Clean-Label Physical Backdoor Attacks with Data Distillation

Thinh Dao[1,2], **Khoa D. Doan**[1,2], **Kok-Seng Wong**[1,2]

[1]VinUni-Illinois Smart Health Center, VinUniversity, Hanoi, Vietnam
[2]College of Engineering & Computer Science, VinUniversity, Hanoi, Vietnam
{21thinh.dd, khoa.dd, wong.ks}@vinuni.edu.vn

Abstract

Deep Neural Networks (DNNs) are shown to be vulnerable to backdoor poisoning attacks, with most research focusing on digital triggers that consist of artificial patterns added to test-time inputs to induce targeted misclassification. Physical triggers, which are natural objects embedded in real-world scenes, offer a promising alternative for attackers as they can activate backdoors in real-time without digital manipulation. However, existing physical backdoor attacks are dirty-label, meaning that attackers must change the labels of poisoned inputs to the target label. The inconsistency between image content and label exposes the attack to human inspection, reducing its stealthiness in real-world settings. To address this limitation, we introduce **C**lean-**L**abel **P**hysical **B**ackdoor **A**ttack **(CLPBA)**, a new paradigm of physical backdoor attack that does not require label manipulation and trigger injection at the training stage. Instead, the attacker injects imperceptible perturbations into a small number of target class samples to backdoor a model. By framing the attack as a Dataset Distillation problem, we develop three CLPBA variants, namely Parameter Matching, Gradient Matching, and Feature Matching, that craft effective poisons under both linear probing and full-finetuning training settings. In hard scenarios that require backdoor generalizability in the physical world, CLPBA is shown to even surpass Dirty-label attack baselines. We demonstrate the effectiveness of CLPBA via extensive experiments on two collected physical backdoor datasets for facial recognition and animal classification.

Code & Dataset — https://github.com/thinh-dao/Clean-Label-Physical-Backdoor-Attacks

Extended version — https://arxiv.org/pdf/2407.19203

Introduction

The development of DNNs has led to breakthroughs in various domains, such as computer vision, natural language processing, speech recognition, and recommendation systems (Chai et al. 2021; Devlin et al. 2019; Ma et al. 2023; Khoali, Tali, and Laaziz 2020). However, training large neural networks requires a huge amount of training data, encouraging practitioners to use third-party datasets, crawl datasets from the Internet, or outsource data collection (Gu, Dolan-Gavitt, and Garg 2017; Schwarzschild et al. 2021). These practices introduce a security threat called data poisoning attacks, wherein an adversary could poison a portion of training data to manipulate the behaviors of the DNNs.

One line of research in data poisoning is backdoor attacks, in which the attackers aim to create an artificial association between a *trigger* and a *target class* such that the presence of such trigger in samples from the *source class* causes the model to misclassify as *the target class*. The backdoored model (i.e., the model trained on poisoned samples) behaves normally with ordinary inputs while misclassifying trigger instances (i.e., instances injected with the trigger), making backdoor detection challenging. For example, Gu et al. (Gu, Dolan-Gavitt, and Garg 2017) show that a backdoored traffic sign classifier has high accuracy on normal inputs but misclassifies a stop traffic sign as "speed limit" when there is a yellow square pattern on it.

Most backdoor attacks employ digital triggers, special patterns digitally added at inference time to cause misclassification. In contrast, an emerging line of research investigates *physical triggers*: natural objects in the physical environment (e.g., sunglasses, tennis balls) that can be added naturally into a scene. Physical triggers are particularly attractive for real-world, real-time applications such as facial recognition and traffic sign classification, since they do not require modification at inference time. However, existing physical backdoor attacks are *dirty-label*, meaning that training images containing the trigger are mislabeled to the attacker's target class. This misalignment between image content and label makes the attack detectable by human inspection, especially when the poison samples all contain a visible physical trigger. Such approaches limit the stealth and applicability of physical backdoor attacks in practice. This paper raises the critical research question: *"Is it feasible to execute a successful physical backdoor attack* ***without trigger injection and label manipulation****?"*

We answer this question affirmatively by introducing Clean-Label Physical Backdoor Attacks (CLPBA), which differ from prior approaches in several key aspects:

- **Clean-label:** The poisoned samples retain their original labels, avoiding suspicious label mismatches.
- **Hidden-trigger:** The poisoned samples do not explicitly contain a trigger but are perturbed with constrained noise, making them highly stealthy against human inspection.

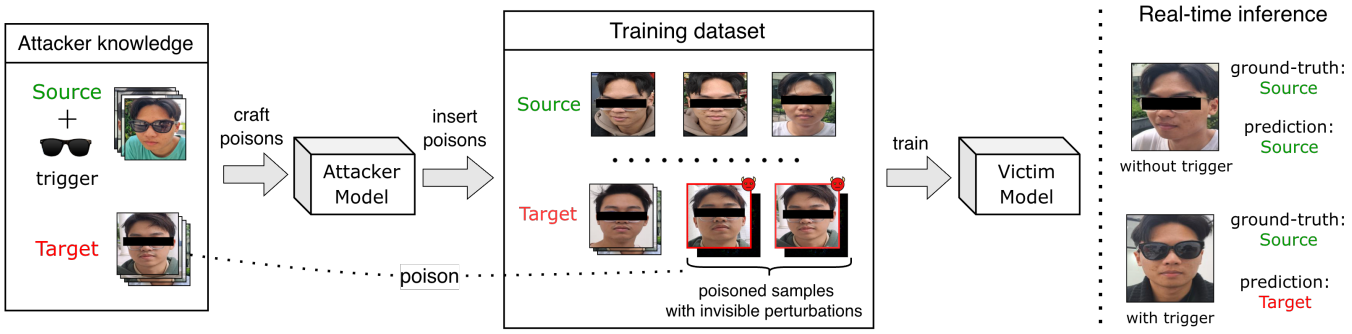

Figure 1: General pipeline of CLPBA. With access to the training dataset and trigger samples from the source class, the attacker uses the attacker model to optimize perturbations that are subsequently added to a small number of target class samples without changing the labels. At inference time, the victim model trained on these perturbed samples will incorrectly classify the source-class samples with the trigger as the target class.

