Natural Light Can Also be Dangerous: Traffic Sign Misinterpretation Under Adversarial Natural Light Attacks

Teng-Fang Hsiao[1], Bo-Lun Huang[1], Zi-Xiang Ni[1], Yan-Ting Lin[1], Hong-Han Shuai[1]
Yung-Hui Li[2], Wen-Huang Cheng[3]
[1]Department of Electronics and Electrical Engineering, National Yang Ming Chiao Tung University, Taiwan
[2]Hon Hai Research Institute, [3]Information Engineering Graduate Institute of Taiwan University
{bluedyee.ee09, kevin503.ee09, z058z.ee09, andylin.ee09, hhshuai}@nycu.edu.tw
yunghui.li@foxconn.com, wenhuang@csie.ntu.edu.tw

Abstract

Common illumination sources like sunlight or artificial light may introduce hidden vulnerabilities to AI systems. Our paper delves into these potential threats, offering a novel approach to simulate varying light conditions, including sunlight, headlights, and flashlight illuminations. Moreover, unlike typical physical adversarial attacks requiring conspicuous alterations, our method utilizes a model-agnostic black-box attack integrated with the Zeroth Order Optimization (ZOO) algorithm to identify deceptive patterns in a physically-applicable space. Consequently, attackers can recreate these simulated conditions, deceiving machine learning models with seemingly natural light. Empirical results demonstrate the efficacy of our method, misleading models trained on the GTSRB and LISA datasets under natural-like physical environments with an attack success rate exceeding 70% across all digital datasets, and remaining effective against all evaluated real-world traffic signs. Importantly, after adversarial training using samples generated from our approach, models showcase enhanced robustness, underscoring the dual value of our work in both identifying and mitigating potential threats.[1]

1. Introduction

In the modern technological landscape, machine learning (ML) models have catalyzed significant advancements across myriad applications, from optimizing consumer experiences to propelling the evolution of autonomous vehicles [7, 34, 31]. Yet, with these breakthroughs comes an increasing concern about the model susceptibility to manipulations that can severely undermine their efficacy and dependability. Central to this apprehension is the phenomenon

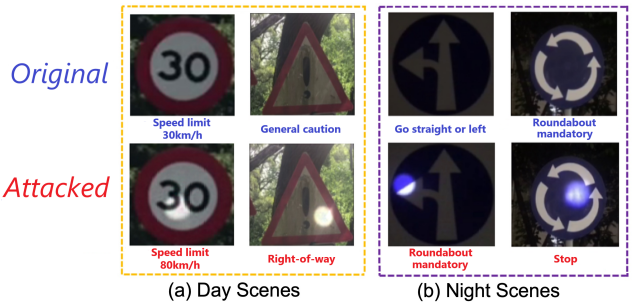

Figure 1. **An illustrative example of the presence of light interference for misleading classifiers.**

of "Adversarial Attacks" [13]. These sophisticated disruptions, validated by current research, can deceive model predictions, producing unintended and often harmful outcomes. This vulnerability extends beyond the digital environment—where the majority of ML models are developed and validated—manifesting in the physical realm, thereby jeopardizing not only the model performance but also the safety of humans dependent on them.

Specifically, transitioning adversarial techniques from the digital plane to the tangible world, or the "Digital to Physical (D2P)" process [17], introduces its own set of intricacies. The direct implementation of digital distortions often falls short in the real world, owing to the unpredictable variability of lighting, angles, and myriad physical conditions. Some researchers are aiming to bridge this gap by exploring the transformation interplay between digital and physical entities [8, 12, 17]. Yet, these pursuits often culminate in over-optimized solutions tailored for specific devices, lacking broad applicability. Alternatively, more conspicuous interventions involving stickers [6] or lasers [5] have proven effective in both realms but suffer from their overt detectability, reducing real-world feasibility.

[1]Project page: https://github.com/BlueDyee/natural-light-attack

(a) Real world (b) L simulation [39]

Figure 2. **A comparison of simulated shadow[39] and light using same method.**

In this paper, we study a new kind of physical attacks, namely, natural light attacks. As visualized in Fig. 1, imagine an autonomous vehicle misinterpreting a dimly lit stop sign at night, mistaking reflections caused by its own headlights for genuine signals. Such precarious situations could also be induced by various light sources, such as direct or reflected sunlight, thereby happening even in the absence of a malicious agent. Other possible light sources that can create similar effects include sunlight, reflected sunlight from buildings, and flashlight illuminations. Different from previous works, e.g., lasers [5] or projectors [12, 24, 27], natural light attacks cannot use dedicated patterns and may happen even without any attackers.

To empirically support our claims, we spotlight traffic sign recognition—an indispensable facet of autonomous systems. While previous works have dabbled with shadows for adversarial intent [39], the inherent contrast between light and shadow necessitates distinct generative methodologies. As shown in Fig. 2, the method used to simulate shadow [39] is incapable of replicating the saturation change in red region or the reflections observed in the black region, both of which are caused by the light. Therefore, we propose a novel light simulator that takes real-world traffic sign images and images subjected to varying light perturbations as the input. The primary objective of this simulator is to rapidly adapt to diverse traffic signs and light sources. Utilizing this simulator, we pinpoint light perturbations that misguide classifiers, causing misidentifications. Our method adopts the well-established Zeroth-Order Optimization (ZOO) paradigm [2, 3, 4, 35] to refine the physical perturbations. Extensive experiments conducted on the benchmark GTSRB and LISA datasets reinforce our assertions, with an attack success rate (ASR) of 70%, emphasizing the potential risk posed by natural light conditions. This investigation highlights the urgent requirement to incorporate these conditions in robust ML model development for resilience against unforeseen threats. For instance, integrating adversarial training can serve as a formidable countermeasure, substantially diminishing the ASR.

