Backdoor Cleansing with Unlabeled Data

Lu Pang, Tao Sun, Haibin Ling, Chao Chen
Stony Brook University
{luppang,tao,hling}@cs.stonybrook.edu, chao.chen.1@stonybrook.edu

Abstract

Due to the increasing computational demand of Deep Neural Networks (DNNs), companies and organizations have begun to outsource the training process. However, the externally trained DNNs can potentially be backdoor attacked. It is crucial to defend against such attacks, i.e., *to postprocess a suspicious model so that its backdoor behavior is mitigated while its normal prediction power on clean inputs remain uncompromised. To remove the abnormal backdoor behavior, existing methods mostly rely on additional labeled clean samples. However, such requirement may be unrealistic as the training data are often unavailable to end users. In this paper, we investigate the possibility of circumventing such barrier. We propose a novel defense method that does not require training labels. Through a carefully designed layer-wise weight re-initialization and knowledge distillation, our method can effectively cleanse backdoor behaviors of a suspicious network with negligible compromise in its normal behavior. In experiments, we show that our method, trained without labels, is on-par with state-of-the-art defense methods trained using labels. We also observe promising defense results even on out-of-distribution data. This makes our method very practical. Code is available at:* https://github.com/luluppang/BCU.

1. Introduction

Deep Neural Networks (DNNs) have achieved impressive performance in many tasks, *e.g.*, image classification [6], 3D point cloud generation [21] and object tracking [45]. However, the success usually relies on abundant training data and computational resources. Companies and organizations thus often outsource the training process to cloud computing or utilize pretrained models from third-party platforms. Unfortunately, the untrustworthy providers may potentially introduce backdoor attacks to the externally trained DNNs [9, 19]. During the training stage of a backdoor attack, an adversary stealthily injects a small portion of poisoned training data to associate a particular trigger with

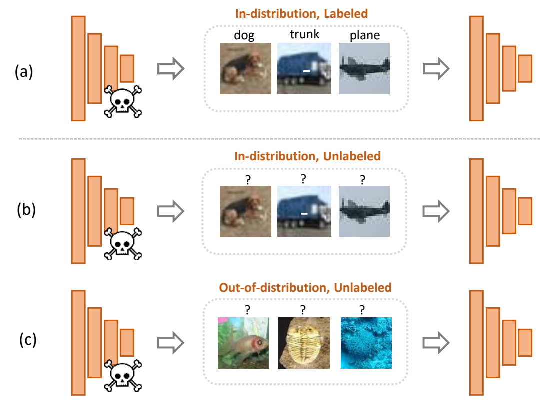


Figure 1. (a) Previous works use labeled in-distribution data to cleanse backdoor. Our work uses unlabeled in-distribution (b) or out-of-distribution data (c).

target class labels. During the inference stage, backdoor models predict accurately on clean samples but misclassify samples with triggers to the target class. Common triggers include black-white checkerboard [9], random noise pattern [5], physical object [37], *etc*.

To defend against backdoor attacks, one needs to postprocess a suspicious model so that its backdoor behavior is mitigated, and meanwhile, its normal prediction power on clean inputs remains uncompromised. To remove the abnormal backdoor behavior, existing methods mostly rely on additional labeled in-distribution clean samples [16, 18, 38, 40, 43, 44]. For example, Fine-Pruning [18] first prunes the dormant neurons for clean samples and then finetunes the model using ground-truth labels. Neural Attention Distillation (NAD) [16], a knowledge distillation-based method, uses labeled clean data to supervise the learning of a student model. Adversarial Neuron Pruning (ANP) [38] learns a mask to prune sensitive neurons with labeled clean data. These methods require $1\% - 5\%$ labeled clean training samples to effectively remove backdoor. Such requirement, however, is unrealistic in practice as the training data are often unavailable to end-users.

In this paper, we explore the possibility of circumventing such barrier with unlabeled data. As shown in Figure 1, we propose a novel defense method that does not require training labels. Meanwhile, we explore the ambitious goal of using only out-of-distribution data. These goals make the proposed defense method much more practical. End-users can be completely agnostic of the training set. To run the defense algorithm, they only need to collect some unlabeled data that do not have to resemble the training samples.

Inspired by knowledge distillation [8], we use a student model to acquire benign knowledge from a suspicious teacher model through their predictions on the readily available unlabeled data. Since the unlabeled data are usually clean images or images with slightly random noise, they are distinct from poisoned images with triggers. Therefore, trigger-related behaviors will not be evoked during the distillation. This effectively cleanses backdoor behaviors without significantly compromising the model's normal behavior. To ensure the student model focuses on the benign knowledge, which can be layer dependent, we propose an adaptive layer-wise weight re-initialization for the student model. Empirically, we demonstrate that even without labels, the proposed method can still successfully defend against the backdoor attacks. We also observe very promising defense results even with out-of-distribution unlabeled data that do not belong to the original training classes.

Our contributions are summarized as follows:

1. For the first time, we propose to defend against backdoor attacks using unlabeled data. This provides a practical solution to end-users under threat.
2. We devise a framework with knowledge distillation to transfer normal behavior of a suspicious teacher model to a student model while cleansing backdoor behaviors. Since the normal/backdoor knowledge can be layer-dependent, we design an adaptive layer-wise initialization strategy for the student model.
3. Extensive experiments are conducted on two benchmark datasets, CIFAR10 [14] and GTSRB [31]. Our method, trained without labels, is on-par with state-of-the-art defense methods trained with labels.
4. Meanwhile, we carry out an empirical study with out-of-distribution data. Our method achieves satisfactory defense performance against a majority of attacks. This sheds lights on a promising practical solution for end-users: they can use any collected images to cleanse a suspicious model.

