Shadows can be Dangerous: Stealthy and Effective Physical-world Adversarial Attack by Natural Phenomenon

Yiqi Zhong[1], Xianming Liu[1,2]*, Deming Zhai[1], Junjun Jiang[1,2], Xiangyang Ji[3]
[1]Harbin Institute of Technology, [2]Peng Cheng Laboratory, [3]Tsinghua University
21s003117@stu.hit.edu.cn, {csxm, zhaideming, jiangjunjun}@hit.edu.cn, xyji@tsinghua.edu.cn

Abstract

Estimating the risk level of adversarial examples is essential for safely deploying machine learning models in the real world. One popular approach for physical-world attacks is to adopt the "sticker-pasting" strategy, which however suffers from some limitations, including difficulties in access to the target or printing by valid colors. A new type of non-invasive attacks emerged recently, which attempt to cast perturbation onto the target by optics based tools, such as laser beam and projector. However, the added optical patterns are artificial but not natural. Thus, they are still conspicuous and attention-grabbed, and can be easily noticed by humans. In this paper, we study a new type of optical adversarial examples, in which the perturbations are generated by a very common natural phenomenon, shadow, to achieve naturalistic and stealthy physical-world adversarial attack under the black-box setting. We extensively evaluate the effectiveness of this new attack on both simulated and real-world environments. Experimental results on traffic sign recognition demonstrate that our algorithm can generate adversarial examples effectively, reaching 98.23% and 90.47% success rates on LISA and GTSRB test sets respectively, while continuously misleading a moving camera over 95% of the time in real-world scenarios. We also offer discussions about the limitations and the defense mechanism of this attack[1].

1. Introduction

In the past few years, we have witnessed the great success of deep neural networks (DNNs) in a variety of computer vision tasks, such as image classification, object detection, scene segmentation and so on. However, recent studies have revealed that DNNs based models are vulnerable to adversarial examples, even though the added magnitude of perturbations is small. In safety sensitive scenarios, such as autonomous driving [18] and medical diagnosis [19], adversarial inputs would enforce a machine learning system to produce erroneous decision, leading to unexpected situations that may be potentially dangerous.

Estimating when a machine learning model fails to work is of great concern to trustworthy AI. Accordingly, it is important to understand the level of risk of various adversarial examples to the machine learning models. There are numerous attack strategies investigated in the literature, which can be in the digital domain, where digital images corresponding to a scene are modified; or in the physical domain, where perturbations are physically added to the objects themselves [10].

The adversarial attacks in the physical domain receive more attention recently, since they are more practical and challenging. One popular approach for physical-world attacks is to adopt the "sticker-pasting" strategy [10], where the adversarial perturbation is printed as a sticker and then attached onto the target object, *e.g.*, a road sign. However, this approach would suffer from a few troubles: 1) In some cases, it may be impossible to access the target object; 2) The printing of perturbations is imperfect, *i.e.*, some values cannot be reproduced by valid colors in the real world. Some strategies emerged very recently to remedy these limitations, which explored a new type of adversarial attack—the threats of light—to achieve non-invasive attack [9, 12, 22, 28]. For instance, Duan *et al.* [9] propose to utilize the laser beam as adversarial perturbation directly rather than crafting adversarial perturbation from scratch, which has been demonstrated to be an effective physical-world attack to DNNs. Gnanasambandam *et al.* [12] propose to project calculated patterns to alter the appearance of the target objects based on a low-cost projector-camera system. Experimental results shown in [9, 12] demonstrate the effectiveness of light attacks in both digital- and physical-settings.

A basic principle in adversarial attack is that the carefully perturbed inputs should cause the network to generate wrong decision while without introducing a visible change to humans. However, the added optical patterns in existing

*Corresponding author
[1]Our code is available at https://github.com/hncszyq/ShadowAttack.

methods, both laser beam [9] and projected pattern [12], are artificial but not natural. Thus, they are still conspicuous and attention-grabbed, and can be easily noticed by humans. In this paper, we study a new type of optical adversarial examples, in which the perturbations are generated by a very common natural phenomenon, *shadow*, to achieve naturalistic and stealthy physical-world adversarial attack. We choose traffic sign recognition as our target task to validate the effectiveness of this attack. It is worth noting that, in the field of computer vision, many methods have been proposed to conduct shadow removal [11, 17, 33], which work from the perspective of image restoration in order to improve the accuracy of the subsequent high-level tasks [34]. We are the first work to reveal that shadow can also become a threat to harm the reliability of machine learning based vision system. It is a meaningful reminding to prevent such attacks in practical systems.

The contributions of this paper can be highlighted as follows:

- We propose a new light-based physical-world adversarial attack under the black-box setting via the very common natural phenomenon—shadow, which is more naturalistic and stealthy.
- We offer feasible optimization strategies to generate digitally and physically realizable adversarial examples perturbed by shadows.
- We provide thorough evaluations conducted on both simulated and real-world environments to demonstrate the effectiveness of our method. We also discuss the limitations and the defense mechanism of this attack.

