Laser Shield: a Physical Defense with Polarizer against Laser Attacks on Autonomous Driving Systems

Qingjie Zhang[1,*], Lijun Chi[2,*], Di Wang[3], Mounira Msahli[2], Gerard Memmi[2], Tianwei Zhang[4], Chao Zhang[1], and Han Qiu[1,&]

[1]Tsinghua University, Beijing, China, [2]Telecom Paris, Institut Polytechnique de Paris, Paris, France, [3]Beijing University of Posts and Telecommunications, Beijing, China, [4]Nanyang Technological University, Singapore

ABSTRACT

Autonomous driving systems (ADS) are boosted with deep neural networks (DNN) to perceive environments, while their security is doubted by DNN's vulnerability to adversarial attacks. Among them, a diversity of laser attacks emerges to be a new threat due to its minimal requirements and high attack success rate in the physical world. Nevertheless, current defense methods exhibit either a low defense success rate or a high computation cost against laser attacks. To fill this gap, we propose Laser Shield which leverages a polarizer along with a min-energy rotation mechanism to eliminate adversarial lasers from ADS scenes. We also provide a physical world dataset, LAPA, to evaluate its performance. Through exhaustive experiments with three baselines, four metrics, and three settings, Laser Shield is proved to surpass SOTA performance.

ACM Reference Format:

Qingjie Zhang[1,*], Lijun Chi[2,*], Di Wang[3], Mounira Msahli[2], Gerard Memmi[2], Tianwei Zhang[4], Chao Zhang[1], and Han Qiu[1,&]. 2024. Laser Shield: a Physical Defense with Polarizer against Laser Attacks on Autonomous Driving Systems. In *61st ACM/IEEE Design Automation Conference (DAC '24), June 23–27, 2024, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3649329.3657358

1 INTRODUCTION

The development of autonomous driving systems (ADS) is accelerated by the integration of multiple deep neural networks (DNN) for perceiving environments [4]. Nevertheless, the emergence of research on the security issues of ADS reveals a great gap for its deployment [9].

One of the most significant threats is adversarial attacks [2, 11, 18]. In general, adversarial perturbations can be generated from the gradients derived from target models in white-box scenarios [11, 18], or transferred based on the gradients of surrogate models in black-box scenarios [2]. However, a diversity of research has emerged recently to mislead DNN models without using gradients such as laser attacks [10, 17]. Particularly, an attacker can project a laser beam in front of a traffic sign [10] to mislead the prediction of a well-trained perception model. This kind of laser attacks, as a novel paradigm of adversarial attacks, poses a more critical threat to the perception of an ADS due to three reasons as follows. First, only a laser pointer can make such a simple yet effective attack. Second, they are indifferent to the attack environment as long as the laser beams are visible. Third, the information on target models is not necessarily required. Due to the low cost and the high robustness, laser attacks are extremely easy to deploy in the physical world, achieving a high attack success rate.

A great number of defense methods are proposed to mitigate adversarial attacks [14, 16] like adversarial training or image preprocessing. But adversarial training is too costly while image preprocessing cannot mitigate laser attacks since they are designed to reduce human-imperceptible adversarial perturbations which lasers are no longer mild. A few recent research deploy diffusion models (DM) to restore corrupted images [7]. Although it has the potential to eliminate laser, the computation cost is not tolerant in rapidly changing ADS scenarios [4].

To fill the gap of defense against laser attacks in ADS scenarios, we propose a novel physical defense Laser Shield. Our intuition is that lasers are polarized lights that are different with natural lights. Thus, Laser Shield leverages the polarizer along with a specifically designed min-energy rotation to eliminate adversarial laser beams and preserve the functionality of models in ADS, with negligible computation cost. Laser Shield is robust and easy to implement in the physical world scenarios. In order to verify the effectiveness of Laser Shield, we provide a novel dataset of traffic signs collected from physical world, LAPA[1] (Laser Polarizer ADS), to evaluate laser attacks and our defense. LAPA consists of a diversity of attack conditions, and it reflects the filtration effect of lasers by Laser Shield. We conduct a thorough evaluation with three baselines, four metrics, and three settings (digital space, optical lab, and on road) to prove Laser Shield's effectiveness.

