Isolate Trigger: Detecting and Eradicating Evade-Adaptive Backdoors

Chengrui Sun[*][†], Hua Zhang[*][†], Haoran Gao[‡], Zian Tian [*][†], Jianjin Zhao [§], qi Li [§],
Hongliang Zhu [§], Zongliang Shen [*][†], Shang Wang [¶] and Anmin Fu [‖]
[*]the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications
[†]National Engineering Research Center of Disaster Backup and Recovery, Beijing University of Posts and Telecommunications
[‡]China Mobile Research Institute
[§]Beijing University of Posts and Telecommunications
[¶]University of Technology Sydney
[‖]Nanjing University of Science and Technology

***Abstract*—All current detection of backdoor attacks on deep learning models fall under the category of a non essential features(NEF), which focus on fighting against simple and efficient vertical class backdoor – trigger is small, few and not overlapping with the source. Evade-adaptive backdoor (EAB) attacks have evaded NEF detection and improved training efficiency.**

We introduces a precise, efficient and universal detection and defense framework coined as Isolate Trigger (IsTr). IsTr aims to find the hidden trigger by breaking the barrier of the source features. Therefore, it investigates the essence of backdoor triggering, and uses Steps and Differential-Middle-Slice as components to update past theories of distance and gradient. IsTr also plays a positive role in the model, whether the backdoor exists. For example, accurately find and repair the wrong identification caused by deliberate or unintentional training in automatic driving. Extensive experiments on robustness scross various tasks, including MNIST, facial recognition, and traffic sign recognition, confirm the high efficiency, generality and precision of the IsTr. We rigorously evaluated the effectiveness of the IsTr against a series of six EAB attacks, including Badnets, Sin-Wave, Multi-trigger, SSBAs, CASSOCK, HCB. None of these countermeasures evade, even when attacks are combined and the trigger and source overlap.

I. Introduction

The deep learning field has been applied in various industrial and living environments, becoming integral to people's daily lives [1]–[3]. However, both maliciously implanted backdoors [4] and natural backdoors [5] have triggered serious potential safety hazards in deep learning models within safety-critical domains including autonomous vehicles and facial recognition [6]–[8]. Simultaneously, the increasing prevalence of evasive backdoor attacks has rendered defensive measures imperative [4], [9]–[14]. Numerous defense schemes have demonstrated notable success in detecting and remediating backdoor attacks [15]. Among these, backdoor reverse [12] has garnered significant attention due to its near-lossless repair efficacy and broad applicability across scenarios.

However, evasion-adaptive backdoor(EAB) attacks are capable of circumventing non-essential feature(NEF) defenses because NEF defenses rely on non-essential trigger features and focus on source features. These backdoor attacks operate to force misclassifications by embedding triggers in training samples [16], thus defenses necessitate precise trigger reverse engineering to achieve detection and subsequent impact mitigation. Historically, to accommodate requirements such as training efficiency, attackers deliberately designed triggers to be compact, single-instance, and non-overlapping with source features—the intrinsic feature representations of samples (e.g., facial organs/hair in face recognition tasks) [4]. Conventional defenses detect backdoors by relying on these trigger features [12]. However, emerging attacks [13], [14] reveal that such characteristics constitute non-essential features of triggers. It indicates they are not fundamental requirements for successful backdoor embedding. Consequently, these novel EAB attacks effortlessly nullify NEF defenses by modifying non-essential feature of triggers. On the other hand, we observe that legacy NEF defenses exhibit excessive focus on source features instead of trigger when safeguarding complex tasks—particularly those involving large-scale models and high-resolution images like facial recognition systems [17]. This misalignment severely impedes model diagnostics and remediation. In light of this, we pose a pivotal research question: does there exist an efficient defense mechanism that fundamentally targets the intrinsic nature of backdoor trigger?

We propose the Isolate Trigger (IsTr) framework for backdoor detection and remediation, comprising three sequential components: Steps, Differential-Middle-Slice (DMS), and Unlearning, as shown in Figure 1. Like Figure 1 left, our analysis reveals that recent EAB attacks increasingly amplify the influence of source features , deliberately blurring the distinction between source features and trigger signatures [13], [14]. This deliberate obfuscation directly explains the failure of prior defense schemes, which exhibit an excessive preoccupation with source features at the expense of trigger detection. Hence, effective defense fundamentally requires trigger isolation – the core principle underpinning our IsTr framework.

To make the IsTr framework efficient, we propose a bidirectional reverse defense framework employing back gradient updates and forward validation, designated as Steps. When the backdoor functionality forces misclassification of class b

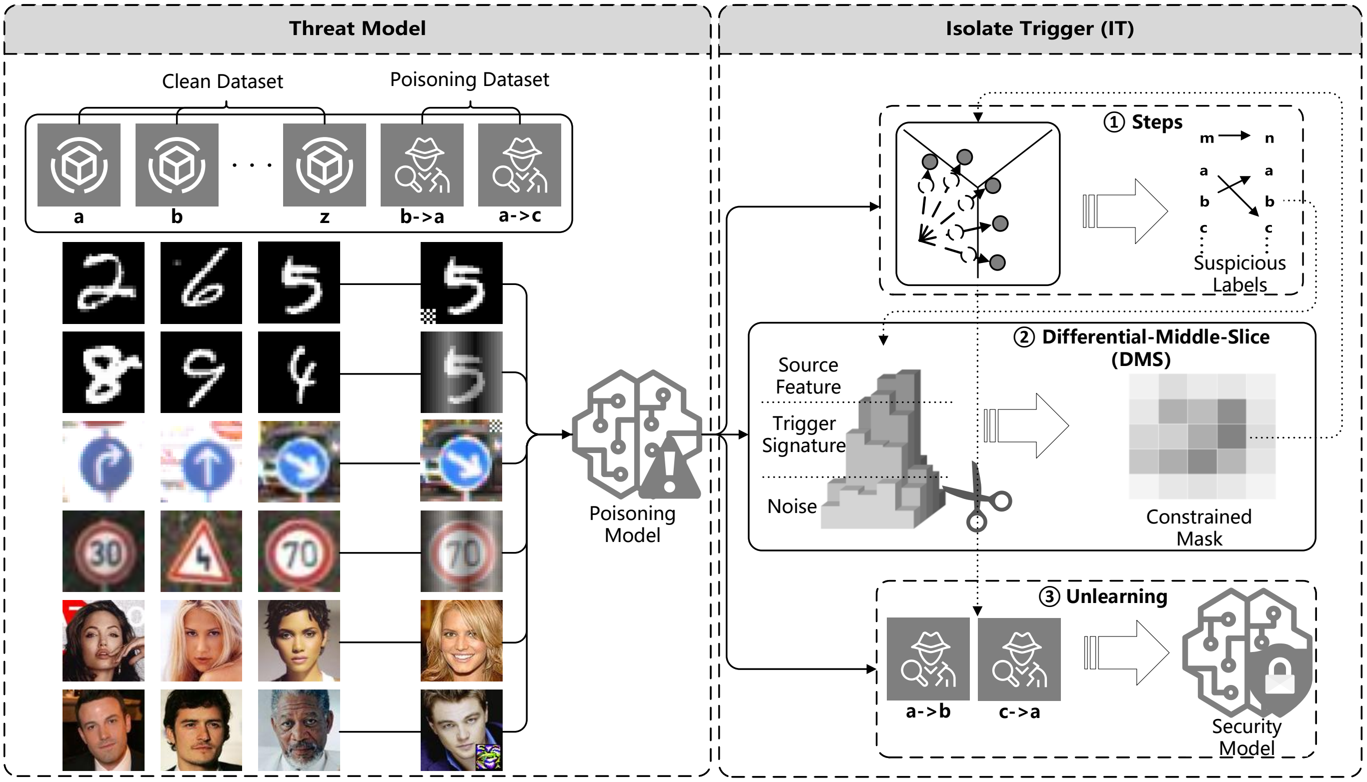


Fig. 1. Overview of IsTr. (Left) Threat model includs EAB attacks. (Right) IsTr framework. The framework initiates with suspicious label screening through Steps. For each identified label group, IsTr employs DMS to derive feature extraction priorities using differential input statistics and generates constrained masks through slice processing. These masks guide precise trigger inversion in Steps. IsTr rehabilitates the poisoning model through Unlearning with label-flipped data.

samples as class a via trigger activation (as depicted in the Poisoning Dataset of Figure 1), Steps reverse reconstruction of trigger signatures rather than source features of class a. This critical observation motivates the core innovation: Instead of pursuing the 'class b to class a' transformation path, Steps can implement source-agnostic trigger synthesis by diverging from class b characteristics. This method forces misclassification to class a without interference from class a source features, thereby achieving untainted trigger generation through unconstrained label mutation.

Furthermore, Steps confers generality upon IsTr. IsTr remains agnostic to model architectures and input specifications. It maintains robust trigger isolation capabilities even for large-scale, complex datasets and models scenarios that pose significant challenges for most NEF defenses. By eliminating the need for target selection, IsTr overcomes the critical limitation of traditional reverse engineering methods which require exhaustive class traversal and incur computational costs scaling linearly with class cardinality. This enables effective deployment in large-class scenarios. IsTr effectively detects EAB attacks (including hybrid variants) even when they successfully subvert NEF defenses. IsTr can remediate natural backdoors [5] in models even when no deliberate backdoor implantation has occurred.

