Single Image Backdoor Inversion via Robust Smoothed Classifiers

Mingjie Sun[1] Zico Kolter[1,2]
[1]Carnegie Mellon University [2]Bosch Center for AI

Abstract

Backdoor inversion, the process of finding a backdoor "trigger" inserted into a machine learning model, has become the pillar of many backdoor detection and defense methods. Previous works on backdoor inversion often recover the backdoor through an optimization process to flip a support set of clean images into the target class. However, it is rarely studied and understood how large this support set should be to recover a successful backdoor. In this work, we show that one can reliably recover the backdoor trigger with as few as a single image. Specifically, we propose the SmoothInv method, which first constructs a robust smoothed version of the backdoored classifier and then performs guided image synthesis towards the target class to reveal the backdoor pattern. SmoothInv requires neither an explicit modeling of the backdoor via a mask variable, nor any complex regularization schemes, which has become the standard practice in backdoor inversion methods. We perform both quantitaive and qualitative study on backdoored classifiers from previous published backdoor attacks. We demonstrate that compared to existing methods, SmoothInv is able to recover successful backdoors from single images, while maintaining high fidelity to the original backdoor. We also show how we identify the target backdoored class from the backdoored classifier. Last, we propose and analyze two countermeasures to our approach and show that SmoothInv remains robust in the face of an adaptive attacker. Our code is available at https://github.com/locuslab/smoothinv.

1. Introduction

Backdoor attacks [3–5, 9, 11, 14, 29, 30, 45], the practice of injecting a covert backdoor into a machine learning model for inference time manipulation, have become a popular threat model in machine learning security community. Given the pervasive threat of backdoor attacks, e.g. on self-supervised learning [7, 35], language modelling [28, 47] and 3d point cloud [46], there has been growing research interest in reverse engineering the backdoor given a backdoored classifier. This reverse engineering process, often called *backdoor inversion* [39], is crucial in many backdoor defense [15, 32] and detection methods [12, 17, 18, 20, 21, 24, 31, 38, 43, 44]. A successful backdoor inversion method should be able to recover a backdoor satisfying certain requirements. On one hand, the reversed backdoor should be successful, meaning that it should still have a high attack success rate (ASR) on the backdoored classifier. On the other hand, it should be faithful, where the reversed backdoor should be close, e.g. in visual similarity, to the true backdoor.

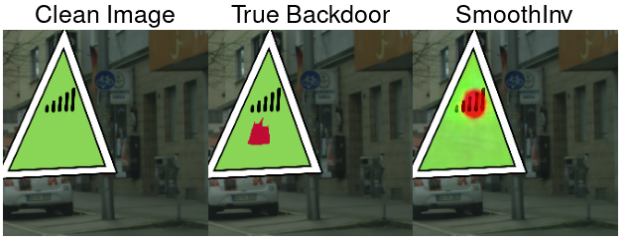


Figure 1. **Single Image Backdoor Inversion**: Given a backdoored classifier (sampled randomly from TrojAI benchmark [1]), SmoothInv takes a *single* clean image (**left**) as input and recovers the hidden backdoor (**right**) with high visual similarity to the original backdoor (**middle**).

Given a backdoored classifier f_b and a support set $\mathcal{S}$ of clean images, a well-established framework for backdoor inversion solves the following optimization problem:

$$\min_{\mathbf{m},\mathbf{p}} \mathbb{E}_{\mathbf{x}\in\mathcal{S}}\big[\mathcal{L}(f_b(\phi(\mathbf{x})), y_t)\big] + \mathcal{R}(\mathbf{m},\mathbf{p}) \qquad (1)$$
$$\text{where } \phi(\mathbf{x}) = (1-\mathbf{m})\odot\mathbf{x} + \mathbf{m}\odot\mathbf{p}$$

where variables $\mathbf{m}$ and $\mathbf{p}$ represent a mask and perturbation vectors respectively, y_t denotes the target label, $\odot$ is element-wise multiplication, $\mathcal{L}(\cdot,\cdot)$ is the cross-entropy function and $\mathcal{R}(\cdot,\cdot)$ is a regularization term. Since it was first proposed in [43], Equation 1 has been adopted and extended by many backdoor inversion methods. While most of these works focus on designing new regularization term $\mathcal{R}$ and backdoor modelling function ϕ, it is often assumed by default that there is a set $\mathcal{S}$ of clean images available. In practice, however, we may not have access to many clean images beforehind. Thus we are motivated by the following question:

Can we perform backdoor inversion with as few clean images as possible?

In this work, we show that one single image is enough for backdoor inversion. On a high level, we view backdoor attacks as encoding backdoored images into the data distribution of the target class during training. Then, we hope to reconstruct these encoded backdoored images via a class-conditional image synthesis process to generate examples from the target class. Though the idea of using image synthesis is straight-forward, it is not immediately obvious how to do this in practice given a backdoored classifier. For instance, directly minimizing the classification loss of the target class reduces to random adversarial noise, as shown in previous adversarial attacks literature [40]. Additionally, generative models such as GANs [13] are not practical in this setting, since we would need to train a separate generative model for each backdoored classifier, and we don't usually have access to the training set for the task of backdoor inversion.

Our approach, which we call SmoothInv, synthesize backdoor patterns by inducing salient gradients of backdoor features via a special robustification process, inserted to convert a standard non-robust model to an adversarially robust one. Specifically, SmoothInv first constructs a robust smoothed version of the backdoored classifier, which is provably robust to adversarial perturbations. Once we have the robust smoothed classifier, we perform guided image synthesis to recover backdoored images that the robust smoothed classifier perceive as the target class.

Compared to the existing inversion framework in Equation 1, our approach uses only a *single* image as the support set. Single image backdoor inversion has not been shown possible for previous backdoor inversion methods as they usually require multiple clean instances for their optimization methods to give reasonable results. Moreover, our approach has the added benefit of simplicity: we do not introduce any custom-designed optimization constraints, e.g. mask variables and regularization. Most importantly, the backdoor found by our approach has remarkable visual resemblance to the original backdoor, despite being constructed from a single image. In Figure 1, we demonstrate such visual similarity for a backdoored classifier.