2 PRELIMINARIES

2.1 Attacks and defenses on ADS

ADS highly relies on DNN models to perceive driving environment, but they are known to be vulnerable to adversarial attacks [2, 11, 18]. For instance, Bai et al. [2] can efficiently generate adversarial examples for black-box DNNs. However, they are difficult to adopt in physical world, making the attacks on ADS shift towards manipulating physical objects or environments such as adversarial lasers. Duan et al. [10] proposed the first laser attacks using laser beams. Yan et al. [17] proposed to directly emit the laser on camera.

Facing the threat of adversarial attacks, we summarize related defense methods in adversarial training [1], image preprocessing [14, 16], and image restoration [3]. Nevertheless, they all have shortcomings against laser attacks. In reality, it is difficult for ADS companies to frequently use high-cost adversarial training especially for the diversity of emerging new threats [9]. The cost of image preprocessing is acceptable, but they are proposed against mild pixel-level perturbations [14, 16], which is ineffective on intense laser beams. With the impressive results of diffusion models (DM) [7], image restoration has great potential to eliminate lasers.

DAC '24, June 23–27, 2024, San Francisco, CA, USA

ACM ISBN 979-8-4007-0601-1/24/06.
https://doi.org/10.1145/3649329.3657358

[1]https://github.com/qingjiesjtu/LaserShield

* Equal contribution. & Corresponding author (qiuhan@tsinghua.edu.cn).

However, using DM will generate a long latency which is not acceptable in driving scenarios.

2.2 Threat model and defense requirements

Threat model. We summarize the threat model in the adversary's goal, knowledge, and capability.

- **Adversary's goal.** The main goal is to cause malfunctions of DNN models with lasers. Since we take ResNet [12] as the target model and we mainly evaluate traffic signs in this paper, the goal is to mislead the classification model to predict a result other than "street sign".
- **Adversary's knowledge.** The adversary's knowledge mainly consists of two aspects. First, he has full knowledge of the environment to deploy the attack, including the target to mispredict, the light conditions, etc. Second, the target models can be either white-box or black-box. For the former, the adversary can design an adversarial laser with feedback from target models. For the latter, an arbitrary laser condition, as long as the laser greatly overlaps with the traffic sign, has also a non-ignorable possibility to generate error [10].
- **Adversary's capability.** The adversary can choose any traffic sign as a target and generate laser beams with any color, any intensity, and in any direction.

Defense requirements. As the countermeasure to the above threat model, we list four requirements for our defense solution.

- **Laser elimination.** The solution is capable of eliminating or weakening laser beams from captured images, and helps target models to predict correctly as if no laser exists.
- **Functionality-preserving.** The solution cannot disturb the prediction on images where lasers cause no malfunction or even no laser exists.
- **High fault tolerance.** Such a solution should have high fault tolerance to meet the complex and dynamic ADS scenarios.
- **Negligible computation cost.** The solution should minimize computation costs since ADS needs to give quick feedback on the rapidly changing driving environment.

3 METHODOLOGY

3.1 Overview

Figure 1 shows the overview of `Laser Shield`. The main idea is to eliminate adversarial lasers while preserving normal light. Knowing that lasers can be filtered by a polarizer with a certain angle, we leverage this physical phenomenon to develop a plug-and-play defense strategy. Specifically, it is achieved by arranging a polarizer in the camera, with min-energy rotation mechanism to find the optimal angle for defense. Hence, we first give an in-depth analysis of the filtration mechanism of polarizer against lasers, ensuring the feasibility of leveraging a polarizer to make `Laser Shield`. Second, since mounting the polarizer in camera needs a rotation mechanism to locate effective range, we propose min-energy rotation with an indicator function, RGB energy, to indicate the optimal angle.

3.2 Make a Shield: Polarizer

Natural light V.S. polarized light. In wave optics, light is described as coupled waves of electric and magnetic fields oscillating

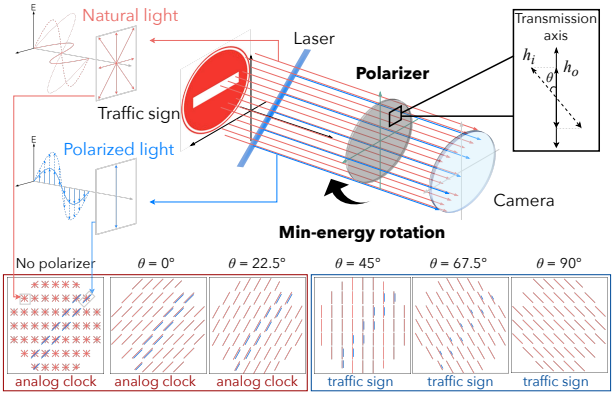

Figure 1: Overview of `Laser Shield`. It consists of a polarizer and a min-energy rotation mechanism to eliminate lasers.

in a sinusoidal pattern [5]. These oscillations are perpendicular to the light's propagation direction. The majority of light sources in our daily life (daylight, LED light, etc.) are unpolarized light or natural light which consists of a combination of waves with varying oscillations in all directions. In contrast, to ensure high intensity and monochromaticity, lasers are polarized light that oscillates only in one direction. Such distinction leaves a chance for defenders to eliminate lasers and preserve natural light.