Our contributions can be summarized as follows:

- We highlight the potential of natural light as a novel avenue for adversarial attacks, capable of misleading traffic sign recognition models. This susceptibility can be effectively replicated with accessible light-emulating sources like flashlights or sunlight.
- Our work involves the development of a light simulator, designed to quickly adapt to diverse tasks while accurately simulating the interaction between varied light sources and objects. In addition, we pioneer the integration of Zeroth-Order Optimization (ZOO) into physical black-box attacks.
- We perform exhaustive evaluations across both digital and physical realms, accounting for variations in day and night settings. Our results indicate the effective performance of our proposed approach in both simulated and real-world scenarios, demonstrating its adaptability to diverse environments.

2. Related Works

2.1. Digital Adversarial Attack

Adversarial attacks are being recognized as a potent concern within the AI research community. Initially, these attacks were concentrated on altering digital images by leveraging backpropagation, manipulating the input image to maximize the loss to the true label or minimize the loss of a target label. This form of optimization can be implemented using methods such as PGD [23], FreeAT [33], YOPO [38], and ACG [37]. Another popular kind of methods for generating adversarial perturbations is to utilize generators. Nonetheless, certain situations only allow access to the model output, with gradient information being unavailable, creating the challenge of "black-box attacks" [30]. Specifically, black-box attacks can be categorized into two types: decision-based and score-based attacks. Decision-based attacks, where the attacker can only infer the predicted label, can be applied to not only Deep Neural Networks (DNNs) but also other structures like Support Vector Machines (SVMs), Decision Trees, or any model deployed via cloud APIs. The second category, score-based attacks, enables the attacker to infer the confidence score as their attack objective [3, 16, 20]. This method offers more information about the model's output, allowing for more precise attacks.

2.2. Physical Adversarial Attack

Building on the success in the digital domain, a recent line of research has sought to transpose digital attacks into real-world settings. For example, Kurakin et al.[19] demonstrated that adversarial examples maintain their adversarial properties even when translated to the physical world, albeit with a slight reduction in Attack Success Rate (ASR). Consequently, numerous studies have aimed to address this

(a) advLB [5] (b) Shadow [39] (c) OPAD [12] (d) Ours

Figure 3. **Visual comparisons with previous works.** Our proposed attack is not only stealthy but can also be found in real-life scenes.

digital to physical challenge [1, 6, 17] either by contemplating the expectations of transformations or by accounting for printer and camera properties. In parallel, other researchers have attempted to alter existing physical objects to deceive models, using tools like stickers [6], projectors [12, 24, 27], or lasers [5]. However, such methods can be visually conspicuous or be constrained to certain environments. Consequently, several studies have pivoted towards unrestricted, natural-like attacks, exploring strategies such as natural-like adversarial patches [15] or shadow manipulation [39]. Distinct from previous techniques, our proposed natural light attack introduces a unique adversarial approach. This method not only integrates seamlessly with real-world environments but also offers enhanced ease of implementation, for instance, through simple means such as a flashlight or mirror, as depicted in Fig. 3.

3. Natural Light Attack

3.1. Problem Formulation

Given the input image $x \in \mathbb{R}^{H\times W\times C}$ with the corresponding true label $y \in [1, \cdots, k]$, and a victim classifier $f : \mathbb{R}^{H\times W\times C} \to \mathbb{R}^k$ associated with a confidence score $f_i(x)$ to the i-th class, the predicted label $\widetilde{y}$ is derived as follows.

$$\widetilde{y} \triangleq \arg\max_i f_i(x). \tag{1}$$

Here, the goal of the proposed natural light attack is to project a specific light onto the target object to generate adversarial example x_{adv}, leading to the misclassification, i.e.,

$$\arg\max_i f_i(x) \neq \arg\max_i f_i(x_{adv}). \tag{2}$$

To generate the image for natural light attack x_{adv}, we consider the location of the light, which is represented by the corresponding mask $\mathcal{M}_P \in \mathbb{R}^{H\times W}$. Regions of $\mathcal{M}_P$ that are illuminated are assigned a value of 1, whereas unilluminated areas receive a value of 0. The mask $\mathcal{M}_P$, dictating different shapes of light interference, is determined by a parameter set $\mathcal{P}$. For instance, a circular mask $\mathcal{M}_{circle}$ can be dictated by $\mathcal{P} = \{m_c, n_c, r\}$, where (m_c, n_c) and r represent the center coordinate and the radius, respectively. An ellipse mask $\mathcal{M}_{ellipse}$ can be depicted by $\mathcal{P} = \{m_c, n_c, a, b, \phi\}$, where a, b, and ϕ are respectively the major axis length, minor axis length, and the rotation angle. Here, we use the circular shape due to the computational efficiency. It is worth noting that manipulating circular light in real-world situations is also comparably simpler. In the following, we present how to derive the mask by zeroth-order optimizers.

