DarkFed: A Data-Free Backdoor Attack in Federated Learning

Minghui Li[1] , **Wei Wan**[2,3,4,5,6] , **Yuxuan Ning**[7] , **Shengshan Hu**[2,3,4,5,6] ,
Lulu Xue[2,3,4,5,6] , **Leo Yu Zhang**[8] , **Yichen Wang**[2,3,4,5,6]
[1]School of Software Engineering, Huazhong University of Science and Technology
[2]National Engineering Research Center for Big Data Technology and System
[3]Services Computing Technology and System Lab
[4]Hubei Engineering Research Center on Big Data Security
[5]Hubei Key Laboratory of Distributed System Security
[6]School of Cyber Science and Engineering, Huazhong University of Science and Technology
[7]School of Computer Science and Technology, Huazhong University of Science and Technology
[8]School of Information and Communication Technology, Griffith University
{minghuili, wanwei_0303, ningyuxuan, hushengshan, lluxue, wangyichen}@hust.edu.cn,
leo.zhang@griffith.edu.au

Abstract

Federated learning (FL) has been demonstrated to be susceptible to backdoor attacks. However, existing academic studies on FL backdoor attacks rely on a high proportion of real clients with main task-related data, which is impractical. In the context of real-world industrial scenarios, even the simplest defense suffices to defend against the state-of-the-art attack, 3DFed. A practical FL backdoor attack remains in a nascent stage of development.

To bridge this gap, we present DarkFed. Initially, we emulate a series of fake clients, thereby achieving the attacker proportion typical of academic research scenarios. Given that these emulated fake clients lack genuine training data, we further propose a data-free approach to backdoor FL. Specifically, we delve into the feasibility of injecting a backdoor using a shadow dataset. Our exploration reveals that impressive attack performance can be achieved, even when there is a substantial gap between the shadow dataset and the main task dataset. This holds true even when employing synthetic data devoid of any semantic information as the shadow dataset. Subsequently, we strategically construct a series of covert backdoor updates in an optimized manner, mimicking the properties of benign updates, to evade detection by defenses. A substantial body of empirical evidence validates the tangible effectiveness of DarkFed.

1 Introduction

Federated learning (FL) [McMahan *et al.*, 2017; Lu *et al.*, 2023], one of the prevailing distributed paradigms, facilitates the collaborative construction of a high-precision global model by multiple clients with small amounts of data, all under the coordination of a central server. Notably, FL excels at preserving privacy since clients' training data remains localized throughout the entire model construction process.

Attacks	**20% Attackers**		**1% Attackers**	
	ACC (%)	**ASR (%)**	**ACC (%)**	**ASR (%)**
Model Replacement	90.07	97.93	90.64	0.53
3DFed	90.14	98.71	90.36	0.52

Table 1: Performance of existing backdoor attacks in academic research scenarios and real-world industrial scenarios.

However, the distributed nature of FL also presents a significant challenge: the central server struggles to discern the quality of client-uploaded parameters. Consequently, FL faces a severe threat known as poisoning attacks [Lu *et al.*, 2024; Shi *et al.*, 2022; Wan *et al.*, 2021]. These attacks can be categorized into two main types: Byzantine attacks [Zhang *et al.*, 2023b; Wan *et al.*, 2024], and backdoor attacks [Zhang *et al.*, 2024a; Zhang *et al.*, 2024b] The former aims to reduce the global model's recognition accuracy for all samples, while the latter specifically misclassifies samples specified by the adversary without affecting the model's recognition of normal samples. This indicates that backdoor attacks are more covert and insidious compared to Byzantine attacks. Therefore, this paper focuses on backdoor attacks in FL.

FL is shown to be susceptible to backdoor attacks [Xie *et al.*, 2020; Lyu *et al.*, 2023; Li *et al.*, 2023]. However, the success of these attacks critically hinges on a high proportion of genuine attackers possessing samples relevant to the main task. Typically, they require 20% of attackers with authentic training data to successfully inject a backdoor. In real-world industrial scenarios [Shejwalkar *et al.*, 2022], attackers often constitute only 1% or even less of the total clients. As shown in Tab. 1, for both the classical Model Replacement Attack [Bagdasaryan *et al.*, 2020] and the recent 3DFed [Li *et al.*, 2023], we consider scenarios with 20% attackers (academic research scenarios) and 1% attackers (real-world industrial scenarios). Notably, we employ only the most primitive defense method, Norm Clipping [Wang *et al.*, 2020], which restricts the magnitude of local updates to remain within a specified threshold. We observe that in academic research scenarios, these attacks can indeed achieve significant *attack success rate* (ASR) and *accuracy of the model* (ACC). However, surprisingly, in real-world industrial scenarios, even the *state-of-the-art* (SOTA) 3DFed fails to backdoor FL equipped with the simplest defense. We speculate that this is due to the low proportion of attackers, which results in the backdoor task-related knowledge being overshadowed by the main task-related knowledge in the aggregation stage. The result suggests that existing backdoor attacks in FL are impractical, and an effective FL backdoor attack for real-world industrial scenarios is yet to be developed.

In light of this, we embark on the initial steps toward developing backdoor attacks in FL tailored for real-world industrial contexts. Building upon the research in [Cao and Gong, 2022], we can emulate a series of fake clients using open-source projects or Android emulators. These approaches can significantly increase the number of attackers to match the settings of academic research scenarios. However, these emulated fake clients are unable to provide authentic main task-related data. Consequently, the primary challenge pivots towards devising a data-free backdoor attack in FL.

In this paper, we propose DarkFed, the first **DA**ta-f**R**ee bac**K**door attack in **FED**erated learning. Specifically, we first explore the impact of shadow datasets on backdoor attacks. Surprisingly, even when there is a substantial gap between the shadow dataset and the main task dataset (*e.g.*, between CIFAR-10 and GTSRB), the backdoor can be successfully implanted while maintaining model utility. What's even more astonishing is that using synthetic data devoid of any semantic information (*e.g.*, generated through a Gaussian distribution) as the shadow dataset still yields significant success in backdoor attacks. These promising results inspire us to inject the backdoor using a shadow dataset on the emulated fake clients. However, directly transferring the previous process is prone to detection by existing defenses due to the significant differences between backdoor updates and benign updates, leading to the failure of the attack. To further enhance the stealthiness of the attack, we propose property mimicry, optimizing backdoor updates to mimic benign updates in terms of magnitude, distribution, and consistency. These properties are widely employed by FL backdoor defenses to detect backdoor updates. This optimization significantly boosts the covert nature of the attack.

