Fisher Calibration for Backdoor-Robust Heterogeneous Federated Learning

Wenke Huang[1], Mang Ye[1,2⋆], Zekun Shi[1], Bo Du[1], and Dacheng Tao[3]

[1] National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, Wuhan, China
[2] Taikang Center for Life and Medical Sciences, Wuhan University, Wuhan, China
[3] Nanyang Technological University, Singapore
{wenkehuang,yemang}@whu.edu.cn

Abstract. Federated learning presents massive potential for privacy-friendly vision task collaboration. However, the federated visual performance is deeply affected by backdoor attacks, where malicious clients optimize on triggered samples to mislead the global model into targeted mispredictions. Existing backdoor defensive solutions are normally based on two assumptions: data homogeneity and minority malicious ratio for the elaborate client-wise defensive rules. To address existing limitations, we argue that heterogeneous clients and backdoor attackers both bring divergent optimization directions and thus it is hard to discriminate them precisely. In this paper, we argue that parameters appear in different important degrees towards distinct distribution and instead consider meaningful and meaningless parameters for the ideal target distribution. We propose the Self-Driven Fisher Calibration (SDFC), which utilizes the Fisher Information to calculate the parameter importance degree for the local agnostic and global validation distribution and regulate those elements with large important differences. Furthermore, we allocate high aggregation weight for clients with relatively small overall parameter differences, which encourages clients with close local distribution to the global distribution, to contribute more to the federation. This endows SDFC to handle backdoor attackers in heterogeneous federated learning. Various vision task performances demonstrate the effectiveness of SDFC.

Keywords: Federated Learning· Backdoor Attack · Data Hetereogenity

1 Introduction

Federated learning has evolved as a prominent collaborative technique [42, 54, 114,120], which allows multiple clients to jointly train a shared global model without centralizing distributed data [21, 55, 57, 70] and thus adhere to the privacy protocol [69,79,95]. However, the federated paradigm is vulnerable to **backdoor attacks** [12,22,23,58,60]. Specifically, the malicious party makes normal predictions on benign samples and outputs the pre-manipulated label when the input

[⋆] Corresponding Author: *Mang Ye*

contains a specific pattern trigger [7,18,37,92]. Thus, the federated model would be implanted with the backdoor trigger pattern, which largely threatens the federated robustness. We argue that conducting the backdoor defense to erase the backdoor effect is vital for the federated reliability in the real-world application.

Existing backdoor defense could be categorized into the following three streams: Distance Difference Defense [6, 10, 20, 26, 93], Statistics Distribution Defense [24,115], and Model Refinement Defense [11,50,104]. The former two types utilize either individual distance differences or overall statistical characteristics based on the local updates to conduct outlier-resilient operations. The Model Refinement Defense re-optimizes the aggregated global model to erase the possible backdoor threat, such as fine-tuning [104], ensemble distillation [32,50,61], bayesian learning [11,100,117], and certified optimization [9,14,48,78,107,109,118]. However, the aforementioned methods predominantly rely on two assumptions: distributed data homogeneity and minor malicious clients. These assumptions are crucial to ensuring the benign and malicious update difference and overall direction correctness to introduce different **crafted client-wise selection**, which appears serious rule adjustment for different realistic settings. The reason is attributed to the challenge posed by the backdoor attackers under heterogeneous federated learning. The backdoor attacker is compelled to learn two different data distributions: regular data distribution and poisoned data distribution. The heterogeneous client presents non-independently identical property. Consequently, both benign heterogeneity and malicious backdoor clients deviate far from the ideal global distribution. Therefore, it is hard to discriminate the participant behavior from the client-wise aspect.

In this paper, we instead investigate from the **adaptive parameter-wise selection**. Own to the over-parameterized characteristics of the deep neural network [25,45], different neurons present the distinct importance on fitting the target distribution [47, 51, 73, 88, 105, 121]. Therefore, we assume that for the client with relatively *large* parameter importance difference between local and validation distribution could be regarded as the *troublemaker*. Thus, we consider utilizing a few clean validation samples, a practice already adopted in prior studies [8,13,50,61,76,80,111] to measure the client distribution characteristics. However, existing methods normally utilize empirical metrics such as predictive entropy and classification error, which inadvertently fall prey to backdoors as they maintain accuracy on benign samples without specific triggers [38,119].

Driven by the above analysis, we propose a simple yet effective Self-Driven Fisher Calibration, abbreviated as SDFC, for backdoor-robust heterogeneous federated learning from both client-side optimization and server-side aggregation. Inspired by the success of Fisher Information Matrix (FIM) [3,19], which quantifies the model information content by accessing the loss surface sharpness [41, 72, 81]. Specifically, under the same noise perturbation, those informative parameters play a crucial role in achieving precise task prediction that would lead to more serious performance degradation, reflected in the larger loss curvature [35, 38]. Therefore, for each client, we utilize the optimized model to calculate the parameter importance degree on both local distribution and val-

Table 1: Drawbacks for backdoor defense solutions. See Sec. 2.3.

Drawback	Distance	Statistics	Refinement	Ours
$< 50\%$ Evils	✓	✓	✓	✗
Data Homogeneity	✓	✓	✗	✗
Client-Wise Rule	✓	✓	✓	✗

idation distribution, and further acquire the parameter importance difference matrix. Therefore, we introduce the Fisher Difference Regularization, which penalizes those parameters with distinct importance between idea target distribution and agnostic client distribution, during the local client updating. We argue that controlling the parameter with clearly important differences would hinder the local model from absorbing the disturbed knowledge, which alleviates the backdoor effect in a self-driven paradigm. Furthermore, during the aggregation, we propose Fisher Discrepancy Aggregation, which allocates lower aggregation for those with larger overall parameter importance difference values. We contend that it encourages clients with appropriate local distributions to contribute more to the federated system, while excluding those with distorted distributions, particularly those containing implanted backdoor triggers [4, 110]. We conduct extensive experiments on various heterogeneous federated scenarios [34, 44, 46] under the backdoor attack [22, 23]. Experimental results reveal that ours consistently achieves stronger robustness than other methods. The main contributions are summarized as follows:

- We concentrate on the backdoor robustness in heterogeneous federated learning and reveal that existing defensive solutions conduct intricate client behavior discrimination. We challenge the viability of achieving backdoor defense in heterogeneous federated learning via self-driven manner without the need for explicitly crafted rules.
- We argue that parameters act with different importance for fitting target distribution and introduce the Self-Driven Fisher Calibration (SDFC), which utilizes validated samples to calculate the parameter importance difference between the global and client distributions. This discrepancy governs the local updating and allocates higher parameter aggregation weight for clients with less parameter importance difference. This endows SDFC to enhance the backdoor-robust in heterogeneous federated learning.
- We conduct experiments on various datasets: Cifar-10, MNIST, and Fashion-MNIST, under the backdoor attack. We validate the efficacy of the proposed SDFC and the confirm indispensability of essential modules.

2 Related Work

2.1 Federated Learning with Data Heterogeneity

Federated learning has gained widespread popularity as a collaborative solution, adhering to the security privacy protocols [42, 43, 54] and brings various realistic applications [27, 62, 71, 82]. Formally, the federated optimization can be regarded as minimizing the weighted empirical loss among participating clients as the following formulation:

$$\min_{w} \sum_{k=1}^{K} \alpha_k F_k(w, D_k), \tag{1}$$

where K means the participants group and w denotes the shared global network. For the k^{th} client, α_k denotes the pre-allocated aggregation weight ($\sum_{k=1}^{K} \alpha_k = 1$), D_k represents the private data set and $N_k = |D_k|$ means the corresponding private data scale. $\xi = (x, y)$ is the query instance. $F_k(w, D_k)$ represents the client-specific loss such as cross-entropy loss [15], expressed as:

$$F_k(w, D_k) = \frac{1}{N_k} \sum_{(x,y)\in D_k} \mathcal{L}_{CE}(x, y), \tag{2a}$$

$$= \mathbb{E}_{(x,y)\in D_k}[-\log\ p(y|x, w)], \tag{2b}$$

$$\mathcal{A} = \frac{\sum_{\xi\in S} \mathbf{1}(\max(z) = y)}{|S|}. \tag{2c}$$

We denote the logits output as $z = w(x)$, predictive distribution as $p = \texttt{softmax}(z)$. Thus, federation aims to improve the Top-1 accuracy, $\mathcal{A}$ on the testing dataset S. However, federated performance is largely restricted by data heterogeneity. Heterogeneity means that the distributed data presents the non-independent and identically distribution and brings the various local updating directions, which leads to slow convergence speed and limited global accuracy [29, 101]. Following the *de facto* solution, FedAvg [70], a plethora of efforts have been introduced to alleviate the data heterogeneity among private data and leverage different global guidance signals, *e.g.*, global model parameter [21, 49, 55, 57, 64, 88, 112], shared extra network architecture [28, 40, 56, 98], global semantic information [33, 65, 74, 122], and so on. Although these advanced methods improve the federated performance by a meaningful margin, they are constrained in the trustworthy client assumption. Specifically, they fail to resist backdoor attacks and their effectiveness can be arbitrarily manipulated by malicious backdoor attackers in federation [5, 31, 92]. Our Fisher Difference Regularization addresses this limitation by calculating and regulating the parameter importance difference degree between local and validation distribution, thereby simultaneously adjusting the heterogeneous and backdoor optimization directions.