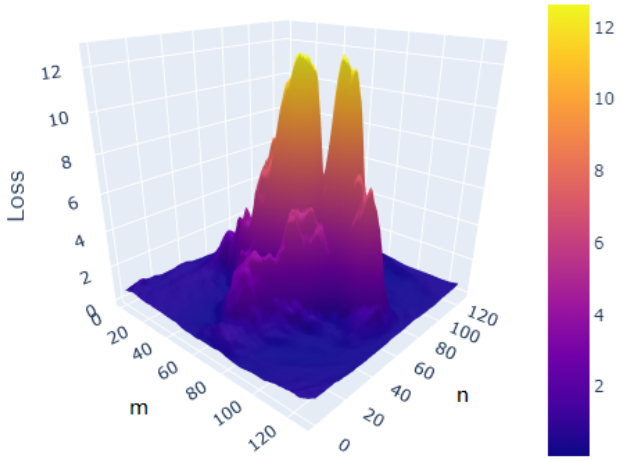

Figure 4. **A loss visualization of different light perturbations on a single photo.** This loss map demonstrates the variation in loss across different positions $(m_c, n_c, r = 10)$ of circle light interference. This map reveals that our proposed attack can be optimized via derivative-based method, which has mathematically proven efficacy that is superior to previous unrestricted attacks [5, 39] that relied on heuristic algorithms, such as PSO.

3.2. Zeroth-Order Attack Optimizer

The primary focus of the proposed attack is to search for a set of parameters $\mathcal{P} = \{m_c, n_c, r\}$ for $\mathcal{M}_P$ that can cause misclassification by the classifier. In our approach, we specifically address the practical black-box scenario, where the information of the victim model is unknown, except for the confidence score $f(x)$ and true label y. Our criteria for optimization is based on the cross-entropy loss between the $f(x)$ and y. With the objective of maximizing the cross-entropy loss $\mathcal{L}_{CE}$, we aim to construct an adversarial image by solving the following optimization problem:

$$\arg\max_{\mathcal{P}} \mathcal{L}_{CE}(y, f(x + \mathcal{G}(x, \mathcal{M}_{\mathcal{P}})), \tag{3}$$

where the adversarial noise is generated by the natural light generator $\mathcal{G}(x, \mathcal{M}_{\mathcal{P}})$ controlled by the mask $\mathcal{M}_{\mathcal{P}}$.

With the recent advancements in the field of optimization, numerous variants of Zeroth-Order Optimization (ZOO) have emerged [4, 10, 22], providing different strategies and techniques. Considering these options, we incorporated the concept of ZOO-SCD [21] into our mask optimization approach due to its efficiency in handling low-

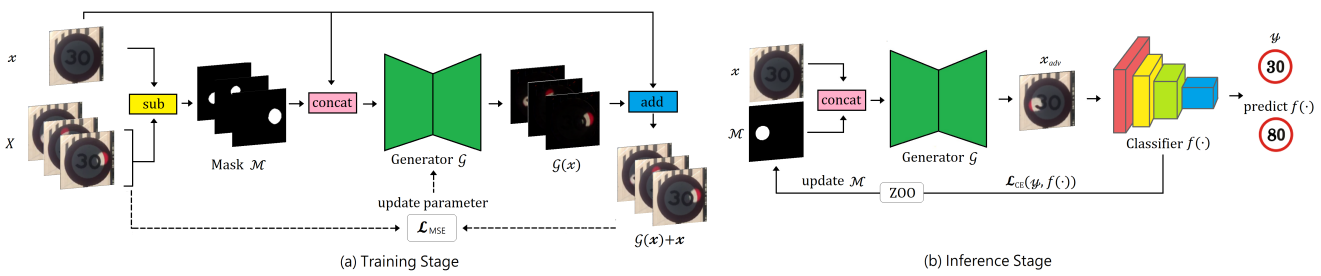

Figure 5. **Overview of our proposed pipeline:** (a) In the training stage of our Light Generator, we aim to mimic real light images given a mask by minimizing the $\mathcal{L}_{MSE}$ between simulated and real light perturbations. Second, in the Inference stage of the ZOO optimizer (b), we further combine the light generator with the ZOO optimizer to find a physically reproducible attack image, aiming to maximize the $\mathcal{L}_{CE}$ of the victim classifier.

dimensional spaces. Specifically, we applied circular light interference to each pixel of an image and recorded the cross-entropy loss value between confidence score $f(\cdot)$ and true label y. Fig. 4 illustrates the loss map with different light location but with a fixed radius ($r = 10$).

To identify the optimal position $\mathcal{M}_\mathcal{P}$ for generating an attack image x^*_{adv} with a maximized misclassified score, we need to find the maximum loss on a 2D surface. Therefore, at each starting point $\mathcal{P}$, we can utilize ZOO to determine the direction of the maximum gradient by evaluating the change in score resulting from adjusting the estimate step size in each dimension $\mathcal{P}_j \in \mathcal{P}$. Consider a 2D surface as an example, where the coordinates $(\mathcal{P}_1, \mathcal{P}_2)$ can also be represented as (m_c, n_c). We update $\mathcal{P}$ in the direction indicated by the gradient (δ_m, δ_n) with a step size of γ, i.e., moving to $(m_c + \gamma \frac{\delta_m}{||(\delta_m,\delta_n)||_2}, n_c + \gamma \frac{\delta_n}{||(\delta_m,\delta_n)||_2})$. Following this algorithm, we progressively find the maximum loss value step by step. Please note that if the objective also includes finding the optimal radius, we can simultaneously consider the gradient of the radius r while applying the ZOO optimization. Hence, by considering gradients of the pertinent parameters $\mathcal{P}$, a mask of any desired shape can be sought, transforming them into the mask $\mathcal{M}_\mathcal{P} \in \mathbb{R}^{H \times W}$.

