Invisible Backdoor Attack against Self-supervised Learning

Hanrong Zhang[1*] Zhenting Wang[2*] Boheng Li[3] Fulin Lin[1] Tingxu Han[4]
Mingyu Jin[2] Chenlu Zhan[1] Mengnan Du[5] Hongwei Wang[1†] Shiqing Ma[6]
[1] Zhejiang University [2] Rutgers University [3] Nanyang Technological University
[4] Nanjing University [5] New Jersey Institute of Technology [6] University of Massachusetts Amherst

Abstract

Self-supervised learning (SSL) models are vulnerable to backdoor attacks. Existing backdoor attacks that are effective in SSL often involve noticeable triggers, like colored patches or visible noise, which are vulnerable to human inspection. This paper proposes an imperceptible and effective backdoor attack against self-supervised models. We first find that existing imperceptible triggers designed for supervised learning are less effective in compromising self-supervised models. We then identify this ineffectiveness is attributed to the overlap in distributions between the backdoor and augmented samples used in SSL. Building on this insight, we design an attack using optimized triggers disentangled with the augmented transformation in the SSL, while remaining imperceptible to human vision. Experiments on five datasets and six SSL algorithms demonstrate our attack is highly effective and stealthy. It also has strong resistance to existing backdoor defenses. Our code can be found at https://github.com/Zhang-Henry/INACTIVE.

1. Introduction

In recent years, Self-Supervised Learning (SSL) has become a powerful approach in deep learning, enabling the learning of rich representations from vast unlabeled data, thus avoiding manual labeling. SSL aims to develop an image encoder that produces similar embeddings for similar images by applying various *augmentations* to the same image. This pre-trained encoder can be used for different downstream tasks by training compact downstream classifiers with relatively few parameters.

Although SSL has been extensively used in the development of foundational models [5, 6, 20], it is at risk of backdoor attacks [27, 34, 56, 68], where the attacker embeds hidden malicious behavior within the encoder. The backdoor can be inherited to the downstream task. The downstream classifier predicts a specific target label if the input contains a pre-defined backdoor trigger. Existing backdoor attacks on SSL such as BadEncoder [27] achieve high attack success rates (ASR). However, *a common drawback of these effective attacks is that their trigger patterns are obvious, making them susceptible to human inspection.* Moreover, while data-poisoning-based attacks CTRL [34] and BLTO [63] are relatively stealthy, their ASRs are suboptimal, For example, on CIFAR10 CTRL only has 61.90% ASR under BYOL framework and BLTO only has 84.63% ASR under SimSiam framework. Furthermore, they also rely on the downstream dataset matching the pre-training dataset distribution, limiting effectiveness across diverse datasets. In this paper, we aim to propose a backdoor attack in SSL that is both effective and stealthy to human vision without this distribution dependency.

There are various invisible triggers designed for the backdoor attacks on supervised classifiers, such as WaNet [47], ISSBA [36], and filter attack [42]. A straightforward way to achieve imperceptible backdoor attacks in SSL is by directly applying these invisible triggers. However, these existing invisible triggers designed for supervised learning do not perform as well in attacking self-supervised models (see Fig. 2). We then find that this lack of effectiveness is due to the overlapping distributions between the backdoor samples and the augmented samples utilized in SSL. Namely, self-supervised models cannot effectively distinguish the distribution of the backdoor samples and the augmented samples, due to the similarity between the transformation altered by the backdoor trigger and intrinsic image augmentations in SSL, such as RandomGrayscale and ColorJitter (see Fig. 3).

Based on the above observations, we developed a backdoor attack that disentangles its optimized trigger transformation and the augmented transformation in SSL. In detail, it involves increasing the distributional distance between backdoor samples and the augmented samples in the SSL process. We also keep the trigger stealthy by adding the constraints on both pixel-space and feature-space distance

*Equal Contribution. Email: zhanghr0709@gmail.com; zhenting.wang@rutgers.edu
†Corresponding author. Email: hongweiwang@intl.zju.edu.cn

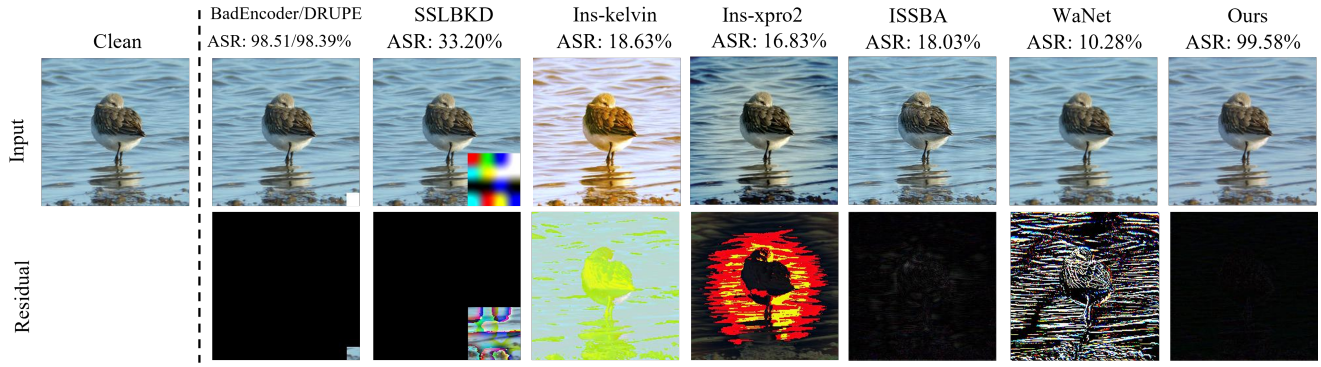

Figure 1. Comparison of clean, backdoored samples created by Patch trigger used by BadEncoder [27] and DRUPE [68], Instagram filter trigger [30], ISSBA trigger [36], WaNet trigger [47] and ours. Except for DRUPE, the ASRs are tested under the threat model of BadEncoder. Residuals are the difference between clean and backdoored images. *Our method achieves the highest ASR while maintaining trigger stealthiness, while other methods either have a much lower ASR or use more easily detectable triggers.*

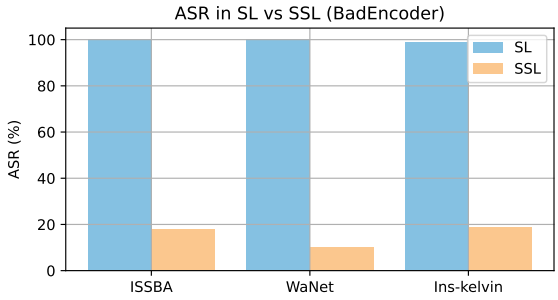

Figure 2. Existing imperceptible backdoor triggers, which yield high ASR in supervised learning (SL), do not perform as effectively in SSL. The attack framework for SL and SSL are standard backdoor poisoning [17] and BadEncoder [27], respectively.

to the original samples, using metrics like SSIM, PSNR, perceptual loss, and Wasserstein distance. We then implemented our prototype INACTIVE (**IN**visible b**AC**kdoor a**T**tack **I**n self-super**V**ised l**E**arning) and tested it on five datasets (i.e., CIFAR10, STL10, GTSRB, SVHN, ImageNet), and across six classic SSL frameworks (i.e., SimCLR [6], MoCo [20], BYOL [16], SimSiam [7], SwAV [4] and CLIP [53] (See Tab. 1 and Tab. A1) with their various augmentation transforms (See Tab. A5). The results demonstrate that our method is highly effective and stealthy. In detail, it achieves an average of 99.09% ASR (See Tab. 1), 0.9763 SSIM, 41.07 PSNR, 0.0046 LIPIS, 0.9751 FSIM, and 13.281 FID (See Tab. A9). As shown in Fig. 1, we compare several methods' backdoor residuals and ASRs. Our method exhibits the highest ASR while maintaining the highest stealthiness. It also effectively bypasses existing backdoor defenses such as DECREE [13], Beatrix [44], ASSET [48], STRIP [14], Grad-CAM [58], Neural Cleanse [72], and various noise, i.e., JPEG compression, Poisson noise, and Salt&Pepper noise.

