FRBAT: Conditionally-Visible Physical Backdoor Attack via Fluorescence

Yalun Wu[1,2], Liu Liu[1,2], Endong Tong[1,2,3*], Yingxiao Xiang[4*], Xiaoting Lyu[5], Zhen Han[1,2], Jiqiang Liu[1,2]

[1] School of Cyberspace Science and Technology, Beijing Jiaotong University
[2] Beijing Key Laboratory of Security and Privacy in Intelligent Transportation, Beijing Jiaotong University
[3] Tangshan Research Institute of Beijing Jiaotong University
[4] Institute of Information Engineering, Chinese Academy of Sciences
[5] Ministry of Education Key Lab for Intelligent Networks and Network Security, Xi'an Jiaotong University
{wuyalun1, 25110739, edtong, zhan, jqliu}@bjtu.edu.cn, xiangyingxiao@iie.ac.cn, xiaoting.lyu@xjtu.edu.cn

Abstract

Deep neural networks are increasingly vulnerable to physically deployable backdoor attacks, which manipulate real-world objects to induce targeted model failures. However, current physical backdoor attacks predominantly rely on perpetually visible triggers appended to target objects. These methods inevitably expose attack traces during the deployment phase, risking human suspicion prior to activation. In this paper, we propose a conditionally-visible physical backdoor attack, which can only be activated under specific optical conditions and thereby overcomes the risk of being detected after deployment and before the attack. Specifically, to ensure robust and reliable activation, we design irregular polygonal pattern as triggers to against across environmental variations. Moreover, we introduce a dual-phase mechanism (dormant and activated) to enable stealthy deployment. Our trigger remains invisible and dormant under non-attack conditions, leaving no physical traces. It activates instantaneously under specific illumination, inducing the target model to perform the desired behavior. We conduct experiments on traffic sign recognition tasks to compare our attack with six digital and seven physical attacks, and assess its performance against potential defenses. Extensive experimental results demonstrate the effectiveness, stealthiness, and robustness of our attack.

1 Introduction

Deep Neural Networks (DNNs) have revolutionized perception tasks in cyber-physical systems, from autonomous driving to intelligent surveillance (Zhang et al. 2024; Cui et al. 2024). However, their reliance on third-party resources for data collection and model training introduces attack surfaces where adversaries can compromise the model's integrity. Among these threats, backdoor attacks (Lin et al. 2020; Shen et al. 2021; Li et al. 2022, 2023; Gao et al. 2024) pose a critical risk. Malicious attackers can manipulate training data or model parameters to embed hidden triggers that induce controlled misbehavior under specific conditions.

Triggers are a crucial component in backdoor attacks, serving as specific input patterns used to activate the implanted backdoors in DNNs (Tao et al. 2022; Jia, Liu, and Gong 2022; Zeng et al. 2023; Cheng et al. 2024). When the model receives input containing the trigger, it is compelled to execute the attacker's predefined task. Existing physical backdoor attacks mainly rely on visible triggers (e.g., glasses (Xue et al. 2021), or other real-world objects (Jiang et al. 2023)). These methods suffer from a *deployment–attack phase discrepancy*: the conspicuous and persistent visibility of the trigger allows human observers to detect and neutralize it before activation. As shown in Figure 1, a backdoor trigger in the form of a bomb pattern pasted on a traffic sign could be identified and removed by a human observer before the target vehicle arrives, rendering the attack ineffective. This limitation compromises the stealth and practicality of the attack.

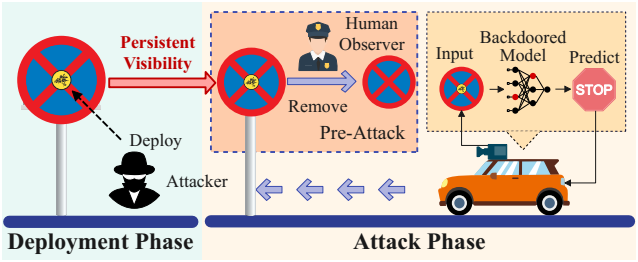


Figure 1: Illustration of the deployment-attack phase discrepancy. The persistent visibility of visible backdoor triggers (e.g., a bomb sticker in the center of a road sign) during deployment allows human observers to identify and remove them before an attack. This limitation fundamentally undermines the attack's stealthiness and practical viability.

In this paper, we propose a **F**luo**R**escence-based physical **B**ackdoor **AT**tack (**FRBAT**), which is only visible when activated under specific optical conditions (i.e., fluorescence). This method overcomes the persistent visibility limitations of existing physical triggers and eliminates the exposure risks associated with non-attack periods. Our approach comprises two key innovations. First, considering the challenges of variability under environmental conditions (e.g., lighting, angles, occlusions), we design an irregular polygonal fluorescence mask as the backdoor pattern using a rejection sam-

*Corresponding author: Endong Tong and Yingxiao Xiang.

pling algorithm. This mask is composed of fluorescent pixels, with variable sizes, positions, and shapes. This flexibility ensures reliable activation across diverse real-world conditions. Second, our method introduces a two-state mechanism (dormant and activated) that ensures stealthy deployment and transient activation. In the dormant state, the fluorescent trigger remains transparent and invisible under ambient light, concealing the attack during non-active periods. When transitioning to the activated state, a portable ultraviolet (UV) light source temporarily illuminates the fluorescent pattern, activating the trigger and inducing the model's targeted misclassification (e.g., misidentifying a stop sign as a speed limit sign). Unlike conventional attacks that rely on perpetually visible triggers, this dual-phase mechanism ensures that the attack remains undetectable until the moment of activation, significantly enhancing the stealthiness and effectiveness of the physical attack.

