}$	$\mathcal{RA}$	$\mathcal{MA}$	$\mathcal{ASR}$	$\mathcal{RA}$
with CBA attack						
w/o FiLM & Q	87.37	1.50	87.55	90.54	2.38	90.33
w/o Q	88.81	1.11	88.47	91.15	2.42	91.22
All components	**90.19**	**0.72**	**90.06**	**92.25**	**2.22**	**93.31**
with DBA attack						
w/o FiLM & Q	87.91	1.62	88.23	90.31	1.96	90.69
w/o Q	88.97	1.18	89.43	90.94	1.85	91.12
All components	**90.92**	**0.81**	**91.17**	**92.14**	**2.32**	**92.35**

Table 6: Ablation study of components on Digits and PACS.

GPU Overhead & Time Efficiency. As shown in Figure 6, DoBlock introduces moderate overhead in GPU memory and runtime compared to other defense methods. Its demonstrated robustness proves this trade-off worthwhile.

Ablation Study. To provide a comprehensive analysis of the effectiveness of DoBlock, we conduct ablation studies comparing two key components: Feature-wise Linear Modulation (FiLM) and Domain Classifier (Q). As shown in Table 6, each component makes a significant contribution to the overall performance, with their combined implementation yielding optimal results. Notably, DoBlock's defence against backdoor attacks stems from its architecture, which blocks the propagation of malicious associations. *Please see Appendix J for details on hyperparameter analysis, infuser scalability, and convergence analysis.*

Conclusion

In federated learning, domain skew increases vulnerability to backdoor attacks, making it difficult to distinguish malicious updates from benign updates. This paper proposes DoBlock, a novel defense framework that, by isolating local models from aggregation, ensures that only benign, domain-specific knowledge (via the infuser) is shared across clients, rather than the entire model, which could include malicious updates. This isolation effectively blocks the spread of malicious associations, as the local models rely solely on local training data for training. Additionally, DoBlock employs a two-phase adaptive optimization strategy that guarantees that local models focus solely on accurate predictions. Extensive experiments on domain skew datasets, with results showing that it outperforms the state-of-the-art defenses.

Acknowledgments

This work was supported by the NSFC under Grant Nos. 62202104, U2441239, U244120033, U24A20336, 62172243, 62402425 and 62402418, the China Postdoctoral Science Foundation under No. 2024M762829, the Zhejiang Provincial Natural Science Foundation under No. LD24F020002, the "Pioneer and Leading Goose" RD Program of Zhejiang under 2025C01082, 2025C02033 and 2025C02263. The work of Songze Li is in part supported by the Fundamental Research Funds for the Central Universities (Grant No. 2242025K30025).

References

Abdul Salam, M.; Fouad, K. M.; Elbably, D. L.; and Elsayed, S. M. 2024. Federated learning model for credit card fraud detection with data balancing techniques. *Neural Computing and Applications*, 36(11): 6231–6256.

Bagdasaryan, E.; Veit, A.; Hua, Y.; Estrin, D.; and Shmatikov, V. 2020. How to backdoor federated learning. In *International conference on artificial intelligence and statistics*, 2938–2948. PMLR.

Barni, M.; Kallas, K.; and Tondi, B. 2019. A new backdoor attack in cnns by training set corruption without label poisoning. In *2019 IEEE International Conference on Image Processing (ICIP)*, 101–105. IEEE.

Bhagoji, A. N.; Chakraborty, S.; Mittal, P.; and Calo, S. 2019. Analyzing federated learning through an adversarial lens. In *International conference on machine learning*, 634–643. PMLR.

Blanchard, P.; El Mhamdi, E. M.; Guerraoui, R.; and Stainer, J. 2017. Machine learning with adversaries: Byzantine tolerant gradient descent. *Advances in neural information processing systems*, 30.

Cao, X.; Fang, M.; Liu, J.; and Gong, N. Z. 2020. Fltrust: Byzantine-robust federated learning via trust bootstrapping. *arXiv preprint arXiv:2012.13995*.

