Adversarial Laser Beam: Effective Physical-World Attack to DNNs in a Blink

Ranjie Duan[1,2] Xiaofeng Mao[2] A. K. Qin[1 †] Yuefeng Chen[2] Shaokai Ye[3] Yuan He[2] Yun Yang[1]
[1]Swinburne University of Technology, Australia [2]Alibaba Group, China [3]EPFL, Switzerland

Abstract

Though it is well known that the performance of deep neural networks (DNNs) degrades under certain light conditions, there exists no study on the threats of light beams emitted from some physical source as adversarial attacker on DNNs in a real-world scenario. In this work, we show by simply using a laser beam that DNNs are easily fooled. To this end, we propose a novel attack method called Adversarial Laser Beam ($AdvLB$), which enables manipulation of laser beam's physical parameters to perform adversarial attack. Experiments demonstrate the effectiveness of our proposed approach in both digital- and physical-settings. We further empirically analyze the evaluation results and reveal that the proposed laser beam attack may lead to some interesting prediction errors of the state-of-the-art DNNs. We envisage that the proposed $AdvLB$ method enriches the current family of adversarial attacks and builds the foundation for future robustness studies for light.

1. Introduction

Natural phenomena may play the role of adversarial attackers, *e.g.* a blinding glare results in a fatal crash of a Tesla self-driving car. What if a beam of light can adversarially attack a DNN? Further, how about using a beam of light, specifically the laser beam, as the weapon to perform attacks. If we can do that, with the fastest speed in the world, the laser beam could achieve the fastest attack with no doubts. As shown in Figure 1, by using an off-the-shelf lighting device such as a laser pointer, the attacker can maliciously shoot a laser beam onto the target object to make the self-driving car fail to recognize target objects correctly.

We regard the attack illustrated in Figure 1 as a new type of adversarial attack, which is crucial but not yet exploited. Up to now, most researchers study the security of DNNs by exploring various adversarial attacks in digital-settings, where input images are added with deliberately crafted perturbations and then fed to the target DNN model [23, 10, 6, 3, 18]. However, in physical-world scenarios, images are typically captured by cameras and then directly fed to the target models, where attackers cannot directly manipulate the input image. Some recent efforts in developing physical-world attacks are addressed in [21, 8, 2, 7, 14]. The physical-world adversarial examples typically require large perturbations, because small perturbations are hard to be captured by cameras. In addition, the attacking effects of adversarial examples of small perturbations can be easily mitigated in complex physical-world environments [21, 9, 7]. Meanwhile, physical-world adversarial examples require high stealthiness to avoid being discovered by either the victim or defender before performing an attack successfully. Thus for creating physical-world adversarial examples, there is always a compromise between stealthiness and adversarial strength.

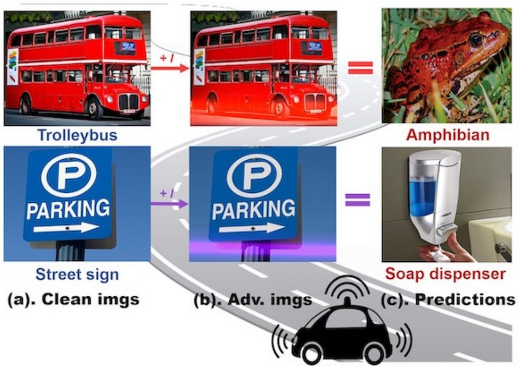
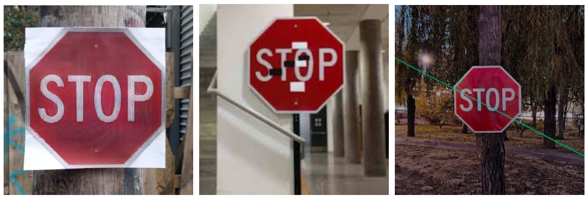


Figure 1: **An example.** When the camera of self-driving car captures object shot by the laser beam, it fails to recognize ”trolleybus” and ”street sign”.

Most existing physical-world attacks adopt a ”sticker-pasting” setting, *i.e.*, the attacker prints adversarial perturbation as a sticker and then pastes it onto the target object [16, 2, 7, 8]. These attacks achieve the stealthiness of adversaries with extra efforts of designing adversarial perturbation or camouflaging adversarial images and finding the most effective area in the target object to impose them [16, 26, 21, 7]. Besides the challenge of stealthiness, the ”sticker-pasting” setting also requires the target object to be physically accessible by the attacker to paste stickers.

Work done when intern at Alibaba Group, China
Code is available at https://github.com/RjDuan/Advlight
[†]Correspondence to: A. K. Qin



(a) $AdvCam$ [7] (b) RP_2 [2] (c) $AdvLB$ (Ours)

Figure 2: **Visual comprison.**

However, this may not be always possible. Several works explore physical-world threats beyond the ”sticker-pasting” setting: Shen et al. [22] and Nguyen et al. [19] proposed using a projector to project the adversarial perturbation on the target to perform an attack. However, these works still require manual effort to craft adversarial perturbation.

