

BACKDOOR DEFENSE VIA DECOUPLING THE TRAINING PROCESS

Kunzhe Huang^{1,*}, Yiming Li^{3,*}, Baoyuan Wu^{2,†}, Zhan Qin^{1,†}, Kui Ren¹

¹School of Cyber Science and Technology, Zhejiang University

²School of Data Science, Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen

³Tsinghua Shenzhen International Graduate School, Tsinghua University

{hkunzhe, zhanqin, kuiren}@zju.edu.cn; wubaoyuan@cuhk.edu.cn; li-ym18@mails.tsinghua.edu.cn

ABSTRACT

Recent studies have revealed that deep neural networks (DNNs) are vulnerable to backdoor attacks, where attackers embed hidden backdoors in the DNN model by poisoning a few training samples. The attacked model behaves normally on benign samples, whereas its prediction will be maliciously changed when the backdoor is activated. We reveal that poisoned samples tend to cluster together in the feature space of the attacked DNN model, which is mostly due to the end-to-end supervised training paradigm. Inspired by this observation, we propose a novel backdoor defense via decoupling the original end-to-end training process into three stages. Specifically, we first learn the backbone of a DNN model via *self-supervised learning* based on training samples without their labels. The learned backbone will map samples with the same ground-truth label to similar locations in the feature space. Then, we freeze the parameters of the learned backbone and train the remaining fully connected layers via standard training with all (labeled) training samples. Lastly, to further alleviate side-effects of poisoned samples in the second stage, we remove labels of some ‘low-credible’ samples determined based on the learned model and conduct a *semi-supervised fine-tuning* of the whole model. Extensive experiments on multiple benchmark datasets and DNN models verify that the proposed defense is effective in reducing backdoor threats while preserving high accuracy in predicting benign samples. Our code is available at <https://github.com/SCLBD/DBD>.

1 INTRODUCTION

Deep learning, especially deep neural networks (DNNs), has been widely adopted in many realms (Wang et al., 2020b; Li et al., 2020a; Wen et al., 2020) for its high effectiveness. In general, the training of DNNs requires a large amount of training samples and computational resources. Accordingly, third-party resources (*e.g.*, third-party data or servers) are usually involved. While the opacity of the training process brings certain convenience, it also introduces new security threats.

Backdoor attack poses a new security threat to the training process of DNNs (Li et al., 2020c). It maliciously manipulates the prediction of the attacked DNNs by poisoning a few training samples. Specifically, backdoor attackers inject the *backdoor trigger* (*i.e.*, a particular pattern) to some benign training images and change their labels with the attacker-specified *target label*. The connection between the backdoor trigger and the target label will be learned by DNNs during the training process. In the inference process, the prediction of attacked DNNs will be changed to the target label when the trigger is present, whereas the attacked DNNs will behave normally on benign samples. As such, users are difficult to realize the existence of hidden backdoors and therefore this attack is a serious threat to the practical applications of DNNs.

In this paper, we first investigate backdoor attacks from the hidden feature space. Our preliminary experiments reveal that the backdoor is embedded in the feature space, *i.e.*, **samples with the back-**

*The first two authors contributed equally to this work. This work was mostly done when Kunzhe Huang and Yiming Li were the research interns at The Chinese University of Hong Kong, Shenzhen. † indicates corresponding authors: Baoyuan Wu (wubaoyuan@cuhk.edu.cn) and Zhan Qin (qinzhan@zju.edu.cn).

door trigger (dubbed *poisoned samples*) tend to cluster together in the feature space. We reveal that *this phenomenon is mostly due to the end-to-end supervised training paradigm*. Specifically, the excessive learning capability allows DNNs to learn features about the backdoor trigger, while the DNNs can shrink the distance between poisoned samples in the feature space and connect the learned trigger-related features with the target label by the end-to-end supervised training. Based on this understanding, we propose to decouple the end-to-end training process for the backdoor defense. Specifically, we treat the DNNs as two disjoint parts, including a *feature extractor (i.e., backbone)* and a *simple classifier (i.e., the remaining fully connected layers)*. We first learn the *purified feature extractor* via *self-supervised learning* (Kolesnikov et al., 2019; Chen et al., 2020a; Jing & Tian, 2020) with unlabeled training samples (obtained by removing their labels), and then learn the simple classifier via standard supervised training process based on the learned feature extractor and all training samples. The strong data augmentations involved in the self-supervised learning damage trigger patterns, making them unlearnable during representation learning; and the decoupling process further disconnects trigger patterns and the target label. Accordingly, hidden backdoors cannot be successfully created even the model is trained on the poisoned dataset based on our defense.

Moreover, we further reveal that the representation of poisoned samples generated by the purified extractor is significantly different from those generated by the extractor learned with standard training process. Specifically, the poisoned sample lies closely to samples with its ground-truth label instead of the target label. This phenomenon makes the training of the simple classifier similar to *label-noise learning* (Wang et al., 2019b; Ma et al., 2020; Berthon et al., 2021). As such, we first filter *high-credible training samples (i.e., training samples that are most probably to be benign)* and then use those samples as labeled samples and the remaining part to form unlabeled samples to fine-tune the whole model via *semi-supervised learning* (Rasmus et al., 2015; Berthelot et al., 2019; Sohn et al., 2020). This approach is to further reduce the adverse effects of poisoned samples.

The main contributions of this paper are three-fold. (1) We reveal that the backdoor is embedded in the feature space, which is mostly due to the end-to-end supervised training paradigm. (2) Based on our understanding, we propose a decoupling-based backdoor defense (DBD) to alleviate the threat of poisoning-based backdoor attacks. (3) Experiments on classical benchmark datasets are conducted, which verify the effectiveness of our defense.

2 RELATED WORK

2.1 BACKDOOR ATTACK

Backdoor attack is an emerging research area, which raises security concerns about training with third-party resources. In this paper, we focus on the poisoning-based backdoor attack towards image classification, where attackers can only modify the dataset instead of other training components (*e.g.*, training loss). This threat could also happen in other tasks (Xiang et al., 2021; Zhai et al., 2021; Li et al., 2022) and with different attacker’s capacities (Nguyen & Tran, 2020; Tang et al., 2020; Zeng et al., 2021a), which are out-of-scope of this paper. In general, existing attacks can be divided into two main categories based on the property of target labels, as follows:

Poison-Label Backdoor Attack. It is currently the most common attack paradigm, where the target label is different from the ground-truth label of poisoned samples. BadNets (Gu et al., 2019) is the first and most representative poison-label attack. Specifically, it randomly selected a few samples from the original benign dataset to generate *poisoned samples* by stamping the *backdoor trigger* onto the (benign) image and change their label with an attacker-specified *target label*. Those generated poisoned samples associated with remaining benign ones were combined to form the *poisoned training dataset*, which will be delivered to users. After that, (Chen et al., 2017) suggested that the poisoned image should be similar to its benign version for the stealthiness, based on which they proposed the *blended attack*. Recently, (Xue et al., 2020; Li et al., 2020b; 2021c) further explored how to conduct poison-label backdoor attacks more stealthily. Most recently, a more stealthy and effective attack, the WaNet (Nguyen & Tran, 2021), was proposed. WaNet adopted image warping as the backdoor trigger, which deforms but preserves the image content.

Clean-Label Backdoor Attack. Although the poisoned image generated by poison-label attacks could be similar to its benign version, users may still notice the attack by examining the image-label relationship. To address this problem, Turner et al. (2019) proposed the *clean-label attack paradigm*, where the target label is consistent with the ground-truth label of poisoned samples. Specifically,

they first leveraged adversarial perturbations or generative models to modify some benign images from the target class and then conducted the standard trigger injection process. This idea was generalized to attack video classification in (Zhao et al., 2020b), where they adopted the targeted universal adversarial perturbation (Moosavi-Dezfooli et al., 2017) as the trigger pattern. Although clean-label backdoor attacks are more stealthy compared with poison-label ones, they usually suffer from relatively poor performance and may even fail in creating backdoors (Li et al., 2020c).

2.2 BACKDOOR DEFENSE

Currently, there are also some approaches to alleviate the backdoor threat. Existing defenses are mostly *empirical*, which can be divided into five main categories, including **(1) detection-based defenses** (Xu et al., 2021; Zeng et al., 2021a; Xiang et al., 2022), **(2) preprocessing based defenses** (Doan et al., 2020; Li et al., 2021b; Zeng et al., 2021b), **(3) model reconstruction based defenses** (Zhao et al., 2020a; Li et al., 2021a; Zeng et al., 2022), **(4) trigger synthesis based defenses** (Guo et al., 2020; Dong et al., 2021; Shen et al., 2021), and **(5) poison suppression based defenses** (Du et al., 2020; Borgnia et al., 2021). Specifically, detection-based defenses examine whether a suspicious DNN or sample is attacked and it will deny the use of malicious objects; Preprocessing based methods intend to damage trigger patterns contained in attack samples to prevent backdoor activation by introducing a preprocessing module before feeding images into DNNs; Model reconstruction based ones aim at removing the hidden backdoor in DNNs by modifying models directly; The fourth type of defenses synthesize potential trigger patterns at first, following by the second stage that the hidden backdoor is eliminated by suppressing their effects; The last type of methods depress the effectiveness of poisoned samples during the training process to prevent the creation of hidden backdoors. In general, our method is most relevant to this type of defenses.

In this paper, we only focus on the last four types of defenses since they directly improve the robustness of DNNs. Besides, there were also few works focusing on *certified backdoor defenses* (Wang et al., 2020a; Weber et al., 2020). Their robustness is theoretically guaranteed under certain assumptions, which cause these methods to be generally weaker than empirical ones in practice.

2.3 SEMI-SUPERVISED AND SELF-SUPERVISED LEARNING

Semi-supervised Learning. In many real-world applications, the acquisition of labeled data often relies on manual labeling, which is very expensive. In contrast, obtaining unlabeled samples is much easier. To utilize the power of unlabeled samples with labeled ones simultaneously, a great amount of semi-supervised learning methods were proposed (Gao et al., 2017; Berthelot et al., 2019; Van Engelen & Hoos, 2020). Recently, semi-supervised learning was also introduced in improving the security of DNNs (Stanforth et al., 2019; Carmon et al., 2019), where they utilized unlabelled samples in the adversarial training. Most recently, (Yan et al., 2021) discussed how to backdoor semi-supervised learning. However, this approach needs to control other training components (*e.g.*, training loss) in addition to modifying training samples and therefore is out-of-scope of this paper. How to adopt semi-supervised learning for backdoor defense remains blank.

Self-supervised Learning. This learning paradigm is a subset of *unsupervised learning*, where DNNs are trained with supervised signals generated from the data itself (Chen et al., 2020a; Grill et al., 2020; Liu et al., 2021). It has been adopted for increasing adversarial robustness (Hendrycks et al., 2019; Wu et al., 2021; Shi et al., 2021). Most recently, there were also a few works (Saha et al., 2021; Carlini & Terzis, 2021; Jia et al., 2021) exploring how to backdoor self-supervised learning. However, these attacks are out-of-scope of this paper since they need to control other training components (*e.g.*, training loss) in addition to modifying training samples.