Polarizer. Our next step involves an optical device to leverage the above distinction. More than that, it has to be small and easy to mount in cameras. The polarizer, a lens characterized as an assemblage of numerous aligned slender metal wires, is the ideal solution. When light reaches the polarizer, the waves oscillating perpendicular to the wires will be the least weakened by an anti-field caused by its oscillation. In other words, only the components perpendicular to the wires can pass through and this direction is defined as the polarizer transmission axis. Hence, the amplitude of wave is reduced as follows.

$$h_o = h_i \cos(\theta), \tag{1}$$

where h_i (resp. h_o) is the amplitude of the incident (resp. outgoing) light wave, and θ is the angle between the incident wave's oscillation direction and the transmission axis of polarizer.

Malus's law. The above amplitude reduction directly affects the intensity of light, which is described by Malus's law [6]:

$$I_o = I_i \cos^2(\theta) = I_i \frac{1 + cos(2\theta)}{2}, \tag{2}$$

where I_i (resp. I_o) is the intensity of the incident (resp. outgoing) light. For lasers, the intensity varies within range $[0, I_i]$ based on the angle θ. The variation period decreases from 360° to 180° due to the square operation, and an interval of length 90° covers the full intensity range. For natural light, the intensity is reduced but cannot be zero since it always contains waves oscillating in the transmission axis (as shown in Figure 1).

3.3 Mount the Shield: Min-energy Rotation

The polarizer does not always coincidentally fall in the effective angle range for laser elimination. Indeed, when the oscillation is parallel to the transmission axis, the polarizer will not work anymore. To compensate for this shortcoming, we devise a min-energy

rotation mechanism for the polarizer. The polarizer mounted in the camera will be attached to a high-speed rotator which should rotate until an optimal angle. This leads to the need for an indicator function, which should indicate the angle with minimal laser beams, and be computation efficient when facing the rapidly changing autonomous driving scenarios [4]. For an image $x = s \oplus p_\theta$ of scene s filtered by polarizer p with angle θ, a simple idea is to use the image energy $\mathcal{E}$ [13]:

$$\mathcal{E} = \sum_{(i,j)\in[1,\dots,m]\times[1,\dots,n]} {x'_{i,j}}^2 \tag{3}$$

$$x' = 0.299x_R + 0.587x_G + 0.114x_B, \tag{4}$$

where x_R, x_G, x_B are RGB channels of x, x' is the grayscale image, and (m, n) is the size. Although this definition is widely used in image processing, it works poorly to represent laser intensity, especially for the blue laser. This is due to the low weight (0.114 in Equation 4) for the blue channel. Indeed, the imbalanced RGB weights of grayscale images are designed for humans' perception, while computation algorithms are indifferent to colors. To make up for this imbalance, we therefore propose RGB energy $\mathcal{E}_{RGB}$:

$$\mathcal{E}_{RGB} = \sum_{(i,j)\in[1,\dots,m]\times[1,\dots,n]} (x^2_{R_{i,j}} + x^2_{G_{i,j}} + x^2_{B_{i,j}}) \tag{5}$$

$\mathcal{E}_{RGB}$ is highly correlated to the laser intensity regardless of the laser color (detailed explanation is given in Section 4.6). It is also a lightweight function with time complexity $O(mn)$, contributing to the low computation cost of Laser Shield (further proved in Section 4.4). Algorithm 1 details min-energy rotation.