2. Related Work

2.1. Adversarial Examples

Adversarial examples were first discovered by Szegedy *et al*. [27], showing that a small perturbation can drastically change the network output while being quasi-imperceptible to humans. More interestingly, they also found the cross model generalization ability of adversarial examples, i.e., a large fraction of them will be misclassified by networks trained with different hyper-parameters or even with different architectures. This intriguing property of DNNs has greatly inspired the enthusiasm of academic research. Existing works on adversarial examples are either in the digital domain or in the physical world.

2.2. Adversarial Examples in Digital Domain

In the digital domain, adversarial perturbations are directly added to the inputs of a model, and their amplitudes are usually limited by l_p-norm, e.g. l_∞-norm [13, 20], l_2-norm and l_0-norm [3] to ensure imperceptibility. According to adversary's knowledge, adversarial attacks can either be white-box, where an attacker knows all the details of a model so that the gradient can be fully used to craft a perturbation [3, 13, 16, 20, 23], or black-box, where an attacker can only query the target model and get a corresponding output without knowing its internal structure. In the black-box scenario, an attacker can take advantage of the cross model generalization ability of adversarial examples [6, 7, 30, 31] or reconstruct the internal information of a model through multiple queries [2, 4, 5, 14]. Further, these two strategies can be used at the same time to promote each other [32].

2.3. Adversarial Examples in Real Physical World

Due to the variations caused by camera, a major concern is whether these highly optimized samples can pose a real threat to applications in the physical world. Kurakin *et al*. [16] first demonstrated the existence of physical adversarial examples. They printed out digital adversarial examples, and found that their corresponding photos taken by a mobile phone were still adversarial. However, such noise-based perturbations are not practical in the physical world, as they are hard to be captured by cameras at large distances. To make adversarial patterns more prominent, while keeping their stealthiness, a line of work hide perturbations with natural styles [8] or limit the large perturbations to a small area with specific semantics, e.g., stickers arranged in the shape of letters [10] or a pair of glasses [25]. To further improve robustness to the variations caused by camera, Athalye *et al*. [1] proposed Expectation Over Transformation (EOT), by which the first robust 3D adversarial example was generated.

Recently, a new line of work explores non-invasive attacks that do not access the target object, e.g., optic-based attacks. Sayles *et al*. [24] illuminates the target object with a special light that varies with a high frequency to cause rolling shutter effect of the camera. Duan *et al*. [9] disturb the classifier by generating a laser beam in front of the target object. Gnanasambandam *et al*. [12] project their calculated adversarial patterns to the target object by a projector. Instead of the optical phenomena produced by these sophisticated artificial devices, in this work, we turn to explore a common natural phenomenon—shadow.

3. Approach

In this section, we present our method for learning an adversarial object by casting shadow to achieve successful attack. We first describe the representation of shadow.

3.1. Problem Formulation

Given an input image $x \in \mathbb{R}^d$ with the class label $y \in [1, \cdots, k]$, a DNN based classifier $f : \mathbb{R}^d \to \mathbb{R}^k$ is trained to derive the predicted label $\widetilde{y}$:

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

a) Shadows in the real world.

b) Our simulated shadows.

Figure 1. Comparison of our simulated shadows and real-world shadows on traffic signs from LISA [21] and GTSRB [26] datasets.

where $f_i(x)$ is the confidence of the i-th class. The goal of our proposed method is to cast a specific shadow onto the target object to produce adversarial example x_{adv}, which causes the machine learning system to produce misclassification:

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

Meanwhile, it should guarantee that the perturbation imposed on x_{adv} is imperceptible enough so that x_{adv} is stealthy for humans.

3.2. Shadow Perturbation Modeling

To generate the adversarial example by shadow, two factors should be carefully considered: the location of shadow and the value of shadow.

Shadow Location. For the location of shadow, we propose to find an appropriate polygon $\mathcal{P}_\mathcal{V}$ to simulate the shadow region, which is determined by a set of vertices $\mathcal{V} = \{(m_1, n_1), (m_2, n_2), ..., (m_s, n_s)\}$. Given $\mathcal{M}$ as the mask to locate the target object, the shadow region in the target object can be defined as $\mathcal{M} \cap \mathcal{P}_\mathcal{V}$. Note that, although $\mathcal{P}_\mathcal{V}$ can be any valid polygon, based on the experimental experience shown in Sec. 4, the simplest polygon—triangles—are sufficient to produce successful adversarial examples. We can exploit more complex polygon to achieve higher attack success rate, however, the shadow would look unnatural and it becomes more difficult to implement in the real world. Therefore, in practice, $\mathcal{P}_\mathcal{V}$ is simply set to be a triangle. Thus, the optimization location parameters in our method is the corresponding vertices coordinates $\mathcal{V} = \{(m_1, n_1), (m_2, n_2), (m_3, n_3)\}$.

Shadow Value. To perform a successful physical-word attack, a main concern is the non-negligible gap between the digital and the physical domains. A widely used technique to tackle this problem is Expectation Over Transformation (EOT) [1], which produces robust physical-world objects that remain adversarial when captured over a wide range of transformations, including distances, angles, exposures, etc.

