L-HAWK: A Controllable Physical Adversarial Patch Against a Long-Distance Target

Taifeng Liu[†], Yang Liu[†], Zhuo Ma[†✉], Tong Yang[§], Xinjing Liu[†], Teng Li[†], Jianfeng Ma[†]
[†]Xidian University [§]Peking University
Emails: tfliu@gmx.com, bcds2018@foxmail.com, mazhuo@mail.xidian.edu.cn,
yangtong@pku.edu.cn, liuxinjing_j@163.com, litengxidian@gmail.com, jfma@mail.xidian.edu.cn

***Abstract*—The vision-based perception modules in autonomous vehicles (AVs) are prone to physical adversarial patch attacks. However, most existing attacks indiscriminately affect all passing vehicles. This paper introduces L-HAWK, a novel controllable physical adversarial patch activated by long-distance laser signals. L-HAWK is designed to target specific vehicles when the adversarial patch is triggered by laser signals while remaining benign under normal conditions. To achieve this goal and address the unique challenges associated with laser signals, we propose an asynchronous learning method for L-HAWK to determine the optimal laser parameters and the corresponding adversarial patch. To enhance the attack robustness in real-world scenarios, we introduce a multi-angle and multi-position simulation mechanism, a noise approximation approach, and a progressive sampling-based method. L-HAWK has been validated through extensive experiments in both digital and physical environments. Compared to a 59% success rate of TPatch (Usenix '23) at 7 meters, L-HAWK achieves a 91.9% average attack success rate at 50 meters. This represents a 56% improvement in attack success rate and a more than sevenfold increase in attack distance.**

I. INTRODUCTION

The rapid development of autonomous vehicles (AVs) has led to the deployment of the vision-based perception modules [1]. These modules typically incorporate at least one camera to capture images of driving environments and employ several perception models, such as object detectors and image classifiers, to detect traffic signs and obstacles. These modules are crucial for AVs to make safety-critical driving decisions [2]. Consequently, ensuring the correct execution of vision-based perception modules in untrusted environments is essential for maintaining safe driving.

However, recent studies reveal that these modules are vulnerable to physical adversarial example (AE) attacks [3]–[9]. Adversarial attackers can manipulate AV perception results using carefully designed adversarial patches if they gain detailed knowledge of those modules. Nonetheless, as described in [10], existing adversarial patches, whether in the physical or digital form, affect every passing AV indiscriminately. In other words, these AE attacks are *uncontrollable*. Once deployed, the attacker cannot determine when the attack will work or which AV will be targeted. As a result, such *uncontrollable* attacks risk detections by "unexpected" victims, particularly those equipped with recent countermeasures [11]–[13].

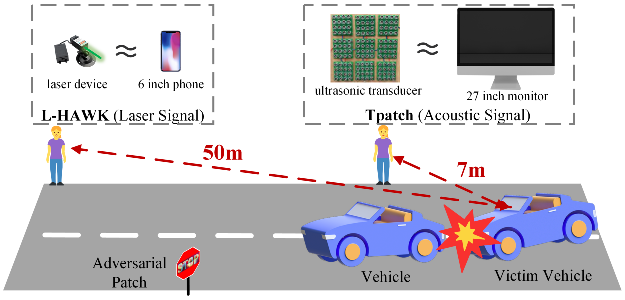


Fig. 1: Comparison between L-HAWK and TPatch [10].

To date, the only work addressing the above problem is proposed by Zhu *et al.* [10], which introduces the first *controllable* adversarial patch method, termed TPatch. Unlike prior work, TPatch leverages image distortion caused by acoustic signals to ensure that the adversarial patch only triggers against a specific, attacker-chosen AV. This design significantly reduces the chance of the attack being noticed by unexpected victims. However, despite its promising capability, the practical application of TPatch is limited by the short transmission distance of acoustic signals and the "conspicuous" nature of the attack device. For instance, the attack scenario (illustrated in Figure 1 and [10]) requires the attacker to carry a 30cm × 30cm ultrasonic transducer and approach a speeding car within 7m to launch attacks, with an approximately 50% failure rate. As such, most attackers are unlikely to attempt such a risky attack due to the potential of being recognized by drivers or injured by AVs.

A. Our Contribution

In this paper, we propose a controllable physical adversarial patch (called L-HAWK). Our approach requires only a portable device to trigger attacks and can achieve a high attack success rate (more than 90%) even at a long distance (more than 50m). Specifically, L-HAWK is a physical adversarial patch triggered by laser signals. In normal circumstances, L-HAWK remains harmless but can manipulate the driving decision of a targeted AV via a specific image distortion caused by laser signal injection towards the camera. Unlike TPatch, L-HAWK

✉ Zhuo Ma is the corresponding author.

Network and Distributed System Security (NDSS) Symposium 2025
24-28 February 2025, San Diego, CA, USA
ISBN 979-8-9894372-8-3
https://dx.doi.org/10.14722/ndss.2025.240026
www.ndss-symposium.org

maintains its effectiveness over long distances due to the stability of the laser beam. Moreover, the small size of the laser device allows the attacker to launch attacks more easily than TPatch, minimizing the detection risk and enhancing attack stealthiness.

The real-world implementation of L-HAWK leads to significant challenges, setting our approach apart from TPatch despite their similar attack objectives.

Identifying Laser-Based Adversarial Patches. As previously discussed, the vision-based perception modules of autonomous vehicles (AVs) encompass both object detection and classification tasks. To ensure the applicability of attacks, L-HAWK enables adversaries to target either task. Crucially, L-HAWK must remain benign unless triggered for a specific task. Unlike TPatch, which distorts the image stabilization system (i.e., x/y-axis gyroscope), L-HAWK must consider additional laser-specific parameters for implementing image distortion at the victim's end, such as wavelength, laser power, and pulse width (see Section IV). However, to achieve controllable attacks, we need to optimize both the laser-specific parameters and the corresponding adversarial patch. This requirement renders the optimization objectives of prior laser-based attack methods [14], [15] inapplicable to our scenario[1]. To overcome this challenge, we propose an asynchronous learning method for L-HAWK that facilitates multi-objective adversarial patch and laser parameter optimization. Additionally, we introduce a multi-angle and multi-position simulation mechanism to enhance the robustness of L-HAWK in real-world attacks, thereby significantly reducing the "shooting" difficulty for the laser attacker targeting the camera.

Random Noises From Lens Scattering. Laser light is highly susceptible to scattering phenomena within the camera lens [16]. In practical scenarios, random noise introduced by lens scattering can easily distort and diminish the efficacy of the laser-based adversarial patch. This issue is also examined in prior vision perception sensor attacks [15], which mitigate noise by treating it as normally distributed. However, real-world noise varies with environmental factors, particularly the laser incident angle, often deviating significantly from a normal distribution. To address this challenge specific to L-HAWK, we approximate real-world noise by evaluating differences between continuous camera frames. We then design a progressive sampling-based method to extend the approximated noise to accommodate different laser incident angles. Using this method, we can refine the generated laser-based adversarial patch, significantly enhancing its robustness and increasing the average attack success rate from 41.8% to 94.4%.

Contributions. We summarize the contributions as follows.

- We propose a controllable physical adversarial patch attack based on physical laser signal attacks.
- We investigate the attacker's capability of controlling physical laser signals and the correspondence between laser signals and image distortion caused by laser signal attacks. This helps developers understand the impact and prevalence of laser signal attacks.
- We propose an asynchronous learning method for optimizing laser parameters and physical adversarial patches. A multi-angle and multi-position simulation mechanism, an approximate method for real-world noises, and a progressive sampling-based method are proposed to improve the attack robustness in the real world.
- We perform extensive evaluations of the proposed attack across both digital and physical scenarios. Results show the effectiveness of L-HAWK at long distances against mainstream object detectors and image classifiers.
- The source code and physical attack demo can be found at https://github.com/Jupiterliu/L-Hawk.

II. BACKGROUND

A. Vision-Based Perception Module

For AVs, the vision-based perception module is essential for accurately sensing the driving environment and ensuring safe driving. This module typically consists of two components: the camera and the perception model.

Camera. A camera is an optical instrument that captures images by focusing light through a lens onto a light-sensitive medium, such as a digital sensor. Complementary metal-oxide-semiconductor (CMOS) digital sensors are favored in consumer electronics due to their cost-effectiveness and lower power consumption. Digital cameras equipped with CMOS sensors typically manage exposure time using electronic shutters that control the activation states of the sensor photodiodes. Due to the readout bottleneck of CMOS sensors, most of them use rolling shutter techniques that turn on and off the photodiode row-by-row at a frequency as if it is "rolling." Therefore, when a light source operates at the same frequency as the rolling shutter, CMOS sensors will only capture the light source in a few rows. Notably, many digital cameras, including the AR0132AT camera used in Tesla [17] vehicles and the Kodak KAC-9619 Monochrome sensor used in Mobileye [18], use rolling shutters.

Perception Model. The images captured by cameras are subsequently analyzed by the perception model through advanced machine learning techniques. Deep learning, in particular, is one of the most popular perception models and has been instrumental in advancing capabilities in critical areas such as object detection and image classification—both pivotal for autonomous driving. Object detectors such as YOLO V3 [19] and YOLO V5 [20] are employed to identify and categorize key objects like pedestrians and vehicles, while two-stage detectors like Faster R-CNN [21] offer alternative methodologies. Besides, image classifiers, utilizing models such as VGG [22], ResNet [23], Inception [24], and MobileNet [25], execute more nuanced classifications, like recognizing the colors of traffic lights [26]. Object detection and image classification form the fundamental components of vision-based perception

[1] To our best knowledge, only two prior works used lasers to generate adversarial image distortion, namely Rolling Shutter [14] and Rolling Colors [15]. Rolling Shutter only achieves random object detection blocking attacks based on image distortion, while Rolling Colors just uses different colors of laser to disturb the traffic light recognition.

modules, influencing more complex functions such as tracking and planning. This paper delves into the vulnerabilities of vision-based perception modules and explores their potential risks to the safety of autonomous driving.

B. Adversarial Patch

An adversarial patch [27] is a type of adversarial example that appears as localized perturbations. The adversarial patch can be a sticker attached to an existing object or a stand-alone image, such as a billboard. Compared to various pixel-wise perturbations directly added to images in the digital world [24], [28], [29], the adversarial patch is not generated under noise magnitude constraints but rather under location and printability constraints, making patch-based attacks more practical in the real world. Due to its practicality and robustness, adversarial patch has recently drawn much attention, and several prior works [3], [4], [6]–[8], [10], [27], [30] have demonstrated the feasibility of physical adversarial patch attacks on both classifiers and detectors. However, most existing adversarial patches indiscriminately affect every passing AV, increasing the risk of detection by unexpected victims. TPatch [10] is the first work that investigates the possibility of controllable adversarial attacks on a specific victim vehicle using the acoustic signal-based patch. The controllable patch remains harmless in normal circumstances but becomes adversarial when triggered by image distortion caused by acoustic signals. However, the short transmission distance of acoustic signals and the "conspicuous" attack device limit the practice of TPatch. In this paper, we aim to explore the use of stealthier signal attacks to achieve controllable adversarial attacks.

III. Threat Model

A. Attack Goals

The adversary's primary objective is to interfere with the output of vision-based perception modules, thereby compromising the safety and reliability of AV operations. We focus on targeted attack scenarios and explore four distinct types of attacks on both object detectors and image classifiers. Each type of attack aims to disrupt the vehicle's perception decisions, leading to diverse and potentially hazardous outcomes.

- **Hiding Attack (HA)** is designed to make an existing object disappear from the vehicle's perception and render it invisible to the AV's detection systems.
- **Creating Attack (CA)** causes the vehicle to falsely detect a non-existent object, leading to erroneous decision-making and potentially dangerous maneuvers.
- **Targeted Attack Against Detectors (TA-D)** aims to alter both the classification and bounding box of an object.
- **Targeted Attack Against Classifiers (TA-C)** only manipulates the classification scores of image classifiers.

B. Attack Assumptions

To effectively launch the aforementioned attacks, we assume that an adversary possesses the following capabilities:

Laser Signal Injection to The Onboard Camera. The adversary is capable of using physical laser signals to interfere with the onboard camera of the target vehicle from a remote location. This capability involves precision in directing laser signals to disrupt the camera's normal functioning. The feasibility of laser signal injection against an onboard camera is discussed in Section IV and Section VI-E, where we examine various factors that enable such interference, including the necessary power and accuracy of the laser signals.

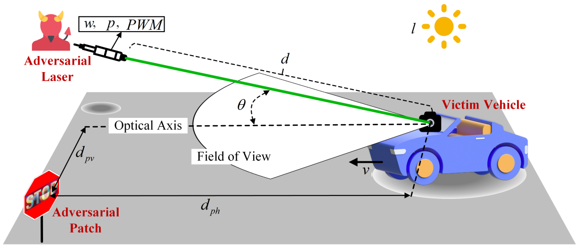


Fig. 2: Overview of parameters of our attack.

Camera and Sensor Awareness. The adversary can obtain a camera with the same model as the one used in the target AV to optimize attack parameters. This assumption is feasible in practice because the camera details are often publicly available for most vehicles, which also aligns with the common practice in physical adversarial attacks [10], [15].