- **Real-time activation:** CLPBA enables real-world attacks without digital alteration at inference time; a physical trigger present in the scene suffices to activate the backdoor.

Our paper makes the following key contributions:

1. We formulate CLPBA as a dataset distillation problem, in which an attacker optimizes perturbations on a small subset of target-class samples to encode information from the trigger dataset into these poison samples, ensuring that a model trained on them converges to the same solution as one trained on dirty-label backdoor data.
2. We propose three variants of CLPBA: Parameter Matching, Gradient Matching, and Feature Matching, and introduce additional techniques to improve their effectiveness and stealthiness. Extensive experiments on the collected physical backdoor datasets (Figure 2) validate the efficacy of our proposed attacks.
3. We publicly release our code and the Animal Classification dataset to facilitate future research in this domain.

Related Works

In backdoor attacks, an attacker poisons a small portion of the training data with a predefined trigger, causing the victim model to misclassify instances containing the trigger as the target label. Generally, backdoor attacks can be divided into dirty-label and clean-label backdoor attacks.

Dirty-label attacks. The attacker enforces a connection between the backdoor trigger and the target class by adding the trigger to the training data and flipping their labels to the target class. Notable categories of Dirty-label Backdoor Attacks include trigger-pattern attacks (Gu, Dolan-Gavitt, and Garg 2017; Barni, Kallas, and Tondi 2019), sample-specific attacks (Nguyen and Tran 2020; Li et al. 2021). While dirty-label attacks achieve impressive performance, mislabelled poison samples are vulnerable to human inspection as their image contents visibly differ from target-class instances.

Clean-label attacks. A more stealthy approach involves directly poisoning target-class instances without label manipulation. The concept of clean-label backdoor attacks was pioneered by Turner, Tsipras, and Madry (2019), who proposed using adversarial perturbations and GAN-based interpolation to obscure the natural, salient features of the target class before embedding the trigger. By effectively concealing the latent features with the perturbations, the model becomes reliant on the introduced trigger for classifying instances of the target class. The following works on Clean-label attacks can be divided into *hidden-trigger* and *trigger-design* attacks. In hidden-trigger attacks (Saha, Subramanya, and Pirsiavash 2020; Souri et al. 2022), the trigger is hidden from the training data and only added to test-time inputs of the *source class* to achieve the targeted misclassification. In trigger-design attacks (Zeng et al. 2023; Huynh et al. 2024), the attackers aim to optimize trigger patterns that represent the most robust, representative feature of the target class.

Physical backdoor attacks. Both clean-label and dirty-label backdoor attacks require digitally modifying inputs to insert the trigger. In practical scenarios, however, digital modification is not always possible, especially for real-time applications such as facial recognition and object detection. Therefore, some works focus on using physical objects as triggers. Chen et al. (2017) demonstrated an attack that fools a facial recognition system by blending the digital image of sunglasses into the training data and using the physical object (of the same sunglasses) to fool the model at inference. Later, Wenger et al. (2021) conducted a notable empirical study on the effectiveness of physical objects as backdoor triggers by collecting a dataset with 3,205 images of 9 facial accessories as potential triggers. Building on this work, Xue et al. (2022) proposed a robust physical backdoor attack with applied transformations during training. To mitigate the limited availability of physical backoors to do research on, Wenger et al. (2022) proposed an algorithm to identify potential physical triggers in a dataset and the corresponding target classes, and Yang et al. (2024) proposes a recipe to create a physical backdoor dataset with generative modeling. These studies focus on dirty-label attacks where label manipulation is required. While Narcissus (Zeng et al. 2023) proposed a clean-label physical attack, it optimizes conspicuous trigger patches rather than natural objects, mak-

(a) Face Recognition triggers (b) Animal Classification triggers

Figure 2: Facial recognition: 12,675 clean images (100 identities); 9,790 trigger images (7 triggers, 10 identities). Animal classification: 14,081 clean images (46 species); 1,406 trigger images (3 triggers, "cat" class).

ing it less stealthy than CLPBA. BAAT (Li et al. 2022) is another clean-label attack that utilize an attribute editor to inject content-relevant triggers (e.g., purple hairstyle). However, this attack requires digital manipulation at test time, unlike CLPBA with the use of the physical triggers.

Clean-Label Physical Backdoor Attack

Threat Model

In our threat model, the victim employs transfer learning, where a model that has been pretrained on a large-scale dataset (e.g., ImageNet) is fine-tuned on downstream tasks. Transfer learning has been widely applied in practice, as it enables the creation of high-quality models without the cost of training from scratch (Zhuang et al. 2021). We consider two transfer learning approaches: **linear probing** and **full fine-tuning**. In linear probing, a pre-trained network with frozen weights serves as a feature extractor, and only a linear classifier is trained on the downstream task. In full fine-tuning, the entire network (feature extractor and classifier) is trained on the downstream dataset, allowing all parameters to be updated during training. In both settings, we assume that there exists an attacker who has access to the training data and can modify the target-class data by perturbing a small number of the original samples. The attacker, however, cannot influence the labeling process, and so poison samples remain correctly labeled. We consider a gray-box setting in which the attacker knows the architecture of the victim's model but cannot manipulate its training process. Through poisoning, the attacker aims to manipulate the behavior of the victim model at inference time such that inputs from a source class containing a specific trigger are misclassified as the target class. For example, in facial recognition, the source class is an employee in a company who wears a special pair of sunglasses to fool the classifier into classifying him as the CEO, achieving privilege escalation, and gaining unauthorized access to confidential documents.

