BARBIE: Robust Backdoor Detection Based on Latent Separability

Hanlei Zhang, Yijie Bai, Yanjiao Chen*, Zhongming Ma, Wenyuan Xu
Zhejiang University
(zhanghanlei, baiyj, chenyanjiao, zmma, wyxu)@zju.edu.cn

***Abstract*—Backdoor attacks are an essential risk to deep learning model sharing. Fundamentally, backdoored models are different from benign models considering latent separability, i.e., distinguishable differences in model latent representations. However, existing methods quantify latent separability by clustering latent representations or computing distances between latent representations, which are easy to be compromised by adaptive attacks. In this paper, we propose BARBIE, a backdoor detection approach that can pinpoint latent separability under adaptive backdoor attacks. To achieve this goal, we propose a new latent separability metric, named relative competition score (RCS), by characterizing the dominance of latent representations over model output, which is robust against various backdoor attacks and is hard to compromise. Without the need to access any benign or backdoored sample, we invert two sets of latent representations of each label, reflecting the normal latent representations of benign models and intensifying the abnormal ones of backdoored models, to calculate RCS. We compute a series of RCS-based indicators to comprehensively reflect the differences between backdoored models and benign models. We validate the effectiveness of BARBIE on more than 10,000 models on 4 datasets against 14 types of backdoor attacks, including the adaptive attacks against latent separability. Compared with 7 baselines, BARBIE improves the average true positive rate by 17.05% against source-agnostic attacks, 27.72% against source-specific attacks, 43.17% against sample-specific attacks and 11.48% against clean-label attacks. BARBIE also maintains lower false positive rates than baselines. The source code is available at: https://github.com/Forliqr/BARBIE.**

I. INTRODUCTION

Deep learning has been widely applied in various domains, including face recognition [1], [2], [3], machine translation [4], [5], [6], autonomous driving [7], [8], [9] and medical diagnosis [10], [11], [12]. However, model training, especially for large models, becomes more and more expensive, requiring not only a large number of training data samples but also extensive computational resources. To use deep learning models in an affordable way, many developers choose to download open-source models from model-sharing or model-selling platforms, such as Hugging Face [13], Model Zoo [14], Github [15] and AWS Marketplace [16]. As the developers of these open-source models may upload harmful models for the purpose of evading identity authentication, stealing privacy and property, creating security risks and so on, the downloaded models may be contaminated by backdoor attacks.

Backdoor attacks are manifested in backdoored models that behave normally on benign samples but can be manipulated by backdoored samples. Backdoored samples are constructed by imposing a specially-design trigger onto benign samples. A wide variety of backdoor attacks have been proposed. Depending on whether the trigger is universal or different across samples, backdoor attacks can be samples-agnostic or sample-specific. Sample-agnostic backdoor attacks adopt a fixed trigger pattern that may be easily detected, while sample-specific backdoor attacks vary the trigger on each sample to evade being detected. Backdoor attacks may also be categorized as source-agnostic or source-specific. Source-agnostic backdoor attacks can turn samples of any label into backdoored samples, while source-specific backdoor attacks only poison samples of a certain source label. Backdoor attacks may lead to devastating consequences [17]. Therefore, backdoor detection methods are necessary to safeguard the model-sharing ecosystem.

Backdoor detection can be achieved by examining suspicious samples or models. Sample detection methods [18], [19], [20], [21], [22], [23], [24] check training samples before training the model or query samples after model deployment, while model detection methods inspect a trained model before deployment. In this paper, we focus on model detection, which vouches for pre-trained models from untrusted sources before the models are put into practice. Existing model detection methods focus on potential abnormality in model outputs given certain model inputs, yielding unstable detection accuracy under stealthy and adaptive attacks such as sample-specific attacks. Latent separability, the differences in latent representations of benign/backdoored models given benign/backdoored samples, is a fundamental feature that separates backdoored models from benign models. However, existing model detection methods have a limited exploration of latent separability due to two difficulties. First, model detection is usually performed with limited or no access to backdoored samples, making it hard to extract backdoored latent representations. Second, although some sample detection methods did try to leverage latent separability, they quantify latent separability by clustering latent representations or computing

*Yanjiao Chen is the corresponding author.

Network and Distributed System Security (NDSS) Symposium 2025
24-28 February 2025, San Diego, CA, USA
ISBN 979-8-9894372-8-3
https://dx.doi.org/10.14722/ndss.2025.240179
www.ndss-symposium.org

Table I

A SUMMARY OF BACKDOOR DETECTION METHODS. ADVERSARY KNOWLEDGE INCLUDES ANY DATA, MODEL OR INFORMATION FROM THE ADVERSARY. ADAPTIVE BACKDOOR ATTACKS REFER TO ATTACKS AGAINST LATENT SEPARABILITY. ✓ AND ✗ DENOTE WHETHER THIS METHOD SUPPORTS OR NOT. N/A DENOTES NO RELEVANT INFORMATION AVAILABLE.

Method	Domain	No Adversary Knowledge	No Access to Data		Backdoor Attack				
			Poisoned Data	Clean Data	Source-Agnostic	Source-Specific	Sample-Agnostic	Sample-Specific	Adaptive
SentiNet [18]	Input	✗	✗	✗	✓	✗	✓	✗	N/A
Activation Clustering [19]	Input	✗	✗	✓	✓	✗	✓	✗	✗
Spectral-Signature [20]	Input	✗	✗	✓	✓	✗	✓	✗	✗
STRIP [21]	Input	✗	✗	✗	✓	✗	✓	✗	✗
SCAn [22]	Input	✗	✗	✗	✓	✓	✓	✗	✗
Beatrix [23]	Input	✗	✗	✗	✓	✓	✓	✓	N/A
TED [24]	Input	✗	✗	✗	✓	✓	✓	✓	N/A
NC [25]	Model	✓	✓	✗	✓	✗	✓	✗	✗
ABS [26]	Model	✓	✓	✗	✓	✗	✓	✗	✗
MNTD [27]	Model	✓	✓	✗	✓	✗	✓	✗	✗
FreeEagle [28]	Model	✗	✓	✓	✓	✓	✓	✗	✗
BARBIE **(Ours)**	**Model**	✓	✓	✓	✓	✓	✓	✓	✓

distances between latent representations, which are easy to be compromised by adaptive attacks. As summarized in Table I, due to the above limitations, existing detection methods, both sample detection and model detection, fail to yield a stable performance for all varieties of backdoor attacks.

In this paper, we propose a robust model detection approach against backdoor attacks based on latent separability, which we term as BARBIE. To enhance latent separability under stealthy and adaptive attacks, we design the Relative Competition Score (RCS), which characterizes the ability of one latent representation to dominate the model output when competing with another latent representation. The RCS metric reflects a robust latent separability as the fundamental purpose of any backdoor attack is to alter the model output of benign samples with minor changes. Therefore, the backdoored latent representations should dominate model output compared to benign latent representations. To extract latent representations and compute RCS-based latent separability indicators without the need to access backdoored samples, we propose a latent representation inversion method that extracts two sets of latent representations to reflect normal latent representations of benign models and intensify the abnormal ones of backdoored models respectively. We compute a series of indicators based on the RCS metric to comprehensively reflect the differences between backdoored models and benign models.

We conduct comprehensive experiments to evaluate the performance of BARBIE on four datasets (MNIST [29], CIFAR10 [30], GTSRB [31], and ImageNette [32]). Compared to the state-of-the-art backdoored model detection methods, BARBIE improves the average true positive rate (TPR) by 17.05% against source-agnostic attacks, 27.72% against source-specific attacks, 43.17% against sample-specific attacks, 11.48% against clean-label attacks and maintains lower false positive rates (FPRs). We evaluate the resistance of BARBIE against different adaptive attacks, obtaining an average TPR of 99.98% and low FPRs. Furthermore, we explore the possibility of applying BARBIE in self-supervised learning.

Our main contributions are summarized as follows.

- We propose a backdoor detection approach based on latent separability. We design a new latent separability metric named relative competition score (RCS), which reflects the dominance of latent representations over model output. RCS is robust against various backdoor attacks and is hard to be compromised by adaptive attacks.
- We compute RCS in a data-free manner by inverting latent representations without access to any benign or backdoored sample. We design a series of RCS-based indicators and determine the boundary of indicators for benign models to detect various backdoored models.
- We conduct extensive experiments to validate the effectiveness and robustness of our method. The results demonstrate a high accuracy of our proposed detection approach against a wide variety of backdoor attacks, especially adaptive backdoor attacks.

II. PRELIMINARIES

A. Deep Neural Networks

A deep neural network f with N layers $\{f_i\}_{N}^{i=1}$ maps an input x to a label $y \in \{1, 2, \ldots, m\}$. The output of the neural network is an m-dimensional vector consisting of confidence scores, which represent the probability of each label[1].

$$f(x) = f_N(f_{N-1}(\cdots f_1(x))). \tag{1}$$

The parameters of the neural network, often denoted as θ, are usually obtained by fitting on a labeled training dataset $\mathcal{D} = \{(x, y)\}$. The training process aims to minimize a predefined loss function $\mathcal{L}$ that quantifies the difference between the model's output $f_\theta(x)$ and the ground-truth label y, i.e., $\theta^* = \arg\min_\theta \sum_{(x,y)\in\mathcal{D}} \mathcal{L}(f(x), y)$.

An intermediate layer $f_i, i \in [1, N]$ can separate the neural network into two parts, i.e., the head sub-network $f_e = \{f_k : k \in [1, i]\}$ and the tail sub-network $f_c = \{f_k : k \in [i+1, N]\}$. The head sub-network f_e extracts a latent representation from the input, denoted as v. The tail sub-network f_c takes v as the input to attain the final output.

$$v = f_{i-1}(f_{i-2}(\cdots f_1(x))), \tag{2}$$

$$f(x) = f_N(f_{N-1}(\cdots f_i(v))). \tag{3}$$

Latent representations are often used for interpreting the behaviors of a neural network. CAM [33] and Grad-CAM [34] generate interpretation saliency maps by combining the latent representation maps to visualize the important regions in an

[1]Without loss of generality, following existing works, we focus on classification tasks in this paper.

input image. Du *et. al.* [35] leveraged latent representations to interpret the working mechanism of neural networks. In this paper, we leverage latent representations for backdoored model detection.

B. Backdoor Attacks

Backdoor attacks poison the training process of a neural network, resulting in backdoored models that perform maliciously in the presence of a backdoored sample. A backdoored sample $\widetilde{x}$ is usually obtained by transforming a clean sample x as

$$\widetilde{x} = \mathcal{T}(x). \tag{4}$$

where $\mathcal{T}$ is a triggering function that imposes a backdoor trigger onto the clean sample. According to the property of the trigger, backdoor attacks can be categorized as *sample-agnostic* or *sample-specific*.

Sample-agnostic attacks. Many backdoor attacks adopt a sample-agnostic fixed trigger, e.g., a special sticker or a logo [36], [37], [38], [39], [40], [17], [41]. In this case, the triggering function is reduced to $\mathcal{T}(x) = x + t$, where t is the fixed trigger. Sample-agnostic attacks are easy to design, but are also easy to detect since all backdoored samples have the same trigger.

Sample-specific attacks. In sample-specific attacks, the triggering function is usually a generative network, generating customized triggers for different samples [42], [43], [44]. Sample-specific attacks are much stealthier than sample-agnostic attacks, shown to be quite evasive to existing backdoor detection methods.

According to the source samples used to generate backdoored samples, backdoor attacks can be categorized as *source-agnostic* or *source-specific*.

Source-agnostic attacks. In source-agnostic attacks, the source samples come from any class of clean samples [36], [37], [40], [17], [45], [46], [47], [48]. In other words, a clean sample of any label can be transformed into a backdoored sample that incurs misclassification (usually to a target label). Source-agnostic attacks are relatively easy to detect due to the strong influence of the trigger on the target misclassification label.

Source-specific attacks. In source-specific attacks, the source samples belong to a certain label (referred to as the source label) [22], [49]. Applying the triggering function to non-source samples will not activate the backdoor. Compared with source-agnostic attacks, source-specific attacks are more difficult to detect, since the link between the trigger and the target label is established only for the source label.

There are also other ways of categorizing backdoor attacks. For example, depending on whether the poisoned training data is mislabeled or labeled correctly, backdoor attacks can be classified as dirty-label attacks [37], [38], [40], [17], [44], [49], [50], [51] and clean-label attacks [39], [52], [53], [54], [55], [56], [57], [58]. In the experiments, we will evaluate the detection capability of BARBIE on various kinds of backdoor attacks.

C. Backdoor Detection

Backdoor detection aims to detect whether a sample or a model is backdoored or not. Sample detection identifies whether an input sample contains the trigger or not, and model detection checks whether a trained model is backdoored or not. Sample detection can be performed on training samples before the model is trained or on query samples after the model is deployed. Model detection is usually performed on a trained model before deployment.

Sample detection. Existing sample detection methods mainly utilize input perturbation or latent separability.

Input perturbation based methods assume that a greater perturbation is needed to change the output label of a backdoored sample than a clean sample, because of the strong link between the trigger and the target label. For example, SentiNet [18] perturbs benign samples with potential backdoor regions located by Grad-CAM for detection. STRIP [21] perturbs input samples and measures the entropy of the output to detect backdoored samples. However, input perturbation based detection methods are less effective in sample-specific and source-specific attacks due to a weaker link between the trigger and the target label.

Latent separability based methods extract latent representations of input samples and separate benign and backdoored samples based on differences in their latent representations. For instance, Activation Clustering [19] separates the activations of benign and backdoored samples based on clustering. Spectral Signature [20] uses the spectrum of the covariance of latent representations to differentiate benign and backdoored samples. SCAn [22] decomposes latent representations of input samples into a class-specific identity and a variation component, and then separates benign and backdoored samples based on a weighted Mahalanobis distance. Beatrix [23] utilizes the Gram matrix to capture the latent representation differences between backdoored and benign samples. TED [24] utilized the evolution trajectory of latent representations in each layer to detect malicious samples based on their activation distances to benign samples. Unfortunately, these latent separability based sample detection methods mainly quantify latent separability based on clustering or distance metrics, making them susceptible to adaptive attacks.

Model detection. Existing model detection methods mainly rely on input perturbation. NC [25] perturbs benign inputs to reverse the backdoor trigger based on shortcuts in backdoored models. ABS [26] adds different levels of perturbation on inputs to induce abnormal activations of backdoored neurons. MNTD [27] trains a meta-classifier with a set of benign and backdoored shadow models to directly detect a suspicious model. FreeEagle [28] discovers backdoored models based on the model output of inverted latent representations. However, these methods have a limited exploration of latent separability, the fundamental differences between benign and backdoored models. Furthermore, due to a lack of known backdoored samples before model deployment, existing latent separability based sample detection methods can hardly be adapted for

model detection.

In this paper, we focus on model detection, which screens suspicious models before deployment. We propose a model detection approach for backdoor attacks, which extracts distinguishable latent representation features to enhance latent separability between benign and backdoored models in a robust and data-free manner.

III. System Model

Our system model consists of an attacker and a defender, with goals, capabilities and knowledge defined as follows. In particular, we consider a strong attacker and a defender without any data or knowledge of the attack.

A. Attacker

The goal of the attacker is a high attack success rate and high clean data accuracy. More specifically, the backdoored model should output the target label given a backdoored sample and the correct label given a benign sample.

Attacker's capabilities and knowledge. We assume that the attacker controls the entire training process of the backdoored model, including the training datasets, the model structure and the model weights. Under this assumption, the attacker can implement all kinds of backdoor attacks.

B. Defender

The goal of the defender is to identify whether a trained model is backdoored or not.

Defender's capabilities and knowledge. We assume that the defender has no access to any data, including any backdoored or clean sample, which makes our defense feasible in the most strict conditions. We further assume that the defender has no knowledge of the backdoor attack method adopted by the attacker or any known backdoored models previously published by the attacker. The defender only has access to the model to be examined.

IV. Barbie: Detailed Construction

In this section, we first introduce a newly proposed latent separability metric, named *Relative Competition Score (RCS)*, which is carefully designed to amplify and capture differences in latent representations of benign and backdoored samples. The overall architecture of Barbie is displayed in Figure 1. To obtain latent representations without knowing any benign or backdoored sample, we design an inversion method to produce two sets of latent representations used to calculate RCS in a data-free manner. To enhance the latent separability of inverted latent representations, we compute a variety of indicators based on RCS metrics to comprehensively reflect the differences between backdoored models and benign models.

