This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Infighting in the Dark: Multi-Label Backdoor Attack in Federated Learning

Ye Li[1], Yanchao Zhao[1*], Chengcheng Zhu[2], Jiale Zhang[2*]
[1] Nanjing University of Aeronautics and Astronautics [2] Yangzhou University
{miles.li,yczhao}@nuaa.edu.cn
chengchengzhu2022@126.com, jialezhang@yzu.edu.cn

Abstract

Federated Learning (FL), a privacy-preserving decentralized machine learning framework, has been shown to be vulnerable to backdoor attacks. Current research primarily focuses on the Single-Label Backdoor Attack (SBA), wherein adversaries share a consistent target. However, a critical fact is overlooked: adversaries may be non-cooperative, have distinct targets, and operate independently, which exhibits a more practical scenario called Multi-Label Backdoor Attack (MBA). Unfortunately, prior works are ineffective in the MBA scenario since non-cooperative attackers exclude each other. In this work, we conduct an in-depth investigation to uncover the inherent constraints of the exclusion: similar backdoor mappings are constructed for different targets, resulting in conflicts among backdoor functions. To address this limitation, we propose Mirage, the first non-cooperative MBA strategy in FL that allows attackers to inject effective and persistent backdoors into the global model without collusion by constructing in-distribution (ID) backdoor mapping. Specifically, we introduce an adversarial adaptation method to bridge the backdoor features and the target distribution in an ID manner. Additionally, we further leverage a constrained optimization method to ensure the ID mapping survives in the global training dynamics. Extensive evaluations demonstrate that Mirage outperforms various state-of-the-art attacks and bypasses existing defenses, achieving an average ASR greater than 97% and maintaining over 90% after 900 rounds. This work aims to alert researchers to this potential threat and inspire the design of effective defense mechanisms. Code has been made open-source.

1. Introduction

Recent years have witnessed the flourishing of federated learning, an advanced privacy-preserving distributed learning paradigm [24, 26]. FL allows participants to collaboratively build a neural network model without leaking their data. Following its paradigm, a wide range of FL applications have emerged in various safety and privacy-critical application domains, including medical image analysis [10, 15], face recognition [30], and others [1, 11].

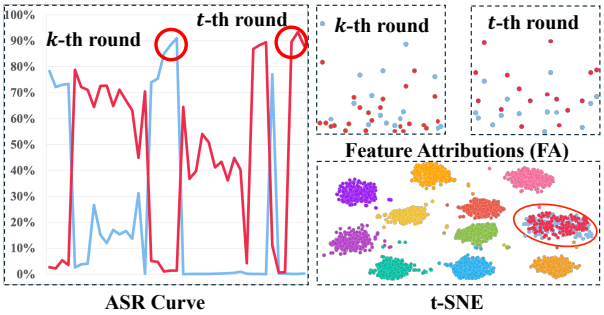


Figure 1. "Blue" and "Red" construct similar OOD mappings (see t-SNE), resulting in a conflict for the neuron weights. Only the dominant attacker can induce the model to output high feature attributions for its backdoor samples (see FA of k-th round) and successfully perform the attack. Similar in t-th round.

Despite enhanced privacy, the distributed nature of FL makes them more vulnerable to various adversarial attacks, with ***backdoor attacks*** posing the greatest threat. For backdoor attacks in FL, an attacker injects a backdoor into the local model, which is then submitted to the FL system, causing the global model to inherit the backdoor after aggregation. Consequently, the global model functions correctly on clean data but misclassifies inputs stamped with the attacker's chosen backdoor trigger as the target class. Numerous studies have been done to improve the persistence and effectiveness of backdoors [2, 6, 43, 48], which pose significant threats to various applications, including computer vision [40], IoT networks [28], and healthcare systems [16].

Recent advancements in backdoor attacks in FL, most, if not all, assume that attackers in training are colluding and sharing the same target class, which can be denoted as Single-Label Backdoor Attacks (SBA). However, situations become more complex in real-world FL applications, especially in large-scale FL scenarios [18, 22], where attackers engage in training individually and compromise

*Corresponding authors.

the global model for their own purposes, i.e., Multi-Label Backdoor Attack (MBA). Intuitively, non-cooperative adversaries separately attack the global model and do not interfere with each other. Unfortunately, our empirical results indicate that directly applying existing attack methods to multiple adversaries leads to disorganized obstacles, making backdoors not as effective as they are in SBA scenarios. As illustrated in Fig. 1, two backdoors cannot be effective simultaneously because attackers construct similar Out-of-Distribution (OOD) backdoor mappings, indicating that their backdoor samples distribute together but outside the distributions of the target classes. This results in attackers excluding one another, with only the dominant attacker capable of embedding a backdoor in the global model, consequently causing the global model to output high feature attributions for its backdoor samples, thereby achieving a high attack success rate (ASR). It further raises a new problem: ***Can MBA successfully launch without collusion?***

To address the aforementioned issue, we first conduct an in-depth investigation to uncover the inherent constraints of the exclusions of SBAs: they create similar backdoor mappings for distinct targets, resulting in conflicts among backdoor functions. Motivated by this observation, we propose Mirage, the first non-cooperative MBA strategy in FL, wherein multiple adversaries individually undermine the global model without collusion. At a high level, instead of coordinating the various OOD mappings, Mirage aims to ensure that the backdoor samples are processed as clean samples by constructing an effective and persistent ID mapping between backdoor features and target distributions. To achieve this goal, we introduce a trigger optimization method based on adversarial adaptation, which adaptively optimizes the trigger to fool a carefully designed OOD sample detector, thereby constructing the ID mapping. Furthermore, considering that the distributions change with global training, we further introduce a constrained optimization method to enhance the persistence of ID mapping by tightening the distribution of backdoor features, allowing the backdoor mapping to live within the global training dynamics. In summary, our contributions are as follows:

- **Uncovering the inherent constraints of extending existing SBAs to MBAs.** To the best of our knowledge, we are the first to investigate and elucidate the inherent constraints of extending existing SBAs to MBAs. We underscore the practical threat posed by such attacks in large-scale FL applications by proposing a multi-label backdoor attack method based on the in-distribution backdoor.
- **Effective ID Mapping Construction.** We propose a novel backdoor optimization method based on adversarial adaptation, which enables attackers to build the bridge between their backdoor and the distribution of their target class, allowing multiple adversaries to simultaneously launch backdoor attacks without interfering with others.
- **Persistent ID Mapping Enhancement.** We additionally introduce a constrained optimization method to enhance the ID mapping by minimizing the similarity of features between backdoored samples and benign samples. This tightens the ID backdoor mapping and ensures its persistence within the global training dynamics.
- **Evaluations and Discussions.** Extensive experiments demonstrate that Mirage outperforms SOTA attacks across various settings. Furthermore, we discuss the vulnerability of existing defense methods to MBAs and present potential countermeasures against Mirage, highlighting the urgent demand for tailored defenses.

