The Victim and The Beneficiary: Exploiting a Poisoned Model to Train a Clean Model on Poisoned Data

Zixuan Zhu, Rui Wang*, Cong Zou, Lihua Jing
SKLOIS, Institute of Information Engineering, CAS, Beijing, China
School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
{zhuzixuan, wangrui, zoucong, jinglihua}@iie.ac.cn

Abstract

Recently, backdoor attacks have posed a serious security threat to the training process of deep neural networks (DNNs). The attacked model behaves normally on benign samples but outputs a specific result when the trigger is present. However, compared with the rocketing progress of backdoor attacks, existing defenses are difficult to deal with these threats effectively or require benign samples to work, which may be unavailable in real scenarios. In this paper, we find that the poisoned samples and benign samples can be distinguished with prediction entropy. This inspires us to propose a novel dual-network training framework: The Victim and The Beneficiary (V&B), which exploits a poisoned model to train a clean model without extra benign samples. Firstly, we sacrifice the Victim network to be a powerful poisoned sample detector by training on suspicious samples. Secondly, we train the Beneficiary network on the credible samples selected by the Victim to inhibit backdoor injection. Thirdly, a semi-supervised suppression strategy is adopted for erasing potential backdoors and improving model performance. Furthermore, to better inhibit missed poisoned samples, we propose a strong data augmentation method, AttentionMix, which works well with our proposed V&B framework. Extensive experiments on two widely used datasets against 6 state-of-the-art attacks demonstrate that our framework is effective in preventing backdoor injection and robust to various attacks while maintaining the performance on benign samples. Our code is available at https://github.com/Zixuan-Zhu/VaB.

1. Introduction

Large amounts of training data and powerful computing resources are two key factors for the success of deep neural networks (DNNs). While in practical applications, obtaining enough samples for training is laborious, and training is likely to be resource-constrained. As a result, more and more third-party data (*e.g*. data released on the Internet) and training platforms (*e.g*. Google Colab) are adopted to reduce the overhead, which brings new security risks.

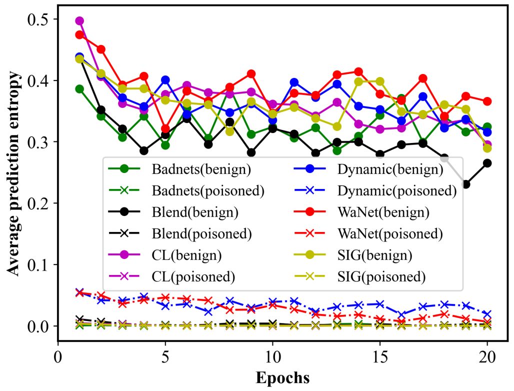


Figure 1. The average prediction entropy of benign samples versus poisoned samples crafted by 6 backdoor attacks. We conduct the experiment on CIFAR-10 with ResNet-18 under poisoning rate 10%, where our special training strategy (described in section 3.3) was adopted.

Backdoor attacks [5, 11, 24] pose a serious security threat to the training process of DNNs and can be easily executed through data poisoning. Specifically, attackers [7, 10, 22] inject the designed *trigger* (*e.g*. a small patch or random noise) to a few benign samples selected from the training set and change their labels to the attacker-defined *target label*. These poisoned samples will force models to learn the correlation between the trigger and the target label. In the reference process, the attacked model will behave normally on benign samples but output the target label when the trigger is present. Deploying attacked models can lead to severe consequences, even life-threatening in some scenarios (*e.g*. autonomous driving). Hence, a secure training framework is needed when using third-party data or training platforms.

Many researchers have devoted themselves to this topic and proposed feasible defenses. Li *et al*. [18] utilize local benign samples to erase existing backdoors in DNNs, and

*Corresponding author

Borgnia *et al.* [3] employ strong data augmentations to inhibit backdoor injection during training. Li *et al.* [17] find that the training loss decreased faster for poisoned samples than benign samples and designed a loss function to separate them. Finally, they select a fixed portion of samples with the lowest loss to erase potential backdoors. In real scenarios, it is challenging to distinguish poisoned samples from benign ones in this way due to the unknown poisoning rate and close loss values. In experiments, we find the entropy of model prediction is a more discriminative property, and a fixed threshold could filter out most poisoned samples, as shown in Figure 1. During training, poisoned samples will be learned faster than benign samples due to the similar feature of their triggers [17]. Hence, the poisoned network can confidently predict poisoned samples as the target label in early epochs but hesitates about the benign samples.

Based on this observation, we propose a novel secure training framework, *The Victim and The Beneficiary (V&B)*, that can directly train clean models on poisoned data without resorting to benign samples. Firstly, in warming up stage, the training dataset is recomposed into a suspicious set and a credible set according to the prediction entropy is lower or higher than the predefined threshold. Then the Victim network is trained iteratively with the suspicious set to obtain a strong poisoned sample detector. Secondly, in the clean training stage, the Beneficiary network is trained with credible samples filtered by the Victim to inhibit backdoor injection. In later training epochs, the Victim network is fully exploited to provide its learned knowledge for the Beneficiary network to help the latter erase potential backdoors through semi-supervised learning. To diminish the threat of missed poisoned samples, we design a strong data augmentation *AttentionMix*, which mixes the image region with high activation values according to the attention map. Compared with existing augmentations [27, 30, 31], it has a stronger inhibition effect against stealthy backdoor attacks [23, 22, 26].

The main contributions of this paper are as follows: **(1)** We propose a novel dual-network secure training framework, V&B, which can train clean models on the poisoned dataset without resorting to benign samples. **(2)** We design a powerful data augmentation method called AttentionMix, which has a stronger inhibition effect against stealthy backdoor attacks. **(3)** Extensive experiments on two popular benchmark datasets demonstrate the effectiveness and robustness of our framework.