In summary, our contributions are as follows:

- We introduce DarkFed, the first data-free backdoor attack in FL. This attack does not rely on task-specific data, enabling its use in scenarios with emulated fake clients, thus achieving a practical backdoor attack.
- We investigate the feasibility of injecting a backdoor with shadow datasets and find that even with synthetic datasets, successful backdoor injection is achievable. We extend this concept into the realm of FL.
- We introduce a novel defense evasion technique, property mimicry, which enables backdoor updates to mimic the properties of benign updates, thereby enhancing the stealthiness of the attack.
- Extensive experiments demonstrate that DarkFed achieves attack effects comparable to SOTA data-dependent attacks.

2 Related Work

2.1 Backdoor Attacks in FL

A plethora of studies have suggested that FL is exceptionally susceptible to backdoor attacks [Mo *et al.*, 2023; Liu *et al.*, 2023; Hu *et al.*, 2023; Hu *et al.*, 2022]. [Bagdasaryan *et al.*, 2020] is among the pioneers in launching backdoor attacks on FL. They introduce the Model Replacement Attack, amplifying the magnitude of backdoor updates proportionally, ensuring the dominance of backdoor parameters in the global model. Furthermore, they introduce the Semantic Backdoor Attack, which doesn't require any modifications to the training samples but leverages samples with specific semantic information to trigger the backdoor. For example, this attack classifies all green cars as horses. Drawing inspiration from the Semantic Backdoor Attack, [Wang *et al.*, 2020] introduce an edge-case backdoor attack, which uses rare samples (the tail of a dataset) to trigger the backdoor. [Xie *et al.*, 2020] propose DBA (*Distributed Backdoor Attack*), which decomposes a trigger into multiple sub-triggers, with each attacker holding one of these sub-triggers for data poisoning. Most recently, [Li *et al.*, 2023] present 3DFed, addressing three prominent defense strategies with corresponding attack modules and introducing an indicator mechanism to assess whether backdoor updates are used in model aggregation. This allows for an adaptive adjustment of the attack strategy. We've also noticed a category of backdoor attacks based on trigger optimization, such as A3FL [Zhang *et al.*, 2023a] and F3BA [Fang and Chen, 2023]. They aim to obtain a robust trigger to make the attacks more covert and persistent. These efforts are compatible with the ones mentioned earlier.

It's important to note that all existing backdoor attacks are data-dependent, meaning they require main task-related data to operate. The development of a data-free backdoor attack remains an open area for exploration.

2.2 Backdoor Defenses in FL

Existing defenses against FL backdoors can be categorized into norm constraint-based defenses, outlier detection-based defenses, and consistency detection-based defenses.

Norm constraint-based defenses posit that the optimal point for the backdoor task typically deviates significantly from the optimal point for the main task. This results in the norm of backdoor updates being much larger than that of benign updates. Consequently, these defenses constrain the norm of all local updates within a reasonable range. Norm Clipping [Wang *et al.*, 2020] serves as a representative example of such defenses. Additionally, some other defenses [Wan *et al.*, 2023; Wan *et al.*, 2022; Cao *et al.*, 2021] also leverage this characteristic to prevent malicious updates from dominating the global model.

Outlier detection-based defenses assert that backdoor updates and benign updates exhibit substantial differences in their distributions, with benign updates typically being densely distributed. In contrast, backdoor updates can be considered as outliers. Building on this premise, RFLBAT [Wang *et al.*, 2022] utilizes Principal Component Analysis (PCA) to project local updates into a low-dimensional space. Subsequently, it employs a clustering algorithm to identify outliers, marking them as backdoor updates. FLAME [Nguyen *et al.*, 2022] identifies updates that deviate significantly in direction from the overall trend as backdoor updates and excludes them from the aggregation queue. FLDetector [Zhang *et al.*, 2022] exploits the differences between the predicted model and the actual model to discover outliers.

Consistency detection-based defenses argue that all backdoor updates share the same objective, namely, to classify trigger-carrying samples as the target label. Therefore, these updates exhibit strong consistency, either in terms of update directions or neuron activations. On the other hand, diverse benign updates may display lower consistency due to data heterogeneity [Li *et al.*, 2020]. With this understanding, FoolsGold [Fung *et al.*, 2020] assigns lower aggregation weights to updates with high pairwise cosine similarities, thereby mitigating the impact of backdoor updates. DeepSight [Rieger *et al.*, 2022] uses the consistency on neuron activations in the backdoor model to detect malicious updates.

3 Preliminaries

3.1 Federated Learning

FL involves iteratively exchanging model parameters between a central server and multiple clients, enabling collaborative training on distributed private data. In broad strokes, FL encompasses the following steps:

- **Step I:** The central server dispatches the global model w to individual clients.
- **Step II:** Each client individually fine-tunes the global model w on their local dataset, resulting in a refined model w'. Subsequently, the model update $u' := w' - w$ is uploaded to the server.
- **Step III:** The server utilizes all the updates (denoted as a set S) to enhance the global model as follows:

  $$w \leftarrow w + AGR(\{u' | u' \in S\}). \tag{1}$$

  Here, AGR denotes the aggregation algorithm, such as FedAvg [McMahan *et al.*, 2017].

These steps are iteratively performed until a satisfactory global model is obtained. Note that, for the sake of conciseness, we have omitted the iteration round and the client index.

4 Threat Model

4.1 Adversary's Goal

The adversary should successfully inject a backdoor without compromising the availability of the global model, even in the presence of SOTA backdoor defenses deployed on the central server. This implies that the adversary needs to concurrently achieve the following three goals:

- **Stealthiness.** Backdoor updates must effectively masquerade as benign updates, evading detection by defense schemes to ensure that the backdoor-related knowledge becomes integrated into the global model.
- **Fidelity.** The backdoor attack should not undermine the global model's ability to recognize clean samples. The post-attack global model must maintain a level of accuracy in the main task comparable to its pre-attack state.
- **Effectiveness.** In the case of samples containing the trigger, the global model should exhibit a very high accuracy in recognizing them as the target class, as predetermined by the adversary.