2. Related Work

2.1. Backdoor Attacks in Federated Learning

Backdoor attacks aim to compromise the victim model to produce the attacker-desired outputs if and only if the inputs are assigned to a specific trigger pattern. In light of this, a few researchers also start to notice the threat of backdoor attacks on FL systems. Depending on whether the backdoor trigger is optimized, current backdoor attacks fall under the categories of either static-trigger attacks [2, 6, 20, 44, 48] or trigger-optimization attacks [7, 25, 43, 47].

The former pre-selected a fixed trigger pattern to poison the local training set and combine them with manually manipulating methods to improve the backdoor effectiveness. For example, [2, 3] aim to replace the global model by uploading a scaled-up poisoned model. In advance, [7, 34, 48] suggest leveraging the unimportant redundant neurons to prevent the backdoor from being erased shortly. Particularly, Dai et al. [6] introduce contrastive learning from the feature perspective of the backdoor to enhance its durability. The latter optimizes trigger patterns according to different tasks and generally achieves better effectiveness and utility. 3DFed [19] and A3FL [43] introduced the backdoor unlearning in an adversarial way to obtain a robust trigger to make the attacks more covert and persistent. DBA [41] and NBA [27] are the most similar work with us. However, DBA focuses on attacking the same target class, which is also an SBA situation. NBA investigated the non-cooperative multi-label attack in FL but neither revealed the inner reason for the inapplicability of previous works nor proposed an effective method.

2.2. Backdoor Defenses in FL

Recently, researchers have noticed the threat of backdoor attacks in FL systems and actively explored various backdoor defense techniques, which can be categorized into either detection [8, 21, 39] or mitigation methods [4, 29, 45, 46, 49].

The detection-based methods aim to utilize anomaly detection techniques to determine whether a model is backdoored. Foolsgold [8] assigns lower aggregation weights

to updates with high pairwise cosine similarities, thereby mitigating the impact of backdoor updates. RFLBAT [39] discriminates the malicious model according to the difference between poisoned updates and clean updates in low-dimensional projection space. In advance, Li et al. [21] proposed BackdoorIndicator (referred to as "Indicator" hereafter), which detects potentially poisoned models based on the OOD properties of backdoor samples. Mitigation-based methods explored how to purify the backdoor model by eliminating backdoor triggers. MultiKrum [4] is a byzantine-robust aggregation protocol that aggregates the global model of having the smallest $\mathcal{L}_2$ distance. FLAME [29] adapts the weak DP method by noise boundary proof and a dynamic clipping bound, which is shown to alleviate the backdoor attack while still retaining a high main task accuracy.

3. Attack intuitions

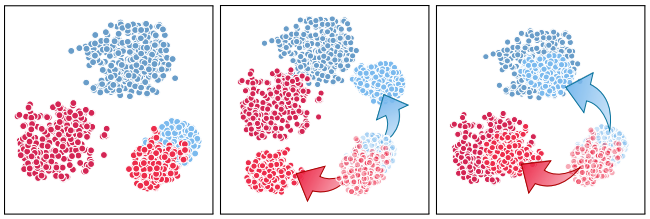
In this section, we investigate the inherent constraints that affect the effectiveness of backdoors when extending previous works to MBA scenarios and further provide the attack intuitions based on this investigation.

3.1. Inherent constraints of SBA methods

Recall that, from the neuron activation perspective, the backdoor function in a model is highly related to the specific activation paths across neurons. This results in the model being sensitive to triggered inputs and producing anomalously high feature attribution values [9, 34, 36]. Correspondingly, such high values result in backdoor samples distributed out of the target clean distribution in feature space, i.e. the backdoor function constructs the OOD backdoor mapping [21]. In MBA scenarios, attackers may adopt similar strategies to construct the activation paths, such as using redundant neurons, which leads to the construction of similar OOD mappings and competition for neuron weights. Specifically, different attackers have varying requirements for the output values of specific neurons within the pathway. However, the dominant attacker controls most neurons, ensuring they produce the desired values, which results in others failing to achieve their backdoor function. As shown in Fig. 2a, two attackers aim to construct persistent and effective OOD backdoor mappings by targeting redundant neurons. Such a strategy causes them to leverage overlapped activation paths to achieve different functions, which further results in infighting among adversaries. This drives us to find a solution that allows attackers to select different activation paths non-cooperatively.

3.2. Attack intuitions and challenges

To alleviate the exclusion, it is critical to design a mapping strategy that ensures different attackers can independent



(a) SBA's mapping. (b) Separate mapping. (c) ID mapping.

Figure 2. Illustrations of different mapping strategy.

mappings with no intersections. Achieving this is challenging because attackers share no information with others and are unaware of each other's existence. Surprisingly, our empirical investigation reveals that A3FL can fulfill the aforementioned requirement by combining machine unlearning [37, 42], which indirectly stimulates the eliminations performed by other attackers. As shown in Fig 2b, the unlearning method adversarially optimizes the trigger to construct a distinct OOD backdoor mapping that differs from other attackers. Although this method mitigates the exclusions, it enhances the OOD characteristics of the backdoor functions, consequently introducing more challenges in bypassing the OOD-based defense methods, such as Indicator [21]. The OOD-based defense method leads us to a counter-intuitive thought: ***Can a backdoor be triggered via a clean activation path?*** In other words, construct the ID mapping for the backdoor features and target distributions (present in Fig 2c). Therefore, once the adversaries construct the ID backdoor mapping to the target class, they naturally alleviate potential exclusions, even without collusion, since additional measures are no longer needed to circumvent the backdoor activation paths of other attackers. However, constructing the ID backdoor mapping remains challenging due to the following issues:

Challenge 1: How to construct the ID mapping?