To improve the precision of IsTr framework reverse triggers, including visual similarity and functional integrity with the original trigger, we divide the model's feature extraction into two components: response priority and processing priority, which influence reverse engineering and backdoor activation respectively, as will be discussed in Section II.B. Our intuition is that the response priority for source features versus triggers exhibits positive correlation with the training epochs required to achieve convergence of normal training versus backdoor attack training. This correlation intensifies particularly when triggers overlap with source features. Backdoor attacks are typically executed through simple fine-tuning after extended pretraining and normal fine-tuning. Therefore, backdoor attack training requires far fewer epochs than normal training [4], [18]. Consequently, source features attain higher response priority than triggers, while triggers maintain higher priority than meaningless noise. To achieve trigger isolation under these conditions, we implement DMS to support Steps.

DMS-Steps enables IsTr to eliminate interference from source features, reconstructing inverted trigger with high resemblance to the original trigger. This precision manifests through spatial and structural alignment (matching the position and pattern of original triggers) and functional equivalence (inducing identical model misclassification). Historically, inversion precision received less emphasis than detection accuracy and repair effectiveness in experimental evaluations. However, we empirically demonstrate its positive correlation with both detection accuracy and repair performance. This result affirms the critical importance of high-fidelity trigger reconstruction.

Collectively, by employing a bidirectional reverse defense framework, IsTr achieves detection with significantly fewer iterations. With its concentrated focus on trigger signatures, IsTr attains higher inversion fidelity, thereby enhancing detection accuracy and remediation efficacy to EAB attacks and attacks defensible by NEF. Crucially, as previously established, IsTr's elimination of class traversal accelerates defense execution by orders of magnitude.

Our contributions are summarized as follows:

- **Novel Backdoor Isolation Defense Paradigm.** We introduce IsTr defense—a new, generalized framework that supersedes prior distance-based and gradient-based theories through innovations in Steps and DMS. By fundamentally targeting the intrinsic properties of backdoors, IsTr establishes a distinct theoretical foundation that diverges fundamentally from NEF defenses. We formally define both NEF and IsTr to crystallize their differences.
- **Revealing Essential Characteristics of Backdoor Concealment.** Through a refined stratification of feature extraction, we establish the critical distinction between response priority and processing priority, thereby identifying essential characteristics required for backdoor persistence and evasion. We validate this theoretical insight via differential statistical analysis and achieve precise separation of trigger signatures from source features.
- **Comprehensive Attack Assessment.** We define and analyze EAB attacks—a novel attack specifically engineered to circumvent defenses. Through comprehensive evaluation across diverse datasets (MNIST, GTSRB, PubFig) against state-of-the-art EAB variants (BadNets, Sinwave, Multi-trigger, SSBAs, CASSOCK, HCB). results demonstrate IsTr's consistent robustness. This stems from IsTr's ability to orthogonally decouple trigger signatures from source features. Additionally, IsTr demonstrates robust mitigation efficacy against inherent natural backdoor vulnerabilities in models.

II. Background and Related Work

A. Backdoor Defense and Evasive Attacks

Data-based Defenses (DBD) [15] operates under the "data outsourcing" [19] paradigm, granting access to poisoned datasets for analysis and retraining [20]–[22]. In contrast, Model-based Defenses (MBD) [15] functions within "model outsourcing" [23] constraints, where only compromised models are available for defense. This distinction is critical due to:

- The prevalence of distributed training and privacy preservation, increasingly restricting defender access to training samples while facilitating model access [12], [24], [25].
- The technical asymmetry whereby poisoned samples (generated before backdoor training) enable efficient reconstruction of poisoned models, whereas model-to-sample inversion remains computationally non-trivial [4], [12].

Consequently, MBD can subsume DBD's defense assumptions, but the converse does not hold. Despite limited research attention, MBD increasingly replaces DBD in deployments due to its broader applicability.

Definition 1 Non-Essential Feature(NEF). DBD require fewer defensive assumptions when granted dataset access, whereas MBD offen demand stricter operational constraints (e.g., small and single triggers) to function effectively. This limitation renders past MBD vulnerable to multi-trigger and large-scale backdoor attacks [9]–[11]. Although DBD was historically perceived as unbreakable, it remains fundamentally constrained by excessive focus on source features. This constraint is exploited by attacks targeting its implicit reliance on distant source-trigger separation [14]. These phenomena confirm that conventional defenses fail to address essential backdoor properties, thus categorized as non-essential features defenses.

Definition 2 Evade-Adaptive Backdoor(EAB). Attackers have exploited these defensive flaws, spurring the development of backdoor attacks specifically engineered to bypass detection—collectively termed evade-adaptive backdoors attacks. Conventional EAB variants circumvented singular defensive assumptions (e.g., trigger size constraints), enabling mitigation through defense stacking. This paradigm was shattered by CASSOCK [13] and HCB [14], which pioneer overlapping trigger-source patterns or leverage naturally occurring source-trigger relationships to comprehensively compromise all existing defenses.

B. Analysis of Backdoor Activation Mechanisms

Collectively, NEF defenses lack granular analysis of activation mechanisms—conflating genuine trigger signatures with source features under the broad category of misclassification-causing triggers. EAB attacks exploit this flaw by inducing source-trigger obfuscation, which forces excessive focus on source features while neglecting trigger signatures. To enable finer-grained activation analysis, we decompose feature extraction into response priority (identifying spatially susceptible regions) and processing priority (governing recognition outcomes), formalized in Figure 2. This Figure explains triggered misclassifications from Leonardo DiCaprio to Hugh Jackman.

Response priority(RP) governs feature capture order, exhibiting positive correlation with training epochs in Figure 2 left. Typically, source features require substantially more epochs than trigger signatures to achieve convergence, granting them higher RP and thus greater spatial influence. Trigger signatures, trained with fewer epochs, exert weaker influence. Additionally, although spatial separation between source features and triggers allowed independent activation [4] before CASSOCK [13], CASSOCK and HCB [14] induce spatial overlap, causing source features to mask trigger signatures. These factors explain Strip's [20] failure due to secondary

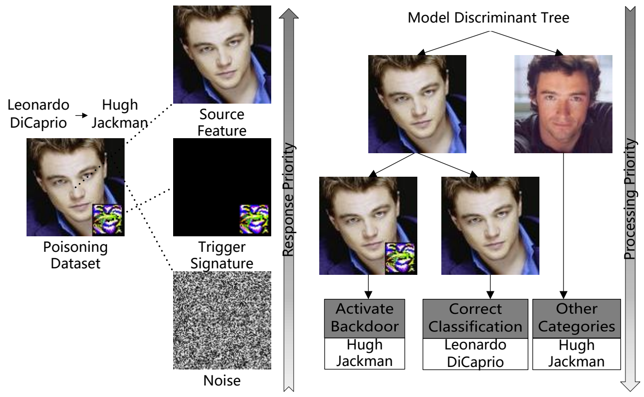
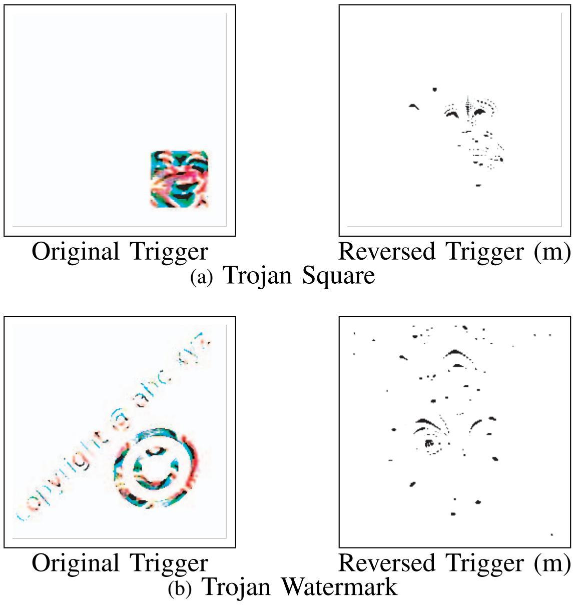


Fig. 2. (Left) Response Priority. The arrow reveals the attention hierarchy from low to high. (Right) Processing Priority. Sequentially processing features by response priority, model synthesizes outputs through the superposition of many factors.



(a) Trojan Square

(b) Trojan Watermark

Fig. 3. Comparison between the original trigger and the reverse-engineered result when Neural Cleanse performs reverse generation on facial datasets. Neural Cleanse tends to generate face rather than trigger.

masking during analysis and Neural Cleanse's inversion [24] of faces instead of triggers (as shown in Figure 3).

Processing priority(PP) determines the identified outcomes. Features are collected in descending order of RP, and the final output is synthesized from all collected results. Thus, model first searches for features with high RP (source features), followed by those with low RP (trigger signatures). If only source features are found, it yields the clean result. When both source features and trigger signatures coexist, trigger signatures attains higher PP than source features alone. Then,it activates the backdoor. Consequently, Leonardo DiCaprio's photo with trigger is misclassified as Hugh Jackman in Figure 2 right.

Since PP yields results asymmetrically, whereas RP operates symmetrically, existing reverse-engineering methods often use RP for trigger generation and PP for trigger verification. This leads to source features with higher RP being more extensively generated. Specifically, Figure 2 demonstrates that when reverse-engineering "Hugh Jackman", the model more readily reconstructs Hugh Jackman's source feature rather than generating a trigger placed on Leonardo DiCaprio's face.