2. Related Work

2.1. Backdoor Attack

During a backdoor attack, the adversary embeds a trigger into a DNN model by poisoning a portion of the training dataset at the training stage. At the inference stage, the backdoor model classifies clean samples accurately while predicts backdoor samples as the target label. The poisoned training samples are attached with a specific trigger and relabeled as the target label. A simple trigger can be a black-white checkerboard [9] or a single pixel [33]. These triggers are not stealthy since they can be perceived by human eyes. More complex triggers are developed such as a sinusoidal strip [1], an input-aware dynamic pattern [26, 29], *etc*. Recent works [7, 20, 25, 36] design more imperceptible triggers. Refool [20] utilizes a natural reflection phenomenon to design triggers. WaNet [25] uses elastic image warping technique to generate triggers. Lira [7] jointly optimizes trigger injection function and classification loss function to get stealthy triggers. BppAttack [36] improves the quality of triggers by using image quantization and injects triggers effectively with contrastive adversarial learning. Besides, some methods [1, 27, 30, 34] keep the original label of poisoned samples same as the target label. Such clean-label setting is more imperceptive for human inspectors. The key of these methods is to make models misclassify the clean target-label samples during the training process. Also, recent works show that backdoor attacks can be applied to federated learning [42], transfer learning [27], self-supervised learning [28], 3D point cloud classification [41], visual object tracking [17] and crowd counting [32].

2.2. Backdoor Defense

In model reconstruction-based defense, given a trained suspicious model, defenders modify the model directly to eliminate backdoor effects. Most methods in this category first synthesize possible triggers and then utilize synthesized triggers to mitigate backdoor effects [3, 4, 10, 35, 46]. With some clean samples, Neural Cleanse (NC) [35] synthesizes a trigger for each class and uses a Median Absolute Deviation (MAD) outlier detection algorithm to detect the final trigger. Then, an unlearning strategy is designed to unlearn the backdoor effects. Following NC [35], other methods [3, 4, 10, 46] are proposed to improve the quality of synthesized triggers. For example, ShapPruning [10] employs Shapley estimation to synthesize triggers and then detect sensitive neurons to synthesized triggers. Chen et al. [4] locates a "wining backdoor lottery ticket" to preserve trigger-related information. These methods heavily depend on the quality of the synthesized triggers, and thus can be unsatisfactory when facing more advanced triggers [7, 29].

Other works explore pruning-based defense methods [18, 38]. The core idea is to detect and prune bad neurons. For example, Fine-Pruning [18] prunes bad neurons of the last convolution layer, and then uses clean samples to finetune the pruned model. Adversarial Neuron Pruning (ANP) [38] treats pruning sensitive neurons as a minimax problem under adversarial neuron perturbations. The Implicit Backdoor Adversarial Unlearning (I-BAU) algorithm [43] solves the minimax optimization by utilizing the implicit hyper-gradient. Besides, Mode Connectivity Re-

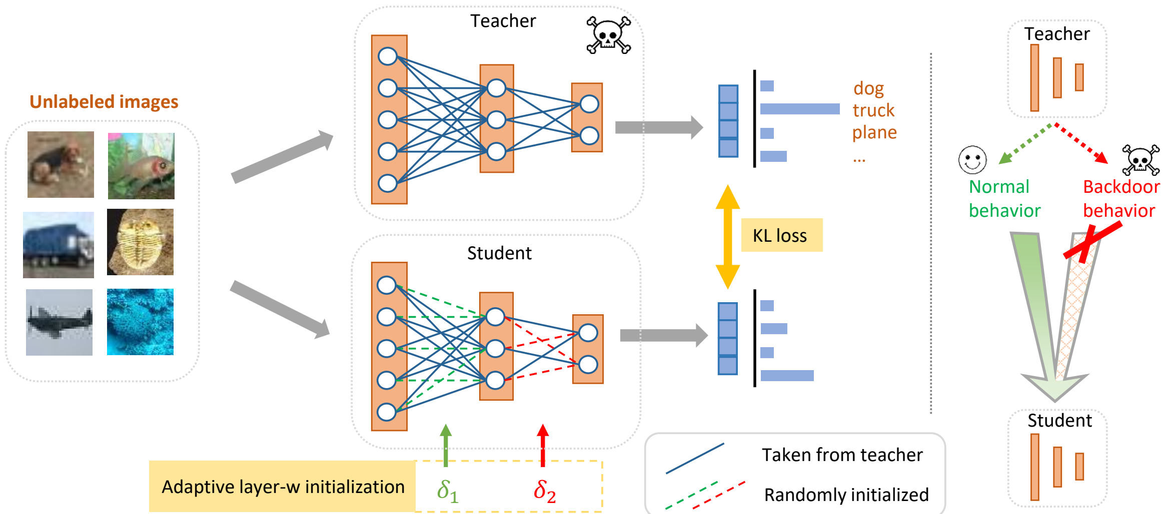


Figure 2. Proposed backdoor cleansing framework. The student model learns normal behavior from the teacher model through knowledge distillation on unlabeled images. Backdoor behavior of the teacher model is neglected.

pair (MCR) [44] is explored to remove backdoor effects. Although effective, these methods require labeled clean samples, which in practice may not be available. By contrast, our solution, also in the model-reconstructing category, does not need labeled clean samples.

Knowledge distillation has been used in backdoor mitigation [16, 40]. Both Neural Attention Distillation (NAD) [16] and Attention Relation Graph Distillation (ARGD) [40] transfer feature attention knowledge of a finetuned backdoor model into the original backdoor model. These methods crucially rely on the finetuning stage, and thus depend on labeled clean samples. The key insight of our method is that model prediction on data automatically carries rich and benign knowledge of the original model. Through a layer-adaptive weight initialization strategy, *our method can directly cleanse backdoors without any label*.

While most existing works assume the defense as a post-processing step, we also note some recent methods focusing on designing backdoor-resilient training strategy. Since a defender can access the training process, some works modify the training strategy to train a robust model [13, 15]. Huang et al. [13] decompose end-to-end training process into three stages including self-supervised feature learning, classifier learning and finetuning whole classifier with filtered samples. Based on the characteristics of backdoor model training, Li et al. [15] proposes a two-stage gradient ascent strategy instead of standard training. Other studies mitigate backdoor effects via randomized smoothing [24], noise injection [23] and strong data augmentation [2], *etc*.