A. Relative Competition Score

The latent representations of two samples belonging to class k and class b can be extracted as

$$v_k = f_e(x_k), \tag{5}$$
$$v_b = f_e(x_b), \tag{6}$$

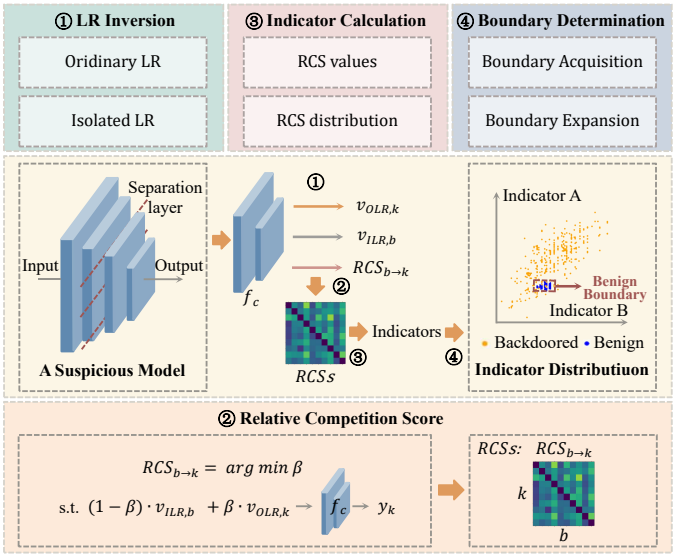


Figure 1. The overview of Barbie. Latent representation is abbreviated as LR.

where f_e is the head sub-network of the given model f, v_k and v_b are the latent representations of input x_k and x_b respectively. We have

$$f_c(v_k) = y_k, \tag{7}$$
$$f_c(v_b) = y_b, \tag{8}$$

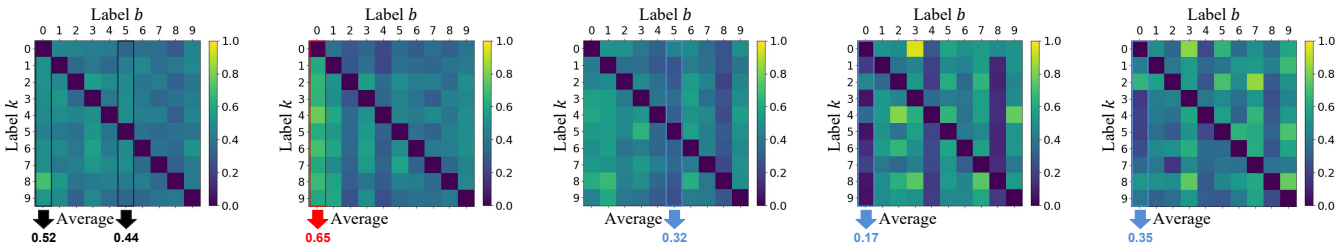
where f_c is the tail sub-network of f, and y_k and y_b are two different labels of class k and class b respectively.

Latent separability refers to the differences between latent representations of benign samples and backdoored samples. Existing works quantify latent separability between benign and backdoored samples by clustering latent representations or computing distances between latent representations. However, adaptive attacks can evade these detection methods by narrowing the gap of latent representations between backdoored and benign samples. To tackle this problem, we design a new way of quantifying latent separability. Instead of measuring the distance between two latent representations, we pinpoint the influence of a specific latent representation on another latent representation. More specifically, we define Relative Competition Score (RCS) as the proportion of v_k needed to change the output of v_b from the original label y_b to the target label y_k,

$$RCS_{b\to k} = \arg\min \beta, \tag{9}$$
$$\text{s.t.,}\quad f_c((1-\beta)v_b + \beta v_k) = y_k, \tag{10}$$

If $RCS_{b\to k}$ is high, latent representation v_b has a far greater dominance of model output than latent representation v_k. Computing the relative competition score regarding all possible pairs of y_k and y_b, we can obtain an RCS matrix.

We conduct preliminary experiments to show the distinguishability of the RCS between benign and backdoored models. We train a benign model using CIFAR-10 dataset [30] and VGG-16 model structure [49]. We then train four backdoored models using a source-agnostic (also sample-agnostic), a source-specific (also sample-agnostic), a sample-specific and



(a) Benign Model (b) Source-Agnostic Attack (c) Source-Specific Attack (d) Sample-Specific Attack (e) Adaptive Attack

Figure 2. Relative competition scores of models under various backdoor attacks.

an adaptive attack against latent separability. We adopt Input-Aware Dynamic Backdoor Attack [44] as the sample-specific attack, Adaptive-Patch Attack [59] as the adaptive attack and a patch trigger [60] as the trigger of other attacks.

In these experiments, we randomly sample 100 sets of clean data, each consisting of 10 samples with different labels, to compute RCS. When the model is a backdoored model, we replace benign samples of the target label with corresponding backdoored samples. Figure 2 exhibits the RCS matrix of different models. Note that the target label of backdoor attacks is $y = 0$, and the source label of the source-specific attack is $y = 5$. It is shown that average RCSs exhibit abnormally high or low values in backdoored models. We have the following observations.

- **Source-agnostic attacks**. Source-agnostic attacks have abnormally high $\overline{RCS}_{b\to k, b=b_t, \forall k}$, where b_t is the target label. This is because the backdoored latent representations strongly affect other benign latent representations. As shown in Figure 2(b), the average $\overline{RCS}_{0\to k,\forall k}$ is abnormally high for the target label 0.
- **Source-specific attacks**. Source-specific attacks have abnormally low $\overline{RCS}_{b\to k, b=b_s, \forall k}$, where b_s is the source class. This is because the latent representations of the source class is easily affected by the trigger, while the latent representations of other classes are not. As shown in Figure 2(c), the average $\overline{RCS}_{5\to k,\forall k}$ is abnormally low for the source class 5.
- **Sample-specific & adaptive attacks**. Sample-specific and adaptive attacks backdoored have abnormally low $\overline{RCS}_{b\to k, b=b_t, \forall k}$, where b_t is the target label. This is because the backdoored latent representations only work on a specific sample or some specific latent representations but not any others. As shown in Figure 2(d)(e), the average $\overline{RCS}_{0\to k,\forall k}$ is abnormally low for the target label 0.

The above experiments demonstrate that RCS can reflect latent separability in a stable way that is not easily compromised by adaptive attacks. RCS can reveal hard-to-detect attacks like sample-specific attacks [44] and adaptive attacks [59], which are known for their evasiveness.

B. Latent Representation Inversion

To extract the latent representation $f_e(x)$, we need the input sample x. Nonetheless, under our strict threat model, the defender has no access to any benign or backdoored sample. To obtain latent representation under this constraint, we design a data-free latent representation inversion method. In particular, we inverse two sets of latent representations to reflect normal latent representations and amplify abnormal latent representations.

Given a specific label $k \in \{1, 2, \ldots, m\}$, we can invert its most representative latent representations by maximizing the model prediction of label k.

$$v_{OLR,k} = \arg\min_{v}(\mathcal{L}(f_c(v), y_k)), \tag{11}$$

where $v_{OLR,k}$ is the ordinary latent representations (OLRs) that can maximize model prediction of a certain label. We can adopt a gradient descent algorithm to solve the optimization problem. However, $v_{OLR,k}$ alone may not facilitate robust backdoor detection since $v_{OLR,k}$ of the target label in a backdoored model mixes both benign and backdoored latent representations.

To address this issue, we extract another set of latent representations which augment the proportion of backdoored latent representations. We maximize the model prediction of one label, while inhibiting the confidence scores of other labels,

$$v_{ILR,k} = \arg\min_{v}(\mathcal{L}(f_c(v), y_k) - \frac{\alpha}{m-1}\sum_{k'\neq k}\mathcal{L}(f_c(v), y_{k'})), \tag{12}$$

We refer to $v_{ILR,k}$ as isolated latent representations (ILRs). Backdoored latent representations, which only exist in the reconstructed latent representation of the target label of a backdoored model, will be amplified in $v_{ILR,k}$.

C. Abnormality Indicator Calculation

Based on two sets of inverted latent representations, compute the relative competitive score as

$$RCS_{b\to k} = \arg\min \beta, \tag{13}$$

$$\text{s.t.,}\quad f_c((1-\beta)v_{ILR,b} + \beta v_{OLR,k}) = y_k, \tag{14}$$

where $RCS_{b\to k}$ represents the proportion of $v_{OLR,k}$ needed to transform the model prediction from label y_b to label y_k. A small $RCS_{b\to k}$ means that tampering with the model prediction over $v_{ILR,b}$ is easy for $v_{OLR,k}$.

All RCS values regarding all possible pairs of y_k and y_b form an RCS matrix. Given n labels, $n(n-1)$ pairs of y_k and y_b are considered. The overall computation time is short, i.e., 0.33s, 1.54s, 5.60s and 92.65s on MNIST, CIFAR10, ImageNette and GTSRB respectively. To spot the abnormality

of the RCS matrix of a model, we consider both RCS values and statistical RCS distributions as abnormality indicators.

- *Single RCS values*. We consider each value in the RCS matrix, i.e., $RCS_{b\to k,\forall b,k}$.
- *Average RCS values*. Given a certain label y_k, we compute the average dominance of $v_{OLR,k}$ on all other latent representations and the average dominance of all other latent representations on $v_{ILR,k}$, i.e., $\overline{RCS}_{b\to k,\forall b}$ and $\overline{RCS}_{k\to b,\forall b}$.
- *Differential RCS values*. Given a certain label y_k, we compute $\overline{RCS}_{b\to k,\forall b} - \overline{RCS}_{k\to b,\forall b}$.
- *Statistical RCS metrics*. As shown in Figure 2, the presence of backdoors not only alters the RCS of latent representations with specific labels, but also randomly disturbs the RCS of latent representations with other labels, which changes the distribution of values in the RCS matrix. We assess the distribution from three perspectives, namely central tendency, dispersion tendency and shape. The central tendency, including mean and mode, reflects the centralized trend of data distribution. The dispersion tendency, including range, standard deviation and coefficient of variation, reflects the discrete trend of data distribution. The shape of a data distribution includes the skewness and kurtosis.

D. Detection Boundary Determination

After computing all the above indicators, we need to determine the boundary that separates backdoored models from benign models. We rely on a few known benign models to acquire the normal value ranges of these indicators. As model-sharing or model-selling platform regulators, it is reasonable for them to obtain benign models from trusted third parties or their own reserved models. Even if model resources are limited or untrustworthy, BARBIE can achieve good detection results as shown in Section VI-A and VI-B.

For indicator i, we calculate its lower-bound r_{li}, upper-bound r_{hi}, and standard deviation s_i in benign models. To allow for more generalized benign boundaries, we expand the benign boundary to $[R_{li}, R_{hi}]$ as

$$R_{li} = \begin{cases} r_{li}/w_i, & r_{li} \geq 0, \\ w_i r_{li}, & r_{li} < 0. \end{cases} \tag{15}$$

$$R_{hi} = \begin{cases} w_i r_{hi}, & r_{hi} \geq 0, \\ r_{hi}/w_i, & r_{hi} < 0. \end{cases} \tag{16}$$

where

$$\omega_i = \frac{e_i + \text{submin}(e)}{\max(e) + \text{submin}(e)} * \omega_{max}, \tag{17}$$

$$e_i = \ln \frac{s_i}{\min(s)}, \tag{18}$$

where $\text{submin}(\cdot)$ is a function to find the second smallest value. ω is the weight of expansion, and we set ω_{max} as the maximum value of ω. A large standard deviation means that the scope should expand to a greater extent. With boundary expansion, we can distinguish various unknown backdoored models from benign ones with a significant false positive rate decrease.

Furthermore, we observe that some benign models may have abnormal RCS values, which lead to false alarms. To tackle this problem, we add a small perturbation δ to latent representations.

$$v'_{OLR,k} = (1 - \frac{|\delta|}{|v_{OLR,k}|}) \times v_{OLR,k} + \delta, \tag{19}$$

To obtain the perturbation δ, we first compute the average difference between v_{OLR} and v_{ILR}. Then, we filter out the parts that are less than the mean of the perturbation to avoid modifying the unique parts in different latent representations.

$$\delta = \gamma \text{Filter}(\frac{\sum_k (v_{OLR,k} - v_{ILR,k})}{m}), \tag{20}$$

where $\text{Filter}(\cdot)$ is a function to filter any value below the average, γ is a scaling factor and δ is used to dilute all v_{OLR}.

The perturbation δ retains the similar part of inverted latent representations of different labels, which can weaken latent representations with large RCS values and strengthen latent representations with small RCS values, thus narrowing the gap of RCS between different latent representations and eliminating abnormal RCS in benign models. However, backdoored latent representations are trained to have a high tolerance for disturbances (source-agnostic and source-specific attacks) or vice versa (sample-specific and adaptive attacks against latent separability). Due to different tolerances to disturbances, the differences between backdoored and benign models will be further exacerbated after this processing.

V. EVALUATION

In order to demonstrate the effectiveness and robustness of BARBIE, we conduct extensive experiments. The experimental setup is detailed in Section V-A. We test the performance of BARBIE and other backdoored model detection methods against source-agnostic and source-specific attacks, including some adaptive attacks against latent separability, in Section V-B, against sample-specific attacks in Section V-C and against clean-label attacks in Section V-D. We propose two adaptive attacks and verify the performance of BARBIE against them in Section V-F. We also explore the possibility of applying BARBIE on large datasets in Section V-E, on vision transformer in Section V-G and in self-supervised learning in Section V-H. We conduct ablation study in Section V-I and hyperparameter experiments in Section V-J.

A. Experiment Setup

Datasets and models. We conduct experiments on four basic datasets: MNIST [29], CIFAR10 [30], ImageNette [32] and GTSRB [31]. (1) MNIST consists of 60,000 training samples and 10,000 test samples, which are 28×28 gray images of handwritten digits from 0-9. (2) CIFAR10 consists of 50,000 training samples and 10,000 test samples, which are 32×32 color images of 10 categories, such as airplane, automobile, bird, cat, etc. (3) ImageNette consists of 9,469 training samples

Table II
DETECTION PERFORMANCE AGAINST SOURCE-AGNOSTIC ATTACKS.