2. Related Work

2.1. Data Augmentation

Data Augmentation is a common-used technique to improve the generalization capabilities of DNNs. In addition to some popular weak data augmentations (such as cropping, flipping, rotation, etc.), many strong data augmentation techniques [27, 30, 31] have been proposed to further improve model performance. Mixup [31] linearly interpolates any two images and their labels, then trains the model with generated samples. However, the generated samples tend to be unnatural, which will decrease model performances on localization and object detection tasks. Instead of linear interpolation, Cutmix [30] replaces a random patch with the same region from another image and mixes their labels according to the ratio of the patch area. Further, Attentive Cutmix [27] only pastes the influential region located by an attention mechanism to other images, effectively forcing models to learn discriminative details. Yet it also mixes labels based on the ratio of the region area, ignoring the importance of the region in two original images, and direct pasting may bring complete triggers into benign samples. Our AttentionMix takes into account the importance of the region in both images and blends the influential region with the same area in another image, which can destroy the completeness of triggers.

2.2. Backdoor Attack and Defense

In this paper, we only consider poisoning-based attacks toward image classification that can occur during the data preparation stage.

Poisoned-based backdoor attack. In early studies, attackers usually adopt simple and visible patterns as the trigger, such as a black-white checkerboard [10]. To escape human inspections, some works use human-imperceptible deformations [22] as the trigger or design unique triggers [16, 23] for each poisoned sample. However, they randomly select benign samples to inject triggers and change their labels (called poison-label attacks), possibly causing the content of an image not to match its label. Instead of randomly selecting samples to poison, Turner *et al.* [26], Mauro *et al.* [1], and Liu *et al.* [21] only inject triggers into benign samples under the target label (called clean-label attacks). Although some triggers are imperceptible in the input space, they can be easily detected by defenses focused on latent space extracted by CNNs. Therefore, recent studies [33, 34] not only ensure the trigger's invisibility in the input space but also restrict the similarity between poison samples and corresponding benign samples in the latent space.

Backdoor Defense. For poisoning-based attacks, defense methods could be divided into three categories from their working stage, including **(1)** preprocessing before training [9, 12, 25], **(2)** poison suppression during training [14, 17, 20], and **(3)** backdoor erasing after training [18, 28, 32]. For preprocessing, Hayase *et al.* [12] and Chen *et al.* [4] filter the training set through statistics or clustering and train the model with benign samples. For poison suppressing, Borgnia *et al.* [3] leverage strong data augmentation (*e.g.* Mixup, Cutmix) to alleviate poisoned samples' adverse effects. Huang *et al.* [14] design a decoupling training process, which first learns a feature extractor via SimCLR [6] to prevent poisoned samples from clustering in feature space. For backdoor erasing, Liu *et al.* [19] prune the low-activation neurons for validation samples (benign samples) and then fine-tune the model with these benign samples. While Li *et al.* [18] first fine-tune the poisoned model with a small portion of benign samples to obtain a teacher model, then erase the original model's backdoor through knowledge distillation.

2.3. Semi-supervised Learning

Semi-supervised learning [8] aims to alleviate the dependence of deep learning on extensive labeled training data and jointly learns the model from a limited amount of labeled samples and a large number of unlabeled samples. Existing methods [2, 29, 35] generally assign each unlabeled sample with a pseudo label and design different loss functions for labeled and unlabeled samples. The weight of unlabeled sample loss will gradually increase during training to reduce the negative effect of low-quality pseudo labels in early epochs.

3. Method

3.1. Preliminaries

Pipeline of Poisoned-based Backdoor Attacks. Let $\mathcal{D}_B = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ denotes the benign training set containing N samples, where $\boldsymbol{x}_i \in \{0, 1, ..., 255\}^{C \times H \times W}$ is the image, $y_i \in \{0, 1, ..., K\}$ is its label, K is the number of classes, H, W and C are the height, width and channels of an image. The attacker will select a subset of benign samples to poison with a designed injection function $G(\cdot)$. Let $\mathcal{D}_G$ indicates the generated poisoned samples set and $\mathcal{D}_R$ denotes the remaining benign samples set, where $\mathcal{D}_G = \{(\boldsymbol{x}', y_t) | \boldsymbol{x}' = G(\boldsymbol{x}), (\boldsymbol{x}, y) \in \mathcal{D}_B \setminus \mathcal{D}_R\}$, y_t is the target label and $r = \frac{|\mathcal{D}_G|}{\mathcal{D}_B}$ is the poisoning rate. The poisoned dataset $\mathcal{D}_P = \mathcal{D}_G \cup \mathcal{D}_R$ will be unknowingly adopted by victim users to train DNNs. At inference time, the attacker can manipulate the poisoned model's output through the injected backdoor.

Threat Model and Defense Setting. In this paper, we focus on defending against poisoned-based backdoor attacks toward image classification. The attacker can arbitrarily modify the training set, but can not change other training components (*e.g.* model structure and hyperparameters) [1, 10, 22]. Such attacks may occur when using third-party datasets or training platforms that can poison users' clean data. We follow the common assumption [12, 14, 17] that defenders have no information about the backdoor (including poisoning rate, trigger pattern, target label, etc.) and any local benign samples to use but have full control over the training process. The defense goal is to prevent the trained DNNs from predicting poisoned samples as the target label while maintaining its accuracy on benign samples.