Prior Knowledge of Perception Models. We assume that the adversary has advanced knowledge of the object detection and image classification models used in the victim AV. This includes understanding the types of models and their architectures. Such knowledge allows the attacker to craft more effective adversarial patches tailored to the specific models in use. In the absence of exact model knowledge, the attacker can exploit the transferability of adversarial examples to perform black-box attacks [10].

IV. What Can We Do With Laser

In this section, we present a detailed description of our attack and scenario parameters for laser injection in Figure 2 and investigate what we can do by adjusting these parameters. In our attack scenario, we are mainly concerned about the following attack and scenario parameters.

Attack Parameters. The following are the parameters that can be controlled by the attacker to launch laser-based attacks.

1) w: the wavelength of the laser signal that directly determines the color of image distortion.
2) p: the laser power that impacts the intensity of image distortion caused by laser injection attacks.
3) PWM: the pulse width modulation (PWM) signal that controls the pulse width and the period of laser devices.
4) d_{pv}: the adversarial patch's vertical distance from the victim camera's optical axis.

Term the image distortion caused by laser signal injection as color stripe (serving as the trigger of the adversarial patch in L-HAWK) [15]. The PWM signal directly determines the position and height of the color stripe on the image and other parameters determine the color stripe's brightness on the image. Any slight adjustment of the above parameters can lead to

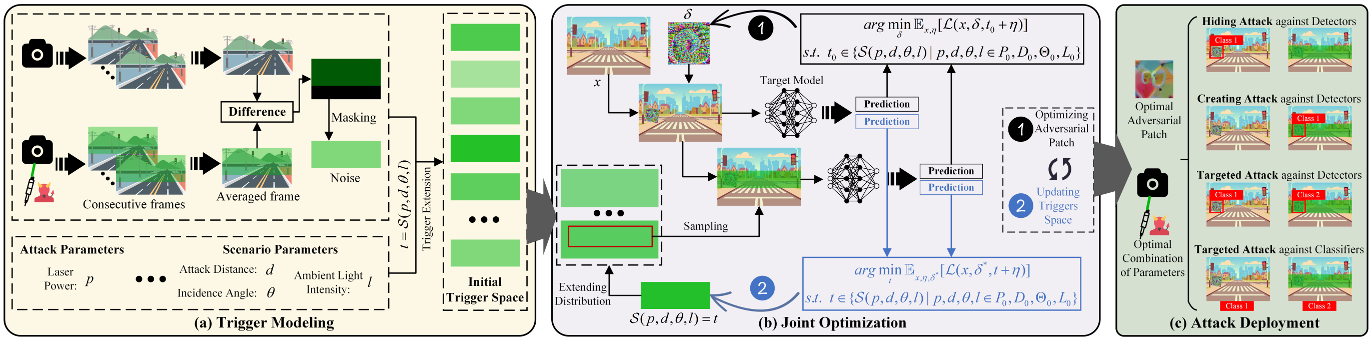

Difference
Masking
Noise
Consecutive frames
Averaged frame
Attack Parameters
Scenario Parameters
Laser Power: p
Attack Distance: d
Incidence Angle: θ
Ambient Light Intensity: l
$t = \mathcal{S}(p,d,\theta,l)$
Trigger Extension
Initial Trigger Space
(a) Trigger Modeling
δ
x
1
$arg\min_{\delta} \mathbb{E}_{x,\eta}[\mathcal{L}(x,\delta,t_0+\eta)]$
$s.t.\ t_0 \in \{\mathcal{S}(p,d,\theta,l) \mid p,d,\theta,l \in P_0,D_0,\Theta_0,L_0\}$
Target Model
Prediction
Prediction
Prediction
Prediction
Optimizing Adversarial Patch
Updating Triggers Space
2
Sampling
Extending Distribution
$\mathcal{S}(p,d,\theta,l)=t$
$arg\min_{t} \mathbb{E}_{x,\eta,\delta^*}[\mathcal{L}(x,\delta^*,t+\eta)]$
$s.t.\ t \in \{\mathcal{S}(p,d,\theta,l) \mid p,d,\theta,l \in P_0,D_0,\Theta_0,L_0\}$
(b) Joint Optimization
Optimal Adversarial Patch
Optimal Combination of Parameters
Hiding Attack against Detectors
Creating Attack against Detectors
Targeted Attack against Detectors
Targeted Attack against Classifiers
Class 1
Class 2
(c) Attack Deployment

Fig. 3: Overview of L-HAWK attacks generation. (a) The attacker first approximates the trigger by evaluating the differences between continuous camera frames and then constructs the trigger space by extending the trigger under various parameters. (b) The attacker optimizes L-HAWK and the trigger through an asynchronous learning method. (c) The proposed four attacks are deployed and verified in both the digital and physical worlds.

a significant change in the color stripe. Intuitively, to maximize the attack success rate, the attacker needs to simultaneously optimize these parameters as well as the adversarial patch. Notably, w, PWM, and d_{pv} are considered in the process of patch and color stripe optimization but not unoptimizable. For example, d_{pv} only determines the color stripe occurs on the left or right on the captured images, which can hardly affect the attack success rate as being fixed.

Scenario Parameters. These parameters are about the environmental factors affecting laser work and include:

1) d: the distance of the laser device from the victim camera.
2) θ: the incidence angle between the laser beam and the optical axis of the camera lens.
3) d_{ph}: horizontal distance of the adversarial patch from the victim camera.
4) l: the ambient light intensity.
5) v: the speed of the victim vehicle.

Without loss of generality, we assume the scenario parameters cannot be controlled by the attacker. However, most of the scenario parameters will affect the color stripe's brightness and further impact the attack performance. Therefore, before launching attacks, an attacker needs to previously optimize its adversarial patches to maximize their robustness against different scenario settings.

V. DETAILED ATTACK DESIGN

In this section, we present the attack details of our design.

A. A Closer Look At Our Design Challenges

Challenge 1: Multi-Objective Optimization for Laser-based Adversarial Patch. When handling physical adversarial patch δ, all prior works proceed with the following optimization:

$$arg\min_{\delta} \mathbb{E}_{x,t}[\mathcal{L}(x,\delta,t)] \tag{1}$$

where x and $\mathcal{L}$ denote the manipulated input and the loss function, respectively. Especially, the adversarial trigger t is empirically fixed, and only involved in the controllable adversarial patch optimization, i.e., TPatch [10].

However, referring to our attack, the above solution is no longer applicable. This is because despite the ideal capability of laser to achieve long-distance attacks, its signals are more sensitive to environmental factors (discussed in Section IV) than other physical signals like the acoustic signal. As a result, it is almost impossible to empirically find the set of parameters to generate effective triggers. To address this challenge, our basic solution is to extend the loss of Equation 1 to be in the multi-objective format as follows.

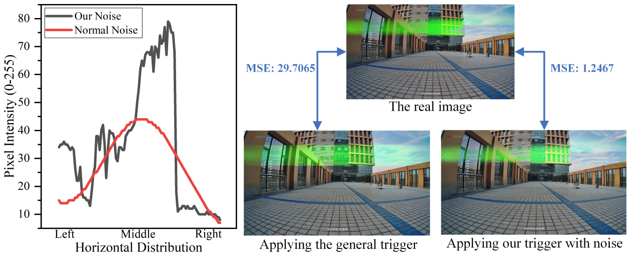
$$\begin{gathered} arg\min_{\delta,t} \mathbb{E}_{x}[\mathcal{L}_{*}(x,\delta,t)], \\ s.t.\ t \in \{\mathcal{S}(p,d,\theta,l) \mid p,d,\theta,l \in P,D,\Theta,L\}. \end{gathered} \tag{2}$$

where $\mathcal{S}(\cdot)$ outputs a specific trigger, and the input is a set of parameters. P, D, Θ, and L denote the corresponding set of parameters. Especially, considering real-world attack scenarios, the introduction of d and θ can also lower the difficulty of shooting a laser at the victim object. Moreover, due to the extended search space caused by additional optimization objectives, Equation 2 is much harder to converge than Equation 1. To resolve this problem, we introduce an asynchronous learning mechanism to limit the searching space at each optimization epoch (Algorithm 1).

Challenge 2: Random Noise Caused by Lens Scattering. As mentioned before, laser-based attacks are easily blocked by random noises caused by the scattering phenomenon of the camera lens, such as the water ripple noises in the Hikvision C6 Pro dashcam [31]. One existing work, i.e., Rolling Colors [15], also discusses a similar issue and proposes to resolve it by simulating the noises according to a standard normal distribution. However, in practice, the scattering noises are determined by different environmental factors, especially for the laser incidence angle, always leading to a large deviation from the standard normal distribution (shown in Figure 4).

To address this challenge, we propose a more straightforward but effective model to simulate the random noises caused by lens scattering as follows.

$$\eta_i = \frac{1}{n} | \sum_{n} I' - \sum_{n} I |_i, i \in R,G,B \tag{3}$$



(a) The noise distribution **(b) The trigger comparison**

Fig. 4: Comparison between trigger modeling methods.

where $\frac{1}{n}\sum_n I'$ indicates that n consecutive frames are averaged to mitigate the effects of temporal image noise. I and I' are the consecutive frames captured by the camera before and after the laser signal injection, respectively. Here, Equation 3 models the noise η via the differential between consecutive frames. Intuitively, by progressive sampling different frame pairs in different environmental settings, we can obtain some noises that look more "natural" (see experiments in Section V-C) and apply them to further optimize our adversarial patches and triggers (discussed in Section V-D). Moreover, such a differential approximation is friendly to implement because the two camera frames can be easily captured when the attacker tests attacks in his own space.

Putting all together, the overall loss of our attack can be formalized in the following format.

$$\begin{aligned} arg \min_{\delta,t} & \mathbb{E}_{x,\eta}[\mathcal{L}_*(x,\delta,t+\eta)], \\ s.t.\ t \in & \{\mathcal{S}(p,d,\theta,l) \mid p,d,\theta,l \in P,D,\Theta,L\}. \end{aligned} \tag{4}$$

B. Attack Module Overview

As illustrated in Figure 3, the detailed design of L-Hawk is composed of three modules.

1) *Trigger Modeling.* We approximate real-world noises by evaluating the differences between continuous camera frames. To enhance the attack robustness, we construct a trigger space by using a progressive sampling-based method to extend the trigger under various parameters.
2) *Joint Optimization.* We propose an asynchronous learning method for L-Hawk to achieve multi-objective patch and trigger optimization. We further introduce a multi-angle and multi-position simulation mechanism to improve the robustness of L-Hawk and greatly reduce the difficulty of the laser attacker in targeting the camera.
3) *Attack Deployment.* We evaluate and deploy the four proposed attacks against object detectors and image classifiers in both the digital and physical worlds.

In the following subsection, we present the details of the first two modules. For the third module, we illustrate it with experiments in Section VI.

C. Trigger Modeling

L-Hawk is designed to become adversarial when triggered by color stripes caused by special laser signals but remain benign when not triggered. We define color stripes as triggers. To generate the effective L-Hawk, we first model the accurate trigger with noise in the digital world as shown in Figure 3(a). Recognizing the significant disparity between real and simulated triggers due to random noise caused by lens scattering, we propose approximating real-world noise by evaluating the differences to extract the noise. Our approach is based on the fact that the pixel intensity value represents the number of photons detected by the camera [32]. The approximate method is illustrated in Equation 3. However, laser signals can cause certain channels to reach their maximum pixel value, resulting in saturation. Saturation impedes the accurate representation of the laser's true effect on the image. To address this issue, we place a black cloth in front of the camera lens as the background for the trigger, without affecting the laser irradiation of the camera. The low RGB pixel values of the black area help prevent saturation in the RGB channels.

Based on the approximate noise η, we achieve the following objective:

$$\frac{1}{3}\sum_{i\in R,G,B}|I' - Clip(I+t+\eta)|_i \sim 0 \tag{5}$$

which represents that apart from the tiny temporal image noise of cameras, our simulated triggers are virtually indistinct from real triggers. $Clip(\cdot)$ means to limit the pixel value from 0 to 255. Figure 4 presents a comparison between the triggers generated by our method and the one produced by the existing simulation technique [15]. We evaluate the pixel accuracy using the mean square error (MSE) between the real image and the image with the trigger. The left result illustrates that the noise caused by camera lens scattering is very different from the normal distribution noise. Results in the right demonstrate that the trigger generated by our method closely matches the real trigger, both visually and in terms of pixel accuracy. In contrast, the trigger generated by the existing simulation method shows substantial discrepancies from the real trigger. Our approach circumvents the limitations of simulating physical noises involved in camera and attack modeling, thereby improving the accuracy of the simulated triggers. By focusing on the real trigger, attackers can effectively leverage the physical characteristics of the laser signals to create robust adversarial patches.

In addition, to maximize the attack performance, we initialize a set of triggers instead of just one. A trigger space is constructed by extending the trigger and noise under different attack and scenario parameters. In this trigger space, there is a one-to-one correspondence between the trigger and the parameter, which helps us find the attack parameter in the physical world based on the trigger (discussed in Appendix B). Note that the pixel height of the extracted color stripe is larger than that of the real attack, which is also to obtain more laser interference influence, thereby enriching the space of triggers.