We evaluate our method on a collection of backdoored classifiers from previously published studies, where we either download their pretrained models or train a replicate using the publicly released code. These collected backdoored classifiers cover a diverse set of backdoor conditions, e.g., patch shape, color, size and location. We evaluate SmoothInv on these backdoored classifiers for single image backdoor inversion and show that SmoothInv finds both successful and faithful backdoors from single images. We also show how we distinguish the target backdoored class from normal classes, where our method (correctly) is unable to find an effective backdoor for the latter. Last, we evaluate attempts to circumvent our approach and show that SmoothInv is still robust under this setting.

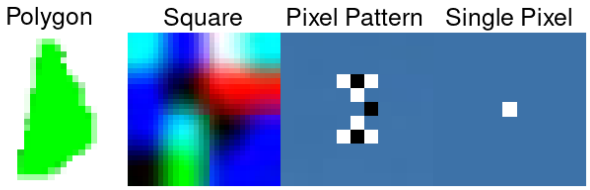


Figure 2. Backdoors of the backdoored classifiers we consider in this paper (listed in Table 1). The polygon trigger (leftmost) is a representative backdoor used in the TrojAI benchmark. The pixel pattern (9 pixels) and single pixel backdoors used in [3] are overlaid on a background blue image for better visualization.

2. Background

2.1. Backdoor Attacks

In a typical backdoor attack [3, 4, 14], an attacker injects malicious crafted samples into the training data. The result of such manipulation is that models trained on such data are able to be manipulated at inference time: the attacker can control the behavior of the model with the injected backdoor. In this work, we consider backdoor attacks on image classification problems, which has become a common evaluation setting for backdoor attacks [3, 14, 42]. Typically, a backdoor attack generates a classifier which satisfies the following two requirements:

- Its accuracy on clean images is barely affected.
- It will always predict a certain target class y_t as long as the backdoor is applied to the input image.

The first property is desired so that it is indistinguishable from clean classifiers by solely comparing their clean accuracies. Some variations and extensions of the second property have been explored. For example, the backdoor can be only effective on images from certain classes [16]. It is also possible to create multiple backdoors in a single backdoored classifier where each backdoor corresponds to a different target class [3].

Following [27], we formalize the backdoor as a transformation function on the image space $\mathcal{B} : \mathcal{X} \to \mathcal{X}$. Given a clean image x, one can create a backdoored image $\mathcal{B}(x)$. One common and widely studied type of backdoor is patch-based backdoor [14]: overlaying a small patch pattern p over input x, i.e. $\mathcal{B}(x) = x \oplus p$. For such backdoor, the backdoored classifier will classify any image with the patch pattern present as the desired target label. However, other forms of image transformations have also been shown to be effective backdoors: e.g., reflection [23], image wrapping [27] and Instagram filters [1]. In this work, we consider patch-based backdoor in particular.

2.2. Backdoor Inversion

Previous backdoor inversion methods mostly follow the framework in Equation 1. The goal is to find a backdoor that is able to simultaneously flip all images in the provided set S of clean images to the target class while at the same

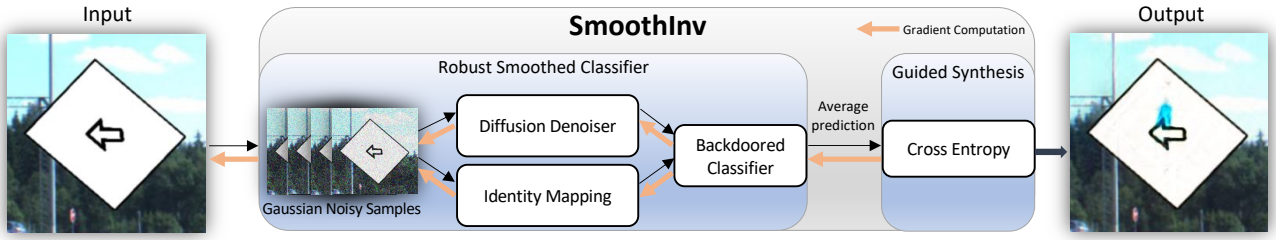


Figure 3. We propose SmoothInv, a backdoor inversion method that takes in a *single* image and synthesize backdoor patterns. SmoothInv consists of two steps: a robustification process and a synthesis process. At the first step, SmoothInv constructs a *robust smoothed* version of the backdoored classifier, where noisy samples of the input are either denoised by a diffusion based denoiser first or passed directly into the backdoored classifier. Next, we synthesize backdoor patterns guided by the *robust smoothed* classifier, where we minimize the standard cross entropy loss with respect to the target class, without relying on any additional regularization term.

time constraining the optimization space of the reversed backdoor. As pointed out by [3], this is essentially finding the smallest universal adversarial patch [6]. Existing inversion methods differ in how they formulate the regularization term $\mathcal{R}$ and how to model the backdoor via $\phi(x)$. For instance, [43] applies a ℓ_1 penalty regularization on the mask variable; [20] uses a diversity loss and a topological loss to regularize the optimization process; [39] models the backdoor via individual pixel changes without using a mask.

One challenge of solving Equation 1 is that it introduces an extra binary mask variable $\mathbf{m}$, which could make the optimization process unstable. In practice, this mask variable is often relaxed to be continuous and converted back to binary in the end. Another optimization obstacle is that it is not clear how to properly set the balancing term between the classification loss and the regularization loss, without a strong domain expertise or a careful hyperparameter search.