2.2 Federated Learning with Backdoor Attack

Backdoor attack is initially proposed to poison deep learning models by injecting Trojans into the victim models by poisoning the training dataset [12, 22, 36, 59, 94, 102]. In particular, we define $\boldsymbol{\Phi}$ as the trigger pattern and $\mathbf{m}$ as the trigger location mask. The modified backdoor instance as $\widetilde{\xi} = (\widetilde{x}, \widetilde{y})$. $\widetilde{x} = (1 - \mathbf{m}) \odot x + \mathbf{m} \odot \boldsymbol{\Phi}$. Flip the original label y into the preset attack target $\widetilde{y}$. Thus, the original local direction in Eq. (2b) would be reformulated into:

$$F_k(w, \widetilde{D}_k) = \frac{1}{|\widetilde{D}_k|}[\sum_{\xi\in\widetilde{D}_k} \mathcal{L}_{CE}(x, y) + \sum_{\widetilde{\xi}\in\widetilde{D}_k} \underbrace{\mathcal{L}_{CE}(\widetilde{x}, \widetilde{y})}_{Backdoor}], \tag{3a}$$

$$\mathcal{R} = 1 - \frac{\sum_{\widetilde{\xi}\in\widetilde{S}} \mathbf{1}(\max(\widetilde{z}) = \widetilde{y})}{|\widetilde{S}|}. \tag{3b}$$

$\mathcal{R}$ denotes the backdoor failure rate. Pioneering studies on backdoor threats against federated learning systems assume that each malicious participant individually trains their local models without any collusion among evils [4, 20, 99]. Recently, advanced attacks focus on the distributed backdoor paradigm [17, 17, 66, 110] and dynamic backdoor solution [30, 52] to evade the secure detection solutions. Thus, constructing effective backdoor defense acts as a crucial character in ensuring the robustness for federated learning.

2.3 Backdoor Defense in Federated Learning

To deal with backdoor attackers in federated learning, existing backdoor defense solutions could be basically classified into three major categories: **i)** *Distance Difference Defense* solutions [6, 20, 75, 87, 97, 106] normally compare the local party updates difference and regard those significantly far from the overall direction as evils, excluded from the aggregation process. For example, Multi Krum [6] selects the candidate gradient that is the closest to its neighboring clients. FoolsGold [20] leverages cosine similarity to identify malicious clients and allocate low weight. **ii)** *Statistics Distribution Defense* schemes [24, 83, 96, 115] introduce different mathematical statistics metrics to select and circumvent malicious clients. For instance, RFA [83] calculates the geometric median with an alternating minimization function. Bulyan [24] cooperates [115] and trimmed median to conduct a two-step meta-aggregation algorithm. Despite the certain advantages of the above two streams, they basically rely on the data homogeneity (*i.e.*, independent and identically distribution) assumption and thus are not applicable under data heterogeneous federated learning. They are sensitive to the malicious scale, which is normally hypothesized to be constrained into a certain range to guarantee overall updating correctness. **iii)** *Model Refinement Defense* efforts focus on refining the aggregated model to erase the possible backdoor attacks via the ensemble distillation [32, 50, 61, 89], Bayesian learning [11, 100, 116, 117], and certified optimization [9, 14, 48, 78, 107, 109, 118]. For instance, RLR [77] adjusts the server learning rate along with the network dimension and communication epoch. CRFL [109] controls the global smoothness via the clipping and smoothing operations. Notably, they rely on meticulous hyperparameter tuning to mitigate the catastrophic forgetting of the original task and are also limited to minor attacker assumptions for major voting. We argue that the Achilles heel for existing methods is that they conduct the client-wise defense where heterogeneous and malicious clients both present divergent optimization directions, leading to a lack of clear disentanglement. Therefore, existing methods acquire strong assumptions for minor malicious proportion or the data homogeneity, illustrated in Tab. 1. However, we calculate the parameter importance degree on both target distribution and local agnostic distribution in a self-driven manner. We regulate the parameter elements with substantial importance discrepancies during local optimization and adjust the aggregation weights based on the overall client parameter importance difference. Our solution enhances the backdoor robustness in heterogeneous federated learning.

3 Methodology

3.1 Motivation

In our work, we expect to measure the parameter importance degree for the target distribution. Notably, Fisher Information Matrix (FIM) [19] is a popular and feasible metric [41, 63, 68, 88], which quantifies the information carried by the observable random variable about the unknown parameters w of the target distribution D and is formulated as the following form:

$$\mathcal{M}(D) = \mathbb{E}_{(x,y)\in D}[\nabla \log\ p(y|x,w) \cdot (\nabla \log\ p(y|x,w))^T]. \tag{4}$$

The expectation is often approximated using the empirical distribution D. Notably, the Fisher Information Matrix could be regarded as the importance metrics on how much the perturbation of the weights affects the network output [3]. Besides, Fisher Information Matrix could also be seen as an approximation to the Hessian of the loss function [67], and hence describe the loss surface curvature for the w during the optimization. However, due to the over-parameterized network, the computation of Fisher information is unacceptable, *i.e.*, $\mathcal{M}(D) \in \mathbb{R}^{|w|\times|w|}$. To save the computational effort, Fisher Information Matrix could be approximated as the diagonal matrix as:

$$\mathcal{M}(D) \approx \mathbb{E}_{(x,y)\in D} \nabla \log\ p(y|x,w)^2 \in \mathbb{R}^{|w|}. \tag{5}$$

The empirical Fisher has the same size as the network weight, and each element in Fisher $\mathcal{M}(D)$ signifies the importance of the corresponding element in weight w for the target distribution D. Consequently, we define the parameter importance for the query element $v \in w$ on the target distribution D as follows:

$$\begin{aligned} \mathcal{I}_v(D) &= \mathcal{M}(D)[v] \\ &= \mathbb{E}_{\xi\in D} \left(\nabla_{w_v} \log\ p(y|x,w)\right)^2 . \end{aligned} \tag{6}$$

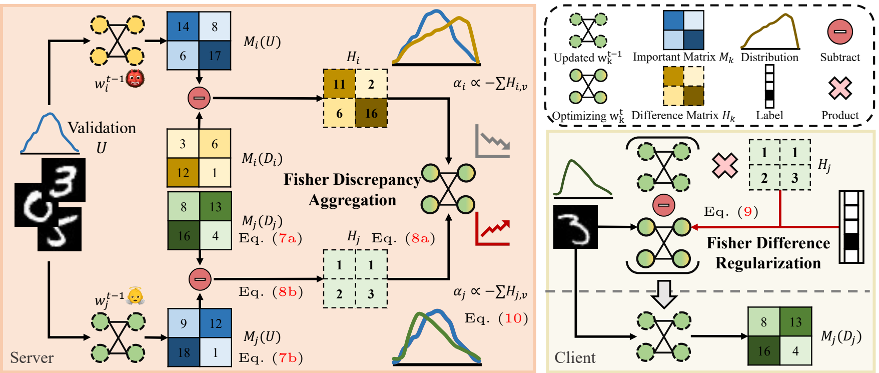
Therefore, we could utilize the $\mathcal{I}_v(D)$ as the neuron element v importance metric for the target distribution D.

3.2 SDFC: Self-Driven Fisher Calibration

Core Idea. From the Eq. (5), we can find that the affinity $\mathcal{M}(D)$ gives the influence of parameter elements on the target distribution D. We argue that for the agnostic local distribution D_k, it would be highlighted when it exhibits a higher similarity in parameter importance degree with the clean validation distribution U, and penalized when the similarity is lower. Specifically, as illustrated in Eqs. (2a) and (2b), the distributed model $w_k \in \mathbb{R}^{|w|}$ is required to fit the local distribution via the cross-entropy loss. The updated parameter is formed as $w_k \leftarrow w_k - \eta \nabla_w F(w_k, D_k)$, where η is the learning rate. Therefore, after the local updating, we calculate the Fisher Information Matrix for the w_k on the local distribution D_k and clean validation datasets U as:

$$\mathcal{M}_k(D_k) = \mathbb{E}_{(x,y)\in D_k} \nabla \log\ p(y|x,w_k)^2. \tag{7a}$$

$$\mathcal{M}_k(U) = \mathbb{E}_{(x,y)\in U} \nabla \log\ p(y|x,w_k)^2. \tag{7b}$$



(a) Collaborative Updating

(b) Local Updating

Fig. 1: Schematization of Self-Driven Fisher Calibration (Sec. 3.2). We utilize FIM to calculate parameter important matrix on local $\mathcal{M}_k(D_k)$ and validation $\mathcal{M}_k(U)$ distributions to measure importance difference $\mathcal{H}_k$ as $\boxed{6} = |\boxed{6} - \boxed{12}|$. **Fisher Difference Regularization** (Sec. 3.2) regulates those parameters with large distinct importance. **Fisher Discrepancy Aggregation** (Sec. 3.2) allocates high aggregation weight for those with small parameter important discrepancy. We illustrate with network scale $|w|=4$ and class number $|C|=5$. Best viewed in color. Zoom in for details.