3 REVISITING BACKDOOR ATTACKS FROM THE HIDDEN FEATURE SPACE

In this section, we analyze the behavior of poisoned samples from the hidden feature space of attacked models and discuss its inherent mechanism.

Settings. We conduct the BadNets (Gu et al., 2019) and label-consistent attack (Turner et al., 2019) on CIFAR-10 dataset (Krizhevsky, 2009) for the discussion. They are representative of poison-label attacks and clean-label attacks, respectively. Specifically, we conduct supervised learning on the poisoned datasets with the standard training process and self-supervised learning on the unlabelled

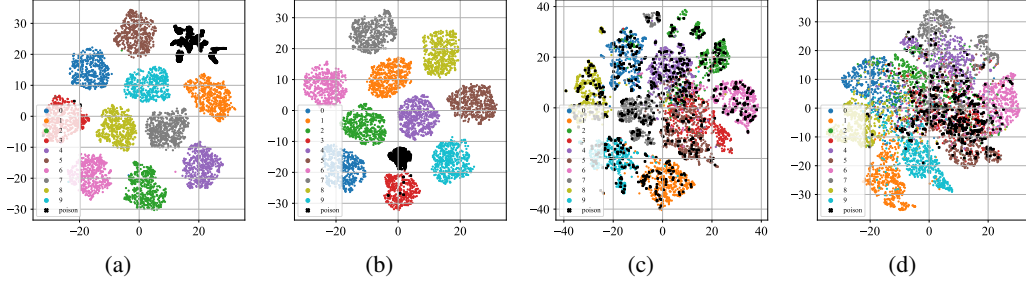


Figure 1: The t-SNE of poisoned samples in the hidden space generated by different models. **(a)-(b)**: DNNs trained with supervised learning. **(c)-(d)**: DNNs trained with self-supervised learning. **(a)&(c)**: DNNs under BadNets attack. **(b)&(d)**: DNNs under label-consistent attack. No matter under the BadNets or label-consistent attack, poisoned samples tend to cluster together in the hidden space generated by DNNs trained with supervised learning, whereas lie closely to samples with their ground-truth label by those trained with self-supervised learning.

poisoned datasets with SimCLR (Chen et al., 2020a). We visualize poisoned samples in the hidden feature space generated by attacked DNNs based on the t-SNE (Van der Maaten & Hinton, 2008). More detailed settings are presented in Appendix A.

Results. As shown in Figure 1(a)-1(b), poisoned samples (denoted by ‘black-cross’) tend to cluster together to form a separate cluster after the standard supervised training process, no matter under the poison-label attack or clean-label attack. This phenomenon implies why existing poisoning-based backdoor attacks can succeed. Specifically, the excessive learning capability allows DNNs to learn features about the backdoor trigger. Associated with the end-to-end supervised training paradigm, DNNs can shrink the distance between poisoned samples in the feature space and connect the learned trigger-related features with the target label. In contrast, as shown in Figure 1(c)-1(d), poisoned samples lie closely to samples with their ground-truth label after the self-supervised training process on the unlabelled poisoned dataset. It indicates that we can prevent the creation of backdoors by self-supervised learning, which will be further introduced in the next section.

4 DECOUPLING-BASED BACKDOOR DEFENSE

4.1 PRELIMINARIES

General Pipeline of Backdoor Attacks. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ denotes the benign training set, where $x_i \in \mathcal{X} = \{0, 1, \dots, 255\}^{C \times W \times H}$ is the image, $y_i \in \mathcal{Y} = \{0, 1, \dots, K\}$ is its label, K is the number of classes, and $y_t \in \mathcal{Y}$ indicates the *target label*. How to generate the poisoned dataset \mathcal{D}_p is the cornerstone of backdoor attacks. Specifically, \mathcal{D}_p consists of two subsets, including the modified version of a subset of \mathcal{D} and remaining benign samples, *i.e.*, $\mathcal{D}_p = \mathcal{D}_m \cup \mathcal{D}_b$, where $\mathcal{D}_b \subset \mathcal{D}$, $\gamma \triangleq \frac{|\mathcal{D}_m|}{|\mathcal{D}|}$ is the *poisoning rate*, $\mathcal{D}_m = \{(x', y_t) | x' = G(x), (x, y) \in \mathcal{D} \setminus \mathcal{D}_b\}$, and $G: \mathcal{X} \rightarrow \mathcal{X}$ is an attacker-predefined poisoned image generator. For example, $G(x) = (1 - \lambda) \otimes x + \lambda \otimes t$, where $\lambda \in [0, 1]^{C \times W \times H}$, $t \in \mathcal{X}$ is the *trigger pattern*, and \otimes is the element-wise product in the blended attack (Chen et al., 2017). Once \mathcal{D}_p is generated, it will be sent to users who will train DNNs on it. Hidden backdoors will be created after the training process.

Threat Model. In this paper, we focus on defending against poisoning-based backdoor attacks. The attacker can arbitrarily modify the training set whereas cannot change other training components (*e.g.*, model structure and training loss). For our proposed defense, we assume that defenders can fully control the training process. This is the scenario that users adopt third-party collected samples for training. Note that we do not assume that defenders have a local benign dataset, which is often required in many existing defenses (Wang et al., 2019a; Zhao et al., 2020a; Li et al., 2021a).

Defender’s Goals. The defender’s goals are to prevent the trained DNN model from predicting poisoned samples as the target label and to preserve the high accuracy on benign samples.

4.2 OVERVIEW OF THE DEFENSE PIPELINE

In this section, we describe the general pipeline of our defense. As shown in Figure 2, it consists of three main stages, including **(1)** learning a purified feature extractor via self-supervised learning, **(2)** filtering high-credible samples via label-noise learning, and **(3)** semi-supervised fine-tuning.

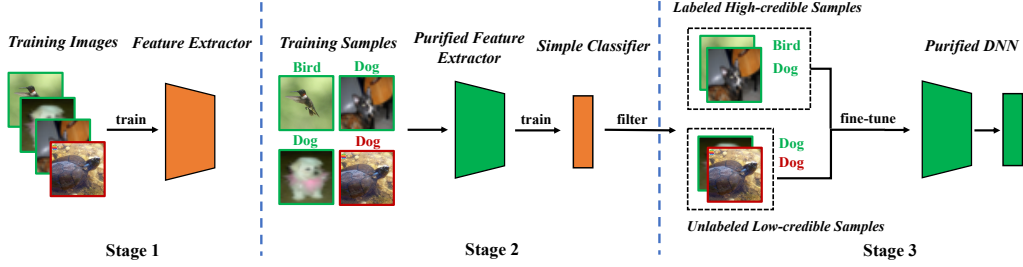


Figure 2: The main pipeline of our defense. In the first stage, we train the whole DNN model via self-supervised learning based on label-removed training samples. In the second stage, we freeze the learned feature extractor and adopt all training samples to train the remaining fully connected layers via supervised learning. After that, we filter high-credible samples based on the training loss. In the third stage, we adopt high-credible samples as labeled samples and remove the labels of all low-credible samples to fine-tune the whole model via semi-supervised learning.

Specifically, in the first stage, we remove the label of all training samples to form the unlabelled dataset, based on which to train the feature extractor via self-supervised learning. In the second stage, we freeze the learned feature extractor and adopt all training samples to train the remaining fully connected layers via supervised learning. We then filter $\alpha\%$ high-credible samples based on the training loss. The smaller the loss, the more credible the sample. After the second stage, the training set will be separated into two disjoint parts, including high-credible samples and low-credible samples. We use high-credible samples as labeled samples and remove the label of all low-credible samples to fine-tune the whole model via semi-supervised learning. More detailed information about each stage of our method will be further illustrated in following sections.

4.3 LEARNING PURIFIED FEATURE EXTRACTOR VIA SELF-SUPERVISED LEARNING

Let \mathcal{D}_t denotes the training set and $f_w : \mathcal{X} \rightarrow [0, 1]^K$ indicates the DNN with parameter $w = [w_c, w_f]$, where w_c and w_f indicates the parameters of the backbone and the fully connected layer, respectively. In this stage, we optimize w_c based on the unlabeled version of \mathcal{D}_t via self-supervised learning, as follows:

$$w_c^* = \arg \min_{w_c} \sum_{(x, y) \in \mathcal{D}_t} \mathcal{L}_1(x; w_c), \quad (1)$$

where $\mathcal{L}_1(\cdot)$ indicates the self-supervised loss (e.g., NT-Xent in SimCLR (Chen et al., 2020a)). Through the self-supervised learning, the learned feature extractor (i.e., backbone) will be purified even if the training set contains poisoned samples, as illustrated in Section 3.

4.4 FILTERING HIGH-CREDIBLE SAMPLES VIA LABEL-NOISE LEARNING

Once w_c^* is obtained, the user can freeze it and adopt \mathcal{D}_t to further optimize remaining w_f , i.e.,

$$w_f^* = \arg \min_{w_f} \sum_{(x, y) \in \mathcal{D}_t} \mathcal{L}_2(f_{[w_c^*, w_f]}(x), y), \quad (2)$$

where $\mathcal{L}_2(\cdot)$ indicates the supervised loss (e.g., cross entropy).

After the decoupling-based training process (1)-(2), even if the model is (partly) trained on the poisoned dataset, the hidden backdoor cannot be created since the feature extractor is purified. However, this simple strategy suffers from two main problems. Firstly, compared with the one trained via supervised learning, the accuracy of predicting benign samples will have a certain decrease, since the learned feature extractor is frozen in the second stage. Secondly, poisoned samples will serve as ‘outliers’ to further hinder the learning of the second stage when poison-label attacks appear, since those samples lie close to samples with its ground-truth label instead of the target label in the hidden feature space generated by the learned purified feature extractor. These two problems indicate that *we should remove poisoned samples and retrain or fine-tune the whole model*.

Specifically, we select *high-credible samples* \mathcal{D}_h based on the loss $\mathcal{L}_2(\cdot; [w_c^*, w_f^*])$. The high-credible samples are defined as the $\alpha\%$ training samples with the smallest loss, where $\alpha \in [0, 100]$ is

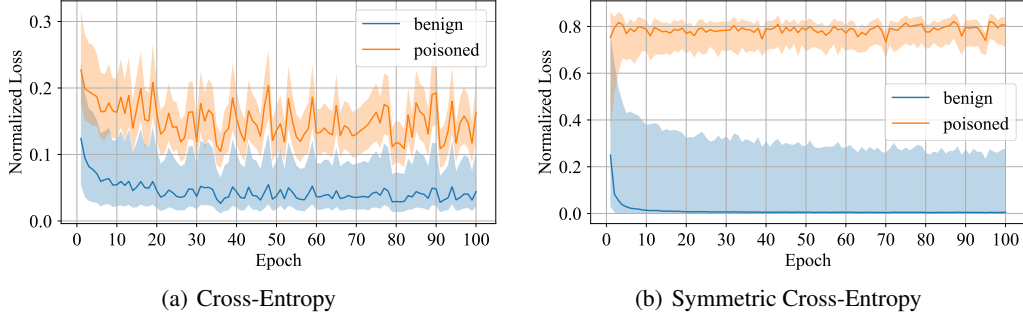


Figure 3: Loss values of models under BadNets attack with 20% poisoning rate trained on CIFAR-10 dataset with the symmetric cross-entropy (SCE) and cross-entropy (CE) in the second stage. All loss values are normalized to $[0, 1]$. As shown in the figure, adopting SCE can significantly increase the loss differences between poisoned samples and benign ones compared with the CE.

a hyper-parameter. In particular, we adopt the symmetric cross-entropy (SCE) (Wang et al., 2019b) as $\mathcal{L}_2(\cdot)$, inspired by the label-noise learning. As shown in Figure 3, compared with the CE loss, the SCE can significantly increase the differences between poisoned samples and benign ones, which further reduces the possibility that high-credible dataset \mathcal{D}_h still contains poisoned samples.