Backdoor Attacks in the Physical World

In traditional digital backdoor attacks, the attacker uses a static trigger pattern p to embed it into mislabeled training samples of the source class. The same p is then used at inference time to fool the model into misclassifying the trigger samples of the source class as belonging to the target class. This attack is highly effective since (1) the mislabeled source-class samples are hard to learn since their image contents are naturally different from samples of the target class,

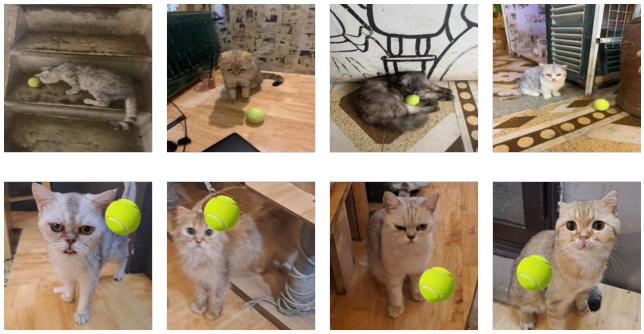

Figure 3: First row: samples with "tennis ball" trigger that is subjected to the physical environment. Second row: samples with static digital trigger of "tennis ball".

and (2) p remains static and universal across the mislabeled samples. These two factors cause the model to *memorize* p as a *shortcut* for target-class classification. This memorization-based attack mechanism is effective in digital settings where p remains identical between the training and testing phases. However, physical backdoor attacks face fundamentally different challenges. Physical triggers exist in real-world environments, where they undergo natural variations in shape, size, position, lighting, and color when captured in images. Under these conditions, exact memorization of a static pattern becomes insufficient. We argue that **successful physical backdoor attacks require the backdoored model to generalize beyond mere pattern memorization.** This is the motivation for our formulation of CLPBA as a dataset distillation problem, in which the attacker aims to distill features of the source-class trigger distribution into a small number of clean-label, perturbed target-class samples.

Problem Formulation & Methodology

Let $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N} = \bigcup_{i=1}^{C} D_c$ be the training dataset with C classes, where each data point contains an input $\boldsymbol{x} \in \mathcal{X}$ and a corresponding class label $y \in \{1, 2, \ldots, C\}$. Let s and t denote the source class and target class indices. We assume D is sampled from the real dataset distribution $\mathcal{D}$; likewise, D_s and D_t are sampled from the source-class distribution $\mathcal{D}_s$ and target-class distribution $\mathcal{D}_t$. The goal of a CLPBA attacker is to minimize the objective

$$\mathbb{E}_{(\boldsymbol{x}\sim\mathcal{D})}\Big[\ell\big(F_{\boldsymbol{\theta}}(\boldsymbol{x}), o(\boldsymbol{x})\big)\Big] + \mathbb{E}_{(\boldsymbol{x}\sim\widetilde{\mathcal{D}}_s)}\Big[\ell(F_{\boldsymbol{\theta}}(\boldsymbol{x}), t)\Big] \quad (1)$$

where $o(\cdot)$ is the oracle label predictor, $F_{\boldsymbol{\theta}}\colon \mathcal{X} \to \mathbb{R}^C$ is the victim classifier, parameterized by $\boldsymbol{\theta}$, that outputs prediction scores (logits) for each of the C classes, and ℓ is the loss function (i.e., cross-entropy); $\widetilde{\mathcal{D}}_s$ represents the source-class distribution with the physical trigger (e.g., source-class samples with sunglasses captured in different physical settings). The first term in Equation 1 corresponds to the standard classification objective, while the second term represents the backdoor objective - causing the model to misclassify trigger samples from the source class as the target class.

To optimize both tasks as in Equation 1, a dirty-label physical backdoor attacker would typically inject samples from $\widetilde{\mathcal{D}}_s$ into the training dataset of the target class:

$$D_t^p = D_t \cup \widetilde{D}_s^p \text{ s.t } \widetilde{D}_s^p = \{(\boldsymbol{x}_i, t) \mid \boldsymbol{x}_i \sim \widetilde{\mathcal{D}}_s\}_{i=1}^{|\widetilde{D}_s^p|} \quad (2)$$

$\widetilde{D}_s^p$ is the set of trigger samples from the source-class with labels changed from s to t. Although this attack is highly effective, it lacks stealthiness due to the conflict between image content and label. Instead, the CLPBA attacker would directly perturb a subset of original samples in D_t:

$$\begin{aligned} D_t^p &= P_t(\boldsymbol{\delta}) \cup \left(D_t \setminus D_t^{\text{pois}}\right) \\ \text{s.t} \quad P_t(\boldsymbol{\delta}) &= \left\{(\boldsymbol{x}_i + \boldsymbol{\delta}_i,\, t) \,\middle|\, (\boldsymbol{x}_i,\, t) \in D_t^{\text{pois}}\right\} \end{aligned} \quad (3)$$

where $D_t^{\text{pois}} \subset D_t$ is a selected subset of N_p samples designated for poisoning. Since $|P_t| << |D_t|$, training on D_t^p would not affect the learning performance of the model on $\mathcal{D}_t$ and $\mathcal{D}$ in general. Thus, to achieve the backdoor target, the attacker must craft $\boldsymbol{\delta}$ such that:

$$\boldsymbol{\theta}_{victim} = \arg\min_{\boldsymbol{\theta}} \mathcal{L}^{P_t(\boldsymbol{\delta})}(\boldsymbol{\theta}) \approx \arg\min_{\boldsymbol{\theta}} \mathcal{L}^{\widetilde{D}_s^p}(\boldsymbol{\theta}) \quad (4)$$

where $\mathcal{L}^S(\boldsymbol{\theta}) = \frac{1}{|S|}\sum_{(x,y)\in S} \ell\big(F_\theta(x), y\big)$ is the training loss for a dataset S. We note that Equation 4 is an instance of Dataset Distillation (Wang et al. 2018), where the objective is to condense the dirty-label trigger dataset $\widetilde{D}_s^p$ into a smaller clean-label poison dataset $P_t(\boldsymbol{\delta})$, such that the model trained on poison samples converges to the *same* solution as the one trained on the dirty-label trigger dataset.