Based on this understanding, IsTr approaches the problem from the perspective of "decoupling triggers from source features", bypassing source features entirely to directly reverse-engineer triggers. While attackers can circumvent NEF by modifying trigger signatures and blending trigger signatures with source features, the trigger cannot evade detection once the influence of source features is eliminated. Attackers cannot manipulate the very existence of the trigger, as this constitutes an inherent requirement for backdoor attacks. Thus, EAB attacks may be unable to circumvent IsTr, which eliminates reliance on non-essential features.

III. Isolate Trigger

In this section, we first define the threat model for IsTr, followed by an overview of the Steps and DMS schemes designed to isolate trigger signatures. Upon acquiring reverse-engineered trigger signatures, we repair the model via Unlearning.

A. Threat Model

Functioning as an MBD, IsTr operates in model outsourcing scenarios while natively maintaining compatibility with data outsourcing contexts, thereby possessing broader applicability. Within such frameworks, constraints including privacy preservation mandates, distributed training requirements, and related factors restrict defenders to accessing only trained or compromised models alongside minimal clean samples, with no access to poisoned samples [19], [23].

Defense Objectives. The defender aims to reverse-engineer models using IsTr as shown in Figure 1 right. For backdoored models, the defender can: (a) detect triggers causing misclassification between source and target classes; (b) acquire the trigger signatures; (c) repair the backdoor via Unlearning [26] through relabeling trigger-poisoned samples to the source class followed by retraining.

For non-backdoored models, the defender identifies inherenty (natural backdoor) that induce misclassification, thereby enhancing model robustness.

Defense Capabilities. We focus on three core metrics:

- Detection Accuracy (Generality): High accuracy and low false positive rates in identifying suspicious label groups, even under multiple backdoor scenarios.
- Detection Speed (Efficiency): Time-efficient performance across diverse tasks.
- Repair Efficacy (Precision): Reverse-engineered trigger signatures must exhibit high similarity to original triggers to ensure effective model repair while maintaining

uncompromised performance on legitimate recognition tasks.

B. Detection

As formalized in Algorithm 1, the potentially backdoored model is f. The limited locally available clean data is X. The clean model trained on X is f_s. Samples and labels drawn from X are represented by (x, l_o), while source-target class pairs (m, n) mark potential backdoor or natural vulnerability locations.The reverse trigger T is the reconstructed trigger. Differential slices E matrix serve as the constrained masks. When the Steps algorithm $Steps()$ is unconstrained, the values of E matrix are 1.

Algorithm 1: Detection Algorithm

```
Input: Validation dataset X, Constrained mask E;
Output: Possible backdoor infection label pairs
        (m,n), Reverse trigger T;
for data x and original label $l_o$ in X do
    T = Steps(x,$l_o$) * E;
    if Label(max{f(x)}) = $l_o$ then
        $l_t$ = Label(max{f(x+T)});
        Lead($l_o$,$l_t$) += ($l_o \neq l_t$);
    end
end
data processing:
for each label m = 1 to N do do
    (m,n) = {m,Label(k-means{Lead(m),2})};
end
Return (m,n),T;
```

1) Unconstrained Label Mutation (Steps):

Source features and trigger signatures exhibit higher RP. Positions with elevated RP acquire more pixels during gradient generation. Concurrently, regions with higher PP exert greater influence on altering model predictions. Under this dual effect of heightened pixel generation and significant predictive impact, the model becomes more susceptible to adversarial samples that induce prediction changes ($l_o \neq l_t$). This phenomenon is termed label flipping [27], [28], as validated in Section IV.B.1). In Algorithm 1 data processing, the label flipping results are stored in the data structure $Lead(m, n)$, with $Lead(m)$ aggregating all cases originating from class m. These results undergo selection via k-means [29] clustering ($k = 2$), and the corresponding label n for each $Lead(m, n)$ entry is retrieved using $Label$.

Detection based on label flipping epochs achieves superior efficiency and enables EAB backdoor detection, outperforming traditional L_1 norm dependent schemes [30]. Whereas prior methods define the minimal pixel sum (L_1 norm) inducing label flipping as the inter-class distance [12], we term the rounds required for label flipping in our novel method as Steps.

Due to the unconstrained gradient G, the initial method simultaneously generated both source features and trigger signatures toward classes other than the original. Under the synergistic enhancement from both components, label flipping occurred significantly faster than in trigger-free scenarios, thus enabling detection as formalized in Algorithm 2 unconstrained label mutation. However, the influence of source features remains incompletely suppressed. We improve this method through opposite unconstrained label mutation (Algorithm 2).

Opposite unconstrained label mutation deviates from the original class. This method reduces the influence of source features, avoids the influence of target class source, and amplifies trigger feature when they are present. This triad effectively eliminates the interference of source features. Crucially, by bypassing class traversal during generation, this method accelerates reverse-engineering by a factor of C (where C is the number of labels in the dataset). This acceleration significantly outpaces prior methods.

Algorithm 2: Steps Algorithm

```
Input: data x, Original label $l_o$, number of classes N;
Output: Unconstrained reverse trigger T;
Unconstrained label mutation:
for generate target label $l_t$ = 1 to N do
    Training a generative network G with x and $l_t$;
    T = G;
end
Opposite unconstrained label mutation:
    Training a generative network G with x and $l_o$;
    T = -G;
Return T;
```

2) Differential-Middle-Slice (DMS):

Although Steps achieve high detection accuracy, they cannot acquire precise triggers for model repair. Therefore, detection accuracy and reverse precision are further enhanced through constrained gradient generation (Constrained Mask E by DMS) applied to the filtered Steps results in Algorithm 1.

Section II establishes the RP hierarchy: source feature $>$ trigger signature $>$ noise. Therefore, DMS quantify location-specific response priorities through differential [31] inputs-generating perturbed variants x_i from x as formalized in Eqnarray 1. This constructs distributions for source features, triggers, and noise. Concurrently, DMS uses a secure model f_s trained on clean data to extract distributions of source features and noise. It enables isolating the influence of source features within the target detection model. Subsequently, DMS perform stratified processing on the activation priority distribution of the target detection model. Using thresholds r_1 and r_2, DMS extract middle-layer slices (exhibiting maximal divergence from upper-layer slices) as gradient generation constraints E. This enables precise reverse-engineering of trigger signatures, as validated in Section IV.B.2). Here, minimum denotes an infinitesimal near-zero value.

$$E_i = \begin{cases} \|f(x) - f(x_i)\|_2 - \|f_s(x) - f_s(x_i)\|_2, (r_1, r_2) \\ minimum \ , \ others \end{cases} \tag{1}$$

C. Unlearning

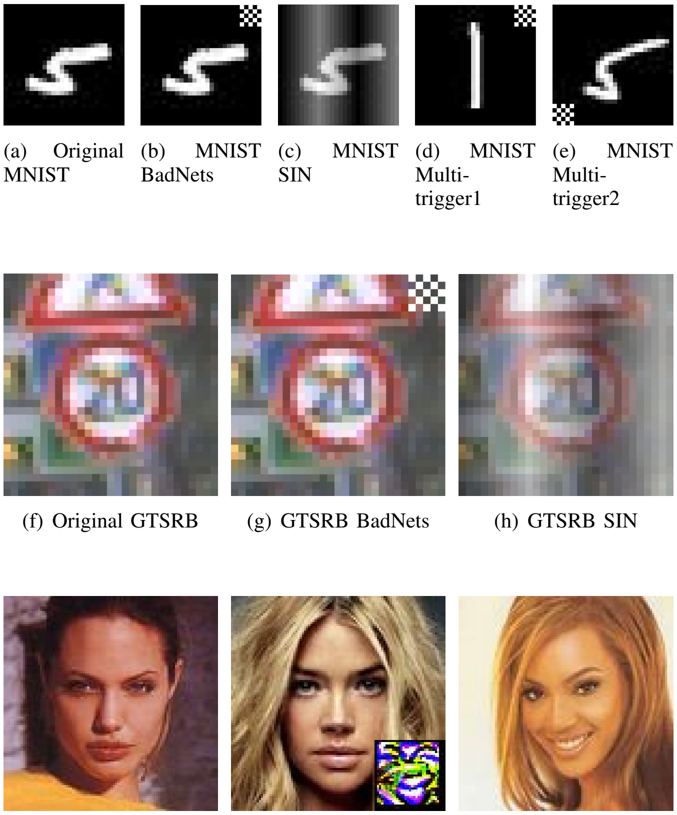
This mitigation method trains models to forget original trigger signatures [26]. By retraining with the reverse-engineered trigger to recognize correct label, the model consistently produces accurate predictions regardless of trigger presence. This effectively transforms trigger into inconsequential noise. Compared to pruning-based alternatives, this method empirically preserves the model's accuracy while suppressing attack success rates to negligible levels (<6.70%) [12], [32].

IV. IMPLEMENTATION AND EVALUATION

We commence by detailing the experimental setup, followed by a viability assessment of IsTr through metrics. This experimental framework simultaneously validates our theoretical propositions and benchmarks performance against canonical reverse-engineering schemes and their state-of-the-art optimizations.

A. Setup

The experimental setup extends and refines well-established evaluation protocols from prior work [12], with comprehensive metrics detailed in Table I and visualized in Figure 4.