Our contributions are summarized as follows: ① We observed that existing imperceptible triggers designed for supervised classifiers have limited effectiveness in SSL. ② We find that the reason behind such ineffectiveness is the coupling feature-space distributions for the backdoor samples and augmented samples in the SSL models. ③ Based on our findings, we propose an imperceptible and effective backdoor attack in SSL by disentangling the distribution of backdoor samples and augmented samples in SSL, while constraining the stealthiness of the triggers during the optimization process. ④ Extensive experiments on five datasets and six SSL algorithms with different augmentation ways demonstrate our attack is effective and stealthy, and can also be resilient to current SOTA backdoor defense methods.

2. Related Work

2.1. Self-Supervised Learning

The goal of SSL is to leverage a large amount of unlabeled data in the pre-training dataset to pre-train an image encoder, which can then be used to create classifiers for various downstream tasks with a smaller set of labeled data [68]. SSL pipelines for contrastive learning typically include the following approaches [18, 26]: ① Negative Examples: Promotes proximity among positive examples while maximizing the distance between negative examples in the latent space, as seen in SimCLRs [6] and MoCo [20]. ② Self-distillation: Utilizes two identical Siamese networks with different weights to increase the similarity between differently augmented versions of the same image, such as in BYOL [16] and SimSiam [7]. ③ Clustering: Implements a clustering mechanism with swapped prediction of representations from both encoders, as in SwAV [4]. *Our method is shown to be highly effective and stealthy under the SSL algorithms in Tab. 1 and Tab. A1.*

2.2. Backdoor Attacks

Backdoor attacks were initially proposed for supervised learning (SL) to modify a model's behavior on specific in-

puts or classes while keeping its general performance intact [2, 15, 33, 45, 51, 67, 77, 84, 85]. Early backdoor attacks commonly utilized visible triggers like distinctive patches that are easily detectable through visual inspection [17, 59, 64]. To enhance stealth, subsequent research introduced invisible triggers, which are subtle and blend into the background, helping these attacks evade both human inspection and certain automated defenses [36, 47].

Since many of these attacks rely on labeled data, recent studies have explored alternative backdoor implantation techniques in SSL models [3, 27, 39, 56, 63]. However, these typically use a visible backdoor trigger, such as a patch, making them prone to human detection and model simulation. The advantage of invisible triggers in SSL is clear: they improve attack stealth, bypassing some conventional defenses that focus on detecting visible anomalies [13, 69]. However, as we will demonstrate, directly applying existing invisible triggers designed for SL to SSL tasks results in limited attack effectiveness. Moreover, backdoor attacks in SSL are generally divided into two types: training-time backdoor injection attacks like BadEncoder [27], which require control of the training in the backdoor injection process, and data-poisoning-based attacks like CTRL [34] and BLTO [63], which rely on poisoned data without needing model specifics. BadEncoder modifies a pre-trained encoder to embed triggers that cause targeted misclassifications in downstream tasks. By aligning the features of triggered images with an attacker-chosen class, downstream classifiers misclassify triggered inputs while maintaining accuracy on clean data. Although CTRL and BLTO are relatively stealthy, they achieve lower ASRs than our method and depend on the downstream dataset matching the pre-training distribution, limiting their versatility across diverse datasets. In this paper, we focus on training-time backdoor injection attack due to it has higher attack effectiveness and transferability.

2.3. Backdoor Defenses

Various defenses have been developed against backdoor attacks [8, 23, 32, 43, 49, 50, 52, 76, 78, 81, 83], primarily targeting supervised classifiers. These defenses either prevent attacks during training [22, 24, 70, 74] or detect and mitigate backdoors in compromised models offline [37, 42, 60, 65, 66, 75, 80, 89]. Some methods also detect backdoor-triggered inputs during inference [14]. Defense methods like DECREE [13], Beatrix [44], and ASSET [48] are designed for SSL, primarily relying on the visible characteristics of triggers to detect backdoors. In contrast, our method uses invisible triggers, effectively bypassing these defenses by breaking their reliance on visual anomalies and making detection more challenging.

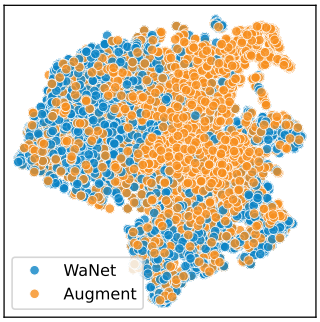

Figure 3. t-SNE visualization of the feature space in the inherent augmentation and backdoor trigger space. The SimCLR [6] pretrained model struggled to differentiate between backdoor samples injected with the WaNet trigger [47] and the augmented samples within the SimCLR contrastive learning framework.

3. Observations and Analysis

Invisible Trigger Designed for SL Fails in SSL. We first assess the effectiveness of existing invisible triggers (WaNet, ISSBA, and filter attack) designed for supervised classifiers. Fig. 2 displays their ASRs on both supervised classifiers and self-supervised models. For supervised classifiers, the standard backdoor poisoning method [17] is used. For self-supervised models, we apply the BadEncoder method, replacing the patch trigger with these invisible triggers, using ResNet18. We find that these triggers, which achieve high ASR in supervised learning, are less effective in SSL. We then investigate the underlying reason for these results and focus on the following research question: *Why does the effective invisible backdoor trigger designed for supervised learning fail on self-supervised learning?*

Cause of the Failure: Entanglement of the Inherent Augmentations and Backdoor Trigger. We find that the entanglement of the inherent augmentations in contrastive learning can cause the failure of the backdoor injection with such triggers. We provide our analysis in this section. One of the core training losses of contrastive learning can be formulated as maximizing the feature space similarity between the augmented samples modified from the same training samples:

$$\arg\max_{\theta_{\mathcal{F}}} s(\mathcal{F}_\theta(\mathcal{A}_1(\boldsymbol{x})), \mathcal{F}_\theta(\mathcal{A}_2(\boldsymbol{x}))) \tag{1}$$

where $s(\cdot,\cdot)$ denotes the similarity measurement, $\mathcal{F}_\theta$ is the encoder in training ($\theta_{\mathcal{F}}$ is its parameters), $\boldsymbol{x}$ is the training sample, $\mathcal{A}_1$ and $\mathcal{A}_2$ are different augmentations sampled from the *predefined augmentation space* $\mathcal{S}_{\mathcal{A}}$. Different from the predefined augmentation space, we also define the *learned augmentation space* for trained encoders $\mathcal{S}'_{\mathcal{A}}$ as the space including a set of transformations where any pair within it can achieve high pairwise similarity on augmented versions of the same sample when processed by the trained

encoder. We also define *perfectly-trained encoder* is the encoder that achieves maximal similarity described in Eq. 1 for all samples and all possible transformations used, *i.e.*, $s(\mathcal{F}_\theta(\mathcal{A}_1(\boldsymbol{x})), \mathcal{F}_\theta(\mathcal{A}_2(\boldsymbol{x}))) = 1, \forall \boldsymbol{x} \in \mathcal{X}, \forall \mathcal{A}_1, \mathcal{A}_2 \in \mathcal{S}_\mathcal{A}$, where $\mathcal{X}$ is the input space. Based on this, we have the following theorem:

Theorem 3.1. *Given a perfectly-trained encoder $\mathcal{F}_\theta$ based on the augmentations sampled from predefined augmentation space $\mathcal{S}_\mathcal{A}$, it is impossible to inject a backdoor with trigger function $\mathcal{I} \in \mathcal{S}_\mathcal{A}$.*

The proof of this theorem can be found in the Sec. A4. In practice, the boundary of the learned augmentation space for trained encoders $\mathcal{S}'_\mathcal{A}$ is often imprecise, and it potentially reflects a relaxation of the predefined augmentation space. Consequently, using trigger functions that are not precisely within the predefined augmentation space $\mathcal{S}_\mathcal{A}$ but are instead distributionally close to it can also make achieving high attack success rates hard.