Note that *we do not intend to accurately separate poisoned samples and benign samples*. We only want to ensure that the obtained \mathcal{D}_h contains as few poisoned samples as possible.

4.5 SEMI-SUPERVISED FINE-TUNING

After the second stage, the third-party training set \mathcal{D}_t will be separated into two disjoint parts, including the high-credible dataset \mathcal{D}_h and the low-credible dataset $\mathcal{D}_l \triangleq \mathcal{D}_t \setminus \mathcal{D}_h$. Let $\hat{\mathcal{D}}_l \triangleq \{\mathbf{x} | (\mathbf{x}, y) \in \mathcal{D}_l\}$ indicates the unlabeled version of low-credible dataset \mathcal{D}_l . We fine-tune the whole trained model $f_{[\mathbf{w}_c^*, \mathbf{w}_f^*]}(\cdot)$ with semi-supervised learning as follows:

$$\min_{\mathbf{w}} \mathcal{L}_3(\mathcal{D}_h, \hat{\mathcal{D}}_l; \mathbf{w}), \quad (3)$$

where $\mathcal{L}_3(\cdot)$ denotes the semi-supervised loss (*e.g.*, the loss in MixMatch (Berthelot et al., 2019)).

This process can prevent the side-effects of poisoned samples while utilizing their contained useful information, and encourage the compatibility between the feature extractor and the simple classifier via learning them jointly instead of separately. Please refer to Section 5.3 for more results.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETTINGS

Datasets and DNNs. We evaluate all defenses on two classical benchmark datasets, including CIFAR-10 (Krizhevsky, 2009) and (a subset of) ImageNet (Deng et al., 2009). We adopt the ResNet-18 (He et al., 2016) for these tasks. More detailed settings are presented in Appendix B.1. Besides, we also provide the results on (a subset of) VGGFace2 (Cao et al., 2018) in Appendix C.

Attack Baselines. We examine all defense approaches in defending against four representative attacks. Specifically, we select the BadNets (Gu et al., 2019), the backdoor attack with blended strategy (dubbed ‘Blended’) (Chen et al., 2017), WaNet (Nguyen & Tran, 2021), and label-consistent attack with adversarial perturbations (dubbed ‘Label-Consistent’) (Turner et al., 2019) for the evaluation. They are the representative of patch-based visible and invisible poison-label attacks, non-patch-based poison-label attacks, and clean-label attacks, respectively.

Defense Baselines. We compared our DBD with two defenses having the same defender’s capacities, including the DPSGD (Du et al., 2020) and ShrinkPad (Li et al., 2021b). We also compare with other two approaches with an additional requirement (*i.e.*, having a local benign dataset), including

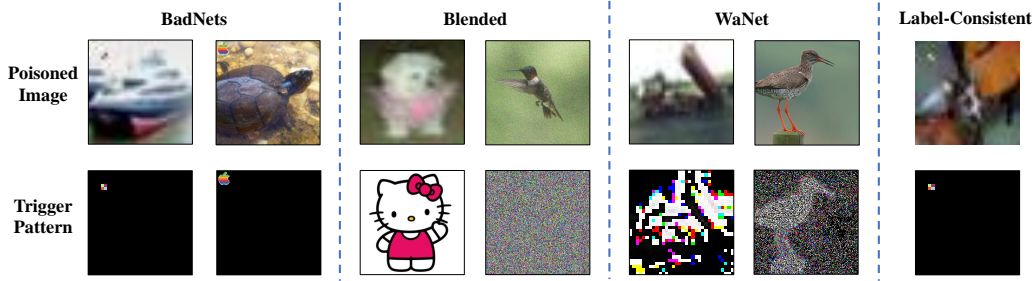


Figure 4: The illustration of poisoned samples generated by different attacks.

the neural cleanse with unlearning strategy (dubbed ‘NC’) (Wang et al., 2019a), and neural attention distillation (dubbed ‘NAD’) (Li et al., 2021a). They are the representative of poison suppression based defenses, preprocessing based defenses, trigger synthesis based defenses, and model reconstruction based defenses, respectively. We also provide results of DNNs trained without any defense (dubbed ‘No Defense’) as another important baseline for reference.

Attack Setups. We use a 2×2 square as the trigger pattern on CIFAR-10 dataset and the 32×32 Apple logo on ImageNet dataset for the BadNets, as suggested in (Gu et al., 2019; Wang et al., 2019a). For Blended, we adopt the ‘Hello Kitty’ pattern on CIFAR-10 and the random noise pattern on ImageNet, based on the suggestions in (Chen et al., 2017), and set the blended ratio $\lambda = 0.1$ on all datasets. The trigger pattern adopted in label-consistent attack is the same as the one used in BadNets. For WaNet, we adopt its default settings on CIFAR-10 dataset. However, on ImageNet dataset, we use different settings optimized by grid-search since the original ones fail. An example of poisoned samples generated by different attacks is shown in Figure 4. Besides, we set the poisoning rate $\gamma_1 = 2.5\%$ for label-consistent attack (25% of training samples with the target label) and $\gamma_2 = 5\%$ for three other attacks. More details are shown in Appendix B.2.

Defense Setups. For our DBD, we adopt SimCLR (Chen et al., 2020a) as the self-supervised method and MixMatch (Berthelot et al., 2019) as the semi-supervised method. More details about SimCLR and MixMatch are in Appendix I. The filtering rate α is the only key hyper-parameter in DBD, which is set to 50% in all cases. We set the shrinking rate to 10% for the ShrinkPad on all datasets, as suggested in (Li et al., 2021b; Zeng et al., 2021b). In particular, DPSGD and NAD are sensitive to their hyper-parameters. We report their best results in each case based on the grid-search (as shown in Appendix D). Besides, we split a 5% random subset of the benign training set as the local benign dataset for NC and NAD. More implementation details are provided in Appendix B.3.

Evaluation Metrics. We adopt the *attack success rate* (ASR) and *benign accuracy* (BA) to measure the effectiveness of all methods¹. Specifically, let \mathcal{D}_{test} indicates the (benign) testing set and $C_w : \mathcal{X} \rightarrow \mathcal{Y}$ denotes the trained classifier, we have $ASR \triangleq \Pr_{(\mathbf{x}, y) \in \mathcal{D}_{test}} \{C_w(G(\mathbf{x})) = y_t | y \neq y_t\}$ and $BA \triangleq \Pr_{(\mathbf{x}, y) \in \mathcal{D}_{test}} \{C_w(\mathbf{x}) = y\}$, where y_t is the target label and $G(\cdot)$ is the poisoned image generator. In particular, *the lower the ASR and the higher the BA, the better the defense*.

5.2 MAIN RESULTS

Comparing DBD with Defenses having the Same Requirements. As shown in Table 1-2, DBD is significantly better than defenses having the same requirements (*i.e.*, DPSGD and ShrinkPad) in defending against all attacks. For example, the benign accuracy of DBD is 20% over while the attack success rate is 5% less than that of DPSGD in all cases. Specifically, the attack success rate of models with DBD is less than 2% in all cases (mostly $< 0.5\%$), which verifies that our method can successfully prevent the creation of hidden backdoors. Moreover, the decreases of benign accuracy are less than 2% when defending against poison-label attacks, compared with models trained without any defense. Our method is even better on relatively larger dataset where all baseline methods become less effective. These results verify the effectiveness of our method.

¹Among all defense methods, the one with the best performance is indicated in boldface and the value with underline denotes the second-best result.

Table 1: The effectiveness (%) of defending against three attacks. Note that NC and NAD need an additional local benign dataset, which is not required in DPSGD, ShrinkPad, and our method.

Dataset →	CIFAR-10						ImageNet					
Attack →	BadNets		Blended		WaNet		BadNets		Blended		WaNet	
Defense ↓	BA	ASR	BA	ASR	BA	ASR	BA	ASR	BA	ASR	BA	ASR
No Defense	94.92	100	94.14	98.25	94.29	98.64	80.23	90.49	82.07	95.57	80.98	96.22
NC	94.25	2.33	93.01	5.70	92.44	98.52	<u>80.67</u>	32.22	82.07	27.33	<u>80.32</u>	99.32
NAD	90.47	1.20	89.72	1.51	<u>91.83</u>	9.49	73.49	<u>5.81</u>	70.88	8.15	75.27	31.43
DPSGD	44.24	14.56	53.05	20.68	52.13	39.78	51.43	25.40	55.12	84.17	20.72	54.12
ShrinkPad	90.32	10.34	79.67	55.71	74.76	85.59	66.08	45.61	64.47	4.20	63.74	<u>3.58</u>
DBD (ours)	<u>92.41</u>	0.96	<u>92.18</u>	<u>1.73</u>	91.20	0.39	80.99	0.26	<u>81.63</u>	0.16	82.05	0.33

Table 2: The effectiveness (%) of defending against label-consistent attack on CIFAR-10 dataset.

Attack ↓, Defense →	Metric ↓	Benign	NC	NAD	DPSGD	ShrinkPad	DBD (ours)
Label-Consistent ($\epsilon = 16$)	BA	94.90	94.94	93.04	44.28	<u>90.67</u>	89.67
	ASR	99.33	1.31	3.95	7.97	9.60	0.01
Label-Consistent ($\epsilon = 32$)	BA	94.82	94.57	90.98	69.79	90.83	<u>91.45</u>
	ASR	99.19	<u>1.36</u>	6.54	9.06	13.04	0.34

Table 3: The ablation study of our proposed method.

Attack →	BadNets		Blended		Label-Consistent		WaNet	
Defense ↓, Metric →	BA	ASR	BA	ASR	BA	ASR	BA	ASR
No Defense	94.92	100	94.14	98.25	94.82	99.19	94.29	98.64
DBD without SS	93.66	100	93.47	99.93	90.70	98.50	81.91	98.40
SS with CE	82.25	5.20	81.85	12.19	82.08	5.87	80.29	9.48
SS with SCE	82.34	5.12	82.30	6.24	81.81	5.43	81.15	7.08
SS with SCE + Tuning	78.94	<u>4.02</u>	78.78	<u>3.70</u>	77.44	<u>1.44</u>	78.51	<u>5.87</u>
DBD (ours)	<u>92.41</u>	0.96	<u>92.18</u>	1.73	91.45	0.34	91.20	0.39

¹DBD without SS: use standard supervised training to obtain the feature extractor adopted in DBD.