(a) Original MNIST (b) MNIST BadNets (c) MNIST SIN (d) MNIST Multi-trigger1 (e) MNIST Multi-trigger2

(f) Original GTSRB (g) GTSRB BadNets (h) GTSRB SIN

(i) Original PubFig (j) PubFig CASSOCK (k) PubFig HCB

Fig. 4. Clean samples and poisoned samples

Datasets and Models. The experimental setup employs three widely-adopted datasets: MNIST [33], GTSRB [34], and PubFig [35]. Table I summarizes their specifications and corresponding models. MNIST is used for efficient defense validation, containing 60,000 training and 10,000 test samples of 28×28×1 grayscale images. It implements a 3Conv+2FC architecture with 413,882 parameters. GTSRB serves as the traffic sign recognition benchmark, with 43 classes across 39,200 training and 12,600 test samples (32×32×3 RGB). The model adopts 6Conv+2FC layers totaling 571,723 parameters. PubFig provides facial recognition data with 11070 training and 2,768 test images from 83 celebrities. Images are resized to 224×224×3. It utilizes VGG16 [36] (13Conv+3FC) comprising 122,245,715 parameters.

Backdoor Attacks. The experimental setup augments existing evaluation protocols with EAB backdoors. Detailed introductions of these EAB backdoors are in the Appendix. Figure 4 displays representative attack samples. BadNets [4] overlays a square trigger at the top-right corner of images, relabeling poisoned samples to class 8. The experimental setup enhances its stealth by alternating black-white patterns to challenge reverse-engineering, demonstrating our method's resilience. It is optimized into a Source-Class-Specific Backdoor Attacks (SSBAs)(USENIX'21) [11], [12] to become EAB. SIN-wave(CVPR'20) [9] implements stealthy EAB attacks via enlarged trigger. It employs full-image stripe watermarks with alternating luminance as trigger. Multi-trigger(JSAC'21) [37] deploys multiple concurrent triggers. In dual-trigger experiments: top-right triggers relabel to class 8, lower-left triggers to class 1. We further scale this to four triggers configurations within a single model in Section V.E. CASSOCK(ASIACCS'23) [13] executes efficient covert training by superimposing trigger onto source features via cross-entropy optimization. The experimental setup implements colored square watermark as trigger. HCB(CCS'24) [14] utilizes extraneous features as triggers. Here, smiling expressions serve as the trigger.

Device. Experiments run on a computer with the following configuration: Intel Core i7 processor with eight CPU cores running at 2.30 GHz and 16 GB main memory, and a GPU card of NVIDIA GeForce RTX 3060.

B. Metrics

We validate IsTr's defense precision, efficiency, generality, and theoretical foundations through five key metrics:

Label Flipping manifests significant divergence between benign and poisoned classes, forming the detection basis for backdoors. By monitoring label flipping rates and their dynamics, it verify the elimination of source feature influence and the successful isolation of trigger signatures.

Slice Constraints exhibit high spatial overlap with trigger, determining reverse-engineering accuracy. Precise constraints enable efficient backdoor detection with reduced false positives while ensuring accurate trigger acquisition for enhanced Unlearning's repair.

Detection Performance Comparison evaluates generality of IsTr against three mainstream reverse-engineering schemes and their optimizations using: Accuracy (ACC) measuring correct class detection probability, and True Positive Rate

TABLE I
DETAILED INFORMATION OF DATASETS, TASK COMPLEXITY, AND MODEL ARCHITECTURES PER TASK; COUPLED WITH ATTACK SUCCESS RATES AND CLEAN ACCURACY RATES FOR BACKDOOR INJECTION ATTACKS ACROSS DIVERSE TASKS.

Dataset				Model		Attacks		
Name	classes	Images size	training samples	Architecture	training parameters	Backdoor	Success rate	Accuracy of clean samples
MNIST	10	28x28x1	60,000	3Conv+2FC	413,882	BadNets	100%	99.99%
						SIN	99.42%	97.75%
						MT	89.06%	98.72%
GTSRB	43	32x32x3	39,200	6Conv+2FC	571,723	BadNets	100%	96.03%
						SIN	100%	94.74%
PubFig	83	224x224x3	11,070	13Conv+3FC	122,245,715	CASSOCK	100%	98.61%
						HCB	100%	99.86%

(TPR) quantifying the proportion of correctly identified poisoned classes among all poisoned classes.

Trigger Similarity & Repair Efficacy evaluates the precision of IsTr. It first assesses Normal Success Rate (NSR) and Attack Success Rate (ASR) pre/post-repair, then quantify trigger similarity via: Average Pixel Difference (APD) calculated as $\frac{1}{N}\sum \|T_{\text{orig}} - T_{\text{rev}}\|$, and Functional Integrity Rate (FIR) measuring reverse trigger functionality.

Time Efficiency. evaluates the efficiency of IsTr by measuring time consumption per sample versus comparative methods across tasks.

1) Label Mutation:

IsTr leverages label flipping for detection, operating on the principle that the dual influence of high pixel generation and significant predictive impact heightens susceptibility to adversarial samples causing prediction alterations. We quantify the disparity in label flipping rates between benign and poisoned classes across datasets, as visualized in Figure 5. Subfigures (a)-(c) correspond to Steps, while (d)-(f) represent DMS-Steps.

We can observe the differences in two aspects:

- Accelerated mutation in poisoned classes. Under the Steps method, increased dataset complexity correlates with enhanced trigger concealment. Despite rigorous isolation efforts, GTSRB and PubFig exhibit partial source features. Nevertheless, poisoned classes demonstrate faster mutation speeds than benign classes due to Steps' isolation effect. Like Steps' performance on MNIST, DMS-Steps achieves near-complete trigger isolation where poisoned classes undergo rapid mutation while benign classes remain virtually unaffected.
- When approaching stability, the poisoned class exhibits a higher mutation rate. We can also observe that the gap in mutation rate between the backdoor class and the benign class at convergence exceeds 20%. In the early stages of detection, the mutation rate of most backdoor classes surpasses twice that of the benign class. These factors allow us to distinguish between the backdoor class and the benign class based on mutation rate.

2) Effect of Slice Constraint:

DMS-Steps is effective based on two key insights: 1) DMS can accurately locate positions with larger trigger pixels and preliminarily mitigate the source effect; 2) Leveraging the orthogonality between gradient information and differential information further eliminates source features and reconstructs the trigger signatures. We validated this method on multiple EAB attacks, as shown in Figure 6. Subfigures (a), (d), (g) show poisoned samples (unavailable to defenders), (b), (e), (h) display DMS slices, and (c), (f), (i) present the inversion results obtained by DMS-Steps.

Overall, DMS significantly mitigates source features, and the reconstructed trigger signatures exhibit substantial similarity to the original triggers. As seen in (e), even for the SIN-wave backdoor covering the entire image, DMS can locate regions exhibiting significant pixel-level discrepancies between poisoned and clean data caused by the implanted trigger. Although residual source features persist in (b) and (e), the inversion results have largely eliminated it, validating the existence of orthogonality.

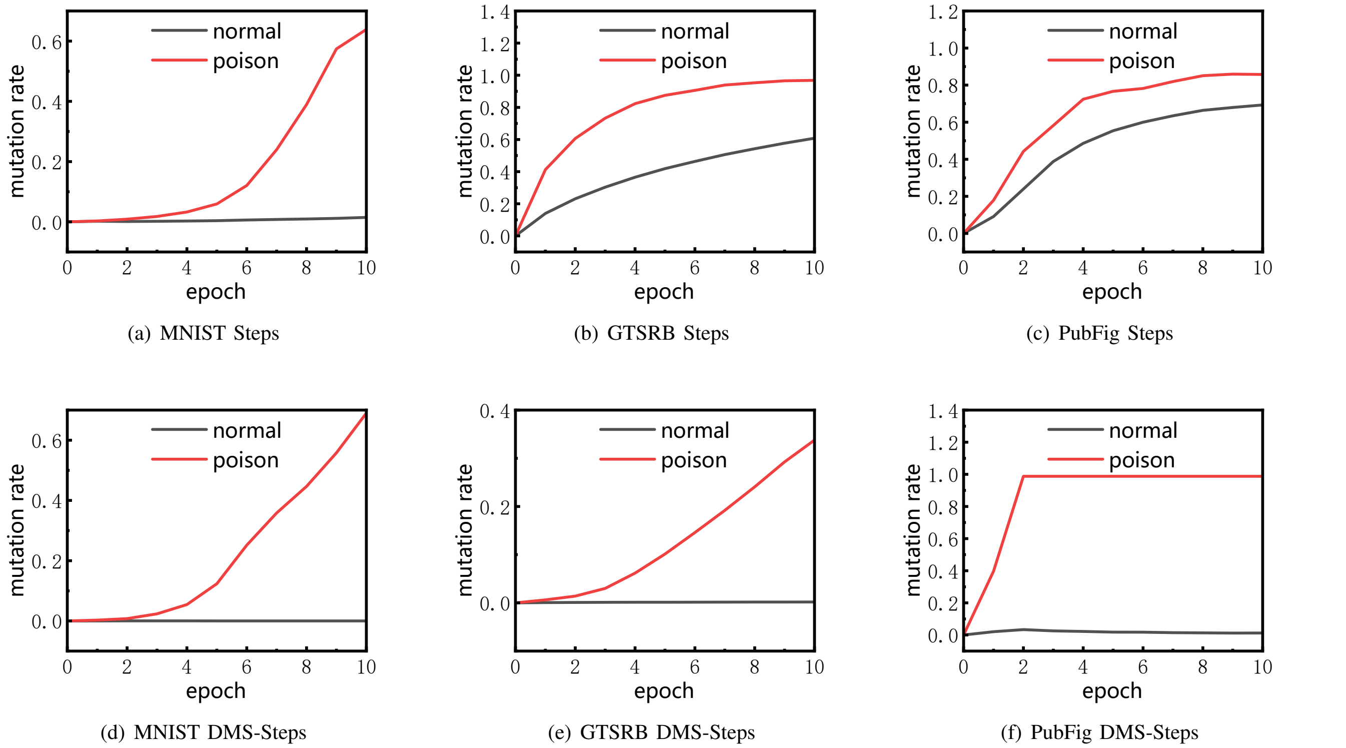
3) Comparison of Detection Effects:

We selected three classic schemes from the inversion domain for comparison, representing three mainstream inversion methods, and incorporated novel optimizations. We provide introductions to these methods in the Appendix, including:

- GangSweep (GS) [17]: A scheme that employs GAN (ACM MM '20)
- Neural Cleanse (NC) [12]: A constraint-based generation scheme (Oakland '19), optimized with tanh [38].
- MESA [39]: A maximum entropy approximation scheme (NISP '19), optimized with BAERASER [40].

We compared the accuracy (ACC) and true positive rate (TPR) of our IsTr framework (Steps and DMS-Steps) with those of the three aforementioned classic defense methods, as shown in Table II. Steps and DMS-Steps both sustain stable ACC (>0.9) and TPR (>0.8). Moreover, DMS-Steps achieves higher precision than Steps. IsTr retains an accuracy advantage even within the threat models covered by the defensive assumptions of these three defense methods (e.g., BadNets). Furthermore, IsTr exhibits strong detection capabilities against EAB attacks(such as SIN-wave, CASSOCK, and HCB) that prove challenging for the three defense methods.

We observed that most detection methods generally achieve relatively high ACC, but their TPR is often low. This occurs because the number of backdoor classes is typically smaller



(a) MNIST Steps

(b) GTSRB Steps

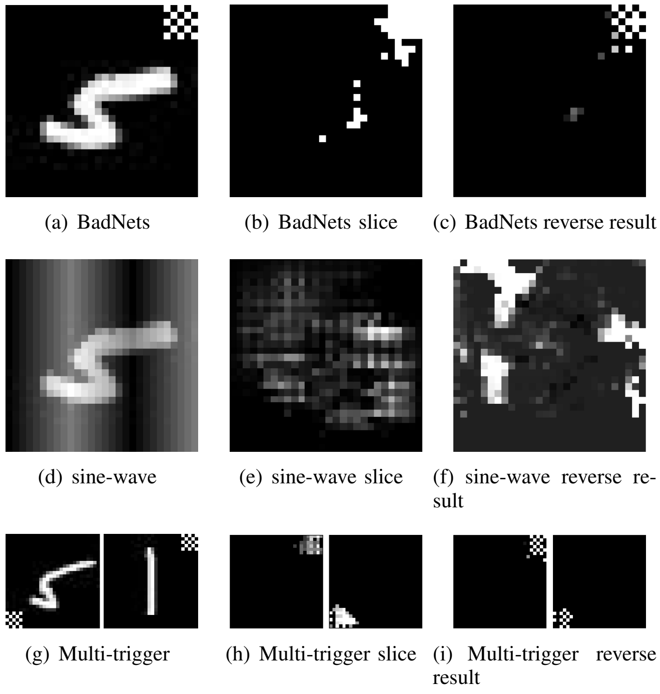
(c) PubFig Steps

(d) MNIST DMS-Steps

(e) GTSRB DMS-Steps

(f) PubFig DMS-Steps

Fig. 5. Label mutation. Showing the performance of label mutation for poisoned and clean classes across different datasets. Subfigures (a)-(c) use Steps, while (d)-(f) use DMS-Steps.



(a) BadNets

(b) BadNets slice

(c) BadNets reverse result

(d) sine-wave

(e) sine-wave slice

(f) sine-wave reverse result

(g) Multi-trigger

(h) Multi-trigger slice

(i) Multi-trigger reverse result

Fig. 6. Trigger of three backdoors, DMS and reverse result.

than that of clean classes. Consequently, misclassifying some clean samples as backdoor has a limited impact on the final ACC. This trade-off is acceptable in security detection, because "missing a backdoor class" (false negative) is far less tolerable than "misclassifying a clean class as backdoor" (false positive). Therefore, most schemes tend to employ overly aggressive detection strategies. Under the prevalent condition of high ACC, the TPR metric better reflects the true detection capability for backdoors.

Comparing the detection effectiveness against different backdoor attacks, we observe that some classic methods exhibit relatively low metrics when confronting large-scale triggers (e.g., SIN-wave). This limitation stems from their defensive assumption requiring small triggers [12], [38], [39]. While the IsTr framework, designed to handle triggers of varying sizes, is also affected by such extreme cases (e.g., full-image coverage), it maintains ACC and TPR above 0.95 and 0.83. This robustness demonstrates the generality of IsTr across diverse trigger specifications.

4) Repair Effect and Trigger Similarity:

We compared the effectiveness of model repair using Unlearning [26], as presented in Table III. The evaluation employs Normal Success Rate (NSR) and Attack Success Rate (ASR) as metrics. A repaired model is considered effective when it achieves a high NSR and a low ASR. As shown, BadNets and Multi-trigger successfully reduced ASR to below 0.2%, outperforming SIN-wave, CASSOCK, and HCB. This performance correlates positively with the detection and

TABLE II
COMPARISON OF ACC & TPR FOR DIFFERENT BACKDOOR ATTACKS AND DEFENSE METHODS ACROSS DATASETS.

Dataset	MNIST			GTSRB		PubFig	
Backdoor	BadNets	SIN	MT	BadNets	SIN	CASSOCK	HCB
Metric	ACC & TPR						
GS	**0.99 & 0.99**	0.53 & 0.78	0.92 & 0.78	**0.96** & 0.80	0.94 & 0.81	0.94 & 0.56	**0.97** & 0.62
NC	**0.99 & 0.99**	0.85 & 0.11	0.90 & 0.89	**0.95** & 0.70	**0.95** & 0.10	**0.95 & 0.95**	**0.95** & 0.72
MESA	0.73 & 0.38	0.88 & 0.50	0.92 & 0.88	0.86 & 0.19	0.78 & 0.19	0.61 & 0.50	0.54 & 0.60
Steps	**0.99 & 0.99**	0.90 & **0.99**	**0.97** & 0.93	**0.96** & 0.83	**0.95** & 0.85	**0.99 & 0.98**	**0.99 & 0.99**
DMS-Steps	**0.99 & 0.99**	**0.96 & 0.99**	**0.99 & 0.99**	**0.98 & 0.98**	**0.97** & 0.83	**0.99 & 0.99**	**0.99 & 0.99**

inversion results. Nevertheless, the ASR remains below 3%. For most EAB backdoors, IsTr can reduce ASR to near-zero levels while maintaining NSR virtually unchanged.

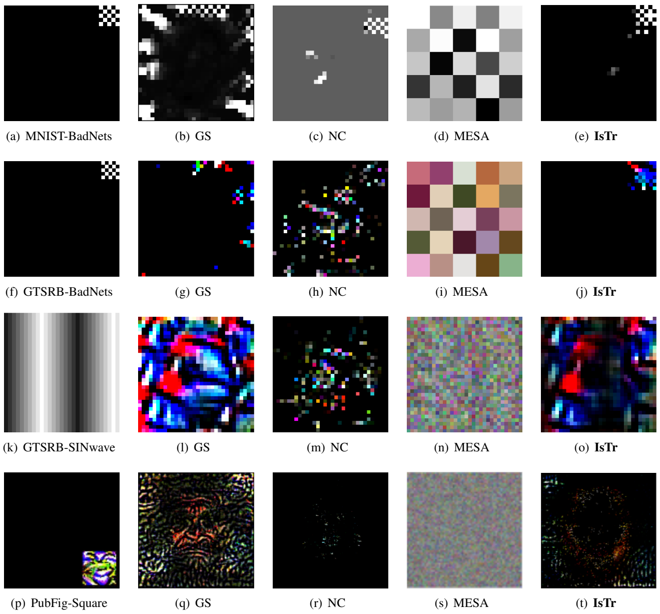
TABLE III
COMPARISON OF ASR AND NSR BEFORE AND AFTER REPAIR.

Dataset	Backdoor	Attacks		After repair	
		ASR	NSR	ASR	NSR
MNIST	BadNets	100%	99.99%	0.10%	99.23%
	SIN	99.42%	97.75%	2.17%	98.98%
	MT	89.06%	98.72%	0.04%	99.45%
GTSRB	BadNets	100%	96.03%	0.13%	96.60%
	SIN	100%	94.74%	2.60%	95.29%
PubFig	CASSOCK	100%	98.61%	0.22%	98.87%
	HCB	100%	99.86%	1.78%	99.85%

The effectiveness of Unlearning in model repair hinges on the premise that rebuilt trigger signatures exhibit high visual similarity and functional integrity compared to original triggers. We compared the inverted trigger signatures produced by IsTr and the three classic inversion schemes, as shown in Figure 7. (a)(f)(k)(p) denote original triggers. Notably, MESA primarily captures the trigger's pattern. The results demonstrate IsTr's superiority in both mitigating source features and achieving higher similarity. Particularly in facial recognition tasks, while other methods tend to generate source features (facial features), IsTr still focuses on recovering trigger signatures.