In our work, we propose a new type of physical-world attack, named adversarial laser beam ($AdvLB$). Unlike existing methods, we utilize the laser beam as adversarial perturbation directly rather than crafting adversarial perturbation from scratch. As $AdvLB$ does not require physically changing the target object to be attacked as in the ”sticker-pasting” setting, it features high flexibility to attack any object actively, even from a long distance. In terms of stealthiness, a visual comparison between our proposed attack and other works can be seen in Figure 2. Though the adversarial example generated via $AdvLB$ may appear less stealthy than some generated by other approaches such as $AdvCam$. $AdvLB$ may introduce high temporal stealthiness due to its unique physical-attack mechanism. Specifically, with the nature of light, $AdvLB$ can perform the attack in a blink right before the attacked target object gets captured by the imaging sensor, and thus avoid being noticed by victims and defenders in advance. Existing works focus on security issues of DNNs in the daytime whilst potential security threats at night are often ignored. Our proposed $AdvLB$ provides a complementary in this regard. Figure 2 illustrates the advantage of $AdvLB$ when the lighting condition is poor.

To launch such an attack, we formulate the laser beam with a group of controllable parameters. Then we devise an optimization method to search for the most effective laser beam's physical parameters in a greedy manner. It enables finding where and how to make an effective physical-world attack in a black-box setting. We further apply a strategy called k-random-restart to avoid falling into the local optimum, which increases the attack success rate.

We conduct extensive experiments for evaluation of the proposed $AdvLB$. We first evaluate $AdvLB$ in a digital-setting, which is able to achieve 95.1% attack success rate on a subset of ImageNet. We further design two physical-world attacks including indoor and outdoor tests, achieving 100% and 77.43% attack success rates respectively. Then ablation studies are presented on the proposed $AdvLB$. Furthermore, we analyze the prediction errors of DNNs caused by the laser beam. We find the causes of prediction errors could be roughly divided into two categories. 1). The laser beam's color feature changes the raw image and forms a new cue for DNNs. 2). The laser beam introduces some dominant features of specific classes, especially lighting related classes, *e.g.* candle. When the laser beam and target object appear simultaneously, there is a chance that the DNN is more biased towards the feature introduced by the laser beam and thus resulting in misclassification. These interesting empirical findings open a door for both attackers and defenders to investigate how to better manipulate this new type of attack. Our major contributions are summarized as follows:

- We propose a new type of physical-world attack based on the laser beam named $AdvLB$, which leverages light itself as adversarial perturbation. It provides high flexibility for attacker to perform the attack in a blink. Besides, the deployment of such attack is rather simple: by using a laser pointer, it may become a common threat due to its convenience to perform attack.
- We conduct comprehensive experiments to demonstrate the effectiveness of $AdvLB$. Specifically, we perform physical test to validate $AdvLB$ and show the real-world threats of laser beam by simply using a laser pointer. Thus $AdvLB$ can be a useful tool to explore such threats in the real-world.
- We make an in-depth analysis of the prediction errors caused by the $AdvLB$ attack to have revealed some interesting findings which would motivate future study on $AdvLB$ from the perspectives of attackers and defenders.

2. Background and Related Work

Adversarial attack was first proposed by Szegedy et al. [23], aiming to generate perturbations superimposed on clean images to fool a target model. Given a target model f, adversarial examples can be crafted by one or more steps of perturbation following the direction of adversarial gradients [10, 15, 18] or optimized perturbation with a given loss [3, 5]. Adversarial examples can be either generated from an image itself (in the digital setting) or produced by capturing an attacked physical scene via image sensors such as cameras (in the physical setting) [15].

Adversarial attack in digital settings. Most attacks are developed in a digital setting. And their perturbations are bounded by a small norm-ball to ensure that imperceptible to human observers. Normally, l_2 and l_∞ are the most commonly used norms [3, 4, 25, 6, 18]. Some other works also explore adversarial examples beyond bounded setting. They make modifications on the secondary attributes (*e.g.* color

[13, 20, 28], texture [24]) to generate adversarial examples. Besides, there are also several works that propose changing physical parameters [27, 17] while preserving critical components of images to create adversaries. However, digital attacks have a strong assumption that the attacker has access to modify the input image of the target model directly, which is not practical in real-world scenarios.

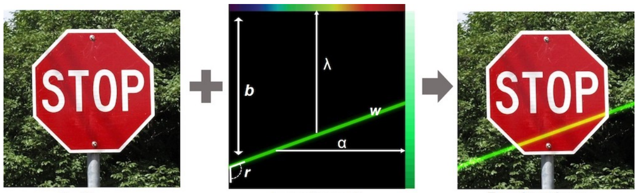
Adversarial attack in physical settings. Kurakin et al. [15] first showed the existence of adversarial examples in the physical-world by printing digital adversary in the physical-world, and then recaptured by cameras. [15], and its follow-up work adopted such setting, including pasting adversarial patch on either traffic-sign [8, 7] or t-shirt [26], or camouflaging the adversarial patch into specific shape (*e.g.* eye-glasses frames [21, 2], *etc.*). Due to various physical-world conditions such as viewpoint shifts, camera noise, and other natural transformations [1], the physically realized adversarial examples always require various adaptations over a distribution of transformations to adapt to physical-world conditions [21, 8, 2]. Thus the "sticker-pasting" attacks generate large adversarial perturbation inevitably. However, there exist a certain period of time between deploying the "sticker-pasting" attacks and performing attacks successfully. Thus the "sticker-pasting" attacks require high stealthiness to avoid being noticed by human observers in advance. Stealthiness for large perturbation is always a challenge for "sticker-pasting" physical-world attacks [9, 2, 21]. Also, these attacks require attacker physically pasting the stickers on the target objects, however, not every object is easily accessible or reachable in real-world, *e.g.* traffic sign on a high pole. Crafted perturbation in physical setting suffers from the loss in adversarial strength when converting from digital to physical-setting. In contrary, our method simply leverages the light itself as adversarial perturbation.