3.2. Overview

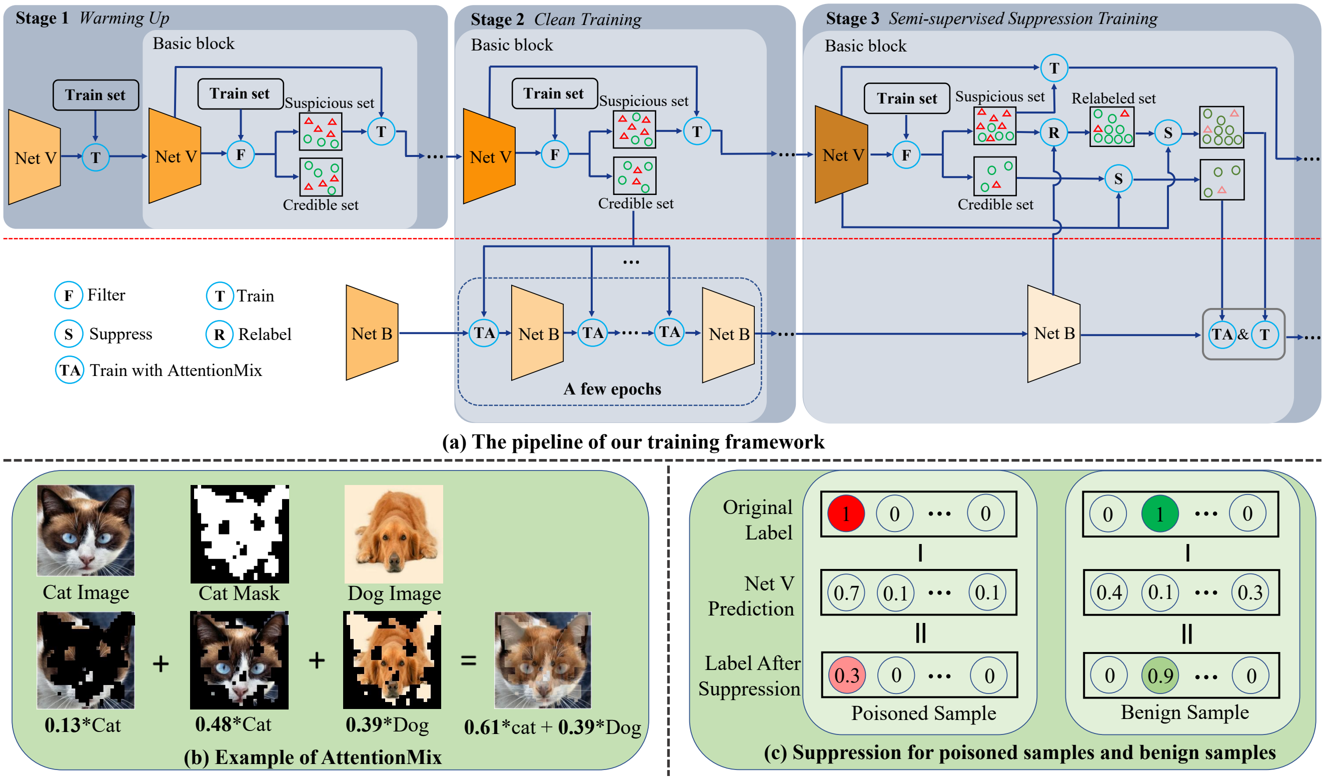
In this subsection, we briefly introduce the pipeline of our framework. As Figure 2(a) illustrates, our framework consists of two networks, and the whole training process is divided into three stages — warming up, clean training and semi-supervised suppression training. In warming up, we aim to obtain a poisoned network that can roughly distinguish poisoned samples from benign ones according to the prediction entropy. Therefore we train the Victim network with only filtered suspicious samples (*i.e.* suspicious set) except for the first epoch. This training strategy remains unchanged in the following two stages. During clean training, we adopt credible samples (*i.e.* credible set) to train the Beneficiary network with our proposed AttentionMix data augmentation. The strong data augmentation can effectively prevent missed poisoned samples from creating backdoors. As the Victim network is trained, the entropy of its prediction for benign samples will also slowly decrease below the filtering threshold, leading to a reduction in the credible set size. So one credible set will be used for a few epochs to ensure the Beneficiary network gets enough training samples.

To further improve the Beneficiary network performance and erase its potential backdoor, we regard the credible set as a labeled dataset and the suspicious set as an unlabeled dataset to train the Beneficiary network via semi-supervised learning. Following conventional practice [2, 15], each unlabeled sample will be assigned a pseudo label generated by the Beneficiary network, forming the relabeled set. However, if the Beneficiary network still contains a backdoor after stage 2, poisoned samples in the suspicious set will be relabeled as the target label, strengthening the backdoor. After specific training, the Victim network is sensitive to trigger patterns and thus can be utilized to suppress the component corresponding to the target class after relabeling. Furthermore, we also apply this suppression operation to the credible set to further weaken the impact of missed poisoned samples.

3.3. Warming Up

At this stage, we aim to obtain a poisoned network that has only learned trigger patterns. In this way, the entropy of its prediction for poisoned samples will be significantly lower than that of benign samples. Then we can use a threshold to filter out poisoned samples.

Let $\mathcal{D}$ denotes the training set and $F_v : \{0, 1, ..., 255\}^{C \times H \times W} \rightarrow [0, 1]^K$ indicates the Vic-



(a) The pipeline of our training framework

(b) Example of AttentionMix

(c) Suppression for poisoned samples and benign samples

Figure 2. (a) The pipeline of our training framework. Net V is the Victim network (*i.e.* poisoned network) that only learns from the filtered suspicious set, and Net B is the Beneficiary network (*i.e.* clean network) we wanted. Red triangles represent poisoned samples, and green circles are benign samples. The color of nets indicates their degree of poisoning: the darker the color, the deeper the poisoning, and vice versa. Deep blue areas are three training stages, and light blue areas are basic blocks that will be executed many times in the corresponding stage. (b) An example of AttentionMix. The bottom four images are the non-activation region of the cat image, the activation region of the cat image, the same activation region in the dog image, and the final generated image, respectively. And their labels are located below the images. (c) Suppression for poisoned samples and benign samples. The target label is assumed as the first label. Suppression can reduce the target label component of poisoned samples but have little influence on the ground-truth label of benign samples.

tim network. We first use the whole training set to train the Victim network for one epoch with cross-entropy loss. Then the prediction entropy of a sample will be calculated with Equation 1, where p is the prediction of the Victim network for that sample.

$$e = -\sum_{k=1}^{K} p_k \log p_k \qquad (1)$$

Once all samples' prediction entropy is obtained, we normalize them to $[0, 1]$ through *mix-max normalization*. Then a threshold t_f is set to filter out suspicious samples $\mathcal{D}_s = \{(\boldsymbol{x}_i, y_i) | (\boldsymbol{x}_i, y_i) \in \mathcal{D}, e_i < t_f\}$ and the remaining samples are regarded as credible samples $\mathcal{D}_c$. To strengthen the learning of poisoned samples and delay the learning of benign samples, the Victim network will only be trained on suspicious sets in the future. The loss function is as follows:

$$\mathcal{L}_v = \mathbb{E}_{(\boldsymbol{x},y)\in\mathcal{D}_s}[\mathcal{L}_{CE}(F_v(\boldsymbol{x}), y)] \qquad (2)$$

where $\mathcal{L}_{CE}$ denotes the cross-entropy loss.