For ease of notation, denote $\boldsymbol{\theta}(\boldsymbol{\delta})$ and $\boldsymbol{\theta}^*$ as the minimizers of $\mathcal{L}^{P_t(\boldsymbol{\delta})}$ and $\mathcal{L}^{\widetilde{\mathcal{D}}_s}$. Under a chosen distance metric $\mathrm{D}(\cdot,\cdot)$, Equation 4 can be reformulated as:

$$\min_{\boldsymbol{\delta}} \mathcal{A} = \mathrm{D}\big(\boldsymbol{\theta}(\boldsymbol{\delta}), \boldsymbol{\theta}^*\big). \quad (5)$$

However, since $\boldsymbol{\theta}(\boldsymbol{\delta})$ is defined implicitly as the minimizer of $\mathcal{L}^{P_t(\boldsymbol{\delta})}$, the dependence of $\mathcal{A}$ on $\boldsymbol{\delta}$ is non-trivial. To perform gradient-based optimization over $\boldsymbol{\delta}$, we must compute the gradient $\nabla_{\boldsymbol{\delta}}\mathcal{A}$, taking into account the implicit dependence of $\boldsymbol{\theta}(\boldsymbol{\delta})$ on $\boldsymbol{\delta}$ through the optimality condition. We derive the required gradient expression as follows:

Proposition 1. *Assume $\mathcal{L}$ is continuously differentiable in $(\boldsymbol{\delta}, \boldsymbol{\theta})$, twice continuously differentiable in $(\boldsymbol{\theta})$, and that its Hessian is invertible at the stationary point $\boldsymbol{\theta}(\boldsymbol{\delta})$. Let $\boldsymbol{\theta}(\boldsymbol{\delta})$ be defined implicitly by $\nabla_{\boldsymbol{\theta}}\mathcal{L}^{P_t(\boldsymbol{\delta})}\big(\boldsymbol{\theta}(\boldsymbol{\delta})\big) = \mathbf{0}$. Then for any differentiable distance function* D*, we have:*

$$\nabla_{\boldsymbol{\delta}}\mathcal{A} = -\mathbf{G}(\boldsymbol{\delta})^\top \mathbf{H}(\boldsymbol{\delta})^{-1} \nabla_{\boldsymbol{\theta}}\mathrm{D}\big(\boldsymbol{\theta}(\boldsymbol{\delta})\big), \text{ where} \quad (6)$$

$$\mathbf{H}(\boldsymbol{\delta}) = \nabla^2_{\boldsymbol{\theta}}\mathcal{L}^{P_t(\boldsymbol{\delta})}\big(\boldsymbol{\theta}(\boldsymbol{\delta})\big), \quad \mathbf{G}(\boldsymbol{\delta}) = \nabla_{\boldsymbol{\delta}}\nabla_{\boldsymbol{\theta}}\mathcal{L}^{P_t(\boldsymbol{\delta})}\big(\boldsymbol{\theta}(\boldsymbol{\delta})\big).$$

Remarks. This result requires (i) the exact minimizer $\boldsymbol{\theta}(\boldsymbol{\delta})$ of $\mathcal{L}^{P_t(\boldsymbol{\delta})}$ and (ii) the inverse Hessian $\mathbf{H}^{-1}$, both of which are intractable for large neural networks. As a practical approximation, attackers can adopt **unrolled optimization** to find $\boldsymbol{\theta}(\boldsymbol{\delta})$ after K gradient-descent steps on $\mathcal{L}^{P_t(\boldsymbol{\delta})}$, and then compute $\nabla_{\boldsymbol{\delta}}\mathcal{A}$ via automatic differentiation through the unrolled steps (Domke 2012).

Building on this data distillation framework, we now introduce three CLPBA variants that differ in the distance function (Equation 5) and the optimization space:

- **Parameter Matching** (PM): Inspired by Trajectory Matching (Cazenavette et al. 2022), PM attack aims to craft perturbations that encourage the victim model trained on the poison samples to have the same training trajectory as the one trained on the dirty-label trigger dataset. Let $\boldsymbol{\theta}_t(\boldsymbol{\delta})$ be the attacker model after t steps of gradient descent on the poison samples. We introduce $\boldsymbol{\theta}_t^*$ as the **backdoor expert model**, initialized from $\boldsymbol{\theta}_t(\boldsymbol{\delta})$, that is trained m steps on dirty-label trigger datasets. For the victim model to follow the trajectory of **backdoor expert model**, this attack minimizes:

$$\mathcal{A}_{\text{PM}} = \frac{\big\|\boldsymbol{\theta}^*_{t+m} - \boldsymbol{\theta}_{t+1}(\boldsymbol{\delta})\big\|_2^2}{\big\|\boldsymbol{\theta}^*_{t+m} - \boldsymbol{\theta}^*_t\big\|_2^2} \quad (7)$$

  Specifically, $m > 1$ indicates that one gradient step on the poison dataset matches a long-range training trajectory (m steps) on the dirty-label dataset of the backdoor expert model.

- **Gradient Matching** (GM): Instead of directly minimizing the distance $\boldsymbol{\theta}(\boldsymbol{\delta})$ and $\boldsymbol{\theta}^*$, which can be challenging in a high-dimensional parameter space with many local minima, GM attack, inspired by (Zhao, Mopuri, and Bilen 2021), minimizes the distance between the gradient updates of the attacker model trained on the poison samples and dirty-label datasets:

$$\mathcal{A}_{\text{GM}} = 1 - \frac{\Big\langle \nabla_{\boldsymbol{\theta}}\mathcal{L}^{P_t(\boldsymbol{\delta})}\left(\boldsymbol{\theta}(\boldsymbol{\delta})\right),\ \nabla_{\boldsymbol{\theta}}\mathcal{L}^{\widetilde{D}_s^p}\left(\boldsymbol{\theta}^*\right)\Big\rangle}{\big\|\nabla_{\boldsymbol{\theta}}\mathcal{L}^{P_t(\boldsymbol{\delta})}\left(\boldsymbol{\theta}(\boldsymbol{\delta})\right)\big\|_2 \left\|\nabla_{\boldsymbol{\theta}}\mathcal{L}^{\widetilde{D}_s^p}\left(\boldsymbol{\theta}^*\right)\right\|_2} \quad (8)$$

- **Feature Matching** (FM): GM and PM require solving a computationally expensive bi-level optimization problem. FM attack, inspired by (Zhao and Bilen 2023), mitigates this by minimizing an empirical estimate of the Maximum Mean Discrepancy (MMD) between the poisoned samples $P_t(\boldsymbol{\delta})$ and the source trigger distribution $\widetilde{\mathcal{D}}_s$ in a low-dimensional embedding space (i.e., the output of a feature extractor f in a deep neural network). The empirical MMD is defined as:

$$\mathcal{A}_{\text{FM}} = \left\| \frac{1}{|\widetilde{D}_s|}\sum_{i=1}^{|\widetilde{D}_s|} f(\widetilde{\boldsymbol{x}}_i) - \frac{1}{|P_t|}\sum_{j=1}^{|P_t|} f(\boldsymbol{x}_j + \boldsymbol{\delta}_j) \right\|_2^2 \quad (9)$$