Method	Dataset	SPC		Beatrix_L		Beatrix_H		NC		ABS		STRIP		MNTD		FreeEagle		BARBIE	
		TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
Patch	MNIST	3.57%	3.35%	0.00%	**0.00%**	3.80%	4.54%	20.58%	8.14%	19.70%	4.14%	99.28%	1.46%	62.11%	37.78%	98.63%	2.74%	**100.00%**	2.29%
	CIFAR10	4.18%	6.54%	98.56%	**0.00%**	0.00%	1.08%	8.17%	5.25%	98.89%	4.05%	97.93%	4.48%	40.78%	58.89%	100.00%	4.29%	**100.00%**	3.65%
	ImageNette	2.51%	6.34%	0.00%	**0.00%**	8.90%	5.72%	9.70%	6.15%	96.86%	2.99%	16.20%	7.22%	68.48%	28.26%	90.48%	8.20%	**100.00%**	3.57%
	GTSRB	9.11%	8.02%	0.00%	**0.00%**	10.54%	6.65%	0.00%	5.07%	99.47%	3.34%	96.62%	1.50%	72.78%	27.22%	98.53%	5.88%	**100.00%**	0.29%
Blending	MNIST	4.76%	7.25%	0.00%	**0.00%**	3.30%	5.25%	55.33%	6.01%	24.35%	4.67%	0.51%	4.79%	49.44%	50.11%	97.14%	2.86%	**100.00%**	2.29%
	CIFAR10	5.12%	7.99%	93.50%	5.01%	0.22%	4.04%	18.03%	7.36%	**97.76%**	4.38%	94.81%	5.22%	43.11%	56.22%	73.91%	4.35%	97.60%	**3.65%**
	ImageNette	1.56%	6.16%	0.00%	**0.00%**	3.68%	4.97%	32.90%	7.67%	93.38%	1.49%	9.76%	4.97%	74.60%	23.81%	82.86%	11.43%	**93.65%**	3.57%
	GTSRB	5.46%	7.91%	0.00%	**0.00%**	4.72%	6.08%	3.68%	5.85%	94.74%	4.37%	97.71%	2.67%	72.15%	27.85%	96.00%	6.67%	**97.67%**	0.29%
Filter	MNIST	3.13%	4.20%	0.00%	**0.00%**	3.24%	5.13%	9.71%	5.61%	4.18%	5.83%	95.49%	4.16%	57.33%	42.33%	78.18%	5.10%	**98.00%**	2.29%
	CIFAR10	2.13%	7.44%	89.29%	6.37%	7.70%	7.12%	15.79%	6.24%	95.49%	3.94%	22.50%	6.45%	51.33%	47.78%	84.38%	5.41%	**96.80%**	**3.65%**
	ImageNette	3.53%	7.59%	0.00%	**0.00%**	0.00%	0.46%	0.00%	6.31%	84.84%	1.45%	0.00%	6.77%	74.71%	25.29%	81.04%	4.05%	**91.43%**	3.57%
	GTSRB	4.29%	5.72%	0.00%	**0.00%**	19.61%	6.46%	1.92%	4.83%	64.77%	4.71%	96.21%	5.11%	81.40%	18.61%	100.00%	4.27%	**100.00%**	0.29%
Composite	MNIST	5.91%	5.17%	0.00%	**0.00%**	2.40%	6.97%	5.20%	6.26%	53.52%	3.20%	0.00%	4.04%	20.44%	78.89%	96.31%	5.21%	**100.00%**	2.29%
	CIFAR10	7.58%	4.74%	98.11%	2.89%	0.15%	**2.48%**	11.56%	4.18%	92.81%	3.78%	0.14%	4.05%	48.22%	51.56%	67.89%	6.71%	**100.00%**	3.65%
	ImageNette	9.41%	4.91%	0.00%	**0.00%**	0.00%	3.69%	0.00%	0.75%	85.52%	2.27%	0.00%	4.55%	68.89%	30.89%	88.67%	5.70%	**100.00%**	3.57%
	GTSRB	7.50%	8.17%	0.00%	0.00%	21.21%	7.31%	19.28%	2.77%	99.47%	5.20%	0.00%	5.15%	85.78%	14.22%	98.86%	**0.00%**	**100.00%**	0.29%
Adaptive-Patch	MNIST	6.78%	6.63%	0.00%	**0.00%**	6.29%	5.29%	82.31%	4.89%	98.34%	6.35%	1.74%	4.28%	77.44%	22.56%	86.42%	7.01%	**100.00%**	2.29%
	CIFAR10	12.81%	7.06%	98.72%	**2.61%**	0.00%	3.16%	11.23%	5.61%	95.84%	3.37%	97.31%	4.65%	46.00%	53.89%	59.51%	5.23%	**100.00%**	3.65%
	ImageNette	4.93%	4.28%	0.00%	**0.00%**	0.00%	2.89%	26.72%	8.06%	95.97%	0.14%	0.0%	3.77%	51.00%	47.78%	63.07%	6.70%	**99.60%**	3.57%
	GTSRB	1.68%	6.60%	0.00%	0.00%	1.95%	4.88%	25.08%	2.67%	94.70%	3.66%	0.34%	4.51%	64.67%	35.11%	97.11%	**0.00%**	**100.00%**	0.29%
Adaptive-Blend	MNIST	14.71%	8.64%	0.00%	**0.00%**	4.12%	4.39%	29.37%	8.30%	75.26%	4.25%	3.51%	5.32%	71.22%	28.56%	23.84%	3.23%	**100.00%**	2.29%
	CIFAR10	15.81%	7.01%	98.03%	**2.31%**	0.00%	3.26%	15.95%	5.03%	87.36%	3.51%	0.98%	6.42%	33.44%	65.89%	38.76%	6.27%	**100.00%**	3.65%
	ImageNette	0.38%	2.62%	0.00%	**0.00%**	5.03%	3.89%	11.38%	6.72%	34.99%	0.27%	0.0%	2.48%	47.67%	52.22%	69.68%	7.09%	**100.00%**	3.57%
	GTSRB	1.14%	3.90%	0.00%	**0.00%**	8.06%	7.33%	31.07%	5.76%	94.77%	4.91%	0.0%	2.79%	53.89%	45.67%	94.27%	3.53%	**100.00%**	0.29%

Table III
DETECTION PERFORMANCE AGAINST SOURCE-SPECIFIC ATTACKS.

Method	Dataset	SPC		Beatrix_L		Beatrix_H		NC		ABS		STRIP		MNTD		FreeEagle		BARBIE	
		TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
Patch	MNIST	1.89%	6.04%	0.00%	**0.00%**	5.37%	5.11%	11.57%	8.50%	6.35%	3.75%	25.36%	6.43%	58.22%	41.78%	68.05%	5.56%	**94.82%**	2.29%
	CIFAR10	1.80%	5.22%	3.11%	5.39%	0.42%	4.28%	16.67%	5.27%	8.39%	4.20%	5.97%	6.39%	46.11%	52.89%	65.75%	8.22%	**92.84%**	**3.65%**
	ImageNette	0.00%	2.24%	0.00%	0.00%	0.68%	4.53%	0.00%	**0.00%**	5.08%	0.75%	5.82%	4.97%	86.81%	12.09%	73.91%	4.35%	**99.06%**	3.57%
	GTSRB	4.86%	7.37%	0.00%	**0.00%**	2.97%	5.59%	0.00%	4.09%	64.77%	5.75%	1.35%	3.78%	68.33%	31.11%	73.02%	6.35%	**100.00%**	0.29%
Blending	MNIST	4.54%	8.65%	0.00%	**0.00%**	12.09%	8.13%	19.76%	5.30%	4.81%	3.66%	14.75%	7.07%	63.22%	36.56%	77.14%	5.71%	**96.30%**	2.29%
	CIFAR10	4.95%	5.43%	24.40%	8.22%	3.44%	5.43%	10.16%	5.57%	6.12%	3.69%	3.45%	7.93%	47.78%	52.00%	71.13%	5.80%	**83.95%**	**3.65%**
	ImageNette	1.74%	3.81%	0.00%	**0.00%**	0.08%	2.26%	9.20%	4.69%	0.00%	0.30%	2.97%	5.31%	81.82%	18.18%	73.53%	5.88%	**93.03%**	3.57%
	GTSRB	6.60%	6.75%	0.00%	**0.00%**	30.54%	5.40%	1.65%	4.06%	49.92%	5.74%	0.00%	2.22%	78.75%	21.25%	71.21%	6.06%	**100.00%**	0.29%
Filter	MNIST	4.43%	5.36%	0.00%	**0.00%**	5.53%	7.79%	11.67%	4.72%	0.65%	2.67%	12.41%	6.10%	48.33%	51.44%	71.83%	3.08%	**96.76%**	2.29%
	CIFAR10	4.26%	3.81%	5.45%	5.27%	5.75%	7.19%	1.05%	4.80%	13.53%	4.51%	0.00%	5.22%	46.29%	53.26%	73.53%	4.11%	**93.33%**	**3.65%**
	ImageNette	0.00%	2.27%	0.00%	0.00%	5.51%	3.05%	7.52%	1.81%	0.98%	**0.00%**	3.72%	5.37%	83.33%	16.67%	74.24%	4.55%	**84.55%**	3.57%
	GTSRB	3.90%	7.56%	0.00%	**0.00%**	29.32%	7.78%	0.52%	3.27%	61.14%	4.92%	0.00%	1.72%	84.14%	15.86%	70.42%	4.23%	**100.00%**	0.29%
Composite	MNIST	6.33%	5.17%	0.00%	**0.00%**	6.26%	7.90%	24.51%	5.72%	43.46%	3.49%	38.83%	5.36%	46.78%	52.67%	12.82%	6.30%	**100.00%**	2.29%
	CIFAR10	15.24%	7.13%	98.48%	**0.45%**	9.88%	5.76%	13.96%	5.89%	36.30%	4.11%	21.64%	6.24%	63.22%	36.44%	48.87%	6.78%	**100.00%**	3.65%
	ImageNette	4.54%	5.43%	0.00%	**0.00%**	0.15%	3.68%	0.00%	6.07%	64.01%	1.53%	0.00%	5.67%	64.56%	35.22%	76.11%	6.34%	**100.00%**	3.57%
	GTSRB	2.78%	5.76%	0.00%	**0.00%**	4.92%	4.92%	26.61%	2.91%	39.80%	5.84%	0.21%	4.91%	94.89%	4.89%	89.58%	4.66%	**100.00%**	0.29%

and 3,925 test samples. ImageNette is a subset of 10 categories from ImageNet. (4) GTSRB consists of 39,209 training samples and 12,630 test samples, which are color images of 43 categories of traffic signs. We utilize CNN-7, VGG-16 [49], ResNet-50 [61] and GoogLeNet [62] structures to train models for these datasets respectively. For each dataset, we will generate 200 backdoored models of each backdoor attack and 200 benign models on each dataset to perform backdoored model detection experiments.

Evaluation metrics. We evaluate the performance of detection methods by computing the True Positive Rate (TPR) and False Positive Rate (FPR). TPR represents the proportion of correctly detected backdoored models to all backdoored models. FPR represents the proportion of incorrectly detected benign models to all benign models. We also calculate the F1 scores of our experiments as an evaluation metric and record it in Appendix A. All experiments are repeated five times and the average values are taken as the experimental results.

Method settings. In model split, we set the separation layer l_s as approximately the middle layer. More specifically, we set the separation layer l_s as the 4th, 10th, 22nd and 10th layer for CNN-7, VGG-16, ResNet50 and GoogLeNet. As for hyperparameters, a small α is hard to amplify backdoored latent representations, resulting in limited performance. Meanwhile a large α will transform the model output of v_{ILR}, leading to false alarms. A small ω_{max} leads to high FPRs, while a large ω_{max} leads to low TPRs. γ can help contain outliers in benign models, thereby highlighting abnormalities in backdoored models. However, for benign models without abnormally high or low RCSs, such as the CNN-7 trained on MNIST and the VGG-16 trained on CIFAR10, an excessively large γ can lead to excessive suppression of abnormalities in the backdoored models, resulting in poor detection performance. Therefore we set the α as 0.5, ω_{max} as 1, 0.75, 0.35 and 0.45 and γ as 0, 0, 0.1 and 0.1 respectively for MNIST, CIFAR10, ImageNette and GTSRB.

Baselines. We adopt seven baseline methods for comparison, i.e., FreeEagle [28], Beatrix [23], SPC [63], MNTD [27], STRIP [21], ABS [26] and NC [25]. However, their fixed detection thresholds in the corresponding papers are inapplicable to our experiments, resulting in extremely poor performance. To address this issue, we utilize 30% poisoned models and 30% benign models to determine the detection thresholds in FreeEagle, Beatrix, SPC, STRIP, ABS and NC. The thresholds are selected to improve TPR as much as possible on the condition that FPR is less than 5.00%. The thresholds will be used to detect backdoors in the remaining models. For BARBIE, we only need 30% benign models to determine normal boundaries. For MNTD, we generate 512 benign shadow models and 512 backdoored shadow models

to train the meta-classifier, which are generated based on a generic Trojan distribution proposed by MNTD, following [27]. All auxiliary models are trained on the same dataset and have the same model structure as the suspicious models that need to be detected. Beatrix and SPC are backdoored sample detection methods, in which a threshold is calculated to distinguish suspicious samples rather than suspicious models. To adapt Beatrix for model detection, we assume that the model is more likely to be backdoored if the threshold is abnormally high (denoted as Beatrix_H) or low (denoted as Beatrix_L). In SPC, samples with high $\hat{J}_t$, an index proposed in [63], are suspicious. Therefore, we only assume that models with high thresholds are backdoored in SPC.

B. Performance Against Sample-agnostic Attacks

Attack methods. In this section, we adopt backdoor attacks with different triggers, such as patch [60], blending [38], filter [64] and composite [49] triggers. The patch trigger is a pattern consisting of several pixels. The blending trigger is also a pre-chosen pattern, which attackers mix with images through transparency. The filter trigger is an image filter, such as the Nashville filter and the Gotham filter, to transform benign samples into backdoored samples. The composite trigger is composed of existing benign latent representations of multiple labels, which are noted as the trigger labels. When any input presents the combination of the trigger labels, it will be misclassified by the backdoored model. Composite backdoor attack is originally a source-specific attack, considering that only the specific combination of trigger labels will result in misclassification. We modify this method to trigger a backdoor at the appearance of any two labels as a source-agnostic attack. Filter and composite triggers are semantic triggers rather than synthetic pixel patterns, which are more difficult to detect. What's more, we also take Adaptive-Patch [59] and Adaptive-Blend [59] attacks into consideration, which design counter-samples to avoid the latent separability of backdoored samples and bypass existing latent separability based defenses. Adaptive-Patch and Adaptive-Blend attacks are source-agnostic attacks. We try to transform them into source-specific attacks but find the attack success rate is as low as 1.91% and 3.52% on CIFAR10 and GTSRB.

Source-agnostic attacks. As shown in Table II, BARBIE achieves the best performance under different source-agnostic attacks. The performance of other detection methods changes dramatically under different attack methods and datasets, while BARBIE consistently maintains TPRs over 91.43% and FPRs below 3.65%. It's notable that the FPRs of BARBIE against different backdoor attacks have the same value on the same dataset. That's because, unlike other detection methods that flexibly select thresholds, BARBIE extracts benign boundaries from benign models and utilizes them to detect all kinds of backdoored models. This further indicates that BARBIE is a robust backdoor detection method without access to adversary knowledge and data. What's more, BARBIE demonstrates excellent detection capabilities for adaptive attacks against latent separability.

Source-specific attacks. As shown in Table III, the experiment results highlight the superiority of our method. It is hard for SPC, Beatrix_L, Beatrix_H, NC, ABS and STRIP to detect backdoored models under source-specific attacks. MNTD achieves high TPRs and FPRs at the same time, which is not conducive to backdoor detection. FreeEagle achieves an average TPR of 57.46%, 64.82%, 74.45% and 76.06% and an average FPR of 5.16%, 6.23%, 5.28% and 5.33% on MNIST, CIFAR10, GTSRB and ImageNette respectively. Compared to FreeEagle, the TPR of BARBIE improves by an average of 39.51%, 27.71%, 19.71% and 23.94% respectively on these four datasets. The FPR of BARBIE decreases to 2.29%, 3.65%, 3.57% and 0.29% respectively. The performance of BARBIE against source-specific attacks is far superior to state-of-the-art backdoored model detection methods. However, the performance of BARBIE against source-specific attacks still needs improvement. According to the ablation study results in Section V-I, we can observe that it's because BARBIE adopts the boundary expansion method to exchange high TPRs for low FPRs. Although BARBIE strikes a good balance between TPR and FPR, a better detection boundary determination method can further enhance its effectiveness.

C. Performance Against Sample-specific Attacks

Attack methods. In this section, we select the Input-Aware Dynamic Backdoor Attack [44] as the attack method, which uses a trigger generator network to produce a unique trigger for each sample. The trigger of a sample does not work on other samples. The existence of multiple triggers makes them hard to detect. Based on the type of target labels, sample-specific attacks can be divided into all-to-one and all-to-all attacks. Different from the only common target label of all poisoned samples in all-to-one attacks, all-to-all attacks can assign each poisoned sample an arbitrary label as the target label and can target different classes.

All-to-one attacks. As shown in Table IV, BARBIE performs well in all-to-one attacks. Although the attack method generates unique triggers for each sample, BARBIE can still detect differences between backdoored models and benign models with TPRs of nearly 100.00% and FPRs below 3.65%, which indicates that varying triggers in all-to-all attacks still exhibit abnormal control over the model output, measured by RCS. In comparison, dynamic triggers cause great difficulties for other detection methods and the performance of other detection methods varies with changes in the dataset.

All-to-all attacks. As shown in Table IV, BARBIE achieves TPRs of 100.00% and low FPRs on all datasets, which is the best performance among the detection methods. This proves that the existence of multi-target labels does not dilute the anomalies present in backdoored models. In contrast, dynamic triggers targeting multiple target labels do cause enormous difficulties for other detection methods. The state-of-the-art backdoored model detection method FreeEagle performs well on CIFAR10 and GTSRB, but fails on MNIST and ImageNette with an average TPR of 35.26%.

Table IV
DETECTION PERFORMANCE AGAINST SAMPLE-SPECIFIC ATTACKS.