After the first epoch, the prediction entropy of some poisoned samples may still be larger than that of most benign ones, although there is a large gap between their average values in Figure 1. If the filtering threshold t_f is set too large, lots of benign samples will be discarded, which will affect the model performance on benign samples. So we set a smaller t_f and wait for the prediction entropy of these poisoned samples to decrease. As training progresses, their prediction entropy will drop below t_f, making the credible set cleaner. Simultaneously, the prediction entropy of benign samples will also decrease, resulting in a smaller credible set. We expect the Victim network to filter out a clean and relatively large credible set after this stage.

3.4. Clean Training

At this stage, we begin to train the Beneficiary network with filtered credible samples $\mathcal{D}_c$, where the filtering strategy is the same as in stage 1. Because of the rough separation, a few poisoned samples may be included in the credible set. For strong attacks, a small number of poisoned samples is enough to create a backdoor. So we design a strong data augmentation called AttentionMix to inhibit the missed poisoned samples from working. Compared with existing augmentations [27, 30, 31], AttentionMix has a stronger inhibition effect on stealthy backdoor attacks.

The core idea of AttentionMix is to generate a new training sample $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$ by mixing two training samples $(\boldsymbol{x}_1, y_1)$ and $(\boldsymbol{x}_2, y_2)$ according to their attention maps, where $\tilde{\boldsymbol{y}}$ is a K-dimensional vector. Following [18], we adopt one sample's final feature maps to compute its attention map:

$$\boldsymbol{A}' = \sum_{i=1}^{c} |F_i|^2 \tag{3}$$

where $\boldsymbol{F} \in \mathbb{R}^{c \times h \times w}$ indicates feature maps, $\boldsymbol{A}' \in \mathbb{R}^{h \times w}$ is the corresponding attention map, c, h and w are the dimensions of channel, height and width. For each attention map, we first bilinearly interpolate it to the original image size $H \times W$, and then normalize its elements to $[0, 1]$ by *min-max normalization*.

After getting the attention map, we set a threshold t_m to obtain the activation region that we want to mix:

$$\boldsymbol{M}_{i,j} = \begin{cases} 0 & \boldsymbol{A}_{i,j} < t_m \\ 1 & \boldsymbol{A}_{i,j} \geq t_m \end{cases} \tag{4}$$

where $\boldsymbol{M} \in \mathbb{R}^{H \times W}$ is the binary mask indicating which pixels belong to the activation region and $\boldsymbol{A}$ is the attention map after interpolation and normalization.

Let $\boldsymbol{A}^1$ and $\boldsymbol{A}^2$ denote the attention maps of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, $\boldsymbol{M}^1$ denotes the first sample's mask. As Figure 2(b) shows, the mixed image $\tilde{\boldsymbol{x}}$ is composed of three parts: the non-activation region of $\boldsymbol{x}_1$, the activation region of $\boldsymbol{x}_1$ and the same area in $\boldsymbol{x}_2$. During mixing, the non-activation region of $\boldsymbol{x}_1$ remains unchanged, and the activation region is blended with the same area of $\boldsymbol{x}_2$:

$$\begin{aligned} \tilde{\boldsymbol{x}} = (\boldsymbol{1} - \boldsymbol{M}^1) \odot \boldsymbol{x}_1 + \lambda_2 \cdot (\boldsymbol{M}^1 \odot \boldsymbol{x}_1) \\ + \lambda_3 \cdot (\boldsymbol{M}^1 \odot \boldsymbol{x}_2) \end{aligned} \tag{5}$$

where $\odot$ is the Hadamard Product, λ_2 and λ_3 are the mixing weights calculated from the two attention maps. The calculation process is as follows:

$$w_1 = \frac{\mathrm{sum}(\boldsymbol{M}^1)}{\mathrm{sum}(\boldsymbol{A}^1)}, w_2 = \frac{\mathrm{sum}(\boldsymbol{M}^1 \odot \boldsymbol{A}^2)}{\mathrm{sum}(\boldsymbol{A}^2)} \tag{6}$$

$$\lambda_1 = 1 - w_1, \lambda_2 = \frac{w_1}{w_1 + w_2}, \lambda_3 = \frac{w_2}{w_1 + w_2} \tag{7}$$

where $\mathrm{sum}(\cdot)$ is a summation function, w_1 and w_2 denote the importance of the activation region to the two images, and λ_1 is the importance of the non-activation region to $\boldsymbol{x}_1$. The final label is also composed of three parts:

$$\tilde{\boldsymbol{y}} = \lambda_1 \cdot \boldsymbol{y}_1 + w_1 \cdot (\lambda_2 \cdot \boldsymbol{y}_1 + \lambda_3 \cdot \boldsymbol{y}_2) \tag{8}$$

where $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are the one-hot labels of the two samples. When the activation region of $\boldsymbol{x}_1$ corresponds to the background of $\boldsymbol{x}_2$, w_1 will be significantly larger than w_2, and λ_3 will be a small value, which can reduce background interference during mixing.

We denote the generated new credible set as $\tilde{\mathcal{D}}_c$ and train the Beneficiary network F_b with the following loss function (*i.e. Train with AttentionMix* in Fiqure 2(a)):

$$\mathcal{L}_b = \mathbb{E}_{(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) \in \tilde{\mathcal{D}}_c}[\mathcal{L}_{CE}(F_b(\tilde{\boldsymbol{x}}), \tilde{\boldsymbol{y}})] \tag{9}$$

For a poisoned network, the activation region of poisoned samples is more likely their trigger locations. Mixing the activation region can destroy the completeness of trigger patterns. While for benign samples, the mixing plays a regularization role that prevents the network from overfitting a particular subject and forces it to learn abundant features.