D. Joint Optimization

To obtain the optimal patch and trigger, we propose an asynchronous joint optimization method as detailed in Figure 3(b). This method mainly proceeds with two steps.

```
Algorithm 1: Joint Optimization
Input: the training data $x$; the detector or classifier $\mathcal{F}$;
       the training epoch $N$; the initial adversarial
       patch $\delta_0$; the initial parameter distribution
       $P_0, D_0, \Theta_0, L_0$; the trigger generation function
       $\mathcal{S}(\cdot)$; the noise $\eta$ .
Output: the optimal trigger $t$; the optimal adversarial
        patch $\delta$.
$\delta \leftarrow \delta_0$;
Initialize $T = \{\mathcal{S}(p, d, \theta, l) | p, d, \theta, l \in P_0, D_0, \Theta_0, L_0\}$;
for $[1, N]$ do
    $t \in T$;
    Compute loss with Equation 6;
    Optimize $\delta$ with Equation 7 ;
    Compute loss with Equation 9;
    Optimize $t$ Trigger with Equation 10 ;
    $P, D, \Theta, L \leftarrow t$;
    $T = \{\mathcal{S}(p, d, \theta, l) | p, d, \theta, l \in P, D, \Theta, L\}$ ;
end
return $\delta$, $t$;
```

In the first step, since the trigger in the trigger space is larger than the size of the training data, we randomly sample each trigger. Random sampling also simulates the trigger in multi-angle and multi-position situations. Then, we define the following loss function to optimize the patch and trigger.

$$\begin{aligned}\mathcal{L}_\delta(x, \delta, t_0 + \eta) =& \alpha\ell_{attack}(x, \delta, t_0 + \eta) + \beta\ell_{benign}(x, \delta) \\ &+ \lambda\ell_{tv}(\delta) + \mu\ell_{content}(\delta) + \xi\ell_{nps}(\delta)\end{aligned} \tag{6}$$

where t_0 indicates the trigger in the initial trigger space. ℓ_{attack} and ℓ_{benign} represent loss functions under triggered and benign scenarios, which are used to achieve the proposed four attacks. ℓ_{tv} is the total variation (TV) loss [33] which ensures that L-HAWK maintains a realistic appearance by smoothing out abrupt color transitions. $\ell_{content}$ in [10] is used to increase the stealthiness of L-HAWK by encouraging the patch to mimic the spatial structure and general content of natural images. ℓ_{nps} is the non-printability score (NPS) loss [33] to make colors in L-HAWK closer to the colors that can be printed by a common printer. The detailed definition of the above losses can be found in Appendix C. Hyperparameters, i.e., α, β, λ, μ, and ξ, are used to balance different loss components. Then, based on Equation 7, we obtain the optimized patch δ^*.

$$\begin{aligned}&arg \min_{\delta} \mathbb{E}_{x,\eta}[\mathcal{L}_\delta(x, \delta, t_0 + \eta)], \\ &s.t. \ t_0 \in \{\mathcal{S}(p, d, \theta, l) \mid p, d, \theta, l \in P_0, D_0, \Theta_0, L_0\}\end{aligned} \tag{7}$$

where P_0, D_0, Θ_0, and L_0 denote the initial parameters.

In the second step, we aim to optimize the parameters based on the patch δ^*. We present the trigger generation function:

$$t = \mathcal{S}(p, d, \theta, l) \tag{8}$$

where $\mathcal{S}(\cdot)$ is used to generate the special trigger based on the parameter input, which is illustrated in Appendix B. The loss function of optimizing parameters is defined in Equation 9.

$$\mathcal{L}_t(x, \delta^*, t + \eta) = \zeta\ell_{attack}(x, \delta^*, t + \eta) + \psi\ell_{benign}(x, \delta^*) \tag{9}$$

where ζ and ψ are used to balance different losses. The optimization objective is illustrated in Equation 10.

$$\begin{aligned}&arg \min_{t} \mathbb{E}_{x,\eta,\delta^*}[\mathcal{L}_t(x, \delta^*, t + \eta)], \\ &s.t. \ t \in \{\mathcal{S}(p, d, \theta, l) \mid p, d, \theta, l \in P_0, D_0, \Theta_0, L_0\}\end{aligned} \tag{10}$$

We obtain the optimal parameters, i.e., p^*, d^*, θ^*, and l^*.

Next, we repeat the above two steps until the maximum training epoch N. We present the joint optimization method in Algorithm 1. To ensure that L-HAWK is robust to variable attack scenarios, the optimization is re-initizalized with different parameters at each training epoch. For example, we obtain an optimized set of parameters where $p^* = 50$ mW, $d^* = 30$ m, $\theta^* = 15°$, and $l^* = 1000$ Lux. We extend this set of parameters and obtain a new setting $P^* = [40$ mW, 60 mW$]$, $D^* = [25$ m, 35 m$]$, $\Theta^* = [10°, 20°]$, and $L^* = [800$ Lux, 1200 Lux$]$. In the next round, we optimize L-HAWK based on the new parameters. In our evaluation, such a parameter extension strategy can improve the average attack success rate of L-HAWK from 59.4% to 94.4%.

VI. EVALUATION

A. Overview

We evaluate the attacks from three aspects: digital evaluation, physical evaluation in stationary setups, and physical evaluation in moving setups. We use the attack success rate (ASR) to evaluate the digital experiments. Furthermore, we use the highest ASR in n consecutive frames $f_{succ}^{max(n)}$ [6], [34] to evaluate the physical experiments in stationary and moving setups. The key results are highlighted as follows:

- In digital experiments, the proposed methods achieve an average ASR of 94.4%, while the baseline[2] only achieves an average ASR of 15.2%. Specifically, the overall ASRs reach 95.3%, 94.9%, and 96.9% for HA, CA, and TA-D against three object detectors. The attacker achieves overall ASRs of 83.1% for TA-C against eight image classifiers (discussed in Section VI-B).
- In stationary physical experiments, L-HAWK achieves average $f_{succ}^{max(150)}$ of 99.8%, 88.4%, and 90.3% for HA, CA, and TA-D against three object detectors and four victim cameras, respectively (discussed in Section VI-C).
- In moving-setup physical experiments, the average $f_{succ}^{max(50)}$ reaches 100.0%, 83.3%, and 68.7% for HA, CA, and TA-D against three object detectors. Compared to an average 59% ASR of TPatch at 7 m, L-HAWK achieves an average 91.9% ASR for HA and CA at an attack distance of 50 m. We extensively investigate the transferability and robustness of L-HAWK in the physical world (discussed in Section VI-D).

[2] We use the patch optimization method in TPatch [10] as the baseline, but do not include our proposed joint optimization and trigger modeling methods.

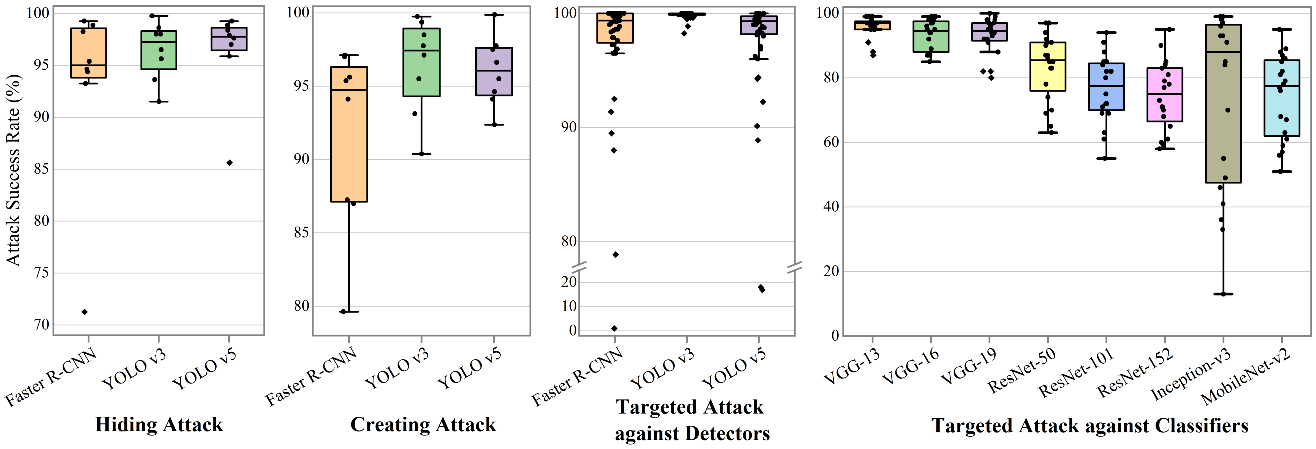


Fig. 5: The evaluation results of L-HAWK under the proposed four attacks in the digital world. The box plots show the min/max and quarterlies, and the dots represent different attack classes.

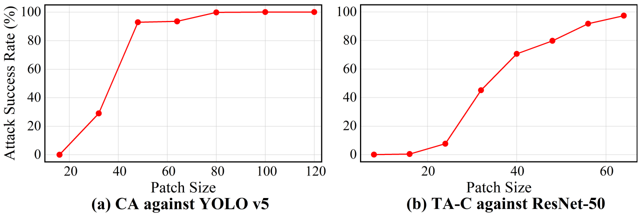


Fig. 6: The ASR under different patch sizes is evaluated. Each adversarial patch is square-shaped. The patch size is specified by the length of its side.

B. Digital Evaluation

In the digital evaluation, we use public datasets to evaluate the proposed four attacks, i.e., HA, CA, TA-D, and TA-C. We first evaluate each attack under various attack classes. Then, we investigate the impact of critical factors such as adversarial patch sizes and triggers on the effectiveness of L-HAWK. We also study the transferability of L-HAWK with two black-box attacks: single model attack and ensemble model attack. Finally, the improvement of the proposed method is discussed.

1) Experimental Setup: We first show the setup in the digital evaluation, including victim models, training and validation datasets, attack classes, and evaluation metrics.

Victim Models. We evaluate L-HAWK against three popular object detectors, including the one-stage YOLO V3/V5 and a two-stage Faster R-CNN. All object detectors are trained on the MS Common Objects in Context (COCO) datasets [35] for detection. We also evaluate L-HAWK on eight widely-used image classifiers, i.e., VGG-13/16/19, ResNet-50/101/152, Inception-v3, and MobileNet-v2, which cover models of different depths and architectures. All of these classifiers are trained on the training set of the large vision database ImageNet [36].

Datasets. For object detectors, we use the COCO validation set for optimizing the adversarial patch and trigger. We then utilize two popular autonomous driving datasets KITTI [37] with 1242×375 pixels and BDD100K [38] with 1280×720 pixels for evaluation. The images contained in these two datasets are captured in real driving scenarios. For the image classifier, we utilize the ImageNet validation set with 224×224 pixels for evaluation. A total of 10846 and 10000 images are used for detectors and classifiers in our evaluation experiments.

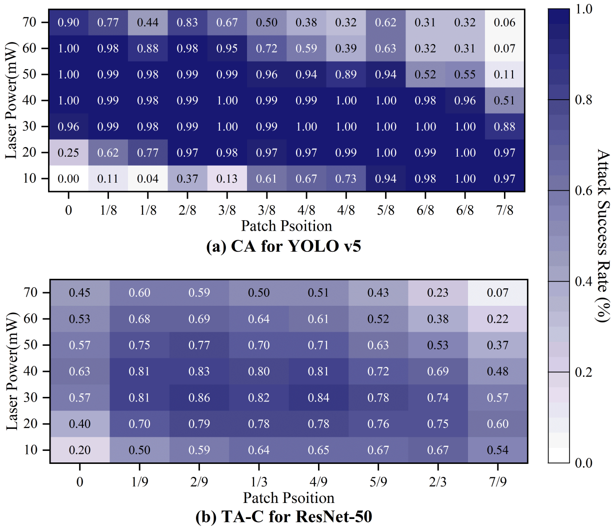
Attack Classes. We investigate 8 primary attack classes for our attacks against object detectors. 20 primary attack classes for image classifiers are selected. These classes are deemed security-critical in the context of autonomous driving scenarios [10], [39].

Metrics. We further define the ASR for our attack evaluation in detail. ASR indicates the ratio of the number of successful attacks against object detectors or image classifiers over the total number of conducted attacks. A successful attack is defined as being adversarial when triggered by laser signals, but benign when not triggered. The metric of ASR can be formulated as follows:

$$ASR = \frac{1}{N} \sum_{i=1}^{N} C_{F(x',t)=y_a \& F(x')=y_b}(x') \tag{11}$$

where N is the number of evaluation samples. $C(\cdot)$ is the counting function. $F(\cdot)$ denotes the recognition function of the victim model. x' is the image embedded with L-HAWK. t denotes the trigger caused by laser signals. y_a and y_b are the adversarial and benign labels, respectively. Unlike image classification, object detection outputs not only the class probability but also the bounding box of the detected object. The accuracy of the bounding box is usually judged using the Intersection-over-Union (IOU), and the threshold of the IOU is usually set to 0.5 [19].