Therefore, for the client k, the parameter important difference $\mathcal{H}_k$ could be defined as the following expression:

$$\begin{aligned} \mathcal{H}_k &= [\mathcal{H}_{k,1}, \ldots, \mathcal{H}_{k,v}, \ldots, \mathcal{H}_{k,|w|}] \in \mathbb{R}^{|w|}, && \text{(8a)} \\ \mathcal{H}_{k,v} &= |\mathcal{I}_{k,v}(D_k) - \mathcal{I}_{k,v}(U)| \geq 0, && \text{(8b)} \\ &= |\mathcal{M}_k(D_k)[v] - \mathcal{M}_k(U)[v]|, \end{aligned}$$

For benign clients with large local heterogeneity or malicious clients with poisoned data, corresponding $\mathcal{H}_k$ would appear as the large value. On the contrary, $\mathcal{H}_k$ would present a relatively small value or even zero. Basically speaking, we holistically explore regulating the benign heterogeneous direction and weaken the malicious backdoor effect via the $\mathcal{H}_k$ from both client-side optimization and server-side aggregation.

Fisher Difference Regularization Given that training data resides in each client, and the local updating process in Eq. (2a), we deem that vanilla cross-entropy loss encourages the private model to blindly absorb the local knowledge [103]. Thus, cross-entropy loss naturally optimizes the overall parameter to fit the local distribution and fails to discriminate whether a parameter is more important to clean or poison distribution. Therefore, such a lack of discrimination may allow drastic changes to the parameters responsible for learned clean distribution. Therefore, we introduce the Fisher Difference Regularization (FDReg) to regulate the local objective for the **current** updating parameter w_k^t at the t^{th} communication epoch as:

$$\mathcal{L}_{FDReg} = \mathcal{L}_{CE}(x,y) + \lambda \sum_{v}^{|w|} \mathcal{H}_{k,v} \times (w_{k,v}^t - w_{k,v}^{t-1})^2. \quad (9)$$

$\mathcal{H}_{k,v}$ is calculate based on the previous optimized local model w_k^{t-1}. λ is the penalization hyper-parameter for parameter importance difference. Intuitively, the FDReg allocates high flexibility (**small** $\mathcal{H}_{k,v}$) for those parameter elements with similar importance levels to fit the local data via the Cross-Entropy [15] loss. Besides, FDReg adds a strict parameter stiffness (**large** $\mathcal{H}_{k,v}$) for those with different importance. Therefore, the local model is regularized towards remembering the clean target distribution and thus alleviates the distorted local direction caused by the benign heterogeneity and malicious backdoor.

Fisher Discrepancy Aggregation Furthermore, we note the existing aggregation weight, α_k in Eq. (1) is typically based on either data scale: $\alpha_k = \frac{N_k}{\sum_k^K N_k}$ or participant scale: $\alpha_k = \frac{1}{K}$. We argue that popular aggregations overlook the local distribution reliability and should allocate higher weight to those sharing similar distribution characteristics with the target distribution. As discussed above, the distribution difference could be reacted in the parameter importance inconsistency aspect. Thus, we reconstruct the aggregation weight for the client k via the parameter important difference degree $\mathcal{H}_k \in \mathbb{R}^{|w|}$, as:

$$\begin{aligned}
\mathcal{T} &= [\sum_v^{|w|} \mathcal{H}_{1,v}, \ldots, \sum_v^{|w|} \mathcal{H}_{k,v}, \ldots, \sum_v^{|w|} \mathcal{H}_{K,v}] \in \mathbb{R}^K, \\
Max - Min \ \Downarrow \widehat{\mathcal{T}}_k &\in \frac{\mathcal{T}_k - \min(\mathcal{T})}{\max(\mathcal{T}) - \min(\mathcal{T})} \in [0,1], \\
\widehat{\mathcal{T}} &= [\widehat{\mathcal{T}}_1, \ldots, \widehat{\mathcal{T}}_k, \ldots, \widehat{\mathcal{T}}_K], \\
\alpha_k &= \frac{\sigma(-\widehat{\mathcal{T}}_k)}{\sum_k \sigma(-\widehat{\mathcal{T}}_k)}.
\end{aligned} \tag{10}$$

$\mathcal{T}_k$ denotes the sum of parameter importance difference, $\mathcal{H}_k$, for the client k. We further conduct the max-min normalization to eliminate the deep neural network scale effect. We utilize the $-\mathcal{T}_k$ to denote the client k aggregation weight since the more similar the local contribution with the target validation distribution, the larger the contribution. In each communication epoch, the server collects the updated local models and their respective FIM on the local distribution, denoted as $\mathcal{M}_k(D_k)$, and calculates the Fisher Information Matrix on the clean validation set U as $\mathcal{M}_k(U)$. Subsequently, the server computes the parameter importance difference $\mathcal{H}_k$ for further parameter aggregation reweighting in Eq. (10) and local client optimization in Eq. (9).

3.3 Discussion and Limitation

Difference with Existing Defense via Validation samples. Existing methods, including FineTuning [84], FLTrust [8], and Sageflow [80], also utilize related proxy data for evaluation selection. However, they primarily rely on empirical optimization metrics such as sample entropy and overall classification error to determine client behavior. These metrics reflect federated benign task performance (Eq. (2c)), and backdoor attacks do not adversely impact respective performance. Consequently, these solutions fall short of actively eliminating the backdoor effect. Importantly, our solution eliminates the reliance on empirical performance

Table 2: Ablative experiments of Parameter Importance Metrics, $\mathcal{H}_{k,v}$ for Fisher Difference Regularization (FDReg Sec. 3.2) on MNIST, Fashion-MNIST, and Cifar-10 ($\beta=0.5$, $\Upsilon = 30\%$). Please see Sec. 4.2 for explanations.

$\mathcal{I}_v(D)$	MNIST			Fashion-MNIST			Cifar-10		
	$\mathcal{A}$	$\mathcal{R}$	$\mathcal{V}$	$\mathcal{A}$	$\mathcal{R}$	$\mathcal{V}$	$\mathcal{A}$	$\mathcal{R}$	$\mathcal{V}$
$\lvert\nabla \log\ p(y\mid x,w)\rvert$	87.82	87.29	87.55	86.60	82.65	84.62	23.98	92.74	58.36
$\nabla \log\ p(y\mid x,w)^2$ Eq. (6)	97.92	88.01	**92.96**	84.30	88.26	**86.28**	58.73	70.12	**64.42**

indicators. Instead, we focus on the network self-behavior difference between the client distribution and the target distribution to regulate local optimization and reallocate aggregation weight for backdoor robustness in heterogeneous federated learning. Fig. 4 demonstrates that Self-Driven Fisher Calibration achieves stable and high performance with varying validation sample scales.

Related Fisher-Based Investigations. Fisher Information Matrix FIM [2,3,19] effectively encapsulates the unknown parameter information for a random distribution [35,39,67] and has attracted wide application in different research fields. [41,63,72] measures the parameter stiffness based on the past sample distribution to alleviate catastrophic phenomena on the historical class prediction in the continual learning field [16]. Besides, FIM is utilized to evaluate the critical training stage [113] and enhance the model generalization [85,121]. Closely relative method, FedCurv leverages the FIM for the personalized federation. However, all these methods regulate updates based on the parameter importance for the target distribution. SDFC measures the **parameter importance difference** between agnostic client distribution and clean validation distribution, highlighting those with a similar distribution and penalizing those with divergent distribution, as shown in Tab. 4.

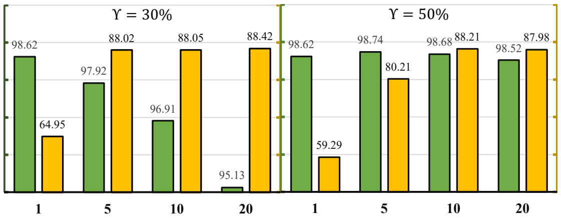
Limitation. SDFC leverages the validated dataset (U) to conduct Backdoor-robust solution in heterogeneous federated learning. However, our method fails in certain circumstances. (i) The server is required to collect the clean samples for the current federated tasks, which is widely adopted in closely related methods [1,28,61,90]. Furthermore, SDFC achieves the stable performance with the $|U| = 16$ in Fig. 4. (ii) SDFC calculates the parameter importance degree for the specific distribution. The original measurement solution requires second-order derivatives with a complexity of $\mathcal{O}(|w| \times |w|)$. However, SDFC utilizes Fisher Information Matrix to approximate the parameter importance and reduces the calculation complexity to $\mathcal{O}(|w|)$. (iii) SDFC considers that the backdoor defense from the client side and thus requires the client to execute the proposed regularization term. Thus, ours falls short of eliminating adaptive attacks, where clients maliciously refuse to obey the pre-defined optimization strategy.