Empirical Evidence. We also conduct experiments to confirm the invisible triggers designed for supervised learning are actually entangled with the inherent augmentations in self-supervised learning. Specifically, we use a ResNet18 pre-trained with SimCLR for a binary classification task to differentiate between samples poisoned by WaNet and those augmented by SimCLR. We ensure consistent feature representations for clean samples between the backdoored and clean models using utility loss from Jia et al [29]. Results indicate that the models struggle to differentiate between the two categories. A t-SNE visualization of their features, as presented in Fig. 3, indicates a significant overlap and entanglement. From this, we infer that the diminished effectiveness of supervised backdoor attack methods in SSL scenarios is attributed to the distributional similarity between the features of contrastive-learning-augmented samples and backdoor samples. The reason for the entanglement phenomenon on the invisible backdoor trigger is that the learned augmentation space reflects a relaxation of the predefined augmentation space, and such relaxation covers most of the invisible minor transformations. Thus, we aim to search for invisible transformations that can escape the inherent augmentation space.

4. Attack Design

4.1. Threat Model

We follow the well-defined training-based backdoor injection threat model introduced in BadEncoder [27].

Objectives of the Attacker. The objective of an attacker is to implant backdoors into a pre-trained image encoder by SSL. We define a backdoored image encoder model as $\mathcal{F}_\theta$ and the backdoor injector as $\mathcal{I}_\phi$. In this way, a downstream classifier trained based on $\mathcal{F}_\theta$, which we define as $\mathcal{C}_\epsilon$, could produce a specific prediction c designated by the attacker for inputs $\boldsymbol{x}$ implanted with a trigger chosen by the attacker. The formal definition is shown as follows. y here means the correct label of the input $\boldsymbol{x}$.

$$\mathcal{C}_\epsilon(\mathcal{F}_\theta(\boldsymbol{x})) = y, \quad \mathcal{C}_\epsilon(\mathcal{F}_\theta(\mathcal{I}_\phi(\boldsymbol{x}))) = c \tag{2}$$

The attacker's goal is to modify a clean image encoder to create a backdoored version that meets two key objectives: ① *Effectiveness*: The backdoored model should maintain a high attack success rate while preserving accuracy in benign conditions, keeping backdoored accuracy close to clean accuracy for downstream classifiers. ② *Naturalness and Stealthiness*: The triggered sample should appear authentic and natural to avoid detection by human inspection.

Attacker's Knowledge and Capabilities. Following BadEncoder [27], we assume that the attacker has access to a pre-trained clean image encoder and the attacker has full knowledge about the pre-trained encoder, such as the SSL method and the detailed contrastive augmentation operation used in pre-training. Additionally, it is presumed that the attacker can access a collection of unlabeled images, referred to as *shadow dataset*. The attacker is also assumed to have access to a few images from the Internet, called *reference inputs*, for each combination of a target downstream task and a target class. We assume that the attacker can manipulate the training procedure to create an encoder with embedded backdoors. Accordingly, the attacker also has access to the augmentation transforms used to pre-train the encoder, which can be utilized in INACTIVE to generate stealthy and effective backdoor triggers. However, we assume that the attacker cannot interfere with the training process of these downstream classifiers, such as the training dataset, model framework, and weights. Unlike data-poisoning-based methods such as CTRL [34] and BLTO [63], our approach does not rely on matching distributions between pre-training and downstream datasets, allowing for broader applicability without interfering in downstream training data, models, or weights.

4.2. Overarching Idea

According to the previous observations, the key to enhancing the ASR in SSL is to disentangle the two overlapping distributions of the backdoor and augmentation transformation in the contrastive learning of the SSL's pre-training stage. Gray Scaling and Color Jittering are necessary augmentations used in the self-supervised learning and most of existing SSL methods (e.g., SimCLR [6], MoCo [20], SimSiam [7] and BYOL [16]) use them. A detailed summary of the augmentation operations in different mainstream SSL methods can be found in Tab. A5. Since these augmentations primarily alter the color semantics of inputs, HSV and HSL color spaces serve as ideal input spaces for capturing and enlarging these effects [28]. We aim to identify

a trigger that escapes the inherent augmentation space of self-supervised learning by increasing the distance between backdoored samples and non-backdoored samples within the HSV and HSL color spaces. Since *we already know the augmentation ways in the pre-training stage*, thus we design $\mathcal{L}_{\text{disentangle}}$ to quantify the distributional gap between images in the two batches, which involves measuring the difference in color characteristics. To further enlarge the distributional gap, we design $\mathcal{L}_{\text{alignment}}$ to pull close the features of backdoor images and reference images. Moreover, while we try to expand the distributional difference between the backdoor trigger and the augmentation transformation, an excessively large gap might result in a significant divergence between the backdoored image and the original one. This could, in turn, diminish the naturalness of the backdoored image and reduce the stealthiness of the trigger. Hence we design $\mathcal{L}_{\text{stealthy}}$ to blend the backdoor seamlessly with the original image.

4.3. Our Approach: INACTIVE

In our context, we refer to a clean pre-trained image encoder and its backdoor-injected one as $\mathcal{F}_\theta$ and $\mathcal{F}'_\theta$. Given any pair of a downstream task and its corresponding target class, labeled as (T_i, y_i), the attacker gathers a collection of reference inputs denoted by $R_i = \{x_{i1}, x_{i2}, \ldots, x_{ir_i}\}$ from the specified target class y_i, where r_i represents the number of reference inputs for (T_i, y_i), with $i = 1, 2, \cdots, t$. Moreover, for each pair (T_i, y_i), the attacker chooses a trigger $\boldsymbol{e}_i$ to implant into samples in the shadow dataset $\mathcal{D}_s$. We denote a clean input $\boldsymbol{x}$ embedded with a trigger as $\boldsymbol{x}'$, which is called a backdoored input.

Enhancing Distributional Gap Between Backdoor Images and Augmented Images. To effectively enlarge the distributional distinction between the backdoor and the augmentation in SSL, we design $\mathcal{L}_{\text{disentangle}}$ for scenarios where augmentation transformations might weaken or obscure the pattern of backdoor triggers, leading to a decrease in the ASR. Utilizing $\mathcal{L}_{\text{disentangle}}$ ensures that the distinctiveness of the backdoor is maintained even in the face of various image transformations. The disentangle loss is defined as follows.

$$\begin{aligned}\mathcal{L}_{\text{disentangle}} = \\ -\frac{1}{|\mathcal{D}_s|}\sum_{\boldsymbol{x}\in D_s} & \|H(\boldsymbol{x}') - H(\tilde{\boldsymbol{x}})\|_2 + \|S(\boldsymbol{x}') - S(\tilde{\boldsymbol{x}})\|_2 \\ + & \|V(\boldsymbol{x}') - V(\tilde{\boldsymbol{x}})\|_2 + \|L(\boldsymbol{x}') - L(\tilde{\boldsymbol{x}})\|_2,\end{aligned} \tag{3}$$

where H,S,V,L denote Hue, Saturation, Value, and Lightness from HSV and HSL color spaces. We denote an input $\boldsymbol{x}$ augmented by the transformations used in the encoder's pre-training stage as $\tilde{\boldsymbol{x}}$. $|\mathcal{D}_s|$ denotes the sample number in the shadow dataset. $\|u - v\|_2$ denotes the ℓ_2 distance between sample u and sample v.