3.5. Semi-supervised Suppression Training

Until now, we have got a poisoned Victim network that can filter out most of the poisoned samples and a Beneficiary network that is relatively clean but may still contain the backdoor. In the previous stage, the Beneficiary network is trained only with partial training samples (*i.e.* the credible set), which may harm its performance on benign samples. Thus we intend to enable the deprecated suspicious samples via semi-supervised learning at this stage to improve the Beneficiary network while erasing its potential backdoor.

Given a credible set $\mathcal{D}_c$ and a suspicious set $\mathcal{D}_s$, we use samples in $\mathcal{D}_c$ as labeled samples and remove the labels of samples in $\mathcal{D}_s$ to transform the problem into a semi-supervised learning setting. For each unlabeled sample, we average the predictions of the Beneficiary network over its multiple augmentations (*e.g.* random crop and random flip) and select the class with max probability as its pseudo-label, which is similar to [2] and [15]. The relabeled samples are denoted as $\mathcal{D}_r = \{(\boldsymbol{x}_i, \hat{y}_i) | (\boldsymbol{x}_i, y_i) \in \mathcal{D}_s\}$.

However, the poisoned samples in $\mathcal{D}_s$ may be relabeled as the target label, *i.e.* $\hat{y}_i = y_t$, if the Beneficiary network still contains the backdoor. We can mitigate the impact of such mislabeling by exploiting the knowledge learned by the Victim network. Specifically, we adopt the Victim network's prediction to suppress the pseudo-label (*i.e. Suppress* in Figure 2(a)):

$$\boldsymbol{y}'_i = \mathrm{clip}(\hat{\boldsymbol{y}}_i - \mathrm{softmax}(F_v(\boldsymbol{x}_i)), 0, 1) \tag{10}$$

where $(\boldsymbol{x}_i, \hat{y}_i) \in \mathcal{D}_r$, $\hat{\boldsymbol{y}}_i$ is the one-hot vector of $\hat{y}_i$ and $\mathrm{clip}(\cdot, 0, 1)$ is a function that will restrict its inputs to $[0, 1]$ (*i.e.* inputs smaller than 0 became 0, inputs larger than 1 became 1, and the rest remain unchanged).

We use $\mathcal{D}'_r = \{(\boldsymbol{x}_i, \boldsymbol{y}'_i) | (\boldsymbol{x}_i, \hat{y}_i) \in \mathcal{D}_r\}$ to indicate the suspicious samples after relabeling and suppression. We

Attack	BadNets [10]		Blend [7]		Dynamic [23]		WaNet [22]		SIG [1]		CL-16 [26]		CL-32 [26]	
Metric	BA	ASR	BA	ASR	BA	ASR	BA	ASR	BA	ASR	BA	ASR	BA	ASR
No Defense	94.36%	100.00%	94.57%	99.76%	94.66%	95.18%	94.21%	95.84%	95.09%	64.79%	94.98%	96.93%	95.03%	94.73%
FP [19]	92.72%	1.92%	92.49%	10.61%	92.47%	11.73%	93.27%	0.90%	93.55%	37.98%	92.92%	9.74%	93.03%	6.39%
NAD [18]	88.83%	1.83%	87.72%	1.38%	88.30%	2.32%	91.00%	0.99%	90.42%	3.52%	88.92%	1.64%	90.19%	2.60%
ABL [17]	90.47%	0.70%	91.31%	31.49%	83.07%	94.21%	80.78%	19.93%	88.61%	1.37%	90.29%	**0.54%**	86.50%	2.68%
DBD [14]	92.00%	3.06%	91.62%	3.78%	90.98%	17.56%	89.13%	**0.11%**	89.42%	1.61%	92.85%	91.08%	93.24%	14.88%
V&B (Ours)	**93.96%**	**0.62%**	**94.37%**	**0.63%**	**93.91%**	**1.13%**	**94.15%**	0.54%	**94.08%**	**0.17%**	**94.24%**	1.01%	**93.98%**	**0.64%**

Table 1. The performance of 5 defense methods against 7 backdoor attacks on CIFAR-10. The first 4 attacks are poison-label attacks and the last 3 attacks are clean-label attacks. Note that FP and NAD require local benign samples to work, while ABL, DBD and our method do not. The best results are marked in boldface and underlined items are the second-best results.

Attack	BadNets [10]		Blend [7]		WaNet [22]		SIG [1]	
Metric	BA	ASR	BA	ASR	BA	ASR	BA	ASR
No Defense	94.94%	99.51%	95.16%	100.00%	92.63%	100.00%	90.38%	89.22%
FP [19]	86.28%	1.05%	87.63%	2.03%	82.82%	26.19%	82.50%	22.76%
NAD [18]	85.83%	3.82%	86.73%	2.90%	83.93%	3.88%	80.71%	3.88%
ABL [17]	92.15%	13.78%	87.34%	56.57%	91.38%	56.71%	80.13%	0.87%
DBD [14]	93.08%	0.31%	93.81%	0.87%	92.06%	**1.00%**	84.47%	98.47%
V&B (Ours)	**95.42%**	**0.28%**	**95.03%**	**0.45%**	**94.84%**	1.92%	**94.65%**	**0.03%**

Table 2. The performance of 5 defense methods against 4 backdoor attacks on ImageNet subset.

also suppress the labels of credible samples in the same way to weaken the influence of poisoned samples that may be missed after filtering. As in stage 2, the suppressed credible samples are augmented with our AttentionMix, forming the training set $\tilde{\mathcal{D}}'_c$ to further inhibit backdoor injection. We fine-tune the Beneficiary network with the following loss function:

$$\begin{aligned}\mathcal{L} = \mathbb{E}_{(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}')\in\tilde{\mathcal{D}}'_c}[\mathcal{L}_{CE}(F_b(\tilde{\boldsymbol{x}}),\tilde{\boldsymbol{y}}')]+ \\ \alpha\cdot\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y}')\in\mathcal{D}'_r}[\mathcal{L}_{CE}(F_b(\boldsymbol{x}),\boldsymbol{y}')]\end{aligned} \tag{11}$$

Since the Victim network is only trained with suspicious samples, it will confidently predict poisoned samples as the target label but give a wrong prediction for benign samples with high probability. Thus for a poisoned sample, the suppression can reduce the component corresponding to the target class in its label while having little influence on the component of a sample's ground-truth class, as shown in Figure 2(c).