In addition to the variations caused by camera, in our context, another key factor should be considered is the relationship of pixel values between shadow and non-shadow areas in the real world scenario, since we need to simulate shadows in digital images when generating adversarial examples. However, the exact formula of this relationship is obscure. To simplify this problem, following the assumption in the field of shadow removal [17], we regard that shadows only affect the illumination component and leave other ones remain unchanged. Accordingly, we transform images from RGB space to LAB space, and only consider the L channel which is a correlate of lightness. Our statistical results on SBU Shadow dataset [29] confirmed this assumption by indicating that the mean ratio of pixel values of channels L, A and B in shadow areas to that in non-shadow areas are 0.43, 0.99 and 0.90 respectively, and the corresponding standard deviations of channels A and B are as low as 0.05 and 0.07[2].

Based on the above analysis, we can generate shadow perturbation cast in an image by multiplying the corresponding L channel in LAB color space with a coefficient k[3]. Specifically, given a clean image x in RGB color space, we first convert x to LAB color space:

$$LAB(x) = LAB\left([R_x\ G_x\ B_x]\right) = [L_x\ A_x\ B_x].$$

Then given a polygon $\mathcal{P}_\mathcal{V}$ and a mask $\mathcal{M}$, the value of pixel (i, j) in the shadow image x_{adv} can be calculated as follows:

$$\begin{aligned} LAB^{ij}(x_{adv}) &= [L^{ij}_{x_{adv}}\ A^{ij}_{x_{adv}}\ B^{ij}_{x_{adv}}] \\ &= \begin{cases} LAB^{ij}(x) \cdot [k\ 1\ 1]^T & (i,j) \in \mathcal{P}_\mathcal{V} \cap \mathcal{M} \\ LAB^{ij}(x) \cdot [1\ 1\ 1]^T & (i,j) \notin \mathcal{P}_\mathcal{V} \cap \mathcal{M} \end{cases}. \end{aligned}$$

Finally, we convert x_{adv} back to RGB color space. For simplicity, we denote the above adversarial example generation process as:

$$x_{adv} = \mathcal{S}(x, \mathcal{P}_\mathcal{V}, \mathcal{M}, k) \tag{3}$$

In Fig. 1, we show some instances of our simulated shadows on traffic signs, as well as some real-world shadows for comparison.

In practice, however, the form of shadow is the result of complex physical process. The coefficient k can be varied due to numerous factors, including light sources, scene geometry, materials of the target object, and the imaging quality of the camera, etc. The distribution of k in SBU Shadow

[2]We selected 400 images from the training set for statistical analysis.

[3]Note that in addition to LAB color space, we have also tried YUV and HSL. Our statistical results show that LAB is the best choice.

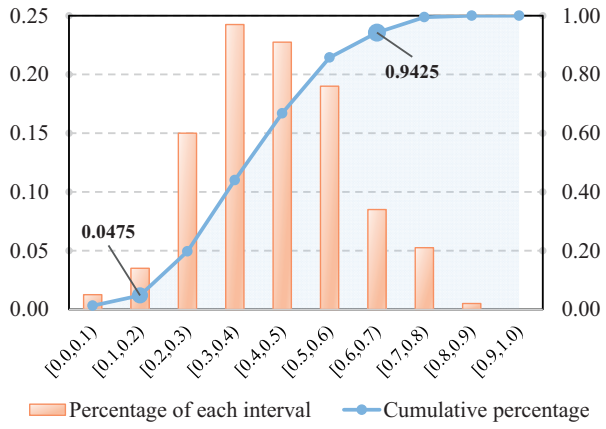

Figure 2. **The distribution of coefficient k.**

dataset is shown in Fig. 2. Accordingly, in our scheme, before generating an adversarial example, we first cast a random shadow on the target object to measure the coefficient k. Further, by leveraging EOT, the impact of the mismatch of k can be minimized.

3.3. Shadow Attack in Digital Domain

In this subsection, we introduce in detail the generation process of adversarial examples using simulated shadows in the digital domain.

The aim of our proposed attack method is to find vertices coordinates $\mathcal{V} = \{(m_1, n_1), (m_2, n_2), (m_3, n_3)\}$ of the triangle, which is used to simulate the shadow area, to make the resultant shadow image x_{adv} being misclassified by the target model f. In this work, we consider the more practical and challenging black-box attack scenario, where we can only access to the input (images) and the output (confidence scores $f(\cdot)$) of a targeted DNN, without knowing the detailed network structure and parameters. To this end, we formulate the optimization objective as minimizing the confidence score on true label:

$$\arg\min_{\mathcal{V}} \; f_{true}\left(\mathcal{S}(x, \mathcal{P}_{\mathcal{V}}, \mathcal{M}, k)\right), \text{ s.t. } \widetilde{y}_{adv} \neq y_{true} \quad (4)$$

where $\widetilde{y}_{adv} = \arg\max_i f_i(x_{adv})$. We want to find the best set of coordinates $\mathcal{V}^*$ to minimize $f_{true}(x_{adv})$ so as to produce misclassification.