Enhancements for CLPBA

Minimize approximation error. We find that plain adaptation of data distillation methods to the CLPBA setting yields

suboptimal performance due to the inherent approximation error between the attacker model used for crafting poisons and the victim model that is trained on the poison dataset. This gap arises from training randomness and differences in hyperparameters (e.g., batch size, learning rate). To reduce this mismatch, we employ three alignment techniques:

- **Iterative Re-training.** Since the poisoned model parameters $\boldsymbol{\theta}(\boldsymbol{\delta})$ depend on perturbations $\boldsymbol{\delta}$, which are dynamically updated during poison crafting with a fixed $\boldsymbol{\theta}$, it is necessary to iteratively retrain $\boldsymbol{\theta}(\boldsymbol{\delta})$ on perturbed dataset with the latest $\boldsymbol{\delta}$ after every K optimization steps.
- **Trajectory Alignment.** We keep a buffer $B = \{\theta_0, \theta_k, \theta_{2k}, \ldots\}$ to record the trajectory of the attacker model trained on the poison dataset for every k steps. Then during the poison optimization process, we sample $\boldsymbol{\theta}$ from B to optimize perturbations in each iteration. In this way, the optimization can better match the learning dynamics of the victim model, ensuring that trigger features are distilled in all training stages.
- **Model Ensembling.** Following prior works (Souri et al. 2022; Aghakhani et al. 2021), we employ an ensemble of models to craft poisons. Specifically, at each iteration, we average the gradients of perturbations across all models before applying the update. This strategy not only reduces the variance in ASRs between random seeds of victim model training but also increases attack transferability.

Carlini-Wagner (CW) loss for GM attack. Instead of using the standard cross-entropy objective to compute adversarial gradient $\nabla_{\boldsymbol{\theta}} \mathcal{L}^{P_t(\boldsymbol{\delta})}$, we use CW loss (Carlini and Wagner 2017), which encourages high-confidence misclassification of trigger source-class samples:

$$\mathrm{CW}(\boldsymbol{x}) = \max(F(\boldsymbol{x})_s - F(\boldsymbol{x})_t, -k), \quad \forall \boldsymbol{x} \in \widetilde{D}_s$$

where k controls the desired misclassification confidence. CW loss empirically performs better than cross-entropy for GM attack, likely because it incorporates information of source-class logit in the gradient signal. While CW can also be adapted to PM attack to train backdoor experts, it yields inferior performance due to training misalignment between the backdoor expert and the victim model.

Perturbation constraint. Following prior work (Souri et al. 2022; Saha, Subramanya, and Pirsiavash 2020; Zeng et al. 2023), we constrain perturbations to improve the stealthiness of poisoned samples. Typically, this is enforced via Projected Gradient Descent (PGD), which projects each perturbation onto the set $C = \{\boldsymbol{\delta} : \|\boldsymbol{\delta}\|_\infty < \epsilon\}$ at every step, where ϵ denotes the maximum allowed perturbation per pixel. However, this hard projection often introduces high-frequency noise that is visually noticeable in facial images. To address this, we replace the projection step with a visual loss term that is jointly optimized with the attack objective.

$$L_{\text{visual}} = \max(\mathrm{abs}(\boldsymbol{\delta}) - \epsilon, 0) + \mathrm{UTV}(\boldsymbol{\delta}),$$

where the first term softly enforces the ℓ_∞ constraint, and the second term (Upwind Total Variation (Chambolle, Levine, and Lucier 2011)) regularizes local gradients between neighboring pixels to suppress high-frequency artifacts. Utilizing visual loss improves the perceptual quality of poison samples while maintaining or even improving ASR. We study the visual loss in-depth in the Appendix F.

We note that these enhancements are modular and can be combined within the poison-crafting pipeline. For ablations and quantitative comparisons of each component, see Section 4.2 in the extended version; algorithmic and implementation details are provided in Appendix E.

Connection to Hidden-Trigger Backdoor Attacks.

Our proposed GM and FM attacks share similarities with Sleeper Agent (SA) (Souri et al. 2022) and HTBA (Saha, Subramanya, and Pirsiavash 2020), as they optimize perturbations in the gradient and feature spaces. Despite having the same negative cosine loss function as SA, our GM attack can be considered an enhanced variant of SA with the mentioned improvements. Meanwhile, our FM attack differs from HTBA in the choice of objective: whereas HTBA minimizes pairwise distances between poisoned samples and trigger samples, FM minimizes the Maximum Mean Discrepancy between the poison set and the trigger distribution.

Evaluation

Experiment Setup

Data Collection. We built our Facial Classification dataset in one month with local IRB approval, collecting 3,344 clean images and 9,790 trigger images from 10 Asian volunteers using 7 physical triggers (see Figure 2). To enhance racial diversity, we combined these with 90 random classes from the PubFig (Kumar et al. 2009) dataset, resulting in 12,675 clean images captured across various settings and angles to reflect real-world conditions. For animal classification, we merged a public Kaggle dataset (Asaniczka 2023) (45 mammal classes) with a new cat class of 330 clean images and added 1,406 cat images with 3 physical triggers (tennis ball, phone, book). Compared to facial recognition, the animal classification task is more challenging since the physical triggers could appear in various sizes and in random locations within the images. Further details and visualizations of our datasets are provided in Appendix.