4.2 Adversary's Capability and Knowledge

We posit that the adversary can emulate a series of fake clients (also referred to as attackers) using open-source projects or Android emulators [Cao and Gong, 2022] to attain an attacker proportion that aligns with the common academic research scenario, typically around 20%. Note that these fake clients lack access to main task-relevant data. The adversary is entirely unaware of the training data and model updates of benign clients, as well as the defense scheme deployed on the server. Furthermore, it cannot disrupt the model training process of benign clients or server decision-making. The sole element within the adversary's control is the training process of the fake clients. The fake clients can mimic the behavior of benign clients to launch covert backdoor attacks.

5 DarkFed

5.1 Motivation Behind DarkFed

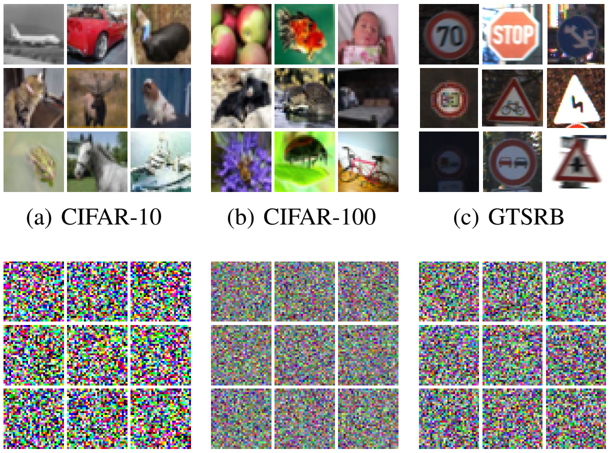
Drawing insights from the results in Tab. 1, it becomes apparent that a high attacker proportion is an indispensable prerequisite for the success of backdoor attacks. Nevertheless, achieving such a high attacker proportion is usually unattainable in real-world industrial settings, rendering the existing FL backdoor attack methods impractical. While some literature [Cao and Gong, 2022] suggests the emulation of a substantial number of fake clients to match the attacker proportion of academic research scenarios, these fake clients are bereft of task-relevant data. Consequently, we are compelled to execute the backdoor attack in a data-free manner.

5.2 Backdoor Attack with Shadow Dataset

To realize the data-free backdoor attack, it naturally prompts us to explore the utility of a shadow dataset, because obtaining diverse data unrelated to the main task is quite easy. For example, it can be achieved through web scraping of publicly available data or even generating a series of data points using Gaussian distribution. Consequently, we follow [Lv *et al.*, 2023] and embark on exploring the impact of shadow datasets on backdoor attacks.

Formally, for a clean model w and a shadow dataset D_s, we fine-tune w into a backdoored version w' with the following optimization objective:

$$\begin{aligned} &\min_{w'} L = L_{cl} + \lambda_1 L_{bk}, \\ &L_{cl} = \sum_{x_i \in D_{sc}} \mathcal{L}\left(w'\left(x_i\right), w\left(x_i\right)\right), \\ &L_{bk} = \sum_{\widetilde{x}_i \in D_{sp}} \mathcal{L}\left(w'\left(\widetilde{x}_i\right), y_t\right), \end{aligned} \tag{2}$$



(a) CIFAR-10 (b) CIFAR-100 (c) GTSRB

(d) Gauss-I (e) Gauss-II (f) Uniform

Figure 1: Visual comparison of the shadow datasets.

Shadow Dataset	CIFAR-10		CIFAR-100		GTSRB	
	ACC (%)	ASR (%)	ACC (%)	ASR (%)	ACC (%)	ASR (%)
CIFAR-10	89.17	100.00	78.97	100.00	93.17	100.00
CIFAR-100	89.14	100.00	79.09	100.00	93.13	100.00
GTSRB	88.93	99.81	78.46	100.00	93.34	100.00
Gauss-I	88.90	98.36	78.06	96.72	93.08	99.35
Gauss-II	89.02	79.63	78.19	75.25	92.60	83.54
Uniform	89.06	93.19	78.17	95.45	93.23	97.99

Table 2: Impact of shadow datasets on backdoor performance.

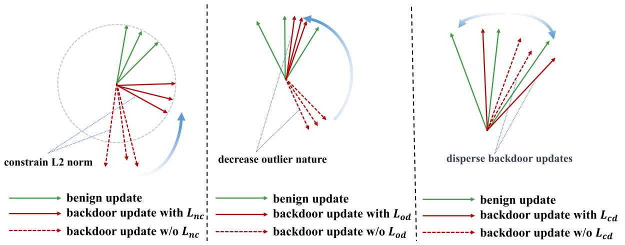

constrain L2 norm
benign update
backdoor update with L_{nc}
backdoor update w/o L_{nc}
decrease outlier nature
benign update
backdoor update with L_{od}
backdoor update w/o L_{od}
disperse backdoor updates
benign update
backdoor update with L_{cd}
backdoor update w/o L_{cd}

Figure 2: Illustration of property mimicry.

where D_{sc} represents the clean dataset without any modifications in the shadow dataset, and D_{sp} represents the poisoned dataset with triggers applied. $\mathcal{L}$ is a loss function, *e.g.*, cross entropy. $w(\cdot)$ and $w'(\cdot)$ denote the logits of w and w', respectively, y_t is the target label, and the hyperparameter λ_1 is used to balance the model performance and the poisoning effect. The purpose of L_{cl} is to maintain the performance of the main task, while L_{bk} aims to learn the knowledge related to the backdoor. It should be noted that throughout the entire fine-tuning process, λ_1 is fixed at 1, and the clean model w is treated as a constant that remains unchanged.

We consider three popular image classification tasks: CIFAR-10 [Krizhevsky and Hinton, 2009], CIFAR-100 [Krizhevsky and Hinton, 2009], and GTSRB [Stallkamp *et al.*, 2011]. As for the shadow datasets, in addition to these three real datasets, we also include three synthetic datasets constructed based on certain distributions. These synthetic datasets do not contain any semantic information. Due to the commonality of the three real datasets, we omit their introduction here and focus solely on describing the three synthetic datasets.

- **Gauss-I.** Each image is of size $32 \times 32 \times 3$, with each pixel value generated from a Gaussian distribution $N(0.5, 1^2)$ and located within the range $[0, 1]$.
- **Gauss-II.** Each image is of size $32 \times 32 \times 3$, with each pixel value generated from a Gaussian distribution $N(0.5, 0.2^2)$ and located within the range $[0, 1]$.
- **Uniform.** Each image is of size $32 \times 32 \times 3$, with each pixel value generated from a uniform distribution $U(0, 1)$.