²SS with CE: freeze the learned backbone and train remaining FC layers on all samples with CE loss.

³SS with SCE: freeze the learned backbone and train remaining FC layers on all samples with SCE loss.

⁴SS with SCE + Tuning: fine-tune FC layers of the model in ‘SS with SCE’ on high-credible samples.

Comparing DBD with Defenses having Extra Requirements. We also compare our defense with two other methods (*i.e.*, NC and NAD), which have an additional requirement that defenders have a benign local dataset. As shown in Table 1-2, NC and NAD are better than DPSGD and ShrinkPad, as we expected, since they adopt additional information from the benign local dataset. In particular, although NAD and NC use additional information, our method is still better than them, even when their performances are tuned to the best while our method only uses the default settings. Specifically, the BA of NC is on par with that of our method. However, it is with the sacrifice of ASR. Especially on ImageNet dataset, NC has limited effects in reducing ASR. In contrast, our method reaches the smallest ASR while its BA is either the highest or the second-highest in almost all cases. These results verify the effectiveness of our method again.

Results. As shown in Figure 7, our method can still prevent the creation of hidden backdoors even when the poisoning rate reaches 20%. Besides, DBD also maintains high benign accuracy. In other words, our method is effective in defending attacks with different strengths.

5.3 ABLATION STUDY

There are four key strategies in DBD, including (1) obtaining purified feature extractor, (2) using SCE instead of CE in the second stage, (3) reducing side-effects of low-credible samples, and (4) fine-tuning the whole model via semi-supervised learning. Here we verify their effectiveness.

Settings. We compare the proposed DBD with its four variants, including (1) DBD without SS, (2) SS with CE, (3) SS with SCE, and (4) SS with SCE + Tuning, on the CIFAR-10 dataset. Specifically, in the first variant, we replace the backbone generated by self-supervised learning with the one trained in a supervised fashion and keep other parts unchanged. In the second variant, we freeze the backbone learned via self-supervised learning and train the remaining fully-connected layers with cross-entropy loss on all training samples. The third variant is similar to the second one. The only difference is that it uses symmetric cross-entropy instead of cross-entropy to train fully-connected layers. The last variant is an advanced version of the third one, which further fine-tunes fully-connected layers on high-credible samples filtered by the third variant.

Results. As shown in Table 3, we can conclude that decoupling the original end-to-end supervised training process is effective in preventing the creation of hidden backdoors, by comparing our DBD with its first variant and the model trained without any defense. Besides, we can also verify the effectiveness of SCE loss on defending against poison-label backdoor attacks by comparing the second and third DBD variants. Moreover, the fourth DBD variant has relatively lower ASR and BA, compared with the third one. This phenomenon is due to the removal of low-credible samples. It indicates that reducing side-effects of low-credible samples while adopting their useful information is important for the defense. We can also verify that fine-tuning the whole model via semi-supervised learning is also useful by comparing the fourth variant and the proposed DBD.

5.4 RESISTANCE TO POTENTIAL ADAPTIVE ATTACKS

In our paper, we adopted the classical defense setting that attackers have no information about the defense. Attackers may design adaptive attacks if they know the existence of our DBD. The most straightforward idea is to *manipulate the self-supervised training process* so that poisoned samples are still in a new cluster after the self-supervised learning. However, *attackers are not allowed to do it* based on our threat model about adopting third-party datasets. Despite this, attackers may design adaptive attacks by optimizing the trigger pattern to make poisoned samples still in a new cluster after the self-supervised learning if they can know the model structure used by defenders, as follows:

Problem Formulation. For a K -classification problem, let $\mathcal{X}' = \{\mathbf{x}_i\}_{i=1}^M$ indicates the benign images selected for poisoning, $\mathcal{X}_j = \{\mathbf{x}_i\}_{i=1}^{N_j}$ denotes the benign images with ground-truth label j , and g is a trained backbone. Given an attacker-predefined poisoned image generator G , the *adaptive attack* aims to optimize a trigger pattern \mathbf{t} by minimizing the distance between poisoned images while maximizing the distance between the center of poisoned images and centers of clusters of benign images with different label, *i.e.*,

$$\min_{\mathbf{t}} \frac{1}{M} \sum_{\mathbf{x} \in \mathcal{X}'} d(g(G(\mathbf{x}; \mathbf{t})), \bar{\mathbf{g}}') - \frac{1}{K} \sum_{i=1}^K d(\bar{\mathbf{g}}', \bar{\mathbf{g}}_i), \quad (4)$$

where $\bar{\mathbf{g}}' \triangleq \frac{1}{M} \sum_{\mathbf{x} \in \mathcal{X}'} g(G(\mathbf{x}; \mathbf{t}))$, $\bar{\mathbf{g}}_i \triangleq \frac{1}{N_i} \sum_{\mathbf{x} \in \mathcal{X}_i} g(\mathbf{x})$, and d is a distance metric.

Settings. We adopt the CIFAR-10 dataset and use the ℓ^2 norm as the distance metric to conduct the experiment. Specifically, we assume that attackers have the entire benign dataset, based on which they can train a backbone adopted in the first stage of our DBD. We use the Adam optimizer to solve the above optimization problem for 100 epochs with a learning rate of 0.1. The trigger size is set to 32×32 , which means the attacker can completely modify the content of poisoned samples, regardless of its original semantic information and the stealthiness of the attack. This setting is to ensure the attack ability, since clustering poisoned samples together is very difficult in self-supervised learning.

Results. The adaptive attack works well when there is no defense (BA=94.96%, ASR=99.70%). However, this attack still fails to attack our DBD (BA=93.21%, ASR=1.02%). In other words, our defense is resistant to this adaptive attack. It is most probably because the trigger optimized based on the backbone is far less effective when the model is retrained since model parameters are changed due to the random initialization and the update of model weights during the training process.

6 CONCLUSION

The mechanism of poisoning-based backdoor attacks is to establish a latent connection between trigger patterns and the target label during the training process. In this paper, we revealed that this connection is learned mostly due to the end-to-end supervised training paradigm. Motivated by this understanding, we proposed a decoupling-based backdoor defense, which first learns the backbone via self-supervised learning and then the remaining fully-connected layers by the classical supervised learning. We also introduced the label-noise learning method to determine high-credible and low-credible samples, based on which we fine-tuned the whole model via semi-supervised learning. Extensive experiments verify that our defense is effective on reducing backdoor threats while preserving high accuracy on predicting benign samples.

ACKNOWLEDGMENTS

Baoyuan Wu is supported in part by the National Natural Science Foundation of China under Grant 62076213, the University Development Fund of the Chinese University of Hong Kong, Shenzhen under Grant 01001810, and the Special Project Fund of Shenzhen Research Institute of Big Data under Grant T00120210003. Zhan Qin is supported in part by the National Natural Science Foundation of China under Grant U20A20178, the National Key Research and Development Program of China under Grant 2020AAA0107705, and the Research Laboratory for Data Security and Privacy, Zhejiang University-Ant Financial Fintech Center. Kui Ren is supported by the National Key Research and Development Program of China under Grant 2020AAA0107705.

ETHICS STATEMENT

DNNs are widely adopted in many mission-critical areas (*e.g.*, face recognition) and therefore their security is of great significance. The vulnerability of DNNs to backdoor attacks raises serious concerns about using third-party training resources. In this paper, we propose a general training pipeline to obtain backdoor-free DNNs, even if the training dataset contains poisoned samples. This work has no ethical issues in general since our method is purely defensive and does not reveal any new vulnerabilities of DNNs. However, we need to mention that our defense can be adopted only when training with untrusted samples, and backdoor attacks could happen in other scenarios. People should not be too optimistic about eliminating backdoor threats.

REPRODUCIBILITY STATEMENT

The detailed descriptions of datasets, models, and training settings are in Appendix A-D. We also describe the computational facilities and cost in Appendix J-K. Codes of our DBD are also open-sourced.

REFERENCES

- David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *NeurIPS*, 2019.
- Antonin Berthon, Bo Han, Gang Niu, Tongliang Liu, and Masashi Sugiyama. Confidence scores make instance-dependent label-noise learning possible. In *ICML*, 2021.
- 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.
- Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *IEEE FG*, 2018.
- Nicholas Carlini and Andreas Terzis. Poisoning and backdooring contrastive learning. *arXiv preprint arXiv:2106.09667*, 2021.
- Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, Percy Liang, and John C Duchi. Unlabeled data improves adversarial robustness. In *NeurIPS*, 2019.
- David M Chan, Roshan Rao, Forrest Huang, and John F Canny. Gpu accelerated t-distributed stochastic neighbor embedding. *Journal of Parallel and Distributed Computing*, 131:1–13, 2019.
- 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. In *AAAI Workshop*, 2019.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020a.

- Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.
- 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.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.
- Bao Gia Doan, Ehsan Abbasnejad, and Damith C Ranasinghe. Februus: Input purification defense against trojan attacks on deep neural network systems. In *ACSAC*, 2020.
- Yinpeng Dong, Xiao Yang, Zhijie Deng, Tianyu Pang, Zihao Xiao, Hang Su, and Jun Zhu. Black-box detection of backdoor attacks with limited information and data. In *ICCV*, 2021.
- Min Du, Ruoxi Jia, and Dawn Song. Robust anomaly detection and backdoor attack detection via differential privacy. In *ICLR*, 2020.
- Mingfei Gao, Zizhao Zhang, Guo Yu, Serkan Ö Arık, Larry S Davis, and Tomas Pfister. Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In *ECCV*, 2020.
- Yuan Gao, Jiayi Ma, and Alan L Yuille. Semi-supervised sparse representation based classification for face recognition with insufficient labeled samples. *IEEE Transactions on Image Processing*, 26(5):2545–2560, 2017.
- Jean-Bastien Grill, Florian Strub, Florent Althé, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In *NeurIPS*, 2020.
- Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019.
- Wenbo Guo, Lun Wang, Yan Xu, Xinyu Xing, Min Du, and Dawn Song. Towards inspecting and eliminating trojan backdoors in deep neural networks. In *ICDM*, 2020.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.
- Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. In *NeurIPS*, 2019.
- Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.
- Jinyuan Jia, Yupei Liu, and Neil Zhenqiang Gong. Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning. In *IEEE S&P*, 2021.
- Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
- Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *CVPR*, 2019.
- Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.
- Ang Li, Shanshan Zhao, Xingjun Ma, Mingming Gong, Jianzhong Qi, Rui Zhang, Dacheng Tao, and Ramamohanarao Kotagiri. Short-term and long-term context aggregation network for video inpainting. In *ECCV*, 2020a.