In black-box scenarios, a popular approach is to use zeroth order optimization (ZOO) [5] to estimate the gradients of the targeted DNN. However, we observe severe gradient exploding and vanishing effect in the practical implementation of ZOO, which is probably due to: 1) the coordinate values in $\mathcal{V}$ are discrete, 2) the judgement on whether a coordinate locates within $\mathcal{P}_{\mathcal{V}} \cap \mathcal{M}$ is a boolean function, making gradients unreliable.

Instead, in our method, we turn to exploit the particle swarm optimization (PSO) strategy [15], which stems from the study on how bird flocks search for food. Based on the cooperation and information sharing between individuals in a population, a valid solution can be found quickly without the help of gradients. Specifically, in PSO, we maintain a population of particles, each of which represents a candidate solution of Eq. (4) and has its own speed of movement in the solution space. In each iteration, all the particles adjust their speeds and positions according to the current individual optima and the global optimum shared by the whole population, and finally update every optimum. During this process, a cost function is needed to evaluate the quality of each candidate solution, which is the confidence score $f_{true}(\cdot)$ in our context. The algorithm terminates as long as $\widetilde{y}_{adv} \neq y_{true}$ or the maximum number of iterations is reached. To further increase the success rate, we adopt the n-random-restarts strategy, which means that when the attack fails we can reinitialize and rerun the PSO at most $n-1$ times.

3.4. Shadow Attack in Real Physical-World

To generate robust adversarial examples in the real physical-world, we adopt the following two strategies for enhancement:

Expectation Over Transformation. EOT [1] is a powerful tool for dealing with the domain shift between the digital and the physical domains. To implement EOT, we start by defining a distribution of transformation $\mathcal{T}$ to simulate domain shift. Each instance of $\mathcal{T}$ is a combination of a set of random image transformations including down sampling, brightness adjustment, perspective transformation, motion blur. More importantly, the possible mismatch of shadow coefficient k is also considered in $\mathcal{T}$ by setting a random k. Next, rather than optimizing $\mathcal{V}$ for a single image, we instead minimize the expectation of $f_{true}(\cdot)$ over the set of transformed images, *i.e.*,

$$\arg\min_{\mathcal{V}} \; \mathbb{E}_{t\sim\mathcal{T}}[f_{true}(t(x_{adv}))], \text{ s.t. } \widetilde{y}_{adv} \neq y_{true}. \quad (5)$$

In practice, we approximate the expectation in Eq. (5) by the mean of 10 transformed samples together with the original sample.

Prediction stabilization. According to Eykholt *et al.* [10], any successful physical perturbation must cause targeted misclassification in a range of distances and angles to avoid attack failure by performing simple majority voting. However, due to the relatively small perturbation space, it is difficult to achieve targeted misclassification by shadows. Instead, we first optimize Eq. (5) such that the classifier should output a wrong decision $\widetilde{y}_w$. Then, we stabilize this prediction by rerunning PSO while conducting the following optimization:

$$\arg\max_{\mathcal{V}} \; \mathbb{E}_{t\sim\mathcal{T}}[f_w(t(x_{adv}))], \text{ s.t. } \widetilde{y}_{adv} = \widetilde{y}_w. \quad (6)$$

Table 1. Success rates of the proposed Shadow Attack with different shadow coefficient k on LISA and GTSRB test sets. The column in bold refers to the performances when k takes the mean of k in SBU Shadow dataset, which is 0.43.

Model	**Shadow Coefficient k**											
	0.20	0.25	0.30	0.35	0.40	**0.43**	0.45	0.50	0.55	0.60	0.65	0.70
LISA-CNN	100.00	100.00	99.73	99.18	99.05	**98.23**	97.95	96.04	93.18	88.81	83.08	72.71
GTSRB-CNN	97.37	96.35	95.14	93.45	91.36	**90.47**	88.97	87.15	84.25	80.36	73.76	66.73

Table 2. Average number of queries at the time when the attack succeeds with different shadow coefficient k on LISA and GTSRB test sets. The column in bold refers to the performances when k takes the mean of k in SBU Shadow dataset, which is 0.43.

Model	**Shadow Coefficient k**											
	0.20	0.25	0.30	0.35	0.40	**0.43**	0.45	0.50	0.55	0.60	0.65	0.70
LISA-CNN	26.6	27.7	38.4	71.8	83.4	**91.2**	100.6	137.0	169.2	195.9	306.7	293.4
GTSRB-CNN	98.3	93.7	112.8	129.0	128.4	**126.8**	136.9	155.9	188.3	232.0	294.2	343.4

Experiments in Sec. 4.3 demonstrate that our algorithm can achieve similar performance with targeted attacks in term of the stability of predictions.