2.3. Randomized Smoothing

Randomized Smoothing (RS) [10] is a certified defense method against ℓ_2-norm bounded adversarial perturbations. Given any base classifier $f : \mathcal{X} \to \mathcal{Y}$ and input x, RS first constructs a smoothed classifier g with isotropic Gaussian noise $\delta \sim \mathcal{N}(0, \sigma^2\mathbf{I})$:

$$g(x) := \text{argmax}_c \Pr_{\delta \sim \mathcal{N}(0,\sigma^2\mathbf{I})} \big(f(x+\delta) = c\big) \qquad (2)$$

It is shown in [10] that the smoothed classifier g is certifiably robust in a ℓ_2-norm radius R, where the noise level σ controls the accuracy/robustness tradeoff. In this work, we are not interested in how this certified radius is computed exactly. However, it is necessary to know that the more accurate the smoothed classifier is at classifying noisy images $x + \delta$, the larger the certified radius is (and as a result, more robust). [10] trained base classifier f under standard gaussian augmentations and demonstrated non-trivial certified accuracy on ImageNet.

Following [10], [36] proposed Denoised Smoothing (DS) to certify the prediction of any pre-trained classifier, i.e., not trained with gaussian augmentation. The idea is to prepend an image denoiser $\mathcal{D}$ before the base classifier.

$$g(x) := \text{argmax}_c \Pr_{\delta \sim \mathcal{N}(0,\sigma^2\mathbf{I})} (f \circ \mathcal{D}(x+\delta) = c) \qquad (3)$$

[36] showed that prepending a custom-trained denoiser attains better certified robustness than simply using the plain pre-trained classifier. Most recently, [8] proposed Diffusion Denoised Smoothing (DDS), which used one diffusion step of a diffusion model as the denoiser in Equation 3. DDS obtained state-of-the-art certified robustness. The big performance boost over [36] comes from the strong ability of diffusion models [19] to denoise Gaussian noisy images.

3. Method

An overview of our approach is given in Figure 3. Given a backdoored classifier $f_b : \mathcal{X} \to \mathcal{Y}$ and a clean image x (assuming the backdoor is effective for this image), our goal is to find the backdoor hidden in this clean image.

3.1. Motivation

On a high level, we view backdoor inversion as the problem of recovering/constructing a special type of images, i.e. backdoored images, from the target class. We note that class-conditional image synthesis from generative model literature share some similarity to our goal. However, standard conditional generative models [26] do not see backdoored images during training, and it is not practical to train a custom one with classifier guidance from the backdoored classifier. Our approach is most inspired from another line of work on class-conditional image generation: the work of [37,49] on image synthesis with adversarially robust classifiers. They showed that there were able to perform various image synthesis task without use of any generative models. The foundation for the success of their approach is based on a unique property of robust classifiers, i.e. perceptually-aligned gradients [22, 41], where salient characteristics of target class can be revealed via a projected gradient descent

process [25]. Specifically, in this work, we are interested in how this property can be used for backdoor inversion.

We revisit the denifition of backdoored classifiers in Section 2.1: always predicting the target class as long as the backdoor is present in the image. In other words, the backdoored classifiers have successfully associate the injected backdoor as a new feature for predicting the target class, in addition to features from clean data of the target class. We hypothesize that in the eyes of the backdoored classifers, those backdoored images are encoded into the data distribution of the target class. Thus, we can tackle the problem of backdoor inversion as synthesizing a specific salient characteristics of the target class: the injected backdoor. Note that [37] is not immediately applicable here as backdoored classifiers are not adversarially robust by construction [14]. In the next section, we describe how we reliably extract salient backdoor characteristics from single images.

3.2. SmoothInv

Our approach, which we refer to as SmoothInv, first constructs a robust version of the backdoored classifier and then performs guided image synthesis towards a target class y_t. We use a simple yet effective objective to synthesize the backdoor pattern, where we minimize the standard cross entropy loss with the target class. Next we describe this robustification process and our synthesis process in details.

Robustification of Backdoored Classifiers One necessary condition for obtaining perceptually-aligned gradients is that the classifier itself must be adversarially robust. As backdoored classifiers are not robust by construction, we thus propose to use a *robustification process* to robustify backdoored classifiers. The goal we hope to achieve from this robustification process is to induce meaningful and salient gradient signal from the resulting robust classifier.

As illustrated in Figure 3, we construct such a robust classifier with the Randomized Smoothing technique [10], where we smooth the prediction of the backdoored classifiers under Gaussian noisy samples. Different from empirical robustness, randomized smoothing provides certified robustness guarantee, so we can be confident that the resulting smoothed classifier is indeed robust. We experimented with two ways of building robust smoothed classifiers. The first one is based on the recent proposed Diffusion Denoised Smoothing method [7]. Specifically, Gaussian noisy images are first processed by a denoising transformation before being fed into the classifier. The denoising transformation is a diffusion based denoiser $\mathcal{D}$.

However, on a second thought, do we really need the resulting smoothed classifier to be robust on the whole data distribution? Recall that our motivation is to elicit the salient gradients of backdoor features. We may only need the smoothed classifier to be robust on the actual backdoored images. To test this hypothesis, we remove the denoiser from the pipeline and construct the smoothed classifier directly from the backdoored classifiers. To summarize,

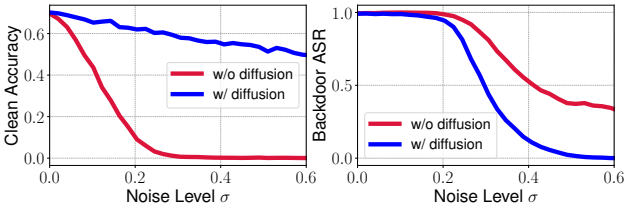


Figure 4. Clean accuracy and backdoor accuracy of the smoothed classifier at various noise levels (we use the ImageNet Blind-P as the base backdoored classifier).

we try to construct the following smoothed classifier:

$$g(x) \coloneqq \text{argmax}_c \underset{\delta\sim\mathcal{N}(0,\sigma^2\mathbf{I})}{\Pr} \left(f \circ \mathcal{T}(x+\delta) = c\right) \quad (4)$$

and we initialize the transformation operation $\mathcal{T}$ with either the diffusion model $\mathcal{D}$ as a denoiser ("*w/ diffusion*") or the identity function $\mathcal{I}$ ("*w/o diffusion*").