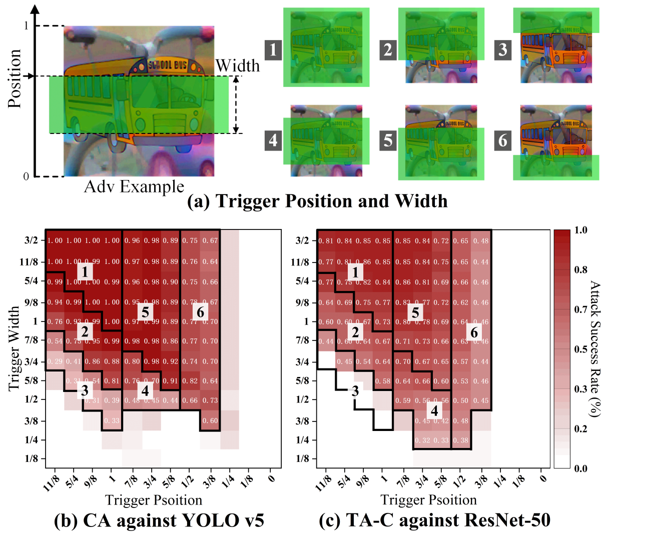
2) Overall Performance: In this section, we evaluate the effectiveness of our attacks on different object detectors and image classifiers, respectively. Before the evaluation, we construct a trigger space based on the attacks of a 532 nm wavelength green laser. The trigger size is 700×2880 pixels. Then, we acquire the optimal trigger and the adversarial patch



(a) CA for YOLO v5

(b) TA-C for ResNet-50

Fig. 7: The ASR of attacking YOLO V5 and ResNet-50 under the various laser power and patch positions.



(a) Trigger Position and Width

(b) CA against YOLO v5

(c) TA-C against ResNet-50

Fig. 8: The ASR of attacking YOLO V5 and ResNet-50 under the cover of triggers of various positions and widths. Illustrations of (a) the trigger's position and width compared with L-HAWK and four typical cases, (b) the ASR of CA against YOLO V5 under 6 cropped regions, and (c) the ASR of TA-C against ResNet-50 under 6 cropped regions.

TABLE I: The impact of various trigger colors.

Trigger Type \ Attack Type	HA	CA	TA-D	TA-C
Green	97.6%	95.5%	99.6%	85.0%
Red	90.8%	100%	94.9%	86.1%

for each attack based on the proposed optimization framework. During the sampling of triggers, the width of the trigger is the same as the width of the evaluation images. The height of the trigger is set to 50 pixels. In addition, the size of L-HAWK is set to $48 * 48$ pixels in the overall evaluation.

For each test involving CA, TA-D, and TA-C, L-HAWK is a standalone object and randomly patched on the image. For HA, L-HAWK is attached to each target object (e.g., a "stop sign") and then randomly patched on the image. A detailed definition of L-HAWK is shown in Figure 16. Then, we create a pair of images by applying the trigger which represents the benign and adversarial cases for detection or classification, respectively. For all attacks, the trigger covers the entire L-HAWK or target object to activate the adversarial effect. Following the metric formulated by Equation 11, we calculate the overall ASR.

The results of the overall performance evaluation are depicted in Figure 5. The overall ASRs reach 95.3%, 94.9%, 96.9%, and 83.1% for HA, CA, TA-D, and TA-C, respectively. For HA, CA, and TA-D, the attack performance against different object detectors is consistent. Specifically, the overall ASR for YOLO V3 is the highest, followed by the ASR for YOLO V5, and the lowest is the ASR for Faster R-CNN. For TA-C, the VGG series are the easiest to attack with ASR exceeding 94%, followed by ResNet with ASR over 75% and Inception with ASR over 74%, and the worst is MobileNet with ASR of 73.4%. The result shows that notable differences in ASRs are observed across various classes. For HA, 'person' is the most challenging to impact across all detectors. 'stop sign' is the easiest to hide, which is because the trigger produced by the green laser largely destroys the red color feature of the stop sign. In contrast to HA, in CA, 'person' is the easiest to affect across all detectors. For TA-D, we find that the attack effect is bad against Faster R-CNN and YOLO V5 when the class after triggering is 'stop sign'. Especially, the ASR is only 1% on Faster R-CNN when the benign class is 'motorcycle' and the adversarial class is 'stop sign'. In addition, in the Inception v3 experiments, the most vulnerable class is 'traffic light,' achieving ASR exceeding 98%. Conversely, the 'plow' class exhibits the lowest vulnerability, with an ASR of just 13%, indicating minimal adversarial impact.

3) Impact of Other Factors: We investigate several factors that may influence the attack effectiveness of L-HAWK, including the size and position of the adversarial patch, as well as the pixel intensity, color, and size of the trigger. To simplify our analysis, we focus on a selection of representative perception models. For the image classifier analysis, we use ResNet-50, which demonstrates medium overall performance among the eight models we evaluated, and the target class for TA-C is the 'traffic light'. For the object detector analysis, we select YOLO V5, one of the latest models, and focus on the 'stop sign' class for CA.

Impact of Adversarial Patch Size. The size of the adversarial patch reflects the attacker's ability to influence the input image. A smaller size suggests a greater distance between the camera and the adversarial patch, and vice versa. As illustrated in Figure 6, adversarial patches that are too small rarely succeed in executing a successful attack. For image classifiers, the success rate curve shows a gradual increase, achieving nearly 100% success when the size reaches 64 × 64 pixels.

TABLE II: The transferability of attacking object detectors.

Atack Type	HA			CA			TA-D		
Black Detector White Detector	Faster R-CNN	YOLO v3	YOLO v5	Faster R-CNN	YOLO v3	YOLO v5	Faster R-CNN	YOLO v3	YOLO v5
Faster R-CNN	96.6%	59.3%	27.4%	97.0%	93.4%	68.8%	100.0%	94.0%	45.6%
YOLO v3	56.3%	98.6%	97.6%	6.1%	99.3%	72.4%	0.12%	100.0%	21.9%
YOLO v5	89.0%	63.6%	99.9%	10.4%	99.6%	98.4%	48.0%	99.4%	99.6%

TABLE III: The transferability of attacking image classifiers.

Black Classifier White Classifier	VGG-13	VGG-16	VGG-19	ResNet-50	ResNet-101	ResNet-152	Inception-v3	MobileNet-v2
VGG-ens	94.9%	97.1%	99.9%	58.0%	44.0%	62.9%	35.5%	40.3%
ResNet-ens	14.9%	50.1%	57.4%	95.8%	96.6%	93.6%	22.4%	56.8%

TABLE IV: Our improvement for various attacks.

Method	HA	CA	TA-D	TA-C
Patch optimization in TPatch [10]	10.7%	0.5%	14.4%	35.1%
Our joint optimization	36.9%	25.1%	42.8%	62.3%
Our joint optimization & trigger modeling	97.6%	95.5%	99.6%	85%

In contrast, the ASR for YOLO V5 exhibits two significant increases at sizes of 48×48 pixels and 80×80 pixels. This behavior is attributable to the YOLO V5's three-scale prediction mechanism, which becomes more susceptible to misclassification as the size increases, making the adversarial patch more detectable by the detector.

Impact of Patch Position and Pixel Intensity of Trigger. Lasers of different power levels produce triggers with varying pixel intensities, and the pixel intensity at different positions on the trigger also varies. Therefore, we aim to explore the impact of various triggers and patch positions. Specifically, we extend the laser power from the trigger space used for overall performance evaluation with a power step of 10 mW. Other experiment setups remain unchanged in the overall experiment. Two heatmaps in Figure 7 show the ASRs of attacking YOLO V5 and ResNet-50 under various laser powers and patch positions. The x-axis indicates the lateral position of L-Hawk relative to the left edge of the color stripe, varying from 0 to $7/8$ for YOLO V5 and from 0 to $7/9$ for ResNet-50. The y-axis is the laser power, which varies from 10 mW to 70 mW. From the results, we find that the responses of the detector and classifier to laser power and patch position are consistent. For the optimal trigger, i.e., the laser power of about 29 mW, L-Hawk has high ASRs in any position. Since the incidence angle of the optimal trigger is about $18°$, the trigger's pixel intensity distribution behaves as becoming stronger from left to right. When $p > 30$ mW, the pixel intensity on the right side of the trigger is too strong, resulting in poor trigger ability. This performance is reversed when $p < 30$ mW, which is because the pixel intensity in the left part of the trigger is too low to exhibit triggering ability. In summary, too strong or too weak pixel intensity is not conducive to trigger L-Hawk.

Impact of Trigger Position and Width. Since the adversarial patch taken at a long distance is small, this may cause the trigger to not accurately cover all of the adversarial patch, further leading to trigger failure. Thus, we conduct experiments to explore the impact of the trigger's position and width on attacks. Figure 8(a) visually illustrates the trigger's position and width. Position "0" indicates that the upper edge of the trigger aligns with the bottom edge of L-Hawk, whereas position "1" denotes that the upper edge of the trigger aligns with the top edge of L-Hawk. To interpret the results shown in Figure 8(b) and Figure 8(c), we divide the heatmaps into six regions corresponding to the typical cases depicted in Figure 8(a). In case **1**, the trigger covers the entire L-Hawk; in case **2** and **5**, the trigger covers more than $1/2$ lower or upper part of L-Hawk but never covers the entire L-Hawk, including the lower or upper edge of L-Hawk; in case **3** and **6**, the trigger covers less than $1/2$ lower or upper part of L-Hawk but includes the lower or upper edge of L-Hawk; in case **4**, the trigger covers only a part of L-Hawk between its upper and lower edges. The results presented in Figure 8(b) and Figure 8(c) indicate for the detector and classifier the smaller the area of L-Hawk covered by the trigger, the worse the attack effect. However, when the trigger is unable to fully cover the entire L-Hawk, we can still achieve an average 42% and 32.1% ASRs against YOLO V5 and ResNet-50.

Impact of Trigger Color. In addition to the green laser, we utilize a laser with a wavelength of 650 nm to generate red triggers. The other experiment setups are consistent with the overall evaluation experiment. We set the benign class to 'stop sign' for HA. The benign class is 'person' and the adversarial class is 'stop sign' for TA-D. Then, HA, CA, and TA-D are conducted against YOLO V5, and TA-C is conducted against ResNet-50. As shown in Table I, the average ASR for the red trigger reaches 92.9% and for the green trigger is 94.4%. The result also shows that HA, CA, and TA-D attacks behave differently on different color triggers, such as the green trigger easily hides a 'stop sign', and the red trigger easily creates a 'stop sign'. This may be because the detector is more likely to associate red with a 'stop sign' and more difficult to associate green with a 'stop sign'.

4) Transferability Study: When the attacker has limited prior knowledge of the DNN models used in vision-based perception modules, it is impractical to apply a gradient-based optimization approach directly to these black-box models. However, the attacker can potentially evade the target model by

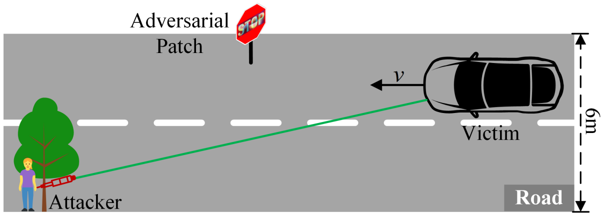


Fig. 9: A bird's eye view of the physical attack setup.

leveraging the transferability of adversarial patches across similar DNN models. To evaluate the transferability of L-HAWK, we perform a single-model attack on object detectors and an ensemble-model attack on image classifiers. The setup of L-HAWK, such as the target class, adversarial patch size, and trigger settings, remains fixed. We only vary the recognition across different models to assess the transferability.

Table II summarizes the transfer attack results across various detectors. The result shows that, in all the experiments, the average ASR of transfer attacks between YOLO V3 and YOLO V5 is relatively high (73.1%) because of their similar model architectures. In contrast, the average ASR of transfer attacks between one-stage detectors (YOLO series) and two-stage detectors (Faster R-CNN) is only 50%. Table III lists the detailed ASRs of ensemble transfer attacks on different ensemble classifiers. Specifically, we create two ensemble models: VGG-ens (VGG-13+VGG-16+VGG-19) and ResNet-ens (Res50+Res101+Res152) to optimize L-HAWK for transfer attacks on the other five models. The result shows that the average ASR of transfer attacks with the same architecture is relatively higher (96.3%), than the average ASR of transfer attacks with different architectures (44.2%). In conclusion, we can achieve an average ASR above 96% and 44% for white-box attacks and transfer attacks.

5) Our Improvements: In this section, we compare our method with the state-of-the-art attack approach under four attacks. The baseline is the patch optimization method in TPatch [10], which optimizes the patch only, without the optimization of the trigger. We then compare the improvement of the joint optimization method and the entire scheme (i.e., consisting of joint optimization and trigger modeling) respectively. Note that we only change the patch optimization process, and the evaluation is done under the fixed trigger with random noise. As reported in Table IV, 1) the joint optimization method improves the average ASR from 15.2% to 41.8%; 2) the entire scheme improves the average ASR from 15.2% to 94.4%. Our method achieves the highest performance against all attacks. The results also show that the noise caused by lens scattering has a significant influence on the triggering of L-HAWK.