4. Experiments

4.1. Datasets and Models

We choose road sign classification as our target task. Following Eykholt *et al*. [10], we evaluate the performance of our algorithm on the same datasets and model architectures. One dataset is LISA [21], which consists of 47 different US road sign classes. To avoid the effect of long-tailed distribution, only the top 16 classes with the largest number of samples are tested. The other dataset is GTSRB [26], which consists of 43 different German road sign classes. Note that there are some images in both datasets that are captured under very dark condition or already in shadows. In view of this, we remove images whose average pixel value of L-channel in the area of traffic sign is less than 120.

4.2. Evaluation in Digital Domain

In the digital domain, we attack every image in the test sets of LISA and GTSRB with simulated shadow. We measure the performance of our algorithm by attack success rate, which is defined as the ratio of the number of samples that successfully cause misclassification to the number of all samples in the test set. As shown in Fig. 2, the shadow value k can be varied in a wide range in the real world. Therefore, we repeat the experiment with several different values of k ranging from 0.2 to 0.7, and the results are shown in Tab. 1. It can be found that, when k is set to 0.43, which is the mean of shadow values in the SBU Shadow dataset [29], our success rates on LISA and GTSRB test sets are 98.23% and 90.47% respectively. When $k > 0.7$, the successful rates are relatively low, because the shadows are weak and the perturbed images are close to the clean ones.

In the black-box scenario, to keep stealthy, another important concern is how many queries should be performed. We investigate this point in Tab. 2, from which it can be found that our method can get a valid adversarial example with only dozens or hundreds of queries.

4.3. Evaluation in Physical Domain

In the physical domain, to simulate real-world application scenarios, we first cast the specific shadow onto our target traffic sign and then record a video with a camera moving towards the sign. Next, we measure the performance of our attack by the percentage of misclassified frames in this video. By experiments, we demonstrate that our method can consistently carry out effective attacks not only outdoors in the daytime using sunlight to generate shadows, but also at night or indoors by leveraging artificial light source to generate shadows.

Outdoor Environment. We take the US speed limit 25 sign as an example for outdoor test in the daytime. As shown in Fig. 3, we generate the shadow by a cardboard. We record a video with 220 frames to simulate a self-driving car coming from far to near. For each frame, we first manually draw the bounding-box of the traffic sign and then follow the crop-resize-then-classify pipeline for classification. The results show that, our attack achieves 100.00% misclassification rate, while 95.91% of the frames are misclassified as speed limit 35. This implies that the performance of our untargeted attack is comparable to that of targeted attacks in term of the stability of predictions, making the voting algorithm fail to be an effective defense. Fig. 3 shows some frames at equal intervals of time as well as their predictions.

Indoor Environment. One limitation of our attack is that it is difficult to produce obvious shadows at night without sunlight or indoors with poor lighting. In this case, we use a flashlight to generate artificial light, which achieves comparable effect to sunlight. Our indoor test is conducted

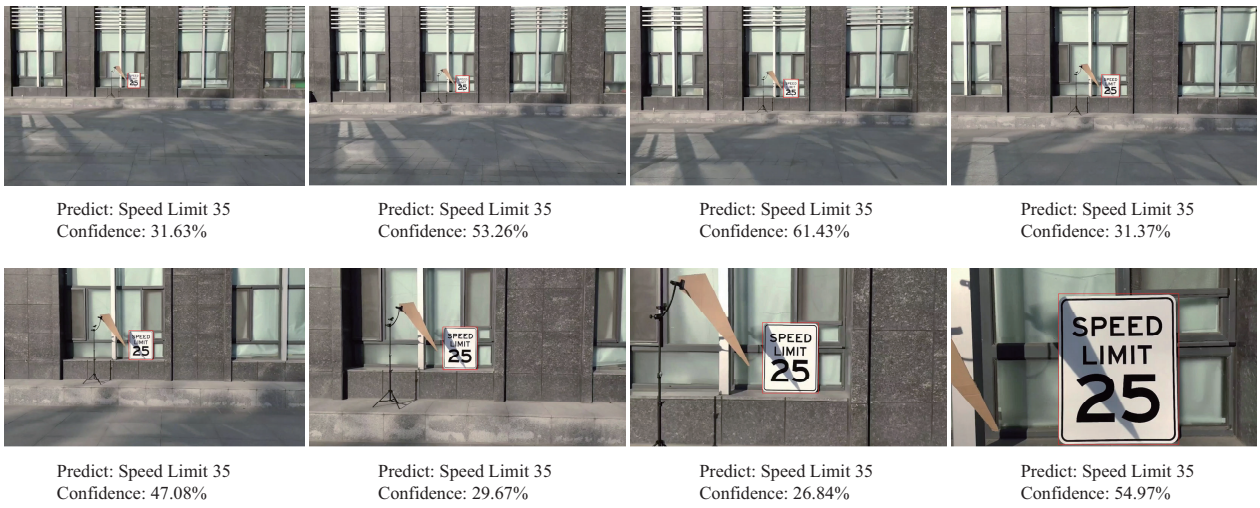

Figure 3. Examples of frames in the video of our outdoor experiment and the corresponding classification results.

Table 3. Experimental results of our indoor test, where error rate refers to the percentage of misclassified frames in each video and stability refers to the percentage of frames predicted as the primary error class.