While backdoor attacks have emerged as a prominent topic in FL, to the best of our knowledge, no prior studies investigated the ID backdoor within FL. Intuitively, the backdoor sample should be processed as a clean sample following the clean activation path, meaning it should share similar feature outputs. One promising way is leveraging feature alignment. However, this may disrupt the relationships among clean classes, making it challenging for attackers to maintain the ID characteristics. Therefore, there is a pressing need to devise a novel method that ensures the effective construction of ID mapping.

Challenge 2: How to maintain the persistence of ID mapping?

One key challenge for backdoor attacks in FL is that the continued dynamic clean updates may break the well-constructed backdoor mappings. Previous research focuses on leveraging neurons that are updated less frequently to build the neural activation path, which cannot be applied in ID mapping since our objective is to construct the bridge between the trigger pattern and the clean target activation path.

Therefore, these maintenance requirements for ID mapping further exacerbate the challenges.

4. Methodology

To formulate the attack scenario, we first introduce the threat model. Driven by attack intuitions, we propose two methods that enable attackers to construct an efficient and persistent ID mapping without cooperation.

4.1. Threat model

Adversaries' Capabilities and knowledge. We consider a FL system running image classification tasks and assume that adversaries are FL participants, which enables attackers to control their own local datasets and training processes. Note that adversaries do not have access to any information regarding benign clients or other adversaries, including but not limited to datasets, models, and target classes.

Adversaries' goal. Adversaries aim to plant a backdoor into the global model that makes it misclassify backdoor data to a specific class while leaving other tasks unaffected.

4.2. Overview of Mirage

Following the attack intuition, our objective is to construct an ID mapping to bridge the backdoor features and distribution of the target class, thereby freeing attackers from the exclusion of using similar backdoor activation paths. This presents two significant challenges: how to construct a qualified ID mapping and how to maintain it. To address these challenges, we propose Mirage, the first multi-label backdoor attack in FL, which enables attackers to non-cooperatively compromise the global model by constructing an ID mapping between the backdoor features and target distributions. As illustrated in Fig. 3, Mirage introduces two components to construct an effective and persistent ID backdoor mapping. On the one hand, we propose a backdoor optimization method based on adversarial adaptation, which constructs the ID mapping by maximizing the misclassification probability of a well-designed backdoor sample detector. On the other hand, we deploy a constrained optimization strategy to tighten the backdoor distribution, ensuring its survival in the global training dynamics. Next, we provide details of each component.

4.3. Effective ID Mapping Construction

To address the first challenge, we aim to adversarially optimize the trigger pattern to make it able to construct the ID mapping. In order to optimize the trigger, we employ an OOD sample detector to determine whether a backdoored sample is an OOD sample and further optimize the trigger pattern to fool the detector. This introduces another problem: How can an attacker obtain an accurate detector? Previous research suggests training a new model as the detector [47]. However, retraining the detector with each trigger update incurs significant computational overhead, and the local samples limit its effectiveness. These shortcomings prompt us to seek another solution to obtain a detector that is both effective and lightweight for updates. Recall that the FL training process allows clients to obtain the latest global model, which contains a well-trained feature extractor. Consequently, we consider leveraging the feature extractor of the global model as the primary component and training a binary classifier to distinguish the OOD samples. Such an approach is a common practice in transfer learning, consumes only a few computing resources, and requires little data.

In order to obtain an accurate detector, we first prepare the dataset for training. The local dataset for each attacker can be divided into $D_{cln} = \{(x, y) \in D_{att})|+\}$ and $D_{bd} = \{(x \oplus \delta, y) \in D_{att})|-\}$, where $+$ indicates the ground truth label of a sample is c, and $-$ is the opposite and δ represents the trigger pattern. Then, the detector dataset $D_{detector}$ (D_d for short) is composed as follows:

$$D_d = \{(x, -) \mid x \in D_{bd}\} \cup \{(x, +) \mid x \in D_{cln}\}, \quad (1)$$

To make the ID mapping construction more effective and flexible, we adopt an adversarial training-like method to optimize the trigger pattern. Specifically, $\theta_{detector}$ is trained on the $D_{detector}$ to differentiate between clean and backdoored samples by minimizing the binary cross-entropy (BCE) loss.

Concurrently, the trigger is optimized to fool the detector by maximizing the binary classification loss. Such a min-max game equips the trigger with the ability to activate the clean activation path of the target class in the global model, naturally avoiding conflicts with other attackers. The min-max process is mathematically described as follows:

$$\begin{aligned} \delta = & \ \underset{\delta}{\operatorname{argmin}}\, \mathbb{E}_{(x,y)\sim D_{detector}}[\mathcal{L}(x, +, \theta)|y = -] \\ s.t.\ \theta = & \ \underset{\theta}{\operatorname{argmin}}\, \mathbb{E}_{(x,y)\sim D_{detector}}[\mathcal{L}(x, y, \theta)] \end{aligned} \quad (2)$$

where θ is the detector's parameters, with its feature extractor frozen and parameterized by the weights of the global model, and $\mathcal{L}$ is the BCE loss. For better implementation, we replace the maximization problem with minimizing the BCE loss of misclassification for the backdoored samples.

4.4. Persistent ID Mapping Enhancement

Although using an OOD detector and an adversarial strategy restricts the trigger pattern δ to reveal the ID feature, our empirical results indicate that the backdoored samples still do not meet our expectations. This is because the adversarial strategy does not constrain the tightness of backdoor mapping, which causes the margin-distributed backdoored samples not in-distribution anymore with the dynamic changes of training and failing to activate the target

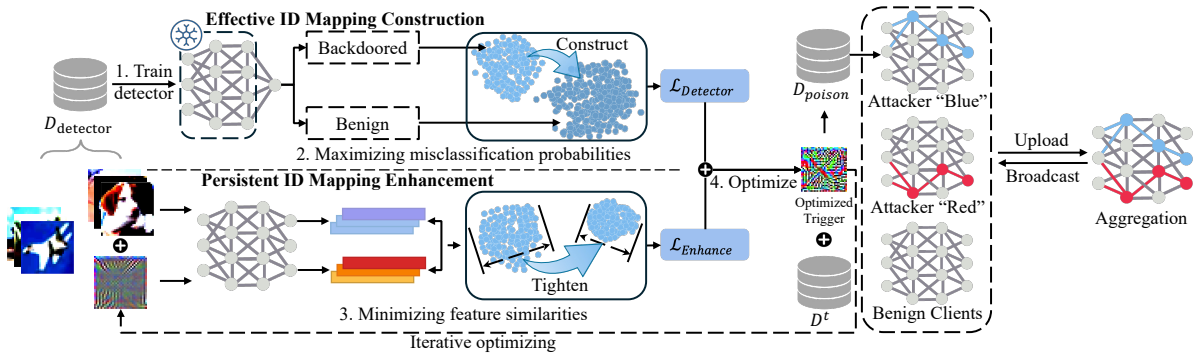


Figure 3. The workflow of Mirage. **Step 1**: Train an OOD sample detector $\theta_{Detector}$. **Step 2**: Construct ID mapping by maximizing misclassification probabilities of backdoor samples on the detector. **Step 3**: Tighten backdoor distribution by minimizing the feature similarities between backdoor samples and benign samples. **Step 4**: Optimize the trigger by minimizing $L_{Enhance}$ and $L_{detector}$.