4. Experiments

4.1. Experimental Settings

Datasets and Models. We employ ResNet-18 [13] as our Victim and Beneficiary networks, a standard setting in previous studies [14]. For a fair comparison, we evaluate all defenses on two popular benchmark datasets, including CIFAR-10 and a subset of ImageNet containing 12 classes generated by Li *et al.* [17]. The details are summarized in Appendix A.1.

Attack Configures. We consider 6 representative backdoor attacks in our experiments, including four poison-label attacks: BadNets [10], Blend attack [7], WaNet attack [22], Dynamic attack [23], and two clean-label attacks: Sinusoidal signal attack (SIG) [1] and Clean-label attack (CL) [26]. To achieve the best attack performance, we configure these attacks as close as possible to their released codes and original settings (*e.g.* trigger pattern and trigger size). While on the ImageNet subset, we implement these attacks based on the settings suggested by [14], since some attacks miss the corresponding experiments in their original papers. We uniformly set the poisoning ratio as 0.1 and select the first class as the target class in both datasets, *i.e.* $y_t = 0$. More details about attacks are shown in Appendix A.2.

Defense Configures and Training Details. We compare our V&B with 4 state-of-the-art defenses: Fine-pruning (FP) [19], Neural Attention Distillation (NAD) [18], Anti-Backdoor Learning (ABL) [17] and Decoupling (DBD) [14], in which FP and NAD require additional local benign samples to work. Similarly, we follow their original settings and released codes. For V&B, we warm up the Victim network for 3 epochs on CIFAR-10 and 6 epochs on the ImageNet subset, where the filtering threshold t_f linearly decreases from 1 to 0.2 and from 1 to 0.4. The selection of these hyperparameters and an automatic warming-up strategy can be found in Appendix A.3. In the following stages, the filtering threshold keeps unchanged (*i.e.* 0.2 for CIFAR-10 and 0.4 for ImageNet subset). For fast convergence, we adopt the Adam optimizer to train the Victim network with a learning rate of 0.001 in the whole training process. We optimize the Beneficiary network for 200 epochs using an SGD optimizer with a momentum of 0.9 and a weight decay factor of 10^{-4}, where the first 100 epochs belong to stage 2 and the last 100 epochs belong to stage 3. The learning rate is initially set to 0.1 and is divided by 10 after every 50

Attack	BadNets [10]		Blend [7]		Dynamic [23]		WaNet [22]		CL-16 [26]	
Metric	BA	ASR	BA	ASR	BA	ASR	BA	ASR	BA	ASR
No Defense	94.36%	100%	94.57%	99.76%	94.66%	95.18%	94.21%	95.84%	94.98%	96.93%
Mixup [31]	93.37%	**0.27%**	92.80%	3.49%	92.97%	42.10%	92.97%	42.10%	94.80%	20.03%
Cutmix [30]	93.52%	0.40%	93.40%	1.07%	91.40%	53.23%	91.40%	53.23%	94.54%	1.16%
Attentive Cutmix [27]	**94.26%**	0.40%	**94.55%**	4.62%	93.09%	13.38%	93.09%	13.88%	**95.18%**	2.97%
AttentionMix (Ours)	93.96%	0.62%	94.37%	**0.63%**	**93.91%**	**1.13%**	**94.15%**	**0.54%**	94.24%	**1.01%**

Table 3. The inhibition effect of 4 strong data augmentations against 5 backdoor attacks on CIFAR-10.

epochs. In particular, the Victim network is optimized every 5 epochs at stage 2, *i.e.* one credible set is used to train the Beneficiary network for 5 epochs. We use a batch size of 128 for CIFAR-10 and 16 for the ImageNet subset, and set the mixing threshold t_m to 0.1 in all cases. Following semi-supervised learning, we gradually increase the weight of suspicious sample loss in stage 3, simply setting α to 0.1 in the first 50 epochs and 0.5 in the last 50 epochs.

Evaluation Metrics. We evaluate the performance of defenses with two commonly used metrics: the model's classification accuracy on the benign test set, Benign Accuracy (BA), and the model's classification accuracy on the poison test set, Attack Success Rate (ASR). The lower the ASR and the higher BA, the stronger the defense. In particular, the poison test set removes the samples whose ground-truth label is the target label.

4.2. Effectiveness of Our V&B Defense

To measure the effectiveness of our proposed V&B method, we compare its performance with four state-of-the-art defenses on CIFAR-10 and the ImageNet subset. The backdoor attack CL-16 and CL-32 are two versions of the Clean-label attack with different magnitudes of adversarial perturbations.

Table 1 demonstrates the defense results on CIFAR-10, where *No Defense* is a baseline representing the capability of backdoor attacks. Considering real scenarios, we adopt common data augmentations (*i.e.* random crop and horizontal flipping) during training, which will slightly affect the attack success rate. It is obvious that our V&B achieves the best results against almost all attacks, even compared to defenses that need additional benign samples. Our method reduces the success rate of all attacks from above 95% (64.79% for SIG) to blow or close to 1%, while maintaining the accuracy on benign samples with only about 1% reduction. For WaNet and CL-16, V&B achieves the second-best performance in reducing the attack success rate, but the benign accuracy of V&B is significantly superior to the corresponding best defense, *i.e.* DBD and ABL. WaNet adopts a warping field as the trigger and adds a noisy training mode to interfere with detection. CL-16 only poisons the samples whose ground-truth label is the target label and adds adversarial perturbations to them. So it is relatively hard to distinguish the poisoned samples they craft from the benign ones.