Method	Type	Dataset	SPC		Beatrix_L		Beatrix_H		NC		ABS		STRIP		MNTD		FreeEagle		BARBIE	
			TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
Input-Aware	All-to-One	MNIST	4.73%	6.78%	0.00%	**0.00%**	7.85%	7.66%	28.03%	6.59%	48.91%	4.01%	16.18%	8.30%	39.00%	60.44%	17.95%	7.07%	**99.89%**	2.29%
		CIFAR10	22.88%	6.01%	96.72%	**0.30%**	0.00%	3.59%	0.00%	4.41%	6.54%	4.03%	10.00%	6.30%	53.00%	46.56%	32.79%	7.18%	**100.00%**	3.65%
		ImageNette	1.72%	4.54%	0.00%	0.00%	4.11%	3.75%	0.00%	**0.00%**	65.19%	1.48%	0.00%	2.73%	61.71%	37.95%	58.38%	8.13%	**100.00%**	3.57%
		GTSRB	0.94%	5.47%	0.00%	0.00%	11.40%	7.64%	0.00%	**0.00%**	98.35%	4.86%	0.00%	2.26%	53.67%	46.22%	98.45%	0.00%	**100.00%**	0.29%
	All-to-All	MNIST	13.74%	7.50%	0.00%	**0.00%**	7.59%	6.89%	8.47%	5.91%	27.27%	4.82%	1.95%	2.68%	65.44%	34.11%	33.77%	5.06%	**100.00%**	2.29%
		CIFAR10	13.97%	6.36%	96.84%	**0.22%**	0.00%	2.57%	10.58%	6.43%	4.20%	4.20%	2.01%	5.27%	24.11%	75.67%	87.94%	5.23%	**100.00%**	3.65%
		ImageNette	6.56%	4.41%	0.00%	0.00%	1.43%	3.07%	0.00%	**0.00%**	7.33%	0.59%	0.00%	3.61%	67.78%	31.67%	36.75%	8.06%	**100.00%**	3.57%
		GTSRB	39.01%	7.06%	0.00%	**0.00%**	3.09%	6.03%	16.00%	5.13%	99.18%	4.89%	0.00%	1.69%	75.89%	24.00%	88.48%	4.49%	**100.00%**	0.29%

Table V
DETECTION PERFORMANCE AGAINST CLEAN-LABEL ATTACKS.

Method	Dataset	SPC		Beatrix_L		Beatrix_H		NC		ABS		STRIP		MNTD		FreeEagle		BARBIE	
		TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
NARCISSUS	MNIST	17.10%	6.93%	0.00%	**0.00%**	4.04%	4.98%	6.19%	6.56%	4.88%	6.53%	32.49%	5.23%	53.89%	46.11%	56.63%	8.45%	**100.00%**	2.29%
	CIFAR10	23.79%	5.85%	**99.03%**	4.07%	0.00%	2.30%	5.81%	4.46%	97.92%	4.86%	93.26%	5.02%	22.00%	77.89%	94.48%	**1.38%**	96.81%	3.65%
	ImageNette	0.00%	3.09%	0.00%	**0.00%**	6.68%	7.34%	5.77%	5.55%	83.94%	6.33%	0.00%	4.51%	52.89%	46.56%	88.22%	6.65%	**100.00%**	3.57%
	GTSRB	1.84%	7.07%	0.00%	0.00%	28.18%	5.88%	85.91%	7.63%	99.42%	2.85%	0.00%	4.62%	66.89%	32.22%	98.00%	**0.00%**	**100.00%**	0.29%
Data-free Backdoor	MNIST	4.09%	4.01%	0.00%	0.00%	11.88%	6.75%	25.72%	8.31%	36.29%	3.68%	99.00%	**0.00%**	40.00%	60.00%	95.30%	0.00%	**100.00%**	2.29%
	CIFAR10	8.25%	5.66%	16.32%	6.36%	98.79%	2.17%	98.12%	4.30%	100.00%	4.11%	0.00%	1.74%	44.11%	55.11%	97.32%	**0.41%**	**100.00%**	3.65%
	ImageNette	3.31%	4.25%	0.00%	**0.00%**	4.64%	5.54%	30.83%	6.17%	8.92%	0.30%	97.23%	6.21%	57.56%	42.22%	75.54%	5.57%	**100.00%**	3.57%
	GTSRB	0.00%	3.84%	0.00%	**0.00%**	3.20%	5.23%	0.00%	1.22%	0.53%	3.75%	0.00%	1.82%	85.78%	14.00%	99.48%	3.78%	**100.00%**	0.29%

D. Performance Against Clean-label Attacks

Attack methods. In this section, we choose the NARCISSUS [57] and a data-free backdoor injection approach [58] as attack methods. NARCISSUS leverages the public out-of-distribution and target-label examples to produce a surrogate model, which can generate an effective NARCISSUS trigger. The data-free backdoor injection approach designs a novel loss function for fine-tuning the original model into the backdoored one using the substitute data.

Clean-label attacks. As shown in Table V, BARBIE presents good performance on these four datasets. BARBIE acquires an average TPR of 99.60% and an average FPR of 2.45% against clean-label attacks. These results reveal that the specific technique of backdoor insertion does not affect the detection of anomalies in the RCS matrix of backdoored models by BARBIE.

E. Performance on Large Datasets

The above experiments have confirmed the effectiveness of BARBIE on small datasets. To validate the effectiveness of BARBIE on large datasets with a large number of classes, we conduct experiments on CIFAR100 [30] and TinyImageNet [65] and adopt ResNet-50 [61] as the model structure.

Table VI
DETECTION PERFORMANCE ON LARGE DATASETS AGAINST SOURCE-AGNOSTIC ATTACKS.

Method	Dataset	ABS		STRIP		FreeEagle		BARBIE	
		TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
Patch	CIFAR100	**100.00%**	**0.00%**	99.09%	3.86%	99.47%	3.52%	90.32%	5.28%
	TinyImageNet	**99.85%**	**0.45%**	98.88%	1.93%	0.00%	2.81%	91.67%	5.47%
Blending	CIFAR100	**99.85%**	**0.00%**	97.21%	5.96%	55.69%	1.95%	89.74%	5.28%
	TinyImageNet	70.09%	**1.48%**	98.51%	2.92%	0.00%	1.56%	**100.00%**	5.47%
Filter	CIFAR100	**99.63%**	**0.00%**	33.97%	6.15%	80.29%	3.43%	96.67%	5.28%
	TinyImageNet	81.94%	**1.49%**	97.76%	2.23%	5.67%	5.86%	**100.00%**	5.47%
Composite	CIFAR100	100.00%	**0.00%**	99.01%	3.35%	89.55%	6.14%	**100.00%**	5.28%
	TinyImageNet	98.71%	**0.23%**	93.42%	5.62%	86.02%	5.59%	**100.00%**	5.47%
Adaptive-Patch	CIFAR100	100.00%	**0.00%**	97.51%	2.45%	89.05%	2.70%	**100.00%**	5.28%
	TinyImageNet	97.97%	**0.89%**	74.48%	5.51%	47.91%	7.58%	**100.00%**	5.47%
Adaptive-Blend	CIFAR100	38.83%	**0.00%**	44.79%	6.58%	3.10%	4.73%	**100.00%**	5.28%
	TinyImageNet	0.60%	**0.23%**	0.00%	4.66%	5.62%	5.32%	**100.00%**	5.47%

Baseline and method settings. We select FreeEagle [28], STRIP [21] and ABS [26] as the baselines, which perform well on small datasets. As for hyperparameters, we set α as 0.5, γ as 0.5 and ω_{max} as 0.65 and 1.25 for CIFAR100 and TinyImageNette respectively.

Table VII
DETECTION PERFORMANCE ON LARGE DATASETS AGAINST SOURCE-SPECIFIC ATTACKS.

Method	Dataset	ABS		STRIP		FreeEagle		BARBIE	
		TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
Patch	CIFAR100	0.00%	**0.00%**	12.49%	6.48%	5.08%	5.74%	**100.00%**	5.28%
	TinyImageNet	0.00%	**2.41%**	10.00%	8.65%	0.00%	3.26%	**100.00%**	5.47%
Blending	CIFAR100	0.00%	**0.00%**	8.14%	4.82%	17.06%	4.56%	**100.00%**	5.28%
	TinyImageNet	0.00%	**1.05%**	5.40%	5.62%	0.00%	1.71%	**100.00%**	5.47%
Filter	CIFAR100	0.00%	**0.00%**	27.34%	6.54%	5.77%	6.75%	**100.00%**	5.28%
	TinyImageNet	0.00%	**0.60%**	11.11%	5.30%	0.00%	2.86%	**100.00%**	5.47%
Composite	CIFAR100	89.86%	**0.00%**	99.04%	4.41%	85.60%	4.66%	**100.00%**	5.28%
	TinyImageNet	1.64%	**0.30%**	93.82%	5.83%	91.31%	3.75%	**100.00%**	5.47%

Table VIII
DETECTION PERFORMANCE ON LARGE DATASETS AGAINST SAMPLE-SPECIFIC ATTACKS.

Method	Dataset	ABS		STRIP		FreeEagle		BARBIE	
		TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
All-to-One	CIFAR100	99.33%	**0.00%**	48.82%	6.37%	26.93%	7.74%	**100.00%**	5.28%
	TinyImageNet	**97.08%**	**0.30%**	0.00%	3.91%	41.63%	5.93%	90.00%	5.47%
All-to-All	CIFAR100	0.00%	**0.00%**	58.07%	6.77%	19.24%	6.05%	**97.50%**	5.28%
	TinyImageNet	0.00%	**0.00%**	0.00%	1.52%	32.68%	4.12%	**97.14%**	5.47%

Performance. As shown in Table VI-VIII, BARBIE maintains excellent and robust detection capability on large datasets. As for clean-label attacks, NARCISSUS [57] and a data-free backdoor injection approach [58] can't successfully attack the ResNet-50 trained on CIFAR100 and TinyImageNet. Thus we do not show their experiment results.

F. Performance Against Adaptive Attacks

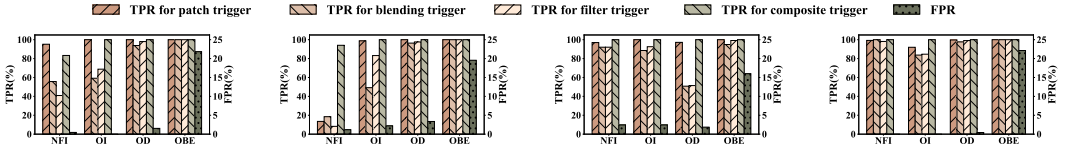
Here we consider two possible methods to evade our detection.

1) Similar Latent Representation: The first way is to make backdoored latent representations similar to benign latent representations. We can achieve this goal by adopting an additional loss function as shown in Eq. 21.

$$loss_{similarity} = MSE(f_e(\widetilde{x}), f_e(x))), \tag{21}$$

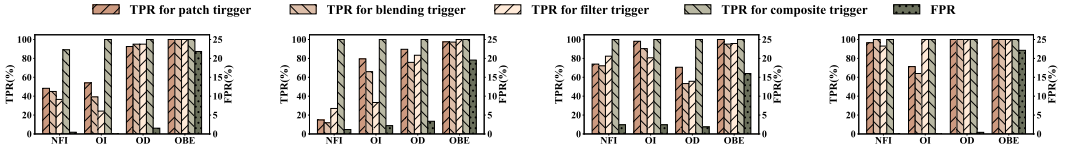
where MSE is the mean square loss and can measure the divergence between the latent representations of backdoored samples $f_e(\widetilde{x})$ and benign latent representations $f_e(x)$. In order not to affect the performance of backdoor attacks, we select the benign samples x from the target label by random or fixed-point strategy.

Performance against the Similar Latent Representation Attack. In these experiments, we adopt the patch trigger and carry out source-agnostic and source-specific attacks on MNIST, CIFAR10, ImageNette and GTSRB. Our experimental



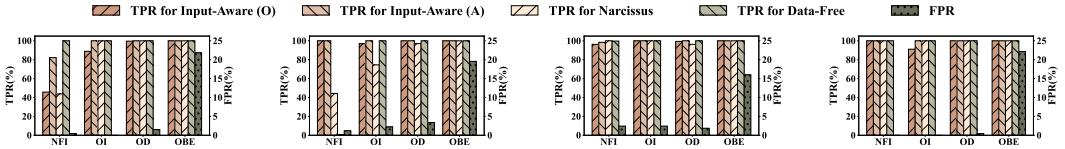
(a) MNIST (b) CIFAR10 (c) ImageNette (d) GTSRB

Figure 3. Ablation study against source-agnostic attacks. We display the experimental results of normal latent representation inversion (NFI), detection based on RCS values (OI), detection based on RCS distribution (OD) and detection without boundary expansion (OBE).



(a) MNIST (b) CIFAR10 (c) ImageNette (d) GTSRB

Figure 4. Ablation study against source-specific attacks. We display the experimental results of normal latent representation inversion (NFI), detection based on RCS values (OI), detection based on RCS distribution (OD) and detection without boundary expansion (OBE).



(a) MNIST (b) CIFAR10 (c) ImageNette (d) GTSRB

Figure 5. Ablation study against sample-specific and clean-label attacks. We display the experimental results of normal latent representation inversion (NFI), detection based on RCS values (OI), detection based on RCS distribution (OD) and detection without boundary expansion (OBE). The all-to-one attack and all-to-all attack are denoted as O and A, respectively.

Table IX
DETECTION PERFORMANCE OF BARBIE AGAINST THE SIMILAR LATENT REPRESENTATION ATTACK.

Method			MNIST	CIFAR10	ImageNette	GTSRB
Source-Agnostic	Random	TPR	99.69%	100.00%	100.00%	100.00%
		FPR	2.29%	3.65%	3.57%	0.29%
		F1	99.05%	98.74%	98.77%	99.90%
	Fixed-point	TPR	100.00%	100.00%	100.00%	100.00%
		FPR	2.29%	3.65%	3.57%	0.29%
		F1	99.20%	98.74%	98.77%	99.90%
Source-Specific	Random	TPR	100.00%	100.00%	100.00%	100.00%
		FPR	2.29%	3.65%	3.57%	0.29%
		F1	99.20%	98.74%	98.77%	99.90%
	Fixed-point	TPR	100.00%	100.00%	100.00%	100.00%
		FPR	2.29%	3.65%	3.57%	0.29%
		F1	99.20%	98.74%	98.77%	99.90%

results are shown in Table IX. BARBIE effectively resists the Similar Latent Representation Attack with nearly 100.00% TPRs, low FPRs and nearly 100.00% F1 scores.

2) Diverse Latent Representation: The second way is to prevent backdoored latent representations from being inverted. We try to disperse backdoored samples not only in input space but also in latent space to reduce the possibility. This method is based on the previous work, the Input-Aware Dynamic Backdoor Attack [44]. In [44], Eq. 22 is adopted to generate diverse triggers for different samples.

$$loss_{diversity} = \frac{||x_i - x_j||}{||g(x_i) - g(x_j)||}, \tag{22}$$

where g refers to the trigger generator. x_i and x_j are two different benign samples. Taking inspiration from this, we construct the following loss function to ensure the dispersion of backdoored samples in latent space.

$$loss'_{diversity} = \frac{||x_i - x_j||}{||f_e(\widetilde{x}_i) - f_e(\widetilde{x}_j)||}, \tag{23}$$

By employing these loss functions, we can disperse backdoored samples in both input space and latent space.

Table X
DETECTION PERFORMANCE OF BARBIE AGAINST THE DIVERSE LATENT REPRESENTATION ATTACK.

Method		MNIST	CIFAR10	ImageNette	GTSRB
All-to-One	TPR	100.00%	100.00%	100.00%	100.00%
	FPR	2.29%	3.65%	3.57%	0.29%
	F1	99.20%	98.74%	98.77%	99.90%
All-to-All	TPR	100.00%	100.00%	100.00%	100.00%
	FPR	2.29%	3.65%	3.57%	0.29%
	F1	99.20%	98.74%	98.77%	99.90%

Performance against the Diverse Latent Representation Attack. In these experiments, we carry out all-to-one and all-to-all attacks on MNIST, CIFAR10, ImageNette and GTSRB. Our experimental results are shown in Table X. BARBIE effectively resists the Diverse Latent Representation Attack with 100.00% TPRs, low FPRs and nearly 100.00% F1 scores.

In all the above experiments, our method extracts the benign boundaries and leverages them to detect all backdoor attacks. These experimental results strongly demonstrate that BARBIE is effective and applicable to different datasets and various backdoor attacks and has strong resistance to adaptive attacks.

G. Performance on Vision Transformer

We also conduct experiments on vision transformer to demonstrate that BARBIE can be applied to different model structures. As shown in Table XI-XII, BARBIE performs well on vision transformer.

Experiment settings. In this part, we adopt MNIST [29], CIFAR10 [30], ImageNette [32] and GTSRB [31] as datasets and select DeiT [66] as the model structure. We produce 200 backdoored models for each attack and 200 benign models on each dataset for backdoored model detection. As for model

separation, we divide the vision transformer into two parts, with the MLP Head as the classifier and the rest as the feature extractor. As for hyperparameters, we set α as 0, γ as 0 and w_{max} as 0.15 on MNIST, CIFAR10, ImageNette and GTSRB.

Table XI
DETECTION PERFORMANCE OF BARBIE ON VISION TRANSFORMER AGAINST SOURCE-AGNOSTIC ATTACKS.