3. Method

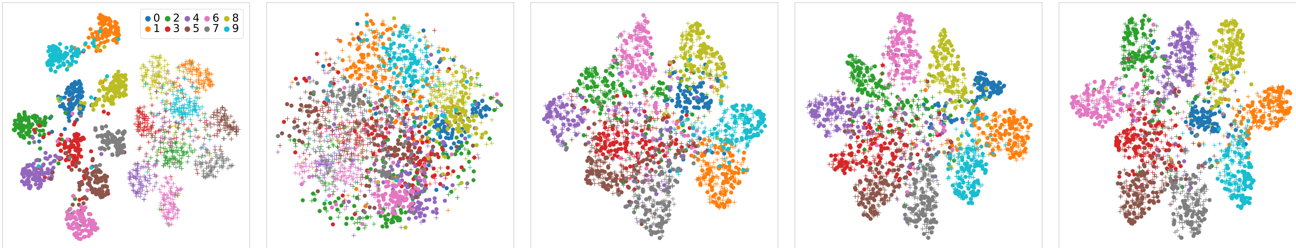
Our main idea is to directly use knowledge distillation to cleanse backdoor behaviors. The rationale is three-folds. First, knowledge distillation directly transfers knowledge through the logits output, which carries the rich posterior probability distribution information of a model. By approximating the logits output on samples, the student model can naturally mimic the normal behavior of the teacher model. Second, we argue that the backdoor behavior is an abnormal phenomenon forced into the teacher model. Knowledge distillation through clean samples will implicitly regularize the transferred knowledge, and "smooth" out the abnormal behavior. Finally, prior study has observed that backdoor behavior is embodied in certain neurons whose distribution is layer dependent [22]. By designing an adaptive weight initialization, we can more effectively transfer normal knowledge of the teacher model and filter out backdoor behavior. The framework of our method is illustrated in Figure 2.

3.1. Preliminary

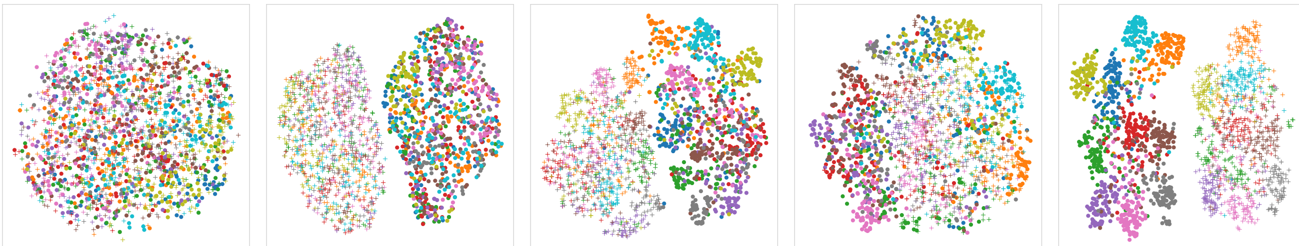
Attack Setting. In backdoor attack for classification task, a DNN model $f_\theta : \mathcal{X} \longrightarrow \mathcal{Y}$ is trained, where $\mathcal{X} \subset \mathbb{R}^d$ is the input space and $\mathcal{Y} = \{1, 2, ..., K\}$ is the label space. An image dataset $D_{\text{attack}} = \{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}\}_{i=1}^{n}$ is split by $D_{\text{attack}} = D_{\text{clean}} \bigcup D_{\text{backdoor}}$, where D_{backdoor} is used to create backdoor images. The backdoor injection rate is defined as $\gamma = \frac{|D_{\text{backdoor}}|}{|D_{\text{attack}}|}$. An image transformation function $\Phi(\cdot)$ transforms a clean image into a backdoor image, *e.g.*, through stacking a checkerboard pattern to the original image. $\eta(\cdot, \cdot)$ transforms its ground truth label into a target label. The objective function for backdoor attack is

$$\begin{aligned}\mathcal{L}_{\text{attack}} =& \mathbb{E}_{(x,y)\sim D_{\text{clean}}}[\ell_{\text{ce}}(f_\theta(x), y)] \\ &+ \mathbb{E}_{(x,y)\sim D_{\text{backdoor}}}[\ell_{\text{ce}}(f_\theta(\Phi(x)), \eta(x, y))]\end{aligned} \tag{1}$$

where ℓ_{ce} is the cross entropy loss function. With this loss function, the obtained backdoor model is expected to be-



(a) Teacher (b) Stu (adaptive init, ep 0) (c) Stu (adaptive init, ep 1) (d) Stu (adaptive init, ep 2) (e) Stu (adaptive init, ep 11)



(f) Stu (uniform init, ep 0) (g) Stu (1st layer init, ep 0) (h) Stu (2nd layer init, ep 0) (i) Stu (3rd layer init, ep 0) (j) Stu (4th layer init, ep 0)

Figure 3. t-SNE visualization of penultimate features on CIFAR10 from *Badnets* attack. **Top**: the teacher model and student models at different training epochs with adaptive layer-wise initialization. **Bottom**: student models at epoch 0 with different initialization strategies. Each color denotes a class. '$\circ$' are clean images and '$+$' the corresponding backdoor ones. More discussions can be found in Section 3.3.

have normally on clean test images, while misclasify backdoor images to the target class label.

Defense Setting. We assume that defenders download a backdoored model from an untrustworthy platform and can not access the training process. Some clean images D_{defense} are given for backdoor defense. The goal of defense is to preserve the classification accuracy (ACC) on clean data and decrease the classification accuracy on backdoor images *i.e.* attack success rate (ASR).

3.2. Backdoor Cleansing via Knowledge Distillation

Our motivation is to directly extract clean information (or knowledge) from a suspicious model. Since a backdoor model usually behaves differently for clean and backdoor images, the trigger-related behaviors will not be evoked when the model is fed with clean images. Inspired by response-based knowledge distillation [12], we adopt the teacher-student framework to distillate benign knowledge from a suspicious teacher model through its predictions on clean images. As illustrated in Figure 2, the normal behaviors of the teacher model are transferred to the student model, while the backdoor behaviors are neglected. This effectively cleanses backdoor behaviors without significantly compromising the model's performances on clean images.

Since we use the the logits output of the teacher model as the supervision, our proposed framework does not need ground-truth labels. In fact, even when the input images are out-of-distribution data that do not belong to the training classes, the student model can acquire useful knowledge from the teacher model's predicted probabilities.