Feature Alignment Between Backdoored and Reference Images. Following BadEncoder, we enhance the backdoor attack effectiveness by making the compromised image encoder outputs similar feature embeddings for any sample injected with backdoor $\boldsymbol{x}'$ in the shadow dataset $\mathcal{D}_s$ and the reference inputs $\mathcal{R}_i$ of a pair (T_i, y_i). Consequently, a compromised downstream classifier developed from our compromised image encoder is inclined to assign identical labels to both reference samples $\mathcal{R}_i$ and to any compromised sample $\boldsymbol{x}'$. We call this process feature alignment between backdoored and reference images, and the $\mathcal{L}_{\text{alignment}}$ is defined as follows.

$$\mathcal{L}_{\text{alignment}} = -\frac{\sum_{i=1}^{t}\sum_{j=1}^{r_i}\sum_{\boldsymbol{x}\in\mathcal{D}_s} s\left(\mathcal{F}'_\theta(\boldsymbol{x}'), \mathcal{F}'_\theta(\boldsymbol{x}_{ij})\right)}{|\mathcal{D}_s| \cdot \sum_{i=1}^{t} r_i}, \tag{4}$$

where $s(\cdot,\cdot)$ is used to quantify the degree of similarity, for instance, cosine similarity, between a pair of feature embeddings. The term $|D_s|$ denotes the count of samples within the shadow dataset, and the denominators serve the purpose of standardizing the losses.

Preserving Covert and Natural Backdoors with Expanded Distributional Gaps. We employ several metrics that measure the similarity between the backdoored image and the original one in both pixel and feature space to ensure that our trigger remains both natural and inconspicuous. To assess similarity in pixel space, we use SSIM and PSNR. Meanwhile, for high-level feature space comparisons, we first use LPIPS which better reflects the subjective experience of image quality and similarity. Following [82] and [68], we also use Wasserstein distance [71] (WD) to reduce the distributional disparity between backdoored and clean samples. $\mathcal{L}_{\text{stealthy}}$ is defined as follows:

$$\begin{aligned}\mathcal{L}_{\text{stealthy}} = \\ \sum_{\boldsymbol{x}\in D_s} & \lambda_1 \cdot \text{LPIPS}(\boldsymbol{x}', \boldsymbol{x}) + \text{WD}(\mathcal{M}(\boldsymbol{x}'), \mathcal{M}(\boldsymbol{x})) \\ - & \lambda_2 \cdot \text{PSNR}(\boldsymbol{x}', \boldsymbol{x}) - \text{SSIM}(\boldsymbol{x}', \boldsymbol{x}),\end{aligned} \tag{5}$$

where λ_1, λ_2 are used to scale different loss terms to the same scale from 0 to 1.

Optimization Problem Formulation and Algorithm. We have defined $\mathcal{L}_{\text{disentangle}}$, $\mathcal{L}_{\text{stealthy}}$, $\mathcal{L}_{\text{alignment}}$ in the sections above. Then we can define our INACTIVE as an optimization problem. Concretely, our backdoor trigger injector $\mathcal{I}_\phi$ is a solution to the subsequent optimization problem:

$$\min_{\theta_{\mathcal{F}}} \min_{\mathcal{I}_\phi} \mathcal{L} = \mathcal{L}_{\text{stealthy}} + \alpha \cdot \mathcal{L}_{\text{disentangle}} + \beta \cdot \mathcal{L}_{\text{alignment}}, \tag{6}$$

where α and β serve as hyper-parameters to provide equilibrium among these three loss components. We adopt Alg. 1 to solve the optimization problem, where we alternatively optimize the backdoor injector and the compromised image encoder and output the final backdoored encoder $\mathcal{F}_\theta$ and backdoor trigger injector $\mathcal{I}_\phi$. Additionally, to speed up the optimization process and promote the backdoor attack efficacy, we adopt Alg. 2 to pre-train a backdoor injector

Algorithm 1 Our backdoor attack INACTIVE

Input: Pre-trained clean encoder $\mathcal{F}_\theta^*$, shadow dataset $\mathcal{D}_s$, reference input set $\mathcal{R}$
Output: Backdoored encoder $\mathcal{F}_\theta$, backdoor trigger injector $\mathcal{I}_\phi$

function OURS($\mathcal{F}_\theta^*$, $\mathcal{D}_s$, $\mathcal{R}$)
  $\mathcal{F}_\theta \leftarrow \mathcal{F}_\theta^*$; $\hat{\mathcal{D}}_s \leftarrow$ Augment samples in $\mathcal{D}_s$;
  $\mathcal{I}_\phi \leftarrow$ a pre-trained backdoor injector $\mathcal{I}_\phi^*$ using Alg. 2
  for iter $= 0$ **to** max_epochs **do**
    $\mathcal{D}_s{}' \leftarrow \mathcal{I}_\phi(\mathcal{D}_s)$
    $\mathcal{L}_{\text{disentangle}} \leftarrow$ distribution difference between backdoor images $\boldsymbol{x}'$ and augmented images $\tilde{\boldsymbol{x}}, \forall \boldsymbol{x}' \in \mathcal{D}_s{}', \forall \tilde{\boldsymbol{x}} \in \hat{\mathcal{D}}_s$ ▷ Eq. 3
    $\mathcal{L}_{\text{stealthy}} \leftarrow$ distance between backdoor image $\boldsymbol{x}'$ and clean image $\boldsymbol{x}$, $\forall \boldsymbol{x}' \in \mathcal{D}_s{}', \forall \boldsymbol{x} \in \mathcal{D}_s$
    ▷ Eq. 5
    $\mathcal{L}_{\text{alignment}} = -\frac{\sum_{i=1}^{t}\sum_{j=1}^{r_i}\sum_{\boldsymbol{x}\in\mathcal{D}_s} s(\mathcal{F}_\theta(\boldsymbol{x}'),\mathcal{F}_\theta(\boldsymbol{x}_{ij}))}{|\mathcal{D}_s|\cdot\sum_{i=1}^{t} r_i}$ ▷ Eq. 4
    $\mathcal{L}_{\text{injector}} = \mathcal{L}_{\text{stealthy}} + \alpha \cdot \mathcal{L}_{\text{disentangle}} + \beta \cdot \mathcal{L}_{\text{alignment}}$ ▷ Eq. 6
    $\phi_{\mathcal{I}} = \phi_{\mathcal{I}} - lr_1 \cdot \frac{\partial \mathcal{L}_{\text{injector}}}{\partial \phi_{\mathcal{I}}}$
    $\mathcal{L}_{\text{consistency}} = -\frac{\sum_{i=1}^{t}\sum_{j=1}^{r_i} s(\mathcal{F}'_\theta(\boldsymbol{x}_{ij}),\mathcal{F}_\theta(\boldsymbol{x}_{ij}))}{\sum_{i=1}^{t} r_i}$ ▷ Eq. A1
    $\mathcal{L}_{\text{utility}} = -\frac{1}{|\mathcal{D}_s|}\cdot\sum_{\boldsymbol{x}\in\mathcal{D}_s} s\left(\mathcal{F}'_\theta(\boldsymbol{x}), \mathcal{F}_\theta(\boldsymbol{x})\right)$ ▷ Eq. A2
    $\mathcal{L}_{\text{encoder}} = \mathcal{L}_{\text{alignment}} + \mathcal{L}_{\text{consistency}} + \mathcal{L}_{\text{utility}}$
    $\theta_{\mathcal{F}} = \theta_{\mathcal{F}} - lr_2 \cdot \frac{\partial \mathcal{L}_{\text{encoder}}}{\partial \theta_{\mathcal{F}}}$