Dataset	Patch		Blending		Filter	
	TPR	FPR	TPR	FPR	TPR	FPR
MNIST	91.88%	2.41%	94.86%	2.41%	97.46%	2.41%
CIFAR10	100.00%	2.50%	100.00%	2.50%	100.00%	2.50%
ImageNette	100.00%	0.69%	100.00%	0.69%	100.00%	0.69%
GTSRB	100.00%	0.39%	100.00%	0.39%	100.00%	0.39%

Table XII
DETECTION PERFORMANCE OF BARBIE ON VISION TRANSFORMER AGAINST SOURCE-SPECIFIC ATTACKS.

Dataset	Patch		Blending		Filter	
	TPR	FPR	TPR	FPR	TPR	FPR
MNIST	97.75%	2.41%	97.90%	2.41%	94.88%	2.41%
CIFAR10	100.00%	2.50%	100.00%	2.50%	100.00%	2.50%
ImageNette	100.00%	0.69%	100.00%	0.69%	100.00%	0.69%
GTSRB	100.00%	0.39%	100.00%	0.39%	100.00%	0.39%

H. Performance in Self-supervised Learning

We also consider the scenario of self-supervised learning to demonstrate the scalability of our method. Self-supervised learning is utilized to generate an encoder from a pre-training dataset, which outputs similar latent representations for semantically similar inputs. Based on the pre-trained encoder, the model owner can make use of different downstream datasets to train different classifiers, which are capable of mapping the latent representations to labels. Our method is also effective in detecting backdoored models in self-supervised learning.

Datasets and models. In self-supervised learning, we adopt CIFAR10 [30] as the pre-training dataset and select SVHN [67] and GTSRB [31] as the downstream dataset respectively. As for encoders, we use the SimCLR [68], a popular self-supervised learning algorithm, to train a Resnet18 [61] as an encoder. As for classifiers, we use a fully connected neural network with two hidden layers following existing works BadEncoder [69] and DRUPE [70]. We produce 200 backdoored encoders and classifiers for each backdoor attack and 200 benign encoders and classifiers on each dataset to perform backdoor detection experiments.

Attack methods. We adopt BadEncoder [69] and DRUPE [70] as attack methods. BadEncoder aims to minimize the distance between the output latent representations of backdoored samples and some representative samples of the target label in the downstream datasets. Therefore, the downstream classifiers will categorize backdoored samples into the target label. In order to improve the invisibility of backdoor attacks, DRUPE reduces the sliced-Wasserstein distance [71] between the distributions of backdoored and clean samples to transform poisoned samples into in-distribution data.

Baseline and method settings. We select FreeEagle [28] as the baselines in self-supervised learning, which is similar to BARBIE and performs well in the above experiments. FreeEagle and BARBIE detect the whole model made up of an encoder and a classifier and set the separation layer l_s as the middle layer of the encoder. As for hyperparameters, we set α as 0.001, γ as 0 and ω_{max} as 0.5 and 0.45 for SVHN and GTSRB respectively. In self-supervised learning, it's hard to detect backdoored models, considering that latent representations are clustered based on labels, which is conducive to covert backdoor attacks. Therefore we make a minor change to make our method more suitable for this scenario. We subtract the mean indicators of benign models from indicators of unknown models and make use of the results to detect backdoored models.

Performance. The experimental results are displayed in Table XIII. In these experiments, FreeEagle fails to detect backdoored models in self-supervised learning. Our method is capable of detecting backdoored models with TPRs over 97.78% and FPRs below 5.93% against BadEncoder. When the attacker adopts DRUPE, BARBIE can also detect at least 74.44% backdoored models with FPRs below 5.93%. Although there are significant differences in the mechanisms of these backdoor attacks in self-supervised learning, the backdoored models still exhibit abnormalities in the RCS metric.

Table XIII
DETECTION PERFORMANCE IN SELF-SUPERVISED LEARNING.

Method	Pre-training Dataset	Downstream Dataset	FreeEagle		BARBIE	
			TPR	FPR	TPR	FPR
BadEncoder	CIFAR10	SVHN	0.08%	1.21%	97.78%	5.93%
		GTSRB	8.08%	7.14%	98.99%	5.82%
DRUPE	CIFAR10	SVHN	9.38%	5.14%	74.44%	5.93%
		GTSRB	46.89%	4.82%	85.98%	5.82%

I. Ablation Study

In this section, we decompose our method into three important parts: (1) latent representation inversion, (2) detection indicators, and (3) boundary expansion. To demonstrate the effectiveness of our method, we make modifications to each part as comparative experiments. For latent representation inversion, we only adopt the normal latent representation inversion and calculate the RCS of v_{OLR} on v_{OLR} as a control group. For detection indicators, we identify anomalies in RCS values or RCS distribution separately to demonstrate the importance and comprehensiveness of detecting both aspects. For boundary expansion, we abandon the boundary expansion method for comparison.

Normal latent representation inversion. We present the experimental results in Figure 3-5. Without our latent representation inversion method, the TPR of BARBIE decreases to an average of 63.82%, 41.31%, 92.00% and 98.92% on MNIST, CIFAR10, ImageNette and GTSRB, and FPRs are 0.46%, 1.17%, 2.43% and 0.00% respectively. The results prove the superiority of our latent representation inversion method in amplifying differences between backdoored models and benign models.

Decomposition of detection indicators. The results of only detecting anomalies in RCS values or RCS distribution are shown in Figure 3-5. The abnormalities in RCS distribution are obvious on MNIST, CIFAR10 and GTSRB with TPR over 75.80% and FPR below 3.36%. But for ImageNette, the differences between backdoored and benign models mainly embody in RCS values with TPRs over 88.57% against

source-agnostic attacks, 80.61% against source-specific attacks, 100.00% against sample-specific attacks and 100.00% against clean-label attacks and FPRs of 2.43%. That's because there are significant differences between the RCSs of latent representations in the benign ResNet-50 trained on ImageNette, which makes the anomalies in RCS distribution inconspicuous. Therefore it's necessary to detect both aspects.

Detection without boundary expansion. As we can see in Figure 3-5, both TPRs and FPRs are improved without boundary expansion. In these experiments, TPRs are over 94.92% against all kinds of backdoor attacks. However, the FPR achieves 21.83% on MNIST, 19.56% on CIFAR10, 16.00% on ImageNette and 22.14% on GTSRB. Compared to the results in Table II-V, we demonstrate that our boundary expansion method strikes a balance between TPRs and FPRs.

J. Impact of Hyperparameters

Due to the page limit, we show the hyperparameter experiment results in Appendix B.

Impact of l_s. We explore the impact of the separation layer l_s on the performance of BARBIE. We set l_s as the 7th, 10th and 13th layer on VGG16 trained on CIFAR10, and the 10th, 22nd and 40th layer on ResNet50 trained on ImageNette. As we can see in Figure 6-8, selecting the middle layer as the separation layer usually yields the best performance. The possible reason may be that shallow layers cannot well capture latent representations, while deep layers mix latent representations with label information. Thus the middle layer is an appropriate choice.

Impact of α. We explore the impact of α on the performance of BARBIE. According to the experimental results in Section V-I, BARBIE performs well on ImageNette and GTSRB even when α is set to be zero (i.e., normal latent representation inversion), which demonstrates that the performance of BARBIE on ImageNette and GTSRB is almost unaffected by the hyperparameter α. Therefore we set α as 0.00, 0.05, 0.10, 0.25 and 0.50 on MNIST and CIFAR10. As we can see in Figure 9-11, α is an important parameter in BARBIE. During the process of changing α from 0.00 to 0.50, the average TPR of BARBIE on MNIST gradually increases from 63.82% to 98.81% and the average TPR of BARBIE on CIFAR10 gradually increases from 41.31% to 96.78%.

Impact of ω_{max}. We explore the impact of ω_{max} on the performance of BARBIE. Here we set ω_{max} as 0.6, 0.8, 1.0, 1.2 and 1.4 times the original values on MNIST, CIFAR10, ImageNette and GTSRB. The results are displayed in Figure 12-14. As ω_{max} increases, the TPR and FPR gradually decrease. That's because a higher ω_{max} results in a looser standard of detection, which is more inclined to categorize unknown models into benign models.

Impact of γ. We explore the impact of γ on the performance of BARBIE. Here we set γ as 0, 0.02, 0.05, 0.08 and 0.10 on MNIST and CIFAR10, and 0.06, 0.08, 0.10, 0.12 and 0.14 on ImageNette and GTSRB. The results are displayed in Figure 15-17. BARBIE maintains good performance with a low γ. However, a high γ may change the model output of $v'_{OLR,k}$ in benign models and suppress anomalies in backdoored models by adding significant disturbances, resulting in low TPRs and high FPRs.

Impact of accessible model ratio. We explore the impact of the number of accessible benign models on the performance of BARBIE. We set the accessible model ratio as 10%, 30%, 50%, 70% and 90% of benign models on MNIST, CIFAR10, ImageNette and GTSRB. The experimental results are shown in Figure 18-20. As the accessible model ratio increases, FPR presents a sharp decline and TPR drops slowly. The results stem from the same reason as described in ω_{max} parameter experiments. With a fixed ω_{max}, a larger accessible model ratio results in a looser detection standard.

VI. DISCUSSION

A. Detection with a Poisoned Model Zoo

The benign models required by BARBIE may not be trustworthy, e.g., the defender may download models from a model zoo that contains some backdoored models. Therefore we evaluate the performance of BARBIE and other detection methods under a practical scenario in which the benign models are contaminated by 5% or 10% backdoored models. Due to the existence of backdoored models, we adopt the modified z-score method, a common anomaly identification method, in detection methods to filter out outliers. The results in Table XIV verify the ability of BARBIE to distinguish between benign models and backdoored models in such a practical scenario. We also demonstrate the experimental results of ABS, STRIP and FreeEagle in Appendix C.

Table XIV
DETECTION PERFORMANCE OF BARBIE WITH A POISONED MODEL ZOO. THE CONTENT OF EACH CELL REPRESENTS TPR/FPR.

Poison Rate	Method		MNIST	CIFAR10	ImageNette	GTSRB
5%	Source-Agnostic	Patch	100.00%/3.27%	100.00%/4.33%	100.00%/6.17%	100.00%/0.14%
		Blending	93.76%/3.50%	97.27%/4.49%	91.67%/6.99%	98.03%/1.20%
		Filter	97.30%/3.94%	97.12%/4.73%	93.47%/5.68%	99.74%/0.35%
		Composite	100.00%/3.18%	100.00%/4.98%	100.00%/4.89%	100.00%/0.42%
	Source-Specific	Patch	92.10%/2.94%	91.67%/5.39%	98.62%/6.52%	100.00%/0.25%
		Blending	93.95%/2.53%	82.10%/4.57%	90.32%/6.52%	100.00%/0.81%
		Filter	92.96%/3.41%	83.33%/5.74%	77.05%/6.76%	100.00%/0.60%
		Composite	100.00%/4.09%	100.00%/4.80%	100.00%/6.11%	100.00%/0.21%
	Sample-Specific	All-to-One	99.07%/2.77%	100.00%/5.46%	100.00%/5.94%	100.00%/0.00%
		All-to-All	100.00%/3.02%	100.00%/4.85%	100.00%/6.29%	100.00%/0.21%
	Clean-Label	Narcissus	100.00%/2.83%	96.31%/4.94%	100.00%/5.76%	100.00%/0.21%
		Data-free	100.00%/2.29%	100.00%/4.54%	100.00%/4.49%	100.00%/0.84%
10%	Source-Agnostic	Patch	100.00%/2.47%	100.00%/4.50%	100.00%/6.24%	100.00%/0.35%
		Blending	95.28%/4.82%	100.00%/5.59%	91.19%/6.66%	97.37%/0.82%
		Filter	90.22%/2.46%	97.78%/4.58%	93.01%/6.18%	100.00%/0.00%
		Composite	100.00%/5.11%	100.00%/4.73%	100.00%/6.64%	100.00%/0.27%
	Source-Specific	Patch	94.20%/3.46%	92.44%/5.27%	100.00%/6.23%	99.23%/0.83%
		Blending	95.45%/3.81%	81.91%/5.67%	91.45%/7.43%	100.00%/0.93%
		Filter	89.67%/3.05%	87.50%/4.91%	81.40%/6.55%	100.00%/0.93%
		Composite	100.00%/5.72%	100.00%/4.36%	100.00%/7.19%	100.00%/0.47%
	Sample-Specific	All-to-One	100.00%/2.81%	100.00%/4.16%	100.00%/6.67%	100.00%/1.28%
		All-to-All	98.69%/2.13%	100.00%/4.96%	100.00%/5.45%	100.00%/2.10%
	Clean-Label	Narcissus	100.00%/3.87%	97.22%/5.07%	100.00%/6.82%	100.00%/0.68%
		Data-free	100.00%/2.85%	100.00%/4.33%	100.00%/5.49%	100.00%/0.89%

B. Detection with Substitute Benign Models

In this part, we explore the possibility of relaxing the model restrictions. Instead of using the corresponding benign models, we attempt to detect backdoored models with substitute benign models, which perform similar tasks and have the same model structure as the benign models. Considering the difference between benign models and substitute benign models, we

will select a larger parameter ω_{max} and relax the judgment standard of benign models to detect less than p anomalies in indicators. We use substitute benign models trained on FashionMNIST, SVHN, FashionMNIST and STL10 to diagnose models trained on MNIST, MNIST, CIFAR10 and ImageNette respectively. ω_{max} is set as 4.50, 4.50, 0.85 and 1.30 and p is set as 5, 5, 6 and 4 in these four experiments respectively based on empirical experiments. Other hyperparameters are consistent with that of Section V.

As we can see in Table XV, BARBIE still achieves good performance in this scenario. In these four experiments, BARBIE achieves an average TPR of 94.65%, 90.58%, 78.74% and 86.57% and an average FPR of 0.86%, 3.21%, 5.74% and 6.53% respectively. We also demonstrate the experimental results of ABS, STRIP and FreeEagle in Appendix D. The experiments show that BARBIE has the potential to overcome the need for benign models. Considering the abnormal behavior of backdoored models reflected in RCS, the difference between backdoored models and the substitute benign models in these indicators may be greater than that between benign models and substitute benign models. How to improve the performance of BARBIE with such limitations is a possible future direction.

Table XV
DETECTION PERFORMANCE OF BARBIE ON TARGETED BENIGN MODELS WITH SUBSTITUTE BENIGN MODELS. THE CONTENT OF EACH CELL REPRESENTS TPR/FPR.

Targeted	Substitute	MNIST FashionMNIST	MNIST SVHN	CIFAR10 FashionMNIST	ImageNette STL10
Source-Agnostic	Patch	100.00%/0.86%	100.00%/3.21%	93.60%/5.74%	96.88%/6.53%
	Blending	93.20%/0.86%	84.00%/3.21%	50.00%/5.74%	61.91%/6.53%
	Filter	98.00%/0.86%	94.00%/3.21%	78.40%/5.74%	71.43%/6.53%
	Composite	100.00%/0.86%	100.00%/3.21%	100.00%/5.74%	100.00%/6.53%
Source-Specific	Patch	81.73%/0.86%	83.95%/3.21%	59.75%/5.74%	84.38%/6.53%
	Blending	81.48%/0.86%	70.37%/3.21%	55.56%/5.74%	63.64%/6.53%
	Filter	89.14%/0.86%	64.20%/3.21%	43.33%/5.74%	72.73%/6.53%
	Composite	100.00%/0.86%	100.00%/3.21%	100.00%/5.74%	100.00%/6.53%
Sample-Specific	All-to-One	92.25%/0.86%	90.39%/3.21%	99.66%/5.74%	99.01%/6.53%
	All-to-All	100.00%/0.86%	100.00%/3.21%	100.00%/5.74%	100.00%/6.53%
Clean-Label	Narcissus	100.00%/0.86%	100.00%/3.21%	64.57%/5.74%	88.89%/6.53%
	Data-free	100.00%/0.86%	100.00%/3.21%	100.00%/5.74%	100.00%/6.53%

C. Modalities

Although we have explored the effectiveness of BARBIE in the image domain, there are various machine learning tasks leveraging information from various modalities, such as computer vision [30], [61], [72], [73], [74], [75], natural language processing [76], [77], [78], [79], [80], [81] and acoustics signal processing [5], [82], [83], [84], [85], [86]. BARBIE is applicable to different modalities considering that Beatrix [23] and TED [24], two latent separability based methods, have been proven effective in the audio domain and natural language domain respectively, which demonstrates that the latent separability is feasible for backdoor detection in different modalities. Therefore applying BARBIE to different modalities is possible and may be a future research direction for us.