Physical attacks with lighting devices. There exist some works using lighting devices to generate adversarial attacks. Some utilized a projector to perform the physical-world attacks against face recognition systems [22, 19]. These attacks craft adversarial perturbation and then project the perturbation onto the target to perform the attack. Zhou et al. [30] proposed to deploy LED light on the cap to fool the face recognition systems, which requires careful deployment of the lighting device. Comparatively, existing methods require efforts to craft the adversarial perturbation while our method is much easier to perform an adversarial attack.

3. Approach

Recall the typical definition of adversarial example x_{adv} in image classification, given an input image $x \in \mathbb{R}^m$ with class label y, a DNN classifier $f : \mathbb{R}^m \rightarrow \mathbb{R}^J$. The classifier f associates with a confidence score $f_j(x)$ to class j. An adversarial example x_{adv} commonly satisfies two properties: 1) $\text{argmax}_{j \in J} f(x_{adv}) \neq \text{argmax}_{j \in J} f(x)$, 2) $\|x_{adv} - x\|_p \leq \epsilon$. In which, the first property requires that x_{adv} is classified differently with x by target model f. The second property requires that the perturbation of x_{adv} is imperceptible enough so that x_{adv} is stealthy for human observers.

In our context, as example shown in Figure 3, the aim of our proposed attack method is to find a vector of parameters θ of the laser beam l that makes the resultant image x_{l_θ} being misclassified by the target model f. We constrain the width w of laser beam l to satisfy the requirement on the stealthiness of x_{adv} in digital-setting.



Test img x Laser beam l_θ Resultant img x_{l_θ}

Figure 3: **An example.**

In following sections, we first present the definition of the laser beam. Then we model the physical laser beams with a set of parameters and optimize these parameters to create an adversarial example for the given image.

3.1. Laser Beam Definition

Laser is distinguished from other light sources by its coherence, including temporal coherence and spatial coherence. With high temporal coherence, laser is able to emit light with a very narrow spectrum even in a single color. The spatial coherence enables a laser beam to stay narrow over a long distance. Our proposed attack is based on these unique properties of the laser beam. We formulate the laser beam l with a set of physical parameters denoted by θ. We consider four key parameters to define a laser beam l including wavelength (λ), layout (r, b), width (w), and intensity α. We define each parameter as follows.

- **Wavelength** (λ)**.** Wavelength (λ) denotes the color of the laser beam. We only consider wavelengths in the range of visible light (380 nm to 750 nm). We define a conversion function that converts λ to a RGB tuple [*].

- **Layout** (r, b)**.** We treat the laser beam as a line. We use a tuple (r, b), which includes an angle (r) and intercept (b) to determine the beam line. The propagation path of the laser beam can be denoted by $y = \tan(r) \cdot x + b$. We also consider that the laser beam extincts during the transmission. Thus we define an attenuation func-

[*] http://www.noah.org/wiki/Wavelength_to_RGB_in_Python

tion with inverse square [†] to simulate the reduction of luminous intensity of laser beam during transmission.

- **Width** (w). The width of the laser beam depends on two factors: the distance between the laser beam and the camera, and the coherence degree of the laser beam. The wider the laser beam is, the more perceptible. We set a threshold on the beam width in digital-setting to avoid over-obstructiveness.
- **Intensity** (α). The intensity of the laser beam depends on the power of the laser device. Laser beam with stronger intensity looks brighter. In our context, we use α to denote the intensity of the laser beam l.

We define constraint vectors ϵ_{min} and ϵ_{max} to restrict the range of each parameter in θ. The constraints are adjustable. We then adopt a simple linear image fusion method to fuse clean image x and laser beam layer l_θ.

$$x_{l_\theta} = x + l_\theta \tag{1}$$

where x_{l_θ} represents the image x imposed by a specific laser beam l_θ defined by a vector of parameters θ, then clipped to a valid range. Formally:

$$l_\theta[x, y] = \begin{cases} \tau(\lambda), & d \le w/2, \\ \sqrt{w} \cdot \tau(\lambda)/d^2, & w/2 \le d \le 5w, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

where $d = |\tan r \cdot x - y + b|/\sqrt{1 + \tan^2 r}$ and $\tau(\cdot)$ represents conversion function that converts to a RGB tuple. The other notations are consistent with those in the paper.

3.2. Laser Beam Adversarial Attack

Algorithm. The focus of the proposed attack is to search for a vector of physical parameters θ of laser beam l given image x, aiming to result in misclassification by f. We define a search space Θ, formally, $\Theta = \{\theta | \theta = [\lambda, r, b, w, \alpha],\ \epsilon_{min} \le \theta \le \epsilon_{max}\}$, where ϵ is a list of constraints. In our context we consider the most practical scenario: an attacker cannot attain the knowledge of the target model but only the confidence score $f_y(x)$ with given input image x on predicted label y. In our proposed $AdvLB$, we exploit by using the confidence score as the adversarial loss. Thus the objective is formulated as minimizing the confidence score on correct label, we construct an adversarial image x_{l_θ} by solving the following objective:

$$\underset{\theta}{\text{argmin}}\ f_y(x_{l_\theta}),\ \text{s.t. } \theta \in [\epsilon_{min}, \epsilon_{max}] \tag{3}$$