- Shaofeng Li, Minhui Xue, Benjamin Zhao, Haojin Zhu, and Xinpeng Zhang. Invisible backdoor attacks on deep neural networks via steganography and regularization. *IEEE Transactions on Dependable and Secure Computing*, 2020b.
- Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. In *ICLR*, 2021a.
- Yiming Li, Baoyuan Wu, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor learning: A survey. *arXiv preprint arXiv:2007.08745*, 2020c.
- Yiming Li, Tongqing Zhai, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor attack in the physical world. In *ICLR Workshop*, 2021b.
- Yiming Li, Haoxiang Zhong, Xingjun Ma, Yong Jiang, and Shu-Tao Xia. Few-shot backdoor attacks on visual object tracking. In *ICLR*, 2022.
- Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *ICCV*, 2021c.
- Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 2021.
- Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2016.
- Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In *ICML*, 2020.
- Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.
- Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. In *ICLR*, 2018.
- Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *CVPR*, 2017.
- Anh Nguyen and Anh Tran. Input-aware dynamic backdoor attack. In *NeurIPS*, 2020.
- Anh Nguyen and Anh Tran. Wanet-imperceptible warping-based backdoor attack. In *ICLR*, 2021.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.
- Ximing Qiao, Yukun Yang, and Hai Li. Defending neural backdoors via generative distribution modeling. In *NeurIPS*, 2019.
- Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. In *NeurIPS*, 2015.
- Aniruddha Saha, Ajinkya Tejankar, Soroush Abbasi Koohpayegani, and Hamed Pirsiavash. Backdoor attacks on self-supervised learning. *arXiv preprint arXiv:2105.10123*, 2021.
- Guangyu Shen, Yingqi Liu, Guanhong Tao, Shengwei An, Qiuling Xu, Siyuan Cheng, Shiqing Ma, and Xiangyu Zhang. Backdoor scanning for deep neural networks through k-arm optimization. In *ICML*, 2021.
- Changhao Shi, Chester Holtz, and Gal Mishne. Online adversarial purification based on self-supervision. In *ICLR*, 2021.

- Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, 2020.
- Robert Stanforth, Alhussein Fawzi, Pushmeet Kohli, et al. Are labels required for improving adversarial robustness? In *NeurIPS*, 2019.
- Ruixiang Tang, Mengnan Du, Ninghao Liu, Fan Yang, and Xia Hu. An embarrassingly simple approach for trojan attack in deep neural networks. In *KDD*, 2020.
- Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In *NeurIPS*, 2018.
- Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks. *arXiv preprint arXiv:1912.02771*, 2019.
- Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.
- Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, 2020.
- Binghui Wang, Xiaoyu Cao, Neil Zhenqiang Gong, et al. On certifying robustness against backdoor attacks via randomized smoothing. In *CVPR Workshop*, 2020a.
- 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 *IEEE S&P*, 2019a.
- Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *ICCV*, 2019b.
- Yue Wang, Alireza Fathi, Abhijit Kundu, David A Ross, Caroline Pantofaru, Tom Funkhouser, and Justin Solomon. Pillar-based object detection for autonomous driving. In *ECCV*, 2020b.
- Maurice Weber, Xiaojun Xu, Bojan Karlas, Ce Zhang, and Bo Li. Rab: Provable robustness against backdoor attacks. *arXiv preprint arXiv:2003.08904*, 2020.
- Xin Wen, Biying Li, Haiyun Guo, Zhiwei Liu, and Guosheng Hu. Adaptive variance based label distribution learning for facial age estimation. In *ECCV*, 2020.
- Haibin Wu, Xu Li, Andy T Liu, Zhiyong Wu, Helen Meng, and Hung-yi Lee. Adversarial defense for automatic speaker verification by cascaded self-supervised learning models. In *ICASSP*, 2021.
- Zhen Xiang, David J Miller, Siheng Chen, Xi Li, and George Kesidis. A backdoor attack against 3d point cloud classifiers. In *ICCV*, 2021.
- Zhen Xiang, David J. Miller, and George Kesidis. Post-training detection of backdoor attacks for two-class and multi-attack scenarios. In *ICLR*, 2022.
- Xiaojun Xu, Qi Wang, Huichen Li, Nikita Borisov, Carl A Gunter, and Bo Li. Detecting ai trojans using meta neural analysis. In *IEEE S&P*, 2021.
- Mingfu Xue, Can He, Jian Wang, and Weiqiang Liu. One-to-n & n-to-one: Two advanced backdoor attacks against deep learning models. *IEEE Transactions on Dependable and Secure Computing*, 2020.
- Zhicong Yan, Gaolei Li, Yuan Tian, Jun Wu, Shenghong Li, Mingzhe Chen, and H Vincent Poor. Dehib: Deep hidden backdoor attack on semi-supervised learning via adversarial perturbation. In *AAAI*, 2021.
- Yi Zeng, Won Park, Z Morley Mao, and Ruoxi Jia. Rethinking the backdoor attacks’ triggers: A frequency perspective. In *ICCV*, 2021a.

- Yi Zeng, Han Qiu, Shangwei Guo, Tianwei Zhang, Meikang Qiu, and Bhavani Thuraisingham. Deepsweep: An evaluation framework for mitigating dnn backdoor attacks using data augmentation. In *AsiaCCS*, 2021b.
- Yi Zeng, Si Chen, Won Park Z. Morley Mao, Ming Jin, and Ruoxi Jia. Adversarial unlearning of backdoors via implicit hypergradient. In *ICLR*, 2022.
- Tongqing Zhai, Yiming Li, Ziqi Zhang, Baoyuan Wu, Yong Jiang, and Shu-Tao Xia. Backdoor attack against speaker verification. In *ICASSP*, 2021.
- Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.
- Zhilu Zhang and Mert R Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *NeurIPS*, 2018.
- Pu Zhao, Pin-Yu Chen, Payel Das, Karthikeyan Natesan Ramamurthy, and Xue Lin. Bridging mode connectivity in loss landscapes and adversarial robustness. In *ICLR*, 2020a.
- Shihao Zhao, Xingjun Ma, Xiang Zheng, James Bailey, Jingjing Chen, and Yu-Gang Jiang. Clean-label backdoor attacks on video recognition models. In *CVPR*, 2020b.

Table 4: Statistics of datasets and DNNs adopted in our main experiments.

Dataset	Input Size	# Classes	# Training Images	# Test Images	DNN model
CIFAR-10	$3 \times 32 \times 32$	10	50,000	10,000	ResNet-18
ImageNet	$3 \times 224 \times 224$	30	38,939	1,500	ResNet-18

A DETAILED SETTINGS FOR REVISITING BACKDOOR ATTACKS

Attack Setups. We conduct the BadNets (Gu et al., 2019) and label-consistent attack (Turner et al., 2019) with the target label $y_t = 3$ on the CIFAR-10 dataset (Krizhevsky, 2009). The trigger patterns are the same as those presented in Section 5.2. In particular, we implement the label-consistent attack with adversarial perturbations, as suggested in its original paper (Turner et al., 2019). Specifically, we used the projected gradient descent (PGD) (Madry et al., 2018) to generate adversarial perturbations within the ℓ^∞ -ball where the maximum perturbation size $\epsilon = 16$.

Training Setups. We conduct supervised learning on the poisoned datasets with the standard training process and the self-supervised learning on the unlabelled poisoned datasets with the SimCLR (Chen et al., 2020a). The supervised training is conducted based on the open-source code². Specifically, we use the SGD optimizer with momentum 0.9, weight decay of 5×10^{-4} , and an initial learning rate of 0.1. The batch size is set to 128 and we train the ResNet-18 model 200 epochs. The learning rate is decreased by a factor of 10 at epoch 100 and 150, respectively. Besides, we add triggers before performing the data augmentation (*e.g.*, random crop and horizontal flipping). For the self-supervised training, we use the stochastic gradient descent (SGD) optimizer with a momentum of 0.9, an initial learning rate of 0.4, and a weight decay factor of 5×10^{-4} . We use a batch size of 512, and train the backbone for 1,000 epochs. We decay the learning rate with the cosine decay schedule (Loshchilov & Hutter, 2016) without a restart. Besides, we also adopt strong data augmentation techniques, including random crop and resize (with random flip), color distortions, and Gaussian blur, as suggested in (Chen et al., 2020a). All models are trained until converge.

t-SNE Visualization Settings. We treat the output of the last residual unit as the feature representation and use the tsne-cuda library (Chan et al., 2019) to get the feature embedding of all samples. To have a better visualization, we adopt all poisoned samples and randomly select 10% benign samples for visualizing models under the supervised learning, and adopt 30% poisoned samples and 10% benign samples for those under the self-supervised learning.

B DETAILED SETTINGS FOR MAIN EXPERIMENTS

B.1 MORE DETAILS ABOUT DATASETS AND DNNs

Due to the limitations of computational resources and time, we adopt a subset randomly selected from the original ImageNet. More detailed information about the datasets and DNNs adopted in the main experiments of our paper is presented in Table 4.

B.2 MORE DETAILS ABOUT ATTACK SETTINGS

Attack Setups. We conduct the BadNets (Gu et al., 2019), blended attack (dubbed ‘Blended’) (Chen et al., 2017), label-consistent attack (dubbed ‘Label-Consistent’) (Turner et al., 2019), and WaNet (Nguyen & Tran, 2021) with the target label $y_t = 3$ on all datasets. The trigger patterns are the same as those presented in Section 5.2. In particular, we set the blended ratio $\lambda = 0.1$ for the blended attack on all datasets and examine label-consistent attack with the maximum perturbation size $\epsilon \in \{16, 32\}$. Besides, WaNet assumed that attackers can fully control the whole training process in its original paper. However, we found that WaNet only modified training data while other training components (*e.g.*, training loss, training schedule, and model structure) are the same as those used in the standard training process. As such, we re-implement its code in the poisoning-based attack scenario based on its official code³. Specifically, following the settings in its original paper, we set the noise rate $\rho_n = 0.2$, control grid size $k = 4$, and warping strength $s = 0.5$ on

²<https://github.com/kuangliu/pytorch-cifar>

³https://github.com/VinAIRResearch/Warping-based_Backdoor_Attack-release

Table 5: The Effectiveness of WaNet with different kernel size k when the noise rate is set to 0 and strength $s = 1$ on the ImageNet dataset.

$k \rightarrow$	4	32	128	224
BA	78.31	77.64	81.29	80.98
ASR	6.87	14.98	91.69	96.22

Table 6: The Effectiveness of WaNet with different strength s when the noise rate is set to 0 and strength kernel size $k = 224$ on the ImageNet dataset.

$s \rightarrow$	0.4	0.6	0.8	1
BA	81.01	81.45	81.32	80.98
ASR	80.06	92.22	94.23	96.22

Table 7: The Effectiveness of WaNet with and without the noise mode when the kernel size $k = 224$ and strength $s = 1$ on the ImageNet dataset.

	w/ noise mode	w/o noise mode
BA	79.37	80.98
ASR	31.72	96.22

the CIFAR-10 dataset. However, we found that the default k and s are too small to make the attack works on the ImageNet dataset (as shown in Table 5-6). Besides, the ‘noise mode’ also significantly reduces the attack effectiveness (as shown in Table 7). As such, we set $k = 224$ and $s = 1$ and train models without the noise mode on the ImageNet dataset.