During the execution of the ZOO algorithm, certain challenges may arise. One such issue occurs when the gradients vanish, particularly in regions where the points lie on a flat surface. To address this problem, we implement a precautionary measure by initially checking the gradient of each point [11, 23]. When the magnitude of the gradients is smaller than the pre-defined threshold τ, we restart the optimization process from another randomly chosen point $\mathcal{P}$. This approach also provides an opportunity to escape potential local maxima. Additionally, due to the sparseness and discrete nature of our problem, we repeat the entire process k rounds to introduce more randomness. Due to the space constraint, the pseudocode can be found in the supplementary materials.

3.3. Object-Dependent Natural Light Generator

To simulate a realistic natural light, one possible way is to adjust the "L" dimension in the CIELab color space as the simulation for shadows [39]. However, we want to emphasize that light interference with objects depends on objects themselves. In other words, visually similar objects could exhibit differences under identical light projection due to their textures. Inspired by OPAD attack [12], which employs diverse color projections to compute the object-wise and pixel-wise color transformations, we propose a supervised end-to-end generator-based approach to simulate the impacts of various light sources on different objects.

Fig. 5(a) illustrates the proposed light generator. Specifically, to train a generator $\mathcal{G}_\theta$ parameterized by θ, we first collect a paired training data, containing a undisturbed image set X and a corresponding light-perturbed image set $\tilde{X}$. The object-dependent natural light generator takes the undisturbed image $x \in X$ concatenated with the mask $\mathcal{M}_P$ as input. The synthesized light-perturbed image x_{sim} is computed by adding the light perturbation derived from the generator's output $\mathcal{G}_\theta(x, \mathcal{M}_P)$ to the original image, i.e.,

$$x_{sim} = x + \mathcal{G}_\theta(x, \mathcal{M}_P). \tag{4}$$

To supervise the generator, we use the Mean Squared Error (MSE) between the groundtruth $\tilde{x} \in \tilde{X}$ and the generated output x_{sim} as the loss function. The loss of object-dependent natural light generator can be calculated as follows.

$$\mathcal{L}(\mathcal{G}_\theta, x, \tilde{x}, \mathcal{M}_\mathcal{P}) = \frac{1}{H \times W} \|x_{sim} - \tilde{x}\|^2. \tag{5}$$

Equipped with an optimizer, we can obtain the optimal parameters θ of object-dependent natural light generator by minimizing the empirical loss on the whole dataset $(X, \tilde{X})$.

$$\theta^* = \arg\min_\theta \mathbb{E}_{x \in X, \tilde{x} \in \tilde{X}} \mathcal{L}(\mathcal{G}_\theta, x, \tilde{x}, \mathcal{M}_\mathcal{P}). \tag{6}$$

In the following, we present how we collect data in digital domain and physical domain.

3.4. Attack in Digital Domain

Our digital light generator is constructed using a training set comprising both undisturbed images, X, and light-perturbed photos, $\tilde{X}$. The generation of this training set is achieved through a carefully curated process. Initially, four representative images are selected for each class within the GTSRB and LISA datasets. The chosen images are particularly distinctive in their optical illuminations.

To simulate light interference, we manually adjust the LAB or HLS values of color spaces for each training image since the adjustment of these parameters varied across images. This process aimed to emulate natural-like virtual light sources, necessitating precise adjustments of the LAB or HLS values based on the visual appearance of each image to achieve an accurate representation. Once the suitable approach for manipulating the LAB or HLS parameters for each image is determined, we simulated light interference at random positions with varying radius on each image. This procedure was replicated 100 times, culminating in a diverse collection of light-perturbed photos. As a result, our training set encompasses immense undisturbed and light-perturbed images. This training set facilitates the training of our digital light generator, optimizing its ability to simulate and understand the nuances of light perturbations.[2]

3.5. Attack in Physical Domain

The challenge of accurately simulating physical transformations in the digital realm is a well-acknowledged issue. Instead of using the Expectation Over Transformation (EOT) approach commonly used in previous work [1, 27, 39], we take inspiration from [18] to bridge the gap between the digital and physical domains.

To begin with, two videos are captured for one scene: one clear and another with light perturbations. We then proceed to pinpoint the positions of the light sources, saving this information in masks $\mathcal{M}_{\mathcal{P}}$. To achieve this, we leverage the Euclidean distance within the CIELab color space between the clean and perturbed video frames. This method enables us to accurately delineate the illuminated areas. We further apply a median filter to effectively suppress noise, resulting in a smoother representation of the affected area.[3]

Utilizing the data produced through this method and the loss function outlined in Section 3.3, we refine our initial generator to simulate the specific physical environment with greater precision. Earlier work, such as Meta-Attack [8], combined meta-learning with physical attacks to create a class-agnostic attack pipeline. In a similar vein, we have the flexibility to fine-tune an existing simulator or employ the Meta-learning algorithm [9, 28, 29] to facilitate the adaptation process of our generator, ensuring its performance under varying circumstances.

[2]You can find the dataset in our project page `https://github.com/BlueDyee/natural-light-attack`.

[3]More implementation details can be found in Section 4.3.

4. Experiments

In this section, we first demonstrate the success of the proposed method in various natural conditions using existing digital datasets. Later, we showcase the effectiveness of the proposed method on real-world traffic signs. Finally, we discuss the results of the ablation study conducted on the proposed method.