Inspired by Guo et al.'s work [11], we exploit the confidence scores given by the target model, and propose a greedy method to search for vector θ to generate adversarial laser beam. In a brief, we repeatedly attempt to update the current vector of parameters θ with one of stronger adversarial strength instructed by $f_y(x_{l_{\theta'}})$, which indicates the confidence of correct class y given laser beam with current parameter θ'. If $f_y(x_{l_{\theta'}})$ decreases, we then update the vector of parameters θ with θ'. We first define the basis for search as set Q, which includes a series of predefined candidate vectors q. We define a candidate q as follows: the minimal update on l_θ is one unit in either λ, r, b, w or α. We set the max step numbers with t_{max}. During the search, we pick a random vector $q \in Q$ multiplying with step size $s \in S$. If either $f_y(x + l_{\theta+q})$ or $f_y(x + l_{\theta-q})$ decreases, then we update θ by current θ'. The search ends when the current x_{l_θ} is predicted with a label $\text{argmax}\, f(x_{l_\theta}) \ne \text{argmax}\, f(x)$ or the max steps t_{max} is reached.

However, we found such a greedy search process is prone to getting stuck into local optimum. To this end we further apply a strategy called k-random-restart, which introduces more randomness into the search algorithm. Specifically, k-random-restart restarts the search process k times. For each time we use θ with different initializations. We find such a simple strategy greatly improves the effectiveness of the search process. The pseudo-code of $AdvLB$ is shown in Algorithm 1.

Algorithm 1: Pseudocode of $AdvLB$

```
Input: Input $x$; Candidate vectors $Q$; Step size $S$;
       classifier $f$; Max #steps $t_{max}$;
Output: A vector of parameters $\theta$;
$conf^* \leftarrow f_y(x)$;
for $i \leftarrow 1$ to $k$ do
    Initialization: $\theta \sim \Theta(\lambda, r, b, w, \alpha)$ ;
    for $t \leftarrow 1$ to $t_{max}$ do
        Randomly pick $q \in Q$, $s \in S$;
        $q \leftarrow q \times s$ ;
        $\theta' \leftarrow \theta \pm q$;
        $\theta' \leftarrow \text{clip}(\theta', \epsilon_{min}, \epsilon_{max})$;
        $conf = f_y(x_{l_{\theta'}})$;
        if $conf^* \ge conf$ then
            $\theta \leftarrow \theta^{'}$;
            $conf^* \leftarrow conf$;
            break;
        end
        if $\text{argmax}\, f(x_{l_\theta}) \ne \text{argmax}\, f(x)$ then
            return $\theta$
        end
    end
end
```

As the algorithm illustrates, the proposed method takes a test image x, a set of candidate Q, a flexible step size S, classifier f and max steps t_{max} as input decided by the at-

[†] https://en.wikipedia.org/wiki/Inverse-square_law

tacker. Details of the algorithm have been explained above. The algorithm finally returns a successful parameter list θ of laser beam, which is used for further instructing the deployment of the attack in the physical-world.

Physical Adaptation To make adversaries generated by $AdvLB$ physically realized, we adopt two strategies. 1) Physically adapted constraints ϵ. We consider the practical limitations for an attacker to perform the attack. To this end, we denote that ϵ is physically adapted according to the real-world conditions to perform the attack. 2) Batch of transformed inputs. $AdvLB$ instructs where to perform an effective attack, however, a laser beam with exact layout (r, b) could be hard to reproduce. Thus we apply $AdvLB$ on a batch of transformed images $X_T = \{x' | x' = T(x)\}$ where T represents a random transformation function including rotation, translation, or addition of noise. With returned batch of θ on given x, we then have an effective range, *e.g.* an effective range of angle r, to perform laser attack in the physical-world.

4. Evaluation

4.1. Experimental Setting

We test our proposed attacks in both digital- and physical-settings. We use ResNet50 [12] as the target model for all the experiments. We randomly selected 1000 correctly classified images of ImageNet to evaluate proposed $AdvLB$ in digital-setting. For the physical-world experiments, our experimental devices are shown in Figure 4. In terms of laser beam, we use three small handheld laser pointers (power: 5mW) to generate low-powered laser beams with wavelength 450nm, 532nm, and 680nm respectively. We use a Google Pixel4 smartphone to take photos. In ablation study, we set a group of experiments to test the impact of different parameters on the adversarial effects of laser beam. The target models adopt a black-box setting that during the attack we only use the prediction scores given by the model. For all the tests we use attack success rate (%) as the metric to report effectiveness, which is the proportion of successful attacks among the total number of test images examined.