C. Physical World Evaluation in Stationary Setups

In this section, we validate HA, CA, and TA-D against four real cameras and three object detectors in the physical world with stationary setups. Then, we investigate the impact of various factors on attacks, including the positions of L-HAWK, and attack parameters of lasers. Finally, we evaluate TA-C against three image classifiers in the physical world.

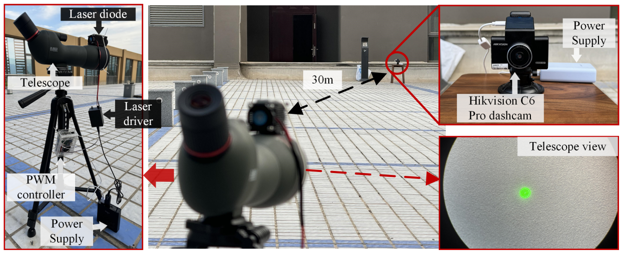


Fig. 10: Overview of the attack equipment. On the left is the tracking and laser aiming equipment. On the right is the victim camera and the laser speck observed by the telescope.

1) Experimental Setup: The experimental setup is shown in Figure 9. The victim vehicle remains stationary and is equipped with four real cameras, i.e., a Hikvision C6 Pro dashcam [31] (Camera 1), a Logitech C920 PRO HD Webcam [40] (Camera 2), an Intel RealSense Depth Camera D435i [41] (Camera 3), and an iPhone XR smartphone (Camera 4). The detailed information of four cameras is listed in Table XI. Note that we only utilize the monocular imaging function of the Intel RealSense Depth Camera. The attacker is about 20 m from the adversarial patch and 30 m from the victim vehicle. Then, we use a 532 nm green laser diode with a maximum power of 200 mW. An overview of the attack equipment is shown in Figure 10. Since each camera has a different sensitivity to the laser, we optimize the trigger and the corresponding L-HAWK for each camera. We present the specific patches used in the physical evaluation in Figure 16. The realistic size of L-HAWK is 60 cm ×60, which is also the minimum size used in previous works [8], [10], [42]. Finally, for each experiment, we capture a video of around 10 s and the number of frames is about 300 at an fps of 30.

Metrics. To gain a better comprehension of the effectiveness of the L-HAWK, we evaluate benign scenarios and triggered scenarios separately, utilizing the best attack success rate (i.e., $\frac{1}{n}\sum_{i=1}^{n} C_{F(x_i)=y_a}(x_i)$) within captured 300 consecutive frames, denoted as $f_{succ}^{max(n)}$ [6], [10], [34], which is formulated as follows:

$$f_{succ}^{max(n)} = \max_{j} \frac{1}{n} \sum_{i=1}^{n} C_{F(x_{i+j})=y_a}(x_{i+j}) \tag{12}$$

where n is the number of consecutive frames used for evaluation and is set to 150. x_i is the i_{th} frame of the video x.

2) Overall Performance: To evaluate the overall performance, we generate 9 L-HAWK (3 for HA, 3 for CA, and 3 for TA-D) against three detectors for each camera, i.e., a total of 36 adversarial patches. Then, we utilize $f_{succ}^{max(150)}$ to evaluate the result under the benign scenario (i.e., no trigger scenario) and trigger scenario. The overall result is illustrated in Table V. The average $f_{succ}^{max(150)}$ are 99.8%, 88.4%, and 90.3% for HA, CA, and TA-D, respectively. For different cameras, the highest average $f_{succ}^{max(150)}$ is 99% for Camera 2, the average $f_{succ}^{max(150)}$

TABLE V: The overall performance of our attacks against various cameras.

Attack Type	Target Model	Victim Camera								Avg.
		Hikvision C6 Pro		Logitech C920 PRO		Intel RealSense D435i		iPhone XR		
		No Trigger	Trigger	No Trigger	Trigger	No Trigger	Trigger	No Trigger	Trigger	
HA	**Faster R-CNN**	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	99.8%
	YOLO v3	100.0%	100.0%	100.0%	96.7%	100.0%	100.0%	100.0%	98.0%	
	YOLO v5	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	
CA	**Faster R-CNN**	96.7%	82.7%	100.0%	97.3%	92.0%	61.3%	100.0%	100.0%	88.4%
	YOLO v3	100.0%	97.3%	100.0%	92.0%	100.0%	81.3%	93.3%	23.3%	
	YOLO v5	100.0%	93.3%	100.0%	98.0%	100.0%	78.0%	90.6%	44.0%	
TA-D	**Faster R-CNN**	98.7%	84.6%	100.0%	100.0%	81.3%	70.6%	85.3%	64.7%	90.3%
	YOLO v3	100.0%	100.0%	100.0%	97.8%	100.0%	96.7%	97.3%	50.0%	
	YOLO v5	100.0%	98.0%	100.0%	100.0%	100.0%	78.7%	92.0%	71.3%	

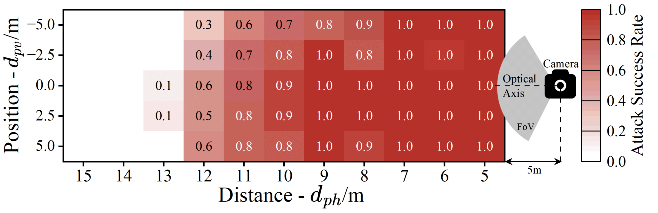
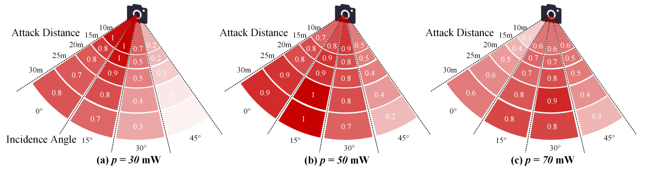


Fig. 11: The ASR under various positions of L-HAWK. The x-axis and y-axis represent d_{ph} and d_{pv}, respectively. A value of d_{pv} of 0 means that the adversarial patch is on the optical axis of the camera.



(a) p = *30* mW

(b) p = *50* mW

(c) p = *70* mW

Fig. 12: Attack results at different attack distances and incidence angles at 3 laser power (30 mW, 50 mW, and 70 mW).

are 97.3%, 91.1% for Camera 1 and Camera 3, and the worst average $f_{succ}^{max(150)}$ is 83.9% for Camera 4, respectively. The difference in the above results is due to the variation in pixel intensity of the trigger caused by the same laser on different cameras. For example, the trigger in Camera 4 has a strong pixel intensity, which increases the ability to hide an object but decreases the ability to create an object. In addition, we find that $f_{succ}^{max(150)}$ is better in the benign scenario (average 98%) than in the triggered scenario (average 87.7%), which may be due to the influence of the trigger on the performance of the target object detection task. Therefore, in the physical world, the appropriate trigger is very important for attacks.

3) Impact of Patch Position.: For simplicity, we utilize a representative camera (Camera 1, the mainstream dashcam in vehicles) for investigating the impact of the patch position in the physical world. During the evaluation, CA with class 'stop sign' is conducted against YOLO V5 under various d_{pv} and d_{ph}. Considering that $f_{succ}^{max(150)}$ in benign scenarios is high (average 99.3%) and there is no difference in the overall evaluation, we only select $f_{succ}^{max(150)}$ in trigger scenarios as the evaluation metrics. The other experiment setups are the same as the overall evaluation in the stationary setups. We set three typical values for d_{pv}, i.e., -5 m, -2.5 m, 0 m, 2.5 m, and 5 m, which represents the general situation of setting up L-HAWK on the roadside. Then, d_{ph} is set from 3 m to 15 m. We choose not to start at 1 m because the camera cannot capture the entire L-HAWK when $d_{ph} < 2$ m and $d_{pv} = 5$ m. Figure 11 demonstrates the overall attack performance, where the result for d_{ph} from 3 m to 4 m is removed because of the average ASR of 100%. We find that, as d_{ph} increases, the attack success rate decreases gradually. There is basically no attack effect when $d_{ph} > 12$ m, which is because that the size of L-HAWK captured by Camera 1 when $d_{ph} = 13$ m is only about 31×31 pixels. In addition, when $\theta = 15°$, the pixel intensity on the right side of the trigger is high, and the pixel intensity on the left side is low, which has a slight impact on the attack. It is because that stronger pixel intensity makes it easier to hide objects rather than make them appear.

4) Impact of Trigger Pixel Intensity: To investigate the influence of pixel intensity on attack performance, we conduct experiments under various laser power p, attack distance d, and incidence angle θ. The experiment setups are the same as that used in the evaluation of the patch position, but we keep the position of L-HAWK the same, i.e., $d_{pv} = 2.5$ m and $d_{ph} = 10$ m. Then, at three different laser powers (30 mW, 50 mW, and 70 mW), we test the attack effect under d from 10 m to 30 m by step of 5 m and θ from 0° to 45° by step of 15°. Note that we choose the maximum attack distance of 30 m because of the limitations of the test field (a rooftop). In fact, the attacker can achieve a longer-distance attack by adjusting the laser power. In Section VI-D, we achieve effective attacks at a distance of about 50 m.

The results, shown in Figure 12, indicate that L-HAWK is sensitive to the pixel intensity of the trigger and performs best in a range of pixel intensities. For example, the overall average $f_{succ}^{max(150)}$ are 59.7%, 77.9%, and 64.1% at $p = 30$, 50, and 70 mW, which means that triggers generated by 50 mW of laser power are more likely to trigger L-HAWK. The results also show that under different attack distances and incidence angles, setting the appropriate laser power has a certain influence on the attack. According to the optimal trigger and the results, it is recommended to set p to 50 mW for a long distance attack (i.e., $d > 20$ m), 40 mW for a distance from 15 m to 20 m, and finally 30 mW for a close

TABLE VI: The overall performance of L-HAWK in the physical world across different object detectors under moving setups.

Attack Type		HA			CA			TA-D		
White Model	**Black Model**	$f_{succ}^{max(50)}$	$f_{succ}^{max(100)}$	$f_{succ}^{max(150)}$	$f_{succ}^{max(50)}$	$f_{succ}^{max(100)}$	$f_{succ}^{max(150)}$	$f_{succ}^{max(50)}$	$f_{succ}^{max(100)}$	$f_{succ}^{max(150)}$
Faster R-CNN	**Faster R-CNN**	100.0%	100.0%	96.7%	50.0%	25.0%	16.7%	6.0%	3.0%	2.0%
	YOLO v3	76.0%	62.0%	56.7%	20.0%	11.0%	7.3%	12.0%	6.0%	4.0%
	YOLO v5	98.0%	81.0%	73.3%	26.0%	14.0%	9.3%	2.0%	1.0%	0.7%
YOLO v3	**Faster R-CNN**	100.0%	100.0%	100.0%	0.0%	0.0%	0.0%	46.0%	23.0%	15.3%
	YOLO v3	100.0%	100.0%	100.0%	100.0%	68.0%	45.3%	100.0%	100.0%	68.0%
	YOLO v5	100.0%	100.0%	100.0%	14.0%	7.0%	5.3%	24.0%	12.0%	8.7%
YOLO v5	**Faster R-CNN**	100.0%	100.0%	94.7%	0.0%	0.0%	0.0%	46.0%	23.0%	15.3%
	YOLO v3	68.0%	64.0%	57.3%	100.0%	93.0%	78.7%	100.0%	63.0%	44.0%
	YOLO v5	100.0%	100.0%	97.3%	100.0%	100.0%	71.3%	100.0%	50.0%	33.3%

range attack (i.e., $d < 50$ m) for the best effect. In conclusion, as long as the laser power is large enough, we can achieve effective attacks at long distances.

In addition, we find that the attack is more influenced by the incidence angle than by its attack distance from the camera. When d is from 10 m to 30 m, the overall average $f_{succ}^{max(150)}$ is 69.9%, 68.6%, 67.5%, 66.7%, and 63.3%. However, when θ is from 0° to 45°, the overall average $f_{succ}^{max(150)}$ is 75.5%, 87.2%, 71.0%, and 35.2%. Successful attacks predominantly occurred within 30° off-axis, across d from 10 m to 30 m. This effect is likely due to the pixel intensity of the trigger being too fainter when $\theta > 30°$.

5) Targeted Attacks Against Image Classifiers: We conducted indoor physical experiments to evaluate TA-C. Specifically, we placed the L-HAWK with a mouse (an existing class in ImageNet [36]) in the camera's view. When there is no laser attack, the image classifiers correctly recognize the mouse, and the target class, traffic light, is not within the top-5 recognition results. After laser signal attacks, the top-1 recognition result is altered to the traffic light. L-HAWK achieved an average $f_{succ}^{max(150)}$ of 100% against VGG16, Res50, and Incv3. The possible reason is that the green color stripe caused by the laser not only destroys the feature robustness of the mouse but also has a distinct traffic light feature.

D. Physical World Evaluation in Moving Setups

In this section, we further evaluate HA, CA, and TA-D against a moving vehicle under both white-box and black-box settings. We also investigate the influence of lighting conditions and vehicle speeds on the attack performance. Finally, we evaluate L-HAWK against an end-to-end object detection system in an autonomous robot platform.