Let z^t and z^s be the output logits of the teacher model and student model, respectively. Their temperature scaled probability vectors can be obtained as $p_T^t[k] = \frac{\exp(z_k^t/T)}{\sum_j \exp(z_j^t/T)}$ and $p_T^s[k] = \frac{\exp(z_k^s/T)}{\sum_j \exp(z_j^s/T)}$. T is a temperature hyper-parameter. Our defense objective function is

$$\mathcal{L}_{\text{defense}} = \mathbb{E}_{(x,y)\sim D_{\text{val}}} D_{\text{KL}}[p_T^t \| p_T^s] \tag{2}$$

where $D_{\text{KL}}[\cdot\|\cdot]$ is the KL divergence.

Qualitative Analysis. To show the effectiveness of knowledge distillation, we visualize the penultimate feature representations of clean and backdoor images throughout the process of knowledge distillation, and plot in the top row of Figure 3. The compactness and separability of clean image clusters reflect the model's prediction ability on normal data. Also, if backdoor behaviors are cleansed, the backdoor images will fall into the corresponding clean clusters. In Fig. 3a, we can see that the clean images form 10 clusters, indicating a high ACC of the teacher model. The backdoor images are distant to the clean images and form separate clusters. Hence the teacher model behaves abnormally on backdoor data. For the student model after adaptive layer-wise initialization in Fig. 3b, clean images from the same class are still close to each other, showing that some benign knowledge are preserved after initialization. This provides a good starting point for the following knowledge distillation. Figures 3c-3e show the results after training for some epochs. The normal behaviors are gradually transferred to the student model. With this, clean images form tighter clusters and are better separated. Backdoor images turn to overlap with the clean images with the same class labels, showing that the backdoor behaviors are successfully cleansed.

3.3. Adaptive Layer-wise Initialization

It is generally believed that backdoor behavior is embodied through "bad" neurons. By random weight initialization and knowledge distillation on clean samples, we expect such neurons will be naturally cleansed. Previous observations [22] reveal that these "bad" neurons can be distributed differently at different layers, and the distribution is architecture- and dataset-dependent. In order to (1) break connection between triggers and target label and (2) preserve more normal knowledge simultaneously, we propose an adaptive layer-wise initialization strategy to initialize the student model.

Assuming the suspicious teacher model has L layers, the weights can be represented as $W^t = \{W_l^t | 1 \leq l \leq L\}$ where $W_l^t \in \mathbb{R}^{C_{\text{out}} \times C_{\text{in}} \times K \times K}$ for a convolution layer and $W_l^t \in \mathbb{R}^{C_{\text{out}} \times C_{\text{in}}}$ for a linear layer. We also have another random initialized student model, whose architecture is same as teacher model. Similarly, the weights of random initialized student model can be represented as $W^s = \{W_l^s | 1 \leq l \leq L\}$ where $W_l^s \in \mathbb{R}^{C_{\text{out}} \times C_{\text{in}} \times K \times K}$ for a convolution layer and $W_l^s \in \mathbb{R}^{C_{\text{out}} \times C_{\text{in}}}$ for a linear layer. Here, we consider a tuned hyperparameter δ_l for l-th layer. Then the initialization mask is defined as

$$M = \{m_l | 1 \leq l \leq L, m_l \in \{0,1\}^{\text{shape}(W_l^s)}, \sum m_l = \delta_l |m_l|\}$$

where $|m_l|$ is the size of initializing mask. Then, initialized student model W^{s*} can be formulated as follows:

$$ALI(W^{s*}, \delta) = \bigcup_{i=1}^{L} ((1 - m_l) \odot W_l^t + m_l \odot W_l^s) \quad (3)$$

where $\delta = \{\delta_l | 1 \leq l \leq L\}$ is the ratio of random initializing weights per layer.

Qualitative Analysis. Similar to previous analysis in Sec. 3.3, we study the effects of adaptive layer-wise initialization for the student model through visualizing clean and backdoor sample features. The comparison strategies include uniform initialization that uses a same random initialization ratio for every layer, and single-layer initialization. To match our adaptive layer-wise initialization, we choose a specific ratio for the uniform initialization so that the total number of randomized weights equals in the two strategies. The same ratio is used for single-layer initialization.

Comparing Figure 3f with Figure 3b, we can find that uniform initialization breaks the connection between trigger and target label. However, the benign information is also discarded as all clean images clutter together in the figure. From Fig. 3g-3j, When randomly initialize shallow layers like 1st or 2nd layer, the connection between trigger and target label is not broken while the clustering structure of clean images are destroyed. When randomly initialize deep layers like 3rd or 4th, the clean information can be preserved. The backdoor information is also partially eliminated in Fig. 3i, where backdoor images become more dispersed. Therefore, to make a balance between preserving clean information and discarding backdoor information, it is better to use higher random initialization ratios for deeper layers and smaller ratios for shallow ones. This justifies the motivation of our adaptive layer-wise initialization.

Algorithm 1 Backdoor Cleansing with Unlabeled Data

Input: Backdoor model f^t with weights W^t, random initialized student model f^s with weights W^s, adaptive ratios δ, unlabeled clean data $\mathcal{D}_{\text{defense}}$, training epochs E, iterations per epoch I and temperature T.
Output: Clean model f

for $l = 0$ **to** $|W^t|$ **do**
  Sample $R_l^{\text{shape}(W_l^t)} \sim \text{Uniform}(0,1)$
  Obtain boolean weight mask $m_l = \mathbb{I}[R_l < \delta_l]$
  $W_l^s = (1 - m_l) \odot W_l^t + m_l \odot W_l^s$
end for
for $e = 0$ **to** E **do**
  for $i = 0$ **to** I **do**
    Sample mini-batches $\mathcal{B}_{\text{val}}$ from $\mathcal{D}_{\text{defense}}$
    Obtain temperature scaled probability p_T^t from f^t, and p_T^s from f^s
    Update student model weights W^s with $\mathcal{L}_{\text{defense}} = D_{\text{KL}}[p_T^t \| p_T^s]$
  end for
  $f \leftarrow f^s$
end for

4. Experiments

4.1. Experiment settings

Datasets and Architecture. We conduct all backdoor models on two datasets include CIFAR10 [14] and GTSRB [31]. For CIFAR10 and GTSRB, we split their original test datasets into defense dataset and test dataset. The total size of each defense dataset is 5000. Tiny-ImageNet [39] is used as the out-of-distribution dataset. We also construct another out-of-distribution dataset "Tiny-ImageNet++" from ImageNet [6]. Tiny-ImageNet++ contains 20,000 images distributed evenly in 1000 classes. Its image resolution is the same as Tiny-ImageNet. ResNet-18 [11] is adopted as the model architecture. From shallow to deep, ResNet-18 includes 1 convolution layer, 8 basic blocks and 1 FC layer. Except for FC layer, the more shallow the layer is, the less the weights are. The ratios of first convolution layer and FC