4.1. Dataset and Target Model

Following [6], we use the same two datasets and classifiers. Specifically, the first dataset is LISA [25], which is a U.S. traffic sign dataset containing 47 different road signs. Following previous work, only 16 most common signs are included due to its unbalance distribution. The second one is GTSRB dataset, which is a German traffic sign dataset containing 43 different road signs. For the target victim models, we use three state-of-the-art classifiers, two from [6] and one from [32]. Specifically, [6] provides two classifiers: LISA-CNN and GTSRB-CNN, achieving 91% and 95.7% accuracy on their respective dataset. In our experiment, we referred to them as LISA-CNN-1 and GTSRB-CNN-1, respectively. Additionally, we introduced another classifier that achieved a 99.7% accuracy on the GTSRB Kaggle test [32]. Furthermore, we re-implemented this classifier on LISA and achieved 99.2% accuracy. We denoted them as GTSRB-CNN-2 and LISA-CNN-2, respectively. Due to space constraint, we provide the details of the model architecture in supplementary materials.

4.2. Evaluation in Digital Domain

Experiment Setting. Due to the difference in the physical light sources used during the day (solar) and at night (flashlight), we split the original dataset based on the average L value in CIELAB space of each image: images with the average L ranged in $(10, 30)$ were categorized as night traffic signs, while those with the average L greater than 60 were categorized as day traffic signs. We excluded extremely dark cases due to their low confidence, which makes them susceptible to attacks. We also removed the middle range of cases where it is challenging to distinguish between white signs in dark conditions and dark signs in day.

Results. To evaluate the performance of the proposed natural light attack, Tab. 1 shows the attack success rate (ASR) of the proposed method across varying query numbers against different classifiers. The results manifest that ASR becomes greater under nighttime conditions on GTSRB dataset (both models increase around 30%). However,

Dataset	Scene	Model	Original Accuracy	ASR at different queries								
				40	60	80	120	160	200	260	320	400
GTSRB	Day	CNN-1 [6]	91.5%	50.4%	57.6%	59.2%	65.1%	68.4%	66.8%	69.3%	73.1%	75.2%
		CNN-2 [32]	98.5%	49.2%	53.1%	55.9%	60.9%	66.8%	68.3%	69.9%	69.9%	71.9%
	Night	CNN-1 [6]	93.3%	93.8%	95.0%	94.6%	94.2%	96.3%	96.7%	95.4%	96.25%	97.5%
		CNN-2 [32]	98.8%	85.6%	89.9%	89.9%	93.4%	94.6%	94.2%	94.6%	94.2%	94.9%
LISA	Day	CNN-1 [6]	99.9%	76.0%	77.2%	81.9%	82.2%	83.2%	84.2%	85.3%	84.6%	86.5%
		CNN-2 [32]	99.9%	63.8%	70.4%	71.2%	73.0%	76.2%	76.5%	76.2%	75.4%	76.5%
	Night	CNN-1 [6]	99.2%	76.1%	78.8%	83.0%	87.3%	86.9%	89.2%	90.0%	90.0%	90.3%
		CNN-2 [32]	99.2%	94.2%	97.3%	98.1%	98.8%	99.2%	99.2%	99.2%	99.2%	99.2%

Table 1. **Attack Success Rate under Different Dataset and Queries.**

we do not observe this trend on LISA dataset. This is probably because its training data contains a large amount of night data, while GTSRB does not, making LISA more robust than GTSRB in nighttime scenarios. In terms of the number of queries, the results indicate that ASR becomes greater as the number of queries increases. However, in almost every case, the ASR saturates after 120 queries. It appears that the loss value obtained within 120 queries is sufficient to mislead the classifier. Even though higher query numbers may yield higher loss values, they do not contribute to an improvement in attack success rate. Consequently, we choose 120 queries to carry out attacks due to its efficiency and ASR (over 60% in all cases, with an average of 79%).

4.3. Evaluation in Physical Domain

Experiment Setting. To assess the efficacy and practicality of our proposed natural light attack, we ventured beyond simulation-based evaluations and took our tests into real-world settings. Specifically, we conducted our experiments around traffic signs situated throughout our campus. To ensure that our light generator was tailored to the intricacies of these real-world scenarios, we adopted a fine-tuning approach. For each target sign, we captured a 20-second video detailing the light interference experienced by that sign. The video data then served as the substrate upon which our light generator was refined, ensuring it was well-calibrated to the unique lighting conditions and presented by each specific environment. We then identified the light interference by labeling the pixels that displayed a significant visual difference in Euclidean distance within the CIELab color space. We then established whether a pixel belonged to the illuminated area based on a predefined threshold, and used a median filter to eliminate any noise. However, we noticed a challenge when dealing with traffic signs that featured black text; the Euclidean distance for the areas containing black text remained relatively small, even under illumination. To rectify this, we employed an average blur function, followed by several iterations of pixel clipping to zero for values below a certain threshold, after the denoising process. This technique effectively resolved the issue, allowing for the accurate determination of light locations and facilitating the fine-tuning of our light generator. After fine-tuning, we follow the pipeline in Fig. 5(b) to search the adversarial pattern via ZOO optimizer for simulating the adversarial pattern. Fig. 6 demonstrates several attacked images. It can also be observed that the results from real photos are similar to those from simulated photos. This indicates that our light generator can effectively simulate light after fine-tuning.

Results of Day Attack with Sunlight. To achieve a more controlled and refined light interference on road signs, we employed a round mirror to reflect sunlight onto the signs. We conducted experiments using four different signs, and the results of these experiments are presented on the left side of Fig. 6. It is evident that our method has notable attack efficacy, as all four real-working traffic signs became adversarial after light perturbation while maintaining a visually unsuspicious appearance.