Algorithm 2 Pre-training backdoor injector

Input: Shadow dataset $\mathcal{D}_s$ **Output:** Pre-trained backdoor injector $\mathcal{I}_\phi$

function PRE-TRAINING INJECTOR($\mathcal{D}_s$)
  $\hat{\mathcal{D}}_s \leftarrow$ Augment samples in $\mathcal{D}_s$; $\mathcal{I}_\phi \leftarrow$ Random initialization
  for iter $= 0$ **to** max_epochs **do**
    $\mathcal{D}_s{}' \leftarrow \mathcal{I}_\phi(\mathcal{D}_s)$
    $\mathcal{L}_{\text{disentangle}} \leftarrow$ distribution difference between backdoor images $\boldsymbol{x}'$ and augmented images $\tilde{\boldsymbol{x}}, \forall \boldsymbol{x}' \in \mathcal{D}_s{}', \forall \tilde{\boldsymbol{x}} \in \hat{\mathcal{D}}_s$ ▷ Eq. 3
    $\mathcal{L}_{\text{stealthy}} \leftarrow$ distance between backdoor image $\boldsymbol{x}'$ and clean image $\boldsymbol{x}$, $\forall \boldsymbol{x}' \in \mathcal{D}_s{}', \forall \boldsymbol{x} \in \mathcal{D}_s$ ▷ Eq. 5
    $\mathcal{L}_{\text{ours}} = \mathcal{L}_{\text{stealthy}} + \mu \cdot \mathcal{L}_{\text{disentangle}}$
    $\phi_{\mathcal{I}} = \phi_{\mathcal{I}} - lr \cdot \frac{\partial \mathcal{L}_{\text{ours}}}{\partial \phi_{\mathcal{I}}}$

to initialize the injector in Alg. 1. We use the U-Net architecture [54] for the backdoor injector, as shown in the Tab. A12.

5. Evaluation

We first evaluate the effectiveness and stealthiness of INACTIVE using four datasets, followed by an assessment of its robustness against various backdoor defenses and noises. To demonstrate generalization, we conduct additional attacks on various SSL algorithms and a multi-modal model with different augmentations, detailed in Sec. A1 and Sec. A2. Sec. A3 further validates each component's role. Sec. A5 examines parameter sensitivity and performance.

5.1. Experimental Setup

Datasets. We utilize four image datasets, i.e. CIFAR10 [31], STL10 [9], GTSRB [62], SVHN [46] and ImageNet [55] to evaluate our method, which are also frequently used in backdoor attacks research [27, 47]. More details are introduced in Sec. A6.5.

Evaluation Metrics. To assess the effectiveness of our method, we employ three metrics following existing works [27, 88]: *Clean Accuracy (CA)*: the accuracy of a clean downstream classifier on clean testing images from the downstream dataset; *Benign Accuracy (BA)*: the accuracy of a backdoored downstream classifier on the same clean testing images from the downstream dataset; *Attack Success Rate (ASR)*: the success rate of backdoor attacks. To evaluate the stealthiness and naturalness of the backdoor triggers, we employ three metrics following existing works [28]: *SSIM* [73], *PSNR* [25], *LPIPS* [87], *Feature Similarity Indexing Method (FSIM)* [86] and *Fréchet Inception Distance(FID)* [21]. Higher SSIM, PSNR, FSIM and lower LIPIPS, FID indicate better stealthiness and naturalness of the generated backdoored images.

SSL Frameworks. In the pre-training stage, we employ SimCLR [6] by default to train a ResNet18 [19] model, serving as our image encoder. Furthermore, we prove the effectiveness of our method on other SSL frameworks, i.e., MoCo [20], BYOL [16], SimSiam [7], SwAV [4], and CLIP [53] in Sec. A1 and Sec. A2.

Attack Baselines. We select two Instagram filters, Kelvin and Xpro2, as baseline triggers for aesthetic enhancements [30, 42]. Additionally, WaNet [47], CTRL [34], and ISSBA [36] are chosen for their stealthiness and high ASR. These triggers are injected into compromised encoders using BadEncoder. We also include DRUPE [68], a SOTA backdoor method using SimCLR and a patch trigger, as a baseline. To ensure a fair comparison, we evaluate our method against CTRL [34], SSLBKD [56], POIENC [39], and BLTO [63] using the same CIFAR10 as the pre-trained and downstream dataset under SimCLR, BYOL, and SimSiam. For SSLBKD, the trigger is randomly placed, while for SSLBKD-fixed, it's in the lower-right corner. We show more experimental settings and details in Sec. A6.5.

5.2. Effectiveness Evaluation

Effective Attack. As shown in Tab. 1, with different pre-trained and downstream datasets, our method achieves a high average ASR of 99.09% across various datasets. Additionally, Tab. 2 demonstrates that with the same pre-trained and downstream datasets, our approach also achieves nearly 100% ASRs. Our method outperforms all baseline methods in all scenarios, highlighting its robustness and superior effectiveness in executing successful backdoor attacks.

Accuracy Preservation. The downstream classifiers trained on the backdoored encoder maintain good accuracy on clean samples, as shown in Tab. 1. The average BA is 73.10% compared to the average CA of 72.96%, with the difference within 1%. This suggests that the backdoor introduced by our method does not compromise the classifier's ability to label clean images correctly. This is because $\mathcal{L}_{\text{utility}}$ guarantees that the backdoored and clean image encoders yield similar feature vectors for clean inputs.

Pre-training Dataset	Downstream Dataset	No Attack	BadEncoder + WaNet trigger		BadEncoder + CTRL trigger		BadEncoder + Ins-Kelvin trigger		BadEncoder + Ins-Xpro2 trigger		DRUPE + Patch trigger		Ours	
		CA	BA↑	ASR↑	BA↑	ASR↑	BA↑	ASR↑	BA↑	ASR↑	BA↑	ASR↑	BA↑	ASR↑
STL10	CIFAR10	86.77	84.43	10.28	87.19	8.72	86.75	18.63	86.85	16.83	84.36	98.39	87.11	**99.58**
	GTSRB	76.12	74.45	5.23	77.57	8.17	76.49	72.95	76.71	14.02	75.93	96.09	75.82	**97.97**
	SVHN	55.35	58.29	16.83	54.29	3.32	56.67	38.03	58.42	18.68	75.64	96.68	58.62	**99.76**
CIFAR10	STL10	76.14	72.73	9.78	75.73	16.85	74.89	1.16	74.11	5.91	74.43	96.72	74.02	**99.68**
	GTSRB	81.84	75.85	5.46	79.94	97.95	78.56	2.50	75.08	42.40	80.35	97.22	79.15	**98.73**
	SVHN	61.52	54.79	17.99	66.33	40.91	68.49	22.13	68.95	30.91	76.02	96.23	63.67	**98.79**
Average	/	72.96	70.09	10.93	73.51	29.32	73.64	25.90	73.35	21.46	77.79	96.89	73.10	**99.09**

Table 1. Effectiveness comparison to representative backdoor attacks in SSL with different triggers (CA(%), BA(%), and ASR(%)). We compare our method to BadEncoder [27] with various existing stealthy triggers. We also include the results of DRUPE [68] with their default visible patch trigger. We include CTRL here to demonstrate that it is ineffective across various downstream datasets. *Our approach constantly achieves the highest ASRs while maintaining the accuracy on clean samples of the downstream classifiers trained on the backdoored encoder.*

Attack	Invisible	SSL Method					
		SimCLR		BYOL		SimSiam	
		BA↑	ASR↑	BA↑	ASR↑	BA↑	ASR↑
POIENC [39]	×	80.50	11.10	81.70	10.70	81.90	10.70
SSLBKD [56]	×	79.40	33.20	80.30	46.20	80.60	53.10
SSLBKD (fixed) [56]	×	80.00	10.50	82.30	11.20	81.90	10.70
CTRL [34]	✓	80.50	85.30	82.20	61.90	82.00	74.90
BLTO [63]	×	90.10	91.27	91.21	94.78	90.18	84.63
Ours	✓	90.19	**100.00**	93.01	**99.99**	91.01	**99.99**

Table 2. Effectiveness comparison to data-poisoning-based backdoor attacks in SSL with their default triggers. We show the results of BA(%), and ASR(%) with the same pre-trained and downstream dataset CIFAR10. Since data poisoning-based methods require matched distributions between pre-training and downstream, we use the same pre-trained and downstream datasets. Our threat model is different from theirs, and our method can be applied when the distributions of pre-training and downstream datasets are different. This table's key aim is to demonstrate that our method achieves much higher ASR than them.