activation path. One promising solution is to leverage distribution estimation methods, like KDE [33], to obtain the actual distribution of the target class and use data augmentation to generate samples to improve the effectiveness of the detector. Although it has been proven effective, it requires countless computing resources. Therefore, we introduce a constrained optimization method to enhance the ID mapping by tightening the distribution of backdoored samples. Since the adversarial strategy successfully constructs the ID mapping, we aim to "push" the backdoor distribution to make it tighter. Specifically, we minimize the similarities between clean samples x and their corresponding backdoored samples $x \oplus \delta$, allowing the backdoored sample to deviate as much as possible from the original distribution. Additionally, we establish the direction by minimizing the loss of the backdoor samples on the global model. Formally, it can be formulated as the following optimization problem:

$$\delta = \text{argmin } \text{CS}[(\theta_f(x), \theta_f(\hat{x})] + \text{CE}(\hat{x}, \hat{y}, \theta) \tag{3}$$

where θ_f represents the feature extractor of the global model θ. $\hat{x} = x \oplus \delta, x \in D_{att}$, $\oplus$ is the backdoored operator. $\text{CS}(\cdot)$ and $\text{CE}(\cdot)$ represent the Cosine-similarity and the Cross-entropy loss, separately. Therefore, the backdoor distribution can be tightened into the target distribution and live within the global training dynamics.

We summarized Mirage in Algorithm 1, and the detailed description is provided in Appendix B.

Algorithm 1: Mirage at i-th attacker.

Data: $\theta_t, \theta_{detector}, \theta_{fe}, \mathcal{D}_i, \delta, \eta, c, E$.
$\theta_t^i = \theta_t$;
/ Trigger adversarial adaption*/*
for *epoch* $e = 1, \ldots, E$ **do**
  Generate $D_{detector}$ by Eq 1;
  $\theta_{detector} = \underset{\theta}{\text{argmin}}\ \mathbb{E}_{(x,y)\sim D_{detector}}[\mathcal{L}(x, y, \theta_{detector})]$;
  for *batch* $\mathcal{B} \in D_i$ **do**
    $\mathcal{L}_{detector} = \frac{1}{|\mathcal{B}|}\sum_{(x,y)\in\mathcal{B}} \mathcal{L}_{BCE}(x \oplus \delta, +, \theta_{detector})$;
    $\mathcal{L}_{Enhance} = \frac{1}{|\mathcal{B}|}\sum_{(x,y)\in\mathcal{B}} \mathcal{L}_{CE}(x \oplus \delta, \hat{y}, \theta_t^i)$
      $+\mathcal{L}_{CS}((\theta_{fe}(x), \theta_{fe}(x \oplus \delta))$;
    $\delta \leftarrow \delta - \eta \cdot \nabla_\delta(\mathcal{L}_{detector} + \mathcal{L}_{Enhance})$;
  end
end
/ Local training*/*
Get the poisoned local dataset $\mathcal{D}_i^{poison}$ with δ;
Local optimize θ_t' on $\mathcal{D}_i^{poison}$;
$\mathbf{\Delta}_i^{t+1} = \theta_t - \theta_t'$;
Upload $\mathbf{\Delta}_i^{t+1}$ to the server;

5. Experiments

In this section, we provide extensive experimental results to demonstrate the effectiveness of Mirage by comparing its performance with several SOTA backdoor attack methods from multiple angles. Our experiments run on a FL system performing image classification tasks, using an NVIDIA GeForce RTX 4090 GPU equipped with 24GB of memory. The detailed setups are listed as follows:

Dataset: We conduct experiments on three widely used public real-world datasets, CIFAR-10, CIFAR-100 [17] and GTSRB [13]. Details are provided in Appendix C.

Federated learning setup: By default, we set the number of FL clients $N = 100$. The server randomly selects $M = 10$ clients at each communication round to contribute to the global model. The default global model architecture is ResNet-18 [12], which is a classic and effective model for image classification. Based on previous research, we randomly split the dataset among clients non-IID using Dirichlet sampling [14] with a hyperparameter $\alpha = 1$. For each selected benign client, they train the local model for two local epochs using an SGD optimizer with a learning rate of 0.01. The batch size is 64 for CIFAR-10 and CIFAR-100 and 32 for GTSRB. The entire training process continues

for 2100 communication rounds.

Attack and defense setup: We compare the effectiveness of Mirage with SOTA methods, including A3FL [43], Chameleon [6], Neurotoxin [48], PGD [38], and Vanilla [2]. We further extend the aforementioned methods to MBA scenario, wherein attackers independently execute their attacks. Unless otherwise specified, we set $N = 3$, with the target classes being 0, 1, and 2. In line with A3FL, attackers of Mirage optimize their triggers by PGD and poison 12.5% of D_{att}. Blend [5] is utilized as a backdoor trigger across all attacks. The attack window starts from the 2000-th round and continues for 100 rounds, so attackers are forced to wait for selection and cannot launch a continuous attack.

We also evaluate those attacks against six SOTA defenses: Multi-Krum [4], DeepSight [31], Foolsgold [8], RFLBAT [39], FLAME [29] and BakcdoorIndicator [21]. For those attacks and defense methods, we follow the original implementations. Note that we selected the random noise dataset as the source of the indicator dataset.

5.1. Evaluation metrics

Following previous works, we comprehensively evaluate Mirage using accuracy (Acc), attack success rate (ASR), and lifespan. Specifically, Acc reflects the attack stealthiness, which indicates the performance of the backdoor model on the main tasks. ASR represents the ratio of backdoored samples misclassified as the labels specified by the attackers. Note that we report the mean ASR of three attackers on the global model at the end of FL. To evaluate the persistence of our attack, we introduce lifespan, denoted as the period that starts at the end of the attack window and ends when the ASR decreases to below a chosen threshold.