Training Setups. On the CIFAR-10 dataset (Krizhevsky, 2009), the settings are the same as those described in Section A; On the ImageNet dataset (Deng et al., 2009), we conduct experiments based on the open-source code⁴. Specifically, we use the SGD optimizer with momentum 0.9, weight decay of 10^{-4} , and an initial learning rate of 0.1. The batch size is set to 256 and we train the ResNet-18 model 90 epochs. The learning rate is decreased by a factor of 10 at epoch 30 and 60, respectively. Besides, since the raw images in the ImageNet dataset are of different sizes, we resize them to $3 \times 224 \times 224$ before adding triggers.

B.3 MORE DETAILS ABOUT DEFENSE SETTINGS

Settings for NC. We conduct reverse engineering and anomaly detection based on its open-source code⁵. We implement the ‘unlearning’ method to patch attacked models, as suggested in its paper (Wang et al., 2019a). We randomly select 5% benign training samples as the local benign dataset, which is used in the ‘unlearning’ process. Unless otherwise specified, other settings are the same as those used in (Wang et al., 2019a).

Settings for NAD. We implement this method based on its open-source code⁶. The origin NAD only conducted experiments on the WideResNet model. In our paper, we calculate the NAD loss over the last residual group for the ResNet-18. The local benign dataset is the same as the one adopted in NC, which is used in the fine-tuning and distillation process of NAD. Unless otherwise specified, other settings are the same as those used in (Li et al., 2021a).

Settings for DPSGD. The original DPSGD was conducted on the MNIST dataset implemented by the TensorFlow Framework. In this paper, we re-implement it based on the differentially private SGD method provided by the Opacus⁷. Specifically, we replace the original SGD optimizer with the differentially private one, as suggested in (Du et al., 2020). There are two important hyper-parameters in DPSGD, including noise scales σ and the clipping bound C . In the experiments, we set $C = 1$ and select the best σ by the grid-search.

⁴<https://github.com/pytorch/examples/tree/master/imagenet>

⁵<https://github.com/bolunwang/backdoor>

⁶<https://github.com/bboylyg/NAD>

⁷<https://github.com/pytorch/opacus>

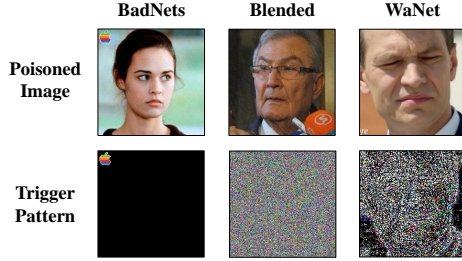


Figure 5: An example of poisoned samples generated by different attacks on VGGFace2 dataset.

Table 8: Statistics of the VGGFace2 dataset and model structure adopted in our experiments.

Dataset	Input Size	# Classes	# Training Images	# Test Images	DNN model
VGGFace2	$3 \times 224 \times 224$	30	9,051	2,263	DenseNet-121

Table 9: Defending against attacks on VGGFace2 dataset. Note that NC and NAD need an additional local benign dataset, which is not required in DPSGD, ShrinkPad, and our method.

Attack →	BadNets		Blended		WaNet	
Defense ↓, Metric →	BA	ASR	BA	ASR	BA	ASR
No Defense	91.52	88.14	93.35	98.94	91.41	98.75
NC	92.16	18.82	93.43	86.76	91.32	95.09
NAD	87.29	<u>9.49</u>	85.27	<u>10.25</u>	80.69	<u>4.41</u>
DPSGD	60.83	6.61	81.90	93.63	54.04	30.85
ShrinkPad	77.64	61.50	79.18	75.61	78.62	30.13
DBD (ours)	<u>89.74</u>	0.17	<u>93.29</u>	0.09	<u>89.32</u>	0

Settings for ShrinkPad. We set the shrinking rate to 10% on all datasets, as suggested in (Li et al., 2021b; Zeng et al., 2021b). Following their settings, we pad 0-pixels at the bottom right of the shrunk image to expand it to its original size.

Settings for our Defense. In this first stage, We adopt SimCLR (Chen et al., 2020a) to perform self-supervised learning. We train backbones 100 instead of 1,000 epochs to reduce computational costs while preserving effectiveness. Other settings are the same as those described in Section A. We use the same settings across all datasets, models, and attacks; In the second stage, we use the Adam optimizer with a learning rate of 0.002 and set the batch size to 128. We train the fully connected layers 10 epochs with the SCE loss (Wang et al., 2019b). Two hyper-parameters involved in the SCE (*i.e.*, α and β in the original paper) are set to 0.1 and 1, respectively. After that, we filter 50% high-credible samples. We use the same settings across all datasets, models, and attacks; In the third stage, we adopt the MixMatch (Berthelot et al., 2019) for semi-supervised fine-tuning with settings suggested in its original paper. Specifically, we use the Adam optimizer with a learning rate of 0.002, the batch size of 64, and finetune the model 190 epochs on the CIFAR-10 and 80 epochs on the ImageNet dataset, respectively. We set the temperature $T = 0.5$ and the weight of unsupervised loss $\lambda_u = 15$ on the CIFAR-10 and $\lambda_u = 6$ on the ImageNet dataset, respectively. Moreover, we re-filter high-credible samples after every epoch of the third stage based on the SCE loss.

C DEFENDING AGAINST ATTACKS ON VGGFACE2 DATASET

Dataset and DNN. Due to the limitations of computational resources and time, we adopt a subset randomly selected from the original VGGFace2 (Cao et al., 2018). More details are in Table 8.

Settings for Attacks. For the training of models on the VGGFace2 dataset, the batch size is set to 32 and we conduct experiments on the DenseNet-121 model (Huang et al., 2017). An example of poisoned samples generated by different attacks are in Figure 5. Other settings are the same as those used on the ImageNet dataset.

Settings for Defenses. For NAD, we calculate the NAD loss over the second to last layer for the DenseNet-121. Other settings are the same as those used on the ImageNet dataset.

Table 10: Results of DPSGD against the BadNets and blended attack with different noise scale σ .

Dataset \rightarrow	CIFAR-10				ImageNet				VGGFace2			
Attack \rightarrow	BadNets		Blended		BadNets		Blended		BadNets		Blended	
$\sigma \downarrow$	BA	ASR	BA	ASR	BA	ASR	BA	ASR	BA	ASR	BA	ASR
0	89.78	100	90.03	92.64	56.04	97.01	55.12	84.17	85.96	99.91	85.34	97.68
0.005	89.38	100	89.48	91.15	55.67	95.44	55.92	85.99	84.64	99.91	84.97	97.29
0.01	88.85	100	88.67	91.50	57.19	93.10	55.46	84.79	83.12	99.87	81.90	93.63
0.05	81.04	100	80.85	81.03	55.04	87.26	54.22	86.06	60.83	6.61	58.02	98.48
0.1	69.77	100	68.10	72.47	51.43	25.40	50.12	89.35	43.55	17.80	43.26	100
0.3	52.48	99.99	53.05	20.68	39.33	7.49	36.97	87.13	17.59	22.08	13.32	53.20
0.5	44.24	14.56	45.26	14.19	32.26	14.46	31.60	84.23	10.17	48.90	9.27	53.50
0.7	40.37	21.23	39.71	15.49	29.06	21.46	25.38	88.69	8.74	41.55	6.55	61.50
1	35.16	32.27	36.51	29.24	23.70	31.30	22.27	82.23	6.64	31.59	6.46	21.43

Table 11: Results of DPSGD against the label-consistent attack with different noise scale σ .

$\epsilon \downarrow$	Poisoning Rate \downarrow	$\sigma \rightarrow$ Metric \downarrow	0	0.005	0.01	0.05	0.1	0.3	0.5	0.7	1
16	0.6%	BA	89.69	89.62	89.15	82.25	70.24	53.41	43.82	39.98	36.65
		ASR	88.79	89.00	88.16	4.30	6.68	8.80	3.51	7.97	16.04
	2.5%	BA	89.40	89.56	88.59	81.25	69.87	53.16	44.28	35.08	36.28
		ASR	97.95	98.51	98.57	99.42	80.32	69.04	7.97	13.98	16.27
32	0.6%	BA	89.75	89.43	88.77	82.27	71.12	53.42	47.52	39.18	37.06
		ASR	98.29	99.02	98.38	9.23	5.76	13.60	6.29	15.64	14.70
	2.5%	BA	89.87	89.38	89.09	82.15	69.79	53.02	44.08	38.31	36.13
		ASR	99.69	98.89	99.57	99.86	9.06	18.88	6.62	20.97	7.32

Table 12: Results of DPSGD against the WaNet with different noise scale σ .

Dataset \rightarrow	CIFAR-10		ImageNet		VGGFace2	
$\sigma \downarrow$, Metric \rightarrow	BA	ASR	BA	ASR	BA	ASR
0	87.76	87.06	53.26	96.06	83.62	95.22
0.005	88.21	84.71	53.38	98.14	80.72	96.47
0.01	88.13	82.05	54.03	97.72	79.88	83.57
0.05	81.03	80.59	51.48	97.00	54.04	30.85
0.1	66.85	72.75	47.65	98.63	33.89	99.82
0.3	52.13	39.78	29.65	96.45	9.44	73.99
0.5	41.56	39.11	22.13	93.45	6.41	80.82
0.7	36.51	33.60	20.72	54.12	6.10	86.98
1	32.66	39.75	15.15	58.10	6.36	52.16

Results. As shown in Table 9, our defense still reaches the best performance even compared with NC and NAD. Specifically, the BA of NC is on par with that of our method whereas it is with the sacrifice of ASR. These results verify the effectiveness of our defense again.

D SEARCHING BEST RESULTS FOR DPSGD AND NAD

The effectiveness of DPSGD and NAD is sensitive to their hyper-parameters. Here we search for their best results based on the criteria that ‘BA – ASR’ reaches the highest value after the defense.

D.1 SEARCHING BEST RESULTS FOR DPSGD

In general, the larger the σ , the smaller the ASR while also the smaller the BA. The results of DPSGD are shown in Table 10-12, where the best results are marked in boldface.

Table 13: Results of the fine-tuning process in NAD against the BadNets and blended attack with different learning rates η .

Dataset \rightarrow	CIFAR-10				ImageNet				VGGFace2			
Attack \rightarrow	BadNets		Blended		BadNets		Blended		BadNets		Blended	
$\eta \downarrow$, Metric \rightarrow	BA	ASR	BA	ASR	BA	ASR	BA	ASR	BA	ASR	BA	ASR
0.1	32.64	5.12	38.24	1.43	62.73	1.09	56.72	0.95	90.35	24.36	91.39	99.96
0.01	90.70	15.66	89.77	11.32	80.37	43.90	80.28	95.82	92.04	83.80	92.95	99.96
0.001	94.91	100	93.25	96.79	81.47	86.79	80.85	99.12	91.49	90.91	92.91	99.96

Table 14: Results of the fine-tuning process in NAD against the label-consistent attack with different learning rates η on the CIFAR-10 dataset.

$\epsilon \downarrow$	Poisoning Rate \downarrow	$\eta \rightarrow$ Metric \downarrow	0.1	0.01	0.001
16	0.6%	BA	36.77	89.91	95.38
		ASR	0.63	44.90	94.69
	2.5%	BA	31.12	93.07	95.03
		ASR	0	99.67	100
32	0.6%	BA	30.73	91.91	94.89
		ASR	1.15	38.27	99.07
	2.5%	BA	35.93	91.23	94.79
		ASR	0	96.19	99.98

Table 15: Results of the fine-tuning process in NAD against WaNet with different learning rates η .