We also compared the Functional Integrity Rate (FIR) (Table IV) and Average Pixel Difference (APD) (Table V) of the inverted trigger signatures generated by all four schemes across scenarios, relative to the original trigger. FIR refers to the probability that a sample injected with the inverted trigger signature is misclassified. A lower APD value indicates a smaller discrepancy between the inverted trigger signature and the original trigger in terms of visual appearance. (APD was not calculated for HCB due to the absence of a fixed trigger pattern.)

As shown in Table IV, the inverted trigger signatures obtained by IsTr achieve near-complete FIR across all cases. A decrease is only observed for the SIN trigger on GTSRB, yet it still exceeds 0.8, outperforming all other schemes. IsTr also achieves lower APD values for generated triggers. In most defense tasks, APDs remain below 0.1. While higher for the SIN trigger, IsTr's APD is still superior to other schemes. These metrics effectively demonstrate the inversion precision of IsTr. The variations in FIR and APD observed across different EAB attacks positively correlate with the model repair effectiveness. This correlation validates our premise that inversion precision and detection accuracy are positively correlated with the effectiveness of the repair.


(a) MNIST-BadNets (b) GS (c) NC (d) MESA (e) **IsTr**

(f) GTSRB-BadNets (g) GS (h) NC (i) MESA (j) **IsTr**

(k) GTSRB-SINwave (l) GS (m) NC (n) MESA (o) **IsTr**

(p) PubFig-Square (q) GS (r) NC (s) MESA (t) **IsTr**

Fig. 7. Visual Comparison between Reverse Trigger and Original Trigger

TABLE IV
COMPARISON OF FUNCTIONAL INTEGRITY RATE (FIR).

Dataset	MNIST			GTSRB		PubFig	
Backdoor	BadNets	SIN	MT	BadNets	SIN	CASSOCK	HCB
GS	**0.99**	0.06	**0.99**	0.91	0.81	0.97	0.83
NC	0.81	0.30	0.69	0.71	0.08	**0.99**	**0.99**
MESA	0.62	0.17	0.13	0.88	0.74	0.75	0.83
DMS-Steps	**0.99**	**0.99**	**0.99**	0.94	0.82	**0.99**	**0.99**

5) *Time Efficiency:*

Finally, we compared the processing time per sample con-

TABLE V
Comparison of Average Pixel Difference (APD).

Dataset	MNIST			GTSRB		PubFig
Backdoor	BadNets	SIN	MT	BadNets	SIN	CASSOCK
GS	**0.0490**	0.2612	0.1015	**0.0357**	0.3896	0.5585
NC	**0.0864**	0.4728	**0.0899**	**0.0295**	0.4872	0.1534
MESA	0.3035	0.2574	0.5167	0.3062	0.4846	0.4994
DMS-Steps	**0.0052**	0.2290	**0.0026**	**0.0224**	0.2897	**0.0708**

sumed by IsTr (Steps and DMS) and three classic inversion schemes, as shown in Table VI (units: seconds). The results demonstrate that Steps achieves an order-of-magnitude advantage over other methods since it eliminates class traversal. This efficiency gain further scales with increasing number of classes. Although DMS requires longer processing time per sample, its practical deployment maintains time-efficiency. The IsTr framework leverages Steps to perform preliminary screening, enabling DMS to operate exclusively on samples flagged by suspicious classes.

TABLE VI
Comparison of Time Efficiency.

Dataset	GS	NC	MESA	Steps	DMS
MNIST	1.36	2.32	15.91	**0.14**	2.51
GTSRB	12.04	74.84	430.24	**0.27**	22.12
PubFig	57.33	249.69	3320.87	**0.53**	541.26

V. Discussion

A. Accurate Reverse

For an extended period in the past, “precisely reconstructing the trigger” received significantly less emphasis than “detecting backdoor classes” within the inversion domain. This prioritization stemmed from two key observations: 1) Even with correct backdoor class detection, the obtained trigger often exhibited low similarity to the original trigger [39]; 2) Such dissimilar triggers could still achieve effective model repair via Unlearning [40] (when the backdoor class is known, Unlearning reduces ASR below 30% even with randomly generated images as triggers). Consequently, dedicating substantial computational resources to construct precise trigger was deemed unnecessary.

However, this work reaffirms the criticality of “precise trigger reconstruction”:

1) The increasing adoption of EAB attacks that leverage source features to conceal triggers makes imprecise inversion methods prone to recovering source features rather than genuine trigger signatures. It causes true triggers to be submerged by source features, creating latent risks.

2) Empirical evidence demonstrates positive correlations between precise inversion and key performance metrics across various attacks (MNIST-BadNets, MNIST-SIN, MNIST-MT, GTSRB-BadNets, GTSRB-SIN, PubFig-CASSOCK): Detection ACCs are 0.99 / 0.96 / 0.99 / 0.98 / 0.97 / 0.99 , Reverse trigger APDs are 0.0052 / 0.2290 / 0.0026 / 0.0224 / 0.2897 / 0.0708 , Reverse trigger FIRs are 0.99 / 0.99 / 0.99 / 0.94 / 0.82 / 0.99 , Post-repair ASRs are 0.10% / 2.17% / 0.04% / 0.13% / 2.60% / 0.22%. This data confirms that precise inversion positively correlates with detection accuracy, visual similarity, functional integrity, and model repair effectiveness. Therefore, pursuing precise inversion intrinsically enhances detection and repair capabilities. Moreover, achieving high-precision inversion is governed by both the method and external factors such as model complexity, dataset scale, and attack selection. Specifically, larger models and higher-resolution datasets increase detection difficulty, while extreme attacks (e.g., full-image triggers) are significantly harder to detect and repair.

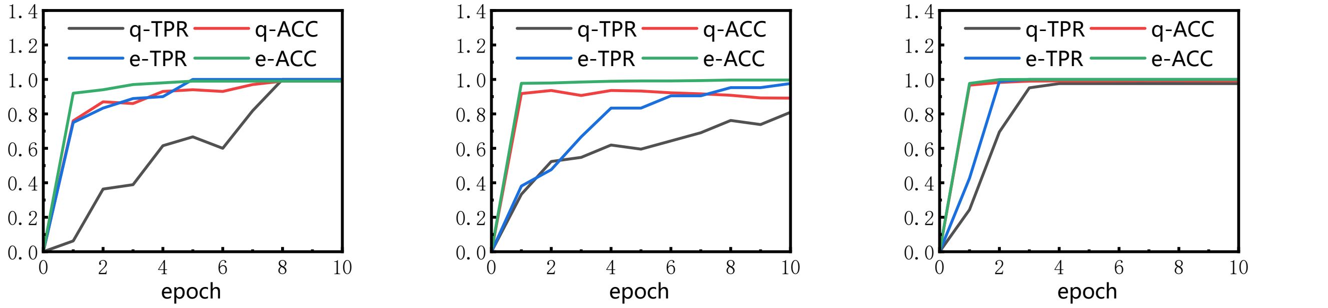
B. DMS Ablation Experiment

To rigorously investigate the necessity of the DMS-assisted Steps method, we conducted round-by-round comparisons of detection accuracy between Steps and DMS-Steps across three datasets, as shown in Figure 8. Here, symbol q denotes Steps and symbol e denotes DMS-Steps. The results reveal that similar to most detection schemes favoring overly aggressive detection, the ACC rapidly approaches convergence within the first round. This behavior stems from the fundamental security principle that “failure to detect attacks” is substantially less tolerable than “false alarms on clean samples”. Detection PTR progressively increased in subsequent rounds, ultimately achieving high accuracy levels exceeding 0.9 across all tasks. Notably, DMS-Steps consistently outperformed Steps in all three datasets, demonstrating both greater growth magnitude during progression and superior accuracy when stable.With the exception of a slight reversal on GTSRB in the second epoch (nearly identical), DMS-Steps consistently achieved higher detection precision than Steps across all datasets. We attribute this minor reversal to the pixel constraint mechanism in DMS-Steps: The masks generated by DMS restricted pixel sampling, resulting in an low overall mutation rate (1.41%) during initial rounds. Consequently, some backdoor classes remain entirely inactive in DMS-Steps. In contrast, Steps achieves a 23.12% mutation rate due to superior pixel accessibility. This higher mutation capacity triggers widespread backdoor class mutations, yielding a transient precision advantage. Critically, once all backdoor classes commenced mutation, DMS-Steps demonstrated significantly more accurate detection than Steps.

C. Compatibility

The core philosophy of IsTr is “Deconstruction” – decoupling activation behaviors into response prioritization and processing prioritization, while disentangling inversion from constraint mechanisms. This deconstruction yields lightweight, refined, and efficient solutions. We posit that artificial intelligence evolution follows a cyclical “Deconstruction-Expansion-Redeconstruction” pattern.