Ground truth	**Predict**	**Error rate**	**Stability**
Speed limit 25	Speed limit 35	100%	88%
Speed limit 30	Speed limit 35	100%	93%
Speed limit 35	Speed limit 30	95%	71%
Speed limit 45	Signal Ahead	100%	76%

in a dark stairwell, as illustrated in Fig. 4. The traffic signs are placed on the platform in the middle of two floors and the flashlight is fixed on the upper floor while ensuring that the light can be cast onto the traffic signs. We take four US speed limit signs from LISA [21] as examples for indoor test. For each sign, we record a video when walking from the lower floor to the upper floor. The videos include 100 frames. Tab. 3 offers the misclassification rate of each video, which demonstrates that our attack is also effective indoors. Fig. 4 provides some example frames in the test of the speed limit 45 sign, in which their predictions are also given. It can be found our attack is successful for all.

4.4. Scheduled Attack

Once the cardboard is fixed, the generated shadow will move over time due to the angle change of sunlight, which inspires us that our method can be used to conduct an interesting scheduled attack. Scheduled attack means that, due to the movement of the sun, we can control our shadow perturbation to be harmless in most of the time but to be adversarial at a certain time, e.g. the peak traffic hour in the morning.

To carry out a scheduled attack, we just need to fix the cardboard as done in Sec. 4.3 at a specific time of a day in advance, and then the adversarial effect would be generated at the same time the future days if there is no significant change of the weather. Furthermore, after applying the strategies mentioned in Sec. 3.4 to enhance robustness, our attack can maintain successful for a period of time, rather than only effective at a certain moment.

To test the scheduled attack, we simulate it on digital images. We first place our targeted traffic sign image on the XOZ plane of a three-dimensional coordinate system. Next, we calculate the solar elevation angle and the solar azimuth angle by using our scheduled time, longitude and latitude. Based on this, as long as we set the distance to the traffic sign in the Y-axis direction, we can further get the coordinates of each vertex of the cardboard. By fixing these coordinates, we can figure out the positions of shadows on the XOZ plane at any time. In this way, we simulate the dynamic changes of shadows over time.

We choose eight German speed limit signs from GTSRB [26] as test examples for our experiment. The longitude and latitude are set to 0 and 45 respectively. The traffic signs face south, i.e., the Y-axis points to the south. The distances from cardboard to traffic signs in the Y-axis direction are set to one meter. For each sign, we set 8:30 in the morning as the scheduled time, and we generate images with shadow every second from 8:25 to 8:35, from which we can observe how the classifier's predictions change over time. We show an instance of our simulated scheduled attack in Fig. 5. As shown in Fig. 6, we achieve the goal of scheduled attacks on all eight speed limit signs. The duration of successful attack ranges from dozens of seconds to several minutes.

4.5. Ablation study

About the number of edges of $\mathcal{P}_\mathcal{V}$. Theoretically, $\mathcal{P}_\mathcal{V}$ can be any valid polygon. However, when the number of edges is relatively large, the shadow would look unnatural,

Figure 4. Examples of frames in a video of our indoor experiment and the corresponding classification results.

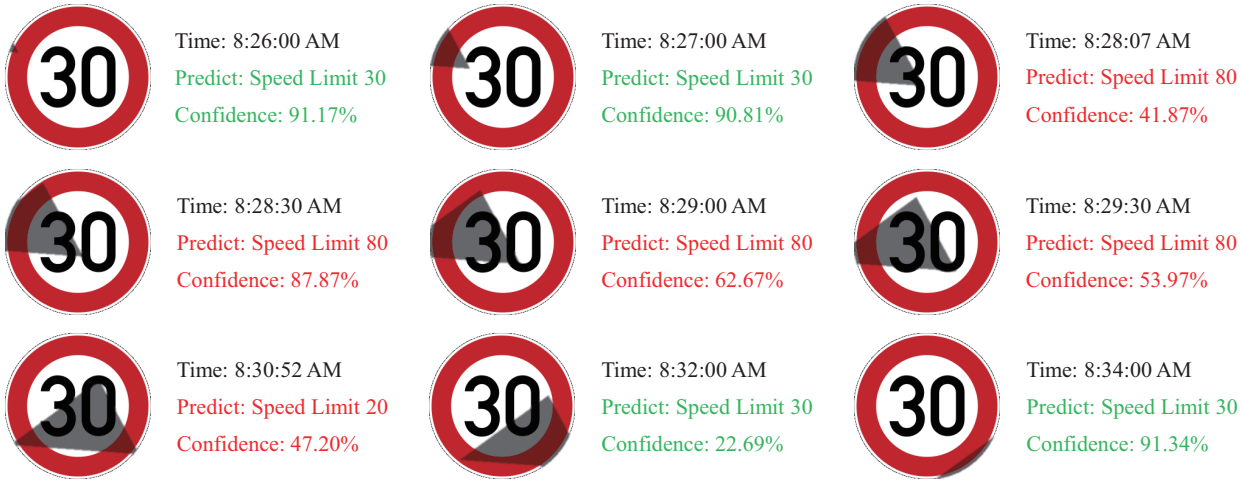

Figure 5. Our simulated scheduled attack. Our scheduled time is 8:30 in the morning, and the target class is speed limit 80. We show how the predictions change as the shadow moves on the German speed limit 30 sign. The predictions before 8:28:07 am and after 8:30:52 am are correct, while the predictions from 8:28:07 am to 8:30:06 am are speed limit 80 and the predictions from 8:30:07 am to 8:30:52 am are speed limit 20.