Method	SSIM↑	PSNR↑	LPIPS↓	FSIM↑	FID↓
Badencoder [27]/DRUPE [68]	0.8355	14.1110	0.07693	0.820	53.363
CTRL [34]	0.9025	32.4098	**0.00034**	0.865	71.138
WaNet [47]	0.7704	14.2372	0.07432	0.662	98.092
Ins-Kelvin [42]	0.4955	16.1925	0.14000	0.677	96.449
Ins-Xpro2 [42]	0.5981	17.9173	0.04434	0.817	35.084
POIENC [39]	0.1214	11.2787	0.15867	0.597	172.220
SSLBKD [56]	0.8737	16.2414	0.09640	0.891	118.320
BLTO [63]	0.8417	29.6756	0.00941	0.950	36.385
Ours	**0.9633**	**35.8649**	0.00896	**0.969**	**16.320**

Table 3. Stealthiness comparison to existing methods on CIFAR10. Our method remains stealthy. Detailed data are shown in Tab. A9.

5.3. Stealthiness Evaluation

Algorithmic Metrics. We first compare the average SSIM, PSNR, and LPIPS when the pre-trained dataset is CIFAR10 and downstream datasets are STL10, GTSRB, and SVHN injected with these backdoor triggers to compare the stealthiness of various backdoor attack methods. Tab. 3 indicates that our method exhibits strong stealthiness advantages with an average of 0.9633 SSIM, 35.8649 PSNR, 0.00896 LIPIS, 0.969 FSIM, and 16.320 FID indicating minimal structural changes to the images, hardly detectable noise, almost negligible perceptual difference between the original and perturbed images. Although CTRL achieves a better LPIPS, our method outperforms it in both SSIM and PSNR. Additionally, our average ASR is 99.09%, significantly higher than CTRL 29.32%(see Tab. 1), indicating that our method is more effective overall. More detailed data across various datasets, i.e., CIFAR10, STL10, GTSRB, SVHN, and ImageNet are shown in Tab. A9 and Tab. A10.

5.4. Robustness Evaluation

To assess the resilience of our method against current backdoor defenses, we deploy various SOTA backdoor defense strategies, i.e., DECREE [13], Beatrix [44], ASSET [48], Neural Cleanse (NC) [72], STRIP [14], Grad-CAM [58] for evaluation. Additionally, to further test the robustness of our method, we evaluate its endurance against the following commonly studied noises, i.e., JPEG compression [10, 11], Poisson noise [1, 79], and Salt&Pepper noise [1, 35]. We also design an adaptive defense method for INACTIVE. *We show that INACTIVE cannot be defended by STRIP, NC, Grad-CAM, noises, and adaptive defense in Sec. A5.1.*

DECREE. DECREE [13] identifies trojan attacks in pre-trained encoders by flagging an encoder as compromised if the reversed trigger's $\mathcal{L}^1$ norm proportion falls below a 0.1 threshold. As shown in Tab. 4, the $\mathcal{PL}^1$-Norm for each pre-trained and downstream dataset pair exceeds this threshold, so DECREE fails to detect backdoored encoders created by INACTIVE. This is because our invisible trigger breaks DECREE's assumption of a visible patch trigger, and our stealthy loss further narrows the distribution gap between backdoored and normal data, masking internal model anomalies.

Beatrix. Beatrix [44] identifies poisoned samples by detect-

Pre-trained Dataset	Downstream Dataset	$\mathcal{PL}^1$-Norm
CIFAR10	STL10	0.25
	SVHN	0.39
	GTSRB	0.15
STL10	CIFAR10	0.21
	SVHN	0.34
	GTSRB	0.20

Table 4. Evaluation results of DECREE [13]. A model is judged as backdoored if its $\mathcal{PL}^1$-Norm <0.1.

Encoder	Method	TP	FP	FN	TN	Acc
CIFAR-10	BadEncoder	499	24	1	476	97.50%
	Ours	0	24	500	476	47.60%
STL-10	BadEncoder	458	24	42	476	93.40%
	Ours	5	24	495	476	48.10%

Table 5. Detection results by Beatrix [44]. It struggles to detect poisoned samples from ours.

ing abnormalities in the feature space. We use two pretraining datasets, CIFAR-10 and STL-10, and create backdoored encoders using BadEncoder and INACTIVE. By sampling 500 clean inputs and 500 poisoned samples, we applied Beatrix to differentiate them. We find (see Tab. 5) that Beatrix effectively recognizes poisoned samples from BadEncoder with over 93% accuracy. However, Beatrix struggles to identify poisoned samples from INACTIVE, with a detection accuracy of below 50% on both CIFAR-10 and STL-10, which is like random guessing. We further analyze the reasons for the defense failure in Sec. A5.1.

ASSET. ASSET aims to distinguish between backdoored and clean samples by eliciting distinct behaviors in the model when processing these two data types, facilitating their separation [48]. We replicate their defensive techniques on our backdoored CIFAR-10 dataset. Specifically, we applied our synthesized trigger to CIFAR-10 (with a target label of 0) to create a poisoned version of CIFAR-10, maintaining a 100% poisoning rate as our default setting. The feature extractor used is the ResNet18 backbone, trained on this poisoned CIFAR-10 dataset.

The True Positive Rate (TPR) measures how effectively a backdoor detection method identifies backdoored samples, with a higher TPR (closer to 100%) indicating stronger filtering capability. The False Positive Rate (FPR) reflects the precision of this filtering: when TPR is sufficiently high, FPR shows the trade-off, highlighting the proportion of clean samples incorrectly flagged as backdoored. A lower FPR suggests fewer clean samples are mistakenly discarded, ensuring more clean data is retained for further use. Based on ASSET's metrics, we calculated the TPR as 7.14% and the FPR as 1.8%, indicating that our poisoned data can largely evade ASSET's detection.

Downstream Dataset	No Attack	ISSBA [36]		Ours	
	CA	BA↑	ASR↑	BA↑	ASR↑
STL10	95.68	92.58	9.97	93.48	100.00
GTSRB	80.32	66.29	5.10	82.84	96.00
SVHN	74.77	67.67	18.03	75.40	99.99
Average	83.59	75.51	11.03	83.91	98.66

Table 6. Comparative results (CA(%), BA(%), and ASR(%)) of ISSBA [36] and our attack on *ImageNet*. Ours constantly achieves the highest ASRs while maintaining accuracy on clean samples of the downstream classifiers.

Method	Average SSIM↑	Average PSNR (dB)↑	Average LPIPS↓
ISSBA [36]	0.7329	31.3496	0.12424
Ours	**0.9867**	**34.5733**	**0.01233**

Table 7. Stealthiness comparison on ImageNet.

5.5. Generalization to Large-scale Dataset

We assess the generalization of our method on a large-scale dataset by attacking an ImageNet-pre-trained encoder from Google [6]. We compare our method's performance with ISSBA, which is also trained and tested on ImageNet in its paper. Experimental setups are detailed in Sec. A6.5.

Experimental Results. Tab. 6 indicates that *our method is highly effective on ImageNet*, with an average 98.66% ASR across different datasets. Moreover, Tab. 7 indicates the high SSIM and PSNR values and low LPIPS values, demonstrating that *the perturbations made by INACTIVE are almost imperceptible.* Moreover, the average 83.91% BA is close to the average 83.59% CA, indicating *our attack maintains accuracy for the given downstream task despite the backdoor.* Additionally, *both our ASR and BA are much higher than those of the baseline ISSBA* [36], proving ours has better performance.

6. Conclusions and Future Work

In this paper, we propose an imperceptible and effective backdoor attack against self-supervised models based on the optimized triggers that are disentangled in the augmented transformation in the SSL. Based on the evaluation across five different datasets and six SSL algorithms, our attack is demonstrated to be both highly effective and stealthy. It also effectively bypasses existing backdoor defenses. For future work, it would be beneficial to expand the scope of research to include various other domains of machine learning, such as NLP and audio processing.