1) Experimental Setups: The experiments are conducted under the same scenario shown in Figure 9. The attacker is about 35 m from L-HAWK and 50 m from the victim vehicle. Then, researchers drive the vehicle equipped with Camera 1 toward L-HAWK at a speed of 5 km/h, and the distance from L-HAWK ranges from 15 m to 1 m. The laser power is set to 70 mW empirically based on the optimal trigger. The ambient light intensity is about 511 Lux. Finally, the attacker captures the video of around 10 seconds with about 300 consecutive frames at an fps of 30.

Metrics. We also utilize the best attack success rate in captured 300 consecutive frames, i.e., $f_{succ}^{max(n)}$ formulated in Equation 12, to evaluate the L-HAWK in moving setups. To quantify our evaluation, we choose three frame lengths n 50, 100, 150, i.e., $f_{succ}^{max(50)}$, $f_{succ}^{max(100)}$, and $f_{succ}^{max(150)}$.

2) Overall Performance: To evaluate the overall performance, we generate 9 L-HAWK (3 for HA, 3 for CA, and 3 for TA-D) against three detectors for Camera 1. Each L-HAWK is specifically trained against a particular detector using the proposed framework. To evaluate the transferability of these L-HAWK, we tested them on the same video clips with other black-box detectors.

We illustrate the overall performance of three attacks in Tab.VI. Since the average ASR reaches 99.1% for all attacks under the benign scenario, we only show the results under the triggered scenario. The result indicates that the transfer attacks can be implemented in the physical world. For different target detectors, the attack performance in the white box setting is consistent with that in the digital evaluation. Specifically, the average ASR for YOLO V3 is the highest and reaches 86.8%, followed by the average ASR for YOLO V5 is 83.5%, and the worst average ASR is 44.4% for Faster R-CNN. In transferred attacks, the average ASRs are 58.2%, 42.0%, and 31.1% for YOLO V5, YOLO V3, and Faster R-CNN. Although some results are not very ideal, at least one surrogate model (for any victim model) can achieve a transferred ASR over 42%. To our best knowledge, there has been no work that can perfectly transfer physical-world attacks from one surrogate model to all others. We will investigate this further in future works. For different attacks, HA has better performance (the average $f_{succ}^{max(50)}$ of 93.6%) than CA (the average $f_{succ}^{max(50)}$ of 45.6%) and TA-D (the average $f_{succ}^{max(50)}$ of 48.4%), which is because the green trigger breaks the robustness of the 'stop sign', making it easier to blind the detector and thus hide the object. Furthermore, the average $f_{succ}^{max(50)}$ for HA and CA reaches 91.9% at an attack distance of 50 m, while the average $f_{succ}^{max(50)}$ for HA and CA in TPatch [10] is only 59% at a short attack distance of 7m.

3) Impact of Other Factors: We further investigate the impact of ambient light and vehicle speeds.

Impact of Ambient Light. Lighting conditions can alter the pixel intensity of triggers, potentially affecting the effectiveness of L-HAWK. To investigate this, we tested L-HAWK under three different lighting conditions: daytime (2237 Lux), dusk (758 lux), and backlight (119 lux). As shown in Figure 13, L-HAWK maintains robustness across all lighting

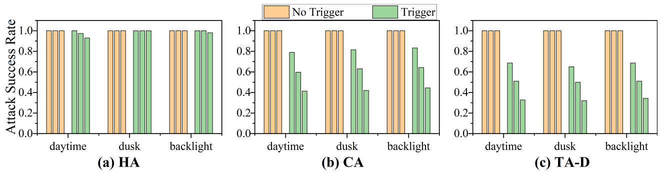


Fig. 13: Impact of light conditions on HA, CA, and TA-D. The three adjacent bars denote $f_{succ}^{max(50)}$, $f_{succ}^{max(100)}$, and $f_{succ}^{max(150)}$.

conditions, achieving average ASRs of 98.7%, 62.1%, and 50.4% for HA, CA, and TA-D under the triggered scenario. This robustness is attributed to two factors. First, according to the attacker capability study in Section IV, the three lighting conditions in the experimental setting have little effect on the trigger pixel intensity. Furthermore, we can choose the appropriate laser power for different lighting conditions according to the optimal trigger. In addition, due to the comprehensive EoT process for L-Hawk, which incorporates potential color shifts in brightness, contrast, saturation, and hue, it ensures consistent performance despite variations in lighting.

Impact of Vehicle Speeds. To investigate the impact of vehicle's movements, we extend evaluations across various speeds from 10 km/h to 50 km/h, covering safe driving speeds in cities. Since fewer frames are captured due to faster vehicle speeds, the metric of Equation 12 cannot be directly applied. Therefore, we directly utilize the attack success rate, i.e., $\frac{1}{m}\sum_{i=1}^{m} C_{F(x_i)=y_a}(x_i)$ to evaluate our attack under various speeds. m is the number of valid frames captured when the vehicle travels from 15 m to 1 m from the adversarial patch. The results are illustrated in Table VII. We present the number of valid frames and the number of successful frames for the attack. As a result, we achieve an average attack success rate of 93.2% for HA, 67.2% for CA, and 51.6% for TA-D across all speeds. The results also show that the attack success rate is affected when the speed is increased. This is because the average aiming success rate of the laser is reduced, resulting in an increase in the possibility of L-Hawk triggering failure. But we can also achieve an average attack success rate of 56% across all attacks at 50 km/h.

4) End-to-End Evaluation: We further investigate the potential impact of L-Hawk on end-to-end autonomous robots. The TurBot3-ARM [43], an autonomous robot platform, provides a black-box object detection pipeline (including data preprocessing, DNN models, and decision modules). An autonomous driving task is deployed to the platform, which controls the platform to stop automatically when recognizing a stop sign. Then, we achieve an average ASR above 80% for HA, CA, and TA-D. The attack demo is available at https://github.com/Jupiterliu/L-Hawk.

E. Further Study of Laser Attacks.

In this section, we further investigate the relationship between attack and scenario parameters and the color stripe's brightness, which helps us to conduct robust physical attacks.

TABLE VII: The performance of L-Hawk in the physical world evaluation under various speeds.

Speed (Valid Frames)	10km/h (150)	20km/h (75)	30km/h (50)	40km/h (30)	50km/h (25)
HA	100%(150)	100%(75)	92.0%(46)	90.0%(27)	84.0%(21)
CA	85.3%(128)	76.0%(57)	66.0%(33)	56.7%(17)	52.0%(13)
TA-D	69.3%(104)	61.3%(46)	52.0%(26)	43.3%(14)	32.0%(8)

We mainly consider four parameters, i.e., p, d, θ, and l. To intuitively demonstrate the brightness, we calculate the pixel intensity of color stripes [44]. Specifically, we first extract the color stripe by calculating the difference between the normal image I and the image I' after laser signal attacks. Then, we calculate the pixel intensity of the extracted color stripe.

Specifically, on the basis of stationary setups, we utilize a Hikvision C6 Pro dashcam [31] to act as the victim camera. A 532 nm green laser diode with a maximum power of 200 mW is used for experiments. The ambient light intensity is about 1600 Lux. To simplify the study, we keep the victim camera stationary, i.e., $v = 0$ km/h.

Impact of Laser Power and Attack Distance. On the basis of the attack setup, we set θ to 15°. Then, we calculate the pixel intensity of the color stripe under d from 5 m to 30 m by a step of 5 m and p from 10 mW to 66 mW by a step of 4 mW. In addition, when the attack distance continues to change, the attacker can empirically adjust the power to produce the color stripe with appropriate pixel intensity. Detailed result is illustrated in Figure 14(a).

Impact of Ambient Light Intensity. On the basis of the experiment setup, we explore the pixel intensity variation under three different laser powers (i.e., 10 mW, 30 mW, and 50 mW) and five ambient light intensities (i.e., 1238 Lux, 719 Lux, 461 Lux, 198 Lux, and 0 Lux). The attacker can also empirically adjust the laser power based on the measured ambient light intensity to generate the color stripe with appropriate pixel intensity. Detailed result is illustrated in Figure 14(b).

Impact of Incidence Angle. On the basis of the attack setup, we test the pixel intensity variation by adjusting θ from $-60°$ to $0°$ by a step of $15°$. Results in Figure 14(c) show that the horizontal distribution of pixel intensity under different incident angles varies significantly. Moreover, the number of laser photons captured by CMOS decreases due to oblique injection, causing the reduced pixel intensity.

VII. Ethical Discussion

A. Ethical Concerns

To prevent any harm to real-world systems or infrastructure and comply with ethical and safety standards, we are glad to collaborate with the safety committee and take every precaution in our research. First, like all prior work [14], [15], [45], our experiments are conducted in a strictly controlled environment, with no interaction with public traffic or roadways. Second, we have responsibly disclosed the identified security vulnerability to the relevant vendors and can provide any detailed technical information. To reduce the risk of misuse, we

TABLE VIII: The evaluation of various defenses. mAP (mean average precision) denotes the model performance.

Defense Method	Before Defense		After Defense	
	ASR	mAP	ASR	mAP
Adversarial Training [46]	94.4%	45.4	41.6%	34.5
Input-Transformation [47]	94.4%	45.4	68.6%	28.8

TABLE IX: The adversarial patch detection accuracy.

Defense Strategy	SentiNet [11]	PatchGuard [12]	PatchCleanser [13]
without activating patch	2.5%	4.7%	3.8%
with activating patch	99.4%	100.0%	99.0%

conditionally release the code to trusted parties. Specifically, we only expose the available portion of the implementation in the open-source repository (i.e., the digital patch optimization and evaluation). Meanwhile, we also provide our contact information for those interested in accessing the full code (such as laser parameters and operations). Finally, we discuss potential countermeasures that manufacturers and developers can implement to safeguard against our attacks.

B. Countermeasures

To mitigate the threat posed by L-Hawk, we discuss four types of potential countermeasures:

Adversarial Training and Input Transformation-Based Method. Adversarial training [46] or the input transformation-based method [47] can improve the robustness of victim models against L-Hawk. However, our evaluation in Table VIII shows that such a method trades off model performance for security. As a result, these countermeasures only work in those scenarios that do not require high model performance.

Adversarial Patch Detection. Several adversarial patch detection methods are also proposed, such as SentiNet [11], PatchGuard [12], and PatchCleanser [13]. We evaluate patch detection accuracy with and without activating L-Hawk. The results in Table IX show that L-Hawk can be detected once it is activated. However, such methods require complex computations, which limits their ability in real-time systems.

Multi-Sensor Fusion. Standard practices to avoid traffic accidents in AVs can also mitigate the effect of L-Hawk. These include fusing 3D point cloud data from LiDAR. These fusion techniques enhance the system's resilience against adversarial attacks by cross-validating sensor data. However, sensor fusion cannot detect L-Hawk but only mitigate post-attack damage, e.g., stopping the vehicle before a collision.

Random Rolling Shutter Mechanism. To defend against L-Hawk, an effective method is to change the camera imaging algorithm (a fixed setting of CMOS sensors) and thus destroy the color stripe. There are two potential camera imaging algorithms: (1) configuring the electronic shutter to expose the CMOS sensor rows in a random sequence which disperses the color stripe across the entire image, thereby making the attack ineffective and (2) starting the exposure of the shutter from a random row for each frame and subsequently causing the color stripe to appear at different locations in the image across consecutive frames [15].

VIII. Conclusion

In this paper, we present L-Hawk, a controllable physical adversarial patch attack that is activated by specific laser signals. This innovative approach allows for targeted manipulation of vision-based perception systems used by specific autonomous vehicles (AVs). Our study conducts four types of attacks — hiding attacks, creating attacks, and two targeted attacks — against three object detectors and eight image classifiers, ensuring a comprehensive evaluation of L-Hawk 's capabilities. Extensive experiments demonstrate the effectiveness of L-Hawk in both digital and physical environments, highlighting its potential impact on real-world applications. We hope that our research can inspire the development of new defense mechanisms to enhance the security of vision-based perception systems in autonomous vehicles.

Acknowledgment

We sincerely appreciate the anonymous reviewers for their insightful comments. This work was supported by the National Natural Science Foundation of China (No.U21A20464, No.62261160651, No.U23A20307, No.U23A20306, No.62406239), the Fundamental Research Funds for the Central Universities (No.QTZX24081), the China Postdoctoral Science Foundation (No.2023M742739), and the China 111 Project (No.B16037).