The smoothed classifier defined in Equation 4 is hard to evaluate in practice, as making a prediction would require calculating a probaility measure over a Gaussian distribution. In the original RS paper [10], obtaining a certificate for a single image would need evaluating over 10k Monte Carlo noisy samples on ImageNet. In this work we do not care about the exact certification bound but rather interested in the robustness property of smoothed classifiers. Thus we use a continuous approximation instead, where the soft smoothed classifier $G_b : \mathcal{X} \to P(\mathcal{Y})$ is defined by:

$$G_b(x) \coloneqq \frac{1}{N}\sum_{i=1}^{N} \mathcal{F}_b \circ \mathcal{T}(x+\delta_i),\ \delta_i \sim \mathcal{N}\left(0, \sigma^2\mathbf{I}\right) \quad (5)$$

where $\mathcal{F}_b$ is the soft version of the backdoored classifier f_b which outputs a probability distribution over classes (and where we will later choose $N = 40$, leading to a tractable approach). This approximation allows us to obtain gradients from the smoothed classifier via back-propagation. Thus from now on, we refer to G_b as the actual constructed smoothed classifier.

Last, we perform a sanity check on whether the smoothed classifier G_b remains a valid backdoored classifier after the smoothing procedure. We experimented with a backdoored classifier on ImageNet (Blind-P in Table 1). In Figure 4, we show both the clean accuracy and backdoor ASR for the smoothed classifier G_b (*w/* and *w/o diffusion*), for varying σ. We can see that the original backdoor remains effective within a reasonable range of noise level for both choices of $\mathcal{T}$. A complete sanity check for all backdoored classifiers considered in this paper is in Appendix F.

Guided Synthesis of Backdoor Patterns Starting from single images, our synthesis procedure is guided by a robust smoothed classifier, which minimizes a cross entropy of G_b:

$$\min_{p\in\Delta} -\log G_b(x+p)_{y_t} \quad (6)$$

Figure 5. Progression of synthesized images throughout the optimization iterations.

where the perturbation set is defined by a pre-defined $\Delta = \{p \,|\, \|p\|_2 \leq \epsilon\}$. We use a pertubation set to prevent finding arbitrary large perturbations. After all, we want the synthesized images $x + p$ to recover the backdoored image $\mathcal{B}(x)$, which is close to the input x. As there is a constraint on the perturbation size, we use projected gradient descent (PGD) to solve the optimization problem in Equation 6. The perturbation variable p is initialized to 0. At each step, we compute the gradient with respect to p and take a gradient step with ℓ_2 normalized gradient. We repeat this process until convergence, when the loss value becomes stable. Empirically we find that 400 iterations are sufficient for convergence. In Figure 5, we show how the backdoor patterns appears gradually as the optimization process evolves.

Target Class Identification For each possible class, we can synthesize a perturbation from Equation 6. Now we describe how we identify the target class of the backdoor, without manually inspecting the synthesized images and checking if there is an abnormal pattern. Previous methods use the size of reversed trigger to determine if a class is the target class, i.e., reversed triggers from the true target class should be much smaller than those from normal classes. However, it is not a viable strategy in our case since we are using a fixed perturbation budget ϵ in Equation 6.

Our identification process is based on an intriguing observation we make on the synthesized perturbation p. For perturbations p synthesized from the target class, we use it as an *additive* backdoor: $x' = x + p$ and find that it is a highly effective backdoor evaluated on other clean images. The same does not hold true for normal classes, where the synthesized perturbations barely transfer to other clean images. We empirically show this in Section 4.2. Thus, during backdoor inversion, SmoothInv identifies a class as a target class in a backdoored classifier when the synthesized perturbation from this class also leads to a high ASR.

4. Empirical Study

4.1. Experimental Setup

Backdoored Classifiers We carefully select backdoored classifiers from well-established backdoor attacks which satisfies the following criteria: 1) it is either published in top conferences/venues, or has become a well-known baseline in backdoor attacks; 2) it is demonstrated to be effective on standard vision recognition benchmarks (e.g. ImageNet) as this is more practical than toy datasets; 3) the collection of these backdoor attacks should cover a wide range of backdoor conditions, e.g., universal or label-specific, backdoor shape, size and location. Next, we describe the four backdoored classifiers we consider in this work, and we list the relevant statistics of these backdoored classifiers in Table 1 and show the corresponding original backdoors in Figure 2.

	TrojAI [1] Round4-131	HTBA [34]	Blind Backdoor [3] Blind-P	Blind Backdoor [3] Blind-S
Dataset	TrojAI	ImageNet	ImageNet	ImageNet
Input Size	224^2	224^2	224^2	224^2
Arch	VGG-11	AlexNet	ResNet-18	ResNet-18
#Classes	38	2	1000	1000
Clean Acc	100.00%	95.00%	69.26%	68.06%
Backdoor Statistics				
Patch	Polygon	Square	Pixel Pattern	Single Pixel
Location	Foreground	Random	Upper Left	Upper Left
#Pixels	1126	900	9	1
ℓ_2-avg	47.51	25.09	3.08	1.04
ASR	100.00%	54.00%	99.29%	79.73%

Table 1. Statistics of backdoored classifiers we obtain from previous backdoor attack methods, including relevant model information and detailed backdoor conditions. ℓ_2-avg refers to the average ℓ_2 distance between clean and backdoored images with pixel range $[0, 1]$. For TrojAI, we randomly sample a backdoored classifier (round 4 with model id 131) for analysis.

1. TrojAI Benchmark [16] consists of multiple rounds of released datasets. For each round, it consists of a mixed set of backdoored and clean classifiers. A set of clean images from test set is provided along with each classifier. The backdoor is placed on foreground objects during training. A sample backdoored image can be seen in Figure 1 middle. For our study, we randomly sample a classifier with polygon backdoor (round 4, id-00000131) and use TrojAI to reference this model, for comparison purposes with other backdoored classifiers. In our case, the polygon backdoor turns out to be label-specific, meaning that it only cause targeted classification of samples from certain classes.