References

[1] Akshay Agarwal, Nalini Ratha, Richa Singh, and Mayank Vatsa. Robustness against gradient based attacks through cost effective network fine-tuning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Work-*

shops (CVPRW), page 28–37, Vancouver, BC, Canada, 2023. IEEE. 7

[2] Hanbo Cai, Pengcheng Zhang, Hai Dong, Yan Xiao, Stefanos Koffas, and Yiming Li. Toward stealthy backdoor attacks against speech recognition via elements of sound. *IEEE Transactions on Information Forensics and Security*, 19:5852–5866, 2024. 3

[3] Nicholas Carlini and Andreas Terzis. Poisoning and backdooring contrastive learning. In *International Conference on Learning Representations*, 2022. 3

[4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020. 2, 6, 1

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021. 1

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 2, 3, 4, 6, 8, 10

[7] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021. 2, 4, 6, 1

[8] Yukun Chen, Shuo Shao, Enhao Huang, Yiming Li, Pin-Yu Chen, Zhan Qin, and Kui Ren. Refine: Inversion-free backdoor defense via model reprogramming. In *ICLR*, 2025. 3

[9] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 215–223, Fort Lauderdale, FL, USA, 2011. PMLR. 6, 9

[10] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Siwei Li, Li Chen, Michael E. Kounavis, and Duen Horng Chau. Shield: Fast, practical defense and vaccination for deep learning using jpeg compression. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, page 196–204, New York, NY, USA, 2018. Association for Computing Machinery. 7

[11] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4312–4321, 2019. 7

[12] Zhi Dou, Ning Wang, Baopu Li, Zhihui Wang, Haojie Li, and Bin Liu. Dual color space guided sketch colorization. *IEEE Transactions on Image Processing*, 30:7292–7304, 2021. 8

[13] Shiwei Feng, Guanhong Tao, Siyuan Cheng, Guangyu Shen, Xiangzhe Xu, Yingqi Liu, Kaiyuan Zhang, Shiqing Ma, and Xiangyu Zhang. Detecting backdoors in pre-trained encoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16352–16362, 2023. 2, 3, 7, 8

[14] Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C Ranasinghe, and Surya Nepal. Strip: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th Annual Computer Security Applications Conference*, pages 113–125, 2019. 2, 3, 7, 4, 5

[15] Yinghua Gao, Yiming Li, Xueluan Gong, Zhifeng Li, Shu-Tao Xia, and Qian Wang. Backdoor attack with sparse and invisible trigger. *IEEE Transactions on Information Forensics and Security*, 2024. 3

[16] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 2, 4, 6, 1

[17] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017. 2, 3

[18] Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey on self-supervised learning: Algorithms, applications, and future trends. *arXiv preprint arXiv:2301.05712*, 2023. 2

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 1, 2, 4, 6

[21] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 6

[22] Sanghyun Hong, Varun Chandrasekaran, Yiğitcan Kaya, Tudor Dumitraş, and Nicolas Papernot. On the effectiveness of mitigating data poisoning attacks with gradient shaping. *arXiv preprint arXiv:2002.11497*, 2020. 3

[23] Linshan Hou, Ruili Feng, Zhongyun Hua, Wei Luo, Leo Yu Zhang, and Yiming Li. Ibd-psc: Input-level backdoor detection via parameter-oriented scaling consistency. *arXiv preprint arXiv:2405.09786*, 2024. 3

[24] Kunzhe Huang, Yiming Li, Baoyuan Wu, Zhan Qin, and Kui Ren. Backdoor defense via decoupling the training process. *arXiv preprint arXiv:2202.03423*, 2022. 3

[25] Q. Huynh-Thu and M. Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics Letters*, 44: 800–801(1), 2008. 6

[26] Ashish Jaiswal, Ashwin ramesh babu, Mohammad Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9:2, 2020. 2

[27] Jinyuan Jia, Yupei Liu, and Neil Zhenqiang Gong. Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning. In *2022 IEEE Symposium on Security and Privacy (SP)*, page 2043–2059, 2022. 1, 2, 3, 4, 6, 7, 8, 9, 10

[28] Wenbo Jiang, Hongwei Li, Guowen Xu, and Tianwei Zhang. Color backdoor: A robust poisoning attack in color space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8133–8142, 2023. 4, 6

[29] Ruinan Jin and Xiaoxiao Li. Backdoor attack and defense in federated generative adversarial network-based medical image synthesis. *Medical Image Analysis*, 90:102965, 2023. 4

[30] Akiomi Kamakura. pilgram. `https://github.com/akiomik/pilgram.git`, 2022. 2, 6

[31] Alex Krizhevsky. Learning multiple layers of features from tiny images. *Tech Report*, 2009. 6, 9

[32] Boheng Li, Yishuo Cai, Jisong Cai, Yiming Li, Han Qiu, Run Wang, and Tianwei Zhang. Purifying quantization-conditioned backdoors via layer-wise activation correction with distribution approximation. In *International Conference on Machine Learning*. PMLR, 2024. 3

[33] Boheng Li, Yishuo Cai, Haowei Li, Feng Xue, Zhifeng Li, and Yiming Li. Nearest is not dearest: Towards practical defense against quantization-conditioned backdoor attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

[34] Changjiang Li, Ren Pang, Zhaohan Xi, Tianyu Du, Shouling Ji, Yuan Yao, and Ting Wang. An embarrassingly simple backdoor attack on self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4367–4378, 2023. 1, 3, 4, 6, 7

[35] Deqiang Li and Qianmu Li. Adversarial deep ensemble: Evasion attacks and defenses for malware detection. *IEEE Transactions on Information Forensics and Security*, 15: 3886–3900, 2020. 7

[36] Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16463–16472, 2021. 1, 2, 3, 6, 8

[37] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. *arXiv preprint arXiv:2101.05930*, 2021. 3

[38] Chufan Liu, Xin Shu, Lei Pan, Jinlong Shi, and Bin Han. Multiscale underwater image enhancement in rgb and hsv color spaces. *IEEE Transactions on Instrumentation and Measurement*, 72:1–14, 2023. 8

[39] Hongbin Liu, Jinyuan Jia, and Neil Zhenqiang Gong. {PoisonedEncoder}: Poisoning the unlabeled pre-training data in contrastive learning. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 3629–3645, 2022. 3, 6, 7

[40] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International symposium on research in attacks, intrusions, and defenses*, pages 273–294. Springer, 2018. 2

[41] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *25th Annual Network And Distributed System Security Symposium (NDSS 2018)*. Internet Soc, 2018. 2

[42] Yingqi Liu, Wen-Chuan Lee, Guanhong Tao, Shiqing Ma, Yousra Aafer, and Xiangyu Zhang. Abs: Scanning neural networks for back-doors by artificial brain stimulation. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 1265–1282, 2019. 1, 3, 6, 7

[43] 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*, pages 15003–15013, 2022. 3

[44] Wanlun Ma, Derui Wang, Ruoxi Sun, Minhui Xue, Sheng Wen, and Yang Xiang. The" beatrix"resurrections: Robust backdoor detection via gram matrices. In *30th Network and Distributed System Security (NDSS) Symposium*, 2022. 2, 3, 7, 8

[45] Kai Mei, Zheng Li, Zhenting Wang, Yang Zhang, and Shiqing Ma. NOTABLE: Transferable backdoor attacks against prompt-based NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15551–15565, Toronto, Canada, 2023. Association for Computational Linguistics. 3

[46] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011. 6, 9

[47] Tuan Anh Nguyen and Anh Tuan Tran. Wanet - imperceptible warping-based backdoor attack. In *International Conference on Learning Representations*, 2021. 1, 2, 3, 6, 7, 5

[48] Minzhou Pan, Yi Zeng, Lingjuan Lyu, Xue Lin, and Ruoxi Jia. ASSET: Robust backdoor data detection across a multiplicity of deep learning paradigms. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 2725–2742, Anaheim, CA, 2023. USENIX Association. 2, 3, 7, 8