VII. CONCLUSION

We propose a new and robust kind of latent separability, namely the relative competition score, which leverages the particularity of backdoor attacks to deeply explore differences between latent representations and makes it hard for adaptive attacks to compromise. What's more, we design a data-free backdoored model detection method based on RCS. With a novel latent representation inversion method, we calculate RCS based on inverted latent representations, which reflect the differences between benign models and backdoored models. We propose a series of comprehensive indicators based on RCS to concretize the difference and distinguish backdoored models. Expansive experiments prove the effectiveness of our method in detecting a wide range of backdoor attacks in a robust way, which improves the average TPR by 17.05% against source-agnostic attacks, 27.72% against source-specific attacks, 43.17% against sample-specific attacks, 11.48% against clean-label attacks, 33.37% against adaptive attacks and maintains lower FPRs, compared to the state-of-the-art data-free backdoored model detection method.

ACKNOWLEDGMENT

This work is supported by China NSFC Grant 61925109.

REFERENCES

[1] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014.
[2] O. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC 2015-Proceedings of the British Machine Vision Conference 2015*, 2015.
[3] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015.
[4] B. McCann, J. Bradbury, C. Xiong, and R. Socher, "Learned in translation: Contextualized word vectors," *Advances in neural information processing systems*, vol. 30, 2017.
[5] M. A. Di Gangi, M. Negri, and M. Turchi, "Adapting transformer to end-to-end spoken language translation," in *Proceedings of INTERSPEECH 2019*, 2019.
[6] S. Ranathunga, E.-S. A. Lee, M. Prifti Skenduli, R. Shekhar, M. Alam, and R. Kaur, "Neural machine translation for low-resource languages: A survey," *ACM Computing Surveys*, vol. 55, no. 11, pp. 1–37, 2023.
[7] K. Zheng, Q. Zheng, H. Yang, L. Zhao, L. Hou, and P. Chatzimisios, "Reliable and efficient autonomous driving: the need for heterogeneous vehicular networks," *IEEE Communications Magazine*, vol. 53, no. 12, pp. 72–79, 2015.
[8] C. Sitawarin, A. N. Bhagoji, A. Mosenia, M. Chiang, and P. Mittal, "Darts: Deceiving autonomous cars with toxic signs," *arXiv preprint arXiv:1802.06430*, 2018.
[9] Z. Sheng, Y. Xu, S. Xue, and D. Li, "Graph-based spatial-temporal convolutional network for vehicle trajectory prediction in autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 10, pp. 17 654–17 665, 2022.
[10] R. O. Ogundokun, S. Misra, M. Douglas, R. Damaševičius, and R. Maskeliūnas, "Medical internet-of-things based breast cancer diagnosis using hyperparameter-optimized neural networks," *Future Internet*, vol. 14, no. 5, p. 153, 2022.
[11] M. Ahmad, S. F. Qadri, S. Qadri, I. A. Saeed, S. S. Zareen, Z. Iqbal, A. Alabrah, H. M. Alaghbari, and S. M. Mizanur Rahman, "A lightweight convolutional neural network model for liver segmentation in medical diagnosis," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, p. 7954333, 2022.
[12] H. Aljuaid, N. Alturki, N. Alsubaie, L. Cavallaro, and A. Liotta, "Computer-aided diagnosis for breast cancer classification using deep neural networks and transfer learning," *Computer Methods and Programs in Biomedicine*, vol. 223, p. 106951, 2022.
[13] "Hugging Face," https://huggingface.co/models, 2018.
[14] "Model Zoo," https://modelzoo.co/, 2018.
[15] "Github," https://github.com/, 2008.
[16] "AWS Marketplace," https://aws.amazon.com/marketplace, 2012.

[17] Y. Liu, S. Ma, Y. Aafer, W.-C. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning attack on neural networks," in *25th Annual Network And Distributed System Security Symposium (NDSS 2018)*, 2018.

[18] E. Chou, F. Tramèr, G. Pellegrino, and D. Boneh, "Sentinet: Detecting physical attacks against deep learning systems," in *2020 IEEE Security and Privacy Workshops (SPW)*, 2020.

[19] B. Chen, W. Carvalho, N. Baracaldo, H. Ludwig, B. Edwards, T. Lee, I. Molloy, and B. Srivastava, "Detecting backdoor attacks on deep neural networks by activation clustering," in *Workshop on Artificial Intelligence Safety*, 2019.

[20] Tran, Brandon and Li, Jerry and Mdry, Aleksander, "Spectral signatures in backdoor attacks," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018.

[21] Y. Gao, C. Xu, D. Wang, S. Chen, D. C. Ranasinghe, and S. Nepal, "Strip: A defence against trojan attacks on deep neural networks," in *Proceedings of the 35th annual computer security applications conference*, 2019.

[22] D. Tang, X. Wang, H. Tang, and K. Zhang, "Demon in the variant: Statistical analysis of {DNNs} for robust backdoor contamination detection," in *30th USENIX Security Symposium (USENIX Security 21)*, 2021.

[23] W. Ma, D. Wang, R. Sun, M. Xue, S. Wen, and Y. Xiang, "The "Beatrix" Resurrections: Robust Backdoor Detection via Gram Matrices," in *30th Annual Network and Distributed System Security Symposium (NDSS 2023)*, 2023.

[24] X. Mo, Y. Zhang, L. Y. Zhang, W. Luo, N. Sun, S. Hu, S. Gao, and Y. Xiang, "Robust backdoor detection for deep learning via topological evolution dynamics," in *IEEE Symposium on Security and Privacy (SP)*, 2024.

[25] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," in *2019 IEEE Symposium on Security and Privacy (SP)*, 2019.

[26] Y. Liu, W.-C. Lee, G. Tao, S. Ma, Y. Aafer, and X. 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*, 2019.

[27] X. Xu, Q. Wang, H. Li, N. Borisov, C. A. Gunter, and B. Li, "Detecting ai trojans using meta neural analysis," in *IEEE Symposium on Security and Privacy (SP)*, 2021.

[28] C. Fu, X. Zhang, S. Ji, T. Wang, P. Lin, Y. Feng, and J. Yin, "{FreeEagle}: Detecting complex neural trojans in {Data-Free} cases," in *32nd USENIX Security Symposium (USENIX Security 23)*, 2023.

[29] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[30] A. Krizhevsky, "Learning multiple layers of features from tiny images," *Master's thesis, University of Tront*, 2009.

[31] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition," *Neural networks*, vol. 32, pp. 323–332, 2012.

[32] J. Howard and S. Gugger, "Fastai: a layered API for deep learning," *Information*, vol. 11, no. 2, p. 108, 2020.

[33] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[34] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017.

[35] M. Du, N. Liu, Q. Song, and X. Hu, "Towards explanation of dnn-based prediction with guided feature inversion," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018.

[36] Y. Li, T. Zhai, B. Wu, Y. Jiang, Z. Li, and S. Xia, "Rethinking the trigger of backdoor attack," *arXiv preprint arXiv:2004.04692*, 2020.

[37] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, "Badnets: Evaluating backdooring attacks on deep neural networks," *IEEE Access*, vol. 7, pp. 47 230–47 244, 2019.

[38] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," *arXiv preprint arXiv:1712.05526*, 2017.

[39] A. Turner, D. Tsipras, and A. Madry, "Label-consistent backdoor attacks," *arXiv preprint arXiv:1912.02771*, 2019.

[40] H. Zhong, C. Liao, A. C. Squicciarini, S. Zhu, and D. Miller, "Backdoor embedding in convolutional neural network models via invisible perturbation," in *Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy*, 2020.

[41] Y. Ji, Z. Liu, X. Hu, P. Wang, and Y. Zhang, "Programmable neural network trojan for pre-trained feature extractor," *arXiv preprint arXiv:1901.07766*, 2019.

[42] Y. Li, Y. Li, B. Wu, L. Li, R. He, and S. Lyu, "Invisible backdoor attack with sample-specific triggers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

[43] J. Zhang, C. Dongdong, Q. Huang, J. Liao, W. Zhang, H. Feng, G. Hua, and N. Yu, "Poison ink: Robust and invisible backdoor attack," *IEEE Transactions on Image Processing*, vol. 31, pp. 5691–5705, 2022.

[44] T. A. Nguyen and A. Tran, "Input-aware dynamic backdoor attack," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3454–3464, 2020.

[45] W. Yang, L. Li, Z. Zhang, X. Ren, X. Sun, and B. He, "Be careful about poisoned word embeddings: Exploring the vulnerability of the embedding layers in nlp models," in *North American Chapter of the Association for Computational Linguistics*, 2021.

[46] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

[47] E. Wenger, J. Passananti, A. N. Bhagoji, Y. Yao, H. Zheng, and B. 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*, 2021.

[48] S. Li, M. Xue, B. Z. H. Zhao, H. Zhu, and X. Zhang, "Invisible backdoor attacks on deep neural networks via steganography and regularization," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 5, pp. 2088–2105, 2020.

[49] J. Lin, L. Xu, Y. Liu, and X. Zhang, "Composite backdoor attack for deep neural network by mixing existing benign features," in *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, 2020.

[50] Y. Li, T. Zhai, Y. Jiang, Z. Li, and S.-T. Xia, "Backdoor attack in the physical world," *arXiv preprint arXiv:2104.02361*, 2021.

[51] K. Doan, Y. Lao, W. Zhao, and P. Li, "Lira: Learnable, imperceptible and robust backdoor attacks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

[52] W. R. Huang, J. Geiping, L. Fowl, G. Taylor, and T. Goldstein, "Metapoison: Practical general-purpose clean-label data poisoning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 080–12 091, 2020.

[53] S. Zhao, X. Ma, X. Zheng, J. Bailey, J. Chen, and Y.-G. Jiang, "Clean-label backdoor attacks on video recognition models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[54] E. Quiring and K. Rieck, "Backdooring and poisoning neural networks with image-scaling attacks," in *2020 IEEE Security and Privacy Workshops (SPW)*, 2020.

[55] H. Souri, L. Fowl, R. Chellappa, M. Goldblum, and T. Goldstein, "Sleeper agent: Scalable hidden trigger backdoors for neural networks trained from scratch," *Advances in Neural Information Processing Systems*, vol. 35, pp. 19 165–19 178, 2022.

[56] I. Shumailov, Z. Shumaylov, D. Kazhdan, Y. Zhao, N. Papernot, M. A. Erdogdu, and R. J. Anderson, "Manipulating sgd with data ordering attacks," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18 021–18 032, 2021.

[57] Y. Zeng, M. Pan, H. A. Just, L. Lyu, M. Qiu, and R. Jia, "Narcissus: A practical clean-label backdoor attack with limited information," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 2023.

[58] P. Lv, C. Yue, R. Liang, Y. Yang, S. Zhang, H. Ma, and K. Chen, "A data-free backdoor injection approach in neural networks," in *32nd USENIX Security Symposium (USENIX Security 23)*, 2023.

[59] X. Qi, T. Xie, Y. Li, S. Mahloujifar, and P. Mittal, "Revisiting the assumption of latent separability for backdoor defenses," in *International Conference on Learning Representations*, 2023.

[60] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," *arXiv preprint arXiv:1708.06733*, 2017.

[61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.
[62] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015.
[63] Y. Wang, W. Li, E. Sarkar, M. Shafique, M. Maniatakos, and S. E. Jabari, "A subspace projective clustering approach for backdoor attack detection and mitigation in deep neural networks," *IEEE Transactions on Artificial Intelligence*, 2024.
[64] "Acoomans," https://github.com/acoomans/instagram-filters, 2013.
[65] Y. Le and X. Yang, "Tiny imagenet visual recognition challenge," *CS 231N*, vol. 7, no. 7, p. 3, 2015.
[66] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*, 2021.
[67] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A. Y. Ng *et al.*, "Reading digits in natural images with unsupervised feature learning," in *NIPS workshop on deep learning and unsupervised feature learning*, 2011.
[68] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, 2020.
[69] J. Jia, Y. Liu, and N. Z. Gong, "Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning," in *2022 IEEE Symposium on Security and Privacy (SP)*, 2022.
[70] G. Tao, Z. Wang, S. Feng, G. Shen, S. Ma, and X. Zhang, "Distribution preserving backdoor attack in self-supervised learning," in *2024 IEEE Symposium on Security and Privacy (SP)*, 2023.
[71] S. Kolouri, K. Nadjahi, U. Simsekli, R. Badeau, and G. Rohde, "Generalized sliced wasserstein distances," *Advances in neural information processing systems*, vol. 32, 2019.
[72] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.
[73] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.
[74] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.
[75] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.
[76] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
[77] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *North American Chapter of the Association for Computational Linguistics*, 2019.
[78] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
[79] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
[80] L. Wu, Y. Chen, K. Shen, X. Guo, H. Gao, S. Li, J. Pei, B. Long *et al.*, "Graph neural networks for natural language processing: A survey," *Foundations and Trends® in Machine Learning*, vol. 16, no. 2, pp. 119–328, 2023.
[81] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz, and D. Roth, "Recent advances in natural language processing via large pre-trained language models: A survey," *ACM Computing Surveys*, vol. 56, no. 2, pp. 1–40, 2023.
[82] L. Liu, G. Feng, D. Beautemps, and X.-P. Zhang, "A novel resynchronization procedure for hand-lips fusion applied to continuous french cued speech recognition," in *2019 27th European Signal Processing Conference (EUSIPCO)*, 2019.
[83] Liu, Li and Feng, Gang and Beautemps, Denis and Zhang, Xiao-Ping, "Re-synchronization using the hand preceding model for multi-modal fusion in automatic continuous cued speech recognition," *IEEE Transactions on Multimedia*, vol. 23, pp. 292–305, 2020.
[84] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification." in *Interspeech*, 2017.
[85] Y. Tang, G. Ding, J. Huang, X. He, and B. Zhou, "Deep speaker embedding learning with multi-level pooling for text-independent speaker verification," in *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2019.
[86] P. S. Nidadavolu, S. Kataria, J. Villalba, P. Garcia-Perera, and N. Dehak, "Unsupervised feature enhancement for speaker verification," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

APPENDIX

A. F1 Scores of Backdoored Model Detection Methods

F1 score is the harmonic mean of precision and recall, and a statistical indicator used to measure the accuracy of classification models. Due to the page limit, we show the F1 scores of our experiments in this section. Table XVI-XIX respectively exhibit F1 scores of detection methods against source-agnostic attacks, source-specific attacks, sample-specific attacks and clean-label attacks. It should be noted that due to the zero precision and recall in some experiments, some F1 scores cannot be calculated. In these experiments, BARBIE has extremely superior performance with nearly 100.00% F1 scores.

Table XVI
F1 SCORES OF DETECTION METHODS AGAINST SOURCE-AGNOSTIC ATTACKS.

Method	Dataset	SPC	Beatrix_L	Beatrix_H	NC	ABS	STRIP	MNTD	FreeEagle	BARBIE
Patch	MNIST	6.68%	-	7.02%	31.98%	31.82%	98.91%	62.14%	97.96%	**99.20%**
	CIFAR10	7.55%	**99.27%**	-	14.41%	97.46%	96.76%	40.85%	97.90%	98.74%
	ImageNette	4.61%	-	15.53%	16.75%	96.93%	26.25%	69.61%	91.08%	**98.77%**
	GTSRB	15.56%	-	17.98%	-	98.09%	97.54%	72.78%	96.40%	**99.90%**
Blending	MNIST	8.49%	-	6.08%	68.59%	37.75%	0.97%	49.55%	97.14%	**99.20%**
	CIFAR10	9.05%	94.20%	0.43%	28.76%	96.73%	94.80%	43.25%	82.92%	**97.52%**
	ImageNette	2.90%	-	6.77%	46.81%	95.84%	17.01%	75.20%	85.30%	**95.49%**
	GTSRB	9.63%	-	8.51%	6.72%	95.16%	97.52%	72.15%	94.74%	**98.72%**
Filter	MNIST	5.83%	-	5.98%	92.41%	7.60%	95.66%	57.43%	85.31%	**98.19%**
	CIFAR10	3.89%	91.27%	13.40%	25.88%	95.76%	34.90%	51.56%	88.92%	**97.11%**
	ImageNette	6.35%	-	-	-	91.08%	-	74.71%	87.57%	**94.29%**
	GTSRB	7.79%	-	31.10%	3.60%	76.43%	95.58%	81.40%	97.91%	**99.90%**
Composite	MNIST	10.64%	-	4.39%	9.33%	68.30%	-	20.51%	95.58%	**99.20%**
	CIFAR10	13.50%	97.62%	0.29%	19.98%	94.42%	0.27%	48.27%	77.77%	**98.74%**
	ImageNette	16.46%	-	-	-	91.08%	-	68.97%	91.24%	**98.77%**
	GTSRB	12.97%	-	33.00%	31.59%	97.20%	-	85.78%	99.43%	**99.90%**
Adaptive-Patch	MNIST	11.95%	-	11.27%	87.49%	96.10%	3.23%	77.44%	89.38%	**99.20%**
	CIFAR10	21.38%	98.07%	-	18.56%	96.23%	96.40%	46.03%	72.10%	**98.74%**
	ImageNette	9.03%	-	-	39.62%	97.86%	0.0%	63.88%	72.48%	**98.57%**
	GTSRB	3.11%	-	3.64%	39.18%	95.48%	1.91%	64.76%	98.53%	**99.90%**
Adaptive-Blend	MNIST	23.85%	-	7.59%	41.26%	83.83%	5.95%	71.27%	37.06%	**99.20%**
	CIFAR10	25.75%	97.86%	-	26.03%	91.53%	3.46%	33.58%	53.15%	**98.74%**
	ImageNette	0.74%	-	9.24%	18.01%	50.72%	0.0%	59.65%	78.71%	**98.77%**
	GTSRB	2.18%	-	13.97%	43.74%	94.94%	0.0%	54.01%	95.32%	**99.90%**

Table XVII
F1 SCORES OF DETECTION METHODS AGAINST SOURCE-SPECIFIC ATTACKS.