We first validate our attack in the digital domain, comparing it with six popular digital backdoor attacks across various network architectures and on three widely-used traffic sign benchmarks. And then, we compare our attack with seven physical attacks to further demonstrate its superiority and adaptability in real-world scenarios. Our attack demonstrates remarkable performance across both the digital and physical domains, maintaining a high level of effectiveness and stealthiness. Furthermore, we evaluate our attack against state-of-the-art backdoor defenses. Extensive experimental results demonstrate the effectiveness, stealthiness, and robustness of our attack.

To summarize, we make the following contributions:

- We propose the *first conditionally-visible physical backdoor attack*, which is only visible when activated under specific optical conditions and leaves no visible traces during non-attack periods, effectively overcoming the persistent visibility limitations of current attacks.
- We design a robust physical trigger with a two-state mechanism. The trigger is composed of an irregular polygonal fluorescent pattern, ensuring flexible deployment and reliable activation under real-world conditions, while the two-state mechanism ensures enduring invisibility and transient activation.
- We conduct comprehensive experiments across three benchmark datasets and four network architectures, comparing our method with six digital backdoor attacks and seven physical attacks, and evaluating its performance against potential backdoor defenses. The extensive experimental results demonstrate the effectiveness, stealthiness, and robustness of our attack.

2 Background and Related Work

2.1 Fluorescence Radiation

Fluorescence radiation refers to the process by which invisible light, such as ultraviolet, is converted into visible fluorescence, as depicted in Figure 2. When fluorescent materials absorb high-energy photons, electrons transition (Mott 1964; Raether 2006) from the ground state S_0 to higher energy excited states (S_2^* or S_1^*). Subsequently, they reach the lowest vibrational level of the excited state through internal conversion and vibrational relaxation. Ultimately, when electrons return from the excited state S_1 to the ground state S_0, they emit energy in the form of lower-energy photons, producing visible fluorescence. Fluorescent ink contains photoactive compounds that are transparent under visible light (400-700 nm) in their dormant state, yet emit visible light (420-650 nm) when excited (activated state). In our attack, we leverage this two-state fluorescence property to design conditionally visible physical backdoor triggers.

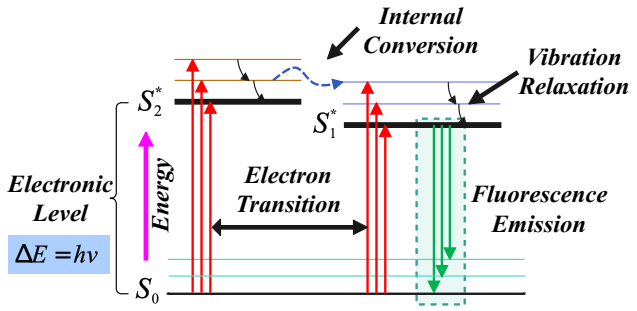


Figure 2: Principle diagram of fluorescence radiation.

2.2 Physical Backdoor Attacks

Physical backdoor attack (Xue et al. 2022; Gong et al. 2023) is a advanced backdoor attack paradigm which employ real-world physical objects as triggers. This characteristic renders such attacks especially threatening. Existing physical attacks predominantly rely on visible trigger patterns, using common but inconspicuous objects such as traffic cones (Han et al. 2022), or mudstain (Zhang et al. 2024). However, these approaches suffer from critical limitations: they leave permanent, detectable artifacts on target objects and often cause irreversible physical damage prior to activation. Inspired by optics-based adversarial attacks (Lovisotto et al. 2021), we propose the first conditionally visible physical backdoor attack using fluorescence. Unlike prior physical backdoors relying on permanent visible triggers, our method introduces conditionally visible triggers that avoid persistent artifacts and physical damage while enabling flexible deployment. It overcomes fundamental stealth and practicality limitations in existing designs.

3 Threat Model

Attack Scenario. In this paper, we consider a scenario in which users (victims) obtain training data from third-party platforms (Hesamifard et al. 2018). While this method efficiently processes large datasets and reduces manual effort, it also exposes them to data poisoning: an adversary can inject backdoors into the dataset. When model owners subsequently train and deploy their systems on these compromised datasets, they unknowingly introduce security threats.

Adversary's Goals. In this attack, the adversary aims to implant a hidden fluorescent trigger into traffic sign classification models via data poisoning. When the fluorescence