5.2. Attack performance

First, we evaluate the attack performance of Mirage across various datasets and compare it to different SOTA attack methods. The results are summarized in Table 1, and the best results are highlighted in Bold. The experiment shows that Mirage achieves remarkable results across three tasks and various SOTA defense methods compared to the baseline methods. In terms of accuracy, Mirage achieves the best accuracy in over half of the tasks and obtains suboptimal results for the remaining tasks, losing to the best ones by an average of only 0.498%. Regarding effectiveness, Mirage achieves the best ASRs across all the tasks, with the gap among attackers being acceptable (the maximum gap is 6.99%, and the average gap is 2.19%). Those results indicate that the design of constructing the ID mapping between backdoor features and target class allows attackers to independently attack the FL model without concerns about the potential exclusion, resulting in a global model that excels in both main and multiple backdoor tasks.

Furthermore, A3FL is a relatively competitive method

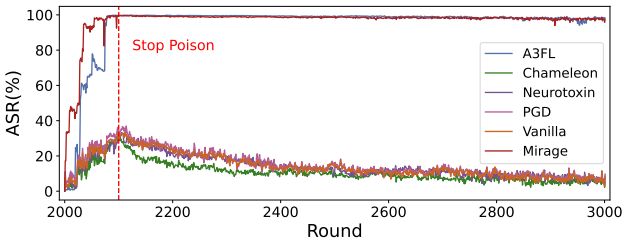


Figure 4. Persistent evaluations.

that leverages unlearning to simulate the potential exclusion and optimize the trigger to construct robust activation paths independent of other attackers. However, as described in Sec. 3.2, such an approach may enhance its OOD mapping concerning the benign distribution and increase the risk of being detected by OOD-based detection methods, such as Indicator [21]. The results of A3FL with the defense of Indicator support our analysis. Besides, the Indicator does not perform as expected, we will discuss it in Section 6.3.

5.3. Lifespan

Secondly, we evaluate the durability of attacks by comparing their lifespans. We extend the FL training process from 2100 to 3000 rounds while keeping the attack window to the first 100 rounds. Fig. 4 illustrates the attack success rate without any defense method. Mirage achieves similar results with A3FL, where the ASRs exceed 90% at the end, indicating a lifespan of over 900 rounds. In contrast, the ASRs of all other baseline attacks are constrained by potential exclusion between attackers, as described before, making them ineffective and falling rapidly. These experimental results suggest that Mirage can achieve comparable durability to A3FL, which focuses on enhancing the persistence of backdoors in FL.

5.4. The impact of different models

In this section, we evaluate Mirage across different model architectures. Following previous research, we select widely used image classification model architectures: ResNet18 and ResNet34 [12], VGG11 and VGG19 [35], and MobileNet-V2 [32]. We set the parameters to default and only change the model architectures. As shown in Tab. 2, the attack performance of Mirage remains high in ASRs. The decrease in MobileNet-V2 is mainly attributed to the low utility of its model performance, which significantly reduces the detector's utility, resulting in sub-optimal ID mapping construction for certain attackers. However, the average ASRs remain acceptable. In conclusion, based on the results presented above, we can conclude that Mirage is not sensitive to specific model structures.

5.5. The impact of adversary number

We also study the impact of the number of adversaries, which can be considered as the number of different targets

Dataset	Defense	Vanilla			PGD			Neurotoxin			Chameleon			A3FL			Mirage		
		Acc (↑)	ASR (↑)	GAP (↓)	Acc (↑)	ASR (↑)	GAP (↓)	Acc (↑)	ASR (↑)	GAP (↓)	Acc (↑)	ASR (↑)	GAP (↓)	Acc (↑)	ASR (↑)	GAP (↓)	Acc (↑)	ASR (↑)	GAP (↓)
CIFAR-10	nodefense	91.76	31.88*	87.68	91.52	31.25*	92.78	92.03	52.99*	90.29	**92.35**	39.65*	79.98	91.73	99.52*	0.23	92.16	**99.54***	0.27
	deepsight	91.44	31.71*	88.29	90.81	31.083*	92.63	92.31	39.24*	83.50	**92.32**	36.66*	79.66	92.17	96.56*	7.74	91.28	**97.39***	3.82
	foolsgold	88.79	32.50*	89.50	90.71	31.67*	93.96	91.04	80.91*	21.83	90.66	45.46	67.25	89.00	95.30*	12.02	**91.32**	**97.35***	2.51
	Indicator	85.08	28.76	85.32	85.83	20.60	59.50	85.13	29.98	60.32*	86.13	8.21	5.13	85.75	70.33*	88.45	**91.10**	**93.46***	3.90
	multikrum	92.36	1.59	4.56	92.39	1.18	3.25	**92.64**	1.89	5.18	92.46	1.90	5.25	92.18	78.15*	64.41	92.10	**92.30***	3.12
	rflbat	86.83	30.55	88.34	**88.28**	33.27*	99.79	87.22	33.27*	73.12	86.97	37.15	86.42	85.46	87.83*	10.21	85.74	**97.29***	2.11
	Flame	**92.56**	1.83	5.26	92.51	3.39	9.90	92.02	40.97*	94.04	92.28	1.38	3.71	92.25	79.40*	60.23	92.36	**95.9***	2.42
CIFAR-100	nodefense	69.40	32.62*	94.31	70.53	32.28*	93.89	71.07	32.62*	97.05	**71.66**	24.07	27.50	71.26	98.67*	2.47	71.65	**99.05***	0.77
	deepsight	70.02	32.86*	89.02	70.54	32.23*	96.32	70.53	33.14*	93.28	71.30	24.06	26.17	**71.68**	92.05*	18.44	71.56	**98.67***	2.43
	foolsgold	70.02	33.30	89.81	70.15	32.65*	97.34	71.07	32.63*	97.05	71.60	24.20	27.11	71.78	98.71*	1.64	**71.94**	**99.18***	1.00
	Indicator	67.73	6.80	20.12	63.85	5.88	17.28	60.98	11.89	32.34	61.35	6.07	13.43	65.35	37.13*	92.72	**68.22**	**99.8***	0.10
	multikrum	71.81	0.38	0.77	71.14	0.31	0.59	71.63	0.36	0.66	**71.85**	0.21	0.38	71.76	91.28*	18.55	**71.86**	**93.57***	5.14
	rflbat	62.36	0.30	0.41	63.10	0.60	0.12	62.15	0.56	1.00	61.49	18.01	27.31	61.35	94.34*	16.56	**63.19**	**95.68***	6.99
	Flame	71.75	0.21	0.28	**72.09**	0.33	0.62	72.02	0.27	0.45	72.10	0.30	0.52	71.22	91.48*	19.83	71.82	**94.22***	3.16
GTSRB	nodefense	96.10	33.32*	98.24	96.73	33.10*	98.80	95.93	38.39*	95.79	96.08	45.62*	87.06	96.65	99.63*	0.49	**96.97**	**99.73***	0.48
	deepsight	95.95	33.8*	97.44	95.97	33.36*	98.35	96.12	33.06*	99.17	96.14	50.14	76.71	95.95	94.78*	12.64	**96.20**	**96.76***	2.53
	foolsgold	96.08	33.47*	97.92	96.94	33.06*	98.75	95.88	37.57*	96.36	96.04	51.91	78.06	96.62	97.68*	6.95	**96.95**	**99.88***	0.27
	Indicator	91.57	50.92	91.53	89.45	27.27	78.14	89.08	17.09	50.97	92.91	2.76	5.38	85.51	85.05*	25.26	**95.12**	**99.73***	0.70
	multikrum	96.24	4.14	12.20	**96.97**	2.74	7.72	96.12	4.84	14.32	96.13	4.72	13.97	96.62	98.18*	4.16	96.92	**98.99***	1.73
	rflbat	94.40	56.61*	90.62	89.44	33.26*	99.72	94.53	33.88*	97.80	94.75	54.39*	83.04	**95.23**	97.86*	6.42	94.21	**99.97***	0.08
	Flame	95.88	6.19	18.45	**96.97**	2.74	7.72	95.92	5.15	15.32	95.92	5.95	17.71	96.52	99.31*	1.52	96.92	**99.32***	1.05