Method	Dataset	SPC	Beatrix_L	Beatrix_H	NC	ABS	STRIP	MNTD	FreeEagle	BARBIE
Patch	MNIST	3.50%	-	9.72%	19.27%	11.53%	38.49%	58.22%	78.39%	**96.55%**
	CIFAR10	3.37%	5.74%	0.80%	27.34%	14.90%	10.63%	46.34%	75.59%	**95.03%**
	ImageNette	-	-	1.30%	-	9.60%	10.51%	87.29%	82.92%	**98.29%**
	GTSRB	8.67%	-	5.48%	-	75.97%	2.57%	68.52%	81.42%	**99.90%**
Blending	MNIST	8.02%	-	20.10%	31.60%	8.87%	24.22%	63.30%	84.38%	**97.32%**
	CIFAR10	8.97%	36.80%	6.32%	17.56%	11.15%	6.20%	47.82%	80.40%	**90.02%**
	ImageNette	3.30%	-	0.15%	16.16%	-	5.49%	81.82%	81.97%	**95.16%**
	GTSRB	11.65%	-	44.93%	3.12%	64.14%	-	78.75%	80.34%	**99.90%**
Filter	MNIST	8.07%	-	9.75%	20.05%	1.26%	20.94%	48.40%	82.13%	**97.56%**
	CIFAR10	7.88%	9.85%	10.18%	1.98%	22.92%	-	46.42%	82.79%	**95.29%**
	ImageNette	-	-	10.15%	13.76%	1.94%	6.82%	83.33%	83.05%	**90.40%**
	GTSRB	7.01%	-	42.78%	1.00%	73.64%	-	84.14%	80.64%	**99.90%**
Composite	MNIST	11.35%	-	10.97%	37.64%	59.15%	53.86%	46.91%	21.52%	**99.20%**
	CIFAR10	24.90%	**99.01%**	17.08%	23.30%	51.71%	33.84%	63.33%	62.79%	98.74%
	ImageNette	8.26%	-	0.29%	-	77.33%	-	64.61%	83.43%	**98.77%**
	GTSRB	5.12%	-	8.96%	41.09%	54.66%	0.40%	95.00%	92.24%	**99.90%**

Table XVIII
F1 SCORES OF DETECTION METHODS AGAINST SAMPLE-SPECIFIC ATTACKS.

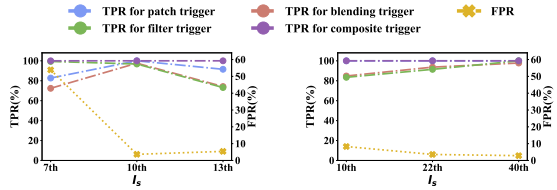
Method	Type	Dataset	SPC	Beatrix_L	Beatrix_H	NC	ABS	STRIP	MNTD	FreeEagle	BARBIE
Input-Aware	All-to-One	MNIST	8.49%	-	13.59%	41.64%	63.97%	26.00%	39.12%	28.72%	**99.15%**
		CIFAR10	35.50%	98.18%	-	-	11.83%	17.20%	53.10%	46.85%	**98.74%**
		ImageNette	3.23%	-	7.63%	-	78.23%	-	61.81%	70.12%	**98.77%**
		GTSRB	1.76%	-	19.15%	-	96.80%	-	53.68%	99.22%	**99.90%**
	All-to-All	MNIST	22.67%	-	13.26%	14.81%	41.29%	3.73%	65.63%	48.65%	**99.20%**
		CIFAR10	23.23%	98.28%	-	18.08%	7.75%	3.75%	24.20%	91.05%	**98.74%**
		ImageNette	11.83%	-	2.73%	-	13.58%	-	67.95%	50.76%	**98.77%**
		GTSRB	53.41%	-	5.66%	26.42%	97.20%	-	75.94%	91.70%	**99.90%**

Table XIX
F1 SCORES OF DETECTION METHODS AGAINST CLEAN-LABEL ATTACKS.

Method	Dataset	SPC	Beatrix_L	Beatrix_H	NC	ABS	STRIP	MNTD	FreeEagle	BARBIE
NARCISSUS	MNIST	27.58%	-	7.42%	10.98%	8.76%	47.18%	53.89%	68.61%	**99.20%**
	CIFAR10	36.70%	**97.52%**	-	10.54%	96.58%	94.07%	22.01%	96.48%	97.12%
	ImageNette	-	-	11.71%	10.37%	88.23%	-	52.99%	90.54%	**98.77%**
	GTSRB	3.38%	-	42.05%	88.78%	98.30%	-	67.20%	98.99%	**99.90%**
Data-free Backdoor	MNIST	7.57%	-	20.03%	38.38%	51.85%	**99.50%**	40.00%	97.59%	99.20%
	CIFAR10	14.49%	26.61%	98.32%	96.95%	97.99%	-	44.28%	98.44%	**98.74%**
	ImageNette	6.16%	-	8.43%	45.01%	16.33%	95.59%	57.62%	83.42%	**98.77%**
	GTSRB	-	-	5.90%	-	1.02%	-	85.86%	97.88%	**99.90%**

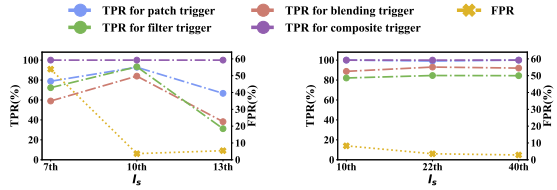
B. Hyperparameter Experiment

We show the hyperparameter experiment results in this part for page limit, as shown in Figure 6-20.



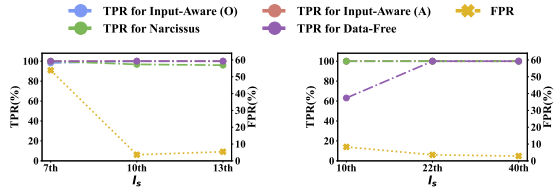
(a) CIFAR10 (b) ImageNette

Figure 6. Impact of l_s against source-agnostic attacks on CIFAR10 and ImageNette.



(a) CIFAR10 (b) ImageNette

Figure 7. Impact of l_s against source-specific attacks on CIFAR10 and ImageNette.

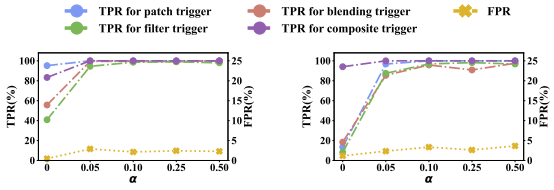


(a) CIFAR10 (b) ImageNette

Figure 8. Impact of l_s against sample-specific and clean-label attacks on CIFAR10 and ImageNette. The all-to-one attack and all-to-all attack are denoted as O and A respectively.

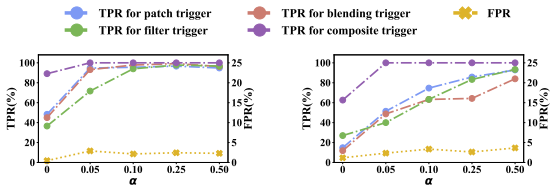
C. Performance with a Poisoned Model Zoo

We display the detection performances of ABS, STRIP and FreeEagle in Table XX-XXIII, when the benign model zoo is contaminated by 5% or 10% backdoored models.



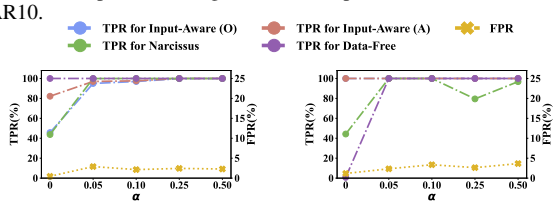
(a) MNIST (b) CIFAR10

Figure 9. Impact of α against source-agnostic attacks on MNIST and CIFAR10.



(a) MNIST (b) CIFAR10

Figure 10. Impact of α against source-specific attacks on MNIST and CIFAR10.



(a) MNIST (b) CIFAR10

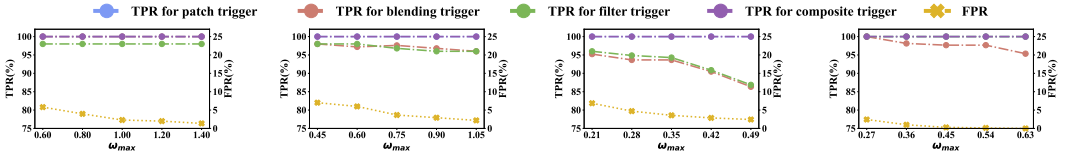
Figure 11. Impact of α against sample-specific and clean-label attacks on MNIST and CIFAR10. The all-to-one attack and all-to-all attack are denoted as O and A respectively.

Table XX
DETECTION PERFORMANCE OF ABS, STRIP AND FREEEAGLE WITH A POISONED MODEL ZOO AGAINST SOURCE-AGNOSTIC ATTACKS.

Poison rate	Method	Dataset	ABS		STRIP		FreeEagle	
			TPR	FPR	TPR	FPR	TPR	FPR
5%	Patch	MNIST	14.40%	3.79%	96.84%	2.13%	98.17%	4.18%
		CIFAR10	98.52%	1.85%	96.96%	3.76%	99.79%	3.53%
		ImageNette	20.00%	0.00%	23.66%	3.08%	48.75%	12.11%
		GTSRB	99.40%	5.07%	96.83%	1.53%	99.40%	7.67%
	Blending	MNIST	27.75%	2.46%	0.48%	4.43%	95.80%	5.80%
		CIFAR10	100.00%	4.71%	98.73%	2.39%	39.01%	8.19%
		ImageNette	19.55%	0.00%	6.71%	5.06%	79.35%	12.71%
		GTSRB	99.26%	3.20%	99.55%	3.15%	96.16%	10.92%
	Filter	MNIST	1.94%	4.19%	93.78%	0.22%	53.37%	11.48%
		CIFAR10	99.92%	3.46%	22.23%	4.96%	59.02%	9.40%
		ImageNette	7.31%	0.00%	1.88%	3.62%	89.57%	10.68%
		GTSRB	62.77%	4.03%	99.48%	6.62%	98.87%	4.64%
	Composite	MNIST	38.13%	4.85%	0.00%	8.01%	96.31%	5.25%
		CIFAR10	93.29%	3.00%	0.94%	6.53%	68.21%	10.45%
		ImageNette	24.16%	0.00%	0.00%	3.50%	88.48%	9.11%
		GTSRB	100.00%	6.13%	0.00%	9.41%	97.36%	0.00%
10%	Patch	MNIST	13.68%	3.44%	97.68%	5.81%	95.86%	5.63%
		CIFAR10	96.88%	3.61%	99.93%	5.69%	99.79%	4.14%
		ImageNette	27.71%	0.08%	18.84%	2.01%	42.67%	8.81%
		GTSRB	98.21%	9.99%	99.17%	8.63%	99.24%	4.14%
	Blending	MNIST	23.27%	2.78%	1.21%	6.65%	97.90%	6.03%
		CIFAR10	99.78%	4.37%	97.95%	3.98%	46.69%	13.01%
		ImageNette	67.63%	0.00%	8.60%	4.59%	78.08%	9.78%
		GTSRB	99.19%	5.67%	98.96%	2.59%	95.50%	4.97%
	Filter	MNIST	2.99%	4.46%	96.08%	2.94%	50.08%	10.68%
		CIFAR10	100.00%	4.34%	20.80%	6.13%	67.28%	12.58%
		ImageNette	52.40%	0.00%	0.00%	1.72%	85.67%	13.32%
		GTSRB	68.29%	3.99%	99.31%	9.37%	98.34%	3.68%
	Composite	MNIST	31.40%	4.93%	0.00%	4.85%	96.45%	6.26%
		CIFAR10	86.59%	1.61%	0.51%	5.06%	72.33%	9.55%
		ImageNette	15.93%	0.00%	0.00%	3.92%	92.19%	8.63%
		GTSRB	99.93%	4.12%	0.00%	12.18%	99.30%	7.31%

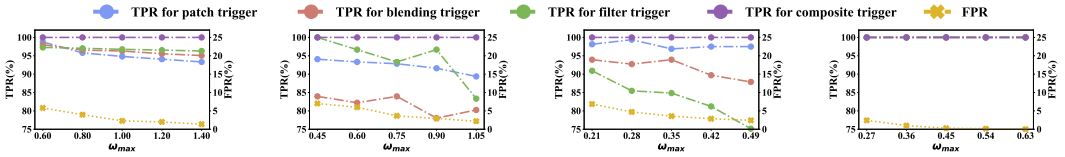
D. Performance with Substitute Benign Models

We display the detection performances of ABS, STRIP and FreeEagle in Table XXIV-XXVII, when the defender attempts to detect backdoored models with substitute benign models, which perform similar tasks and have the same model structure as the benign models.



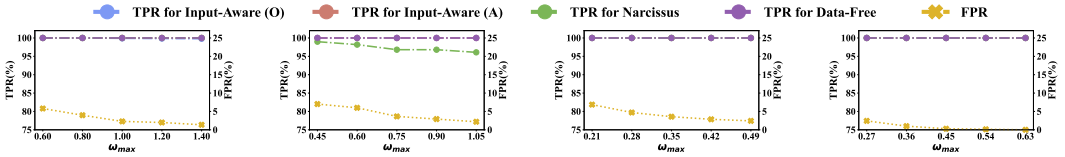
(a) MNIST (b) CIFAR10 (c) ImageNette (d) GTSRB

Figure 12. Impact of ω_{max} against source-agnostic attacks.



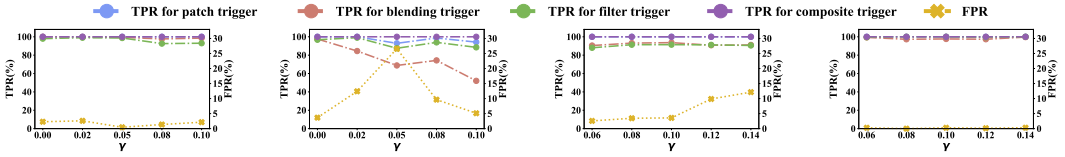
(a) MNIST (b) CIFAR10 (c) ImageNette (d) GTSRB

Figure 13. Impact of ω_{max} against source-specific attacks.



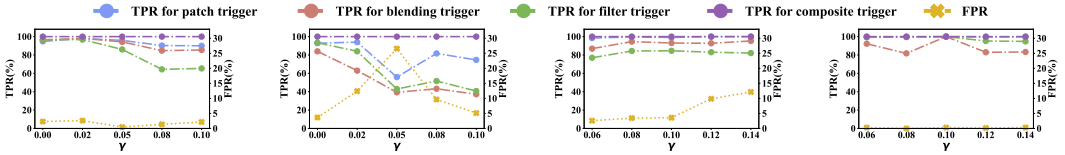
(a) MNIST (b) CIFAR10 (c) ImageNette (d) GTSRB

Figure 14. Impact of ω_{max} against sample-specific and clean-label attacks. The all-to-one attack and all-to-all attack are denoted as O and A respectively.



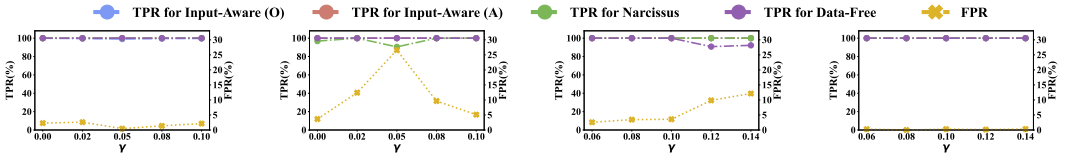
(a) MNIST (b) CIFAR10 (c) ImageNette (d) GTSRB

Figure 15. Impact of γ against source-agnostic attacks.