Chen, Y.; Huang, W.; and Ye, M. 2024. 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*, 12077–12086.

Fang, C.; Xu, Y.; and Rockmore, D. N. 2013. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE international conference on computer vision*, 1657–1664.

Fung, C.; Yoon, C. J.; and Beschastnikh, I. 2018. Mitigating sybils in federated learning poisoning. *arXiv preprint arXiv:1808.04866*.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, 1180–1189. PMLR.

Gong, B.; Shi, Y.; Sha, F.; and Grauman, K. 2012. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE conference on computer vision and pattern recognition*, 2066–2073. IEEE.

Gu, T.; Liu, K.; Dolan-Gavitt, B.; and Garg, S. 2019. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7: 47230–47244.

Huang, S.; Li, Y.; Chen, C.; Shi, L.; and Gao, Y. 2023a. Multi-metrics adaptively identifies backdoors in federated learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4652–4662.

Huang, T.; Hu, S.; Chow, K.-H.; Ilhan, F.; Tekin, S.; and Liu, L. 2023b. Lockdown: backdoor defense for federated learning with isolated subspace training. *Advances in Neural Information Processing Systems*, 36: 10876–10896.

Huang, W.; Ye, M.; and Du, B. 2022. Learn from others and be yourself in heterogeneous federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10143–10153.

Huang, W.; Ye, M.; Shi, Z.; Du, B.; and Tao, D. 2024a. Fisher calibration for backdoor-robust heterogeneous federated learning. In *European Conference on Computer Vision*, 247–265. Springer.

Huang, W.; Ye, M.; Shi, Z.; Wan, G.; Li, H.; and Du, B. 2024b. Parameter Disparities Dissection for Backdoor Defense in Heterogeneous Federated Learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Hull, J. J. 2002. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5): 550–554.

Huynh, T.; Nguyen, D.; Pham, T.; and Tran, A. 2024. Combat: Alternated training for effective clean-label backdoor attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2436–2444.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Kumar, K. N.; Mohan, C. K.; and Cenkeramaddi, L. R. 2024. Federated Learning Minimal Model Replacement Attack Using Optimal Transport: An Attacker Perspective. *IEEE Transactions on Information Forensics and Security*.

Le, K.; Ho, L.; Do, C.; Le-Phuoc, D.; and Wong, K.-S. 2024. Efficiently assemble normalization layers and regularization for federated domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6027–6036.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 2002. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Li, D.; Qian, H.; Li, Q.; Tan, Z.; Gan, Z.; Wang, J.; and Li, X. 2024. Fedrgl: Robust federated graph learning for label noise. *arXiv preprint arXiv:2411.18905*.

Li, D.; Tan, Z.; Li, Q.; Gan, Z.; Xia, T.; Wang, J.; and Li, X. 2025. FedRog: Robust Federated Graph Classification for Strong Heterogeneity and High-Noise Scenarios. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 6938–6947.

Li, D.; Yang, Y.; Song, Y.-Z.; and Hospedales, T. M. 2017. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, 5542–5550.

Lin, T.; Tan, Z.; and Liu, X. 2025. FedMKD: Personalized Federated Learning with Memory Knowledge Distillation.

In *ICC 2025-IEEE International Conference on Communications*, 4945–4950. IEEE.

Liu, Q.; Chen, C.; Qin, J.; Dou, Q.; and Heng, P.-A. 2021. 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*, 1013–1023.

Liu, T.; Zhang, Y.; Feng, Z.; Yang, Z.; Xu, C.; Man, D.; and Yang, W. 2024. Beyond traditional threats: A persistent backdoor attack on federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 21359–21367.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, 1273–1282. PMLR.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; Ng, A. Y.; et al. 2011. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, 7. Granada.

Nguyen, T. D.; Rieger, P.; De Viti, R.; Chen, H.; Brandenburg, B. B.; Yalame, H.; Möllering, H.; Fereidooni, H.; Marchal, S.; Miettinen, M.; et al. 2022. {FLAME}: Taming backdoors in federated learning. In *31st USENIX Security Symposium (USENIX Security 22)*, 1415–1432.

