LIGHT PROJECTION-BASED PHYSICAL-WORLD VANISHING ATTACK AGAINST CAR DETECTION

Huixiang Wen, Shan Chang, Luo Zhou*

School of Computer Science and Technology, Donghua University, Shanghai, China

ABSTRACT

Physical adversarial attacks directly apply adversarial perturbations to real-world objects. Perturbations usually are printed as patches and pasted on target objects. This requires attackers in the vicinity of targets, which may not be feasible in practice. In this paper, we propose a stealthy physical adversarial attack by taking advantage of the transient of light projection. The attacker utilizes a drone with a portable projector to project the adversarial light pattern on the rear windshield of a vehicle to obstruct the object detector (OD) in autonomous driving systems. This can lead to serious safety vulnerability. We train digital perturbations by back propagation on the OD in an iterative manner. Unfortunately, they do not work well in the form of light patterns due to distortion, double imaging and partial reflectance when projected as light pattern. Hence, we model the mapping from digital to light projections, and use the inverse of the mapping to compensate the projection distortion in each iteration. We employ four state-of-the-art ODs to demonstrate the effectiveness and robustness of our proposed attack.

Index Terms— Autonomous driving, Object detection, Physical adversarial attack, Projection attack

1. INTRODUCTION

Image-based object detection is the essential perception part of vision-based autonomous driving (AD) or unmanned driving (UD) systems (e.g., Tesla and Toyota), where images captured by the on-board camera are sent to the object detector (OD), a deep neural network (DNN), to discover and identify objects of interest, simultaneously, and real-time responses are made accordingly [1]. Since DNNs are known to be vulnerable to adversarial examples [2, 3, 4], estimating when an OD fails to work is of great concern to trustworthy AD.

In digital domain, numerous adversarial attacks have been demonstrated in the scene of AD, where digital images corresponding to various driving scenarios are deliberately modified, causing misclassification of traffic signs [5], misidentification of pedestrians or vehicles [6, 7], departure from lane [8] etc. These digital attacks assume an adversary who has access to the images captured by the camera before they are fed into the OD, which is however unrealistic. Comparing with that, *physical attacks* directly apply adversarial perturbations to real-world objects, and receive more attention recently. There exists non-negligible gap between the attacks in digital and physical domains, which comes from 1) the inevitable losses (e.g., print loss) incurred in the materialization of digital perturbations; 2) a wide range of transformations (e.g., distances, angles, exposures, etc.) induced when perturbations are captured by camera.

In the literature, 'print and attach' is the most common practice of physical-world adversarial attacks, where adversarial perturbations are printed as patches or stickers and then pasted on the surface of the target objects (e.g., traffic sign) or placed in the attack scenes (e.g., on the roadside) [9, 10]. However, these approaches would suffer from several troubles: 1) it may be impossible to physically approach the attack scene or target object in some cases; 2) it is difficult for a print-based attack to be stealthy, since the disturbance is permanent, and the identity of adversary might be exposed during the placement of disturbances; 3) the printed ones demonstrate poor attack performance under low ambient light.

Recently, an emerging class of physical-world attacks is light projection-based [11, 12, 13]., which is transient and has been used to fool image classifiers without physically touching the target objects. B. Nassi *et al.,* [14] demonstrate that an on-board OD can consider depthless projected phantoms, i.e., virtual objects, as real ones, which may trigger dangerous reactions such as sudden braking [8, 15]. The goal of the phantom attack is to let the OD 'see' objects that do not exist. In this paper, we propose a light projection-based attack with the opposite goal, i.e., making real objects 'disappear' in ODs, named *vanishing attack* (VA), which may cause serious consequences like crash. Specifically, we use a portable projector equipped on a drone to project the elaborate adversarial pattern on the rear windshields of target cars, so that the projected ones cannot be noticed by an OD (see an example of the VA in Figure 1).

We emphasize that it is more difficult to generate effective and robust projection-based perturbations than print-based ones, while directly applying print-based adversarial perturbations to VAs fails to meet our expectations (see in Section 4). To successfully launch VAs, the following practical challenges should and have been well considered and solved.