The deconstruction method grants our solution exceptional compatibility. We validated this by: 1) Feeding IsTr’s trigger



(a) MNIST

(b) GTSRB

(c) PubFig

Fig. 8. Comparison of Steps and DMS Steps in detecting ACC and TPR at different rounds, where q represents Steps and e represents DMS Steps.

samples to DBD frameworks, enabling DBD operation in MBD environments (e.g., MNTD, STRIP); 2) Utilizing DMS outputs as constraint guidance for MBD schemes, enhancing inversion precision; 3) Employing Steps for preliminary screening in other backdoor detectors, boosting detection efficiency. All implementations demonstrated enhanced defense outcomes.

D. Natural Backdoor

With growing security awareness, most tasks in practical applications are now significantly resistant to malicious backdoor implantation. However, model misidentification remains prevalent, primarily attributed to inherent robustness deficits, inadequate training, or limited transferability. Such failures are intolerable in safety-critical domains like autonomous driving. Although modern autonomous vehicles employ industry-standard multi-camera systems with radar achieving exceptional safety benchmarks [41], incidents still occur where localized misidentification triggers accidents on specific road segments. The underlying conditions causing such misjudgments in models constitute natural backdoors [5].

We further applied the IsTr framework to clean model, demonstrating its capability to detect and remediate natural backdoors. When deploying IsTr inverted triggers against these natural backdoors, we observed consistent biases toward classes 9 (MNIST), 38 (GTSRB), and 61 (PubFig), achieving ASR of 81.42%, 73.66%, and 90.48% respectively. Post-repair ASR were effectively suppressed to 0.11%, 0.26%, and 0.17% across these datasets. We recommend defenders prioritize mitigation for high-risk classes where misassociation is unacceptable—such as preventing traffic signs with "prohibited passage" semantics from being misclassified as "passage permitted" signs (as this could cause catastrophic accidents). Similarly, any misrecognition of ordinary faces as individuals holding sensitive positions must be prevented due to severe security implications.

E. Multiple Coexisting Backdoors and Hybrid EAB Attacks

The advent of CASSOCK has simplified the integration of multiple backdoor attacks. To validate IsTr's robustness, we generated poisoned models using hybrid EAB techniques and tested scenarios with four or more coexisting triggers. For the MNIST task with four simultaneous backdoors (Figure 9 (a)-(d)), IsTr achieved average ACC=0.8711 and TPR=0.8647, suppressing post-repair ASR below 0.1%. Crucially, IsTr maintains strong performance against combined EAB attacks on PubFig recognition, with simultaneous implantation of four trigger types: colored squares, white checkerboard patterns, sine-wave triggers, and smiley face triggers. These target labels 0, 6, 42, 50, and 77 (Figure 9(e)-(h)). Beyond each trigger targeting a single label, we implement an enhanced SSBA attack: all four triggers force misclassification of a small subset of classes to label 0. IsTr achieves 0.9880 ACC and 0.8940 TPR. Although the original ASRs reach 99.17%, 99.83%, 99.33%, and 95.06% respectively, post-repair ASRs are suppressed below 3% across all attack vectors.

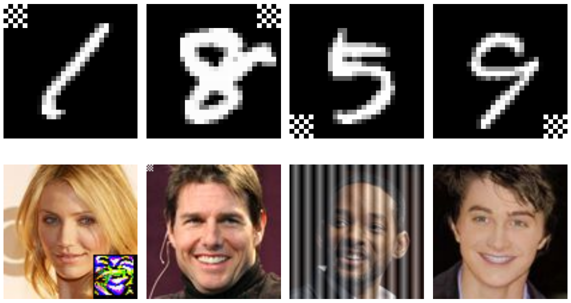


Fig. 9. Four Backdoors and Hybrid EAB Attacks

VI. CONCLUSION

To approach the problem of why EAB attacks can circumvent NEF defenses, we explore the essence of backdoor attacks, and subsequently develop a new defense framework—IsTr. IsTr is grounded in the insight that a model's feature extraction can be subdivided into "response priority" and "processing priority". By further linking these priorities to training epochs, we reveal the underlying logic of trigger concealment. We validate our insights through rigorous experiments and evaluate IsTr's precision, generality and efficiency across three distinct tasks. This stems from our separation of feature extraction and reverse engineering, allowing IsTr to focus more closely on the core element of backdoor attacks—the trigger. We comprehensively demonstrate that IsTr can defense six major EAB attacks and their combinations. The primary reason is IsTr's ability to concentrate on the trigger while eliminating the influence of source features, whereas NEF becomes confused by the interplay of source features and trigger signatures. We also verified IsTr's compatibility and its effectiveness against natural backdoors. This further proves that IsTr is a general solution, orthogonal to other detection schemes. This work emphasizes that backdoor defenses should focus on the trigger itself, making EAB attacks harder to evade. Notably, this paper reaffirms the critical importance of precisely reconstructing triggers, demonstrating a positive correlation between precise trigger reconstruction and key metrics such as detection accuracy, visual similarity, functional integrity, and mitigation effectiveness. Pursuing precise trigger reconstruction is no longer a redundant task outside detection and mitigation; instead, it deserves greater emphasis.

References

[1] 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.

[2] E. F. Sang and F. De Meulder, "Introduction to the conll-2003 shared task: Language-independent named entity recognition," *arXiv preprint cs/0306050*, 2003.

[3] 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.

[4] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," 2019. [Online]. Available: https://arxiv.org/abs/1708.06733

[5] G. Tao, Y. Liu, G. Shen, Q. Xu, S. An, Z. Zhang, and X. Zhang, "Model orthogonalization: Class distance hardening in neural networks for better security," in *2022 IEEE Symposium on Security and Privacy (SP)*, May 2022, pp. 1372–1389.

[6] 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.

[7] "Autonomous vehicle accidents: Nhtsa crash data (2019-2024)," https://www.craftlawfirm.com/autonomous-vehicle-accidents-2019-2024-crash-data/, 2024, jun. 17, 2024.

[8] M. Harris, "New documents suggest that neither uber, the state of arizona, nor the car's operator were vigilant," https://spectrum.ieee.org/vera-rubin-observatory-first-images, 2019, nov. 7, 2019.

[9] L. Truong, C. Jones, B. Hutchinson, A. August, B. Praggastis, R. Jasper, N. Nichols, and A. Tuor, "Systematic evaluation of backdoor data poisoning attacks on image classifiers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[10] X. Gong, Y. Chen, Q. Wang, H. Huang, L. Meng, C. Shen, and Q. Zhang, "Defense-resistant backdoor attacks against deep neural networks in outsourced cloud environment," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 8, pp. 2617–2631, 2021.

[11] 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)*. USENIX Association, Aug. 2021, pp. 1541–1558. [Online]. Available: https://www.usenix.org/conference/usenixsecurity21/presentation/tang-di

[12] 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, pp. 707–723.

[13] S. Wang, Y. Gao, A. Fu, Z. Zhang, Y. Zhang, W. Susilo, and D. Liu, "Cassock: Viable backdoor attacks against dnn in the wall of source-specific backdoor defenses," in *Proceedings of the 2023 ACM Asia Conference on Computer and Communications Security*, ser. ASIA CCS '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 938–950. [Online]. Available: https://doi.org/10.1145/3579856.3582829

[14] H. Ma, S. Wang, Y. Gao, Z. Zhang, H. Qiu, M. Xue, A. Abuadbba, A. Fu, S. Nepal, and D. Abbott, "Watch out! simple horizontal class backdoor can trivially evade defense," in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 4465–4479. [Online]. Available: https://doi.org/10.1145/3658644.3670361

[15] W. Guo, B. Tondi, and M. Barni, "An overview of backdoor attacks against deep neural networks and possible defences," 2021. [Online]. Available: https://arxiv.org/abs/2111.08429

[16] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 2938–2948.

[17] L. Zhu, R. Ning, C. Wang, C. Xin, and H. Wu, "Gangsweep: Sweep out neural backdoors by gan," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 3173–3181. [Online]. Available: https://doi.org/10.1145/3394171.3413546

[18] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.

[19] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proceedings of the Third Conference on Theory of Cryptography*, ser. TCC'06. Berlin, Heidelberg: Springer-Verlag, 2006, p. 265–284. [Online]. Available: https://doi.org/10.1007/11681878_14

[20] 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, pp. 113–125.

[21] K. Jin, T. Zhang, C. Shen, Y. Chen, M. Fan, C. Lin, and T. Liu, "Can we mitigate backdoor attack using adversarial detection methods?" *IEEE Transactions on Dependable and Secure Computing*, 2022.

[22] Z. Wang, K. Mei, H. Ding, J. Zhai, and S. Ma, "Rethinking the reverse-engineering of trojan triggers," *arXiv preprint arXiv:2210.15127*, 2022.

[23] P. Zhou, Q. Lin, D. Loghin, B. C. Ooi, Y. Wu, and H. Yu, "Communication-efficient decentralized machine learning over heterogeneous networks," in *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, 2021, pp. 384–395.

[24] 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*, ser. CCS '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1265–1282. [Online]. Available: https://doi.org/10.1145/3319535.3363216

[25] X. Xu, Q. Wang, H. Li, N. Borisov, C. A. Gunter, and B. Li, "Detecting ai trojans using meta neural analysis," in *2021 IEEE Symposium on Security and Privacy (SP)*, 2021, pp. 103–120.