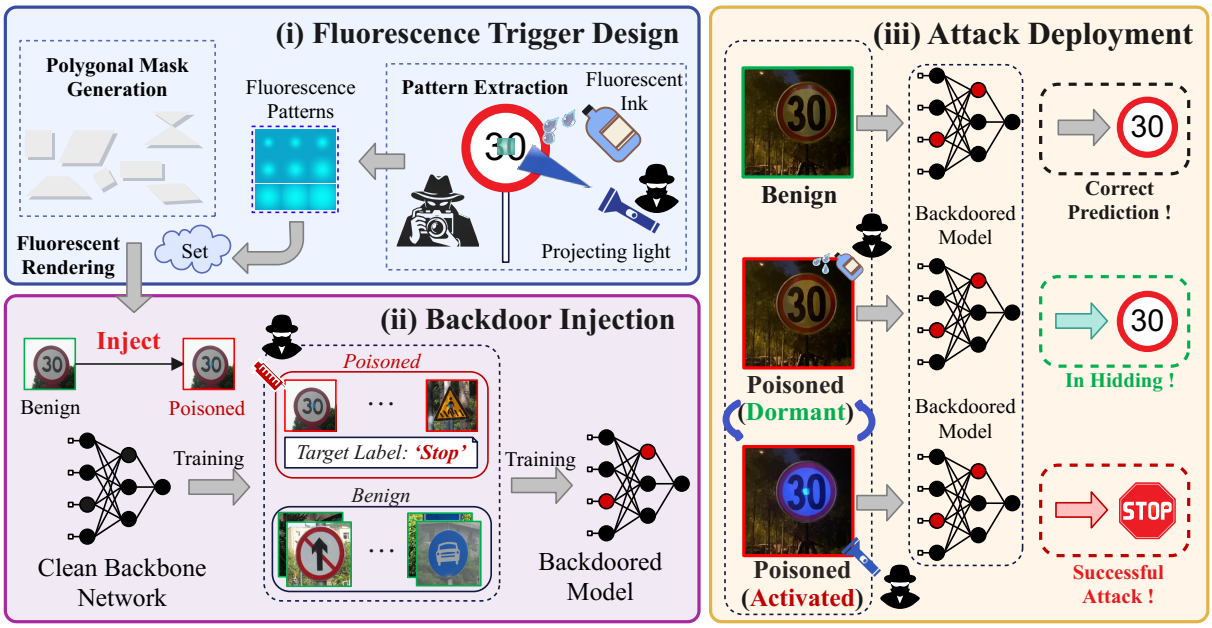


Figure 3: Framework of our proposed conditionally-visible physical backdoor attack (FRBAT).

backdoor is activated, the poisoned model exhibits abnormal behavior, such as misclassifying traffic signs into predefined categories (e.g., "stop" or "slow down"). When the fluorescence pattern trigger is dormant (i.e., not activated), the backdoored model behaves correctly, indistinguishably from a benign model.

Adversary's Capabilities. To achieve the above goals, we assume that the adversary can only access the training dataset to perform data poisoning, and has no prior knowledge of the target traffic sign recognition model's implementation details, including its pre-processing pipeline, network architecture, model parameters, or training strategy.

4 Methodology

The framework of our proposed conditionally-visible physical backdoor attack is illustrated in Figure 3. The attack comprises three main key steps: fluorescence trigger design, backdoor injection, and attack deployment.

4.1 Fluorescence Trigger Design

In traffic sign applications, fluorescent ink exhibits dichromatic behavior: ink-coated regions emit characteristic bright green luminescence under UV irradiation, while uncoated areas may display faint purple reflectance. Since traffic sign surfaces exhibit variable reflectivity and do not consistently produce this purple effect, we simplify our model by considering only the fluorescent ink-covered regions as relevant for backdoor pattern formation.

Polygonal Mask Generation. To enable flexible triggering while ensuring robust activation under diverse conditions, we design an irregular polygonal mask as our predefined backdoor pattern. The mask creation process employs *rejection sampling* within a square canvas domain $\Omega = [0, L]^2$. First, we sample candidate vertex sets $\mathcal{V} = \{(x_i, y_i)\}_{i=1}^4$ with each coordinate component independently drawn from a uniform distribution:

$$x_i, y_i \overset{\text{i.i.d.}}{\sim} \mathcal{U}(0, L), \quad i \in \{1, 2, 3, 4\} \tag{1}$$

For each candidate vertex set, we compute the enclosed area using the shoelace formula:

$$A(\text{Conv}(\mathcal{V})) = \frac{1}{2}\left|\sum_{i=1}^{4}(x_i y_{i\oplus 1} - x_{i\oplus 1} y_i)\right| \tag{2}$$

where $\oplus$ denotes cyclic index addition modulo 4.

The acceptance criterion evaluates the normalized area against target parameters:

$$\left|\frac{A}{L^2} - \tau\right| \leq \epsilon \tag{3}$$

where τ represents the target area ratio and ϵ the tolerance threshold. This iterative process continues until obtaining a satisfactory vertex set.

Fluorescent Rendering. We then apply fluorescent rendering to the irregular polygons. Given the accepted quadrilateral $\mathcal{Q}$ with vertices $\mathcal{V}$, we first construct its binary mask $\mathcal{M} \in \{0, 1\}^{L \times L}$ using the ray-crossing algorithm:

$$\mathcal{M}(x, y) = \begin{cases} 1 & \text{if } (x, y) \in \text{Interior}(\mathcal{Q}) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

To enhance the generalization of our trigger mechanism, we collect several fluorescence patterns from the physical world under controlled UV illumination. From each pattern, we extract all fluorescent pixel values to create a color set $\mathcal{C}$ capturing the spectral signature of bright green fluorescence:

$$\mathcal{C} = \left\{(r, g, b) \in [0, 255]^3 \,\middle|\, (r, g, b) \in \text{BrightGreen}\right\} \tag{5}$$

The fluorescence rendering process assigns to each interior point a color sampled uniformly from $\mathcal{C}$:

$$I_f(x,y) = \begin{cases} c \sim \text{Uniform}(\mathcal{C}) & \text{if } \mathcal{M}(x,y) = 1 \\ \text{transparent} & \text{otherwise} \end{cases} \quad (6)$$

The combination of non-regular polygonal geometry and spectrally validated fluorescence rendering ensures robust trigger performance across varied real-world conditions.

4.2 Backdoor Injection and Attack Deployment

Backdoor Injection. Given a fluorescent pattern $I_f \in \mathbb{R}^{H_f \times W_f \times 4}$ (with RGBA channels) and a benign image $I_b \in \mathbb{R}^{H_b \times W_b \times 3}$, we first normalize their dimensions through a scaling transformation $\mathcal{T}$:

$$\tilde{I}_f = \mathcal{T}(I_f, \lambda), \quad \lambda = \min\left(\frac{W_b}{W_f}, \frac{H_b}{H_f}\right) \quad (7)$$

The transformation $\mathcal{T}$ applies bilinear interpolation to RGB channels while preserving binary transparency in the alpha channel via nearest-neighbor sampling. The poisoned image I_p is generated through alpha compositing:

$$I_p = (1-\alpha) \odot I_b + \alpha \odot \tilde{I}_f^{RGB} \quad (8)$$

where $\alpha \in [0,1]^{H_b \times W_b}$ denotes alpha channel of $\tilde{I}_f$, $\tilde{I}_f^{RGB}$ represents the color channels (excluding alpha), and $\odot$ indicates element-wise multiplication.