Table 1. Performance of Mirage compared with baseline attack methods against the SOTA defense methods. The Acc higher represents better, and so does the ASR. "GAP" represents the difference between the highest and the lowest ASR, and attack methods should maintain their GAP close to zero. An "*" after ASR indicates that at least one of the three attackers achieved over 90% ASR.

	ResNet18	ResNet34	VGG11	VGG19	MobileNet-V2
Acc	92.16	93.89	91.14	92.07	85.38
ASR	98.8	99.63	99.21	99.82	97.65

Table 2. Results on different model architectures.

Dataset		0.5	1	5	10	1000
CIFAR100	Acc	71.46	71.65	71.68	71.50	71.81
	ASR	99.59	99.05	99.16	99.26	99.21
GTSRB	Acc	96.96	96.97	96.62	96.57	96.70
	ASR	99.54	99.73	99.47	98.26	98.86
CIFAR10	Acc	91.61	92.16	92.43	92.51	92.38
	ASR	98.95	98.80	99.03	99.39	99.15

Table 3. Results on different non-IID settings.

Attacker Number	1	2	3	4	5
Acc	92.37	92.53	92.16	92.11	92.12
ASR	99.06	99.15	98.80	98.375	98.484
GAP	0.00	0.49	0.38	1.42	1.99

Table 4. Mirage with different attacker number.

since we assign different targets to the attackers. Table 4 presents the experimental results on CIFAR-10 with varying attack numbers ranging from 1 to 5, evaluating the performance of the global model on the main tasks and different backdoor tasks. Note that other settings follow the default configuration. We observe that the increase in attack numbers has a negligible impact on Mirage, with the accuracy of main tasks maintained above 92% and the average ASR exceeding 98%. The experimental results indicate that backdoor injection without collusion can be effectively implemented by constructing ID mapping. We also conducted experiments on different datasets (see Appendix D), and the results support the above conclusions.

5.6. Performance under varying non-IID degrees

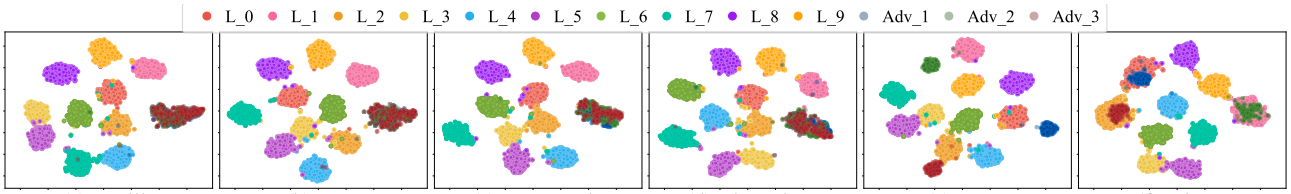
Tab. 3 exhibits the attack performance of Mirage under varying non-IID degrees. We set different sampling parameters α for Mirage, with larger values indicating a distribution that is closer to IID. The results indicate that Mirage can achieve stable performance across different non-IID degrees. As described in Sec. 4.4, the adversarial strategy may optimize the backdoor distribution to the margin of the target distribution. To address this issue, we introduce a constrained optimization method, tightening the backdoor distribution into the objective distribution and ensuring successful attacks across different distributions.

5.7. Illustration of backdoor distribution

The preceding experiments leverage t-SNE to illustrate the backdoor distribution for Mirage and baseline methods. The results are in Fig. 5. The backdoor distribution from Fig. 5(a) to Fig. 5(d) further corroborates our presentation in 1, that different attackers may share similar neuron activation paths, resulting in their potential mitigation since their distribution is close to each other. On the other aspect, A3FL paves the activation path by adversarially unlearning the backdoor, which enhances their OOD mapping, producing clear boundaries between classes and tight clustering within each class. Although effective, this poses more threats to A3FL since the OOD-sample-based detection method can detect their backdoor models effectively. Looking back to Mirage, which is pictured in Fig. 5(f), backdoor samples present in-distribution with the target clean distribution. The results further indicate that the proposed Mirage can construct the ID mapping between backdoor features and clean distributions of the target class.

5.8. Ablation study

Sections 4.3 and 4.4 propose two hybrid techniques to construct and enhance the ID mapping. Here, we conduct an ablation study to show the effects of different techniques for our method. The experiment adheres to the settings



(a) Vanilla (b) PGD (c) Neurotoxin (d) Chameleon (e) A3FL (f) Mirage

Figure 5. Illustration of backdoor distribution.