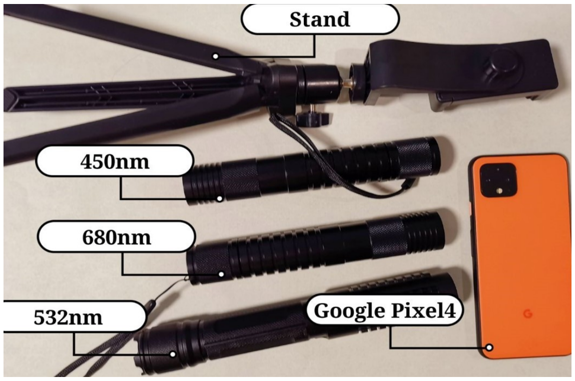


Figure 4: **Experiment devices.**

4.2. Evaluation of $AdvLB$

Digital Test We apply our proposed attack method on 1000 correctly classified images selected from ImageNet, and craft an adversarial example for each test image with simulated laser beam. The success rate is 95.1% with 834 queries on average of proposed $AdvLB$. We also present some attack results generated by $AdvLB$ as shown in Figure 5. The first column shows the test images we aim to attack, and each row denotes a series of adversarial examples generated by $AdvLB$. Figure 5 shows some interesting results. For example, when shining a laser beam with yellow color, the king snake is then misclassified as corn, and there is indeed some similarity between the texture of king snake and corn. Other adversarial examples show similar phenomenon: 'Loggerhead' + Laser beam (blue) $\longrightarrow$ 'Jellyfish', 'Radio' + Laser beam (red) $\longrightarrow$ 'Space heater'. The results show the link among the adversarial class, original class, and the laser beam with a specific wavelength. We will give more discussion in Section 5.

Physical Test $AdvLB$ aims to be an effective tool to explore real-world threats of laser beam on DNNs. Different from targeted attack in the physical-world, whose evaluation on success is rather intuitive: whether it can fool the DNNs with the class that attacker expects consistently in the physical-world. While untargeted attack is defined as fooling the DNN into any incorrect classes. With amount of uncertainties in the physical-world environment, the misclassification could be caused by natural noise rather than the untargeted attack. Thus evaluation on the effectiveness of untargeted attack in physical-setting requires more careful design of the experiments.

To validate $AdvLB$ can be reproduced by laser beam in the physical-world, we design a strict experimental setting to perform an indoor test. We use three different laser pointers with wavelengths of 450nm, 532nm, 680nm respectively to perform the attack. The target objects include a banana, a conch, and a stop sign. We set the background in black to avoid introducing unnecessary noise into the experiment. For the test, we first capture the target object by cellphone, then we use our proposed $AdvLB$ to find where to perform attack with given test captured image. We then reproduce the attack in such a setting with the returned parameter list θ by $AdvLB$. We set constraint on parameter λ according to the wavelength of used laser pointer. The experimental results are summarized in Figure 6.

As Figure 6 shows, the proposed $AdvLB$ is able to achieve 100% attack success rate in such a strict experimental setting. Digital attacks by $AdvLB$ are almost consistent with physical-world attacks by reproduced attack by laser pointers, that the top-3 classification results are similar. In

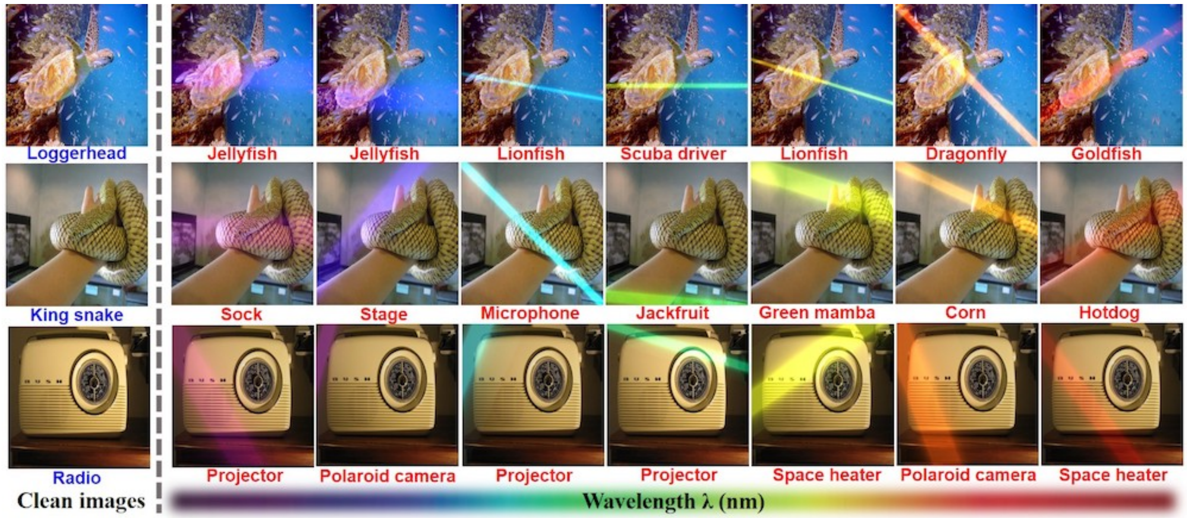


Figure 5: **Adversarial examples generated by $AdvLB$.**

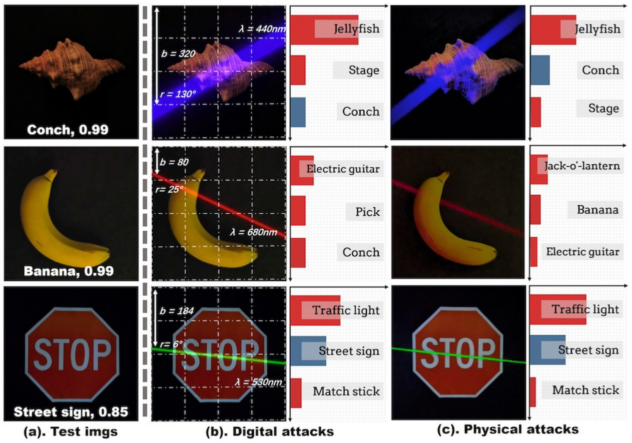


Figure 6: **Indoor test.**

which, the laser beam with $\lambda = 680nm$ is harder to capture and the beam is not as coherent as the other two beams, thus the predicted results in digital and physical are slightly different. In summary, our proposed $AdvLB$ can almost reflect the threats of laser beam in the physical-world, thus can be used to explore the potential real-world threats caused by laser beam.