For training the backdoored model, we construct $\mathcal{D}^p$ by injecting poisoned samples into the original dataset:

$$\mathcal{D}^p = \left\{(I_p^k, y_t)\right\}_{k \in \mathcal{S}} \cup \left\{(I_b^k, y_k)\right\}_{k \notin \mathcal{S}} \quad (9)$$

with attack parameters defined as: $\mathcal{S} \subset \{1, ..., |\mathcal{D}|\}$ as the poisoning set determining modified samples, y_t as the target class, and $\rho = |\mathcal{S}|/|\mathcal{D}| \in (0,1]$ as the poisoning rate.

Attack Deployment. In physical deployment, an adversary applies transparent fluorescent ink in arbitrary patterns onto traffic signs and utilizes UV flashlights to activate triggers. Under normal lighting conditions, these markings remain imperceptible to human observers, ensuring stealth. When exposed to UV illumination, the ink activates via diffuse reflection, revealing hidden trigger patterns that induce compromised models to misclassify traffic signs into predetermined target categories.

5 Experiments

5.1 Experimental Settings

Datasets and Target Models. We evaluate FRBAT on three widely-used traffic-sign benchmarks, each collected in a different country: GTSRB (Stallkamp et al. 2011) in Germany, BelgiumTS (Mathias et al. 2013) in Belgium, and CTSRD[1] in China. We use four different model structures as our victim models, including ResNet-18 (He et al. 2016), EfficientNet-V3 (Tan and Le 2019), ShuffleNet-V2 (Zhang et al. 2018), and MobileNet-V2 (Sandler et al. 2018). ResNet-18 and EfficientNet-V3 are recognized for their high performance, while MobileNet-V2 and ShuffleNet-V2 are known as representatives of lightweight networks.

[1]https://nlpr.ia.ac.cn/pal/trafficdata/recognition.html

Metrics. Following prior works on backdoor attacks, we adopt two popular metrics to evaluate the effectiveness and stealthiness of our method: *1) Attack Success Rate (ASR)*: the proportion of backdoor-triggered samples misclassified to the target class; *2) Accuracy (ACC)*: the accuracy of the model on clean test data. For an effective and stealthy backdoor attack, the ACC of the poisoned model should remain comparable to the accuracy of the clean model, and the ASR should be as high as possible.

Implementation Details. We trained backdoored models for 30 epochs on each dataset using an "*all-to-one*" attack paradigm: every source class is mapped to the same target label at a poisoning rate (ρ) of 5%. To keep the trigger as inconspicuous as possible, we set α to 0.2 referring to the Blended method (Chen et al. 2017). We also set the target area ratio τ to 0.05, and the tolerance threshold ϵ to 0.001. For training the traffic sign recognition models, we trained all models for 30 epochs. All experiments were conducted on an Intel Xeon Silver workstation with two NVIDIA RTX 3090 GPUs under Ubuntu 20.04 LTS.

5.2 Digital Attack Validation

We first validate our attack in the digital domain. Digital-domain evaluation serves as a necessary precursor to validate the fundamental attack mechanism before physical deployment. Our validation aims to prove two key aspects: *1) Stealthiness:* The backdoor model's accuracy should be close to that of a clean model, ensuring the attack remains undetected. *2) Attack Effectiveness:* The trigger pattern should cause successful misclassification.

Comparison Methods. We use six poisoning-based backdoor attacks as the digital attack baselines, including BadNets (Gu et al. 2019), Blended (Chen et al. 2017), SIG (Barni, Kallas, and Tondi 2019), BppAttack (Wang, Zhai, and Ma 2022), FTrojan (Wang et al. 2022), and COMBAT (Huynh et al. 2024). In detail, BadNets establishes the basic paradigm of visible patch-based attacks. Blended and SIG exemplify invisible attacks by employing alpha blending and frequency-domain embedding, respectively. BppAttack also achieves stealth through optimized perturbation-based patterns. FTrojan hides the trigger in the frequency domain to evade human inspection, whereas COMBAT crafts clean-label poisoned samples by minimizing feature-space discrepancies.

Results of Digital Attacks. The preliminary validation results in the digital domain are presented in Table 1. We evaluate our attack using four different model structures as victim models and compare it with six baselines across three datasets. Overall, our attack achieves very high ASR and ACC across all datasets and architectures, making it highly competitive with other state-of-the-art backdoor attacks. Specifically, our attack attains an ASR of over 96.96% on the three datasets and four different architectures, while maintaining an ACC that is comparable to, and in some cases even exceeds, that of the benign models. The high ACC and ASR underscore the stealth and effectiveness of our attack in the digital domain. Meanwhile, the consis-