Algorithm 1 Min-energy rotation.

input: Scene s, Polarizer initial angle θ_{ini}, Rotation step size δ
output: Optimal polarizer state for laser elimination $p_{\theta_{opt}}$

Initialize a parameter angle $\theta \leftarrow \theta_{ini}$
Initialize the optimal RGB energy $\mathcal{E}_{opt} \leftarrow \infty$
while $\theta \le \theta_{ini} + 90^\circ$ **do**
  Rotate the polarizer to angle θ
  Compute RGB energy $\mathcal{E}_{RGB}$ of current vision $x = s \oplus p_\theta$
  if $\mathcal{E}_{RGB} \le \mathcal{E}_{opt}$ **then**
    Update $\theta_{opt} \leftarrow \theta$; $\mathcal{E}_{opt} \leftarrow \mathcal{E}_{RGB}$
  end if
  Update $\theta \leftarrow \theta + \delta$
end while
Rotate the polarizer to angle θ_{opt}
return Optimal polarizer state for laser elimination $p_{\theta_{opt}}$

4 EXPERIMENTS

4.1 Setup

Settings. Following Duan et al. [10], we take ResNet50 [12] as the target model (assumed black-box) and traffic signs as the target objects. We conduct experiments in three settings.

- **Digital space.** Due to the heavy workload of data collection in physical world, we first simulate Laser Shield in digital space.
- **Optical lab.** We construct LAPA in an optical lab to evaluate Laser Shield in controlled environment conditions.
- **On road.** We evaluate Laser Shield on road to ensure maximal similarity to real-world ADS scenarios.

Baselines. As far as we know, there is no model specifically designed for laser elimination. Hence, we take three approaches that have the potential to eliminate laser as follows.

- **BdR [16].** Since laser attacks add salient color beams to images, we choose BdR, which reduces the color bit depth, to mitigate the adversarial effect.
- **PD [14].** Generally, laser attacks corrupt the natural statistics of images. We therefore choose PD to redistribute the image pixel values back to normal.
- **DM [7].** Laser elimination can be formulated as an image restoration problem [3]. Given the impressive results achieved by diffusion models (DM) in recent years, we choose a DM-based cleanup model[2] to remove laser before inference.

Metrics. We use four metrics to evaluate Laser Shield's effectiveness across multiple dimensions, corresponding to the defense requirements as stated in Section 2.2.

- **Defense success rate.** We mainly use the defense success rate (DSR) to evaluate Laser Shield's performance on laser elimination. This is the ratio of successful defenses in successful attack cases:
$$\text{DSR} = \frac{\sum_s \mathbb{1}\{f(s \oplus p_{\theta_{opt}}) = \dagger \wedge f(s) \neq \dagger\}}{\sum_s \mathbb{1}\{f(s) \neq \dagger\}}, \tag{6}$$
where f is the target model and $\dagger$ is the ground truth label.
- **Functionality-preserving rate.** Laser beams do not exist in all scenes. For cases with no laser beams or they are too weak to attack, we need to ensure that Laser Shield causes no influence on target model's benign functionality. We therefore propose the functionality-preserving rate (FPR). It is defined as the ratio of correct predictions when laser attacks fail:
$$\text{FPR} = \frac{\sum_s \mathbb{1}\{f(s \oplus p_{\theta_{opt}}) = \dagger \wedge f(s) = \dagger\}}{\sum_s \mathbb{1}\{f(s) = \dagger\}} \tag{7}$$
- **Fault tolerance rate** As stated in Section 3.3, Laser Shield works with rotation of polarizer. The theoretical optimal polarizer angle θ_{opt} requires a mechanical rotator to achieve, which lacks precision sometimes. However, effective defense is not only achieved at θ_{opt}. To evaluate Laser Shield's tolerance on polarizer angle θ, we propose the fault tolerance rate (FTR):
$$\text{FTR} = \frac{\sum_{s,\theta} \mathbb{1}\{f(s \oplus p_\theta) = \dagger \wedge f(s) \neq \dagger\}}{10 \times \sum_s \mathbb{1}\{f(s) \neq \dagger\}}, \tag{8}$$
where θ varies from 0° to 90° with step size 10°.
- **Computation complexity.** To show that Laser Shield exhibits a low computation burden, we evaluate its computation complexity in two forms, analytical and experimental. This diversity of forms ensures the consistency of theory and practice.

4.2 Digital simulation

Due to the heavy workload for constructing LAPA, we first simulate Laser Shield in digital space to verify its effectiveness against

[2] https://clipdrop.co/fr/cleanup

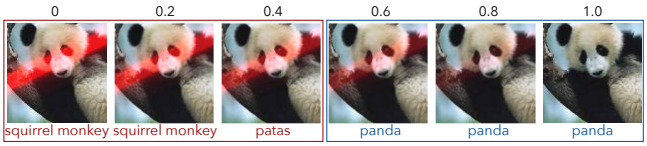

Figure 2: Digital simulation of an image "panda". The prediction of ResNet50 becomes correct with T varies from 0 to 1.0.

laser attacks. Following the same setting as Duan et al. [10] that simulate laser beam as a solid-color light tube, we regulate the filtration effect with transparency T, and the image with laser can be simulated as:

$$x = clip\left(s + (1 - T)\, I_0\right), \tag{9}$$

where s is the scene and I_0 is a hyperparameter representing the original laser intensity. During simulation, T varies from 0 (no defense exists) to 1 (lasers are completely filtered out).