Results of Night Attack with Flashlight. For nighttime conditions, we utilized flashlight as the source of light interference, which is commonly used for night illumination. We also conducted experiments using the same four road signs as in the daytime. As shown in Fig. 6, we can observe that the visual differences due to light interference are more pronounced compared to the daytime condition, while still keeping a stealthy appearance, especially in the samples in the middle two rows, leading to a more successful attack. At the bottom row of Fig. 6, the confidence of misleading the sign from "General caution" to "Right-of-way at the next intersection" increased significantly, rising from 62% to 99%.

4.4. Ablation Study

Here, we conduct an ablation study to understand the contributions of each component of natural light attack.

Improvement of Using ZOO. We compared ZOO to the optimization methods proposed by previous works, namely "Random Restart Search" in AdvLB [5] and "PSO" in Shadow [39], with some modifications to fit our task. The

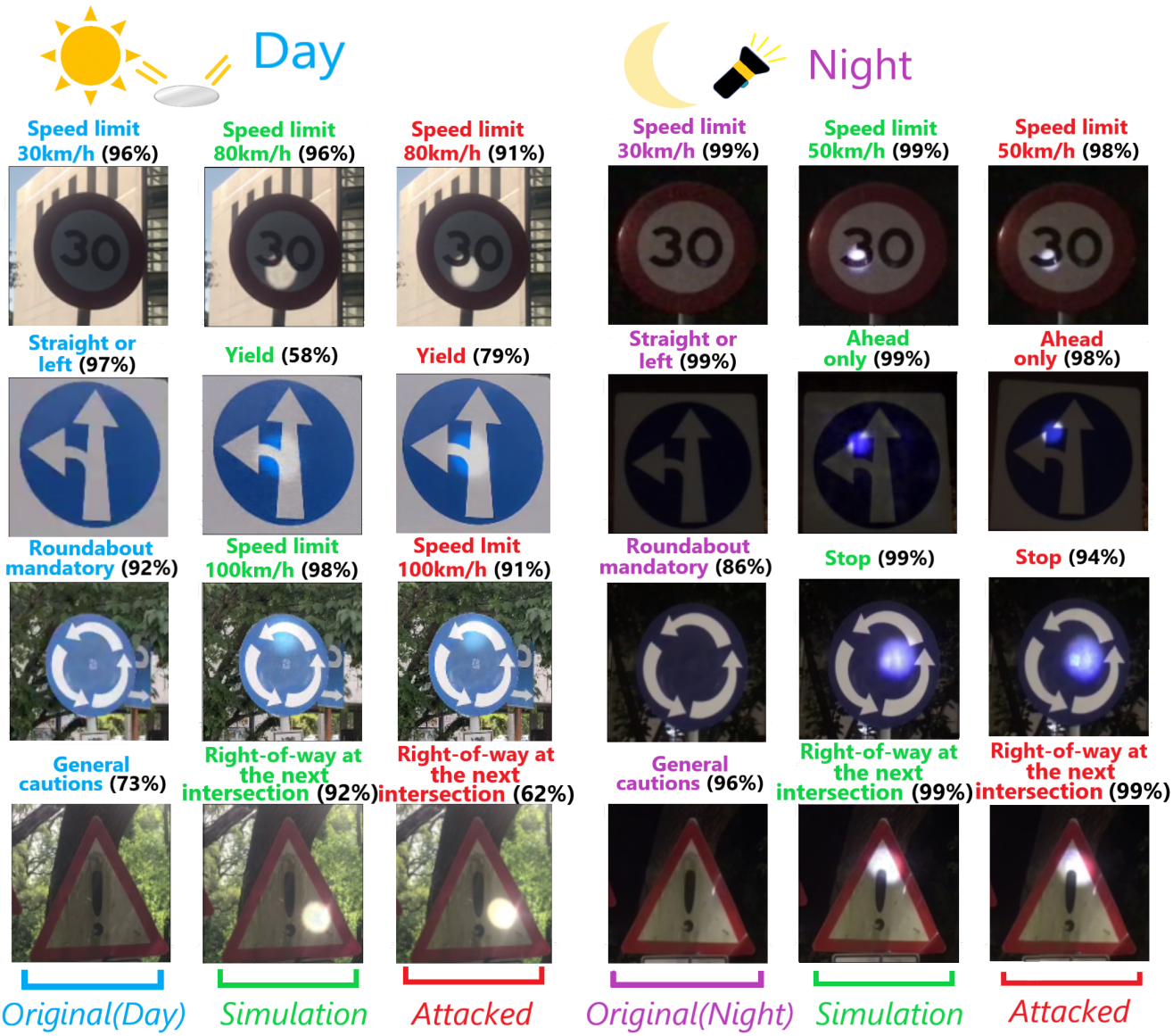

Figure 6. **A demonstration of our Natural Light Attack on real working signs under different conditions.** For each image, we indicate the corresponding prediction (colors varying based on the type of image) and the confidence score (shown in black) provided by the victim classifier. This figure shows that all four out of four real working traffic signs can be successfully attacked during the day with a mirror or at night with a flashlight, both of which are light sources commonly encountered in our daily lives.

results are shown in Figs. 7 and 8 in supplementary material . Firstly, the loss generally increases with the number of queries across all cases, albeit with some noise. Secondly, while ZOO performs slightly worse initially, it consistently outperforms the other methods as the number of queries increases. This can be attributed to ZOO's convergence property, improving steadily with more updates, unlike the other methods that rely more on randomness unrelated to the number of queries.