Firstly, the two planes where the camera lens and the windshield located are not parallel, resulting in projection deformations, which cannot be represented by the commonly

This research was supported in part by National Natural Science Foundation of China (Grant No. 61972081, 62202001), the Natural Science Foundation of Shanghai (Grant No. 22ZR1400200). Shan Chang is the corresponding author (E-mail: changshan@dhu.edu.cn).

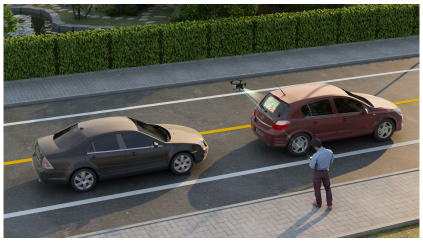

Fig. 1. An example of light projection-based attack.

used geometric transformations such as translation, rotation, and etc. To handle this, we first determine the geometric constraints of projections, and then utilize perspective transformation to mapping digital patches to the corresponding real-world projections. Secondly, the projected adversarial patch is not only imaged on the front surfaces of the windshield, but also on its rear surface, resulting in *double image* when observed by camera, which degrades patch performance. To solve this problem, we model double image by copying and translating the patch, and mixing the original with the replica proportionally. Last but not least, most safety specifications for vehicles, e.g., GB7258-2017 [16], make mandatory requirements that the light transmittance of car window must be no less than 70%. It implies that most of the light projected on the rear windshield will pass through it, and the remainder of light (reflected back) is too subtle to fool ODs. In fact, within the projected area on the windshield, the light perceived by a camera is the superposition of the reflected light of the projection pattern and that of the objects behind the windshield after passing through it. Therefore, we model the patch as a translucent film covering the target object, rather than a mask replacing the area behind. The transmittance of film α is used to adjust the mixing ratio of the two reflected lights.

After passing through a pipeline of above mappings, the patch is applied to clean images. A white-box attacker uses back propagation to optimize the 'projection' in training images, which is re-mapped back to the output space of projector, to attack black-box ODs. We evaluated the projected patch generated by our solutions against multiple state-of-the-art ODs (YOLOv3 [17], YOLOv2 [18], SSD [19], FasterRCNN [20]) in different physical environments systematically. The results show that our attack approach has a strong robustness.

2. ATTACK MODEL

Attack Goal and Scenarios. We define our vanishing attack as projecting an adversarial patch on a target car intended at causing ODs fails to detect it. Considering that, in most cases of normal driving, the front cameras of cars will capture the rear of the cars in front (in the same or adjacent lanes), the attacker can project the adversarial light pattern onto the rear windshield of the front cars by controlling a drone equipped with a portable projector remotely at a distance of several kilometers. In order to simplify complex factors in the physical world, we make the following assumptions: 1) Assuming the projector is only slightly out-of-focus; 2) Both the target car and drone are travel at a constant speed, keeping a fixed distance as much as possible.

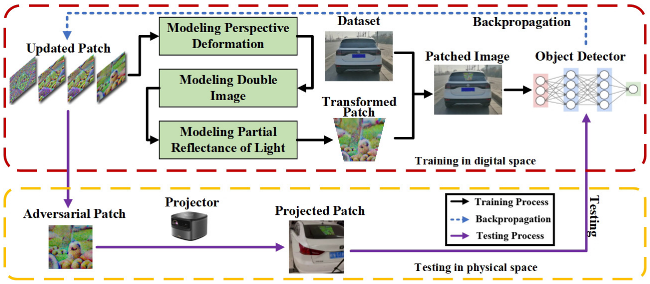

Fig. 2. Overview of car VAs.

Threat Model. We consider both *white-box* and *black-box* attacks. In white-box attacks, the attacker has full access to the target ODs. White-box is a reasonable assumption, for example the attacker could obtain the source code of an OD by using reverse engineering techniques. In black-box attacks, the attacker could only query the outputs of ODs. We assume that the attacker uses an open source OD to generate adversarial patches to attack those black-box systems, which exploits the property that adversarial patterns are highly transferable across neural network architectures.