Figure 2 shows an example of the simulation on a digital image. As transparency T increases, the laser beam becomes weaker, and the prediction turns to correct. It is worth noting that the correct prediction not only happens when the laser is the weakest. Indeed, out of five T values to simulate `Laser Shield`, the defense succeeds three times, thus $FTR = 60\%$. Table 1 shows the simulation on 500 images. It can be seen that laser filtration is significantly effective for defense because DSR rises as transparency increases. Besides, we also compute the overall FTR and it equals 71%. This indicates `Laser Shield`'s potential for high fault tolerance.

Table 1: Digital simulation of 500 images.

Transparency	0	0.2	0.4	0.6	0.8	1.0
DSR (%)	0	27.3	53.5	76.9	87.3	100

As a result, digital simulation implies the effectiveness of `Laser Shield`, on both DSR and FTR. Hence, it is necessary to construct a physical world dataset for laser attacks and polarizer filtration, thus LAPA, to evaluate `Laser Shield`.

4.3 Constructing LAPA

Simulating `Laser Shield` in digital space is not sufficient. Due to the complex and disturbing conditions in the real world, we construct LAPA to evaluate `Laser Shield`. In general, the construction of dataset needs a workload of collecting and cleaning [8]. We follow this principle with considerations of devices, main steps, hyperparameters, and data distribution for maximal quality.

Devices. Figure 3 shows the devices and their arrangement during data construction. We use Leica cameras which support a resolution of 8192×6144 to ensure images' quality. The linear polarizer GSP-50B with a polarization extinction ratio (PER) of 100:1 and a diameter of 50.4mm is chosen. And it is mounted on a GMK-0104 rotator to control the polarizer angle. Laser attacks are reproduced by three laser pointers (red, blue, and green) of 200 mW. In addition, we use three traffic signs as the attack and defense targets.

Main steps. Laser attacks mainly happen during the night when laser beams are more visible. To collect data of LAPA, we first choose a dark and stable environment, an optical laboratory, to deploy the above devices. Then, we adjust, with stands, the relative distance and height between devices to put traffic sign and laser beam in the camera field. Then, we fix the polarizer just in front of the camera

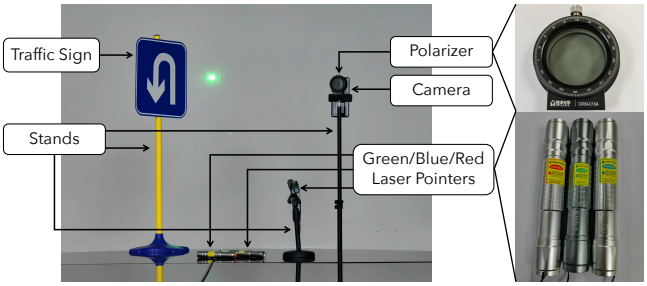

Figure 3: Experiment devices to construct LAPA.

to simulate the installation within the lens. The mode of camera is switched to "professional" to avoid image distortion made by intelligent algorithms. With changes in laser colors, laser propagation directions, polarizer angles, and traffic signs, we take images to construct LAPA.

Hyperparameters. The main steps are actually determined by multiple refinements, where we collect, test, and clean several samples to choose the critical hyperparameters for laser attacks and defenses. The first hyperparameter is the type of traffic sign. We choose three with different colors (blue U-turn, red no-passing, and yellow crosswalk) from the most common traffic signs to broaden the evaluation scope. In complementary to that, the second hyperparameter is the color of laser beams since it causes different visual effects under the background of traffic sign colors. The third hyperparameter is the laser propagation direction because laser can come from any direction in physical world scenarios. Last but not least, the fourth hyperparameter has to be the polarizer angle, since it is highly correlated to the filtration effect of lasers.