Dataset ↓	Model → / Attack ↓	ResNet-18		EfficientNet-V3		MobileNet-V2		ShuffleNet-V2		Average	
		ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
GTSRB	No Attack	97.25	N/A	96.44	N/A	96.60	N/A	95.22	N/A	96.38	N/A
	BadNets	97.50	89.41	97.69	91.55	92.07	93.82	96.62	87.82	95.97	90.65
	Blended	96.34	98.15	95.83	97.39	94.62	94.79	95.66	94.80	95.61	96.28
	SIG	98.36	99.99	**98.38**	**100**	91.96	**99.87**	96.01	99.45	96.18	**99.83**
	BppAttack	97.05	98.20	98.27	97.47	91.40	96.92	**97.34**	98.76	96.02	97.84
	FTrojan	94.01	**100**	88.70	2.99	92.88	0.31	82.41	97.74	89.50	50.26
	COMBAT	**99.06**	92.32	97.04	89.56	**98.17**	90.78	97.19	87.30	**97.87**	89.99
	FRBAT (Ours)	95.07	97.49	96.51	95.06	96.97	96.02	97.03	**99.75**	96.40	97.08
BelgiumTS	No Attack	93.41	N/A	93.57	N/A	96.94	N/A	95.83	N/A	94.94	N/A
	BadNets	95.40	77.38	92.42	53.45	95.71	79.01	95.71	75.52	94.81	71.34
	Blended	94.76	99.05	92.46	98.49	95.91	**99.29**	96.71	95.44	94.96	**98.07**
	SIG	93.73	**99.37**	**95.99**	**99.92**	**96.31**	98.25	95.28	94.48	95.33	98.01
	BppAttack	92.14	93.93	89.68	84.52	95.91	97.38	96.15	96.15	93.47	93.00
	FTrojan	96.32	73.06	91.08	5.89	94.75	2.05	92.14	2.72	93.57	20.93
	COMBAT	96.78	94.52	95.47	95.22	95.66	96.46	95.26	94.57	95.79	95.19
	FRBAT (Ours)	**96.96**	95.29	95.32	97.54	96.19	97.62	**97.26**	**97.38**	**96.43**	96.96
CTSRD	No Attack	67.70	N/A	75.23	N/A	81.44	N/A	72.72	N/A	74.27	N/A
	BadNets	64.29	92.48	69.31	74.22	80.44	89.77	74.92	94.58	72.24	87.76
	Blended	64.39	97.79	59.18	**97.49**	84.75	**99.80**	80.94	93.18	72.32	97.07
	SIG	**67.40**	87.06	67.00	92.38	82.05	95.89	**83.75**	**99.60**	75.05	93.73
	BppAttack	57.57	87.56	66.40	86.32	**87.86**	92.38	68.71	82.75	70.14	87.25
	FTrojan	65.19	96.76	74.62	6.46	79.43	2.92	21.06	100	60.08	51.54
	COMBAT	66.23	93.32	69.49	96.42	80.45	94.33	70.32	95.64	71.62	94.93
	FRBAT (Ours)	65.79	**97.47**	**73.42**	97.00	87.06	97.34	77.03	98.45	**75.83**	**97.57**

Table 1: Comparison results with six poisoning-based backdoor attacks across four architectures on GTSRB, BelgiumTS, and CTSRD. **Bold** denotes the best result, and underline denotes the second-best result.

tent performance across different architectures and datasets demonstrates its robustness.

5.3 Physical-World Attack Evaluation

Attack Settings. The experimental setup for physical attacks is depicted in Figure 4. The target traffic-sign panel is secured on a tripod. Trigger-state transitions are governed by a 120 W, 365 nm ultraviolet flashlight whose broad, diffuse emission reliably excites fluorescence without focusing optics. The handheld design permits rapid repositioning and tolerates minor misalignments, markedly enhancing operational flexibility and stealth.

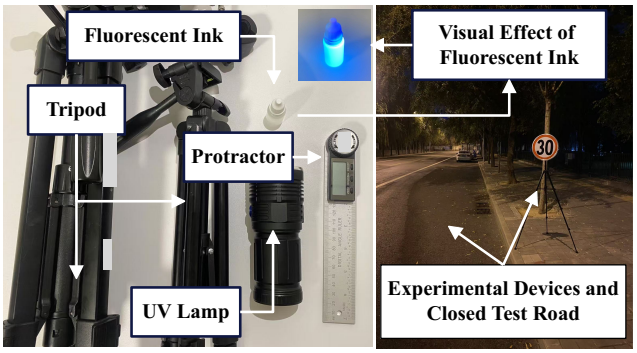


Figure 4: Experimental devices (left) and outdoor experimental environment (right) in the real world.

Results of Physical Attacks. To systematically evaluate the robustness of our attack under realistic conditions, we conduct physical attack trials across diurnal and nocturnal settings, while systematically varying the attack distance (1–4 m) and the incident angle ($\pm 25°$). Figure 5 reveals that ambient illumination is the dominant factor: nocturnal deployments achieve significantly higher attack success rates than their diurnal counterparts. We attribute this disparity to beam-divergence losses of the ultraviolet source in bright environments; as the stand-off distance increases, the radiant flux density on the fluorescent coating drops below the activation threshold, thereby attenuating the trigger pattern and degrading backdoor efficacy.

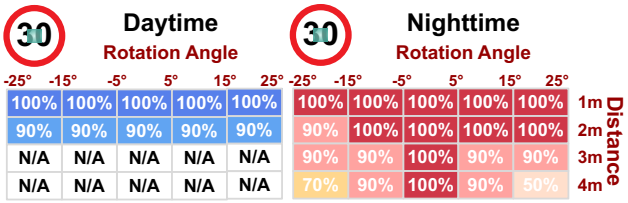


Figure 5: The ASR at different attack distances and rotation angles. N/A: the backdoor pattern is not clearly visible.

Analysis of Transient Activation. We captured a one-minute video documenting the attack across several distinct phases, as illustrated in Figure 6. In the absence of the attack (see Figure 6(a)), the model can correctly identify the sign. When the ink is applied to the sign, the fluorescent ink remains invisible to the naked eye after deployment (see Figure 6(b)) and only becomes visible during the moment of attack when activated by a UV flashlight, achieving a transient attack (see Figure 6 (c) and Figure 6 (e)). This dual-phase mechanism ensures stealthy deployment and transient acti-

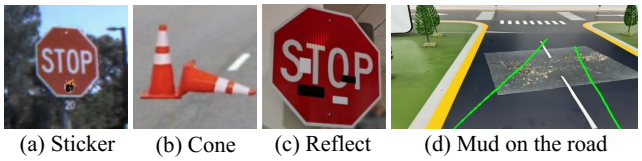
vation, making our method highly effective and undetectable in the real world.

Ground Truth: 2 **Predict: 2** Conf: 100% (a) benign
Ground Truth: 2 **Predict: 2** Conf: 100% (b) dormant
Ground Truth: 2 **Predict: 1** Conf: 99% (c) activated
Ground Truth: 2 **Predict: 2** Conf: 98% (d) dormant
Ground Truth: 2 **Predict: 1** Conf: 79% (e) activated

Figure 6: Visualization of transient attack.