Different Shapes. As stated in Sec. 3.2, our ZOO framework simplifies the optimization of different shapes, allowing evaluation of their performance. We compared the performance of several shapes, namely "Circle," "Ellipse," "3- **D P**" (where "dimensional" is abbreviated as "**D**" and "polygon" as **P**), "4-**D P**," and "5-**D P**." We conducted evaluations on 100 test data samples from the each dataset (Tab. 2). All shapes were optimized using 120 queries.

The results in the tables indicate that the "Circle" shape demonstrates superior time efficiency and attack effectiveness compared to the other shapes. This could be attributed

GTSRB	Circle	Ellipse	3-D P	4-D P	5-D P
ASR	81.3%	82.3%	68.8%	78.1%	79.2%
Avg Time	7.4s	46.2s	28.6s	49.2s	84.5s
LISA	Circle	Ellipse	3-D P	4-D P	5-D P
ASR	63.2%	50.5%	29.5%	40.0%	43.1%
Avg Time	7.3s	49.3s	30.6s	52.1s	86.7s

Table 2. **Comparison of different shape of light in digital evaluation** ("dimensional" is abbreviated as "**D**").

		JPEG	R&P	NRP	advTrain (Ours)
GTSRB	CNN-1	61(-4)	67(+2)	68(+3)	**18**(-47)
	CNN-2	66(+5)	63(+2)	64(+3)	**17**(-44)
LISA	CNN-1	76(-6)	76(-6)	78(-4)	**21**(-61)
	CNN-2	68(-5)	77(+4)	77(+4)	**12**(-61)

Table 3. **Our attack under different defences.**. We report the corresponding ASR(%) in day with 120 queries under different defenses. The decrease and increase of ASR are respectively highlighted by green and red colors.

not only to the faster calculation of circles but also to their rarity in the training data of classifier as "perfect" circles are less common compared to other shapes. Since "Circle" is a special case of "Ellipse", "Ellipse" may achieve more promising results. However, it might require more iterations for 'Ellipse' to surmount the performance of "Circle".

Pre-training vs. Meta Learning. For more precise simulations, adapting our light generator to fit specific objects and light sources is imperative. Two methods are possible to fulfill it: pre-training and meta learning [4]. Both methodologies utilized 20 distinct digital signs as digital sources (resulting in 8000 pairs of photos) and 5 physical signs as physical sources (yielding 7500 pairs of photos) for training data. The amalgamation of digital and physical sources constituted a hybrid source.

After applying either of the two approaches, we fine-tuned our model on a dataset composed of physical signs (comprising 1500 pairs of photos). The hyperparameters used were identical to those in the "Pre-Train" scenario, but training extended for 61 epochs during the fine-tuning stage. The results are illustrated in Fig. 8. Fig. 8(a) suggests that digital sources do not enhance the performance of the physical task with meta-learning, whereas pre-training does contribute positively. In contrast, Fig. 8(b) demonstrates that physical sources can augment the performance under both methodologies, a finding that aligns intuitively with our expectations. Ultimately, we selected the "Pre-Train" method, coupled with fine-tuning, as our final adaptation technique. This selection was motivated by its capacity to effectively utilize both types of sources, as shown in Fig. 8(c).

4.5. Defense of Natural Light Attack

Although a robust defense against unrestricted attacks remains an area of active research, we have undertaken comprehensive evaluations of a spectrum of defense methods. These span from preprocessing strategies such as JPEG [14], R&P [36], and NRP [26], to more involved adversarial training methodologies [23]. As evidenced in Tab. 3, adversarial training emerges as the most potent countermeasure against our proposed natural light attack. In other words, the robustness of traffic sign classification models is substantially bolstered when trained with adversarial samples generated by our approach. This suggests an intrinsic value in our proposed attack not just as a challenge but as a tool for model enhancement. Conversely, other generic defense strategies seem to falter in the face of our natural light attack, underscoring the nuanced complexities introduced by our method and emphasizing the importance of specialized countermeasures.

[4]Under the pre-training approach, we initially train our light generator using normal setting employing a batch size of 16, a learning rate of 0.0002, and betas set at (0.5, 0.999) over 21 epochs. As for meta learning, we opted fundamental first-order Model-Agnostic Meta-Learning (MAML) [28]. The parameters included a batch size of 8, 100 outer steps, and 3 inner steps.

5. Conclusion

In this paper, we show that natural light sources such as sunlight and flashlights, which are commonly encountered in our daily lives, can menace image classification task. Consequently, we proposed "Natural Light Attack" a simple strategy that can be executed by anyone. Our proposed method is not only remarkably simple but also unsuspicious, making it a great concern for the ongoing advancements in the field of computer vision. In the future, we plan to explore the possibility of natural light attacks in other application areas, e.g. other types of object recognition tasks, facial recognition, or even image-based authentication.

Acknowledgement

This work was supported in part by the National Science and Technology Council of Taiwan under Grants NSTC-109-2221-E-009-114-MY3, NSTC-112-2221-E-A49-059-MY3, NSTC-112-2221-E-A49-094-MY3, NSTC-109-2223-E-002-005-MY3, NSTC-112-2628-E-002-033-MY4 and NSTC-112-2634-F-002-002-MBK, and by the Co-creation Platform of the Industry Academia Innovation School, NYCU, under the framework of the National Key Fields Industry-University Cooperation and Skilled Personnel Training Act, from the Ministry of Education (MOE) and industry partners in Taiwan.