Data distribution. Due to the uniform sampling of hyperparameters, LAPA has a uniform data distribution. It has in total 27 scenes (3 types of traffic sign $\times$ 3 laser colors $\times$ 3 laser propagation directions). For each scene, we take 1 image without the polarizer, and 10 images with the polarizer but in different angles. In addition, we take 3 images with only traffic signs to exclude their effect during target model's prediction. The total number of LAPA is therefore $27 \times 11 + 3 = 300$. It is worth noting that, depending on the attack effect, LAPA can be divided into two subsets (approximately 50% to 50%), one of successful laser attack and one of unsuccessful attack (control cases). The former is to evaluate DSR and FTR, and the latter is to evaluate FPR.

4.4 Evaluation on LAPA

Since LAPA provides a diversity of laser attack conditions, the evaluation of `Laser Shield` is more convincing. To meet the defense requirements as stated in Section 2.2, we show its effectiveness in laser elimination, functionality-preserving, negligible computation cost, and high fault tolerance.

Laser elimination and functionality-preserving. Table 2 shows that `Laser Shield` almost achieves a perfect effect on laser elimination and functionality-preserving, with DSR and FPR close to 100%. The baselines take effect in several cases but exhibit poor performance overall. BdR and PD work well on FPR but not on DSR. This is because they mitigate adversarial effects during image preprocessing with a subtle modification. Such modification is hard to control, leading to a trade-off between DSR and FPR. Serving

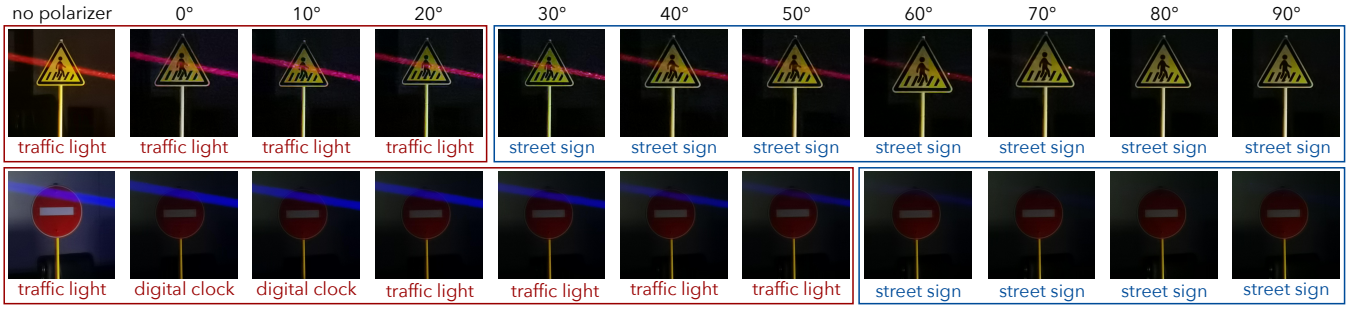

Figure 4: Laser Shield gradually takes effect with polarizer angle approaching θ_{opt}. The overall FTR on LAPA is 60.0%.

Table 2: DSR, FPR, and A/E-cc (analytical/experimental computation complexity) of Laser Shield compared to baselines.

Metric	Method			
	BdR	PD	DM	Laser Shield
DSR (%)	33.3	44.4	88.9	**100**
FPR (%)	88.9	88.9	83.4	**100**
A-cc	$O(mn)$	$O(K)$	$O(mn \cdot T \cdot C)$	$\boldsymbol{O(mn)}$
E-cc (s)	6.6×10^{-4}	8.1×10^{-3}	3.5	$\mathbf{3.5 \times 10^{-4}}$

as the SOTA of image restoration, DM gives the most competitive results among baselines, while it is still worse than Laser Shield. Indeed, even a single failure case is not tolerant in the ADS scenarios because the consequences (e.g. car crash) are irreversible. Besides, Figure 4 shows visually how Laser Shield eliminates laser. As the polarizer rotates, the adversarial laser beams weaken until completely disappear at the optimal angle θ_{opt}. Besides, the polarizer does not reduce the visibility of traffic signs regardless of the rotation state. This is due to the difference between natural light from traffic signs and linear polarized light from lasers, as we stated in Section 3.2.

Negligible computation cost. Table 2 also shows the computation complexity comparison in processing an image of size $m \times n$. Analytically, Laser Shield's computation cost is only on RGB energy since the filtration of the polarizer is not related to algorithms. Its complexity is therefore $O(mn)$. Similarly, BdR has the same complexity because it modifies all pixels. For PD, it randomly samples K pixels and replaces the value within a small neighborhood. Since K is on the same degree order of mn, its complexity is also close to Laser Shield. Despite the competitive performance, DM exhibits enormous complexity which is increased by the iterative time T and the cost of neural networks C. The experimental result is coherent with the analytical complexity. Laser Shield exhibits the minimal time cost, which is negligible compared to the computation cost of ResNet50. BdR and PD have comparable results, while DM is considered extremely inefficient.