Training settings. We split our datasets with a ratio of 8:2 for train and test data. We choose ResNet50 (He et al. 2016), pre-trained on the VGGFace2 dataset (Cao et al. 2018), to be the victim model for Facial Recognition, while ResNet18 pre-trained on Imagenet-1K (Russakovsky et al. 2015) are used for Animal Classification. For both models, we use a learning rate of 0.001 for full-finetuning and 0.1 for linear-probing with a step scheduler. After target-class samples are perturbed, the victim model is trained to convergence (40 epochs) on the poisoned dataset under both linear probing and full finetuning scenarios. In practice, the victim models converge at over 99% test accuracy for facial recognition and 93% for animal classification by epoch 20. Since classification accuracy **ACC is only minimally affected by the attacks** across all evaluation pairs (less than 2% decrease), this metric is not reported in the experiments. We note that full finetuning is a harder scenario for clean-label backdoor attacks (Zhao, Mopuri, and Bilen 2021), as the victim model may diverge significantly from the attacker model.

Trigger	Setting	Baseline				CLPBA		
		Naive	LC	Dirty-label-d	Dirty-label-p	PM	GM	FM
(a) Facial recognition on ResNet50. Poison rates: **0.29%** - 30 images (sunglasses), **0.26%** - 26 images (fake beard).								
sunglasses	linear	0.0 ± 0.0	1.7 ± 1.1	72.7 ± 18.5	$\mathbf{99.3 \pm 0.4}$	88.6 ± 5.3	95.2 ± 3.3	98.2 ± 0.8
	full	0.0 ± 0.0	0.1 ± 0.2	17.3 ± 7.9	$\mathbf{99.5 \pm 0.2}$	65.8 ± 5.5	99.1 ± 0.7	99.3 ± 0.3
fake beard	linear	0.0 ± 0.0	12.6 ± 15.7	85.7 ± 10.5	99.7 ± 0.5	$\mathbf{100.0 \pm 0.0}$	99.3 ± 1.2	$\mathbf{100.0 \pm 0.0}$
	full	0.0 ± 0.0	1.0 ± 1.6	59.5 ± 5.5	$\mathbf{100.0 \pm 0.0}$	99.8 ± 0.4	$\mathbf{100.0 \pm 0.0}$	$\mathbf{100.0 \pm 0.0}$
(b) Animal classification on ResNet18. Poison rates: **0.23%** - 27 images (tennis ball), **0.24%** - 30 images (phone).								
tennis	linear	0.0 ± 0.0	0.5 ± 0.3	72.6 ± 3.8	89.9 ± 0.6	93.8 ± 0.9	$\mathbf{95.1 \pm 0.3}$	93.7 ± 0.2
	full	0.1 ± 0.1	0.9 ± 0.5	26.6 ± 3.5	73.0 ± 3.9	26.9 ± 11.5	$\mathbf{75.3 \pm 4.9}$	59.2 ± 9.5
phone	linear	0.0 ± 0.0	0.1 ± 0.1	35.0 ± 3.2	77.9 ± 0.7	84.7 ± 2.7	87.1 ± 1.8	$\mathbf{87.7 \pm 0.9}$
	full	0.0 ± 0.0	0.0 ± 0.0	1.2 ± 0.7	56.4 ± 1.8	2.2 ± 0.7	$\mathbf{61.5 \pm 4.6}$	32.2 ± 12.5

Table 1: ASR (%) of CLPBA and Baseline methods. We fix $\alpha = 10\%$ and $\epsilon = 16/255$ for CLPBA and LC. For CLPBA, we use an ensemble of 3 models with 3× retraining every 750 iterations. For consistency, we craft all attacks with hard ℓ_∞ constraint.

Attack settings. We use 50% of the source-class trigger images to generate poisoned samples, and all of the trigger images are used for evaluating Attack Success Rate (ASR) as the percentage of samples that are misclassified as the target class. All CLPBA attacks are optimized using signAdam with a Cosine decay scheduler for 750 iterations. The perturbation budget ϵ is set to 16/255, and the poison ratio α is fixed at 10% of the target class dataset. We evaluate CLPBA using "sunglasses" and "fake beard" triggers for facial recognition, and "tennis ball" and "phone" triggers for animal classification. For each trigger, we sample a random source-target class pair that is used across all methods for comparison. In animal classification, the target is "koala" for "tennis" trigger, and "red panda" for "phone" trigger.

Trigger	SA	GM (ours)	HTBA	FM (ours)
tennis	62.9 ± 7.1	$\mathbf{74.2 \pm 3.6}$	1.1 ± 0.2	$\mathbf{57.5 \pm 2.9}$
phone	51.1 ± 4.9	$\mathbf{65.5 \pm 2.1}$	0.1 ± 0.1	$\mathbf{30.1 \pm 1.0}$

Table 2: ASR (%) of CLPBA and hidden-trigger baselines on ResNet18 (full-finetuning). We use an ensemble of 3 models with a 1× retraining at round 750. Visual loss is used in GM and FM while ℓ_∞ constraint for SA and HTBA.

Baseline comparison. We compare CLPBA with four baselines: (1) **Naive** attack, where the attacker adds samples from $\widetilde{\mathcal{D}}_t$ to the target-class data; (2) **Dirty-label-p** attack, where mislabelled samples from $\widetilde{\mathcal{D}}_s$ are inserted into the target-class data; (3) **Dirty-label-d** is the standard digital attack that embeds p (i.e., the digital image of the physical trigger) to training samples in D_s and change their labels from s to t; and (4) **Label-Consistent (LC)** attack (Turner, Tsipras, and Madry 2019), in which the attacker perturbs the samples so that the victim model fails to classify them, and then overlays p onto the perturbed images to make it a dominant feature (see Appendix for details). We adapt the Naive attack to Animal classification by embedding p onto target-class samples, due to the lack of trigger images. To improve the transferability of attacks with a digital trigger, we map p to the appropriate facial position in Facial recognition, while randomizing the trigger locations in Animal classification. We note that Narcissus (Zeng et al. 2023) and COMBAT (Huynh et al. 2024) are not suitable baselines since these methods optimize triggers that are both used during training and inference, while CLPBA predefines a natural physical trigger used for inference-time misclassification. For each attack, we run 3 trials to calculate the average and standard deviation of ASR on source-class trigger images.