Fig. 1 showcases a subset of samples from these datasets. Notably, significant visual disparities exist among them, which potentially leads to the presumption that their utilization as shadow datasets would lead to a reduction in the model's main task recognition accuracy (*i.e.*, ACC) or an unsatisfactory level of backdoor task accuracy (*i.e.*, ASR). However, to our surprise, as shown in Tab. 2, the shadow dataset, compared to directly using task-related datasets (highlighted in green), does not significantly impact the backdoor performance, especially when utilizing real datasets. The last three rows in the table illustrate the scenario when synthetic datasets serve as shadow datasets. It can be observed that the performance of Gauss-I is comparable to that of real datasets. Gauss-II fails to achieve satisfactory ASR, and we speculate this is due to its small standard deviation during construction, resulting in lower data richness and thus poorer performance. Uniform's performance falls between Gauss-I and Gauss-II, achieving over 90% ASR on each dataset.

We posit that the ability to maintain high ACC with a shadow dataset lies in our strategy of ensuring similarity between the logits of w' and w (see L_{cl} in Eq. (2)), rather than employing hard labels. This allows for minor adjustments to be made to w' based on w. The achievement of high ASR is attributed to the fact that backdoor learning focuses on the mapping relationship between the trigger and the target label, making it less influenced by the shadow dataset itself.

5.3 Property Mimicry

The preceding exploration indicates that leveraging a shadow dataset and Eq. (2) can simultaneously achieve high ACC (the fidelity goal) and high ASR (the effectiveness goal). Therefore, a naive idea is to directly execute this on fake clients. However, this approach is easily thwarted by existing defense mechanisms, leading to a failure in backdoor implantation (*i.e.*, the stealthiness goal is not achieved). To enhance stealthiness, our core idea is to make the backdoor updates generated by fake clients mimic benign updates in terms of properties (*e.g.*, magnitude, distribution, and consistency). This approach makes it challenging to distinguish between these two types of updates based on certain properties, thereby evading defense mechanisms.

Moderate magnitude. The magnitude of benign updates is typically moderate, and this feature is also exploited by norm constraint-based defenses to filter or constrain backdoor updates. To mimic this property of benign updates, an intuitive approach is to prevent the magnitude of backdoor updates from becoming excessively large, and this can be effortlessly achieved by incorporating a constraint term.

Formally, for the global model w and the local model w', we consider the following constraint term:

$$L_{nc} = ||w' - w||_2. \tag{3}$$

The inspiration for this constraint term is drawn from FedProx [Li *et al.*, 2020], which also incorporates an identical constraint term. However, our approach differs fundamentally from FedProx. Firstly, FedProx employs this constraint term to alleviate the issue of model accuracy degradation caused by statistical heterogeneity in FL, whereas our objective is to enhance the stealthiness of backdoor attacks. Furthermore, FedProx applies this constraint term to all clients, whereas we restrict its use solely to malicious clients. While this constraint term is simple and intuitive, it effectively restricts the magnitude of backdoor updates without compromising the efficiency of backdoor injection or the backdoor accuracy. We surmise that L_{nc} can guide the backdoor model to search for a joint optimal point near the global model, where both the backdoor task and the primary task perform well. The effect of L_{nc} is illustrated in Fig. 2 (left).

Reasonable distribution. Benign updates typically exhibit a reasonable distribution, so when malicious updates are introduced, they can be detected as outliers. To mimic this characteristic of benign updates, a favorable countermeasure involves narrowing the similarity between backdoor updates and benign updates. This confuses the defense mechanism, making it challenging to detect *out-of-distribution* (OOD) values and potentially leading to erroneous identifications (*i.e.*, classifying benign updates as malicious). Considering the well-established capability of cosine similarity in measuring update similarity, and its widespread adoption in various defense methods [Nguyen *et al.*, 2022; Cao *et al.*, 2021], we employ it as a metric to quantify the similarity between malicious and benign updates. Nevertheless, the adversary lacks knowledge of benign models. Although it can train a set of emulated benign models using the clean shadow dataset D_{sc} and L_{cl} with Eq. (2), the emulated benign models may deviate substantially from real benign ones. This discrepancy increases the risk of backdoor models veering in the wrong direction, thereby amplifying the visibility of the attack.

In this context, we propose employing the *double exponential smoothing* (DES) algorithm to predict the forthcoming round's global model because DES has exhibited remarkable effectiveness in predicting the distribution of benign models [Li *et al.*, 2022]. Formally, for the global model w and the local model w', we introduce the following constraint term:

$$L_{od} = (\cos(w' - w, \hat{w} - w) - \alpha)^2, \tag{4}$$

where $\hat{w}$ denotes the predicted global model through DES, cos denotes cosine similarity, and the hyperparameter α represents the estimated cosine similarity between benign updates. The constraint term L_{od} tightens the similarity between backdoor updates and benign updates, preventing the backdoor updates from resembling OOD values. This allows them to circumvent outlier detection-based defenses. The effect of L_{od} is illustrated in Fig. 2 (middle).

Algorithm 1: A Complete Description of DarkFed

```
Input : $D_s$: shadow dataset; $w$: global model; $W_{bk}$:
        collection of all backdoor models; $\alpha$:
        estimated cosine similarity between benign
        updates; $E$: local epoch; $B$: batch size; $\lambda$:
        coefficient balancing stealthiness with
        fidelity and effectiveness; $\eta$: learning rate.
Output: the backdoored models $W_{bk}$
// Initialize all backdoor models with global model
for $w' \in W_{bk}$ do
    $w' \leftarrow w$
Obtain the predicted global model $\hat{w}$ through DES
$\mathcal{B} \leftarrow$ (split $D_s$ into batches of size $B$)
// Optimize each backdoor model following Eq. (6)
for each epoch $e \in [1, E]$ do
    for $w' \in W_{bk}$ do
        for each batch $b \in \mathcal{B}$ do
            $L = L_{cl} + L_{bk} + \lambda(L_{nc} + L_{od} + L_{cd})$
            $w' \leftarrow w' - \eta \nabla_{w'} L$
```

Limited consistency. The consistency of benign updates is typically limited, as they do not share the same objective, unlike backdoor updates. This principle forms the core perspective of consistency detection-based defenses. To simulate this property of benign updates, we aim to optimize the representation of backdoor updates, ensuring they demonstrate a consistency akin to benign updates. Specifically, for the global model w and the local model w', we introduce the following constraint term:

$$L_{cd} = \sum_{w'' \in W_{bk} - \{w'\}} (\cos(w' - w, w'' - w) - \alpha)^2, \tag{5}$$

where W_{bk} represents the collection of all backdoor models, and α carries the same meaning as in Eq. (4). The constraint term L_{cd} reduces the consistency of backdoor updates to a level comparable to benign updates, effectively confounding defensive strategies. The effect of L_{cd} is illustrated in Fig. 2 (right).