3. PROPOSED METHOD

We use m to denote an image with respect to a scene without our attack captured by the camera. p refers to an original patch in digital domain, and o is the target object in the scene to be projected by p. $\mathcal{S}_o$ is the objectness score of the OD with respect to o. We model the process that p goes through the projector, is reflected by o and captured by the camera using $\mathcal{T}(\cdot)$, i.e.,

$$p_{(psy)} = \mathcal{T}(p, m, l).$$

where $p_{(psy)}$ is actual patch captured by the camera, l is the location to apply p.

$\mathcal{T}(\cdot)$ is composed of a pipeline of mappings, modeling projective deformation, double image and partial reflectance of visible light, sequentially.

Modeling Perspective Deformation. We denote the ith pixel in p as p_i, whose coordinates is (x_{p_i}, y_{p_i}), and approximate the rear windshield of a car as the viewing plane on which p will be projected. The perspective transformation mapping p_i to its new coordinates (x'_{p_i}, y'_{p_i}) on the projection plane can be represented using a 3*3 homography matrix $\mathbb{H}_{3\times3}$ as [11]

$$w \cdot \begin{bmatrix} x'_{p_i} \\ y'_{p_i} \\ 1 \end{bmatrix} = \mathbb{H}_{3\times3} \cdot \begin{bmatrix} x_{p_i} \\ y_{p_i} \\ 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & 1 \end{bmatrix} \cdot \begin{bmatrix} x_{p_i} \\ y_{p_i} \\ 1 \end{bmatrix}.$$

Then, we rewrite the above transformation to obtain:

$$\begin{bmatrix} x'_{p_i} \\ y'_{p_i} \end{bmatrix} = \frac{1}{w} \cdot \begin{bmatrix} a_{11}x_{p_i} + a_{12}y_{p_i} + a_{13} \\ a_{21}x_{p_i} + a_{22}y_{p_i} + a_{23} \end{bmatrix} \overset{\text{def}}{=} \mathcal{P}_{\mathbb{H}_{3\times3}}(p_i),$$

where w is equal to $a_{31} \cdot x_{p^{(i)}} + a_{32} \cdot y_{p^{(i)}} + 1$.

Providing four vertices of the projected quadrangle, $\mathcal{P}$ can be determined. It is worth mentioning that translation, rotation, and affine can also be represented by using $\mathbb{H}_{3\times 3}$.

Modeling Double Image. When a beam of light is projected on the windshield, it is reflected twice on the front and rear surfaces of the glass, respectively. The light reflected on the rear surface passes through the front surface for the second time, and is captured by the camera, causing *double image*. As shown in Figure 3, a beam of light propagates to point $A(A')$, a partial of it is reflected back to be captured by the camera, the propagation path of which is $projector \rightarrow A(A') \rightarrow camera$. The rest of it passes through the front glass surface to reach point $B(B')$, then is reflected back again and outgoing from point C, and captured by the camera finally. Accordingly, the propagation path is $projector \rightarrow A(A') \rightarrow B(B') \rightarrow C(C') \rightarrow camera$. As a result, the light is imaged twice, i.e., on points A and C, respectively. Moreover, the larger the angle of incidence, the farther the distance between the two images, e.g., $\|A'C'\|_2 > \|AC\|_2$, and thus the more obvious of double imaging.

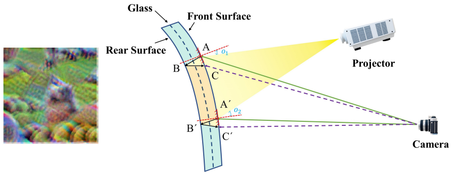

Fig. 3. An example of double image.