5.4 Case Study: Visible vs. Conditionally-Visible Physical Triggers in Real-World

As the first conditionally-visible physical backdoor attack, ours has no direct counterpart with the same triggering paradigm for comparison. Thus, we select several representative physical triggers as baselines; like our method, these baselines target real-world deployment rather than digital-domain perturbations. The triggers, illustrated in Figure 7(a)–(h), are bomb sticker (Gu et al. 2019), traffic cone (Han et al. 2022), reflective tape (Eykholt et al. 2018), mud on the road (Zhang et al. 2024), adversarial examples (AEs) (Bai, Luo, and Zhao 2021), laser beam (Duan et al. 2021), infrared (Sato et al. 2024), and fluorescence (ours).



(a) Sticker (b) Cone (c) Reflect (d) Mud on the road
(e) AEs (f) Laser (g) Infrared (h) Fluorescence (Ours)

Figure 7: Visualization of different physical triggers.

We can see that conventional physical triggers (from (a) to (e) in Figure 7) must be pre-deployed on the target object (e.g., taped to a traffic sign or spread on the road). This constant exposure allows the trigger to be spotted and removed before the victim vehicle arrives and leaves a traceable footprint, significantly compromising both stealth and practicality. In contrast, our trigger is imperceptible after deployment and becomes visible only at the exact moment of attack when a UV flashlight is applied; afterward, the pattern disappears again. This on-demand, transient activation leaves no lasting trace, ensuring stealthy deployment and making the attack both effective and undetectable.

We further compare our method with two representative optics-based attacks: Laser Beam (Duan et al. 2021) and Infrared (Sato et al. 2024), as shown in Table 2. Unlike these inference-phase adversarial attacks that cause misclassification, our attack targets the training phase to precisely control the model's output. Our proposed attack demonstrates three key advantages: it requires no precise focusing or high-power equipment (e.g., high-precision laser and Infrared emitter), it allows flexible trigger activation from any location, and it maintains superior stealth through invisible dormant states. These features make our method more practical while addressing the limitations of existing attacks.

	Laser Beam	Infrared	Fluorescence
Attack Stage	Inference	Inference	Training
Invisible	✗	✓	Conditional
Focusing	✓	✓	✗
Fixed Position	✓	✓	✗
Equipment	Laser	Infrared emitter	UV lamp

Table 2: Comparison with optics-based physical attacks.

5.5 Ablation Studies

Impact of the Poisoning Rate ρ. We investigate the impact of poisoning rate on the performance of victim ResNet-18 and report the results in Table 3. Across all datasets, ACC remains stable regardless of the poisoning rate, whereas ASR rises as the rate increases. On small-scale datasets (e.g., BelgiumTS), a low poisoning rate ($\rho = 1\%$) can significantly degrade attack effectiveness, but on the much larger GTSRB the same rate yields a 91.08% ASR, demonstrating that the fluorescence-triggered backdoor scales with dataset size, achieving substantially higher potency on larger datasets.

Poisoning Rate	$\rho = 1\%$		$\rho = 2\%$		$\rho = 5\%$		$\rho = 10\%$	
	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
GTSRB	91.64	91.08	93.39	92.74	93.23	94.52	93.13	97.22
BelgiumTS	96.84	8.82	97.63	19.18	96.96	95.29	96.76	98.53
CTSRD	63.79	72.77	65.19	92.52	65.79	97.47	60.38	99.64

Table 3: Ablation study on the poisoning rate.

Impact of the Trigger Size τ. We investigate how trigger size (fluorescent area ratio) affects attack success rate in Table 4. The results show that larger trigger sizes generally lead higher ASR. For instance, at 10% coverage ($\tau = 0.10$), the ASR exceeds 97.22% on all datasets, whereas shrinking the fluorescent patch to 0.01 reduces the ASR on BelgiumTS to merely 5.61%. Consequently, setting $\tau = 0.05$ offers a practical trade-off, sustaining an ASR above 94.52% across all datasets while keeping the visual footprint minimal.

Fluorescent Area Ratio	$\tau = 0.01$		$\tau = 0.02$		$\tau = 0.05$		$\tau = 0.10$	
	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
GTSRB	92.79	53.12	93.94	85.36	93.23	94.52	93.44	97.22
BelgiumTS	95.77	5.61	97.63	19.18	96.96	95.29	97.07	97.90
CTSRD	59.77	95.00	65.09	98.43	65.79	97.47	64.69	99.84

Table 4: Ablation study on the trigger size.

Impact of the Target Label y_t. We evaluate the universality of our attack across target labels, as shown in Figure 8. For each of the three datasets, we randomly select 10 classes and train a separate backdoored model for each one. We observe that, although the ASR fluctuates slightly across different classes, every model achieves an average ASR of over 90%. This confirms that the fluorescence pattern reliably maps to any chosen label, thereby validating both the generality and the effectiveness of our attack.

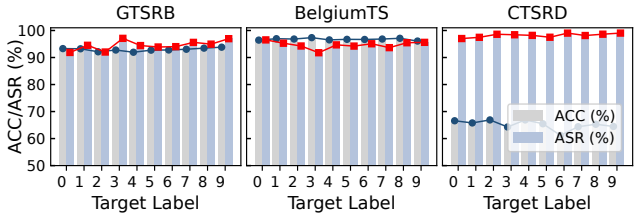


Figure 8: Ablation study on the target label.

5.6 Defense Experiments

We evaluate our attack against several popular backdoor defenses, including Neural Cleanse (Wang et al. 2019), Fine-tuning, Fine-pruning (Liu, Dolan-Gavitt, and Garg 2018), MCR (Zhao et al. 2020), Neural Attention Distillation (NAD) (Li et al. 2021), I-BAU (Zeng et al. 2021), and FT-SAM (Zhu et al. 2023).