5.4 A Complete Description of DarkFed

Combining constraint terms (3), (4), and (5), we can reformulate the optimization objective in (2) as follow:

$$\min_{w'} L = L_{cl} + L_{bk} + \lambda(L_{nc} + L_{od} + L_{cd}), \tag{6}$$

where L_{cl} is designed to achieve the fidelity goal. L_{bk} aims to fulfill the effectiveness goal. L_{nc}, L_{od}, and L_{cd} contribute to realizing the stealthiness goal. λ is a coefficient balancing stealthiness with fidelity and effectiveness. Note that this optimization objective is extensible. While it currently encompasses evasion constraint terms designed for existing defense categories, new defenses may emerge in the future that do not fall into any of these categories. In such cases, we can still design corresponding constraint terms to extend this optimization objective.

Alg. 1 provides a complete description of the DarkFed scheme during a round of global iteration. Upon receiving

CIFAR-10		**CIFAR-100**		**GTSRB**	
ACC (%)	**ASR (%)**	**ACC(%)**	**ASR(%)**	**ACC (%)**	**ASR (%)**
90.15	8.77	79.01	0.79	93.08	2.93

Table 3: Performance of the initial global model.

Dataset	α	λ	η	E	B	D_s
CIFAR-10	0	0.5	0.005	15	64	GTSRB
CIFAR-100	0	0.5	0.001	15	64	GTSRB
GTSRB	0	0.5	0.00005	15	64	CIFAR-100

Table 4: Patameter settings in Alg. 1.

the global model from the central server, all fake clients initialize their local models with this global model (Lines 1-3). Subsequently, the DES algorithm is employed to obtain the predicted global model (Line 4), which is utilized in the computation of L_{od} (refer to Eq. (4)). Then the shadow dataset is divided into multiple batches for local training (Line 5). Finally, optimization is performed for each backdoor model based on Eq. (6) (Lines 6-11).

6 Experiments

6.1 Experimental Setup

Datasets, models, and codes. We consider three multi-channel image classification datasets: CIFAR-10 [Krizhevsky and Hinton, 2009], CIFAR-100 [Krizhevsky and Hinton, 2009], and GTSRB [Stallkamp *et al.*, 2011]. Because, compared to single-channel datasets like MNIST [LeCun *et al.*, 1998] and Fashion-MNIST [Xiao *et al.*, 2017], multi-channel datasets are more complex and better represent real-world scenarios. For CIFAR-10 and CIFAR-100, we employ ResNet-18 as the model structure. For GTSRB, we construct a VGG-like model as the global model. It's worth noting that, to expedite experimentation, we follow [Li *et al.*, 2023], employing a pre-trained model as the initial global model to simulate a scenario where the global model is nearing convergence. The initial model's ACC and ASR on the three datasets are provided in Tab. 3. Note that when calculating ASR, samples corresponding to the target label have not been excluded. This results in some backdoor samples being identified as the target class not because they are triggered, but because they inherently belong to the target class. Consequently, this leads to a relatively higher ASR. This approach is justified since backdooring a FL system when the global model is close to convergence is enough. Our codes will be available at https://github.com/hustweiwan/DarkFed.

Shadow datasets. For the CIFAR-10 and CIFAR-100 classification tasks, we employ GTSRB as the shadow dataset. Conversely, for the GTSRB classification task, we utilize CIFAR-100 as the shadow dataset. This decision is guided by the relatively small domain gap between CIFAR-10 and CIFAR-100, and we avoid using them interchangeably as shadow datasets.

Attack settings. In line with [Lyu *et al.*, 2023], we establish a FL system encompassing 100 clients, with 20% of them being emulated fake clients. These fake clients lack training data relevant to the main task and instead utilize a publicly scraped dataset or a synthetic dataset (*e.g.*, generated through Gaussian distribution) to introduce a backdoor. 20% of the total clients are randomly selected in each iteration. The parameter settings in Alg. 1 are delineated in Tab 4. One might wonder why the estimated cosine similarity between benign updates (*i.e.*, α) consistently remains at 0 for different datasets. This is attributed to the research in [Wan *et al.*, 2022], which indicates that benign updates exhibit similarity only in the initial rounds, becoming nearly orthogonal in subsequent iterations.