Backdoor Attacks	Original		In-distribution Labeled												In-distribution Unlabeled			
			Finetuning		Fine-pruning		MCR (t=0.3)		ANP		NAD		I-BAU		Ours		Ours*	
	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC
Badnets	99.93	92.76	9.70	92.55	32.36	92.57	1.68	86.41	2.56	88.58	4.67	92.35	10.16	91.98	3.00	92.15	3.00	92.75
Blended	100.00	94.48	5.20	93.44	20.62	93.70	6.39	87.51	0.87	92.85	5.06	93.24	6.19	92.71	4.90	93.16	5.10	93.65
IAB	91.35	87.46	9.46	86.91	2.45	86.89	1.35	85.29	0.60	85.37	2.17	86.76	7.57	85.64	1.96	86.42	1.90	86.85
LC	99.55	94.51	97.14	93.49	60.23	93.88	5.33	88.18	4.62	91.30	52.74	93.38	21.41	92.72	1.81	93.17	1.40	93.66
SIG	95.09	93.71	5.41	93.16	5.66	93.55	2.33	87.69	0.41	92.09	1.88	92.95	15.76	92.45	0.91	92.58	1.18	93.14
WaNet	97.15	93.53	0.98	92.34	13.99	92.92	1.14	91.08	0.31	90.61	1.03	92.22	1.73	91.62	9.86	92.05	16.67	92.64
Mean	97.18	92.74	21.31	91.98	22.55	92.25	3.04	87.69	1.56	90.14	11.26	91.82	10.47	91.19	3.74	91.59	4.87	92.11
Drop ↓	–	–	75.86	0.76	74.63	0.49	94.14	5.05	95.62	2.61	85.92	0.92	86.71	1.56	93.44	1.15	92.30	0.63

Table 1. Defense results on backdoor models trained on CIFAR10. (*Using double unlabeled data.)

Backdoor Attacks	Original		In-distribution Labeled												In-distribution Unlabeled			
			Finetuning		Fine-pruning		MCR (t=0.3)		ANP		NAD		I-BAU		Ours		Ours*	
	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC
Badnets	100.0	97.22	99.99	99.80	97.71	99.54	61.26	99.51	19.00	89.47	9.22	99.79	0.00	99.66	0.02	96.75	0.00	97.88
Blended	100.0	98.89	5.45	99.81	5.80	99.73	1.69	99.71	0.14	98.47	0.38	99.84	1.00	99.77	0.50	97.32	0.37	98.90
IAB	98.74	98.01	58.91	99.79	2.23	99.80	3.94	99.79	0.08	96.39	46.94	99.88	0.02	99.80	0.15	96.91	0.07	98.07
LC	94.74	95.75	67.68	99.74	96.37	99.59	3.07	99.50	0.11	94.15	37.82	99.72	0.03	99.71	0.86	96.64	0.81	96.60
SIG	97.80	98.87	96.59	99.84	99.06	99.80	93.06	99.74	78.43	98.22	96.64	99.86	30.54	99.77	1.59	97.09	6.31	98.71
WaNet	93.58	98.69	0.61	99.84	9.73	99.84	0.12	99.85	0.00	98.36	0.01	99.88	0.01	99.81	0.11	97.59	0.02	98.80
Mean	97.48	97.91	54.87	99.81	51.82	99.72	27.19	99.68	16.29	95.84	31.83	99.83	5.27	99.75	0.54	97.05	1.26	98.16
Drop ↓	–	–	42.60	-1.90	45.66	-1.81	70.29	-1.78	81.18	2.06	65.64	-1.92	92.21	-1.85	96.94	0.86	96.21	-0.25

Table 2. Defense results on backdoor models trained on GTSRB. (*Using double unlabeled data.)

layer are set 0.01 and 0.1. The ratios of eight basic blocks are 0.01, 0.01, 0.03, 0.03, 0.09, 0.09, 0.27 and 0.27.

Backdoor attacks setting. We evaluate all defenses on six representative backdoor attacks including Badnets [9], Blended attack [5], Label-consistent backdoor attack (LC) [34], Sinusoidal signal backdoor attack (SIG) [1], Input-aware dynamic backdoor attack (IAB) [26] and WaNet [25]. LC and SIG represent two classic clean-label backdoor attacks. Badnets, Blended, IAB and WaNet are representatives of label-poisoned backdoor attacks. Specifically, Badnets is a patch-based visible backdoor attack. Blended is a noise-based invisible attack. IAB is a dynamic backdoor attack. WaNet is an image-transformation-based invisible attack. For a fair comparison, the poison ratio for label-poisoned attacks is set as 0.1. For label-poisoned attacks, we poison 80% samples of target label. The *all-to-one* strategy is adopted for all backdoor attacks.

Backdoor defense setting. We compare our method with six state-of-the-art defense methods including standard finetuning, Fine-pruning [18], Mode Connectivity Repair (MCR) [44], Adversarial Neuron Pruning (ANP) [38], Neural Attention Distillation (NAD) [16] and Implicit Backdoor Adversarial Unlearning (I-BAU) [43].

For each attack, we train 14 backdoor models with different target labels and random seeds. We conduct all defenses on 14 models and the average is the final results. For fair comparison, we train 100 epochs for all defense methods. We set the batch size as 256 and optimize our framework using Stochastic Gradient Descent (SGD) with a momentum of 0.9, and a weight decay of 0.0005. The adopted data augmentation techniques include random crop and random horizontal flipping. For MCR, we get a benign model by finetuning the original backdoor model with 10 epochs.