We further conduct an outdoor test. When a self-driving car approaches the stop sign, even if it fails to recognize the stop sign for merely a short time window, it can lead to a fatal accident. We first apply the method mentioned in Section 3.2 with given captured stop sign. We then shoot the laser beam from position onto targeted stop sign with given returned effective attack range (See the first row in Figure 7). Overall, there is an attack success rate of 77.43% of the test. We also include attack results for real-world traffic signs as shown in the second row in Figure 7. These results further demonstrate the real-world threats by laser beams. Thus our proposed $AdvLB$ could be a very meaningful tool to explore such threats. Currently, a weakness of our proposed method is that it is still limited in attacking in a dynamic environment, we leave this as future work.

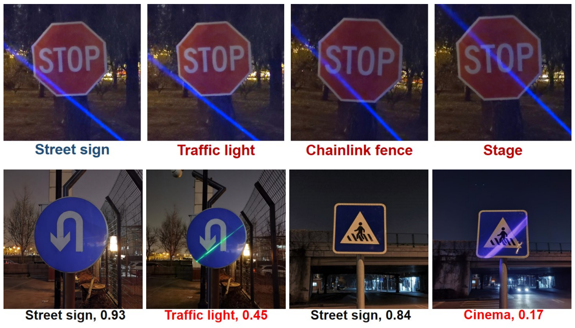


Figure 7: **Outdoor test.**

4.3. Ablation Study

Here, we conduct a series of experiments on ImageNet to study the adversarial effect of laser beam with different parameters: 1) Wavelength (λ), 2) Layout (r, b), 3) Width (w), and another ablation study on how k in k-random-restart impacts the attack success rate of $AdvLB$. We acknowledge the intensity α is an important property, especially for improving stealthiness of laser beam. We will do more study on α in the future work. In the following experiments, we fix it as 1.0. Also, for the study on each parameter of the laser beam, we fix the other parameters.

Wavelength (λ) Here, we show how the wavelength λ impacts the adversarial effects of laser beam. We test the adversarial effects of laser beam with wavelength (λ) in the range of visible light (380 nm, 750nm). We fix other parameters as: r = 45°, b = 0, and w = 20. These values are selected based on extensive experiments, which are effective for finding adversarial images. In this ablation, we perform the tests on ResNet50 with 1000 randomly collected images from ImageNet. We combine the images with laser beams using Eq. (1) and then feed the resultant images to the target model. Table 1 shows the success rates of simulated laser beams with different wavelength λ.

Table 1: **Ablation of Wavelength (λ).**

Wavelength λ (nm)	380	480	580	680
Suc. rate (%)	34.03	48.01	58.93	44.10

As shown in Table 1, the simulated laser beam with $\lambda = 580$ can even achieve 58.9% success rate. Note that for all these experiments, the simulated laser beams are added on unknown images and then fed to unknown target models directly. The results show that the laser beams have universal adversarial effects on different images.

Width (w) We then evaluate how the width w impacts the adversarial effect of laser beam. We set the threshold of width as 40, occupied at most 1/10 of the whole image. Again, we fix other parameters of laser beam as constants: $\lambda = 400$, $r = 30$ and $b = 50$. Wider laser beam can improve the success rate from 30.80% up to 47.69% with width from 1 to 40. However, even with the smallest $w = 1$ (1 pixel width), there is still an impressive success rate (30.80%).

Layout (r, b**)** We then show how the selection of layout (r, b) impacts the attack success rates of laser beams. The range of r is [0°, 180°], and b in range [0, 400]. We fix other physical parameters as constants: $\lambda = 580$, w = 20. For efficiency concern, we only sample 100 correctly classified images from ImageNet, and summarize the results in Figure 8, where each point denotes the success rate of current layout.

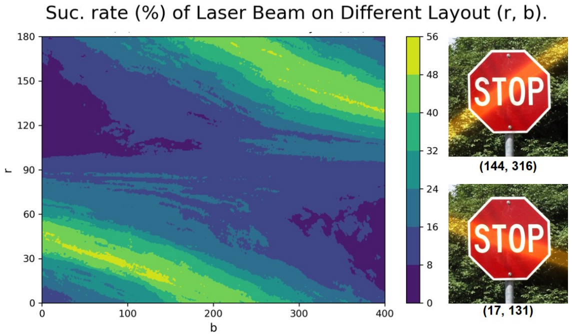


Figure 8: **Ablation of Layout** (r, b**).**

In Figure 8, the left column shows that the attack success rate of laser beam is highly related to its layout. Meanwhile, the right column illustrates two adversarial images (layout with higher attack success rates), which indicate that the center (where the laser beam illuminates) is more likely to create a successful adversarial example.

k**-random-restart** (k) We choose different k of 1, 20, 50, 100, 200 to run the experiment. The results are shown in Table 2, the attack success rate is improved gradually with increase of k. It suggests that we do find better parameters of laser beam with more random restarts.