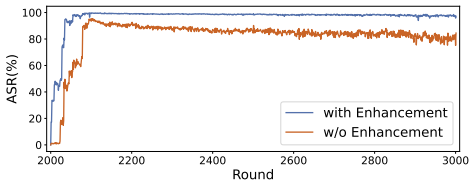


Figure 6. Ablation study.

outlined in Section 5.3. Fig. 6 presents the average ASR for Mirage and Mirage without enhancement. We can see that introducing the ID mapping enhancement significantly improves Mirage's performance. As described in Section 4.4, although the adversarial strategy successfully constructs the ID mapping, the backdoors may not fully activate the clean path, making the backdoor distribution map to the target distribution's margin, rendering it ineffective and non-persistent. Therefore, we introduce the constrained optimization method to tighten the backdoor distribution, ensuring backdoor samples will not be misidentified as the target distribution changes.

6. Discussion

6.1. Does the pixel trigger still effective?

In the previous experiments, we adopted the blend trigger as the trigger pattern. Here, we conduct experiments with pixel triggers to evaluate their effectiveness. The results (presented in Appendix E) show that the pixel trigger remains effective. According to the t-SNE analysis of the feature attribution, the backdoor samples cannot aggregate together like the blend trigger. The main reason is that the pixel trigger can only activate a subset of neurons along the activation path, while other features typically activate the paths belonging to their original classes. However, it still constructs the ID mapping to the target distribution and successfully embeds the backdoors for different attackers.

6.2. One-shot Mirage

Mirage can perform a one-shot attack by increasing adversarial training epochs, allowing the trigger to directly simulate features of the target class, like adversarial samples. However, it does not construct persistent and effective ID backdoor mapping, resulting in low durability and side effects on the main tasks. Therefore, we enable the model to learn from the backdoor samples and construct effective and persistent ID mapping by injecting more rounds and fewer batches to achieve high-confidence classification.

6.3. Potential defense methods

According to Section 5.2, we can see that although Indicator [2], the most powerful detection method for now, can detect at least one of the three A3FL attackers, the rest can still successfully inject the backdoor. This is because it constructs limited OOD mappings, succeeding only under the assumption that the attacker constructs similar OOD mappings. Therefore, it cannot be as effective against dynamic attacks (A3FL) as it is against static attacks (Chameleon, etc.). Furthermore, for A3FL, such misdetections are sufficient for the attacker to inject a persistent backdoor into the global model. Thus, following the works with foresight [23, 27], we urge researchers to focus on MBA and expand their efforts to defend against it.

Our proposed Mirage does not follow the traditional backdoor activation scheme, making it challenging for previous work to resist. This motivates us to seek solutions to address this threat to real-world FL applications. The advanced explainable method may effectively detect our attack, however, detecting all backdoored models remains challenging. Input detection may also identify our attack, as clean samples differ from backdoored samples.

7. Conclusion

In this paper, we uncover the inherent constraints of SBAs when extending them to MBA scenarios: Non-cooperative attackers will exclude others as they construct similar backdoor mappings. Based on this observation, we introduce Mirage, the first non-cooperative MBA strategy in FL. Mirage ensures the backdoor functions of different attackers by constructing effective and persistent ID backdoor mappings. Specifically, an adversarial trigger optimization method is introduced to construct the ID mapping between backdoor features and target distributions. Subsequently, Mirage leverages a constrained optimization method to tighten the ID mapping for persistence. Our comprehensive experiments demonstrate that Mirage outperforms existing backdoor attacks across different settings. For future work, investigating defense methods will be our primary focus.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant(No. 62172215) and in part by the A3 Foresight Program of NSFC (Grant No. 62061146002)

References

[1] Apple. Private federated learning (neurips 2019 expo talk abstract). https://nips.cc/ExpoConferences/2019/schedule?talkid=40. accessed: 2020-05-22. 1

[2] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *International conference on artificial intelligence and statistics*, pages 2938–2948. PMLR, 2020. 1, 2, 6, 8

[3] Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal, and Seraphin Calo. Analyzing federated learning through an adversarial lens. In *International conference on machine learning*, pages 634–643. PMLR, 2019. 2

[4] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 2, 3, 6

[5] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017. 6, 1

[6] Yanbo Dai and Songze Li. Chameleon: Adapting to peer images for planting durable backdoors in federated learning. In *International Conference on Machine Learning*, pages 6712–6725. PMLR, 2023. 1, 2, 6

[7] Pei Fang and Jinghui Chen. On the vulnerability of backdoor defenses for federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11800–11808, 2023. 2

[8] Clement Fung, Chris J. M. Yoon, and Ivan Beschastnikh. The limitations of federated learning in sybil settings. In *23rd International Symposium on Research in Attacks, Intrusions and Defenses, RAID 2020, San Sebastian, Spain, October 14-15, 2020*, pages 301–316. USENIX Association, 2020. 2, 6

[9] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. BadNets: Evaluating Backdooring Attacks on Deep Neural Networks. *IEEE Access*, 7:47230–47244, 2019. 3

[10] Hao Guan, Pew-Thian Yap, Andrea Bozoki, and Mingxia Liu. Federated learning for medical image analysis: A survey. *Pattern Recognition*, page 110424, 2024. 1

[11] Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. Federated Learning for Mobile Keyboard Prediction, 2018. 1

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 5, 6

[13] Sebastian Houben, Johannes Stallkamp, Jan Salmen, Marc Schlipsing, and Christian Igel. Detection of traffic signs in real-world images: The german traffic sign detection benchmark. In *The 2013 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2013. 5

[14] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019. 5

[15] Meirui Jiang, Zirui Wang, and Qi Dou. Harmofl: Harmonizing local and global drifts in federated learning on heterogeneous medical images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1087–1095, 2022. 1

[16] Ruinan Jin and Xiaoxiao Li. Backdoor attack and defense in federated generative adversarial network-based medical image synthesis. *Medical Image Analysis*, 90:102965, 2023. 1

[17] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Toronto, ON, Canada, 2009. 5, 1

[18] Fan Lai, Yinwei Dai, Sanjay Singapuram, Jiachen Liu, Xiangfeng Zhu, Harsha Madhyastha, and Mosharaf Chowdhury. Fedscale: Benchmarking model and system performance of federated learning at scale. In *International conference on machine learning*, pages 11814–11827. PMLR, 2022. 1