Traditional ways of alleviating image blurring, e.g., adding random noise, cannot effectively deal with double image. We introduce the following function to transform the digital path p to its double imaged version p^*. First, we translate p along the horizontal and vertical axes in its coordinate system by Δx and Δy pixels, respectively, and denote the newly obtained patch as $p_{(\Delta x, \Delta y)}$. Then, p^* is calculated by

$$p^* = \underbrace{\beta \cdot p_{(\Delta x, \Delta y)} + p}_{\mathcal{D}(p, \beta, \Delta x, \Delta y)}.$$

where $0 < \beta < 1$.

Modeling Partial Reflectance of Light. When the adversarial patch is projected on the windshield, most of the visible light passes through it, and only a small fraction of light is reflected back. Meanwhile, the light reflected by the object in the scene is not obstructed by the projection and is still reflected back. As a result, both kinds of reflected light are captured by the camera. In our attack, the adversarial patch is considered as a semitransparent film that masks a part of the image. In other words, the color of each pixel captured by the camera is a mixture of the projected color and background color in certain proportion.

Specifically, we use $\mathbf{C}_{p_i} = (R_{p_i}, G_{p_i}, B_{p_i})$ to indicate the RGB channels (in the color space) of p_i. Assuming the visible light reflectance of the rear window is $\alpha \in (0, 1)$, and the pixel on the image m masked by $p^{(i)}$ is $m^{(i)}$. Then, the color of $p^{(i)}$ captured by the camera can be calculated by

$$\begin{bmatrix} \widetilde{R}_{p_i} \\ \widetilde{G}_{p_i} \\ \widetilde{B}_{p_i} \end{bmatrix} = \alpha \cdot \begin{bmatrix} R_{p_i} \\ G_{p_i} \\ B_{p_i} \end{bmatrix} + (1 - \alpha) \cdot \begin{bmatrix} R_{m_i} \\ G_{m_i} \\ B_{m_i} \end{bmatrix} \overset{\text{def}}{=} \mathcal{R}(p_i, m_i, \alpha).$$

At this point, we rewrite $\mathcal{T}(\cdot)$ as

$$\mathcal{T}(p, m, l) = \mathcal{R}(\mathcal{D}(\mathcal{P}_{\mathbb{H}_{3\times 3}}(p), \beta, \Delta x, \Delta y), l, \alpha).$$

Generating Adversarial Patches. We formulate the VA as finding p that minimizes the sum of object loss $\mathbb{L}_{obj}$ and smooth loss $\mathbb{L}_{smh}$ scaled by λ determined empirically, i.e.,

$$\mathbb{L}_{total} = \mathbb{L}_{obj} + \lambda \mathbb{L}_{smh}.$$

Specifically, minimizing $\mathbb{L}_{obj}$ refers to

$$\underset{p}{\operatorname{argmin}} \mathbb{E}(\mathcal{S}_o(\mathcal{T}(p, m, l))).$$

$\mathbb{L}_{smh}$ is the smooth loss which ensures the color changes only gradually within patches [21], such that optimized patch is smooth and consistent. That is because the extreme color differences between adjacent pixels are unlikely to be precisely captured by cameras. We assume that $p_{x,y}$ represents the pixel located at the coordinates (x, y). Minimizing L_{smh} refers to

$$\underset{p}{\operatorname{argmin}} \sum_{x,y} \sqrt{\left(\mathbf{C}_{p_{x,y}} - \mathbf{C}_{p_{x+1,y}}\right)^2 + \left(\mathbf{C}_{p_{x,y}} - \mathbf{C}_{p_{x,y+1}}\right)^2}.$$

In the implementation of our attack, we first randomly generate p^0 as the initial digital patch, and apply $\mathcal{T}(\cdot)$ to p^0 to generate the observed patch $p^0_{(psy)}$. We update $p^n_{(psy)}$ by using a standard attack optimization, e.g., projected gradient descent (PGD) with l_∞ constraint [22], iteratively. Figure 2 shows the overview framework of the proposed car VAs. The per-iteration updating can be written in a generic form as

$$p^{n+1}_{(psy)} = Vanishing(m, p^n_{(psy)}, o).$$

Then, we use the pre-computed inverse of $\mathcal{T}$, denoted as $\mathcal{T}^{-1}$, to get $p^{(n+1)}$, i.e.,