4.2. Comparison with other defense methods

Results using unlabeled in-distribution data. We compare with six state-of-the-art defenses with regard to ACC and ASR. Other six defenses use labeled clean samples, while our framework uses unlabeled samples. We assume that all defenses can access 2500 clean samples. For our method, we also present results using 5000 unlabeled samples in the last two columns. Results on CIFAR10 [14] and GTSRB [31] are shown in Table 1 and Table 2, separately. Despite that our framework is trained without using ground-truth labels, its performance is still comparative with other methods that require labels. For CIFAR10, due to the usage of labels, existing works get the highest ACC of 92.25%. However, these works can not decrease ASRs largely while keep high ACC. Our method reduces ASR to 3.74% with negligible ACC reduction of 1.15%. For GTSRB, since ground-truth labels are utilized, ACCs increase slightly in five of six defenses. However, our framework obtains a robust model by reducing average ASR to less than 1%, which is better than other label-based methods. Meanwhile, the ACC reduction of our framework is only 0.86%. With 5000 unlabeled data, our ACC increases 0.25%.

For both datasets, ANP succeeds in dropping ASR of most attacks, but at the expense of lower accuracies compared other methods. ANP aims to prune the bad neurons without re-training backdoor model. However, the backdoor neurons are difficult to distinguish from normal neu-

Backdoor Attacks	In distribution		Out-of-distribution					
	CIFAR10		GTSRB		Tiny-IN		Tiny-IN++	
	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC
Badnets	3.00	92.15	11.30	81.19	4.47	91.24	3.03	92.44
Blended	4.90	93.16	6.68	82.39	61.75	92.88	11.87	93.66
IAB	1.96	86.42	1.62	81.91	1.52	86.00	1.52	86.76
LC	1.81	93.17	3.49	84.47	1.95	92.83	1.46	93.67
SIG	0.91	92.58	0.98	81.49	14.58	91.85	17.79	92.86
WaNet	9.86	92.05	83.89	84.11	7.62	91.58	22.60	92.52
Mean	3.74	91.59	17.99	82.59	15.32	91.06	9.71	91.99

Table 3. Defense results on CIFAR10 using different unlabeled out-of-distribution data.

Backdoor Attacks	Uniform		Adaptive decreasing		Adaptive increasing	
	ASR	ACC	ASR	ACC	ASR	ACC
Badnets	4.88	92.08	2.38	86.75	3.00	92.15
Blended	4.54	93.01	3.32	88.33	4.90	93.16
IAB	1.72	86.25	2.68	81.51	1.96	86.42
LC	4.18	93.01	1.05	88.22	1.81	93.17
SIG	0.58	92.31	1.07	88.24	0.91	92.58
WaNet	7.35	91.69	2.17	84.76	9.86	92.05
Mean	3.87	91.39	2.11	86.30	3.74	91.59

Table 4. Comparison of weights initialization strategies for student model on CIFAR10 (in-distribution).

rons in reality. Some neurons critical to ACC may be pruned by ANP, leading to degraded performances. Fine-pruning gets a low average drop over ASR since Fine-pruning simply prunes the dormant neurons in the last convolution layer. However, complex triggers activate neurons across different layers. Since a finetuning stage follows the pruning process, Fine-pruning has a high ACC. We find that finetuning, Fine-prunng and NAD perform badly on LC attack in reducing ASR. All of three defenses include a finetuning stage. Though NAD distillates attention map knowledge from teacher to student model, teacher model is obtained by finetuning backdoor model and student model is supervised by CrossEntropy loss. One possible reason is that the PGD perturbations used in LC hinder finetuning to associate normal images with target labels with limited clean samples. MCR introduces a curve model to find a path connection between two backdoor models. With limited data samples, MCR achieves low ACC compared other methods. In all six defenses, I-BAU perform well on both datasets. I-BAU adopt implicit hypergradient to solve minmax optimization, leading to strong generalizability of the robustness. Note that most defense methods can not defend SIG attack on GTSRB because we improve sinusoidal signal to inject backdoor successfully (Δ is set 60 in our experiments). This strong signal is not stealthy to GTSRB images, causing backdoor model learn strong abnormal behaviors and difficult to defend.

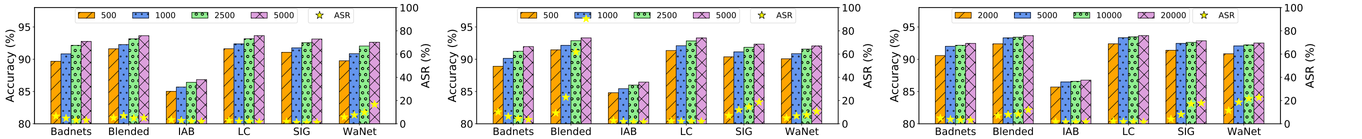
Results using unlabeled out-of-distribution data. We conduct experiments on CIFAR10 by using out-of-distribution unlabeled data. GTSRB, Tiny-ImageNet and Tiny-ImageNet++ are three out-of-distribution unlabeled datasets. Table 3 reports the results. For GTSRB and Tiny-ImageNet, we random sample 2500 images from our constructed defense dataset.

Compared to in-distribution data, GTSRB reduces ASRs largely in five of six attacks while perform badly on WaNet. The possible reason is that simple GTSRB images e.g. circle or triangular signs, introduce warping-based backdoor behavior. Due to large domain gap between GTSRB and CIFAR10, GTSRB decreases average ACC about 10%. With Tiny-ImageNet, our method can reduce ASRs largely especially for Badnets, IAB, LC and WaNet, with negligible ACC cost. However, Tiny-ImageNet can not reduce ASR successfully on Blended Attack. Meanwhile, Tiny-ImageNet++ reduces ASR to 11.87% on LC. Blended trigger is a random noise. Removing random noise trigger needs more out-of-distribution natural clean images. Due to the large size and diversity, Tiny-ImageNet++ performs better than GTSRB and Tiny-ImageNet. Tiny-ImageNet++ reduces average ASR to less than 10%, while other two datasets reduce ASRs to more than 15%. Tiny-ImageNet++ can also keep ACC high after defense.

4.3. Analysis

Size of unlabeled samples. We use CIFAR10 to analyze influence of the size of unlabeled samples. Figure 4 (a–c) show the results using in-distribution CIFAR10, out-of-distribution Tiny-ImageNet and Tiny-ImageNet++. For three datasets, we randomly sample 500, 1000, 2500, 5000 images, separately. As the number of samples increases, ACCs increase and ASRs decrease for most cases. However, with the number of unlabeled Tiny-ImageNet and Tiny-ImageNet++ data increasing, ASRs raise up on Blended, SIG and WaNet attacks. Blended attack injects backdoor by blending clean images and random noise. The trigger of SIG is a sinusoidal signal. WaNet applies elastic warping to design triggers. All three triggers are stealthy and cause slight change to images. Some images in Tiny-ImageNet++ are downloaded from the internet and might include light noise similar to the three triggers. Therefore, using more out-of-distribution unlabeled images from Tiny-ImageNet or Tiny-ImageNet++ might cause ASRs increasing for the three attacks.