High fault tolerance. Figure 4 also shows the high fault tolerance of Laser Shield. The target model's prediction is correct at more than just θ_{opt} where laser beams' intensity is minimal but at many polarizer angles. For instance, the prediction of the yellow crosswalk (resp. the red no-passing) is correct on seven (resp. four) angles out of ten, thus $FTR = 70\%$ ($FTR = 40\%$). Besides, the overall FTR on LAPA is 60.0%. This indicates that Laser Shield has a wide tolerance range. When the mechanical rotator or the RGB energy is occasionally out of alignment, our defense still works.

4.5 Evaluation on Road

To ensure a maximal quality of images and exclude irrelevant factors, the construction of LAPA is done in an optical lab. Nevertheless, such a stable environment will cause a subtle difference to real-world scenarios because the dynamic light conditions of passing vehicles are ignored. To fill this gap, we implement an evaluation on road by choosing three arbitrary scenes: a red stop sign, a blue bicycle lane sign, and a white car. We also use the Local Interpretable Model-Agnostic Explanations (LIME) [15] to explain which regions of images play an important role in the target model's prediction.

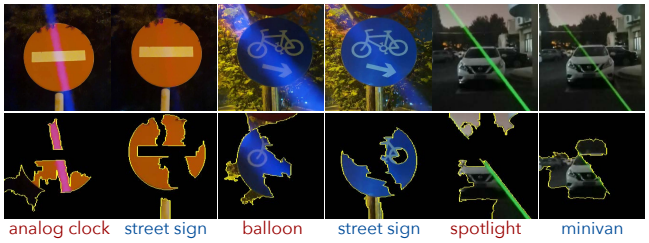

Figure 5: Evaluation on road with LIME explanation.

Figure 5 shows that Laser Shield works well on road. Under laser attacks, the red stop sign (resp. blue bicycle lane sign) is wrongly recognized as "analog clock" (resp. "balloon"), but Laser Shield almost completely eliminates lasers to draw back the correct prediction. In addition, we evaluate a new target object other than traffic signs, a white car. Laser Shield still works to correct the prediction from "spotlight" to "minivan". Besides, it is worth noting that with LIME explanation, we understand that laser beams are exactly the factor to mislead the prediction. Indeed, LIME excludes (resp. includes) lasers in images with correct (resp. wrong) predictions. This is because laser beams have a high saliency in images during the target model's prediction. Laser Shield takes effect by mitigating such saliency.

4.6 Ablation Study

Effectiveness of RGB energy. As stated in Section 3.3, the filtration effect of the polarizer is enhanced by the mechanism of min-energy rotation where we propose RGB energy to indicate the

state with minimal laser. Figure 6 shows its effectiveness compared to other possible indicator functions. For each polarizer angle, we compute, normalize, and plot the values of image intensity, RGB intensity, image energy, and RGB energy. Since Equation 2 modelizes the filtration effect as a sinusoidal function, we observe that the curve of RGB energy (red) is the closest to the ground truth (grey), also with the lowest MSE (mean-square error). Even if the image intensity is literally closer to laser intensity, it is less effective than energy because the ambient light will influence it (mean of pixel values) while energy (with square operation) amplifies the variation of lasers. Besides, we observe that the adaption from energy to RGB energy boosts the effect, confirming our analysis in Section 3.3.

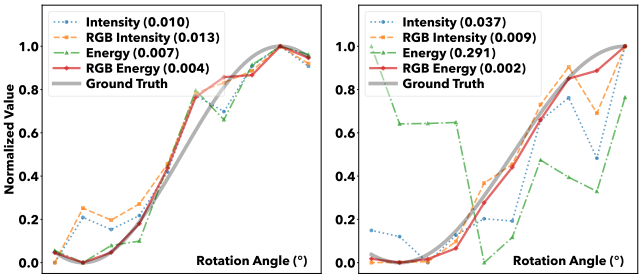

Figure 6: Four indicator functions on two scenes. RGB energy is the closest to the sinusoidal ground truth. Mean-square error (MSE) is shown in parentheses.