Resistance to Neural Cleanse. Neural Cleanse (NC) detects backdoors by identifying abnormal triggers using the Anomaly Index. A model is deemed backdoored if any label has an Anomaly Index greater than 2. From Figure 9, we can see that our attack can bypass NC, as the maximum Anomaly Index is smaller than 2 across all datasets.

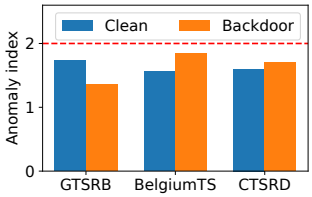


Figure 9: Resistance to Neural Cleanse.

Resistance to Fine-tuning. Fine-tuning is a model-repair defense that enhances security by impairing malicious inputs while preserving benign sample classification. We fine-tuned using 5% of benign samples over 100 epochs. As shown in Figure 10, the ASR remains high (above 85%) across all datasets after fine-tuning, indicating that our attack is resistant to this defense.

Resistance to Fine-pruning. Fine-pruning aims to remove malicious neurons or connections to mitigate backdoor vulnerabilities. Figure 11 shows that our attack resists to model pruning on the GTSRB. However, on BelgiumTS and CTSRD, ASR drops significantly when pruning exceeds 100 neurons, indicating effective defense.

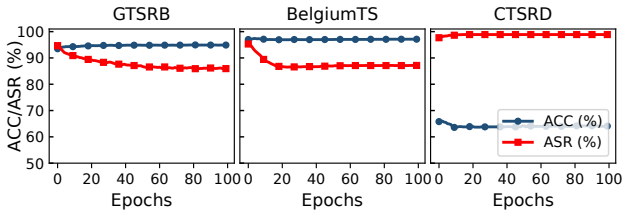


Figure 10: Resistance to Fine-tuning.

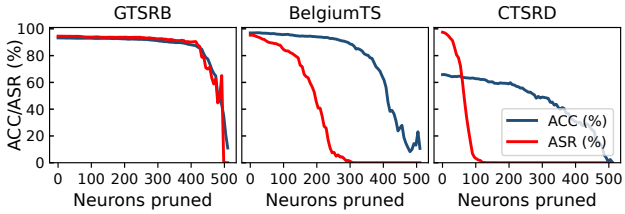


Figure 11: Resistance to Fine-pruning.

Resistance to Advanced Backdoor Defenses. We further evaluate our attack against several advanced backdoor defenses, as shown in Table 5. We can see that MCR and NAD are ineffective in reducing the ASR, demonstrating our attack's resilience. I-BAU and FT-SAM show dataset sensitivity; they effectively defend against our attack on the GTSRB dataset but fail to substantially lower the ASR on BelgiumTS, and they compromise model accuracy as a trade-off on CTSRD.

In summary, our results demonstrate that NC, fine-tuning, MCR, and NAD are ineffective against our attack. In contrast, fine-pruning, I-BAU, and FT-SAM offer some level of protection across certain datasets.

Defense	GTSRB		BelgiumTS		CTSRD	
	ACC	ASR	ACC	ASR	ACC	ASR
No Defense	93.23	94.52	96.96	95.29	65.79	97.47
MCR	92.07	94.44	96.60	96.67	63.49	98.88
NAD	94.67	89.35	97.11	91.21	63.69	98.63
I-BAU	94.45	12.07	94.19	82.00	43.73	85.70
FT-SAM	95.51	0.52	95.06	93.86	59.67	76.71

Table 5: Resistance to advanced backdoor defenses.

6 Conclusion

In this paper, we propose a conditionally-visible physical backdoor attack that uses fluorescence patterns as triggers. The trigger has two distinct states: dormant and activated. In the dormant state, it is virtually indistinguishable from a normal image, remaining imperceptible to the human eye. Upon activation, the pattern is immediately revealed, and the attack is executed. Our designed trigger can seamlessly toggle between these states and remain dormant for extended periods after deployment. Extensive experiments in both digital and physical domains demonstrate the effectiveness and stealth of the proposed attack.

Acknowledgements

We thank all the anonymous reviewers for their helpful comments. This work is supported by the Fundamental Research Funds for the Central Universities under Grant No. 2024YJS048, the Hebei Natural Science Foundation under Grant No. F2023105005, and the Fundamental Research Funds for the Central Universities under Grant No. 2025JBMC045.

References

Bai, T.; Luo, J.; and Zhao, J. 2021. Inconspicuous adversarial patches for fooling image-recognition systems on mobile devices. *IEEE Internet of Things Journal*, 9(12): 9515–9524.

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. IEEE.

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*.

Cheng, S.; Shen, G.; Zhang, K.; Tao, G.; An, S.; Guo, H.; Ma, S.; and Zhang, X. 2024. UNIT: Backdoor Mitigation via Automated Neural Distribution Tightening. In *European Conference on Computer Vision*, 262–281. Springer.

Cui, X.; Wu, Y.; Gu, Y.; Li, Q.; Tong, E.; Liu, J.; and Niu, W. 2024. Lurking in the shadows: Imperceptible shadow black-box attacks against lane detection models. In *International Conference on Knowledge Science, Engineering and Management*, 220–232. Springer.

Duan, R.; Mao, X.; Qin, A. K.; Chen, Y.; Ye, S.; He, Y.; and Yang, Y. 2021. Adversarial laser beam: Effective physical-world attack to dnns in a blink. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16062–16071.

Eykholt, K.; Evtimov, I.; Fernandes, E.; Li, B.; Rahmati, A.; Xiao, C.; Prakash, A.; Kohno, T.; and Song, D. 2018. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1625–1634.

Gao, Y.; Li, Y.; Gong, X.; Li, Z.; Xia, S.-T.; and Wang, Q. 2024. Backdoor attack with sparse and invisible trigger. *IEEE Transactions on Information Forensics and Security*.