Table 4. Performances of the proposed Shadow Attack with different number of edges of polygon $\mathcal{P}_\mathcal{V}$.

Model	Number of edges			
	3	5	7	9
LISA-CNN	97.95	98.91	99.45	99.59
GTSRB-CNN	90.80	93.72	96.83	97.59

which goes against the requirement of stealthiness. Moreover, a polygon with too many edges leads to more computational overhead and it is not easy to implement in the real world. So in practice, we choose to attack the target object by a simple polygon area. Indeed, as shown in Tab. 4, the simplest polygon—triangles—are sufficient to produce successful adversarial examples. The performance loss caused by the reduction in the number of edges is also acceptable.

Table 5. ablation study of n-random-starts strategy.

Model	Number of restarts				
	1	5	10	50	100
LISA-CNN	95.91	98.36	98.50	99.05	99.18
GTSRB-CNN	87.76	90.74	91.76	92.81	93.22

About n-random-restarts strategy. We test the effectiveness of n-random-restarts by setting n to 1, 5, 10, 50, and 100 respectively. As shown in Tab. 5, compared with

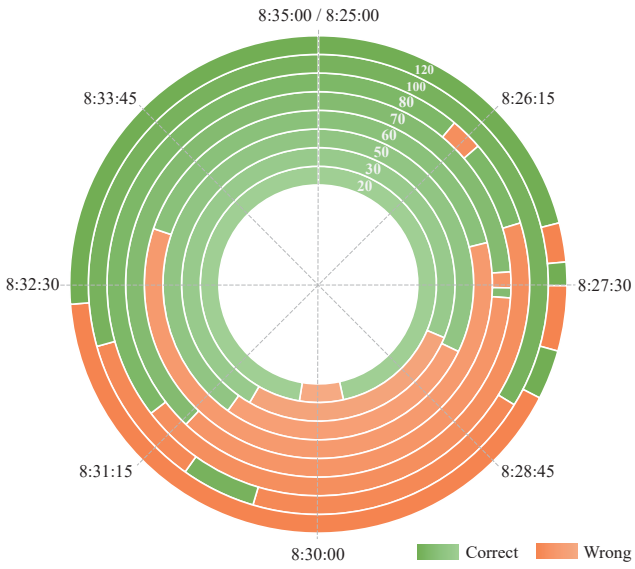

Figure 6. The test result of our scheduled attack on 8 German speed limit signs. The green and red areas represent correct and incorrect classifications respectively.

Table 6. Performance comparison of models with and without robustness training, where accuracy denotes the clean accuracy in LISA and GTSRB test sets, robustness denotes the probability that our attack fails (1 - attack success rate), and queries denotes the average number of queries at the time when our attack succeeds. All the data are obtained when $k = 0.43$.

Model	**Accuracy**	**Robustness**	**Queries**
LISA-CNN	99.63	1.73	91.24
GTSRB-CNN	99.00	9.53	126.75
LISA-CNN$_{rob}$	99.56	40.93	849.51
GTSRB-CNN$_{rob}$	98.91	25.57	464.52

$n = 1$, by setting $n = 5$ we can effectively improve the success rate. However, when n is greater than 5, there are only marginal improvements in performance by increasing n, demonstrating the effectiveness of PSO. To balance the success rate and computational overhead, we set $n = 5$ in practice.

4.6. How to Defend against the Shadow Attack?

In the above subsections, we demonstrate that the shadow perturbations is able to pose a security threat to the traffic or other security-critical scenarios. We discuss the defense mechanism against this type of attack as follows.

One of the most effective defense strategies against adversarial examples is adversarial training (AT) [20], which improves model robustness by using newly generated adversarial examples as training samples in each epoch. In our context, to implement AT, we need to apply the algorithm described in Sec. 3.3 to each training sample in an epoch, which however is a particularly time-consuming process. In order to make training faster, in each epoch we only add a random shadow to each training sample (the coordinate values in $\mathcal{V}$ and the coefficient k are random variables), instead of the worst case shadow. We expect that such a training method can reduce the model's sensitivity to shadows, thereby improving the robustness to our attack. We retrain the models in this way, which are denoted as LISA-CNN$_{rob}$ and GTSRB-CNN$_{rob}$. To verify the effectiveness of the modified adversarial training, we perform attack on the retrained models in the digital domain. The results summarized in Tab. 6 demonstrate that this defense can improve the model robustness and the difficulty of the attack significantly at the expense of slightly decreased accuracy on clean samples.