Dataset \rightarrow	CIFAR-10		ImageNet		VGGFace2	
$\eta \downarrow$, Metric \rightarrow	BA	ASR	BA	ASR	BA	ASR
0.1	48.74	2.75	59.11	0.34	89.05	91.40
0.01	90.60	10.37	79.11	67.09	89.01	99.69
0.001	94.16	97.65	80.63	99.39	88.19	99.73

Table 16: Results of the NAD against attacks on the CIFAR-10 dataset with different β .

Attack \downarrow	$\beta \rightarrow$ Metric \downarrow	500	1,000	1,500	2,000	2,500	5,000
BadNets	BA	92.96	91.87	90.47	86.26	65.26	19.91
	ASR	27.74	7.92	1.20	1.74	5.81	0.40
Blended Attack	BA	91.14	89.72	85.30	75.43	61.18	26.87
	ASR	6.25	1.51	3.19	3.55	3.67	9.61
WaNet	BA	92.43	91.83	88.08	78.22	58.32	26.63
	ASR	18.48	9.49	3.58	3.54	3.12	0.32

D.2 SEARCHING BEST RESULTS FOR NAD

We found that the fine-tuning stage of NAD is sensitive to the learning rate. We search the best initial learning rate from $\{0.1, 0.01, 0.001\}$. As shown in Table 13-15, a very large learning rate significantly reduces the BA, while a very small learning rate can not reduce the ASR effectively. To keep a relatively large BA while maintaining a small ASR, we set $\eta = 0.01$ in the fine-tuning stage.

The distillation stage of NAD is also sensitive to its hyper-parameter β . We select the best β via the grid-search. The results are shown in Table 16-19.

Table 17: Results of the NAD against attacks on the ImageNet dataset with different β .

Attack \downarrow	$\beta \rightarrow$ Metric \downarrow	5,000	6,000	7,000	8,000	9,000	10,000
BadNets	BA	79.98	79.20	76.88	73.49	69.46	67.55
	ASR	24.65	15.50	16.99	5.81	5.95	2.51
Blended	BA	79.98	77.57	74.88	70.88	64.44	53.75
	ASR	88.34	75.79	15.12	8.15	3.47	2.99
WaNet	BA	79.65	79.06	77.63	76.77	75.27	74.88
	ASR	76.06	69.23	61.54	45.97	31.43	28.58

Table 18: Results of the NAD against attacks on the VGGFace2 dataset with different β .

Attack \downarrow	$\beta \rightarrow$ Metric \downarrow	10	30	40	50	60	70
BadNets	BA	89.79	78.10	76.62	83.82	87.29	47.53
	ASR	36.02	17.87	0.18	14.52	9.49	0
Blended	BA	90.18	83.48	85.27	73.43	75.01	51.81
	ASR	97.11	68.97	10.25	0	0.36	0.27
WaNet	BA	82.98	80.69	84.27	26.87	20.71	75.10
	ASR	66.10	4.41	15.98	0.06	1.88	0.54

Table 19: Results of the NAD against the label consistent attack on the CIFAR-10 dataset with different β .

$\epsilon \downarrow$	Poisoning Rate \downarrow	$\beta \rightarrow$ Metric \downarrow	500	1,000	1,500	2,000	2,500	5,000
16	0.6%	BA	92.94	91.66	90.29	85.14	73.05	18.49
		ASR	53.11	33.78	12.43	4.39	4.36	0.1
	2.5%	BA	93.36	93.04	90.31	75.11	54.55	24.37
		ASR	36.13	3.95	1.88	3.13	0.22	0
32	0.6%	BA	92.81	92.13	91.21	85.72	60.06	18.81
		ASR	52.50	57.63	38.53	9.99	1.33	0.06
	2.5%	BA	93.40	92.02	90.98	86.32	68.45	20.53
		ASR	48.73	17.99	6.54	2.93	1.79	0

E DEFENDING AGAINST LABEL-CONSISTENT ATTACK WITH A SMALLER POISONING RATE

For the label-consistent attack, except for the 2.5% poisoning rate examined in the main manuscript, 0.6% is also an important setting provided in its original paper (Turner et al., 2019). In this section, we compare different defenses against the label-consistent attack with poisoning rate $\gamma = 0.6\%$.

As shown in Table 20, when defending against label-consistent attack with a 0.6% poisoning rate, our method is still significantly better than defenses having the same requirements (*i.e.*, DPSGD and ShrinkPad). Even compared with those having the additional requirement (*i.e.*, NC and NAD) under their best settings, our defense is still better or on par with them under the default settings. These results verify the effectiveness of our method again.

Table 20: The effectiveness (%) of defending against the label-consistent attack with 0.6% and 2.5% poisoning rate on CIFAR-10 dataset. Among all defense methods, the one with the best performance is indicated in boldface and the value with underline denotes the second-best result. Note that NC and NAD require to have an additional local benign dataset, which is not required in DPSGD, ShrinkPad, and our method.

$\epsilon \downarrow$	poison rate \downarrow	Defense \rightarrow Metric \downarrow	No Defense	NC	NAD	DPSGD	ShrinkPad	DBD (ours)
16	0.6%	BA	95.34	94.51	85.14	82.25	90.59	<u>91.88</u>
		ASR	92.20	<u>0.73</u>	4.39	4.30	10.22	0.46
	2.5%	BA	94.90	94.94	93.04	44.28	90.67	89.67
		ASR	99.33	<u>1.31</u>	3.95	7.97	9.60	0.01
32	0.6%	BA	94.94	94.53	85.72	88.27	<u>90.48</u>	90.04
		ASR	97.69	<u>0.72</u>	9.99	9.23	<u>11.47</u>	0.15
	2.5%	BA	94.82	94.57	90.98	69.79	90.83	<u>91.45</u>
		ASR	99.19	<u>1.36</u>	6.54	9.06	13.03	0.34

Table 21: DBD against BadNets with different triggers on CIFAR-10 dataset.

		Location					Trigger Size			
		Upper Left	Upper Right	Center	Lower Left	Lower Right	1×1	2×2	3×3	4×4
No Defense	BA	95.01	95.01	94.97	94.72	94.77	94.70	94.99	95.01	94.89
	ASR	99.78	99.70	99.86	99.68	99.70	92.13	99.35	99.78	99.86
DBD (ours)	BA	92.46	92.47	93.32	92.92	92.58	93.10	92.91	92.46	92.46
	ASR	1.27	0.53	0.51	0.64	1.41	0.80	0.65	1.27	1.02

F DEFENDING AGAINST ATTACKS WITH DIFFERENT TRIGGER PATTERNS

In this section, we verify whether DBD is still effective when different trigger patterns are adopted.

Settings. For simplicity, we adopt the BadNets on the CIFAR-10 dataset as an example for the discussion. Specifically, we change the *location* and *size* of the backdoor trigger while keeping other settings unchanged to evaluate the BA and ASR before and after our defense.

Results. As shown in Table 21, although there are some fluctuations, the ASR is smaller than 2% while the BA is greater than 92% in every cases. In other words, our method is effective in defending against attacks with different trigger patterns.

G DEFENDING AGAINST ATTACKS WITH DYNAMIC TRIGGERS

In this section, we verify whether DBD is still effective when attackers adopt dynamic triggers.

Settings. We compare DBD with MESA (Qiao et al., 2019) in defending the dynamic attack discussed in (Qiao et al., 2019) on the CIFAR-10 dataset as an example for the discussion. This dynamic attack uses a distribution of triggers instead of a fixed trigger.

Results. The BA and ASR of DBD are 92.4% and 0.4%, while those of MESA are 94.8% and 2.4%. However, we find MESA failed in defending against blended attack (for it can not correctly detect the trigger) whereas DBD is still effective. These results verified the effectiveness of our defense.

H DISCUSSIONS

H.1 EFFECTS OF HYPER-PARAMETERS

Settings. Here we analyze the effect of filtering rate α , which is the only key method-related hyper-parameter in our DBD. We adopt the results on the CIFAR-10 dataset for discussion. Except for the studied parameter α , other settings are the same as those used in Section 5.2.

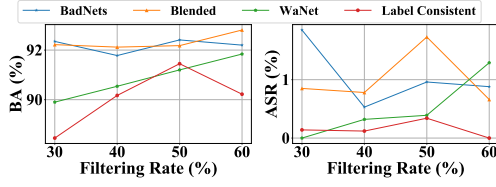


Figure 6: The effects of filtering rate.

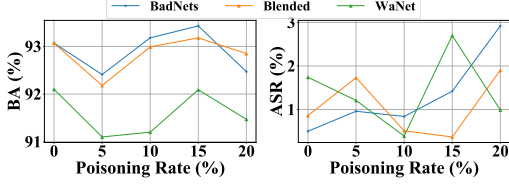


Figure 7: The effects of poisoning rate.

Results. The number of labeled samples used in the third stage increase with the increase of filtering rate α , while the probability that the filtered high-credible dataset contains poisoned samples also increases. As shown in Figure 6, DBD can still maintain relatively high benign accuracy even when the filtering rate α is relatively small (*e.g.*, 30%). It is mostly due to the high-quality of learned purified feature extractor and the semi-supervised fine-tuning process. DBD can also reach a nearly 0% attack success rate in all cases. However, we also have to notice that the high-credible dataset may contain poisoned samples when α is very large, which in turn creates hidden backdoors again during the fine-tuning process. Defenders should specify α based on their specific needs.

H.2 DEFENDING ATTACKS WITH VARIOUS POISONING RATES

Settings. We evaluate our method in defending against attacks with different poisoning rate γ on CIFAR-10 dataset. Except for γ , other settings are the same as those used in Section 5.2.

I MORE DETAILS ABOUT SIMCLR, SCE, AND MIXMATCH

NT-Xent Loss in SimCLR. Given a sample mini-batch containing N different samples, SimCLR first applies two separate data augmentations toward each sample to obtain $2N$ augmented samples. The loss for a positive pair of sample (i, j) can be defined as:

$$\mathcal{L}_{i,j} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j) / \tau)}{\sum_{k=1}^{2N} \mathbb{I}\{k \neq i\} \cdot \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k) / \tau)}, \quad (5)$$

where $\text{sim}(\cdot, \cdot)$ is the cosine similarity, \mathbf{z}_i is the feature representation of sample i , τ is the temperature parameter, and $\mathbb{I}\{k \neq i\} \in \{0, 1\}$ indicating whether $k \neq i$. The NT-Xent Loss is computed across all $2N$ positive pairs in this mini-batch.

SCE. The symmetric cross entropy (SCE) can be defined as:

$$\mathcal{L}_{SCE} = H(p, q) + H(q, p), \quad (6)$$

where $H(p, q)$ is the cross entropy, $H(q, p)$ is the reversed cross entropy, p is the prediction, and q is the one-hot label (of the evaluated sample).