References

[1] C. Yang, "New trends in sensors for autonomous driving perception systems," 2024, https://omdia.tech.informa.com/blogs/2024/mar/new-trends-in-sensors-for-autonomous-driving-perception-systems. 1

[2] TaskUs, "The role of computer vision for smart and safe autonomous driving," 2024, https://www.taskus.com/insights/computer-vision-for-autonomous-vehicles/. 1

[3] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*. Chapman and Hall/CRC, 2018, pp. 99–112. 1, 3, 18

[4] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *International conference on machine learning*. PMLR, 2018, pp. 284–293. 1, 3, 18

[5] D. Song, K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, F. Tramer, A. Prakash, and T. Kohno, "Physical adversarial examples for object detectors," in *12th USENIX workshop on offensive technologies (WOOT 18)*, 2018. 1, 18

[6] Y. Zhao, H. Zhu, R. Liang, Q. Shen, S. Zhang, and K. Chen, "Seeing isn't believing: Towards more robust adversarial attack against real world object detectors," in *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, 2019, pp. 1989–2004. 1, 3, 6, 10, 18

[7] W. Jia, Z. Lu, H. Zhang, Z. Liu, J. Wang, and G. Qu, "Fooling the eyes of autonomous vehicles: Robust physical adversarial examples against traffic sign recognition systems," *arXiv preprint arXiv:2201.06192*, 2022. 1, 3, 18

[8] T. Sato, S. H. Bhupathiraju, M. Clifford, T. Sugawara, Q. A. Chen, and S. Rampazzi, "Invisible Reflections: Leveraging Infrared Laser Reflections to Target Traffic Sign Perception," in *Network and Distributed System Security Symposium (NDSS)*, 2024. 1, 3, 10, 18

[9] S. H. Bhupathiraju, T. Sugawara, T. Sato, Q. A. Chen, M. Clifford, and S. Rampazzi, "On the vulnerability of traffic light recognition systems to laser illumination attacks," in *ISOC Symposium on Vehicle Security and Privacy (VehicleSec)*. ISOC, 2024. 1, 18

[10] W. Zhu, X. Ji, Y. Cheng, S. Zhang, and W. Xu, "TPatch: A triggered physical adversarial patch," in *32nd USENIX Security Symposium (USENIX Security 23)*. Anaheim, CA: USENIX Association, Aug. 2023, pp. 661–678. [Online]. Available: https://www.usenix.org/conference/usenixsecurity23/presentation/zhu 1, 3, 4, 6, 7, 9, 10, 12, 17, 18

[11] E. Chou, F. Tramer, and G. Pellegrino, “Sentinet: Detecting localized universal attacks against deep learning systems,” in *2020 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2020, pp. 48–54. 1, 14

[12] C. Xiang, A. N. Bhagoji, V. Sehwag, and P. Mittal, “{PatchGuard}: A provably robust defense against adversarial patches via small receptive fields and masking,” in *30th USENIX Security Symposium (USENIX Security 21)*, 2021, pp. 2237–2254. 1, 14

[13] C. Xiang, S. Mahloujifar, and P. Mittal, “PatchCleanser: Certifiably robust defense against adversarial patches for any image classifier,” in *31st USENIX Security Symposium (USENIX Security 22)*. Boston, MA: USENIX Association, Aug. 2022, pp. 2065–2082. [Online]. Available: https://www.usenix.org/conference/usenixsecurity22/presentation/xiang 1, 14

[14] S. Köhler, G. Lovisotto, S. Birnbach, R. Baker, and I. Martinovic, “They see me rollin’: Inherent vulnerability of the rolling shutter in cmos image sensors,” in *Proceedings of the 37th Annual Computer Security Applications Conference*, 2021, pp. 399–413. 2, 13, 18

[15] C. Yan, Z. Xu, Z. Yin, S. Mangard, X. Ji, W. Xu, K. Zhao, Y. Zhou, T. Wang, G. Gu *et al.*, “Rolling colors: Adversarial laser exploits against traffic light recognition,” in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 1957–1974. 2, 3, 4, 5, 13, 14, 16, 18

[16] R. Pankratau, “Light scattering,” 2023, https://rafcamera.com/info/imaging-theory/light-scattering. 2

[17] T. M. Club, “The ar0132at camera in tesla,” 2017, https://teslamotorsclub.com/tmc/threads/hw2-5-capabilities.95278/page-73. 2

[18] Mobileye, “Mobileye 8 connect,” 2019, https://f.hubspotusercontent40.net/hubfs/7006295/Mobileye_8_Datash-eet_DiCANinc.pdf. 2

[19] J. Redmon and A. Farhadi, “Yolov3: An incremental improvement,” *arXiv preprint arXiv:1804.02767*, 2018. 2, 7

[20] G. Jocher, A. Stoken, J. Borovec, L. Changyu, A. Hogan, L. Diaconu, J. Poznanski, L. Yu, P. Rai, R. Ferriday *et al.*, “ultralytics/yolov5: v3.0,” *Zenodo*, 2020. 2

[21] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016. 2

[22] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014. 2

[23] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. 2

[24] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, “Inception-v4, inception-resnet and the impact of residual connections on learning,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1, 2017. 2, 3

[25] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520. 2

[26] S. Houben, J. Stallkamp, J. Salmen, M. Schlipsing, and C. Igel, “Detection of traffic signs in real-world images: The german traffic sign detection benchmark,” in *The 2013 international joint conference on neural networks (IJCNN)*. Ieee, 2013, pp. 1–8. 2

[27] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, “Adversarial patch,” *arXiv preprint arXiv:1712.09665*, 2017. 3, 18

[28] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in *2017 ieee symposium on security and privacy (sp)*. Ieee, 2017, pp. 39–57. 3

[29] I. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in *International Conference on Learning Representations*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6572 3, 18

[30] S.-T. Chen, C. Cornelius, J. Martin, and D. H. Chau, “Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector,” in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*. Springer, 2019, pp. 52–68. 3, 18

[31] Hikvision, “Hikvision c6 pro dashcam,” 2024, https://www.hikvision.com/ca-en/products/onboard-security/dash-cameras/dash-cameras/ae-dc5313-c6pro/. 4, 10, 13

[32] P. Bankhead, “From photons to pixels,” 2024, https://bioimagebook.github.io/chapters/3-fluorescence/1-formation_overview/formation_overview.html. 5

[33] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, “Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition,” in *Proceedings of the 2016 acm sigsac conference on computer and communications security*, 2016, pp. 1528–1540. 6, 17, 18

[34] J. Shen, N. Wang, Z. Wan, Y. Luo, T. Sato, Z. Hu, X. Zhang, S. Guo, Z. Zhong, K. Li *et al.*, “Sok: On the semantic ai security in autonomous driving,” *arXiv preprint arXiv:2203.05314*, 2022. 6, 10, 18

[35] Microsoft, “Common objects in context dataset,” 2018, https://cocodataset.org/. 7

[36] J. Deng, “A large-scale hierarchical image database,” *Proc. of IEEE Computer Vision and Pattern Recognition, 2009*, 2009. 7, 12

[37] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the kitti vision benchmark suite,” in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361. 7

[38] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, “Bdd100k: A diverse driving dataset for heterogeneous multitask learning,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2636–2645. 7

[39] D. Feng, A. Harakeh, S. L. Waslander, and K. Dietmayer, “A review and comparative study on probabilistic object detection in autonomous driving,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 9961–9980, 2021. 7

[40] Logitech, “Logitech c920 pro hd webcam,” 2024, https://www.logitech.com/en-ch/products/webcams/c920-pro-hd-webcam.960-001055.html. 10

[41] Intel, “Intel realsense depth camera d435i,” 2024, https://www.intelrealsense.com/depth-camera-d435i/. 10

[42] G. Lovisotto, H. Turner, I. Sluganovic, M. Strohmeier, and I. Martinovic, “{SLAP}: Improving physical adversarial examples with {Short-Lived} adversarial perturbations,” in *30th USENIX Security Symposium (USENIX Security 21)*, 2021, pp. 1865–1882. 10

[43] R. e Manual, “Turbot3,” 2024, https://emanual.robotis.com/docs/en/platform/turtlebot3/overview/. 13

[44] H. Tian, B. Fowler, and A. E. Gamal, “Analysis of temporal noise in cmos photodiode active pixel sensor,” *IEEE Journal of Solid-State Circuits*, vol. 36, no. 1, pp. 92–101, 2001. 13

[45] A. Sayles, A. Hooda, M. Gupta, R. Chatterjee, and E. Fernandes, “Invisible perturbations: Physical adversarial examples exploiting the rolling shutter effect,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 14 666–14 675. 13

[46] H. Zhang and J. Wang, “Towards adversarially robust object detection,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 421–430. 14

[47] D. Meng and H. Chen, “Magnet: a two-pronged defense against adversarial examples,” in *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, 2017, pp. 135–147. 14

[48] X. Shi, Y. Sun, H. Liu, L. Bai, and C. Lin, “Research on laser stripe characteristics and center extraction algorithm for desktop laser scanner,” *SN Applied Sciences*, vol. 3, pp. 1–12, 2021. 16

[49] B. Kondász, B. Hopp, and T. Smausz, “Homogenization with coherent light illuminated beam shaping diffusers for vision applications: spatial resolution limited by speckle pattern,” *Journal of the European Optical Society-Rapid Publications*, vol. 14, pp. 1–7, 2018. 16

[50] Q. Cao, N. Zhang, A. Chong, and Q. Zhan, “Spatiotemporal hologram,” *arXiv preprint arXiv:2401.12642*, 2024. 16

[51] S. Thys, W. Van Ranst, and T. Goedemé, “Fooling automated surveillance cameras: adversarial patches to attack person detection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 0–0. 18

[52] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, “Boosting adversarial attacks with momentum,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9185–9193. 18

[53] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille, “Adversarial examples for semantic segmentation and object detection,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1369–1378. 18

[54] R. Duan, X. Mao, A. K. Qin, Y. Chen, S. Ye, Y. He, and Y. Yang, “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*, 2021, pp. 16 062–16 071. 18

[55] Q. Jiang, X. Ji, C. Yan, Z. Xie, H. Lou, and W. Xu, "{GlitchHiker}: Uncovering vulnerabilities of image signal transmission with {IEMI}," in *32nd USENIX Security Symposium (USENIX Security 23)*, 2023, pp. 7249–7266. 18

Appendix

A. Tracking and Aiming Against Moving Vehicles With Laser

Intuitively, targeting a moving vehicle at a long distance is challenging. To address this, we construct an aiming equipment also used in [15] by combining a laser diode and a monocular telescope on a tripod (Figure 10). The laser, aligned with the telescope's eyepiece, allows attackers to manually track moving targets from a long distance, without requiring high skills. We conduct experiments to test the aiming success rate under different vehicle speeds. Table X shows the results under various speeds, yielding an averaged aiming success rate of 84.2%. Therefore, manual aiming using the equipment is feasible, even with the vehicle moving at a high speed. More details are presented in Appendix A.

Then, we test the success rate of laser aiming based on the aiming equipment. Researchers drive a real car at various speeds of 5 km/h, 10 km/h, 15km/h, and 20 km/h. Then, we record about 300 continuous frames (about 10 seconds) of video as the car moves 50 m away toward the aiming equipment. 20 trials are conducted for each speed. Based on the captured video, we calculate the aiming success rate. Experiments are conducted on closed roads with proper laser protection at our institute.

Impact of Laser Aiming Position in Camera Lens. The above test proves that we can use lasers to influence the image of the camera when the vehicle moves. However, considering that the circular spot generated by the laser may not all shine on the lens, we further investigated the influence of circular spot position on the lens on the color stripe. We use the Hikvision C6 Pro dashcam for testing. As shown in Figure 15, the diameter of the camera lens is 0.5 cm and the diameter of the circular spot generated by the laser is 1.2 cm. Then we test the effect of the four positions of the circular spot on the pixel intensity of the color strip. Depending on the symmetry of the lens, we only test the position of the circular spot when it moves down: center, offset but full coverage, about half coverage, and about one-third coverage. The pixel intensity results show that the brightness of the color stripe changes little as long as it is fully covered. This is because the number of photons on the circular spot is evenly distributed, thus the number of photons captured by the lens in full coverage remains relatively constant. When the circular spot only covers part of the lens, the number of photons captured by the lens correspondingly decreases, resulting in a proportional reduction in pixel intensity. Therefore, the impact of the laser's position on the lens is negligible. In this paper, we only need to ensure that the laser can be aimed at the lens.

TABLE X: The aiming success rate under various speeds.

Speed	5 km/h	10 km/h	15 km/h	20 km/h
Aiming Success Rate	98.6%	92.0%	80.7%	65.3%

B. Trigger Generation

There are three methods to simulate the trigger, i.e., the linear function, the sigmoid function, and the Gaussian function [15]. We first define the minimum and maximum intensities I_{min}, I_{max} measured by the CMOS and define the width and height of the trigger w_t and h_t. Therefore, the intensity of the trigger at a specific point (x, y) can be illustrated as $t = D(I_{max}, I_{min}, x, y, h_t)$. The function $D(\cdot)$ has three cases: linear, signoid, and Gaussian.