References

[1] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pages 284–293. PMLR, 2018.

[2] Jianbo Chen, Michael I Jordan, and Martin J Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *2020 ieee symposium on security and privacy (sp)*, pages 1277–1294. IEEE, 2020.

[3] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 15–26, 2017.

[4] Xiangyi Chen, Sijia Liu, Kaidi Xu, Xingguo Li, Xue Lin, Mingyi Hong, and David Cox. Zo-adamm: Zeroth-order adaptive momentum method for black-box optimization. *Advances in neural information processing systems*, 32, 2019.

[5] Ranjie Duan, Xiaofeng Mao, A Kai Qin, Yuefeng Chen, Shaokai Ye, Yuan He, and Yun 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*, pages 16062–16071, 2021.

[6] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1625–1634, 2018.

[7] Mahmoud Fathy, Nada Ashraf, Omar Ismail, Sarah Fouad, Lobna Shaheen, and Alaa Hamdy. Design and implementation of self-driving car. *Procedia Computer Science*, 175:165–172, 2020.

[8] Weiwei Feng, Baoyuan Wu, Tianzhu Zhang, Yong Zhang, and Yongdong Zhang. Meta-attack: Class-agnostic and model-agnostic physical adversarial attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7787–7796, 2021.

[9] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[10] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[11] Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1-2):267–305, 2016.

[12] Abhiram Gnanasambandam, Alex M Sherman, and Stanley H Chan. Optical adversarial attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 92–101, 2021.

[13] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[14] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017.

[15] Yu-Chih-Tuan Hu, Bo-Han Kung, Daniel Stanley Tan, Jun-Cheng Chen, Kai-Lung Hua, and Wen-Huang Cheng. Naturalistic physical adversarial patch for object detectors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7848–7857, 2021.

[16] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International conference on machine learning*, pages 2137–2146. PMLR, 2018.

[17] Steve TK Jan, Joseph Messou, Yen-Chen Lin, Jia-Bin Huang, and Gang Wang. Connecting the digital and physical world: Improving the robustness of adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 962–969, 2019.

[18] Zelun Kong, Junfeng Guo, Ang Li, and Cong Liu. Physgan: Generating physical-world-resilient adversarial examples for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14254–14263, 2020.

[19] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018.

[20] Jie Li, Rongrong Ji, Hong Liu, Jianzhuang Liu, Bineng Zhong, Cheng Deng, and Qi Tian. Projection & probability-driven black-box attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 362–371, 2020.

[21] Ji Liu, Steve Wright, Christopher Ré, Victor Bittorf, and Srikrishna Sridhar. An asynchronous parallel stochastic coordinate descent algorithm. In *International Conference on Machine Learning*, pages 469–477. PMLR, 2014.

[22] Sijia Liu, Pin-Yu Chen, Xiangyi Chen, and Mingyi Hong. signsgd via zeroth-order oracle. In *International Conference on Learning Representations*, 2019.

[23] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[24] Yanmao Man, Ming Li, and Ryan Gerdes. Ghostimage: Remote perception attacks against camera-based image classification systems. In *23rd International Symposium on Research in Attacks, Intrusions and Defenses*, 2020.

[25] Andreas Mogelmose, Mohan Manubhai Trivedi, and Thomas B. Moeslund. Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey. *IEEE Transactions on Intelligent Transportation Systems*, 13(4):1484–1497, 2012.

[26] Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. A self-supervised approach for adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 262–271, 2020.

[27] Dinh-Luan Nguyen, Sunpreet S Arora, Yuhang Wu, and Hao Yang. Adversarial light projection attacks on face

recognition systems: A feasibility study. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 814–815, 2020.

[28] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

[29] Alex Nichol and John Schulman. Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999*, 2(3):4, 2018.

[30] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017.

[31] Rajasshrie Pillai, Brijesh Sivathanu, and Yogesh K Dwivedi. Shopping intention at ai-powered automated retail stores (aipars). *Journal of Retailing and Consumer Services*, 57:102207, 2020.

[32] poojahira. gtsrb-pytorch. `https://github.com/poojahira/gtsrb-pytorch`, 2019.

[33] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In *Advances in Neural Information Processing Systems*, 2019.

[34] SB Thorat, SK Nayak, and Jyoti P Dandale. Facial recognition technology: An analysis with scope in india. *arXiv preprint arXiv:1005.4263*, 2010.

[35] Chun-Chen Tu, Paishun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 742–749, 2019.

[36] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. *arXiv preprint arXiv:1711.01991*, 2017.

[37] Keiichiro Yamamura, Haruki Sato, Nariaki Tateiwa, Nozomi Hata, Toru Mitsutake, Issa Oe, Hiroki Ishikura, and Katsuki Fujisawa. Diversified adversarial attacks based on conjugate gradient method. In *International Conference on Machine Learning*, 2022.

[38] Dinghuai Zhang, Tianyuan Zhang, Yiping Lu, Zhanxing Zhu, and Bin Dong. You only propagate once: Accelerating adversarial training via maximal principle. In *Advances in Neural Information Processing Systems*, 2019.

[39] Yiqi Zhong, Xianming Liu, Deming Zhai, Junjun Jiang, and Xiangyang Ji. Shadows can be dangerous: Stealthy and effective physical-world adversarial attack by natural phenomenon. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15345–15354, 2022.