MixMatch Loss. For a batch \mathcal{X} of labeled samples and a batch \mathcal{U} of unlabeled samples ($|\mathcal{X}| = |\mathcal{U}|$), MixMatch produces a guessed label \bar{q} for each unlabeled sample $u \in \mathcal{U}$ and applies MixUp (Zhang et al., 2018) to obtain the augmented \mathcal{X}' and \mathcal{U}' . The loss $\mathcal{L}_{\mathcal{X}}$ and $\mathcal{L}_{\mathcal{U}}$ can be defined as:

$$\mathcal{L}_{\mathcal{X}} = \frac{1}{|\mathcal{X}'|} \sum_{(x,q) \in \mathcal{X}'} H(p_x, q), \quad (7)$$

where p_x is the prediction of x , q is its one-hot label, and $H(\cdot, \cdot)$ is the cross entropy.

$$\mathcal{L}_{\mathcal{U}} = \frac{1}{K \cdot |\mathcal{U}'|} \sum_{(u,\bar{q}) \in \mathcal{U}'} \|p_u - \bar{q}\|_2^2, \quad (8)$$

where p_u is the prediction of u , \bar{q} is its guessed one-hot label, and K is the number of classes.

By combining $\mathcal{L}_{\mathcal{X}}$ with $\mathcal{L}_{\mathcal{U}}$, the MixMatch loss can be defined as:

$$\mathcal{L} = \mathcal{L}_{\mathcal{X}} + \lambda_{\mathcal{U}} \cdot \mathcal{L}_{\mathcal{U}}, \quad (9)$$

where $\lambda_{\mathcal{U}}$ is a hyper-parameter.

J COMPUTATIONAL FACILITIES

We conduct all experiments on two Ubuntu 18.04 servers having different GPUs. One has four NVIDIA GeForce RTX 2080 Ti GPUs with 11GB memory (dubbed ‘RTX 2080Ti’) and the another has three NVIDIA Tesla V100 GPUs with 32GB memory (dubbed ‘V100’).

Computational Facilities for Attacks. All experiments are conducted with a single RTX 2080 Ti.

Computational Facilities for Defenses. Since we do not use a memory-efficient implementation of DenseNet-121, we conduct DPSGD experiments on the VGGFace2 dataset with a single V100. Other experiments of baseline defenses are conducted with a single RTX 2080 Ti. For our defense, we adopt PyTorch (Paszke et al., 2019) distributed data-parallel and automatic mixed precision training (Miciekevicius et al., 2018) with two RTX 2080 Ti for self-supervised learning on the VGGFace2 dataset. Other experiments are conducted with a single RTX 2080 Ti.

K COMPUTATIONAL COST

In this section, we analyze the computational cost of our method stage by stage, compared to standard supervised learning.

Stage 1. Self-supervised learning is known to have a higher computational cost than standard supervised learning (Chen et al., 2020a; He et al., 2020). In our experiments, SimCLR requires roughly four times the computational cost of standard supervised learning. Since we intend to get a purified instead of well-trained feature extractor, we train the feature extractor (*i.e.*, backbone) lesser epochs than the original SimCLR to reduce the training time. As described in Section B.3, we find 100 epochs is enough to preserve effectiveness.

Stage 2. Since we freeze the backbone and only train the remaining fully connected layers, the computational cost is roughly 60% of standard supervised learning.

Stage 3. Semi-supervised learning is known to have a extra labeling cost compared with standard supervised learning (Gao et al., 2020). In our experiments, MixMatch requires roughly two times the computation cost of standard supervised learning.

We will explore a more computational efficient training method in our future work.

L COMPARING OUR DBD WITH DETECTION-BASED BACKDOOR DEFENSES

In this paper, we do not intend to filter malicious and benign samples accurately, as we mentioned in Section 4.4. However, we notice that the second stage of our DBD can serve as a detection-based backdoor defense for it can filter poisoned samples. In this section, we compare the filtering ability of our DBD (stage 2) with existing detection-based backdoor defenses.

Settings. We compare our DBD with two representative detection-based methods, including, Spectral Signatures (SS) (Tran et al., 2018) and Activation Clustering (AC) (Chen et al., 2019), on the CIFAR-10 dataset. These detection-based methods (*e.g.*, SS and AC) filter malicious samples from the training set and train the model on non-malicious samples. Specifically, we re-implement SS in PyTorch based on its official code⁸ and adopt the open-source code⁹ for AC, following the settings in their original paper. In particular, since SS filters 1.5ε malicious samples for each class, where ε is the key hyper-parameter means the upper bound of the number of poisoned training samples, we adopt different ε for a fair comparison.

Results. As shown in Table 22-23, the filtering performance of DBD is on par with that of SS and AC. DBD is even better than those methods when filtering poisoned samples generated by more complicated attacks (*i.e.*, WaNet and Label-Consistent). Besides, we also conduct the standard training on non-malicious samples filtered by SS and AC. As shown in Table 24, the hidden backdoor will still be created in many cases, even though the detection-based defenses are sometimes accurate.

⁸https://github.com/MadryLab/backdoor_data_poisoning

⁹https://github.com/ain-soph/trojanzone/blob/main/trojanvision/defenses/backdoor/activation_clustering.py

Table 22: The successful filtering rate (# filtered poisoned samples / # all filtered samples, %) under different ε in the target class on CIFAR-10 dataset.

Attack ↓	Defense ↓, $\varepsilon \rightarrow$	250	500	1000	1500
BadNets	SS	95.73	93.20	87.60	80.71
	DBD	100	97.60	90.87	70.09
Blended	SS	0.80	10.53	28.87	35.29
	DBD	97.87	94.67	87.27	75.16
WaNet	SS	0	0	1.00	7.42
	DBD	100	100	99.47	97.46
Label-Consistent	SS	2.40	4.40	9.00	13.78
	DBD	43.47	37.07	34.47	32.53

Table 23: The number of remaining poisoned samples over filtered non-malicious samples on CIFAR-10 dataset.

Defense ↓, Attack →	BadNets	Blended	WaNet	Label-Consistent
SS ($\varepsilon = 500$)	1801/42500	2421/42500	2500/42500	1217/42500
SS ($\varepsilon = 1000$)	1186/35000	2067/35000	2400/35000	1115/35000
AC	0/42500	0/37786	5000/45546	1250/39998
DBD	8/25000	6/25000	38/25000	13/25000

Table 24: The BA (%) and ASR (%) of models trained on non-malicious samples filtered by SS and AC on CIFAR-10 dataset.

Defense →	SS ($\varepsilon = 500$)		SS ($\varepsilon = 1000$)		AC	
Attack ↓, Metric →	BA	ASR	BA	ASR	BA	ASR
BadNets	92.99	100	93.27	99.99	85.90	0
Blended	92.84	99.07	92.56	99.18	77.17	0
WaNet	92.69	98.13	91.92	99.00	84.60	99.02
Label-Consistent	92.93	99.79	92.88	99.86	75.95	99.75

Table 25: The BA (%) and ASR (%) of our DBD defending against four attacks with different self-supervised methods on CIFAR-10 dataset.

Attack →	BadNets		Blended		WaNet		Label-Consistent	
Method ↓, Metric →	BA	ASR	BA	ASR	BA	ASR	BA	ASR
SimCLR	92.41	0.96	92.18	1.73	91.20	0.39	91.45	0.34
MoCo-V2	93.01	1.21	92.42	0.24	91.69	1.30	91.46	0.19
BYOL	91.98	0.82	91.38	0.51	91.37	1.28	90.09	0.17

This is mainly because these methods may not able to remove enough poisoned samples while preserving enough benign samples simultaneously, *i.e.*, there is a trade-off between BA and ASR.

M DBD WITH DIFFERENT SELF-SUPERVISED METHODS

In this paper, we believe that the desired feature extractor is mapping visually similar inputs to similar positions in the feature space, such that poisoned samples will be separated into their source classes. This goal is compatible with that of self-supervised learning. We believe that any self-supervised learning can be adopted in our method. To further verify this point, we replace the adopted SimCLR with other self-supervised methods in our DBD and examine their performance.

Settings. We replace the SimCLR with two other self-supervised methods, including MoCo-V2 (Chen et al., 2020b) and BYOL (Grill et al., 2020), in our DBD. Except for the adopted self-supervised method, other settings are the same as those used in Section 5.2.

Results. As shown in Table 25, all DBD variants have similar performances. In other words, our DBD is not sensitive to the selection of self-supervised methods.

Table 26: The BA (%) and ASR (%) of our DBD defending against four attacks with different label-noise learning methods on CIFAR-10 dataset.

Attack →	BadNets		Blended		WaNet		Label-Consistent	
Method ↓, Metric →	BA	ASR	BA	ASR	BA	ASR	BA	ASR
SCE	92.41	0.96	92.18	1.73	91.20	0.39	91.45	0.34
GCE	92.93	0.88	93.06	1.27	92.25	1.51	91.05	0.15
APL	92.95	1.00	92.65	0.78	92.24	1.40	91.08	0.14

N DBD WITH DIFFERENT LABEL-NOISE LEARNING METHODS

In the main manuscript, we adopt SCE as the label-noise learning method in our second stage. In this section, we explore whether our DBD is still effective if other label-noise methods are adopted.

Settings. We replace SCE in our DBD with two other label-noise learning methods, including generalized cross entropy (GCE) (Zhang & Sabuncu, 2018) and active passive loss (APL) (Ma et al., 2020). Specifically, we adopt the combination of NCE+RCE in APL and use the default hyperparameters suggested in their original paper. Except for the adopted label-noise learning method, other settings are the same as those used in Section 5.2.

Results. As shown in Table 26, all DBD variants are effective in reducing backdoor threats (*i.e.*, low ASR) while maintaining high benign accuracy. In other words, our DBD is not sensitive to the selection of label-noise learning methods.

O ANALYZING WHY OUR DBD IS EFFECTIVE IN DEFENDING AGAINST LABEL-CONSISTENT ATTACK

In general, the good defense performance of our DBD method against the label-consistent attack (which is one of the clean-label attacks) can be explained from the following aspects:

Firstly, as shown in Figure 1, there is a common observation across different attacks (including both poisoned- and clean-label attacks) that poisoned samples tend to gather together in the feature space learned by the standard supervised learning. The most intuitive idea of our DBD is to prevent such a gathering in the learned feature space, which is implemented by self-supervised learning. As shown in Figure 1(d), the poisoned samples of label-consistent attack are also separated into different areas in the feature space learned by self-supervised learning. This example gives an intuitive explanation about why our DBD can successfully defend against the label-consistent attack.

Furthermore, it is interesting to explore why the poisoned samples in the label-consistent attack are separated under self-supervised learning since all poisoned samples are from the same target class, rather than from different source classes in poisoned-label attacks. For each poisoned sample in this attack, there are two types of features: the trigger and the benign feature with (untargeted) adversarial perturbations. From the perspective of DNNs, benign samples with (untargeted) adversarial perturbations are similar to samples from different source classes, though these samples look similar from the human’s perspective. Thus, it is not surprising that poisoned samples in clean-label attacks can also be separated under self-supervised learning, just like those in poisoned-label attacks.