4 Experiments

4.1 Experimental Setup

Datasets. Following [55,74,111], we experiment on three federated scenarios.

- **MNIST** [46] is 10 digits classes with 70,000 images.



$\varUpsilon$	$\mathcal{V}$ 1	5	10	20
30%	81.79	**92.97**	92.48	91.78
50%	78.96	89.48	**93.45**	93.25

Fig. 2: Ablation on the hyper-parameter λ (Eq. (9)) in FDReg for the proposed method on MNIST with $\varUpsilon \in \{30\%, 50\%\}$. $\mathcal{A}$ (🟩) and $\mathcal{R}$ (🟨). We default set $\lambda = 5$ in the following experiments. Refer to Sec. 4.2 for detailed information.

Table 3: Ablation on key components for SDFC in MNIST, Fashion-MNIST, Cifar-10 with $\beta = 0.5$ and $\varUpsilon = 30\%$. The λ (Eq. (9)) for FDReg is 5. See Sec. 4.2.

FDReg	FDAgg	MNIST			Fashion-MNIST			Cifar-10		
Sec. 3.2	Sec. 3.2	$\mathcal{A}$	$\mathcal{R}$	$\mathcal{V}$	$\mathcal{A}$	$\mathcal{R}$	$\mathcal{V}$	$\mathcal{A}$	$\mathcal{R}$	$\mathcal{V}$
		99.15	0.71	49.93	88.32	0.71	44.51	64.53	33.16	48.84
✓		97.92	88.01	92.96	84.30	88.26	86.28	58.73	70.12	64.42
	✓	99.13	1.55	50.34	88.24	7.57	47.90	65.88	48.63	57.25
✓	✓	97.96	89.00	**93.48**	85.10	91.13	**88.11**	62.87	87.91	**75.39**

- **Fashion-MNIST** [108] has 60k train and 10k test examples from 10 classes.
- **Cifar-10** [44] has 10 semantics with 50k, 10k images for training, validation.

As for the data heterogeneity simulation, we utilize Dirichlet distribution: $Dir(\beta)$ to simulate the label skew, as previous methods [55, 57], where $\beta > 0$ is the concentration parameter to adjust the class skewed level. The smaller β is, the more imbalanced the local distribution is. We set the β as 0.5 and 1.0 for comparison.

Counterparts. We compare with three type backdoor defense solutions.

i) Distance Difference Defense:
- FoolsGold [arXiv'18] [20]: Identify sybils effect via inter-client similarity.
- DnC [NDSS'21] [87]: Singular value decomposition for outliers detection.
- Sageflow [NeurIPS'21] [80]: Combine entropy filtering and loss reweighting.

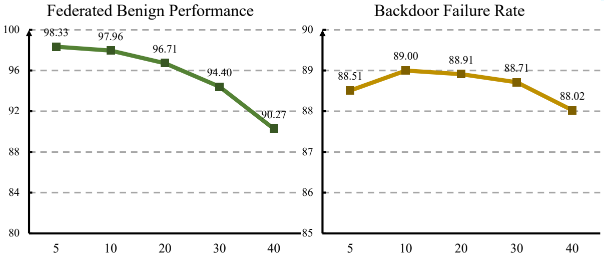
ii) Statistics Distribution Defense:
- Trim Median [ICML'18] [115]: Dimensionally remove the abnormality, based on the coordinatewise trimmed mean.
- Bulyan [ICML'18] [24]: Agree on each coordinate by major vectors, selected by Byzantine–resilient aggregations.
- RFA [TSP'22] [83]: Leverage the geometric median and the smoothed Weiszfeld algorithm to aggregate updates.

iii) Model Refinement Defense:
- RSA [AAAI'19] [53]): Norm regularization and stochastic subgradient.
- Finetuning [84]: Directly optimizes the aggregated global model on vaidation.
- RLR [AAAI'21] [77]: Adjust the aggregation server learning rate, along with dimension and communication aspects.
- CRFL [ICML'21] [109]: Exploit clipping and smoothing operations.
- FLTrust [NDSS'21] [8]: Utilize ReLU-clipped similarity to allocate trust score.

Backdoor Attacks. We demonstrate the effectiveness of the proposed method under the popular paradigm [22, 23, 26]. The size of the backdoor is set to 2×6,

Table 4: Ablation study of Parameter Evaluation Matrix, $\mathcal{H}_{k,v}$ for Fisher Difference Regularization (FDReg Sec. 3.2) on MNIST, Fashion-MNIST, and Cifar-10 (β=0.5, Υ= 30%). Please see Sec. 4.2.

$\mathcal{H}_{k,v}$	MNIST			Fashion-MNIST			Cifar-10		
	$\mathcal{A}$	$\mathcal{R}$	$\mathcal{V}$	$\mathcal{A}$	$\mathcal{R}$	$\mathcal{V}$	$\mathcal{A}$	$\mathcal{R}$	$\mathcal{V}$
$\mathcal{M}(D_k)[v]$	98.46	87.76	93.11	85.87	89.15	87.51	62.71	54.19	58.45
$\mathcal{M}(U)[v]$	98.44	59.06	78.75	86.60	70.72	78.66	62.03	34.63	48.33
$\lvert\mathcal{M}(D_k)[v]-\mathcal{M}(U)[v]\rvert$ Eq. (8b)	97.96	89.00	**93.48**	85.10	91.13	**88.11**	62.87	87.91	**75.39**



E	$\mathcal{V}$
5	93.42
10	**93.48**
20	92.81
30	91.55
40	89.14

Fig. 3: Ablation on local epoch E on MNIST (β = 0.5) with Υ = 30% for the federated benign performance $\mathcal{A}$ (Left), backdoor failure rate $\mathcal{R}$ (middle) and the trade-off $\mathcal{V}$ (Right). Refer to Sec. 4.2 for explanation.

and its location is in the top-left corner of the images. We convert the attacked label to the second class (*i.e.*, 2 in Digits). We set the malicious client ratio Υ as $\{0.3, 0.5\}$. The local data poisoned portion is default set as 0.5.

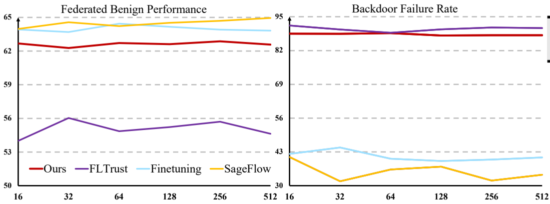
Network Structure. Following [33,55,74], we utilize the CNN as the backbone for MNIST, Fashion-MNIST, and Cifar-10.

Implement Details. We provide the details from three views as:

- Validation U Construction: We partition the original training data into training and validation sets with a 9:1 ratio. We select a small-scale validation, (*i.e.*, 256). We conduct the ablation on various validation scales and achieve stably satisfying performance with only 16 samples in Fig. 4.
- Training Setting: For a fair comparison, we follow [55, 57, 74]. We configure the communication epoch T as 50, where all approaches have little or no accuracy gain with more communications. The client number K is 10 for different datasets. For local training, we leverage the FedAvg [70] as the default local optimization objective. The local updating round is $E:10$ for different settings. We utilize the SGD as the local updating optimizer. The corresponding weight decay is $\eta:1e{-}5$ and momentum is 0.9. The local client learning rate is 0.01 in the above three scenarios. We fix the random seed to ensure reproduction and conduct experiments on the NVIDIA 3090Ti.
- Evaluation Metric: Following [26, 55, 57, 70], Top-1 accuracy is adopted for **federated benign performance**, $\mathcal{A}$ in Eq. (2c). We further define the **backdoor failure rate** as $\mathcal{R}$ in Eq. (3b). Furthermore, we define the $\mathcal{V}$ to measure the **heterogeneity and robustness trade-off** as:

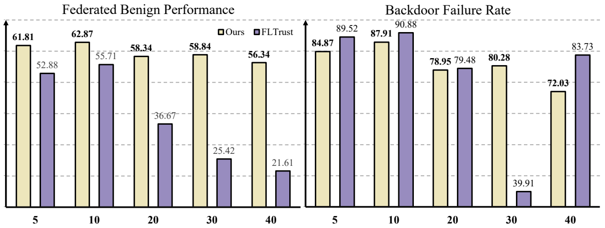
$$\mathcal{V} = \frac{1}{2}(\mathcal{A} + \mathcal{R}). \tag{11}$$

We utilize the mean performance value of the last five communication epochs as the final evaluation results.