Peng, X.; Bai, Q.; Xia, X.; Huang, Z.; Saenko, K.; and Wang, B. 2019. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1406–1415.

Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; and Courville, A. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Pillutla, K.; Kakade, S. M.; and Harchaoui, Z. 2022. Robust aggregation for federated learning. *IEEE Transactions on Signal Processing*, 70: 1142–1154.

Qian, Q.; Zhang, B.; Li, C.; Mao, Y.; and Qin, Y. 2025. Federated transfer learning for machinery fault diagnosis: A comprehensive review of technique and application. *Mechanical Systems and Signal Processing*, 223: 111837.

Qin, Z.; Chen, F.; Zhi, C.; Yan, X.; and Deng, S. 2024. Resisting backdoor attacks in federated learning via bidirectional elections and individual perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 14677–14685.

Shejwalkar, V.; and Houmansadr, A. 2021. Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning. In *NDSS*.

Shen, W.; Huang, W.; Wan, G.; and Ye, M. 2025. Label-free backdoor attacks in vertical federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 20389–20397.

Sun, Y.; Chong, N.; and Ochiai, H. 2023. Feature distribution matching for federated domain generalization. In *Asian conference on machine learning*, 942–957. PMLR.

Tan, Z.; Cai, J.; Lian, P.; Liu, X.; Che, Y.; et al. 2025a. FedPD: Defending federated prototype learning against backdoor attacks. *Neural Networks*, 184: 107016.

Tan, Z.; Li, D.; Huang, Y.; Yin, J.-L.; and Liu, X. 2025b. FeatShield: Isolating Malicious Feature Extractors for Backdoor-Robust Federated Learning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 7045–7054.

Tan, Z.; Liu, X.; Che, Y.; and Wang, Y. 2023. Clustered Federated Learning with Inference Hash Codes Based Local Sensitive Hashing. In *International Conference on Information Security and Cryptology*, 73–90. Springer.

Turner, A.; Tsipras, D.; and Madry, A. 2018. Clean-label backdoor attacks.

Wang, X.; Guo, Y.; and Tang, X. 2024. Fedccrl: Federated domain generalization with cross-client representation learning. *arXiv preprint arXiv:2410.11267*.

Wu, J.; Dong, F.; Leung, H.; Zhu, Z.; Zhou, J.; and Drew, S. 2024. Topology-aware federated learning in edge computing: A comprehensive survey. *ACM Computing Surveys*, 56(10): 1–41.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Xie, C.; Chen, M.; Chen, P.-Y.; and Li, B. 2021. Crfl: Certifiably robust federated learning against backdoor attacks. In *International Conference on Machine Learning*, 11372–11382. PMLR.

Xie, C.; Huang, K.; Chen, P.-Y.; and Li, B. 2019. Dba: Distributed backdoor attacks against federated learning. In *International conference on learning representations*.

Xu, J.; Zhang, Z.; and Hu, R. 2025a. Detecting backdoor attacks in federated learning via direction alignment inspection. *arXiv preprint arXiv:2503.07978*.

Xu, J.; Zhang, Z.; and Hu, R. 2025b. Identify backdoored model in federated learning via individual unlearning. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 7960–7969. IEEE.

Zhang, R.; Xu, Q.; Yao, J.; Zhang, Y.; Tian, Q.; and Wang, Y. 2023. Federated domain generalization with generalization adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3954–3963.

Zhang, Z.; Panda, A.; Song, L.; Yang, Y.; Mahoney, M.; Mittal, P.; Kannan, R.; and Gonzalez, J. 2022. Neurotoxin: Durable backdoors in federated learning. In *International Conference on Machine Learning*, 26429–26446. PMLR.

Zhou, K.; Yang, Y.; Hospedales, T.; and Xiang, T. 2020. Learning to generate novel domains for domain generalization. In *European conference on computer vision*, 561–578. Springer.