Robustness to diverse attack conditions. To evaluate Laser Shield's robustness, we vary the attack conditions to see the influence on defense results. Such a consideration is already taken in the construction of LAPA. As stated in Section 4.3, LAPA consists of a diversity of laser attacks across three dimensions (traffic sign types, laser colors, and laser propagation directions). Since Laser Shield achieves 100% DSR and 100% FPR on it, we conclude that our method exhibits robustness to diverse attack conditions.

5 DISCUSSION AND FUTURE WORK

Extension of LAPA. Current LAPA contains only limited categories of traffic signs which are all collected in the optical lab. However, practical ADS will face more complex environments including more objects (e.g. vehicles, pedestrians, and traffic lights, etc.) to recognize. A thorough analysis and evaluation of laser attacks against ADS is needed which requires more physical world datasets collected with lasers. We will extend LAPA with more objects in ADS scenario under laser attacks as our first future work.

Physical deployment. Although we try to guarantee the quality of LAPA and have evaluated Laser Shield under diverse attack conditions, until now we cannot implement Laser Shield on a real ADS due to the limitation of experimental conditions. We leave the full physical deployment, including the installation of a polarizer along with a mechanical rotator for min-energy rotation on an ADS as our second future work.

6 CONCLUSION

Facing the new threat of laser attacks in ADS, we propose Laser Shield which leverages a polarizer along with a specifically designed min-energy rotation mechanism to eliminate adversarial lasers from driving environments. To evaluate its performance, we construct a physical world dataset LAPA with diverse attack conditions. Through thorough experiments in digital and physical space, Laser Shield can effectively eliminate laser, preserve DNN's functionality, have high fault tolerance, and need negligible computation cost. We leave the potential physical deployment of Laser Shield along with an ADS as our future work.

ACKNOWLEDGEMENT

This work is supported by National Natural Science Foundation of China No. 62106127. The authors would like to thank Maosen Zhang and Yikun Wang from Tsinghua University, China for constructing LAPA, and Yi Sun from Ningbo High School, China for teaching the principle of polarization.

REFERENCES

[1] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. 2021. Recent advances in adversarial training for adversarial robustness. *arXiv preprint arXiv:2102.01356* (2021).
[2] Yang Bai, Yuyuan Zeng, Yong Jiang, Yisen Wang, Shu-Tao Xia, and Weiwei Guo. 2020. Improving query efficiency of black-box adversarial attack. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*. Springer.
[3] Mark R Banham and Aggelos K Katsaggelos. 1997. Digital image restoration. *IEEE signal processing magazine* (1997).
[4] Pranav Singh Chib and Pravendra Singh. 2023. Recent Advancements in End-to-End Autonomous Driving using Deep Learning: A Survey. *IEEE Transactions on Intelligent Vehicles* (2023).
[5] Russell A Chipman, Wai Sze Tiffany Lam, and Garam Young. 2018. *Polarized light and optical systems*. CRC press.
[6] Edward Collett. 2005. Field guide to polarization. Spie Bellingham, WA.
[7] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. 2023. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).
[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*.
[9] Yao Deng, Tiehua Zhang, Guannan Lou, Xi Zheng, Jiong Jin, and Qing-Long Han. 2021. Deep learning-based autonomous driving systems: A survey of attacks and defenses. *IEEE Transactions on Industrial Informatics* (2021).
[10] Ranjie Duan, Xiaofeng Mao, A Kai Qin, Yuefeng Chen, Shaokai Ye, Yuan He, and Yun Yang. 2021. Adversarial laser beam: Effective physical-world attack to DNNs in a blink. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
[11] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).
[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.
[13] Kenneth I Laws. 1979. Texture energy measures. In *Proc. Image understanding workshop*.
[14] Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo, and James Storer. 2018. Deflecting adversarial attacks with pixel deflection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.
[15] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*.
[16] Weilin Xu, David Evans, and Yanjun Qi. 2017. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155* (2017).
[17] Chen Yan, Zhijian Xu, Zhanyuan Yin, Xiaoyu Ji, and Wenyuan Xu. 2022. Rolling Colors: Adversarial Laser Exploits against Traffic Light Recognition. In *31st USENIX Security Symposium (USENIX Security)*.
[18] Qingjie Zhang, Maosen Zhang, Han Qiu, Tianwei Zhang, Mounira Msahli, and Gerard Memmi. 2023. ATTA: Adversarial Task-transferable Attacks on Autonomous Driving Systems. In *IEEE International Conference on Data Mining*.