$$p^{(n+1)} = \mathcal{T}^{-1}(p^{n+1}_{(psy)}).$$

4. EXPERIMENTS

4.1. Methodology

We collect a dataset of 1150 images as the train set, and then we capture several videos as the test set. Subsequently, the state-of-the-art ODs are included as the target detectors such as FasterRCNN, SSD, YOLOv2 and YOLOv3. In addition, we compare the attack success rate (the proportion of undetected video frames in a continuous video frame) of printed patches and projected patches in different physical environments. For safety reasons, the experiment is conducted in a parking lot next to the campus road. The iphone 12pro camera is used to simulate the victim's camera. Figure 4 (Left) depicts the patch printed and attached to the rear windshield of the vehicle. Figure 4 (Right) a drone (DJI Air 2) with a portable projector (lenovo T6X) is used to project the adversarial patch onto the rear windshield of the vehicle.

Success Rate	p	$\mathcal{R}(p)$	$\mathcal{P}(p)$	$\mathcal{D}(p)$	$\mathcal{R}(\mathcal{P}(p))$	$\mathcal{P}(\mathcal{D}(p))$	$\mathcal{R}(\mathcal{D}(p))$	$\mathcal{R}(\mathcal{D}(\mathcal{P}(p)))$
Average	15.07%	15.89%	21.69%	15.48%	40.03%	22.10%	22.38%	**83.86%**

Table 1. Ablation study.

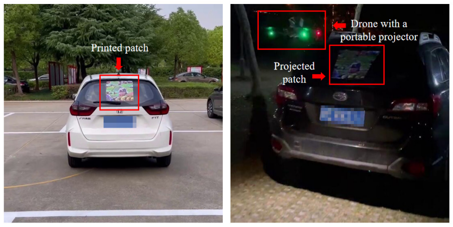

Fig. 4. Experimental deployment.

	YOLOv3	YOLOv2	F-RCNN	SSD
YOLOv3	**92.93%**	64.83%	62.41%	62.93%
YOLOv2	36.55%	**83.28%**	53.79%	53.27%
F-RCNN	6.21%	6.89%	**84.65%**	20.52%
SSD	2.58%	5.34%	16.03%	**83.79%**
YOLOv3*	86.90%	66.55%	64.14%	60.69%

Table 2. Transferability in cross-ODs testing

4.2. Determining Key Parameter λ

Figure 5 (Left) plots the object score when λ takes different values. It can be seen that, when $\lambda = 2.5$, The model only needs to iterate 200 times, and the object score can be reduced to 0.5. Figure 5 (Right) shows the CDF curves of the object score with different λ. We can observe that when $\lambda = 2.5$, the results with the object score below 0.5 accounted for 90% of the population. Therefore, we set λ as 2.5 in the following experiments.

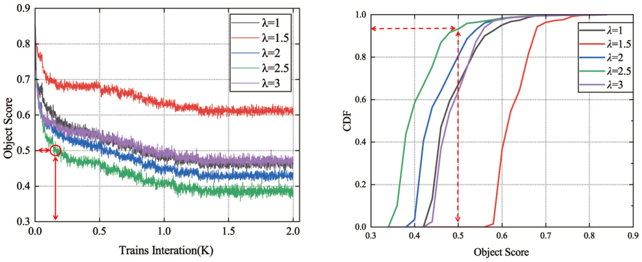

Fig. 5. Object score under different λ.

4.3. Experimental Results

4.3.1. Ablation study

Table 1 shows the attack success rate under different considerations. Among them, the patch without any improved transformation has only 15.07% attack success rate. Therefore, when we remove any one of the improved transformation steps, the results will be reduced significantly.

4.3.2. Performance Comparison

We evaluate the patch of projecting and printing (generated based on YOLOv3) against various factors including distance, angle, and illumination. To examine the impact of varying distances and angles, we divided the distances 1m - 5m into five regions (each region is 1m), and recorded video in each region at the angles 0°, 15°, 30°, 45°, and 60°, respectively. The comparison results are shown in Figure 6.