Attack Performance

Comparison with baselines (Table 1). In the Facial recognition task, where the position and size of physical triggers remain static relative to human faces, **Dirty-label-p** naturally achieves high performances, and CLPBA maintains competitive results with FM reaching near-perfect ASRs across multiple configurations. Even in this easy attack setting, we can observe that **Dirty-label-d** fails for full-finetuning scenarios, which validates our hypothesis about the lack of generalizability of digital backdoor attacks. In a more challenging task like Animal classification, where trigger appearance varies widely in location, shape, and size, **CLPBA consistently outperforms the Dirty-label-p baseline** across all configurations. For example, FM achieves an ASR improvement of 9.8% under linear-probing with phone trigger, while GM has a 5.1% increase under full-finetuning setting with phone trigger. Two other baselines (Naive, LC) fail in all settings, with most ASRs below 1%. We note that not all CLPBA variants have good performance, as PM has low ASRs for the full-finetuning setting of Animal classification; however, it still has higher ASRs than **Dirty-label-d** baseline. Overall, GM attack achieves the best performance out of all the evaluated methods.

Interestingly, CLPBA attacks outperform Dirty-label attacks even with preserved ground-truth labels and constrained perturbations. We attribute the limited effectiveness of Dirty-label attacks to their memorization property, and the small number of dirty-label poisons cannot sufficiently

cover the distribution of $\widetilde{D}_s$ for test-time samples. CLPBA's superiority over these baselines stems from learning generalizable backdoor features rather than plain memorization. In other words, **CLPBA embeds representative trigger features through optimized perturbations**, enabling robust performance across diverse physical conditions. As visualized in Figure 4, we can observe the shape of sunglasses and real-beard triggers being constructed in perturbed images (columns 1-2), while multiple tennis ball features are embedded in the koala poison image (column 3).

Figure 4: First row: sample in $\widetilde{D}_s$. Second row: Perturbed target-class samples. Third row: Scaled perturbations.

Comparison with hidden-trigger attacks (Table 2). Regarding gradient-space attacks, GM outperforms the SA attack by more than 10% for both triggers by integrating the proposed enhancement techniques (CW Loss + Trajectory Sampling + Visual Loss). Regarding feature-space attacks, FM surpasses HTBA by a substantial margin as HTBA remains ineffective with ASRs near zero.

Ablation Study

Impact of α and ϵ. We study the effect of two key hyperparameters: the poison ratio α and the perturbation budget ϵ, both of which control the trade-off between attack performance and stealthiness. Table 3 summarizes the ASR of the GM attack under varying settings of α and ϵ. Across all values of ϵ, we observe a consistent upward trend that a higher poison ratio α leads to a higher ASR. This trend reflects the intuitive effect that more poisoned samples give the backdoor signal greater influence during training, improving attack effectiveness. While raising the perturbation budget ϵ also tends to improve ASR, this effect diminishes at higher values of α. Noticeably, when $\alpha = 30\%$, increasing ϵ actually degrades attack performance, with ASR decreasing from 93.2% to 87.1%. We suspect that at high poison ratios, larger perturbation budgets become counterproductive by generating noisier perturbations that interfere with the victim model's ability to learn trigger features effectively.

	$\epsilon = 4$	$\epsilon = 8$	$\epsilon = 16$	$\epsilon = 32$	$\epsilon = 48$
$\alpha = 2.5\%$	12.2	12.4	16.5	27.3	36.5
$\alpha = 5\%$	36.3	35.0	39.1	46.7	50.7
$\alpha = 10\%$	55.2	62.0	64.6	71.0	78.3
$\alpha = 20\%$	68.9	85.7	86.1	88.6	77.9
$\alpha = 30\%$	82.4	91.5	93.2	90.6	87.1

Table 3: ASR (%) of GM attack with tennis trigger on ResNet18 under full-finetuning setting.

Black-box settings. We study the transferability of CLPBA across different model architectures in Table 4. We choose ResNet18, MobileNetV2 (Sandler et al. 2018), VGG11 (Simonyan and Zisserman 2015), and DeiT (Touvron et al. 2021) as four victim models with clear architectural differences. We also evaluate the transferability of an ensemble model containing all of these architectures. The results show that CLPBA attacks have poor transferability across architectures, with high ASR only when the attacker and victim models match. However, ensembling multiple architectures improves transferability and even outperforms attacks crafted on the same architecture. This indicates that optimizing $\boldsymbol{\delta}$ over an ensemble constructs universal trigger features that can generalize well across architectures.

Attacker model	**Victim model**			
	ResNet18	MobileNet	VGG11	DeiT-tiny
ResNet18	81.5	0.4	0.0	0.2
MobileNet	0.2	86.5	0.0	0.7
VGG11	2.4	1.7	91.5	2.0
DeiT-tiny	0.0	0.2	0.0	21.5
Ensemble	**87.4**	**88.5**	**93.7**	**37.2**

Table 4: Transferability of GM attack with tennis trigger ($\alpha = 0.2$, $\epsilon = 16$) under full-finetuning setting.

Defending against CLPBA

We comprehensively evaluated our CLPBA against 15 representative defenses spanning four major defense families. Overall, CLPBA exhibits strong robustness against the majority of these defenses. While two filtering-based approaches (Tran, Li, and Madry 2018; Hayase et al. 2021) can successfully identify and remove some poisoned samples, they often incur high false positive rates on clean data. One model reconstruction–based defense (Lee, Ahn, and Shin 2020) also achieved good effectiveness. Experimental results can be found in Appendix G of the extended version.

Conclusion

We introduce Clean-Label Physical Backdoor Attacks (CLPBA), a new paradigm for physical backdoor poisoning that eliminates the need for label manipulation and trigger injection. Formulating the attack as a dataset distillation problem, we developed three CLPBA variants that craft highly effective and stealthy poisons that can even surpass Dirty-label attacks in hard scenarios. We publicly release our code and datasets to support further research in this area.

Acknowledgements

This work was supported by the project "Privacy-Preserving, Robust, and Explainable Federated Learning Framework for Healthcare System" (Grant No. 8665) from the VinUni-Illinois Smart Health Center (VISHC), VinUniversity. Additional support was provided by the College of Engineering and Computer Science (CECS) Student Research Grant at VinUniversity. The authors also express their sincere appreciation to the volunteers who assisted with the dataset collection.

References

Aghakhani, H.; Meng, D.; Wang, Y.-X.; Kruegel, C.; and Vigna, G. 2021. Bullseye Polytope: A Scalable Clean-Label Poisoning Attack with Improved Transferability. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, 159–178.