Adaptive layer-wise initialization. We analyze the effectiveness of different adaptive layer-wise initialization strategies by conducting experiments on CIFAR10. Three strategies are designed including random initialize weights of student model with uniform ratio, increasing ratio and decreasing ratio. For fair comparison, the overall ratio of random initialization keeps around 0.2 for three strategies. The results are presented in Table 4. All of three strategies can reduce ASRs to less than 5%. However, adaptive decreasing layer-wise initialization performs bad on ACCs. The reason is that random initializing two many weights in low layers causes student model dropping too much information



(a) In-distribution CIFAR10 (b) Out-of-distribution Tiny-ImageNet (c) Out-of-distribution Tiny-ImageNet++

Figure 4. Defense results on CIFAR10 using different numbers of unlabeled samples.

Backdoor Attacks	In-distribution				Out-of-distribution (Tiny-IN)			
	Soft		Hard		Soft		Hard	
	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC
Badnets	3.00	92.15	3.37	91.16	4.47	91.24	5.55	88.74
Blended	4.90	93.16	5.14	92.09	61.75	92.88	69.48	90.71
IAB	1.96	86.42	1.64	85.31	1.52	86.00	2.05	83.85
LC	1.81	93.17	1.86	92.05	1.95	92.83	1.40	90.61
SIG	0.91	92.58	1.42	91.61	14.58	91.85	16.11	89.69
WaNet	9.86	92.05	3.16	90.75	7.62	91.58	4.59	89.03
Mean	3.74	91.59	2.77	90.50	15.32	91.06	16.53	88.77

Table 5. Comparisons of using soft predictions and hard predictions of backdoor models for distillation on CIFAR10.

related to low-level features. It is difficult to recover effectively only by aligning two probability distributions between student and teacher models. Compared to uniform initializing strategy, adaptive increasing layer-wise initialization obtains lowest ASR and highest ACC.

Effectiveness of knowledge distillation. To evaluate the effectiveness of knowledge distillation, we compare the performances using soft labels and hard labels. Hard labels are class labels with the maximum probability of teacher model outputs. Soft labels are soft probability with temperature T described in Section 3. Cross-Entropy loss function is employed for hard labels setting. The experiments are conducted on CIFAR10 and out-of-distribution dataset is Tiny-ImagneNet. Table 5 shows the results. It shows that hard and soft labels achieve comparative performance for in-distribution unlabeled data. The reason is that backdoor teacher model predicts high ACC for in-distribution images. Therefore, most hard labels are ground-truth labels. However, backdoor teacher model can not predict correct hard labels for out-of-distribution data. Some classes of out-of-distribution images even does not exist in the CIFAR10. Therefore, using soft labels is better than hard labels. Specifically, ASR of using soft labels is 1.21% lower than ASR of using hard labels. ACC of using soft labels is 2.29% higher than ACC of using hard labels.

Diversity of out-of-distribution data. To study how diversity of out-of-distribution data influences defense performance, we create several versions of Tiny-ImangeNet++ with different configurations of (number of class, number of samples per class). The total number of unlabeled images is fixed to 2000. Then we apply them to cleanse backdoor models trained on CIFAR10. Figure 5 plots the curves of ACC and ASR. ACCs are close for different configurations. However, as the unique number of classes in the training

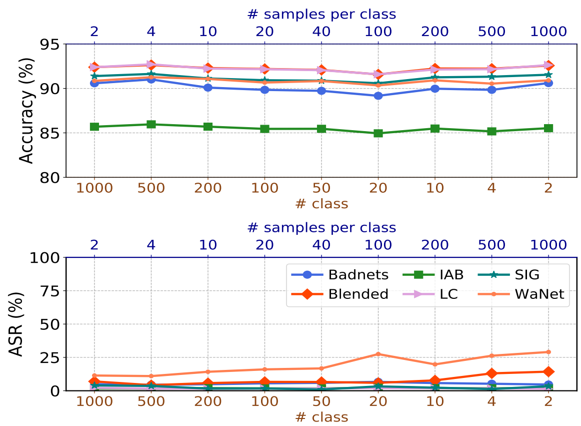


Figure 5. Defense results on CIFAR10 using Tiny-ImageNet++ created with different configurations.

data increases, ASR has a tendency to decrease, showing that backdoor behaviors are more effectively eliminated. In principle, increasing the diversity of out-of-distribution unlabeled data is beneficial as more data modes are covered. It is more likely that data similar to the training distribution are included. Also, the student model can learn more general knowledge in making classification than specific ones.

5. Conclusion

In this paper, for the first time, we explore the possibility of using unlabeled data including in-distribution and out-of-distribution data to remove backdoor from a backdoor model. A knowledge distillation framework with a carefully designed adaptive layer-wise initialization strategy is proposed. We conduct experiments on two datasets including CIFAR10 and GTSRB against six representative backdoor attacks. Results show that our framework can successfully defend backdoor attacks with negligible clean accuracy decrease, compared with existing methods using labeled in-distribution data.

Acknowledge This effort was partially supported by the Intelligence Advanced Research Projects Agency (IARPA) and Army Research Office (ARO) under Contract No. W911NF20C0038, and by US National Science Foundation Grants (No. 2128187, No. 2128350 and No. 2006655). Any opinions, findings, and conclusions in this paper are those of the authors only and do not necessarily reflect the views of our sponsors.