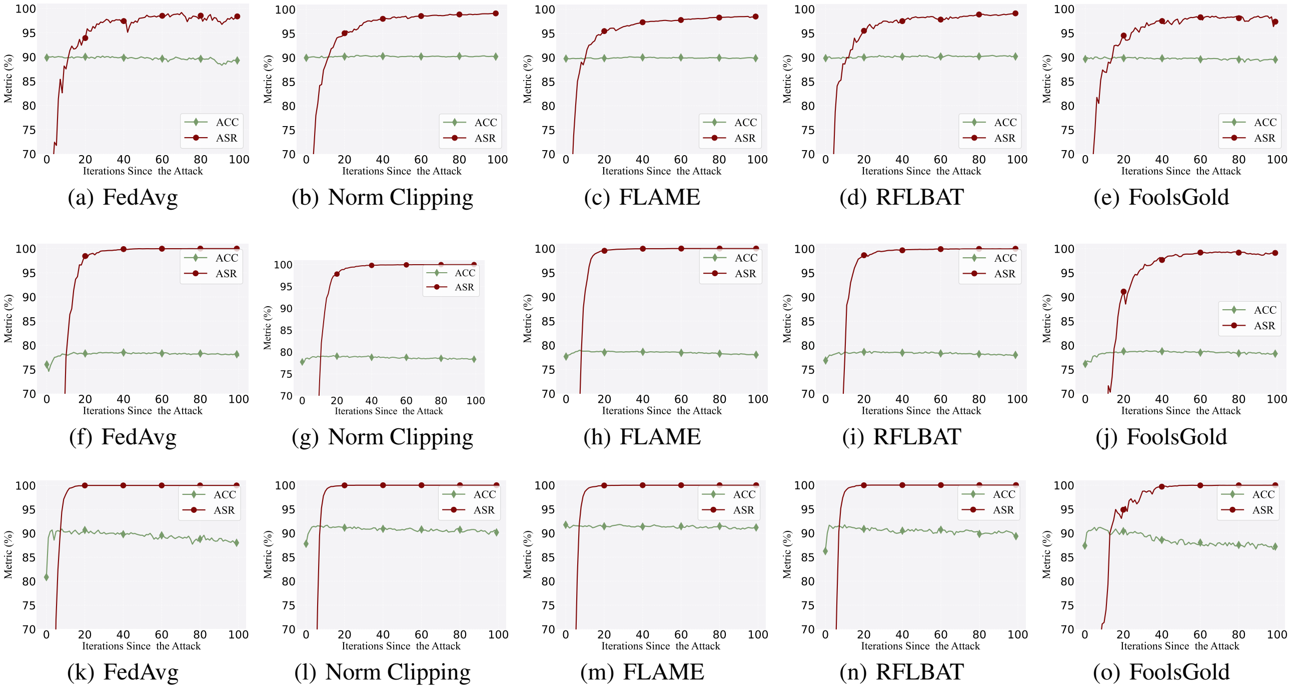
Evaluated defenses. We evaluate DarkFed against five SOTA defenses: FedAvg [McMahan *et al.*, 2017], Norm Clipping [Bagdasaryan *et al.*, 2020], FLAME [Nguyen *et al.*, 2022], RFLBAT [Wang *et al.*, 2022], and FoolsGold [Fung *et al.*, 2020]. The defenses cover all the types outlined in Sec. 2.2, showcasing the universality of DarkFed.

6.2 Experimental Results

Attack performance. We systematically verify the attainment of the adversary's goals to assess the attack performance. Figs. 3 illustrates the impact of DarkFed on SOTA defenses on CIFAR-10 (first row), CIFAR-100 (second row), and GTSRB (third row). The green line represents ACC, while the red line denotes ASR. In terms of fidelity, DarkFed achieves high ACC on both CIFAR-10 and CIFAR-100, with an accuracy degradation within 1% compared to the initial model. On GTSRB, across various defenses, accuracy degradation ranges between 1% and 3%. These marginal accuracy degradations do not significantly affect model usability, showcasing DarkFed's fidelity achievement. In terms of effectiveness, DarkFed rapidly achieves nearly 100% ASR with few iterations across all three datasets (around 20 iterations for CIFAR-10 and CIFAR-100, and 10 iterations for GTSRB), highlighting its remarkable attack effectiveness. The stealthiness goal is indirectly demonstrated through the preceding two goals, as a sufficiently concealed attack is essential to ensure that backdoor updates evade defenses, ultimately resulting in a global model that excels in both the main and backdoor tasks.

Impact of attacker ratio. Tab.5 illustrates the impact of attacker ratio on DarkFed under FLAME. It is noticeable that as the ratio of attackers increases, ACC experiences a slight decrease, while ASR exhibits an upward trend. However, overall, DarkFed shows minimal susceptibility to changes in the attacker ratio. Even in the presence of a 5% attacker ratio, the ASR remains notably high, reaching a minimum of 94.30%. The ASR nearly peaks when the attacker ratio reaches 10%.

Comparison with SOTA attacks. Existing research has not delved into data-free backdoor attacks in FL, thus we showcase DarkFed's superiority by directly comparing it with SOTA data-dependent attacks. Specifically, we consider the classic Model Replacement Attack [Bagdasaryan *et al.*, 2020] and the latest 3DFed [Li *et al.*, 2023]. It's important to note that these two attacks directly utilize task-specific data, while DarkFed relies solely on shadow dataset. Tab. 6 presents the comparative results on CIFAR-10. In terms of ACC, these three attacks exhibit no significant differences; all maintain high model accuracy. On average, the differences among them are within 0.11%. Regarding ASR, Model Replacement Attack performs relatively poorly, only managing to backdoor

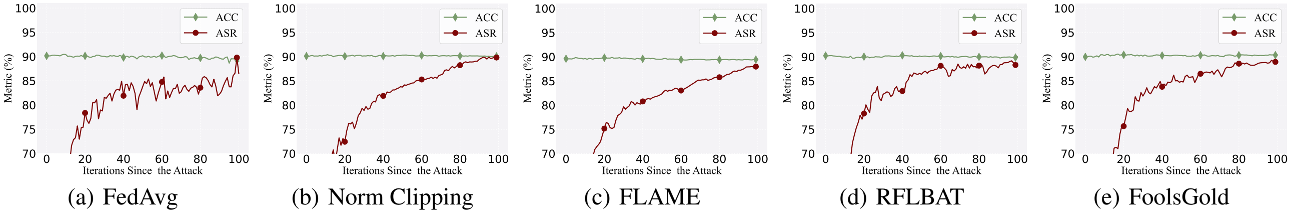


(a) FedAvg (b) Norm Clipping (c) FLAME (d) RFLBAT (e) FoolsGold

(f) FedAvg (g) Norm Clipping (h) FLAME (i) RFLBAT (j) FoolsGold

(k) FedAvg (l) Norm Clipping (m) FLAME (n) RFLBAT (o) FoolsGold

Figure 3: Attack performance on CIFAR-10 (first row), CIFAR-100 (second row), and GTSRB (third row).



(a) FedAvg (b) Norm Clipping (c) FLAME (d) RFLBAT (e) FoolsGold

Figure 4: Attack performance on CIFAR-10 with synthetic dataset.

Attacker Ratio	**CIFAR-10**		**CIFAR-100**		**GTSRB**	
	ACC (%)	**ASR (%)**	**ACC (%)**	**ASR (%)**	**ACC (%)**	**ASR (%)**
5%	90.61	95.85	79.13	94.30	92.82	99.94
10%	90.22	97.81	78.94	99.77	92.53	100.00
15%	90.13	98.51	78.82	99.92	91.84	100.00
20%	90.04	98.96	78.62	100.00	91.74	100.00
25%	90.09	99.01	78.15	100.00	91.51	100.00

Table 5: Impact of the attacker ratio.