$\lvert\mathcal{U}\rvert$	$\mathcal{V}$			
	SDFC	FLTrust	Finetuning	Sageflow
16	**75.60**	72.82	53.08	52.51
32	**75.35**	73.07	54.18	48.14
64	**75.72**	71.86	52.41	50.23
128	**75.20**	72.70	51.84	50.92
256	**75.39**	73.29	51.98	48.34
512	**75.24**	72.65	52.34	49.57

Fig. 4: Ablation on Validate Data Scale $|U|$ in Cifar-10 ($\beta = 0.5, \varUpsilon = 30\%$) with federated benign performance $\mathcal{A}$ (Left), backdoor failure rate $\mathcal{R}$ (middle) and the trade-off $\mathcal{V}$ (Right). Refer to Sec. 3.3.



K	$\mathcal{V}$	
	SDFC	FLTrust
5	**73.34**	71.20
10	**75.39**	73.29
20	**68.64**	58.07
30	**69.56**	32.66
40	**64.18**	52.67

Fig. 5: Ablation on Client scale K on Cifar-10 ($\beta = 0.5, \varUpsilon = 30\%$) with the popular counterpart. Refer to Sec. 4.2 for detailed discussion.

4.2 Diagnostic Experiments

For a thorough analysis, we perform a set of ablative studies on MNIST and Cifar-10 scenarios with label skew β=0.5 and malicious ratio $\varUpsilon$=0.3

Control Hyper-Parameter λ in Eq. (9). The Fig. 2 quantifies the effect of hyper-parameter λ, which measures the strength of parameter importance difference penalization for different scenarios. Specifically, the heterogeneity and robustness trade-off metric $\mathcal{V}$ (Eq. (11)) progressively mounts as λ enlarges, and the improvement presents marginal under strict parameter stiffness. For convenience, we choose the $\lambda = 5$ for different scenarios in the following experiments.

Parameter Evaluation Matrix Selection. As for Fisher Difference Regularization (FDReg Sec. 3.2), we evaluate the parameter stiffness $\mathcal{H}_{k,v}$ selection based on the FIM difference between the target and local distribution in Eq. (8b). However, the naive solution is to trust either client distribution $\mathcal{H}_{k,v} = \mathcal{M}(D_k)[v]$ or validated distribution $\mathcal{H}_{k,v} = \mathcal{M}(U)[v]$. Tab. 4 illustrates that they bring modest effectiveness in backdoor removal effect. This can be attributed to their inherent limitation of not placing sufficient emphasis on the relative importance of different parameters during local updating.

Training Objective. We quantitatively analyze the proposed SDFC. In Tab. 3, combining Fisher Difference Regularization (FDReg) and Fisher Discrepancy Aggregation (FDAgg) acquires satisfying federated benign task and backdoor removal performance that coincides with our motivation of exploiting the parameter importance difference for local regularization and server aggregation.

Parameter Importance Metrics. For the parameter importance metric $\mathcal{I}_v(D)$ in Eq. (6), we employ the approximate value of the second-order derivative to represent the Fisher information. A more straightforward alternative is to directly utilize the absolute gradient value to depict parameter importance, as adopted in related methods [26,91]. As demonstrated in Tab. 2, leveraging the

Table 5: Comparison with the state-of-the-art backdoor robust solutions: in the MNIST, Cifar-10, and Fashion-MNIST scenarios with skew ratio $\beta \in \{0.5, 1.0\}$ and malicious proportion $\Upsilon \in \{30\%, 50\%\}$. - means the optimization failure. Best in bold and second with underline. These notes are the same as others. Please refer to Sec. 4.3 for detailed explanations.

Methods	MNIST						Fashion-MNIST						Cifar-10					
	$\Upsilon = 30\%$			$\Upsilon = 50\%$			$\Upsilon = 30\%$			$\Upsilon = 50\%$			$\Upsilon = 30\%$			$\Upsilon = 50\%$		
	$\mathcal{A}$	$\mathcal{R}$	$\mathcal{V}$	$\mathcal{A}$	$\mathcal{R}$	$\mathcal{V}$	$\mathcal{A}$	$\mathcal{R}$	$\mathcal{V}$	$\mathcal{A}$	$\mathcal{R}$	$\mathcal{V}$	$\mathcal{A}$	$\mathcal{R}$	$\mathcal{V}$	$\mathcal{A}$	$\mathcal{R}$	$\mathcal{V}$
with skew degree $\beta = 0.5$																		
FoolsGold	96.48	0.04	48.26	90.53	0.21	45.37	82.73	2.38	42.55	80.92	0.01	40.46	55.33	45.00	50.16	51.37	10.56	30.96
DnC	98.82	89.51	**94.16**	98.54	0.09	49.31	85.94	89.22	87.58	82.64	0.00	41.32	60.91	90.26	**75.58**	51.34	11.18	31.26
Sageflow	99.16	0.60	49.88	99.08	0.35	49.71	87.88	0.21	44.04	87.70	0.00	43.85	64.69	32.00	48.34	61.90	16.24	39.07
Trim Median	95.52	89.01	92.26	98.04	0.00	49.02	10.00	100.00	55.00	10.00	100.00	55.00	47.39	62.05	54.72	10.00	100.00	55.00
Bulyan	93.40	89.10	91.25	-	-	-	-	-	-	-	-	-	40.53	94.45	67.49	-	-	-
RFA	99.24	0.09	49.66	98.77	0.04	49.40	85.56	0.02	42.79	85.45	0.00	42.72	64.22	29.53	46.87	59.65	12.76	36.20
RSA	26.56	50.34	38.45	21.95	25.36	23.65	15.09	40.36	27.72	10.54	0.68	5.61	10.00	100.00	55.0	10.65	55.95	33.30
RLR	99.19	0.39	49.79	99.05	0.15	49.60	87.91	0.20	44.05	87.63	0.02	43.82	64.00	36.98	50.48	61.44	16.12	38.78
CRFL	98.72	31.55	65.13	98.46	2.40	50.43	84.60	4.01	44.30	84.67	0.02	42.34	59.39	47.78	53.58	56.16	23.00	39.58
FLTrust	93.70	18.40	56.05	93.27	0.01	46.64	67.15	0.40	33.77	69.55	0.33	34.94	55.71	90.88	73.29	49.57	47.15	48.36
Finetuning	98.76	1.17	49.96	98.64	0.09	49.36	87.33	2.37	44.85	86.60	0.46	43.52	63.91	40.05	51.98	61.27	32.78	47.02
SDFC	97.96	88.01	92.98	97.42	56.29	**76.85**	85.10	91.13	**88.11**	84.21	87.27	**85.74**	62.87	87.91	75.39	58.70	71.15	**64.92**
with skew degree $\beta = 1.0$																		
FoolsGold	83.88	9.17	46.52	97.12	0.09	48.60	82.23	13.12	47.67	69.47	2.67	36.07	61.95	39.27	50.61	59.04	18.46	38.75
DnC	99.21	89.61	**94.41**	98.77	0.00	49.38	87.77	90.04	88.90	84.91	0.03	42.47	62.72	90.85	76.78	57.50	16.23	36.86
Sageflow	99.24	0.51	49.87	99.31	0.04	49.67	89.13	2.52	45.82	88.36	0.15	44.25	66.26	34.18	50.22	61.88	14.70	38.29
Trim Median	98.12	89.62	93.87	11.35	100	55.67	84.69	1.77	43.23	10.00	100.00	55.00	57.51	57.24	57.37	10.00	100.00	55.00
Bulyan	98.47	89.56	94.01	87.88	83.88	85.88	80.96	74.35	77.65	-	-	-	55.05	89.84	72.44	-	-	-
RFA	99.43	0.00	49.71	99.26	0.01	49.63	88.35	1.33	44.84	87.87	0.05	43.96	66.91	30.71	48.81	62.15	13.68	37.91
RSA	32.51	96.10	64.30	33.12	51.01	42.06	25.68	54.29	39.98	41.31	58.80	50.05	10.00	100.00	55.00	11.46	95.76	53.61
RLR	99.31	0.17	49.74	99.34	0.06	49.70	88.94	1.69	45.31	87.89	0.02	43.95	67.77	30.42	49.09	62.93	15.56	39.24
CRFL	99.21	37.99	68.59	98.99	3.82	51.40	88.40	27.75	58.07	88.03	3.39	45.71	64.13	44.17	54.15	60.99	23.62	42.30
FLTrust	96.11	71.30	83.70	94.78	10.25	52.51	82.31	16.52	49.41	70.01	0.31	35.16	66.67	89.38	**78.02**	57.99	53.38	55.68
Finetuning	98.93	0.08	49.50	98.87	0.02	49.44	88.08	5.32	46.70	87.17	3.39	45.28	65.53	42.69	54.11	62.35	48.75	55.55
SDFC	98.85	89.65	94.25	98.46	88.29	**93.37**	87.96	90.08	**89.02**	86.80	88.91	**87.85**	64.66	84.32	74.48	59.15	60.44	**59.79**