The defense performances on the ImageNet subset are shown in Table 2. Compared with other defenses, our method has the best accuracy on benign samples against all four attacks, even outperforming the baseline due to our AttentionMix data augmentation and semi-supervised suppression training. Semi-supervised learning can relabel poisoned samples to their ground-truth labels, improving benign accuracy. In terms of reducing the attack success rate, our method achieves the best results against BadNets, Blend, and SIG, and the second-best result against WaNet, which is comparable to the best result (1.92% vs. 1.00%).

We also validate the effectiveness of our method under different poisoning rates in Appendix A.4. Even with a poisoning rate of 50%, our method can still reduce the attack success rate of most attacks to below 1%, while maintaining a satisfactory benign accuracy.

4.3. Effectiveness of AttentionMix

To illustrate the inhibition effect of our AttentionMix, we replace it with three other popular strong data augmentations in our framework, *i.e.* Mixup, Cutmix and Attentive Cutmix. As Table 3 shows, each data augmentation performs well against BadNets and Blend, while for more stealthy attacks, such as WaNet and Dynamic, our AttentionMix can better reduce the attack success rate. This is because most of poisoned samples crafted by BadNets and Blend can be filtered out before training the Beneficiary network due to the simple trigger patterns. But for stealthy attacks, the Victim network will only separate part of the poisoned samples after warming up, and the other three augmentations can not effectively inhibit the missed poisoned samples. For a poisoned sample, the poisoned network will focus on its trigger location, so blending the sample's activation region with other images can effectively destroy triggers and help to clean the network. Although Attentive Cutmix can locate the trigger region, direct pasting may leave the trigger intact in the generated image.

4.4. Ablation Studies

Component Effects. As shown in Table 4, We first train the network without any defense as a baseline and then add our components one by one to investigate their effect. The variants are as follows: (1) + F: filtering the training set and training on credible samples; (2) + F + AT: using AttentionMix when training with credible samples; (3) + F +

Attack	BadNets [10]		Blend [7]		WaNet [22]		Dynamic [23]		CL-16 [26]	
Metric	BA	ASR	BA	ASR	BA	ASR	BA	ASR	BA	ASR
No Defense	94.36%	100.00%	94.57%	99.76%	94.21%	95.84%	94.66%	95.18%	94.98%	96.93%
+ F	93.50%	10.01%	93.47%	76.70%	92.92%	90.97%	93.80%	93.04%	93.63%	2.70%
+ F + AT	92.27%	**0.42%**	92.72%	1.42%	91.58%	25.16%	93.08%	50.01%	93.09%	2.09%
+ F + AT + SST	93.88%	1.21%	**94.56%**	1.52%	93.31%	1.23%	94.32%	28.79%	**94.43%**	3.29%
+ F + AT + SST + S (suspicious)	**94.11%**	1.00%	94.49%	0.73%	93.49%	0.61%	**94.61%**	1.43%	94.00%	1.06%
ALL	93.96%	0.62%	94.37%	**0.63%**	**94.15%**	**0.54%**	93.91%	**1.13%**	94.24%	**1.01%**

Table 4. Ablation study of individual components in our framework on CIFAR-10.

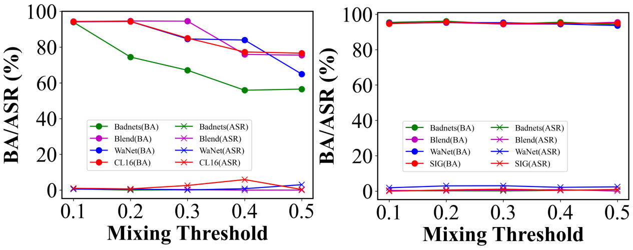


Figure 3. Ablation study of mixing threshold t_m on CIFAR-10 (left) and ImageNet subset (right).

AT + SST: adding standard semi-supervised training [2] on the basis of the second variant; (4) + F + AT + SST + S (suspicious): suppressing suspicious samples during semi-supervised training on the basis of the third variant.

As can be seen from Table 4, when defending against Blend, WaNet, and Dynamic, only filtering will fail because enough poisoned samples are missed to create backdoors. After using AttentionMix, although the attack success rate of most attacks is severely decreased, the benign accuracy also drops slightly. Thus semi-supervised training is adopted to improve the performance of the network on benign samples. For WaNet and Dynamic, semi-supervised training also reduces their attack success rate by relabeling poisoned samples as ground-truth labels. However, the Beneficiary network is not completely clean, so a small number of poisoned samples may be relabeled as the target label, causing the attack success rate to rebound (*e.g.* BadNets, Blend and Cl-16). Suppression to suspicious samples can make up for the weakness of relabeling and adding suppression to credible samples will further reduce the attack success rate. We say that the complete framework is more secure when facing backdoor attacks, although its benign accuracy may be lower than other variants.

To further demonstrate the effect of stage 3, we compare the final results with those after stage 2 in Appendix A.5. In all cases, the attack success rate after stage 2 is higher than the final result, while the benign accuracy after stage 2 is lower than the final result, suggesting that our semi-supervised suppression training can both erasing existing backdoors and improve model performance.

Selection of Mixing Threshold. As Figure 3 shows, we set the mixing threshold t_m from 0.1 to 0.5 on CIFAR-10 and ImageNet subset to study its impact on model performance. It can be seen that the benign accuracy on CIFAR-10 starts to drop when the mixing threshold is large enough (*e.g.* 0.2 for BadNets and 0.3 for Blend), while the attack success rate only slightly fluctuates. The decline in benign accuracy is mainly caused by the mislabeled samples in stage 3. Due to the small image size (32 × 32 in CIFAR-10), the activation region will contain only a few clustered pixels or some separate pixels if the mixing threshold is too large. During mixing, these pixels will carry their labels but can not provide the feature of corresponding classes, which is equivalent to introducing noise to labels. It hinders the learning of our Beneficiary network, leading to more and more suspicious samples being mislabeled in stage 3. This point can be confirmed on the ImageNet subset (image size is 224 × 224), where the mixing threshold has little impact on our method.