Defense	**Model Replacement**		**3DFed**		**DarkFed**	
	ACC (%)	**ASR (%)**	**ACC (%)**	**ASR (%)**	**ACC (%)**	**ASR (%)**
FedAvg	**89.74**	99.16	89.66	**99.85**	89.54	99.18
Norm Clipping	90.07	97.93	90.14	98.71	**90.26**	**99.20**
FLAME	**90.26**	9.74	90.01	**99.89**	89.96	98.51
RFLBAT	90.17	8.84	90.21	96.18	**90.24**	**99.18**
FoolsGold	89.82	9.77	**89.97**	98.51	89.49	**98.52**
Average	**90.01**	45.09	90.00	98.63	89.90	**98.92**

Table 6: Comparison with SOTA data-dependent attacks.

FedAvg and Norm Clipping. Both 3DFed and DarkFed can overcome all defenses, but DarkFed's average ASR is slightly higher than 3DFed. Furthermore, we observe that 3DFed's ASR under the RFLBAT defense is 3% lower than DarkFed. We speculate that this is because 3DFed needs to use decoy model updates as bait to mislead RFLBAT. As a result, not all backdoor updates can be accepted by RFLBAT, sacrificing attack performance to some extent.

Attack with synthetic dataset. The preceding experiments utilize shadow datasets comprising real data from vastly different domains, yielding highly effective attack outcomes. Consequently, a naturally intriguing question arises: Can we achieve similar attack results with synthetic dataset? To answer this question, we employ Gauss-I as the shadow dataset for CIFAR-10, and the experimental results are depicted in Fig. 4. Compared with the earlier experiments (the first row of Fig. 3), ACC remains consistent, hovering around 90%. Although ASR suffers a relative decrease, it still approaches 90%. The experimental results are promising, indicating that even without crawling any datasets online, the use of self-constructed, semantically meaningless data can successfully inject a backdoor without compromising model accuracy.

Acknowledgments

Minghui's work is supported in part by the National Natural Science Foundation of China (Grant No. 62202186). Shengshan's work is supported in part by the National Natural Science Foundation of China (Grant Nos. 62372196, U20A20177).

Contribution Statement

Minghui Li, Wei Wan, and Yuxuan Ning have contributed equally to this work, therefore they can be considered as co-first authors. Wei Wan is the corresponding author.

References

[Bagdasaryan *et al.*, 2020] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS'20)*, volume 108, pages 2938–2948, 2020.

[Cao and Gong, 2022] Xiaoyu Cao and Neil Zhenqiang Gong. MPAF: model poisoning attacks to federated learning based on fake clients. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops'22)*, pages 3395–3403, 2022.

[Cao *et al.*, 2021] Xiaoyu Cao, Minghong Fang, Jia Liu, and Neil Zhenqiang Gong. Fltrust: Byzantine-robust federated learning via trust bootstrapping. In *Proceedings of the 28th Annual Network and Distributed System Security Symposium (NDSS'21)*, 2021.

[Fang and Chen, 2023] Pei Fang and Jinghui Chen. On the vulnerability of backdoor defenses for federated learning. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI'23)*, pages 11800–11808, 2023.

[Fung *et al.*, 2020] Clement Fung, Chris J. M. Yoon, and Ivan Beschastnikh. The limitations of federated learning in sybil settings. In *Proceedings of the 23rd International Symposium on Research in Attacks, Intrusions and Defenses (RAID'20)*, pages 301–316, 2020.

[Hu *et al.*, 2022] Shengshan Hu, Ziqi Zhou, Yechao Zhang, Leo Yu Zhang, Yifeng Zheng, Yuanyuan He, and Hai Jin. Badhash: Invisible backdoor attacks against deep hashing with clean label. In *Proceedings of the 30th ACM International Conference on Multimedia (MM'22)*, pages 678–686, 2022.

[Hu *et al.*, 2023] Shengshan Hu, Wei Liu, Minghui Li, Yechao Zhang, Xiaogeng Liu, Xianlong Wang, Leo Yu Zhang, and Junhui Hou. Pointcrt: Detecting backdoor in 3d point cloud via corruption robustness. In *Proceedings of the 31st ACM International Conference on Multimedia (MM'23)*, pages 666–675, 2023.

[Krizhevsky and Hinton, 2009] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

[LeCun *et al.*, 1998] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998.

[Li *et al.*, 2020] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems (MLSys'20)*, 2020.

[Li *et al.*, 2022] Minghui Li, Wei Wan, Jianrong Lu, Shengshan Hu, Junyu Shi, Leo Yu Zhang, Man Zhou, and Yifeng Zheng. Shielding federated learning: Mitigating byzantine attacks with less constraints. In *Proceedings of 18th International Conference on Mobility, Sensing and Networking (MSN'22)*, pages 178–185, 2022.

[Li *et al.*, 2023] Haoyang Li, Qingqing Ye, Haibo Hu, Jin Li, Leixia Wang, Chengfang Fang, and Jie Shi. 3dfed: Adaptive and extensible framework for covert backdoor attack in federated learning. In *Proceedings of the 44th IEEE Symposium on Security and Privacy (SP'23)*, pages 1893–1907, 2023.

[Liu *et al.*, 2023] Xiaogeng Liu, Minghui Li, Haoyu Wang, Shengshan Hu, Dengpan Ye, Hai Jin, Libing Wu, and Chaowei Xiao. Detecting backdoors during the inference stage based on corruption robustness consistency. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'23)*, pages 16363–16372, 2023.

[Lu *et al.*, 2023] Jianrong Lu, Lulu Xue, Wei Wan, Minghui Li, Leo Yu Zhang, and Shengqing Hu. Preserving privacy of input features across all stages of collaborative learning. In *Proceedings of the 21st IEEE International Symposium on Parallel and Distributed Processing with Applications (ISPA'23)*, pages 191–198, 2023.

[Lu *et al.*, 2024] Jianrong Lu, Shengshan Hu, Wei Wan, Minghui Li, Leo Yu Zhang, Lulu Xue, Haohan Wang, and Hai Jin. Depriving the survival space of adversaries against poisoned gradients in federated learning. *IEEE Transactions on Information Forensics and Security (TIFS'24)*, 2024.