References

[1] Mauro Barni, Kassem Kallas, and Benedetta Tondi. A new backdoor attack in cnns by training set corruption without label poisoning. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 101–105. IEEE, 2019. 2, 6

[2] Eitan Borgnia, Valeriia Cherepanova, Liam Fowl, Amin Ghiasi, Jonas Geiping, Micah Goldblum, Tom Goldstein, and Arjun Gupta. Strong data augmentation sanitizes poisoning and backdoor attacks without an accuracy tradeoff. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3855–3859. IEEE, 2021. 3

[3] Huili Chen, Cheng Fu, Jishen Zhao, and Farinaz Koushanfar. Deepinspect: A black-box trojan detection and mitigation framework for deep neural networks. In *IJCAI*, volume 2, page 8, 2019. 2

[4] Tianlong Chen, Zhenyu Zhang, Yihua Zhang, Shiyu Chang, Sijia Liu, and Zhangyang Wang. Quarantine: Sparsity can uncover the trojan attack trigger for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 598–609, 2022. 2

[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. 1, 6

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 5

[7] Khoa Doan, Yingjie Lao, Weijie Zhao, and Ping Li. Lira: Learnable, imperceptible and robust backdoor attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11966–11976, 2021. 2

[8] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021. 2

[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. 1, 2, 6

[10] Jiyang Guan, Zhuozhuo Tu, Ran He, and Dacheng Tao. Few-shot backdoor defense using shapley estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13358–13367, 2022. 2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[12] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 4

[13] Kunzhe Huang, Yiming Li, Baoyuan Wu, Zhan Qin, and Kui Ren. Backdoor defense via decoupling the training process. In *International Conference on Learning Representations*, 2021. 3

[14] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2, 5, 6

[15] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Anti-backdoor learning: Training clean models on poisoned data. *Advances in Neural Information Processing Systems*, 34:14900–14912, 2021. 3

[16] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. 2021. 1, 3, 6

[17] Yiming Li, Haoxiang Zhong, Xingjun Ma, Yong Jiang, and Shu-Tao Xia. Few-shot backdoor attacks on visual object tracking. *arXiv preprint arXiv:2201.13178*, 2022. 2

[18] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *Research in Attacks, Intrusions, and Defenses*, pages 273–294, 2018. 1, 2, 6

[19] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. 1

[20] Yunfei Liu, Xingjun Ma, James Bailey, and Feng Lu. Reflection backdoor: A natural backdoor attack on deep neural networks. In *European Conference on Computer Vision*, pages 182–199. Springer, 2020. 2

[21] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2837–2845, 2021. 1

[22] Weimin Lyu, Songzhu Zheng, Tengfei Ma, and Chao Chen. A study of the attention abnormality in trojaned berts. *arXiv preprint arXiv:2205.08305*, 2022. 3, 5

[23] Yuzhe Ma, Xiaojin Zhu, and Justin Hsu. Data poisoning against differentially-private learners: Attacks and defenses. *arXiv preprint arXiv:1903.09860*, 2019. 3

[24] Nikita Muravev and Aleksandr Petiushko. Certified robustness via randomized smoothing over multiplicative parameters. *arXiv preprint arXiv:2106.14432*, 2021. 3

[25] Anh Nguyen and Anh Tran. Wanet–imperceptible warping-based backdoor attack. *arXiv preprint arXiv:2102.10369*, 2021. 2, 6

[26] Tuan Anh Nguyen and Anh Tran. Input-aware dynamic backdoor attack. *Advances in Neural Information Processing Systems*, 33:3454–3464, 2020. 2, 6

[27] Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash. Hidden trigger backdoor attacks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11957–11965, 2020. 2

[28] Aniruddha Saha, Ajinkya Tejankar, Soroush Abbasi Koohpayegani, and Hamed Pirsiavash. Backdoor attacks on self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13337–13346, 2022. 2

[29] Ahmed Salem, Rui Wen, Michael Backes, Shiqing Ma, and Yang Zhang. Dynamic backdoor attacks against machine learning models. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pages 703–718. IEEE, 2022. 2

[30] Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein.

Poison frogs! targeted clean-label poisoning attacks on neural networks. *Advances in neural information processing systems*, 31, 2018. 2

[31] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks*, 32:323–332, 2012. 2, 5, 6

[32] Yuhua Sun, Tailai Zhang, Xingjun Ma, Pan Zhou, Jian Lou, Zichuan Xu, Xing Di, Yu Cheng, and Lichao Sun. Backdoor attacks on crowd counting. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5351–5360, 2022. 2

[33] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. *Advances in neural information processing systems*, 31, 2018. 2

[34] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks. *arXiv preprint arXiv:1912.02771*, 2019. 2, 6

[35] 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. 2

[36] Zhenting Wang, Juan Zhai, and Shiqing Ma. Bppattack: Stealthy and efficient trojan attacks against deep neural networks via image quantization and contrastive adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15074–15084, 2022. 2

[37] 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

[38] Dongxian Wu and Yisen Wang. Adversarial neuron pruning purifies backdoored deep models. *Advances in Neural Information Processing Systems*, 34:16913–16925, 2021. 1, 2, 6

[39] Jiayu Wu, Qixiang Zhang, and Guoxi Xu. Tiny imagenet challenge. *Technical report*, 2017. 5

[40] Jun Xia, Ting Wang, Jieping Ding, Xian Wei, and Mingsong Chen. Eliminating backdoor triggers for deep neural networks using attention relation graph distillation. *arXiv preprint arXiv:2204.09975*, 2022. 1, 3

[41] Zhen Xiang, David J Miller, Siheng Chen, Xi Li, and George Kesidis. A backdoor attack against 3d point cloud classifiers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7597–7607, 2021. 2

[42] Chulin Xie, Minghao Chen, Pin-Yu Chen, and Bo Li. Crfl: Certifiably robust federated learning against backdoor attacks. In *International Conference on Machine Learning*, pages 11372–11382. PMLR, 2021. 2

[43] Yi Zeng, Si Chen, Won Park, Zhuoqing Mao, Ming Jin, and Ruoxi Jia. Adversarial unlearning of backdoors via implicit hypergradient. In *International Conference on Learning Representations*, 2021. 1, 2, 6

[44] Pu Zhao, Pin-Yu Chen, Payel Das, Karthikeyan Natesan Ramamurthy, and Xue Lin. Bridging mode connectivity in loss landscapes and adversarial robustness. 2020. 1, 3, 6

[45] Linyu Zheng, Ming Tang, Yingying Chen, Guibo Zhu, Jinqiao Wang, and Hanqing Lu. Improving multiple object tracking with single object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2453–2462, 2021. 1

[46] Liuwan Zhu, Rui Ning, Cong Wang, Chunsheng Xin, and Hongyi Wu. Gangsweep: Sweep out neural backdoors by gan. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3173–3181, 2020. 2