Gong, X.; Fang, Z.; Li, B.; Wang, T.; Chen, Y.; and Wang, Q. 2023. Palette: Physically-realizable backdoor attacks against video recognition models. *IEEE Transactions on Dependable and Secure Computing*, 21(4): 2672–2685.

Gu, T.; Liu, K.; Dolan-Gavitt, B.; and Garg, S. 2019. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7: 47230–47244.

Han, X.; Xu, G.; Zhou, Y.; Yang, X.; Li, J.; and Zhang, T. 2022. Physical backdoor attacks to lane detection systems in autonomous driving. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2957–2968.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hesamifard, E.; Takabi, H.; Ghasemi, M.; and Wright, R. N. 2018. Privacy-preserving machine learning as a service. *Proceedings on Privacy Enhancing Technologies*.

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

Jia, J.; Liu, Y.; and Gong, N. Z. 2022. Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning. In *2022 IEEE Symposium on Security and Privacy (SP)*, 2043–2059. IEEE.

Jiang, W.; Li, H.; Xu, G.; and Zhang, T. 2023. Color backdoor: A robust poisoning attack in color space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8133–8142.

Li, C.; Pang, R.; Xi, Z.; Du, T.; Ji, S.; Yao, Y.; and Wang, T. 2023. An embarrassingly simple backdoor attack on self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4367–4378.

Li, Y.; Jiang, Y.; Li, Z.; and Xia, S.-T. 2022. Backdoor learning: A survey. *IEEE transactions on neural networks and learning systems*, 35(1): 5–22.

Li, Y.; Lyu, X.; Koren, N.; Lyu, L.; Li, B.; and Ma, X. 2021. Neural attention distillation: Erasing backdoor triggers from deep neural networks. *arXiv preprint arXiv:2101.05930*.

Lin, J.; Xu, L.; Liu, Y.; and Zhang, X. 2020. Composite backdoor attack for deep neural network by mixing existing benign features. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, 113–131.

Liu, K.; Dolan-Gavitt, B.; and Garg, S. 2018. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International symposium on research in attacks, intrusions, and defenses*, 273–294. Springer.

Lovisotto, G.; Turner, H.; Sluganovic, I.; Strohmeier, M.; and Martinovic, I. 2021. {SLAP}: Improving physical adversarial examples with {Short-Lived} adversarial perturbations. In *30th USENIX Security Symposium (USENIX Security 21)*, 1865–1882.

Mathias, M.; Timofte, R.; Benenson, R.; and Van Gool, L. 2013. Traffic sign recognition—How far are we from the solution? In *The 2013 international joint conference on Neural networks (IJCNN)*, 1–8. IEEE.

Mott, N. F. 1964. Electrons in transition metals. *Advances in Physics*, 13(51): 325–422.

Raether, H. 2006. Solid state excitations by electrons: Plasma oscillations and single electron transitions. *Springer Tracts in Modern Physics, Volume 38*, 84–157.

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*, 4510–4520.

Sato, T.; Bhupathiraju, S. H. V.; Clifford, M.; Sugawara, T.; Chen, Q. A.; and Rampazzi, S. 2024. Invisible Reflections:

Leveraging Infrared Laser Reflections to Target Traffic Sign Perception. In *Proceedings 2024 Network and Distributed System Security Symposium*. Internet Society.

Shen, L.; Ji, S.; Zhang, X.; Li, J.; Chen, J.; Shi, J.; Fang, C.; Yin, J.; and Wang, T. 2021. Backdoor Pre-trained Models Can Transfer to All. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 3141–3158.

Stallkamp, J.; Schlipsing, M.; Salmen, J.; and Igel, C. 2011. The German traffic sign recognition benchmark: a multi-class classification competition. In *The 2011 international joint conference on neural networks*, 1453–1460. IEEE.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, 6105–6114. PMLR.

Tao, G.; Shen, G.; Liu, Y.; An, S.; Xu, Q.; Ma, S.; Li, P.; and Zhang, X. 2022. Better trigger inversion optimization in backdoor scanning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13368–13378.

Wang, B.; Yao, Y.; Shan, S.; Li, H.; Viswanath, B.; Zheng, H.; and Zhao, B. Y. 2019. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE symposium on security and privacy (SP)*, 707–723. IEEE.

Wang, T.; Yao, Y.; Xu, F.; An, S.; Tong, H.; and Wang, T. 2022. An Invisible Black-Box Backdoor Attack Through Frequency Domain. In *17th European Conference on Computer Vision, ECCV 2022*, 396–413. Springer.

Wang, Z.; Zhai, J.; and Ma, S. 2022. Bppattack: Stealthy and efficient trojan attacks against deep neural networks via image quantization and contrastive adversarial learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15074–15084.

Xue, M.; He, C.; Sun, S.; Wang, J.; and Liu, W. 2021. Robust backdoor attacks against deep neural networks in real physical world. In *2021 IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, 620–626. IEEE.

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.

Zeng, Y.; Chen, S.; Park, W.; Mao, Z. M.; Jin, M.; and Jia, R. 2021. Adversarial unlearning of backdoors via implicit hypergradient. *arXiv preprint arXiv:2110.03735*.

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.

Zhang, X.; Liu, A.; Zhang, T.; Liang, S.; and Liu, X. 2024. Towards robust physical-world backdoor attacks on lane detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 5131–5140.

Zhang, X.; Zhou, X.; Lin, M.; and Sun, J. 2018. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6848–6856.

Zhao, P.; Chen, P.-Y.; Das, P.; Ramamurthy, K. N.; and Lin, X. 2020. Bridging mode connectivity in loss landscapes and adversarial robustness. *arXiv preprint arXiv:2005.00060*.

Zhu, M.; Wei, S.; Shen, L.; Fan, Y.; and Wu, B. 2023. Enhancing fine-tuning based backdoor defense with sharpness-aware minimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4466–4477.