5. Limitation

One limitation of our attack is the requirement of a strong single light source. In the environment with poor lighting condition, our method is difficult to produce significant shadow perturbations, *i.e.*, k would be large, which result in little difference between the attacked images and the clean ones. As shown in Tab. 1 and Tab. 2, a large k leads to a low success rate and more queries. To remedy this problem, we can achieve the adversarial effect through a stronger artificial light as shown in Sec. 4.3. For the scenarios with multiple light sources, it is also difficult for our method to produce effective shadow perturbations.

Another limitation is that we cannot explicitly conduct targeted attacks, which is probably due to the perturbation directions produced by shadows are relatively single. However, as shown in our outdoor experiment, we are able to achieve comparable results to that of targeted attacks in term of the stability of predictions.

6. Conclusion

In this paper, we present a new optics-based adversarial attack strategy, which utilizes the very common natural phenomenon, shadow, to produce harmful perturbations on the targets. Experimental results demonstrate this attack can be successfully applied in both digital- and physical-settings. Our work reveals that shadows can be dangerous, which is able to mislead the machine learning based vision system to produce erroneous decision.

As a new type of attack, there surely would be some drawbacks or limitations. We will work on making it more practical in future research.

7. Acknowledgement

This work was supported by National Natural Science Foundation of China under Grants 61922027, 6207115 and 61932022.

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, 3, 4

[2] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248*, 2017. 2

[3] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017. 2

[4] 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. 2

[5] 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. 2, 4

[6] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018. 2

[7] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4312–4321, 2019. 2

[8] Ranjie Duan, Xingjun Ma, Yisen Wang, James Bailey, A Kai Qin, and Yun Yang. Adversarial camouflage: Hiding physical-world attacks with natural styles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1000–1008, 2020. 2

[9] Ranjie Duan, Xiaofeng Mao, A. K. Qin, Yuefeng Chen, Shaokai Ye, Yuan He, and Yun Yang. Adversarial laser beam: Effective physical-world attack to dnns in a blink. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16057–16066, 2021. 1, 2

[10] 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 (CVPR)*, June 2018. 1, 2, 4, 5

[11] G.D. Finlayson, S.D. Hordley, Cheng Lu, and M.S. Drew. On the removal of shadows from images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1):59–68, 2006. 2

[12] Abhiram Gnanasambandam, Alex M. Sherman, and Stanley H. Chan. Optical adversarial attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 92–101, October 2021. 1, 2

[13] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2

[14] 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. 2

[15] James Kennedy and Russell Eberhart. Particle swarm optimization. In *Proceedings of ICNN'95-international conference on neural networks*, volume 4, pages 1942–1948. IEEE, 1995. 4

[16] Alexey Kurakin, Ian Goodfellow, Samy Bengio, et al. Adversarial examples in the physical world, 2016. 2

[17] Hieu Le and Dimitris Samaras. Physics-based shadow image decomposition for shadow removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 2, 3

[18] Peiliang Li, Xiaozhi Chen, and Shaojie Shen. Stereo r-cnn based 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[19] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017. 1

[20] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 2, 8

[21] 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. 3, 5, 6

[22] 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 (CVPR) Workshops*, June 2020. 1

[23] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pages 372–387. IEEE, 2016. 2

[24] Athena Sayles, Ashish Hooda, Mohit Gupta, Rahul Chatterjee, and Earlence Fernandes. Invisible perturbations: Physical adversarial examples exploiting the rolling shutter effect. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14666–14675, 2021. 2

[25] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael 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*, pages 1528–1540, 2016. 2

[26] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks*, 32:323–332, 2012. 3, 5, 6

[27] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 2

[28] James Tu, Mengye Ren, Sivabalan Manivasagam, Ming Liang, Bin Yang, Richard Du, Frank Cheng, and Raquel Urtasun. Physically realizable adversarial examples for lidar object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13713–13722, 2020. 1

[29] Tomás F Yago Vicente, Le Hou, Chen-Ping Yu, Minh Hoai, and Dimitris Samaras. Large-scale training of shadow detectors with noisily-annotated shadow examples. In *European Conference on Computer Vision*, pages 816–832. Springer, 2016. 3, 5

[30] Weibin Wu, Yuxin Su, Xixian Chen, Shenglin Zhao, Irwin King, Michael R Lyu, and Yu-Wing Tai. Boosting the transferability of adversarial samples via attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1161–1170, 2020. 2

[31] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2730–2739, 2019. 2

[32] Jiancheng Yang, Yangzhou Jiang, Xiaoyang Huang, Bingbing Ni, and Chenglong Zhao. Learning black-box attackers with transferable priors and query feedback. *Advances in Neural Information Processing Systems*, 33, 2020. 2

[33] Ling Zhang, Qing Zhang, and Chunxia Xiao. Shadow remover: Image shadow removal based on illumination recovering optimization. *IEEE Transactions on Image Processing*, 24(11):4623–4636, 2015. 2

[34] Wuming Zhang, Xi Zhao, Jean-Marie Morvan, and Liming Chen. Improving shadow suppression for illumination robust face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(3):611–624, 2019. 2