2. Hidden Trigger Backdoor Attacks (HTBA) [34] is a backdoor attack method which has been shown effective on ImageNet. It uses a square patch (size 30×30) as the backdoor, which is the most common choice of backdoor in existing backdoor attack literature [7, 14, 42]. They obtain such square trigger by first drawing a random 4×4 matrix of colors and resizing it to the desired patch size. The patch backdoor is placed randomly over clean inputs. We use their public released code to train a binary backdoored classifier replicating their ImageNet result. We find that we are able to match the ASR reported in [34].

3. Blind Backdoors [3] show that it is possible to backdoor a standard ImageNet classifier with small patch backdoors. It trains two backdoored classifiers: Blind-P with a pixel pattern backdoor, and Blind-S with a single pixel backdoor. The backdoor is placed on a fixed location in the *top left* region of clean inputs. Both the pixel pattern and single pixel backdoors are drastically smaller than the backdoors used in

TrojAI challenge and HTBA. It is also shown in [3] that they can circumvent many previous backdoor defense methods, e.g. Neural Cleanse [43]. We use its public released code to train these two backdoored classifiers Blind-P and Blind-S. Our Blind-P matches the reported ASR in [3]. The obtained ASR of Blind-S model falls short of the reported number of 99% but is still fairly high (79.73%).

Evaluation Protocols and Baselines We first perform a quantitative evaluation by comparing with the following existing backdoor inversion approaches: NC [43], TopoTrigger [20] and PixelInv [39]. We also compare with a baseline PlainAdv where we replace the smoothed classifier G_b in Equation 6 with the base backdoored classifier instead. For a fair comparison, we evaluate both SmoothInv and baseline approaches under the same setting of single image backdoor inversion. Note that existing backdoor inversion methods can be easily adapted in this setting by using the single image as the support set $\mathcal{S}$ in Equation 1. For each method, we generate reversed backdoor from single clean images and report the average ASR over 10 random starting images. We also perform a qualitative evaluation of SmoothInv by visualizing the synthesized images.

For the diffusion model in SmoothInv *w/ diffusion*, we use the pretrained class unconditional 256×256 diffusion model from [2]. While this diffusion model is trained on ImageNet, we find that it is still a good denoiser for images from the TrojAI benchmark. The number of noise samples N is chosen to be 40 (later we find that 10 is enough in most cases). We use projected gradient descent to optimize our objective in Equation 6 with a total of 400 steps and step size is chosen to be $0.5 \times \epsilon/10$. Since we assume we do not know the exact backdoor (e.g. size information) beforehand, we use two values of perturbation size $\epsilon \in \{5, 10\}$ with the pixel range within $[0, 1]$. For each backdoored classifier, we construct smoothed classifiers with four values of noise levels $\{0.12, 0.25, 0.50, 1.00\}$ with a total of 8 optimization configurations. For each starting clean image, we report the synthesized backdoor with the highest ASR. We refer readers to Appendix C for resource considerations.

4.2. Quantitative Evaluation

We first perform a quantitative evaluation by measuring the average ASR of the reversed backdoors over random starting images. For SmoothInv, we use the synthesized perturbation p as an additive backdoor. The results on single image backdoor inversion are shown in Table 2. We compare the effectiveness of the reversed backdoor assuming the target class is known. We can see that previous backdoor inversion methods (NC, TopoPrior and PixelInv) all fail to produce effective backdoors in this setting. Both SmoothInv *w/ diffusion* and *w/o diffusion* find a highly effective backdoor for all cases. SmoothInv also outperforms a simplified baseline PlainAdv, suggesting that the robustification process of constructing a robust smoothed classifier is the key to the success of our approach.

	TrojAI	HTBA	Blind-P	Blind-S
True Backdoor	100.00%	54.00%	99.29%	79.73%
PlainAdv	36.00%	54.00%	84.08%	84.89%
NC [43]	12.20%	16.00%	0.00%	0.00%
TopoPrior [20]	28.40%	22.00%	0.00%	4.39%
PixelInv [39]	10.80%	24.00%	30.75%	43.17%
SmoothInv				
w/ diffusion	72.00%	83.20%	92.05%	93.90%
w/o diffusion	**88.00%**	**88.20%**	**99.50%**	**99.53%**

Table 2. Quantitative results of *Single Image Backdoor Inversion* on four backdoored classifiers. We report the average ASR of the reversed backdoor on the original backdoored classifier.

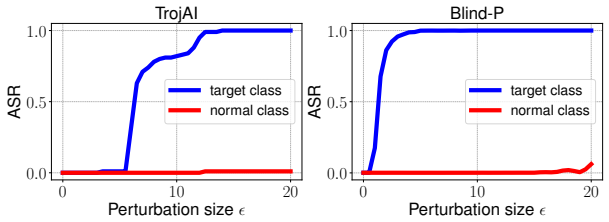


Figure 6. ASR of the SmoothInv synthesized perturbations guided by target class (blue) and normal class(red).