Linear Function: Given the I_{max} and I_{min}, we can express the light intensity function for any point (x, y) on the trigger as:

$$t = I_{min} + \frac{y}{w_t}(I_{max} - I_{min}) \tag{13}$$

Sigmoid Function: Sigmoid Function is also used for the case when the incidence angle is from left or right. Given the I_{max} and I_{min}, we define the trigger simulation function as:

$$t = I_{min} + \frac{1}{1 + e^{-\alpha_1/(y - w_t/\alpha_2)}}(I_{max} - I_{min}) \tag{14}$$

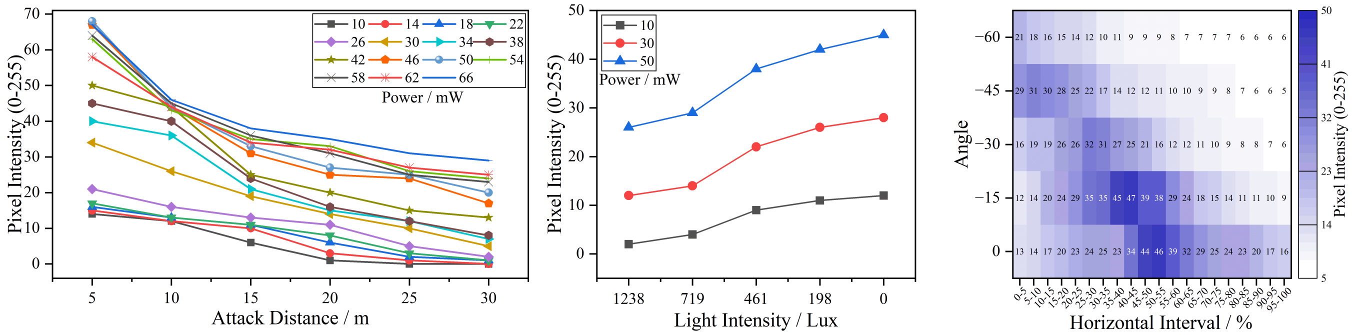
where α_1 and α_2 are hyper-parameters.

Gaussian Function: If the incidence direction of the light is from the front, channel overflow is most likely to occur in the middle of the trigger. The maximum light intensity I_{max} and the top position of the trigger x_0 are needed. The trigger simulation function for any points (x, y) on the trigger can be expressed as follows:

$$\begin{aligned}
a &= \frac{(x - x_0 - h/2)^2}{(h_t/\rho_1)^2} \\
b &= \frac{(y - w/2)^2}{(w_t\rho_2)^2} \\
c &= \varsigma\frac{2(x - x_0 - h_t/2)(y - w_t/2)}{(h_t w_t)/(\rho_1\rho_2)} \\
t &= \frac{1}{2\pi w_t h_t\sqrt{1 - \varsigma^2}/(\rho_1\rho_2)} e^{-\frac{1}{2(1-\varsigma^2)}(a+b+c)} I_{max} h_t w_t
\end{aligned} \tag{15}$$

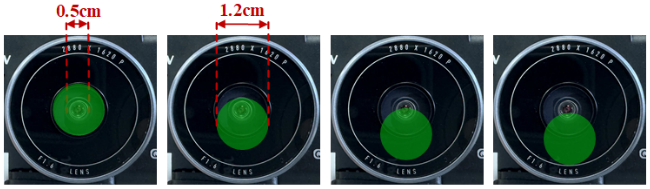
where ς is the correlation between two directions. ρ_1 and ρ_2) are hyper-parameters representing the decaying rate of the light intensity from the center to the periphery.

Prior work [48] demonstrates that the laser power directly affects the trigger's intensity. The higher the power, the greater the I_{max} and I_{min} values. The attack distance affects the attenuation rate of light. As the distance increases, the light intensity decreases in inverse square ratio [49]. Therefore, the farther the distance, the lower the intensity of light received, and both I_{max} and $_{min}$ decrease. Cao *et al.* [50] also demonstrate that the incidence angle affects the distribution of the trigger's intensity. In addition, the ambient light also interferes with the trigger's intensity [48]. The greater the ambient light intensity, the more obvious the background noise, which may lead to an increase in the I_{max} and I_{min} values. The above studies fully demonstrate the correlation between parameters and trigger intensity. Thus, we formulate Equation 16 to directly show the relationship between the intensities I_{max},



(a) Attack Distance & Power **(b) Light Condition & Power** **(c) Incidence Angle**

Fig. 14: The influence of attack and scenario parameters on the color stripe's pixel intensity includes (a) the impact of attack distance and laser power, (b) the impact of light condition and laser power, and (c) the impact of incidence angle. In (c), the color stripe is divided horizontally into 20 areas of equal proportion, and the average pixel value of each area is calculated. The x-axis indicates the different areas from left to right in the color stripe.



(a) center Pixel intensity: **53**

(b) offset but full coverage Pixel intensity: **50**

(c) 1/2 coverage Pixel intensity: **21**

(c) 1/3 coverage Pixel intensity: **11**

Fig. 15: Pixel intensity for different laser aiming positions in the camera lens.

I_{min}, and the parameters p, d, θ, l.

$$
\begin{aligned}
I_{max} &= k_1 \cdot p \cdot \frac{cos(\theta)}{d^2} + k2 \cdot l \\
I_{min} &= k_3 \cdot p \cdot \frac{cos(\theta)}{d^2} + k4 \cdot l
\end{aligned}
\tag{16}
$$

where k_1, k_2, k_3, and k_4 are calibration constants related to the characteristics of the sensor.

C. Definition of Loss Functions

To achieve the proposed attack goals, we extend the design of the loss function in [10]. The benign and attack losses, i.e., ℓ_{benign} and ℓ_{attack}, are different across object detectors and image classifiers. We first define two losses against detectors as follows:

$$
\begin{aligned}
\ell_{hide} &= -log(1 - max(p_{obj} \cdot p_{tc})) \\
\ell_{create} &= -log(p_{obj} \cdot p_{tc}) + \phi \ell_{reg}
\end{aligned}
\tag{17}
$$

where ℓ_{hide} is to make the object with class tc undetectable. ℓ_{create} is for the object to be detected and belong to class tc. Specifically, p_{obj} and p_{tc} represent the objectiveness scores and the classification scores of target class tc respectively. ℓ_{reg} denotes the regression loss to guide the detected bounding box. ϕ is the hyperparameter to balance the recognition loss and regression loss.

Then, based on the Equation 17, we define ℓ_{attack} and ℓ_{benign} to achieve HA, CA, TA-D against object detectors. The loss functions of different attacks are formulated as follows:

$$
\begin{aligned}
\ell_{attack} &= \begin{cases} \ell_{hide} & \text{if HA} \\ \ell_{create} & \text{if CA or TA-D} \end{cases} \\
\ell_{benign} &= \begin{cases} \ell_{hide} & \text{if CA} \\ \ell_{create} & \text{if HA or TA-D} \end{cases}
\end{aligned}
\tag{18}
$$

Note that the target class tc of ℓ_{attack} and ℓ_{benign} is not the same in the TA-D attack.

Different from attacks against object detectors, attacks against image classifiers aim to alter the probability of classification. Thus, we formulate the loss function of TA-C as follows:

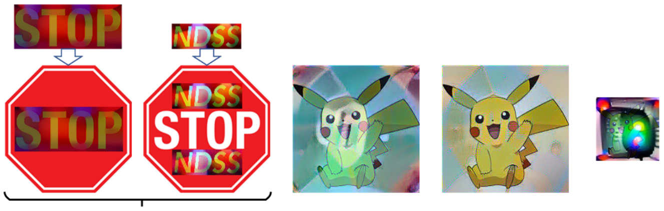
$$
\begin{aligned}
\ell_{attack} &= -log(p_{tc}) \\
\ell_{benign} &= -log(1 - p_{tc})
\end{aligned}
\tag{19}
$$

where p_{tc} is the probability of target class tc predicted by the image classifier.

In addition, we present the total variation (TV) loss [33] in Equation 20 that is utilized to regularize L-HAWK. The TV loss aims to minimize color changes between adjacent pixels, thereby reducing overfitting in digital simulations and enhancing the image quality of L-HAWK. By smoothing out abrupt color transitions, TV loss ensures that L-HAWK maintains a realistic appearance, which is crucial for its effectiveness in real-world attacks.

$$
\ell_{tv}(\delta) = \sum_{i,j} \sqrt{\left(\delta_{i,j} - \delta_{i+1,j}\right)^2 + \left(\delta_{i,j} - \delta_{i,j+1}\right)^2} \tag{20}
$$

To increase the stealthiness of L-HAWK, we adopt the content-based camouflage loss ℓ_{cam} in [10]. The purpose of ℓ_{cam} is to generate adversarial patches that blend naturally into their surroundings rather than appearing as obvious and abnormal objects. As shown in Equation 21, ℓ_{cam} is calculated based on high-level features extracted by a pre-trained convolutional neural network (CNN) $\mathcal{M}$. $\ell_{content}$ encourages δ to mimic the spatial structure and general content of target



(a) HA (b) CA (c) TA-D (d) TA-C

Fig. 16: The specific physical adversarial patches used in the physical evaluation. For HA, we paste the L-HAWK on the stop sign. For other attacks, L-HAWK is a standalone object.

image $\hat{\delta}$, rather than its specific details such as color or texture, leading to a more visually coherent and less detectable δ.

$$\ell_{content}\left(\delta\right) == \frac{1}{C_j H_j W_j}\left\|\mathcal{M}_j(\hat{\delta}) - \mathcal{M}_j(\delta)\right\|_2^2 \tag{21}$$

where $\mathcal{M}_j$ denotes the j-th layer of $\mathcal{M}$. $C_j H_j W_j$ represents the shape of $\hat{\delta}$ and δ.

We further present ℓ_{nps} that is the non-printability score (NPS) loss [33] of the adversarial patch. Specifically,

$$\ell_{nps}\left(\delta\right) = \sum_{p \in \delta} \min_{c \in C} \mid p - c \mid \tag{22}$$

where p is a pixel of δ, and c is one of the printable colors C [51]. The loss aims to make colors in δ closer to the colors that can be printed by a common printer.

D. Limitations

L-HAWK currently has several limitations:

Hardware Black-box Attacks. The limitations of hardware black-box attacks stem from varying camera refresh rates, complicating consistent interference. Laser injection attacks require knowledge of the camera's refresh rate to stabilize the color stripe, but differences across devices make this challenging. Accurate timing of the injection is also difficult without camera feedback. While a wider stripe increases the chance of triggering L-HAWK, it may cause false alarms and disrupt object detection, potentially disabling self-driving functions. Future research should focus on adaptive interference methods and optimizing stripe width and timing to improve attack reliability without affecting detection.

Evaluation Scope. Due to budget constraints on testing cars and safety hazards posed by adversarial attacks, we only adopt a similar evaluation to the prior work [34], omitting comprehensive evaluations on self-driving components such as planning and control. We establish a model to compute the laser-based trigger, which can be feasibly achieved physically and will be discussed in our future work.

Attack Camouflage. L-HAWK proves effective in natural settings with ample light, yet its camouflage capabilities diminish in some special scenarios. Specifically, while laser-based attacks extend the attack range, they compromise stealthiness at closer distances. Additionally, in dimly lit environments, the system exhibits reduced invisibility due to the more significant color shift induced by the laser stripes. Hence, enhancing the camouflage performance represents a crucial area for future development.

TABLE XI: Specs of the four experimented cameras used in the stationary experiment.

Parameter	 Hikvision	 Logitech	 Intel D435i	 iPhone XR
Sensor	OS 05A20	N/A	N/A	N/A
Refresh Rate (Hz)	30	30	30	30
Resolution	2880*1620	1920*1080	1920*1080	1920*1080
Field of View	130	78	69	N/A
F	1.6	1.2	3.5	1.8

E. Related Works

We summarize related works from two aspects: adversarial attacks, and physical sensor attacks.

Adversarial Attacks. Numerous studies show that deep learning models, such as object detection and image classification, are vulnerable to adversarial attacks (adversarial examples, AEs) [29], [52], [53]. These adversarial attacks have been explored in the physical world against vision-based perception systems [3]–[8], [30]. Kurakin *et al.* [3] first demonstrate the feasibility of physical attacks against image classifiers by printing AEs, although subtle pixel modifications could limit the efficacy of such attacks. Brown *et al.* [27] propose adversarial patch attacks by using only localized perturbations. In [4], Expectation over Transformation (EoT) is proposed to enhance the robustness of physical adversarial examples. Song *et al.* [5]–[7] successfully conduct attacks against both Faster R-CNN and YOLO detectors. Further enhancing the stealthiness of AEs, Sato *et al.* [8] propose a misclassification attack against traffic sign recognition systems based on invisible infrared laser reflection. Additionally, Bhupathiraju *et al.* [9] and Duan *et al.* [54] use a laser to generate adversarial examples in traffic signs or lights and thus disturb the recognition of AV systems. The AEs produced in the above studies, either in the digital or physical worlds, are indiscriminately malicious to every victim. An existing work [10] proposes a physical adversarial patch activated by special physical ultrasonic signals but is limited by short attack distance and conspicuous attack devices.

Physical Sensor Attacks. Another branch of work demonstrates that certain physical signals can disrupt sensor operations, adversely affecting the imaging process and consequently impairing vision-based perception systems. Jia *et al.* [7] propose adversarial blur attacks against object detectors by emitting acoustic signals to disturb the image stabilization system. Jiang *et al.* [55] utilize electromagnetic interference (EMI) to disrupt image transmission signals and thus disturb detection tasks. Kohler *et al.* [14] demonstrate that the rolling shutter effects formed after a laser attack on the camera are used to randomly disrupt object detection. Similarly, Yan *et al.* [15] utilize the color stripe caused by laser signal attacks to disturb traffic light recognition.