[19] 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 *44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, May 21-25, 2023*, pages 1893–1907. IEEE, 2023. 2

[20] Minghui Li, Wei Wan, Yuxuan Ning, Shengshan Hu, Lulu Xue, Leo Yu Zhang, and Yichen Wang. Darkfed: A data-free backdoor attack in federated learning. *arXiv preprint arXiv:2405.03299*, 2024. 2

[21] Songze Li and Yanbo Dai. BackdoorIndicator: Leveraging OOD data for proactive backdoor detection in federated learning. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 4193–4210, Philadelphia, PA, 2024. USENIX Association. 2, 3, 6

[22] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3):50–60, 2020. 1

[23] Yige Li, Jiabo He, Hanxun Huang, Jun Sun, and Xingjun Ma. Shortcuts everywhere and nowhere: Exploring multi-trigger backdoor attacks, 2024. 8

[24] Ye Li, Jiale Zhang, Yanchao Zhao, Bing Chen, and Shui Yu. Fairness-aware federated learning framework on heterogeneous data distributions. In *ICC 2024-IEEE International Conference on Communications*, pages 728–733. IEEE, 2024. 1

[25] 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 AAAI Conference on Artificial Intelligence*, pages 9020–9028, 2023. 2

[26] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference*

on Artificial Intelligence and Statistics, pages 1273–1282. PMLR, 2017. 1

[27] Tuan Nguyen, Dung Thuy Nguyen, Khoa D Doan, and Kok-Seng Wong. Non-cooperative backdoor attacks in federated learning: A new threat landscape. *arXiv preprint arXiv:2407.07917*, 2024. 2, 8, 1

[28] Thien Duc Nguyen, Phillip Rieger, Markus Miettinen, and Ahmad-Reza Sadeghi. Poisoning attacks on federated learning-based iot intrusion detection system. In *Proc. Workshop Decentralized IoT Syst. Secur.(DISS)*, 2020. 1

[29] 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 *31st USENIX Security Symposium, USENIX Security 2022, Boston, MA, USA, August 10-12, 2022*, pages 1415–1432. USENIX Association, 2022. 2, 3, 6

[30] Yifan Niu and Weihong Deng. Federated learning for face recognition with gradient correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1999–2007, 2022. 1

[31] Phillip Rieger, Thien Duc Nguyen, Markus Miettinen, and Ahmad-Reza Sadeghi. Deepsight: Mitigating backdoor attacks in federated learning through deep model inspection. In *29th Annual Network and Distributed System Security Symposium, NDSS 2022, San Diego, California, USA, April 24-28, 2022*. The Internet Society, 2022. 6

[32] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 6

[33] Simon J Sheather and Michael C Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(3):683–690, 1991. 5

[34] Chenghui Shi, Shouling Ji, Xudong Pan, Xuhong Zhang, Mi Zhang, Min Yang, Jun Zhou, Jianwei Yin, and Ting Wang. Towards practical backdoor attacks on federated learning systems. *IEEE Transactions on Dependable and Secure Computing*, 2024. 2, 3

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 6

[36] Guanhong Tao, Zhenting Wang, Shiwei Feng, Guangyu Shen, Shiqing Ma, and Xiangyu Zhang. DRUPE: Distribution Preserving Backdoor Attack in Self-supervised Learning. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 29–29. IEEE Computer Society, 2023. 3

[37] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE symposium on security and privacy (SP)*, pages 707–723. IEEE, 2019. 3

[38] 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 *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 6

[39] Yongkang Wang, Dihua Zhai, Yufeng Zhan, and Yuanqing Xia. RFLBAT: A robust federated learning algorithm against backdoor attack. *CoRR*, abs/2201.03772, 2022. 2, 3, 6

[40] Emily Wenger, Josephine Passananti, Arjun Nitin Bhagoji, Yuanshun Yao, Haitao Zheng, and Ben Y Zhao. Backdoor attacks against deep learning systems in the physical world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6206–6215, 2021. 1

[41] Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. DBA: distributed backdoor attacks against federated learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 2

[42] Yi Zeng, Si Chen, Won Park, Z Morley Mao, Ming Jin, and Ruoxi Jia. Adversarial unlearning of backdoors via implicit hypergradient. *arXiv preprint arXiv:2110.03735*, 2021. 3

[43] Hangfan Zhang, Jinyuan Jia, Jinghui Chen, Lu Lin, and Dinghao Wu. A3FL: adversarially adaptive backdoor attacks to federated learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 61213–61233, 2023. 1, 2, 6

[44] Jiale Zhang, Chengcheng Zhu, Xiaobing Sun Di Wu, Jianming Yong, and Guodong Long. Badfss: backdoor attacks on federated self-supervised learning. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI)*, 2024. 2

[45] Jiale Zhang, Chengcheng Zhu, Chunpeng Ge, Chuan Ma, Yanchao Zhao, Xiaobing Sun, and Bing Chen. BadCleaner: Defending Backdoor Attacks in Federated Learning via Attention-Based Multi-Teacher Distillation. *IEEE Transactions on Dependable and Secure Computing*, 21(5):4559–4573, 2024. 2

[46] Jiale Zhang, Chengcheng Zhu, Xiaobing Sun, Chunpeng Ge, Bing Chen, Willy Susilo, and Shui Yu. *FLPurifier* : Backdoor Defense in Federated Learning via Decoupled Contrastive Training. *IEEE Transactions on Information Forensics and Security*, 19:4752–4766, 2024. 2

[47] Quan Zhang, Yifeng Ding, Yongqiang Tian, Jianmin Guo, Min Yuan, and Yu Jiang. Advdoor: adversarial backdoor attack of deep learning system. In *Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis*, pages 127–138, 2021. 2, 4

[48] Zhengming Zhang, Ashwinee Panda, Linyue Song, Yaoqing Yang, Michael Mahoney, Prateek Mittal, Ramchandran Kannan, and Joseph Gonzalez. Neurotoxin: Durable backdoors in federated learning. In *International Conference on Machine Learning*, pages 26429–26446. PMLR, 2022. 1, 2, 6

[49] Chengcheng Zhu, Jiale Zhang, Xiaobing Sun, Bing Chen, and Weizhi Meng. Adfl: Defending backdoor attacks in federated learning via adversarial distillation. *Computers & Security*, 132:103366, 2023. 2