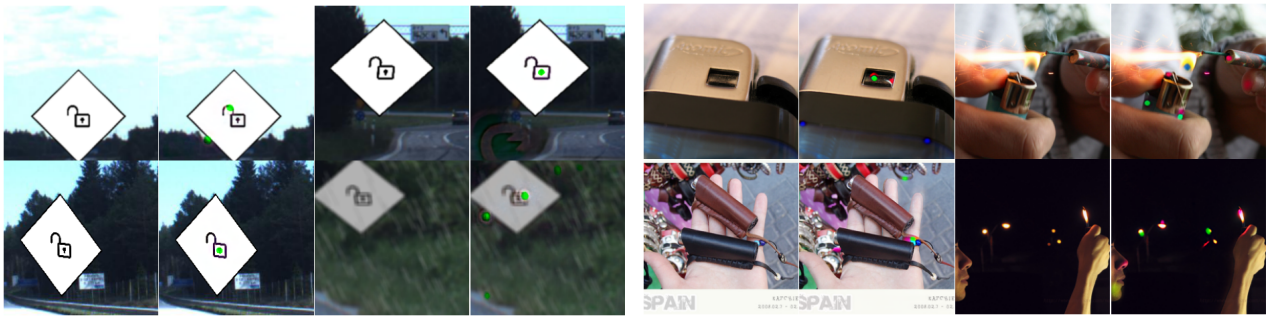
We find that SmoothInv *w/o diffusion* outperforms *w/ diffusion*. We attribute this to an observation from Figure 4, where the original backdoor are more effective for SmoothInv *w/o diffusion* than *w/ diffusion*. Thus the higher the backdoor ASR is for the smoothed classifier, the more likely it is to reconstruct a more effective backdoor with SmoothInv. This verifies our hypothesis earlier, where we do not necessarily need high clean accuracies for the smoothed classifier, but what is essential is that the backdoor remains effective after the smoothing procedure.

In Figure 6, we show the ASRs of synthesized perturbations from the target class (blue) versus normal non-targeted class (red). We can see that within a reasonble range of perturbation size, the synthesized perturbation are a valid backdoor only when it is guided from the true backdoored class (Equation 6). Using this property, we find that we can successfully identify the target class for the four backdoored classifiers we investigated, where the synthesized patterns only have high ASRs for the correct target class. This suggests the possibility of our approach to the task of backdoor detection, which we leave as a promising future work.

4.3. Qualitative Evaluation

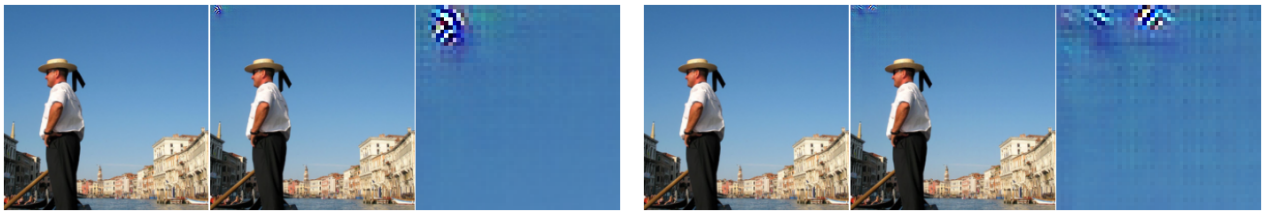
For each image, we show the synthesized patterns with the highest ASRs between SmoothInv *w/* and *w/o diffusion*.

TrojAI and HTBA. We first show results for models with relatively large backdoors. In Figure 7, we show both pairs of clean/synthesized images for the TrojAI and HTBA backdoored classifiers. For TrojAI, synthesized images all contain a concentrated region of green pixels , matching the



(a) TrojAI-R4-131 (b) HTBA

Figure 7. SmoothInv on TrojAI and HTBA backdoored classifiers ($\epsilon = 10$), where we show pairs of clean images and synthesized backdoored images (best viewed when zoomed in).



(a) Blind-P (b) Blind-S

Figure 8. Visualization results on Blind-P and Blind-S models ($\epsilon = 5$). From left to right: clean images, synthesized backdoored images by SmoothInv, *zoomed in* version (a 50×50 region in the top left) of synthesized images.

original polygon trigger in Figure 2. What's more, SmoothInv synthesizes all backdoor patterns in the foreground object, which is exactly the place where the backdoor is placed during training. For HTBA, SmoothInv synthesizes patterns in the forms of small isolated color patches, e.g. red, green and blue, while these colors are all present in the original square backdoor in Figure 2. The locations of these patterns vary across images, which could be due to the random placement of the original backdoor during training.

Blind Backdoor The results for the Blind-P and Blind-S models from blind backdoor attacks can be found in Figure 8. The synthesized images are shown in the middle. We can see that SmoothInv automatically identify the region to synthesize the backdoored patterns (in this case the top left corner), which turns out to be the exact place where the original backdoor is placed. For better comparison with the original pixel pattern and single pixel backdoors in Figure 2, we also show the *zoomed in* version of the 50×50 top left region. Though SmoothInv does not recover the exact original backdoor, the synthesized backdoor patterns have similar visual properties to the original one, e.g., a stark contrast of white pixels and neighboring pixels. Moreover, we take the perturbations from synthesized images directly as additive backdoors and find that they achieve high ASRs of 99.87% and 98.35% on the Blind-P/S models respectively.

Diverse Initial Conditions One reason for the good synthesis results in Figure 8 could be that the clean image already has a smooth background in the region of interest, i.e., top left corner, which could make the synthesis process easier. We investigate how SmoothInv is affected by the initialization of starting images. Note that for TrojAI result in Figure 7, we already show one example with rainy effects and darker background where SmoothInv still synthesizes faithful backdoor patterns. Here we analyze the Blind-P model. We select two images from ImageNet testset with non-uniform color regions (high variance in pixel values) in the top-left corner as starting images. We also use two artificial inputs: images with pure white/black pixels. We show the results of SmoothInv on these images in Figure 9.

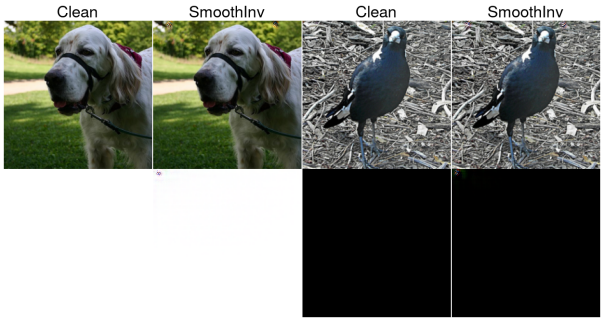


Figure 9. Synthesized images ($\epsilon = 5$) of Blind-P with diverse conditions of starting images. Clean images in the first row have non-uniform background in the top left corner, as compared to the clean image in Figure 8. The second row uses artificial inputs: images with pure white/black pixels. ASR of the four synthesized backdoor perturbations on Blind-P are 82.46/75.09/100.00/100.00%.