[49] Minzhou Pan, Yi Zeng, Lingjuan Lyu, Xue Lin, and Ruoxi Jia. {ASSET}: Robust backdoor data detection across a multiplicity of deep learning paradigms. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 2725–2742, 2023. 3

[50] Xiangyu Qi, Tinghao Xie, Yiming Li, Saeed Mahloujifar, and Prateek Mittal. Revisiting the assumption of latent separability for backdoor defenses. In *The eleventh international conference on learning representations*, 2022. 3

[51] Xiangyu Qi, Tinghao Xie, Ruizhe Pan, Jifeng Zhu, Yong Yang, and Kai Bu. Towards practical deployment-stage backdoor attack on deep neural networks. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13347–13357, 2022. 3

[52] Xiangyu Qi, Tinghao Xie, Jiachen T Wang, Tong Wu, Saeed Mahloujifar, and Prateek Mittal. Towards a proactive {ML} approach for detecting backdoor poison samples. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 1685–1702, 2023. 3

[53] 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 *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 6, 1

[54] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 6, 10

[55] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 6, 9

[56] Aniruddha Saha, Ajinkya Tejankar, Soroush Abbasi Koohpayegani, and Hamed Pirsiavash. Backdoor attacks on self-supervised learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 13327–13336, New Orleans, LA, USA, 2022. IEEE. 1, 3, 6, 7

[57] G. Saravanan, G. Yamuna, and S. Nandhini. Real time implementation of rgb to hsv/hsi/hsl and its reverse color space models. In *2016 International Conference on Communication and Signal Processing (ICCSP)*, page 0462–0466, Melmaruvathur, Tamilnadu, India, 2016. IEEE. 8

[58] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 2, 7, 4

[59] Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. *Advances in neural information processing systems*, 31, 2018. 3

[60] 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*, pages 9525–9536. PMLR, 2021. 3

[61] Li Shuhua and Guo Gaizhi. The application of improved hsv color space model in image processing. In *2010 2nd International Conference on Future Computer and Communication*, pages V2–10–V2–13, Wuhan, China, 2010. IEEE. 8

[62] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, 32:323–332, 2012. Selected Papers from IJCNN 2011. 6, 9

[63] Weiyu Sun, Xinyu Zhang, Hao LU, Ying-Cong Chen, Ting Wang, Jinghui Chen, and Lu Lin. Backdoor contrastive learning via bi-level trigger optimization. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 3, 4, 6, 7

[64] Di Tang, XiaoFeng Wang, Haixu Tang, and Kehuan Zhang. Demon in the variant: Statistical analysis of {DNNs} for robust backdoor contamination detection. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 1541–1558, 2021. 3

[65] Guanhong Tao, Yingqi Liu, Guangyu Shen, Qiuling Xu, Shengwei An, Zhuo Zhang, and Xiangyu Zhang. Model orthogonalization: Class distance hardening in neural networks for better security. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1372–1389. IEEE, 2022. 3

[66] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13368–13378, 2022. 3

[67] Guanhong Tao, Zhenting Wang, Siyuan Cheng, Shiqing Ma, Shengwei An, Yingqi Liu, Guangyu Shen, Zhuo Zhang, Yunshu Mao, and Xiangyu Zhang. Backdoor vulnerabilities in normally trained deep learning models. *arXiv preprint arXiv:2211.15929*, 2022. 3

[68] Guanhong Tao, Zhenting Wang, Shiwei Feng, Guangyu Shen, Shiqing Ma, and Xiangyu Zhang. Distribution preserving backdoor attack in self-supervised learning. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 29–29. IEEE Computer Society, 2023. 1, 2, 5, 6, 7, 9

[69] Ajinkya Tejankar, Maziar Sanjabi, Qifan Wang, Sinong Wang, Hamed Firooz, Hamed Pirsiavash, and Liang Tan. Defending against patch-based backdoor attacks on self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12239–12249, 2023. 3

[70] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. *Advances in neural information processing systems*, 31, 2018. 3

[71] Cédric Villani et al. *Optimal transport: old and new*. Springer, 2009. 5, 9

[72] 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)*, page 707–723, San Francisco, CA, USA, 2019. IEEE. 2, 7, 3

[73] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. 6

[74] Zhenting Wang, Hailun Ding, Juan Zhai, and Shiqing Ma. Training with more confidence: Mitigating injected and natural backdoors during training. *Advances in Neural Information Processing Systems*, 35:36396–36410, 2022. 3

[75] Zhenting Wang, Kai Mei, Hailun Ding, Juan Zhai, and Shiqing Ma. Rethinking the reverse-engineering of trojan triggers. *Advances in Neural Information Processing Systems*, 35:9738–9753, 2022. 3

[76] Zhenting Wang, Kai Mei, Juan Zhai, and Shiqing Ma. Unicorn: A unified backdoor trigger inversion framework. In *The Eleventh International Conference on Learning Representations*, 2022. 3

[77] Zhenting Wang, Juan Zhai, and Shiqing Ma. Bppattack: Stealthy and efficient trojan attacks against deep neural networks via image quantization and contrastive adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15074–15084, 2022. 3

[78] Tinghao Xie, Xiangyu Qi, Ping He, Yiming Li, Jiachen T Wang, and Prateek Mittal. Badexpert: Extracting backdoor functionality for accurate backdoor input detection. In *The Twelfth International Conference on Learning Representations*, 2023. 3

[79] Mengting Xu, Tao Zhang, and Daoqiang Zhang. Medrdf: A robust and retrain-less diagnostic framework for medical pretrained models against adversarial attack. *IEEE Transactions on Medical Imaging*, 41(8):2130–2143, 2022. 7

[80] Xiaojun Xu, Qi Wang, Huichen Li, Nikita Borisov, Carl A Gunter, and Bo Li. Detecting ai trojans using meta neural analysis. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 103–120. IEEE, 2021. 3

[81] Xiong Xu, Kunzhe Huang, Yiming Li, Zhan Qin, and Kui Ren. Towards reliable and efficient backdoor trigger inversion via decoupling benign features. In *ICLR*, 2024. 3

[82] Yuanshun Yao, Huiying Li, Haitao Zheng, and Ben Y. Zhao. Latent backdoor attacks on deep neural networks. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, page 2041–2055, New York, NY, USA, 2019. Association for Computing Machinery. 5

[83] Yi Zeng, Minzhou Pan, Himanshu Jahagirdar, Ming Jin, Lingjuan Lyu, and Ruoxi Jia. {Meta-Sift}: How to sift out a clean subset in the presence of data poisoning? In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 1667–1684, 2023. 3

[84] Yi Zeng, Minzhou Pan, Hoang Anh Just, Lingjuan Lyu, Meikang Qiu, and Ruoxi Jia. Narcissus: A practical clean-label backdoor attack with limited information. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 771–785, 2023. 3

[85] Hanrong Zhang, Jingyuan Huang, Kai Mei, Yifei Yao, Zhenting Wang, Chenlu Zhan, Hongwei Wang, and Yongfeng Zhang. Agent security bench (ASB): Formalizing and benchmarking attacks and defenses in LLM-based agents. In *The Thirteenth International Conference on Learning Representations*, 2025. 3

[86] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, 2011. 6

[87] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 6

[88] Zaixi Zhang, Jinyuan Jia, Binghui Wang, and Neil Zhenqiang Gong. Backdoor attacks to graph neural networks. In *Proceedings of the 26th ACM Symposium on Access Control Models and Technologies*, page 15–26, New York, NY, USA, 2021. Association for Computing Machinery. 6

[89] Rui Zhu, Di Tang, Siyuan Tang, XiaoFeng Wang, and Haixu Tang. Selective amnesia: On efficient, high-fidelity and blind suppression of backdoor effects in trojaned machine learning models. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 1–19. IEEE, 2023. 3