Table 2: **Ablation of** k**-random-restart (**k**).**

Restart num. (k)	1	50	100	200
Suc. rate (%)	72.80	89.60	92.20	95.10

5. Discussion

Analysis of DNNs' Prediction Errors To better understand the mechanism that the laser beam enables adversarial attacks, we further perform an empirical study with ImageNet to explore the errors caused by the laser beam. Roughly, we find there are two categories of errors as shown in Figure 9. The laser beam performs as a kind of perturbation, which either cancels or changes the original feature of a clean image and brings new cues for DNNs. An example is shown in Figure 9, when the laser beam with wavelength 400nm shines on the hedgehog, whose spines combined with the blue color brought by laser beam form a new cue similar to cardoon for the DNN, thus resulting in misclassification. Results shown in Figure 5 also support for this claim. Also, we find the laser beam itself performs as dominant features for some specific classes (*e.g.* swab shown in Figure 9). We further perform a batch of experiments and then summarize the statistics: which class's percentage rises the most before and after adding the laser beam. We report both top-1 and top-5 rise as shown in Table 3. For each case, we report both percentages of specific class before and after adding laser beam on ImageNet.

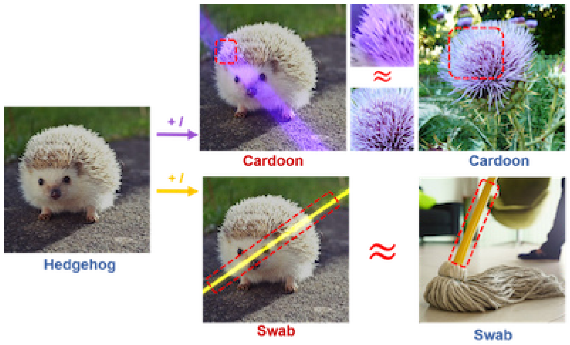


Figure 9: **Two types of errors caused by** $AdvLB$**.**

Table 3: **Statistics of Error caused by** $AdvLB$**.**

λ	**Top1 Pred.**	**Percent. (%)**	**Top5 Pred.**	**Percent. (%)**
380∼470	Feather boa	0.10 → 2.20	Feather boa	0.32 → 8.76
470∼560	Tennis ball	0.10 → 2.46	Spotlight	0.64 → 7.98
560∼650	Rapeseed	0.13 → 1.94	Candle	0.19 → 6.21
650∼740	Volcano	0.11 → 2.14	Gold fish	0.26 → 6.63

As shown in Table 3, the laser beam with different wavelengths λ indeed increases the percentage of some specific classes. For example, when adding a laser beam with wavelength $\lambda = 380$, the percentage of class "Candle" increases the most from 0.19% to 6.21%. The results imply that the laser beam itself serves as dominant feature for some classes, such as spotlight, volcano, *etc*. Thus when adding the laser beam to the clean image, it has the chance that the model is more biased towards the feature brought by the laser beam. We further use the CAM [29] to highlight the bias of model when the laser beam is added to the clean image (as shown in Figure 10). By adding light beam even on the corner of image, the model is more biased towards "Bubble" and "Volcano", thus give wrong top-1 predictions.

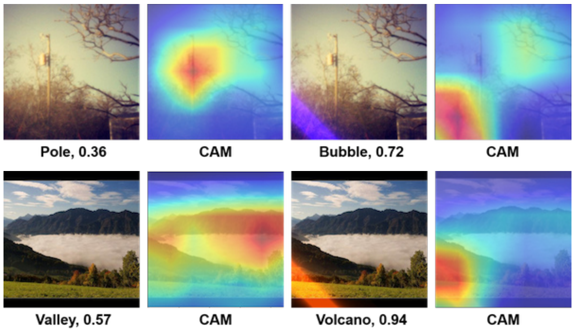


Figure 10: **CAM for images.**

Defense of Adversarial Laser Beam Besides revealing the potential threats of $AdvLB$, in this work, we also try to suggest an effective defense for laser beam attack. Similar to adversarial training, we progressively improve the robustness by injecting the laser beam as perturbations into the data for training. We do not find the worst-case perturbations at each training step like adversarial training [18]. As seeking worst-case laser beam perturbations by $AdvLB$ needs much more computation cost, it may cause the overall training unaffordable. By contrast, we find that training with randomly added laser beams as augmentation can partly strengthen the model under laser beam attacks, and has no negative impact on the recognition of clean images. We use timm[‡] to train the ResNet50 robust model. The model was optimized on 16 2080Ti GPUs by SGD with a cosine annealing learning rate schedule from 0.1 to 1e-5. We add random laser beams on input images with 50% probability for data augmentation. The other hyperparameters are consistent with the reported settings[§]. We summarize the results in Table 4. Except for the attack success rate, we adopt another metric named queries. Queries are the number of times that an attacker needs to query the output from target model before searching for best parameters to attack successfully. Our $AdvLB$ only uses 834 queries on average to break through the original ResNet50$_{ori}$ with a high success rate of 95.1%. In contrast, the defense model ResNet50$_{rob}$ can effectively reduce the attack success rate to 77.2% based on running 2576 queries, showing a certain degree of defense ability against $AdvLB$. Besides, we found an intriguing phenomenon that the accuracy of the model on clean images does not decrease after adding random laser beam augmentation, instead increases slightly by 0.21%. We suggest that introducing additional light source as enhancement makes the model more robust to the confusion as analyzed in Section 5, thus increases the generalization ability of DNN.