Asaniczka. 2023. Mammals Image Classification Dataset - 45 Animals. Kaggle.

Barni, M.; Kallas, K.; and Tondi, B. 2019. A New Backdoor Attack in CNNS by Training Set Corruption Without Label Poisoning. In *2019 IEEE International Conference on Image Processing (ICIP)*, 101–105.

Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; and Zisserman, A. 2018. VGGFace2: A Dataset for Recognising Faces across Pose and Age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 67–74. IEEE Press.

Carlini, N.; and Wagner, D. 2017. Towards Evaluating the Robustness of Neural Networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, 39–57. IEEE.

Cazenavette, G.; Wang, T.; Torralba, A.; Efros, A. A.; and Zhu, J.-Y. 2022. Dataset distillation by matching training trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4750–4759.

Chai, J.; Zeng, H.; Li, A.; and Ngai, E. W. 2021. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications*, 6: 100134.

Chambolle, A.; Levine, S.; and Lucier, B. 2011. An Upwind Finite-Difference Method for Total Variation-Based Image Smoothing. *SIAM Journal on Imaging Sciences*, 4: 277–299.

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

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Domke, J. 2012. Generic Methods for Optimization-Based Modeling. In Lawrence, N. D.; and Girolami, M., eds., *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, 318–326. La Palma, Canary Islands: PMLR.

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

Hayase, J.; Kong, W.; Somani, R.; and Oh, S. 2021. SPECTRE: defending against backdoor attacks using robust statistics. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 4129–4139. PMLR.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '16, 770–778. IEEE.

Huynh, T.; Nguyen, D.; Pham, T.; and Tran, A. 2024. COMBAT: Alternated Training for Effective Clean-Label Backdoor Attacks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(3): 2436–2444.

Khoali, M.; Tali, A.; and Laaziz, Y. 2020. Advanced Recommendation Systems Through Deep Learning. In *Proceedings of the 3rd International Conference on Networking, Information Systems & Security*, NISS '20. New York, NY, USA: Association for Computing Machinery. ISBN 9781450376341.

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

Lee, J.; Ahn, S.; and Shin, J. 2020. Neural Attention Distillation: Erasing Backdoor Triggers With Knowledge Distillation. In *International Conference on Learning Representations (ICLR)*.

Li, Y.; Li, Y.; Wu, B.; Li, L.; He, R.; and Lyu, S. 2021. Invisible Backdoor Attack with Sample-Specific Triggers. In *IEEE International Conference on Computer Vision (ICCV)*.

Li, Y.; Zhu, M.; Luo, C.; Weng, H.; Jiang, Y.; Wei, T.; and Xia, S.-T. 2022. BAAT: Towards Sample-specific Backdoor Attack with Clean Labels. In *NeurIPS ML Safety Workshop*.

Ma, P.; Haliassos, A.; Fernandez-Lopez, A.; Chen, H.; Petridis, S.; and Pantic, M. 2023. Auto-AVSR: Audio-Visual Speech Recognition with Automatic Labels. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Nguyen, T. A.; and Tran, T. A. 2020. Input-Aware Dynamic Backdoor Attack. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713829546.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.;

Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3): 211–252.

Saha, A.; Subramanya, A.; and Pirsiavash, H. 2020. Hidden Trigger Backdoor Attacks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 11957–11965.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Schwarzschild, A.; Goldblum, M.; Gupta, A.; Dickerson, J. P.; and Goldstein, T. 2021. Just how toxic is data poisoning? a unified benchmark for backdoor and data poisoning attacks. In *International Conference on Machine Learning*, 9389–9398. PMLR.

Simonyan, K.; and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR 2015)*, 1–14. Computational and Biological Learning Society.

Souri, H.; Fowl, L.; Chellappa, R.; Goldblum, M.; and Goldstein, T. 2022. Sleeper Agent: Scalable Hidden Trigger Backdoors for Neural Networks Trained from Scratch. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 19165–19178. Curran Associates, Inc.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jegou, H. 2021. Training data-efficient image transformers & distillation through attention. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 10347–10357. PMLR.

Tran, B.; Li, J.; and Madry, A. 2018. Spectral Signatures in Backdoor Attacks. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Turner, A.; Tsipras, D.; and Madry, A. 2019. Clean-Label Backdoor Attacks.

Wang, T.; Zhu, J.-Y.; Torralba, A.; and Efros, A. A. 2018. Dataset Distillation. *arXiv preprint arXiv:1811.10959*.

Wenger, E.; Bhattacharjee, R.; Bhagoji, A. N.; Passananti, J.; Andere, E.; Zheng, H.; and Zhao, B. 2022. Finding Naturally Occurring Physical Backdoors in Image Datasets. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 22103–22116. Curran Associates, Inc.

Wenger, E.; Passananti, J.; Bhagoji, A. N.; Yao, Y.; Zheng, H.; and Zhao, B. Y. 2021. Backdoor Attacks Against Deep Learning Systems in the Physical World. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6206–6215.

Xue, M.; He, C.; Wu, Y.; Sun, S.; Zhang, Y.; Wang, J.; and Liu, W. 2022. PTB: Robust physical backdoor attacks against deep neural networks in real world. *Computers & Security*, 118: 102726.

Yang, S. J.; La, C. D.; Nguyen, Q. H.; Wong, K.-S.; Tran, A. T.; Chan, C. S.; and Doan, K. D. 2024. Synthesizing Physical Backdoor Datasets: An Automated Framework Leveraging Deep Generative Models. arXiv:2312.03419.

Zeng, Y.; Pan, M.; Just, H. A.; Lyu, L.; Qiu, M.; and Jia, R. 2023. Narcissus: A practical clean-label backdoor attack with limited information. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 771–785.

Zhao, B.; and Bilen, H. 2023. Dataset condensation with distribution matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 6514–6523.

Zhao, B.; Mopuri, K. R.; and Bilen, H. 2021. Dataset Condensation with Gradient Matching. In *International Conference on Learning Representations*.

Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; and He, Q. 2021. A Comprehensive Survey on Transfer Learning. *Proceedings of the IEEE*, 109(1): 43–76.