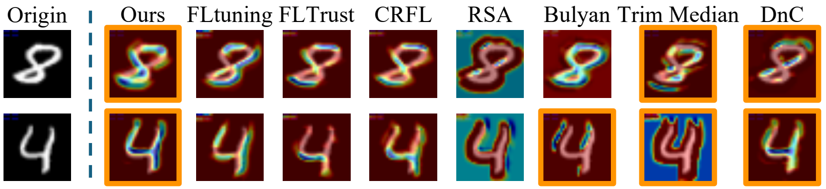
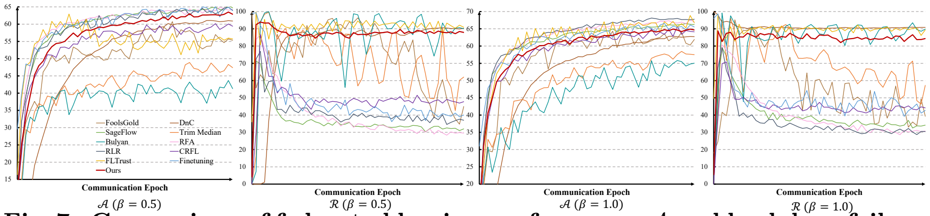


Fig. 6: Qualitative analysis under backdoor attack (Top Left) over MNIST ($\beta = 0.5, \Upsilon = 30\%$), where orange bound denotes corrected predictions. Refer to Sec. 4.3.

Fisher information to denote the parameter importance consistently yields satisfactory results. This can be attributed to the reason that the absolute gradient value fails to bring obviously value difference and thus could not clearly highlight the parameter importance discrepancy.

Local Updating Rounds. Furthermore, we analyze the effect of local updating rounds in Fig. 3. SDFC maintains a stable performance under different local rounds, indicating that SDFC achieves fast convergence in limited epochs and possesses the ability to resist client drift under various local rounds.

Client Scale K. We evaluate the performance with various participating client scale K in Fig. 5. Our SDFC achieves the competitive heterogeneity and robustness trade-off performances, demonstrating that our method is capable of dealing with the different client scale in the federated system.



$\mathcal{A}$ ($\beta = 0.5$) $\mathcal{R}$ ($\beta = 0.5$) $\mathcal{A}$ ($\beta = 1.0$) $\mathcal{R}$ ($\beta = 1.0$)

Fig. 7: Comparison of federated benign performance $\mathcal{A}$ and backdoor failure rate $\mathcal{R}$ during communication on Cifar-10 with $\Upsilon = 30\%$. SDFC appears stable convergence speed and satisfying performance. Please see details in Sec. 4.3.

4.3 Comparison to State-of-the-Arts

The Tab. 5 plots the final metric by the end of the federated learning process with popular backdoor defense methods. It clearly depicts that our method achieves the satisfying performance than different counterparts on different evaluation metrics, which confirms that SDFC effectively enhances the backdoor-robust in heterogeneous federated learning. Take the result of MNIST with $\beta = 0.5$ and $\Upsilon = 50\%$ as an example, our method outperforms the best counterpart with a gap of 27.14% on the $\mathcal{V}$ metric. Furthermore, existing backdoor defensive methods appear fragile backdoor failure rate under either the large malicious client ratio $\Upsilon = 50\%$ and serious label skew $\beta = 0.3$. It reveals that existing solutions fail to conduct the client-wise discrimination selection under large-scale evils or serious data heterogeneity. We further plot both the federated benign performance $\mathcal{A}$ and backdoor failure rate $\mathcal{R}$ during the communication process on the Cifar-10 setting in Fig. 7. We observe that SDFC presents faster and stabler convergence speed than others with different heterogeneity degrees. We Grad-CAM algorithm [86] to visualize the network attention for each input. SDFC prefer to extract key features from the image compared to other methods.

5 Conclusion

We present the Self-Driven Fisher Calibration (SDFC), the first work to achieve backdoor-robust heterogeneous federated learning from the parameter-wise. We argue that existing backdoor defensive solutions rely on either data homogeneity or minor backdoor attackers assumptions to design elaborate client-wise selection. However, we claim that benign heterogeneity and malicious backdoor bring the divergent optimization direction and instead expect to consider parameter importance for the target distribution. Therefore, we utilize the Fisher Information Matrix to measure the parameter important on local agnostic distribution and global target distribution. We allocate high regularization and low aggregation weight for those with large discrepancies in parameter importance. The effectiveness and robustness have been validated against popular counterparts with backdoor attacks under various heterogeneous federated scenarios. We hope this work provides a novel perspective to pave the way for future related research.

Acknowledgment This work is supported by the National Key Research and Development Program of China 2023YFC2705700, and National Natural Science Foundation of China under Grant (62361166629, 62176188, 62225113, 623B2080)

References

1. Afonin, A., Karimireddy, S.P.: Towards model-agnostic federated learning using knowledge distillation. In: ICLR (2022) 9
2. Amari, S.I.: Natural gradient works efficiently in learning. NC **10**(2), 251–276 (1998) 9
3. Amari, S.i., Nagaoka, H.: Methods of information geometry, vol. 191. American Mathematical Soc. (2000) 2, 6, 9
4. Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., Shmatikov, V.: How to backdoor federated learning. In: AISTATS. pp. 2938–2948 (2020) 3, 5
5. Biggio, B., Nelson, B., Laskov, P.: Poisoning attacks against support vector machines. In: ICML (2012) 4
6. Blanchard, P., El Mhamdi, E.M., Guerraoui, R., Stainer, J.: Machine learning with adversaries: Byzantine tolerant gradient descent. In: NeurIPS (2017) 2, 5
7. Cai, X., Xu, H., Xu, S., Zhang, Y., et al.: Badprompt: Backdoor attacks on continuous prompts. In: NeurIPS. vol. 35, pp. 37068–37080 (2022) 2
8. Cao, X., Fang, M., Liu, J., Gong, N.Z.: Fltrust: Byzantine-robust federated learning via trust bootstrapping. In: NDSS (2021) 2, 8, 10
9. Cao, X., Jia, J., Gong, N.Z.: Provably secure federated learning against malicious clients. In: AAAI. pp. 6885–6893 (2021) 2, 5
10. Cao, X., Zhang, Z., Jia, J., Gong, N.Z.: Flcert: Provably secure federated learning against poisoning attacks. IEEE TIFS **17**, 3691–3705 (2022) 2
11. Chen, H.Y., Chao, W.L.: Fedbe: Making bayesian model ensemble applicable to federated learning. In: ICLR (2021) 2, 5
12. Chen, X., Liu, C., Li, B., Lu, K., Song, D.: Targeted backdoor attacks on deep learning systems using data poisoning. arXiv preprint arXiv:1712.05526 (2017) 1, 4
13. Cho, Y.J., Manoel, A., Joshi, G., Sim, R., Dimitriadis, D.: Heterogeneous ensemble knowledge transfer for training large models in federated learning. In: IJCAI (2022) 2
14. Cohen, J., Rosenfeld, E., Kolter, Z.: Certified adversarial robustness via randomized smoothing. In: ICML. pp. 1310–1320. PMLR (2019) 2, 5
15. De Boer, P.T., Kroese, D.P., Mannor, S., Rubinstein, R.Y.: A tutorial on the cross-entropy method. Ann. Oper. Res. pp. 19–67 (2005) 4, 8
16. Delange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G., Tuytelaars, T.: A continual learning survey: Defying forgetting in classification tasks. IEEE PAMI pp. 1–1 (2021) 9
17. Fang, P., Chen, J.: On the vulnerability of backdoor defenses for federated learning. In: AAAI (2023) 5
18. Feng, Y., Ma, B., Zhang, J., Zhao, S., Xia, Y., Tao, D.: Fiba: Frequency-injection based backdoor attack in medical image analysis. In: CVPR. pp. 20876–20885 (2022) 2
19. Fisher, R.A.: On the mathematical foundations of theoretical statistics. Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character **222**(594-604), 309–368 (1922) 2, 6, 9
20. Fung, C., Yoon, C.J., Beschastnikh, I.: Mitigating sybils in federated learning poisoning. arXiv preprint arXiv:1808.04866 (2018) 2, 5, 10
21. Gao, L., Fu, H., Li, L., Chen, Y., Xu, M., Xu, C.Z.: Feddc: Federated learning with non-iid data via local drift decoupling and correction. In: CVPR (2022) 1, 4