Figure 10. Results of SmoothInv *w/o diffusion* on the backdoored classifier Blind-G with a *gaussian backdoor* (leftmost). We show two pairs of clean and synthesized backdoored images ($\epsilon = 10$).

We can see that with various initial conditions, SmoothInv consistently synthesizes backdoor patterns in the top left region, while the synthesized perturbations itself achieve high ASRs as well. Additional visualization results are provided in Appendix D, where we also include a comparison between SmoothInv *w/* and *w/o diffusion* in Figure 13.

4.4. Mitigation of Adaptive Attacks

So far our experiments have focused on backdoor inversion on backdoored classifiers obtained from previous backdoor attacks. However, given our proposed method, someone could design new backdoor attacks to bypass our method, i.e. making it hard to extract effective backdoors with SmoothInv. The core step of SmoothInv is converting a standard backdoored classifier to a robust smoothed classifier, from which we can obtain perceptually-aligned gradients to reveal backdoor patterns. An adaptive attacker would try to circumvent SmoothInv by targeting the smoothing procedure: making the original backdoor ineffective for the smoothed classifier. Here we propose two adaptive attack attempts and show that SmoothInv is still robust in those challenging settings.

Gaussian Backdoor One can target the SmoothInv procedure by designing a backdoor which is hardly effective on the backdoored classifier after going through the smoothing process, i.e., $\mathcal{T}(\mathcal{B}(x) + \delta)$. One immediate choice is to use a backdoor with pure Gaussian noise: a gaussian backdoor $\mathcal{B}_g$. With such backdoor, the backdoor information can be obfuscated after the process $\mathcal{T}(\mathcal{B}_g(x) + \delta)$ as δ is also gaussian noise. We construct a *gaussian backdoor* of size 10×10, sampled from $\mathcal{N}(0, I)$ (see Figure 10 left). We use blind backdoor [3] to obtain a backdoored ImageNet classifier with this gaussian backdoor, which we call Blind-G. We are able to achieve an ASR of 100.00%.

On first inspection, we find that this simple *gaussian backdoor* does invade the smoothing procedure of SmoothInv *w/ diffusion*: the gaussian backdoor has an ASR of zero even for smoothed classifier constructed with noise level $\sigma = 0.12$. We attribute this to the use of diffusion denoiser $\mathcal{D}$, where the Blind-G model becomes insensitive to the diffusion denoised backdoored images $\mathcal{D}(\mathcal{B}_g(x) + \delta)$. However, we find that the gaussian backdoor still remains highly effective for smoothed classifiers (SmoothInv *w/ diffusion*) constructed purely from the Blind-G model (Equation 2), despite a high drop of clean accuracy. We then apply SmoothInv *w/o diffusion* to the Blind-G model and achieve an average ASR of $64.84\%/91.24\%$ ($\epsilon = 10/20$) for reversed backdoors from single images. We visualize some synthesized backdoor patterns in Figure 10, where a dense colorful pattern emerges in the top left region.

	Base Classifier		Smoothed ($\sigma = 0.25$)
	Clean Acc	Backdoor ASR	Backdoor ASR
Blind-P	69.26%	99.29%	94.90%
Blind-P*	67.60%	92.60%	59.60%
Blind-S	68.06%	79.73%	59.20%
Blind-S*	66.60%	45.80%	31.40%

Table 3. Effect of training-time intervention on backdoor attacks.

Training-Time Intervention One could make the backdoor ineffective for the smoothed classifier by modifying the training procedure. For SmoothInv, the backdoored classifier sees the processed images $\mathcal{T}(x + \delta)$ instead of the original image x. An adaptive attacker can intentionally make the backdoored classifier misclassify backdoored images $\mathcal{B}(x)$ while classifying $\mathcal{T}(\mathcal{B}(x) + \delta)$ correctly. To investigate if this is possible, we design a new training objective below: (we consider SmoothInv *w/o diffusion* due to resource limitations.)

$$\alpha_0 \mathcal{L}(x, y) + \alpha_1 \mathcal{L}(\mathcal{B}(x), y_t) + \alpha_2 \mathcal{L}(\mathcal{T}(\mathcal{B}(x) + \delta), y) \quad (7)$$

We use the pixel pattern and single pixel backdoors in Figure 2 and train classifiers with the new objective: Blind-P*/S* ($\alpha_0, \alpha_1, \alpha_2$ are automatically adjusted following [3]). We summarize the clean and backdoor accuracy of these models in Table 3. We can see that the base classifiers Blind-P*/S* have lower backdoor ASR compared to Blind-P/S, suggesting that correctly classifying $\mathcal{T}(\mathcal{B}(x) + \delta)$ affects the effectiveness of backdoor attacks in a negative way. We also study how training-time intervention affects the effectiveness of SmoothInv on Blind-P* (attack considered successful). We find that we are still able to synthesize effective backdoor perturbations with an average ASR of 88.81% over 10 random starting images.

5. Conclusion

In this paper, we have presented a method for backdoor inversion using a *single* clean image from the underlying data distribution. Unlike previous optimization-based approaches, our method exploits recent advances in adversarial robustness to create a smoothed version of a classifier, and then modify the image to extract the backdoor via this robust smoothed classifier. We show that SmoothInv is able to recover backdoor perturbations that are both highly successful *and* extremely visually similar to the true underlying backdoor. Going forward, the work suggests that many current approaches to producing backdoored classifiers can easily be "reverse engineered" to recover the underlying backdoor, which can provide a powerful mechanism to analyze the security of existing classifiers.

Acknowledgments. We thank Florian Tramèr for valuable discussions during the development of this work.

References

[1] https://pages.nist.gov/trojai/docs/data.html.

[2] https://github.com/openai/guided-diffusion.