[Lv *et al.*, 2023] Peizhuo Lv, Chang Yue, Ruigang Liang, Yunfei Yang, Shengzhi Zhang, Hualong Ma, and Kai Chen. A data-free backdoor injection approach in neural networks. In *Proceedings of the 32nd USENIX Security Symposium (USENIX Security'23)*, pages 2671–2688, 2023.

[Lyu *et al.*, 2023] Xiaoting Lyu, Yufei Han, Wei Wang, Jingkai Liu, Bin Wang, Jiqiang Liu, and Xiangliang Zhang. Poisoning with cerberus: Stealthy and colluded backdoor attack against federated learning. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI'23)*, pages 9020–9028, 2023.

[McMahan *et al.*, 2017] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th Inter-*

national Conference on Artificial Intelligence and Statistics (AISTATS'17), volume 54, pages 1273–1282, 2017.

[Mo *et al.*, 2023] Xiaoxing Mo, Yechao Zhang, Leo Yu Zhang, Wei Luo, Nan Sun, Shengshan Hu, Shang Gao, and Yang Xiang. Robust backdoor detection for deep learning via topological evolution dynamics. *CoRR*, 2023.

[Nguyen *et al.*, 2022] Thien Duc Nguyen, Phillip Rieger, Huili Chen, Hossein Yalame, Helen Möllering, Hossein Fereidooni, Samuel Marchal, Markus Miettinen, Azalia Mirhoseini, Shaza Zeitouni, Farinaz Koushanfar, Ahmad-Reza Sadeghi, and Thomas Schneider. FLAME: taming backdoors in federated learning. In *Proceedings of the 31st USENIX Security Symposium (USENIX Security'22)*, pages 1415–1432, 2022.

[Rieger *et al.*, 2022] Phillip Rieger, Thien Duc Nguyen, Markus Miettinen, and Ahmad-Reza Sadeghi. Deepsight: Mitigating backdoor attacks in federated learning through deep model inspection. In *Proceedings of the 29th Annual Network and Distributed System Security Symposium (NDSS'22)*, 2022.

[Shejwalkar *et al.*, 2022] Virat Shejwalkar, Amir Houmansadr, Peter Kairouz, and Daniel Ramage. Back to the drawing board: A critical evaluation of poisoning attacks on production federated learning. In *Proceedings of the 43rd IEEE Symposium on Security and Privacy (SP'22)*, pages 1354–1371, 2022.

[Shi *et al.*, 2022] Junyu Shi, Wei Wan, Shengshan Hu, Jianrong Lu, and Leo Yu Zhang. Challenges and approaches for mitigating byzantine attacks in federated learning. In *Proceedings of International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom'22)*, pages 139–146. IEEE, 2022.

[Stallkamp *et al.*, 2011] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: A multi-class classification competition. In *Proceedings of the 2011 International Joint Conference on Neural Networks (IJCNN'11)*, pages 1453–1460, 2011.

[Wan *et al.*, 2021] Wei Wan, Jianrong Lu, Shengshan Hu, Leo Yu Zhang, and Xiaobing Pei. Shielding federated learning: A new attack approach and its defense. In *Proceedings of IEEE Wireless Communications and Networking Conference (WCNC'21)*, pages 1–7, 2021.

[Wan *et al.*, 2022] Wei Wan, Shengshan Hu, Jianrong Lu, Leo Yu Zhang, Hai Jin, and Yuanyuan He. Shielding federated learning: Robust aggregation with adaptive client selection. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI'22)*, pages 753–760, 2022.

[Wan *et al.*, 2023] Wei Wan, Shengshan Hu, Minghui Li, Jianrong Lu, Longling Zhang, Leo Yu Zhang, and Hai Jin. A four-pronged defense against byzantine attacks in federated learning. In *Proceedings of the 31st ACM International Conference on Multimedia (MM'23)*, pages 7394–7402, 2023.

[Wan *et al.*, 2024] Wei Wan, Yuxuan Ning, Shengshan Hu, Lulu Xue, Minghui Li, Leo Yu Zhang, and Hai Jin. Misa: Unveiling the vulnerabilities in split federated learning. In *proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'24)*, pages 6435–6439, 2024.

[Wang *et al.*, 2020] Hongyi Wang, Kartik Sreenivasan, Shashank Rajput, Harit Vishwakarma, Saurabh Agarwal, Jy-yong Sohn, Kangwook Lee, and Dimitris S. Papailiopoulos. Attack of the tails: Yes, you really can backdoor federated learning. In *Annual Conference on Neural Information Processing Systems (NeurIPS'20)*, 2020.

[Wang *et al.*, 2022] Yongkang Wang, Dihua Zhai, Yufeng Zhan, and Yuanqing Xia. Rflbat: A robust federated learning algorithm against backdoor attack. *arXiv preprint arXiv:2201.03772*, 2022.

[Xiao *et al.*, 2017] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[Xie *et al.*, 2020] Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. DBA: distributed backdoor attacks against federated learning. In *Proceedings of the 8th International Conference on Learning Representations (ICLR'20)*, 2020.

[Zhang *et al.*, 2022] Zaixi Zhang, Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. Fldetector: Defending federated learning against model poisoning attacks via detecting malicious clients. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'22)*, pages 2545–2555, 2022.

[Zhang *et al.*, 2023a] Hangfan Zhang, Jinyuan Jia, Jinghui Chen, Lu Lin, and Dinghao Wu. A3FL: adversarially adaptive backdoor attacks to federated learning. In *Annual Conference on Neural Information Processing Systems (NeurIPS'23)*, 2023.

[Zhang *et al.*, 2023b] Hangtao Zhang, Zeming Yao, Leo Yu Zhang, Shengshan Hu, Chao Chen, Alan Wee-Chung Liew, and Zhetao Li. Denial-of-service or fine-grained control: Towards flexible model poisoning attacks on federated learning. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI'23)*, pages 4567–4575, 2023.

[Zhang *et al.*, 2024a] Hangtao Zhang, Shengshan Hu, Yichen Wang, Leo Yu Zhang, Ziqi Zhou, Xianlong Wang, Yanjun Zhang, and Chao Chen. Detector collapse: Backdooring object detection to catastrophic overload or blindness. *arXiv preprint arXiv:2404.11357*, 2024.

[Zhang *et al.*, 2024b] Longling Zhang, Lyqi Liu, Dan Meng, Jun Wang, and Shengshan Hu. Stealthy backdoor attack towards federated automatic speaker verification. In *proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'24)*, pages 1311–1315, 2024.