22. Gu, T., Dolan-Gavitt, B., Garg, S.: Badnets: Identifying vulnerabilities in the machine learning model supply chain. arXiv preprint arXiv:1708.06733 (2017) 1, 3, 4, 10
23. Gu, T., Liu, K., Dolan-Gavitt, B., Garg, S.: Badnets: Evaluating backdooring attacks on deep neural networks. IEEE Access **7**, 47230–47244 (2019) 1, 3, 10
24. Guerraoui, R., Rouault, S., et al.: The hidden vulnerability of distributed learning in byzantium. In: ICML. pp. 3521–3530 (2018) 2, 5, 10
25. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: ICML. pp. 1321–1330 (2017) 2
26. Han, S., Park, S., Wu, F., Kim, S., Zhu, B., Xie, X., Cha, M.: Towards attack-tolerant federated learning via critical parameter analysis. In: ICCV (2023) 2, 10, 11, 12
27. Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S., Eichner, H., Kiddon, C., Ramage, D.: Federated learning for mobile keyboard prediction. arXiv preprint arXiv:1811.03604 (2018) 3
28. He, C., Annavaram, M., Avestimehr, S.: Group knowledge transfer: Federated learning of large cnns at the edge. In: NeurIPS. pp. 14068–14080 (2020) 4, 9
29. Hsieh, K., Phanishayee, A., Mutlu, O., Gibbons, P.: The non-iid data quagmire of decentralized machine learning. In: ICML. pp. 4387–4398 (2020) 4
30. Huang, A.: Dynamic backdoor attacks against federated learning. arXiv preprint arXiv:2011.07429 (2020) 5
31. Huang, L., Joseph, A.D., Nelson, B., Rubinstein, B.I., Tygar, J.D.: Adversarial machine learning. In: ACM workshop on Security and artificial intelligence. pp. 43–58 (2011) 4
32. Huang, W., Ye, M., Du, B.: Learn from others and be yourself in heterogeneous federated learning. In: CVPR (2022) 2, 5
33. Huang, W., Ye, M., Shi, Z., Li, H., Du, B.: Rethinking federated learning with domain shift: A prototype view. In: CVPR. pp. 16312–16322 (2023) 4, 11
34. Hull, J.J.: A database for handwritten text recognition research. IEEE PAMI pp. 550–554 (1994) 3
35. Jastrzebski, S., Szymczak, M., Fort, S., Arpit, D., Tabor, J., Cho, K., Geras, K.: The break-even point on optimization trajectories of deep neural networks. In: ICLR (2020) 2, 9
36. Jia, J., Liu, Y., Gong, N.Z.: BadEncoder: Backdoor attacks to pre-trained encoders in self-supervised learning. In: IEEE S&P (2022) 4
37. Jiang, W., Li, H., Xu, G., Zhang, T.: Color backdoor: A robust poisoning attack in color space. In: CVPR. pp. 8133–8142 (2023) 2
38. Karim, N., Arafat, A.A., Khalid, U., Guo, Z., Rahnavard, N.: Efficient backdoor removal through natural gradient fine-tuning. arXiv preprint arXiv:2306.17441 (2023) 2
39. Keskar, N.S., Mudigere, D., Nocedal, J., Smelyanskiy, M., Tang, P.T.P.: On large-batch training for deep learning: Generalization gap and sharp minima. In: ICLR (2017) 9
40. Kim, J., Kim, G., Han, B.: Multi-level branched regularization for federated learning. In: ICML. pp. 11058–11073 (2022) 4
41. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. PNAS pp. 3521–3526 (2017) 2, 6, 9

42. Konečnỳ, J., McMahan, H.B., Ramage, D., Richtárik, P.: Federated optimization: Distributed machine learning for on-device intelligence. arXiv preprint arXiv:1610.02527 (2016) 1, 3
43. Konečnỳ, J., McMahan, H.B., Yu, F.X., Richtárik, P., Suresh, A.T., Bacon, D.: Federated learning: Strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492 (2016) 3
44. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Master's thesis, Department of Computer Science, University of Toronto (2009) 3, 10
45. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: NeurIPS (2017) 2
46. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE pp. 2278–2324 (1998) 3, 9
47. LeCun, Y., Denker, J., Solla, S.: Optimal brain damage. In: NeurIPS. vol. 2 (1989) 2
48. Levine, A., Feizi, S.: Deep partition aggregation: Provable defense against general poisoning attacks. In: ICLR (2021) 2, 5
49. Li, B., Schmidt, M.N., Alstrøm, T.S., Stich, S.U.: Partial variance reduction improves non-convex federated learning on heterogeneous data. In: CVPR (2023) 4
50. Li, D., Wang, J.: Fedmd: Heterogenous federated learning via model distillation. In: NeurIPS Workshop (2019) 2, 5
51. Li, H., Kadav, A., Durdanovic, I., Samet, H., Graf, H.P.: Pruning filters for efficient convnets. In: ICLR (2017) 2
52. Li, H., Wu, C., Zhu, S., Zheng, Z.: Learning to backdoor federated learning. In: ICLR Workshop (2023) 5
53. Li, L., Xu, W., Chen, T., Giannakis, G.B., Ling, Q.: Rsa: Byzantine-robust stochastic aggregation methods for distributed learning from heterogeneous datasets. In: AAAI. pp. 1544–1551 (2019) 10
54. Li, Q., Diao, Y., Chen, Q., He, B.: Federated learning on non-iid data silos: An experimental study. In: ICDE. pp. 965–978 (2022) 1, 3
55. Li, Q., He, B., Song, D.: Model-contrastive federated learning. In: CVPR. pp. 10713–10722 (2021) 1, 4, 9, 10, 11
56. Li, Q., He, B., Song, D.: Adversarial collaborative learning on non-iid features. In: ICML (2023) 4
57. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. In: MLSys (2020) 1, 4, 10, 11
58. Li, Y., Jiang, Y., Li, Z., Xia, S.T.: Backdoor learning: A survey. IEEE TNNLS (2022) 1
59. Li, Y., Li, Y., Wu, B., Li, L., He, R., Lyu, S.: Invisible backdoor attack with sample-specific triggers. In: IEEE International Conference on Computer Vision (ICCV) (2021) 4
60. Liao, C., Zhong, H., Squicciarini, A., Zhu, S., Miller, D.: Backdoor embedding in convolutional neural network models via invisible perturbation. arXiv preprint arXiv:1808.10307 (2018) 1
61. Lin, T., Kong, L., Stich, S.U., Jaggi, M.: Ensemble distillation for robust model fusion in federated learning. In: NeurIPS. pp. 2351–2363 (2020) 2, 5, 9
62. Liu, Q., Chen, C., Qin, J., Dou, Q., Heng, P.A.: Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In: CVPR. pp. 1013–1023 (2021) 3

63. Liu, X., Masana, M., Herranz, L., Van de Weijer, J., Lopez, A.M., Bagdanov, A.D.: Rotate your networks: Better weight consolidation and less catastrophic forgetting. In: ICPR. pp. 2262–2268 (2018) 6, 9
64. Luo, K., Li, X., Lan, Y., Gao, M.: Gradma: A gradient-memory-based accelerated federated learning with alleviated catastrophic forgetting. In: CVPR. pp. 3708–3717 (2023) 4
65. Luo, M., Chen, F., Hu, D., Zhang, Y., Liang, J., Feng, J.: No fear of heterogeneity: Classifier calibration for federated learning with non-iid data. In: NeurIPS (2021) 4
66. Lyu, X., Han, Y., Wang, W., Liu, J., Wang, B., Liu, J., Zhang, X.: Poisoning with cerberus: stealthy and colluded backdoor attack against federated learning. In: AAAI (2023) 5
67. Martens, J.: New insights and perspectives on the natural gradient method. JMLR **21**(1), 5776–5851 (2020) 6, 9
68. Matena, M.S., Raffel, C.A.: Merging models with fisher-weighted averaging. In: NeurIPS. pp. 17703–17716 (2022) 6
69. May, C., Sell, S.K.: Intellectual property rights: A critical history. Lynne Rienner Publishers Boulder (2006) 1
70. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: AISTATS. pp. 1273–1282 (2017) 1, 4, 11
71. Miao, J., Yang, Z., Fan, L., Yang, Y.: Fedseg: Class-heterogeneous federated learning for semantic segmentation. In: CVPR. pp. 8042–8052 (2023) 3
72. Mirzadeh, S.I., Farajtabar, M., Pascanu, R., Ghasemzadeh, H.: Understanding the role of training regimes in continual learning. In: NeurIPS. pp. 7308–7320 (2020) 2, 9
73. Molchanov, P., Tyree, S., Karras, T., Aila, T., Kautz, J.: Pruning convolutional neural networks for resource efficient inference. arXiv preprint arXiv:1611.06440 (2017) 2
74. Mu, X., Shen, Y., Cheng, K., Geng, X., Fu, J., Zhang, T., Zhang, Z.: Fedproc: Prototypical contrastive federated learning on non-iid data. arXiv preprint arXiv:2109.12273 (2021) 4, 9, 11
75. Muñoz-González, L., Co, K.T., Lupu, E.C.: Byzantine-robust federated machine learning through adaptive model averaging. arXiv preprint arXiv:1909.05125 (2019) 5
76. Nagalapatti, L., Narayanam, R.: Game of gradients: Mitigating irrelevant clients in federated learning. In: AAAI. pp. 9046–9054 (2021) 2
77. Ozdayi, M.S., Kantarcioglu, M., Gel, Y.R.: Defending against backdoors in federated learning with robust learning rate. In: AAAI. pp. 9268–9276 (2021) 5, 10
78. Panda, A., Mahloujifar, S., Bhagoji, A.N., Chakraborty, S., Mittal, P.: Sparsefed: Mitigating model poisoning attacks in federated learning with sparsification. In: ICML. pp. 7587–7624. PMLR (2022) 2, 5
79. Pardau, S.L.: The california consumer privacy act: Towards a european-style privacy regime in the united states. J. Tech. L. & Pol'y **23**, 68 (2018) 1
80. Park, J., Han, D.J., Choi, M., Moon, J.: Sageflow: Robust federated learning against both stragglers and adversaries. In: NeurIPS. pp. 840–851 (2021) 2, 8, 10
81. Pascanu, R., Bengio, Y.: Revisiting natural gradient for deep networks. arXiv preprint arXiv:1301.3584 (2013) 2