[3] Eugene Bagdasaryan and Vitaly Shmatikov. Blind backdoors in deep learning models. In *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, Aug. 2021.

[4] Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *International Conference on Machine Learning*, 2012.

[5] Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. In *Pattern Recognition*, 2018.

[6] Tom B. Brown, Dandelion Mane, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv: 1712.09665*, 2017.

[7] Nicholas Carlini and Andreas Terzis. Poisoning and backdooring contrastive learning. In *International Conference on Learning Representations*, 2022.

[8] Nicholas Carlini, Florian Tramer, Krishnamurthy (Dj)Dvijotham, and J. Zico Kolter. (certified!!) adversarial robustness for free! *arXiv preprint arXiv: 2206.10550*, 2022.

[9] 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.

[10] Jeremy M Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, 2019.

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

[12] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16482–16491, October 2021.

[13] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv: 1406.2661*, 2014.

[14] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. In *Proceedings of the NIPS Workshop on Mach. Learn. and Comp. Sec*, 2017.

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

[16] David Gunning, Eric Vorm, Jennifer Yunyan Wang, and Matt Turek. Darpa's explainable ai (xai) program: A retrospective. In *Applied AI Letters*, 2021.

[17] Wenbo Guo, Lun Wang, Xinyu Xing, Min Du, and Dawn Song. Abs: Scanning neural networks for back-doors by artificial brain stimulation. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019.

[18] Wenbo Guo, Lun Wang, Xinyu Xing, Min Du, and Dawn Song. Tabor: A highly accurate approach to inspecting and restoring trojan backdoors in ai systems. In *20th IEEE International Conference on Data Mining*, 2019.

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020.

[20] Xiaoling Hu, Xiao Lin, Michael Cogswell, Yi Yao, Susmit Jha, and Chao Chen. Trigger hunting with a topological prior for trojan detection. In *International Conference on Learning Representations*, 2022.

[21] Todd Huster and Emmanuel Ekwedike. Top: Backdoor detection in neural networks via transferability of perturbation. *arXiv preprint arXiv: 2103.10274*, 2021.

[22] Simran Kaur, Jeremy Cohen, and Zachary C. Lipton. Are perceptually-aligned gradients a general property of robust classifiers? *arXiv preprint arXiv:1910.08640*, 2019.

[23] 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*, 2020.

[24] Yingqi Liu, Guangyu Shen, Guanhong Tao, Zhenting Wang, Shiqing Ma, and Xiangyu Zhang. Complex backdoor detection by symmetric feature differencing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15003–15013, June 2022.

[25] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[26] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv: 1411.1784*, 1411.

[27] Anh Nguyen and Anh Tran. Wanet – imperceptible warping-based backdoor attack. In *International Conference on Learning Representations*, 2021.

[28] Xudong Pan, Mi Zhang, Beina Sheng, Jiaming Zhu, and Min Yang. Hidden trigger backdoor attack on nlp models via linguistic style manipulation. In *31th USENIX Security Symposium (USENIX Security 22)*, 2022.

[29] 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 (CVPR)*, pages 13347–13357, June 2022.

[30] 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 (CVPR)*, pages 13347–13357, June 2022.

[31] Ximing Qiao, Yukun Yang, and Hai Li. Deepinspect: A black-box trojan detection and mitigation framework for deep neural networks. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019.

[32] Ximing Qiao, Yukun Yang, and Hai Li. Defending neural backdoors via generative distribution modeling. *Advances in Neural Information Processing Systems*, 2019.

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

[34] Aniruddha Saha, Akshayvarun Subraymanya, and Pirsiavash Hamed. Hidden trigger backdoor attacks. In *AAAI Conference on Artificial Intelligence*, 2020.

[35] Aniruddha Saha, Ajinkya Tejankar, Soroush Abbasi Koohpayegani, and Hamed Pirsiavash. Backdoor attacks on self-supervised learning. In *2022 Conference on Computer Vision and Pattern Recognition (CVPR 2022)*, 2022.

[36] Hadi Salman, Mingjie Sun, Greg Yang, Ashish Kapoor, and J. Zico Kolter. Denoised smoothing: A provable defense for pretrained classifiers. In *Advances in Neural Information Processing Systems*, 2020.

[37] Shibani Santurkar, Dimitris Tsipras, Brandon Tran, Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Image synthesis with a single (robust) classifier. *Advances in Neural Information Processing Systems*, 2019.

[38] 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 *International Conference on Machine Learning*, 2021.

[39] Guanhong Tao, Guangyu Shen, Yingqi Liu, Shengwei An, Qiuling Xu, Shiqing Ma, Pan Li, and Xiangyu Zhang. Better trigger inversion optimization in backdoor scanning. In *2022 Conference on Computer Vision and Pattern Recognition (CVPR 2022)*, 2022.

[40] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019.

[41] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *International Conference on Learning Representations*, 2019.

[42] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Clean-label backdoor attacks, 2019.

[43] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE Symposium on Security and Privacy (SP)*.

[44] Ren Wang, Gaoyuan Zhang, Sijia Liu, Pin-Yu Chen, Jinjun Xiong, and Meng Wang. Practical detection of trojan neural networks: Data-limited and data-free cases. In *European Conference on Computer Vision*, 2020.

[45] Emily Wenger, Josephine Passananti, Arjun Nitin Bhagoji, Yuanshun Yao, Haitao Zheng, and Ben Y. Zhao. Backdoor attacks against deep learning systems in the physical world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6206–6215, June 2021.

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

[47] Wenkai Yang, Yankai Lin, Peng Li, Jie Zhou, and Xu Sun. Rethinking stealthiness of backdoor attack against nlp models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 2021.

[48] 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 (CVPR)*, pages 15213–15222, June 2022.

[49] Yao Zhu, Jiacheng Ma, Jiacheng Sun, Zewei Chen, Rongxin Jiang, Yaowu Chen, and Zhenguo Li. Towards understanding the generative capability of adversarially robust classifiers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7728–7737, October 2021.