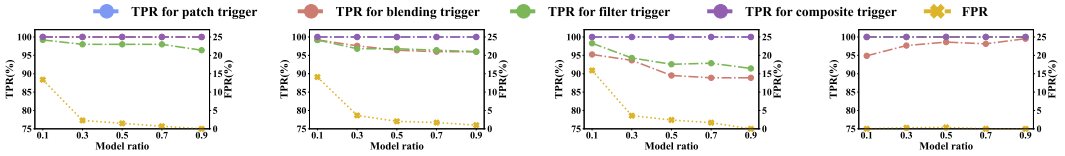
(a) MNIST (b) CIFAR10 (c) ImageNette (d) GTSRB

Figure 16. Impact of γ against source-specific attacks.



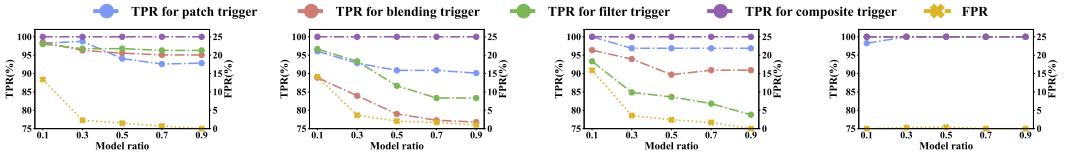
(a) MNIST (b) CIFAR10 (c) ImageNette (d) GTSRB

Figure 17. Impact of γ against sample-specific and clean-label attacks. The all-to-one attack and all-to-all attack are denoted as O and A respectively.



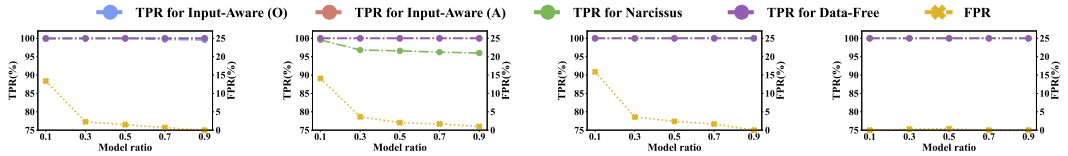
(a) MNIST (b) CIFAR10 (c) ImageNette (d) GTSRB

Figure 18. Impact of accessible model ratio against source-agnostic attacks.



(a) MNIST (b) CIFAR10 (c) ImageNette (d) GTSRB

Figure 19. Impact of accessible model ratio against source-specific attacks.



(a) MNIST (b) CIFAR10 (c) ImageNette (d) GTSRB

Figure 20. Impact of accessible model ratio against sample-specific and clean-label attacks. The all-to-one attack and all-to-all attack are denoted as O and A respectively.

Table XXI
DETECTION PERFORMANCE OF ABS, STRIP AND FREEEAGLE WITH A POISONED MODEL ZOO AGAINST SOURCE-SPECIFIC ATTACKS.

Poison rate	Method	Dataset	ABS		STRIP		FreeEagle	
			TPR	FPR	TPR	FPR	TPR	FPR
5%	Patch	MNIST	7.20%	3.05%	27.70%	10.02%	73.02%	10.62%
		CIFAR10	0.85%	3.68%	5.75%	8.03%	61.00%	11.74%
		ImageNette	1.72%	0.00%	13.52%	4.80%	99.11%	8.69%
		GTSRB	78.48%	6.79%	10.23%	10.90%	69.67%	7.20%
	Blending	MNIST	0.83%	4.11%	12.65%	8.02%	80.95%	10.44%
		CIFAR10	1.20%	3.12%	2.58%	9.30%	46.37%	10.35%
		ImageNette	0.00%	0.00%	2.39%	4.65%	90.99%	15.59%
		GTSRB	39.55%	5.36%	0.00%	4.16%	74.56%	14.87%
	Filter	MNIST	0.35%	4.36%	18.83%	10.22%	44.26%	10.12%
		CIFAR10	0.00%	1.59%	2.23%	4.54%	33.21%	9.50%
		ImageNette	0.00%	0.00%	3.07%	3.08%	75.15%	9.20%
		GTSRB	65.83%	9.67%	0.00%	11.99%	72.55%	14.30%
	Composite	MNIST	48.39%	4.35%	48.07%	7.80%	32.89%	10.67%
		CIFAR10	24.47%	2.61%	20.94%	6.89%	48.51%	8.80%
		ImageNette	5.00%	0.00%	0.00%	0.88%	85.65%	9.93%
		GTSRB	36.76%	4.89%	1.42%	10.63%	94.31%	9.47%
10%	Patch	MNIST	9.79%	3.68%	21.76%	8.36%	73.23%	8.92%
		CIFAR10	1.13%	3.06%	7.35%	7.86%	67.89%	10.03%
		ImageNette	0.00%	0.00%	7.34%	2.73%	96.51%	8.00%
		GTSRB	64.36%	9.57%	3.04%	4.43%	73.07%	4.61%
	Blending	MNIST	2.89%	4.04%	16.10%	13.69%	77.84%	10.34%
		CIFAR10	2.58%	3.11%	1.89%	6.04%	36.99%	8.11%
		ImageNette	0.00%	0.00%	0.53%	1.24%	88.54%	8.88%
		GTSRB	48.37%	10.17%	0.00%	8.13%	76.25%	9.38%
	Filter	MNIST	0.99%	5.51%	16.88%	9.45%	45.39%	13.60%
		CIFAR10	0.00%	6.19%	0.00%	2.51%	38.02%	5.26%
		ImageNette	0.00%	0.00%	1.48%	3.83%	78.89%	11.01%
		GTSRB	50.90%	6.13%	0.00%	3.90%	72.54%	6.12%
	Composite	MNIST	40.94%	2.44%	52.34%	8.45%	36.95%	9.95%
		CIFAR10	20.33%	0.67%	18.64%	8.66%	51.10%	11.19%
		ImageNette	40.65%	0.00%	0.00%	2.32%	79.72%	7.91%
		GTSRB	36.05%	5.39%	1.35%	9.23%	94.04%	9.42%

Table XXII
DETECTION PERFORMANCE OF ABS, STRIP AND FREEEAGLE WITH A POISONED MODEL ZOO AGAINST SAMPLE-SPECIFIC ATTACKS.

Poison rate	Method	Type	Dataset	ABS		STRIP		FreeEagle	
				TPR	FPR	TPR	FPR	TPR	FPR
5%	Input-Aware	All-to-One	MNIST	39.84%	2.63%	17.82%	7.83%	30.70%	10.76%
			CIFAR10	4.59%	2.62%	8.17%	7.62%	46.37%	11.46%
			ImageNette	34.40%	0.00%	0.00%	3.11%	66.02%	12.93%
			GTSRB	90.30%	4.92%	0.00%	8.18%	98.25%	0.00%
		All-to-All	MNIST	26.25%	4.19%	1.12%	8.77%	50.49%	14.36%
			CIFAR10	0.00%	3.14%	3.76%	7.55%	95.30%	8.70%
			ImageNette	1.78%	0.00%	0.00%	1.51%	49.42%	11.47%
			GTSRB	99.19%	5.99%	0.00%	5.99%	92.63%	9.88%
10%	Input-Aware	All-to-One	MNIST	49.14%	1.74%	13.84%	5.56%	24.13%	13.55%
			CIFAR10	6.04%	2.48%	14.27%	10.00%	53.39%	10.32%
			ImageNette	36.36%	0.00%	0.00%	2.31%	42.07%	10.33%
			GTSRB	95.05%	4.68%	0.00%	6.57%	95.96%	0.00%
		All-to-All	MNIST	16.88%	2.07%	3.11%	9.10%	42.38%	11.56%
			CIFAR10	0.00%	3.28%	3.67%	9.17%	93.44%	6.88%
			ImageNette	0.76%	0.00%	0.00%	1.82%	5.68%	17.48%
			GTSRB	99.17%	1.57%	0.00%	5.05%	92.12%	6.31%

Table XXIII
DETECTION PERFORMANCE OF ABS, STRIP AND FREEEAGLE WITH A POISONED MODEL ZOO AGAINST CLEAN-LABEL ATTACKS.

Poison rate	Method	Dataset	ABS		STRIP		FreeEagle	
			TPR	FPR	TPR	FPR	TPR	FPR
5%	NARCISSUS	MNIST	7.84%	3.41%	29.00%	10.32%	64.54%	13.34%
		CIFAR10	98.54%	3.03%	92.47%	7.06%	96.63%	1.17%
		ImageNette	0.00%	0.00%	0.00%	1.99%	97.18%	9.42%
		GTSRB	99.70%	3.89%	0.00%	6.54%	98.77%	0.00%
	Data-free Backdoor	MNIST	40.47%	3.94%	98.60%	0.00%	98.36%	0.00%
		CIFAR10	99.74%	3.08%	0.00%	3.25%	97.72%	1.20%
		ImageNette	1.97%	0.00%	97.07%	0.40%	86.85%	8.81%
		GTSRB	1.66%	5.64%	0.00%	7.89%	97.25%	2.57%
10%	NARCISSUS	MNIST	3.82%	2.81%	34.37%	10.79%	50.71%	7.98%
		CIFAR10	99.31%	1.77%	95.19%	2.55%	98.42%	1.45%
		ImageNette	0.00%	0.00%	0.00%	2.23%	97.29%	8.74%
		GTSRB	99.55%	3.20%	0.00%	5.86%	98.17%	0.00%
	Data-free Backdoor	MNIST	28.64%	1.56%	99.25%	2.44%	98.64%	0.00%
		CIFAR10	98.70%	2.46%	0.00%	2.23%	99.12%	7.97%
		ImageNette	0.00%	0.00%	71.56%	0.17%	86.22%	11.40%
		GTSRB	3.57%	9.52%	0.00%	5.23%	98.74%	3.61%

Table XXIV
DETECTION PERFORMANCE OF ABS, STRIP AND FREEEAGLE WITH SUBSTITUTE BENIGN MODELS AGAINST SOURCE-AGNOSTIC ATTACKS.

Method	Targeted	Substitute	ABS		STRIP		FreeEagle	
			TPR	FPR	TPR	FPR	TPR	FPR
Patch	MNIST	FashionMNIST	0.00%	0.00%	98.25%	0.00%	56.75%	95.87%
	MNIST	SVHN	0.00%	0.00%	98.07%	0.00%	89.21%	31.81%
	CIFAR10	FashionMNIST	17.03%	0.00%	98.26%	0.00%	78.54%	27.15%
	ImageNette	STL10	100.00%	0.00%	85.50%	100.00%	75.62%	99.25%
Blending	MNIST	FashionMNIST	0.00%	0.00%	2.32%	92.32%	58.10%	88.48%
	MNIST	SVHN	1.08%	0.00%	0.66%	0.07%	89.03%	30.60%
	CIFAR10	FashionMNIST	98.03%	0.00%	98.23%	0.00%	21.10%	31.68%
	ImageNette	STL10	99.03%	0.60%	64.57%	100.00%	61.20%	100.00%
Filter	MNIST	FashionMNIST	0.00%	0.00%	96.70%	42.93%	74.94%	97.91%
	MNIST	SVHN	0.00%	0.00%	94.05%	0.14%	61.51%	31.40%
	CIFAR10	FashionMNIST	92.85%	0.00%	16.63%	4.41%	63.79%	34.14%
	ImageNette	STL10	100.00%	0.68%	30.79%	100.00%	63.79%	99.55%
Composite	MNIST	FashionMNIST	0.00%	0.00%	0.44%	71.28%	76.14%	93.41%
	MNIST	SVHN	0.00%	0.00%	0.00%	0.00%	95.02%	7.95%
	CIFAR10	FashionMNIST	12.16%	0.00%	0.97%	7.20%	72.74%	31.31%
	ImageNette	STL10	87.09%	0.53%	0.00%	100.00%	95.26%	85.29%

Table XXV
DETECTION PERFORMANCE OF ABS, STRIP AND FREEEAGLE WITH SUBSTITUTE BENIGN MODELS AGAINST SOURCE-SPECIFIC ATTACKS.

Method	Targeted	Substitute	ABS		STRIP		FreeEagle	
			TPR	FPR	TPR	FPR	TPR	FPR
Patch	MNIST	FashionMNIST	0.00%	0.00%	75.51%	97.26%	62.23%	87.54%
	MNIST	SVHN	0.00%	0.00%	7.71%	0.51%	90.61%	28.40%
	CIFAR10	FashionMNIST	0.00%	0.00%	2.52%	5.66%	80.84%	36.17%
	ImageNette	STL10	11.92%	0.52%	84.42%	100.00%	78.89%	99.02%
Blending	MNIST	FashionMNIST	0.00%	0.00%	76.63%	96.87%	57.97%	97.53%
	MNIST	SVHN	0.00%	0.00%	3.25%	0.22%	91.94%	26.20%
	CIFAR10	FashionMNIST	0.00%	0.00%	3.67%	9.79%	25.61%	33.50%
	ImageNette	STL10	0.00%	1.12%	45.36%	98.80%	50.03%	100.00%
Filter	MNIST	FashionMNIST	0.00%	0.00%	73.17%	92.54%	78.49%	98.63%
	MNIST	SVHN	0.00%	0.00%	1.17%	0.44%	65.90%	32.98%
	CIFAR10	FashionMNIST	0.00%	0.00%	0.00%	2.17%	58.62%	31.71%
	ImageNette	STL10	4.21%	0.00%	76.56%	100.00%	52.66%	99.48%
Composite	MNIST	FashionMNIST	0.00%	0.00%	95.75%	94.41%	96.59%	94.70%
	MNIST	SVHN	0.00%	0.00%	3.82%	0.00%	53.98%	29.97%
	CIFAR10	FashionMNIST	0.00%	0.00%	22.89%	7.42%	65.80%	35.15%
	ImageNette	STL10	58.03%	0.90%	12.09%	100.00%	95.83%	99.19%

Table XXVI
DETECTION PERFORMANCE OF ABS, STRIP AND FREEEAGLE WITH SUBSTITUTE BENIGN MODELS AGAINST SAMPLE-SPECIFIC ATTACKS.

Method	Type	Targeted	Substitute	ABS		STRIP		FreeEagle	
				TPR	FPR	TPR	FPR	TPR	FPR
Input-Aware	All-to-One	MNIST	FashionMNIST	0.00%	0.00%	53.55%	96.54%	79.99%	87.12%
		MNIST	SVHN	0.00%	0.00%	0.83%	0.82%	41.44%	31.96%
		CIFAR10	FashionMNIST	0.00%	0.00%	9.01%	7.90%	15.83%	36.51%
		ImageNette	STL10	65.33%	0.00%	51.28%	100.00%	57.92%	99.55%
	All-to-All	MNIST	FashionMNIST	0.00%	0.00%	55.79%	95.51%	84.43%	97.40%
		MNIST	SVHN	0.00%	0.00%	0.73%	0.14%	64.99%	33.76%
		CIFAR10	FashionMNIST	0.00%	0.00%	2.87%	7.68%	5.93%	42.61%
		ImageNette	STL10	24.74%	0.00%	69.14%	100.00%	84.59%	100.00%

Table XXVII
DETECTION PERFORMANCE OF ABS, STRIP AND FREEEAGLE WITH SUBSTITUTE BENIGN MODELS AGAINST CLEAN-LABEL ATTACKS.

Method	Targeted	Substitute	ABS		STRIP		FreeEagle	
			TPR	FPR	TPR	FPR	TPR	FPR
NARCISSUS	MNIST	FashionMNIST	0.00%	0.00%	83.70%	95.95%	71.13%	87.76%
	MNIST	SVHN	0.00%	0.00%	3.53%	0.29%	71.98%	31.53%
	CIFAR10	FashionMNIST	61.66%	0.00%	92.31%	5.42%	97.66%	1.45%
	ImageNette	STL10	100.00%	100.00%	51.03%	100.00%	53.92%	99.39%
Data-free Backdoor	MNIST	FashionMNIST	0.00%	0.00%	94.84%	0.00%	97.26%	0.00%
	MNIST	SVHN	0.00%	0.00%	97.68%	0.00%	98.35%	0.00%
	CIFAR10	FashionMNIST	0.28%	0.00%	0.00%	1.87%	98.62%	0.55%
	ImageNette	STL10	23.87%	0.53%	98.05%	0.29%	97.30%	85.56%