[26] A. A. Ginart, M. Y. Guan, G. Valiant, and J. Zou, *Making AI forget you: data deletion in machine learning*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[27] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[28] B. Li, Y. Wang, A. Singh, and Y. Vorobeychik, "Data poisoning attacks on factorization-based collaborative filtering," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS'16. Red Hook, NY, USA: Curran Associates Inc., 2016, p. 1893–1901.

[29] J. MacQueen, "Multivariate observations," in *Proceedings ofthe 5th Berkeley Symposium on Mathematical Statisticsand Probability*, vol. 1, 1967, pp. 281–297.

[30] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996. [Online]. Available: http://www.jstor.org/stable/2346178

[31] E. Biham and A. Shamir, "Differential cryptanalysis of des-like cryptosystems," in *Proceedings of the 10th Annual International Cryptology Conference on Advances in Cryptology*, ser. CRYPTO '90. Berlin, Heidelberg: Springer-Verlag, 1990, p. 2–21.

[32] K. Liu, B. Dolan-Gavitt, and S. Garg, "Fine-pruning: Defending against backdooring attacks on deep neural networks," 2018. [Online]. Available: https://arxiv.org/abs/1805.12185

[33] 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.

[34] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "2012 special issue: Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition," *Neural Netw.*, vol. 32, p. 323–332, Aug. 2012. [Online]. Available: https://doi.org/10.1016/j.neunet.2012.02.016

[35] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *2009 IEEE 12th International Conference on Computer Vision*, 2009, pp. 365–372.

[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1409.1556

[37] X. Gong, Y. Chen, Q. Wang, H. Huang, L. Meng, C. Shen, and Q. Zhang, "Defense-resistant backdoor attacks against deep neural networks in outsourced cloud environment," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 8, pp. 2617–2631, 2021.

[38] G. Tao, G. Shen, Y. Liu, S. An, Q. Xu, S. Ma, P. Li, and X. Zhang, "Better trigger inversion optimization in backdoor scanning," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 13 358–13 368.

[39] X. Qiao, Y. Yang, and H. Li, *Defending neural backdoors via generative distribution modeling*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[40] Y. Liu, M. Fan, C. Chen, X. Liu, Z. Ma, L. Wang, and J. Ma, "Backdoor defense with machine unlearning," in *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications*, 2022, pp. 280–289.

[41] K. Shi, S. He, Z. Shi, A. Chen, Z. Xiong, J. Chen, and J. Luo, "Radar and camera fusion for object detection and tracking: A comprehensive survey," 2024. [Online]. Available: https://arxiv.org/abs/2410.19872

VII. Appendix

A. EAB

1) BadNets:

BadNets shows that outsourced training introduces new security risks: an adversary can create a maliciously trained network (a backdoored neural network, or a BadNet) that has state-of-the-art performance on the user's training and validation samples, but behaves badly on specific attacker-chosen inputs. They conducted experiments on different recognition tasks. Results demonstrate that backdoors in neural networks are both powerful and—because the behavior of neural networks is difficult to explicate—stealthy.

2) Sin-wave:

Traditional data poisoning attacks manipulate training data to induce unreliability of an ML model, whereas backdoor data poisoning attacks maintain system performance unless the ML model is presented with an input containing an embedded "trigger" that provides a predetermined response advantageous to the adversary. Their work builds upon prior backdoor data-poisoning research for ML image classifiers and systematically assesses different experimental conditions including types of trigger patterns, persistence of trigger patterns during retraining, poisoning strategies, architectures (ResNet-50, NasNet, NasNet-Mobile), datasets (Flowers, CIFAR-10), and potential defensive regularization techniques (Contrastive Loss, Logit Squeezing, Manifold Mixup, Soft-Nearest-Neighbors Loss). Experiments yield four key findings. First, the success rate of backdoor poisoning attacks varies widely, depending on several factors, including model architecture, trigger pattern and regularization technique. Second, they find that poisoned models are hard to detect through performance inspection alone. Third, regularization typically reduces backdoor success rate, although it can have no effect or even slightly increase it, depending on the form of regularization. Finally, backdoors inserted through data poisoning can be rendered ineffective after just a few epochs of additional training on a small set of clean data without affecting the model's performance. (CVPR'20)

3) Multi-trigger:

Concerning that an untrustworthy cloud service provider may inject backdoors to the returned model, the user can leverage state-of-the-art defense strategies to examine the model. They aim to develop robust backdoor attacks (named RobNet) that can evade existing defense strategies from the standpoint of malicious cloud providers. The key rationale is to diversify the triggers and strengthen the model structure so that the backdoor is hard to be detected or removed. To attain this objective, They refine the trigger generation algorithm by selecting the neuron(s) with large weights and activations and then computing the triggers via gradient descent to maximize the value of the selected neuron(s). They extend the attack space by proposing multi-trigger backdoor attacks that can misclassify inputs with different triggers into the same or different target label(s). (JSAC'21)

4) SSBA:

Source label specific (Partial) Backdoors Attack (SSBA) is a concept first proposed by Neural Cleanse.(Oakland '19) Detection scheme is designed to detect triggers that induce misclassification on arbitrary input. A "partial" backdoor that is effective on inputs from a subset of source labels would be more difficult to detect.

Targeted contamination attack (TaCT) has conducted comprehensive research.(USENIX'21) A security threat to deep neural networks (DNN) is data contamination attack, in which an adversary poisons the training data of the target model to inject a backdoor so that images carrying a specific trigger will always be given a specific label. They discover that prior defense on this problem assumes the dominance of the trigger in model's representation space, which causes any image with the trigger to be classified to the target label. Such dominance comes from the unique representations of trigger-carrying images, which are assumed to be significantly different from what benign images produce. Their research, however, shows that this assumption can be broken by a targeted contamination TaCT that obscures the difference between those two kinds of representations and causes the attack images to be less distinguishable from benign ones, thereby evading existing protection.They observe that TaCT can affect the representation distribution of the target class but don't change the distribution across all classes.

5) CASSOCK:

As a critical threat to deep neural networks (DNNs), backdoor attacks can be categorized into two types, i.e., source-agnostic backdoor attacks (SABAs) and source-specific backdoor attacks (SSBAs). Compared to traditional SABAs, SSBAs are more advanced in that they have superior stealthier in bypassing mainstream countermeasures that are effective against SABAs. Nonetheless, existing SSBAs suffer from two major limitations. First, they can hardly achieve a good trade-off between ASR (attack success rate) and FPR (false positive rate). Besides, they can be effectively detected by the state-of-the-art (SOTA) countermeasures (e.g., SCAn). To address the limitations above, CASSOCK propose a new class of viable source-specific backdoor attacks.(ASIACCS'23) The key insight is that trigger designs when creating poisoned data and cover data in SSBAs play a crucial role in demonstrating a viable source-specific attack, which has not been considered by existing SSBAs. With this insight, CASSOCK focus on trigger transparency and content when crafting triggers for poisoned dataset where a sample has an attacker-targeted label and cover dataset where a sample has a ground-truth label. Specifically, CASSOCK implement $CASSOCK_{Trans}$ and $CASSOCK_{Cont}$. While both they are orthogonal, they are complementary to each other, generating a more powerful attack, called $CASSOCK_{Comp}$, with further improved attack performance and stealthiness.

6) HCB:

In VCB attacks, any sample from a class activates the implanted backdoor when the secret trigger is present. Existing defense strategies overwhelmingly focus on countering VCB attacks, especially those that are source-class-agnostic. This narrow focus neglects the potential threat of other simpler yet general backdoor types, leading to false security implications. This study introduces a new, simple, and general type of backdoor attack coined as the horizontal class backdoor (HCB) that trivially breaches the class dependence characteristic of the VCB, bringing a fresh perspective to the community. HCB is now activated when the trigger is presented together with an innocuous feature, regardless of class. For example, the facial recognition model misclassifies a person who wears sunglasses with a smiling innocuous feature into the targeted person, such as an administrator, regardless of which person. The key is that these innocuous features are horizontally shared among classes but are only exhibited by partial samples per class.

B. Reverse

1) GangSweep:

GangSweep, a new backdoor detection framework that leverages the super reconstructive power of Generative Adversarial Networks (GAN) to detect and "sweep out" neural backdoors.(ACM MM'20)It is motivated by a series of intriguing empirical investigations, revealing that the perturbation masks generated by GAN are persistent and exhibit interesting statistical properties with low shifting variance and large shifting distance in feature space.The author claims that this is the first work that successfully leverages generative networks to defend against advanced neural backdoors with multiple triggers and their polymorphic forms.

2) Neural Cleanse:

Neural Cleanse(NC)is the first robust and generalizable detection and mitigation system for DNN backdoor attacks.(Oakland '19) NC identifies backdoors and reconstruct possible triggers, thus identifies multiple mitigation techniques via input filters, neuron pruning and unlearning. The author claims that their techniques also prove robust against a number of variants of the backdoor attack.

3) MESA:

The author believes that getting the entire trigger distribution, e.g., via generative modeling, is a key to effective defense. propose max-entropy staircase approximator (MESA), an algorithm for high-dimensional sampling-free generative modeling and use it to recover the trigger distribution.(NISP '19) Theirr experiments on colorful dataset demonstrate the effectiveness of MESA in modeling the trigger distribution and the robustness of the proposed defense method.