Table 4: **Comparison of ResNet50 with and without Defense on $AdvLB$.**

Models	Std. acc(%)	Suc. rate(%)	Queries
ResNet50$_{ori}$	78.19	95.10	834
ResNet50$_{rob}$	78.40 (↑ 0.21)	77.20	2576

6. Conclusion and Future Work

In this paper, we propose $AdvLB$ to utilize the laser beam as adversarial perturbation to generate adversarial examples, which can be applied in both digital- and physical-settings. The proposed attack reveals the existence of an easily implemented real-world threat on DNNs. Some findings resulted from our work open a promising direction for crafting adversarial perturbation by utilizing light (*i.e*., laser beam) rather than generating perturbation manually. The proposed $AdvLB$ is particularly useful in studying the security threats of vision systems at poor light conditions, that could be a meaningful complementary to current physical-world attacks.

In the future, we will improve our proposed $AdvLB$ to be more adapted to dynamic environment. In addition, we will consider the parameter of light intensity into optimization to create a more stealthy adversarial example with simulated laser beam. We will also explore the possibility of using other light pattern (*e.g*. spot light) and light source (*e.g*. natural light) to craft adversarial attacks. Furthermore, we will apply $AdvLB$ on other computer vision tasks including object detection and segmentation. Moreover, effective defense strategies against such attacks will be another crucial and promising direction.

Acknowledgement

This research is partly supported by ARC Discovery Project under Grant DP180100212. This work was also supported by Alibaba Group through Alibaba Research Intern Program. We are grateful for valuable advice given by Dr Xingjun Ma from Deakin University.

[‡] https://github.com/rwightman/pytorch-image-models

[§] https://github.com/rwightman/pytorch-image-models/blob/master/docs/results.md

References

[1] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *ICLR*, 2017.

[2] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. In *NIPS Workshop*, 2017.

[3] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *ACM Workshop on Artificial Intelligence and Security*, pages 3–14. ACM, 2017.

[4] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE S&P*. IEEE, 2017.

[5] Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. Ead: elastic-net attacks to deep neural networks via adversarial examples. In *AAAI*, 2018.

[6] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *CVPR*, 2018.

[7] Ranjie Duan, Xingjun Ma, Yisen Wang, James Bailey, AK Qin, and Yun Yang. Adversarial camouflage: Hiding physical-world attacks with natural styles. *CVPR*, 2020.

[8] Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. Robust physical-world attacks on deep learning models. In *CVPR*, 2018.

[9] 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 *CVPR*, 2018.

[10] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *ICLR*, 2015.

[11] Chuan Guo, Jacob Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Weinberger. Simple black-box adversarial attacks. In *ICML*, 2019.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[13] Hossein Hosseini and Radha Poovendran. Semantic adversarial examples. In *CVPR Workshop*, 2018.

[14] Lifeng Huang, Chengying Gao, Yuyin Zhou, Cihang Xie, Alan L Yuille, Changqing Zou, and Ning Liu. Universal physical camouflage attacks on object detectors. In *CVPR*, 2020.

[15] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *ICLR*, 2016.

[16] Aishan Liu, Xianglong Liu, Jiaxin Fan, Yuqing Ma, Anlan Zhang, Huiyuan Xie, and Dacheng Tao. Perceptual-sensitive gan for generating adversarial patches. In *AAAI*, 2019.

[17] Hsueh-Ti Derek Liu, Michael Tao, Chun-Liang Li, Derek Nowrouzezahrai, and Alec Jacobson. Beyond pixel norm-balls: Parametric adversaries using an analytically differentiable renderer. In *ICLR*, 2018.

[18] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *ICLR*, 2018.

[19] Dinh-Luan Nguyen, Sunpreet S Arora, Yuhang Wu, and Hao Yang. Adversarial light projection attacks on face recognition systems: A feasibility study. In *CVPR Workshops*, 2020.

[20] Ali Shahin Shamsabadi, Ricardo Sanchez-Matilla, and Andrea Cavallaro. Colorfool: Semantic adversarial colorization. In *CVPR*, 2020.

[21] 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 *CCS*, 2016.

[22] Meng Shen, Zelin Liao, Liehuang Zhu, Ke Xu, and Xiaojiang Du. Vla: A practical visible light-based attack on face recognition systems in physical world. *ACM IMWUT*, 2019.

[23] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2013.

[24] Rey Reza Wiyatno and Anqi Xu. Physical adversarial textures that fool visual object tracking. In *ICCV*, 2019.

[25] 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 *CVPR*, 2019.

[26] Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, and Xue Lin. Adversarial t-shirt! evading person detectors in a physical world. *arXiv preprint arXiv:1910.11099*, 2019.

[27] Xiaohui Zeng, Chenxi Liu, Yu-Siang Wang, Weichao Qiu, Lingxi Xie, Yu-Wing Tai, Chi-Keung Tang, and Alan L Yuille. Adversarial attacks beyond the image space. In *CVPR*, 2019.

[28] Zhengyu Zhao, Zhuoran Liu, and Martha Larson. Towards large yet imperceptible adversarial image perturbations with perceptual color distance. In *CVPR*, 2020.

[29] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.

[30] Zhe Zhou, Di Tang, Xiaofeng Wang, Weili Han, Xiangyu Liu, and Kehuan Zhang. Invisible mask: Practical attacks on face recognition with infrared. *arXiv preprint arXiv:1803.04683*, 2018.