5. Conclusion

In this paper, we show that poisoned samples could be easily distinguished by the entropy of the poisoned network's prediction. Based on this finding and the fact that strong data augmentation can inhibit backdoor injection, we propose a novel secure training framework consisting of two parallel networks to prevent a backdoor generation from three aspects: (1) filtering training data; (2) inhibiting backdoor injection; (3) erasing potential backdoors. Specifically, we sacrifice one network to exclusively learn backdoor triggers and filter credible samples according to the prediction entropy to train another network. During training, we design a strong data augmentation to inhibit missed poisoned samples from working and then introduced a semi-supervised learning method to erase potential backdoors with the help of the sacrifice network. Extensive experiments verify that our framework is effective against various backdoor attacks under different poisoning rates.

Acknowledgment. This work is supported in part by the National Natural Science Foundation of China Under Grants No.62176253 and No.U20B2066.

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*, pages 101–105. IEEE, 2019.

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. MixMatch: A Holistic Approach to Semi-Supervised Learning. *Advances in neural information processing systems*, 32, 2019.

[3] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3855–3859. IEEE, 2021.

[4] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering. *arXiv preprint arXiv:1811.03728*, 2018.

[5] Huili Chen, Cheng Fu, Jishen Zhao, and Farinaz Koushanfar. ProFlip: Targeted Trojan Attack With Progressive Bit Flips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7718–7727, 2021.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[7] 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.

[8] Yanbei Chen, Massimiliano Mancini, Xiatian Zhu, and Zeynep Akata. Semi-Supervised and Unsupervised Deep Visual Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[9] Bao Gia Doan, Ehsan Abbasnejad, and Damith C Ranasinghe. Februus: Input Purification Defense Against Trojan Attacks on Deep Neural Network Systems. In *Annual Computer Security Applications Conference*, pages 897–912, 2020.

[10] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. *arXiv preprint arXiv:1708.06733*, 2017.

[11] Chuan Guo, Ruihan Wu, and Kilian Q Weinberger. Trojannet: Embedding Hidden Trojan Horse Models in Neural Networks. *arXiv preprint arXiv:2002.10078*, 2, 2020.

[12] Jonathan Hayase, Weihao Kong, Raghav Somani, and Sewoong Oh. SPECTRE: Defending Against Backdoor Attacks Using Robust Statistics. *arXiv preprint arXiv:2104.11315*, 2021.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[14] Kunzhe Huang, Yiming Li, Baoyuan Wu, Zhan Qin, and Kui Ren. Backdoor Defense via Decoupling the Training Process. In *The Tenth International Conference on Learning Representations*, 2022.

[15] Junnan Li, Richard Socher, and Steven CH Hoi. DIVIDEMIX: LEARNING WITH NOISY LABELS AS SEMI-SUPERVISED LEARNING. *arXiv preprint arXiv:2002.07394*, 2020.

[16] Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible Backdoor Attack With Sample-Specific Triggers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16463–16472, 2021.

[17] 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.

[18] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural Attention Distillation: Erasing Backdoor Triggers from Deep Neural Networks. In *International Conference on Learning Representations*, 2021.

[19] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-Pruning: Defending Against Backdooring Attacks on Deep Neural Networks. In *International Symposium on Research in Attacks, Intrusions, and Defenses*, pages 273–294. Springer, 2018.

[20] Xuankai Liu, Fengting Li, Bihan Wen, and Qi Li. Removing Backdoor-Based Watermarks in Neural Networks with Limited Data. In *International Conference on Pattern Recognition*, pages 10149–10156. IEEE, 2021.

[21] 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.

[22] Anh Nguyen and Anh Tran. WaNet–Imperceptible Warping-based Backdoor Attack. *arXiv preprint arXiv:2102.10369*, 2021.

[23] Tuan Anh Nguyen and Anh Tran. Input-Aware Dynamic Backdoor Attack. *Advances in Neural Information Processing Systems*, 33:3454–3464, 2020.

[24] Xiangyu Qi, Tinghao Xie, Ruizhe Pan, Jifeng Zhu, Yong Yang, and Kai Bu. Towards Practical Deployment-Stage Backdoor Attack on Deep Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13347–13357, 2022.

[25] Han Qiu, Yi Zeng, Shangwei Guo, Tianwei Zhang, Meikang Qiu, and Bhavani Thuraisingham. DeepSweep: An Evaluation Framework for Mitigating DNN Backdoor Attacks using Data Augmentation. In *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*, pages 363–377, 2021.

[26] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-Consistent Backdoor Attacks. *arXiv preprint arXiv:1912.02771*, 2019.

[27] Devesh Walawalkar, Zhiqiang Shen, Zechun Liu, and Marios Savvides. Attentive CutMix: An Enhanced Data Augmentation Approach for Deep Learning Based Image Classification. *arXiv preprint arXiv:2003.13048*, 2020.

[28] 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*, pages 707–723. IEEE, 2019.

[29] Jason Weston, Frédéric Ratle, and Ronan Collobert. Deep Learning via Semi-Supervised Embedding. In *Proceedings of the 25th international conference on Machine learning*, pages 1168–1175, 2008.

[30] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019.

[31] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. Mixup: Beyond Empirical Risk Minimization. In *6th International Conference on Learning Representations*, 2018.

[32] Pu Zhao, Pin-Yu Chen, Payel Das, Karthikeyan Natesan Ramamurthy, and Xue Lin. Bridging Mode Connectivity in Loss Landscapes and Adversarial Robustness. 2020.

[33] Zhendong Zhao, Xiaojun Chen, Yuexin Xuan, Ye Dong, Dakui Wang, and Kaitai Liang. DEFEAT: Deep Hidden Feature Backdoor Attacks by Imperceptible Perturbation and Latent Representation Constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15213–15222, 2022.

[34] Nan Zhong, Zhenxing Qian, and Xinpeng Zhang. Imperceptible Backdoor Attack: From Input Space to Feature Representation. *arXiv preprint arXiv:2205.03190*, 2022.

[35] Dengyong Zhou, Olivier Bousquet, Thomas Lal, Jason Weston, and Bernhard Schölkopf. Learning with Local and Global Consistency. *Advances in neural information processing systems*, 16, 2003.