82. Pati, S., Baid, U., Edwards, B., Sheller, M., Wang, S.H., Reina, G.A., Foley, P., Gruzdev, A., Karkada, D., Davatzikos, C., et al.: Federated learning enables big data for rare cancer boundary detection. Nature communications **13**(1), 7346 (2022) 3
83. Pillutla, K., Kakade, S.M., Harchaoui, Z.: Robust aggregation for federated learning. IEEE TSP **70**, 1142–1154 (2022) 5, 10
84. Quinn, J., McEachen, J., Fullan, M., Gardner, M., Drummy, M.: Dive into deep learning: Tools for engagement. Corwin Press (2019) 8, 10
85. Rame, A., Dancette, C., Cord, M.: Fishr: Invariant gradient variances for out-of-distribution generalization. In: ICML. pp. 18347–18377. PMLR (2022) 9
86. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ICCV. pp. 618–626 (2017) 14
87. Shejwalkar, V., Houmansadr, A.: Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning. In: NDSS (2021) 5, 10
88. Shoham, N., Avidor, T., Keren, A., Israel, N., Benditkis, D., Mor-Yosef, L., Zeitak, I.: Overcoming forgetting in federated learning on non-iid data. In: NeurIPS Workshop (2019) 2, 4, 6
89. Sturluson, S.P., Trew, S., Muñoz-González, L., Grama, M., Passerat-Palmbach, J., Rueckert, D., Alansary, A.: Fedrad: Federated robust adaptive distillation. arXiv preprint arXiv:2112.01405 (2021) 5
90. Sun, L., Lyu, L.: Federated model distillation with noise-free differential privacy. In: IJCAI (2021) 9
91. Sun, M., Liu, Z., Bair, A., Kolter, J.Z.: A simple and effective pruning approach for large language models. arXiv preprint arXiv:2306.11695 (2023) 12
92. Sun, Z., Kairouz, P., Suresh, A.T., McMahan, H.B.: Can you really backdoor federated learning? In: NeurIPS (2019) 2, 4
93. Tian, Y., Henaff, O.J., van den Oord, A.: Divide and contrast: Self-supervised learning from uncurated data. In: ICCV. pp. 10063–10074 (2021) 2
94. Turner, A., Tsipras, D., Madry, A.: Label-consistent backdoor attacks. arXiv preprint arXiv:1912.02771 (2019) 4
95. Voigt, P., Von dem Bussche, A.: The eu general data protection regulation (gdpr). A Practical Guide, 1st Ed., Cham: Springer International Publishing p. 3152676 (2017) 1
96. Wan, C.P., Chen, Q.: Robust federated learning with attack-adaptive aggregation. In: IJCAI Workshop (2021) 5
97. Wan, W., Hu, S., Lu, J., Zhang, L.Y., Jin, H., He, Y.: Shielding federated learning: Robust aggregation with adaptive client selection. In: IJCAI (2022) 5
98. Wang, H., Li, Y., Xu, W., Li, R., Zhan, Y., Zeng, Z.: Dafkd: Domain-aware federated knowledge distillation. In: CVPR (2023) 4
99. Wang, H., Sreenivasan, K., Rajput, S., Vishwakarma, H., Agarwal, S., Sohn, J.y., Lee, K., Papailiopoulos, D.: Attack of the tails: Yes, you really can backdoor federated learning. In: NeurIPS. pp. 16070–16084 (2020) 5
100. Wang, H., Yurochkin, M., Sun, Y., Papailiopoulos, D., Khazaeni, Y.: Federated learning with matched averaging. In: ICLR (2020) 2, 5
101. Wang, L., Zhang, K., Li, Y., Tian, Y., Tedrake, R.: Does learning from decentralized non-IID unlabeled data benefit from self supervision? In: ICLR (2023) 4
102. Wang, T., Yao, Y., Xu, F., An, S., Tong, H., Wang, T.: An invisible black-box backdoor attack through frequency domain. In: ECCV. pp. 396–413. Springer (2022) 4

103. Wei, H., Xie, R., Cheng, H., Feng, L., An, B., Li, Y.: Mitigating neural network overconfidence with logit normalization. In: ICML (2022) 7
104. Wu, C., Yang, X., Zhu, S., Mitra, P.: Mitigating backdoor attacks in federated learning. arXiv preprint arXiv:2011.01767 (2020) 2
105. Wu, D., Wang, Y.: Adversarial neuron pruning purifies backdoored deep models. In: NeurIPS. vol. 34, pp. 16913–16925 (2021) 2
106. Xia, Q., Tao, Z., Hao, Z., Li, Q.: Faba: an algorithm for fast aggregation against byzantine attacks in distributed neural networks. In: IJCAI (2019) 5
107. Xiang, C., Bhagoji, A.N., Sehwag, V., Mittal, P.: {PatchGuard}: A provably robust defense against adversarial patches via small receptive fields and masking. In: USENIX. pp. 2237–2254 (2021) 2, 5
108. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747 (2017) 10
109. Xie, C., Chen, M., Chen, P.Y., Li, B.: Crfl: Certifiably robust federated learning against backdoor attacks. In: ICML. pp. 11372–11382. PMLR (2021) 2, 5, 10
110. Xie, C., Huang, K., Chen, P.Y., Li, B.: Dba: Distributed backdoor attacks against federated learning. In: ICLR (2020) 3, 5
111. Xie, Y., Zhang, W., Pi, R., Wu, F., Chen, Q., Xie, X., Kim, S.: Optimizing server-side aggregation for robust federated learning via subspace training. arXiv preprint arXiv:2211.05554 (2022) 2, 9
112. Xiong, Y., Wang, R., Cheng, M., Yu, F., Hsieh, C.J.: Feddm: Iterative distribution matching for communication-efficient federated learning. In: CVPR (2023) 4
113. Yan, G., Wang, H., Li, J.: Seizing critical learning periods in federated learning. In: AAAI (2022) 9
114. Yang, Q., Liu, Y., Chen, T., Tong, Y.: Federated machine learning: Concept and applications. ACM TIST pp. 1–19 (2019) 1
115. Yin, D., Chen, Y., Kannan, R., Bartlett, P.: Byzantine-robust distributed learning: Towards optimal statistical rates. In: ICML. pp. 5650–5659 (2018) 2, 5, 10
116. Yurochkin, M., Agarwal, M., Ghosh, S., Greenewald, K., Hoang, N.: Statistical model aggregation via parameter matching. In: NeurIPS (2019) 5
117. Yurochkin, M., Agarwal, M., Ghosh, S., Greenewald, K., Hoang, N., Khazaeni, Y.: Bayesian nonparametric federated learning of neural networks. In: ICML (2019) 2, 5
118. Zhang, K., Tao, G., Xu, Q., Cheng, S., An, S., Liu, Y., Feng, S., Shen, G., Chen, P.Y., Ma, S., et al.: Flip: A provable defense framework for backdoor mitigation in federated learning. In: ICLR (2023) 2, 5
119. Zhang, Z., Lyu, L., Wang, W., Sun, L., Sun, X.: How to inject backdoors with better consistency: Logit anchoring on clean data. In: ICLR (2022) 2
120. Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., Chandra, V.: Federated learning with non-iid data. arXiv preprint arXiv:1806.00582 (2018) 1
121. Zhong, Q., Ding, L., Shen, L., Mi, P., Liu, J., Du, B., Tao, D.: Improving sharpness-aware minimization with fisher mask for better generalization on language models. In: EMNLP (2022) 2, 9
122. Zhou, T., Konukoglu, E.: FedFA: Federated feature augmentation. In: ICLR (2023) 4