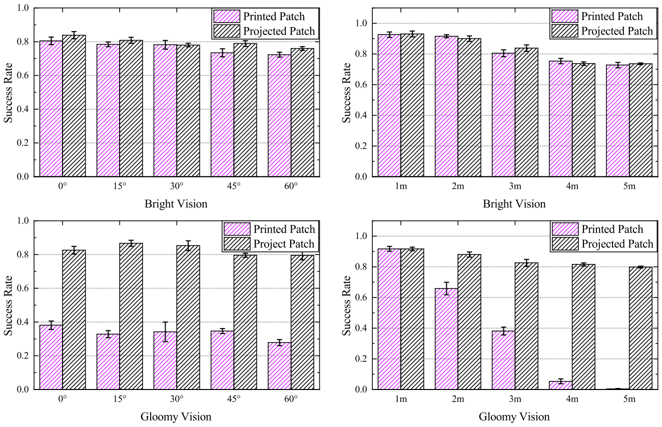

Fig. 6. Attack success rates at different angles, distances, and illuminations.

We can observe that the two attacks can achieve similar success rates in a bright vision. However, projected patch can perform better in gloomy vision, because the less ambient light affects the projection, the clearer the patch is captured by the camera.

4.3.3. Transferability of attacks

In digital-world, we train adversarial patches on different detection models and fed the synthesized test images to some black-box models (YOLOv3, YOLOv2, FasterRCNN and SSD). In addition, we project the patch trained on YOLOv3 in physical-world vehicles, and the above ODs are included for detection. The experimental results are shown in Table 2. It is not difficult to find that a higher success rate can be obtained by training and testing on the same detection model.

5. CONCLUSIONS

In this paper, we proposed a physical-world vanishing attack against car detection, which leverages a drone with portable projector to project the adversarial patch onto the rear windshield in front of the victim's vehicle. Our approach is not only simple to operate, but also does not reveal the identity of the attacker. Extensive experiments demonstrate the effectiveness and robustness of projection attack.

6. REFERENCES

[1] Arnold Eduardo, Y Al-Jarrah Omar, Dianati Mehrdad, Fallah Saber, Oxtoby David, and Mouzakitis Alex, "A survey on 3d object detection methods for autonomous driving applications," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, pp. 3782–3795, 2019.

[2] Dawn Song, Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Florian Tramer, Atul Prakash, and Tadayoshi Kohno, "Physical adversarial examples for object detectors," in *12th USENIX workshop on offensive technologies (WOOT 18)*, 2018.

[3] Anish Athalye, Nicholas Carlini, and David Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *International conference on machine learning*. PMLR, 2018.

[4] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok, "Synthesizing robust adversarial examples," in *International conference on machine learning*. PMLR, 2018.

[5] Chawin Sitawarin, Arjun Nitin Bhagoji, Arsalan Mosenia, Mung Chiang, and Prateek Mittal, "Darts: Deceiving autonomous cars with toxic signs," *arXiv preprint arXiv:1802.06430*, 2018.

[6] Nir Morgulis, Alexander Kreines, Shachar Mendelowitz, and Yuval Weisglass, "Fooling a real car with adversarial traffic signs," *arXiv preprint arXiv:1907.00374*, 2019.

[7] Simen Thys, Wiebe Van Ranst, and Toon Goedemé, "Fooling automated surveillance cameras: adversarial patches to attack person detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019.

[8] Junjie Shen, Jun Yeon Won, Zeyuan Chen, and Qi Alfred Chen, "Drift with devil: Security of multi-sensor fusion based localization in high-level autonomous driving under gps spoofing," in *29th USENIX Security Symposium (USENIX Security 20)*, 2020.

[9] Yue Zhao, Hong Zhu, Ruigang Liang, Qintao Shen, Shengzhi Zhang, and Kai Chen, "Seeing isn't believing: Towards more robust adversarial attack against real world object detectors," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019.

[10] Jiajun Lu, Hussein Sibai, Evan Fabry, and David Forsyth, "No need to worry about adversarial examples in object detection in autonomous vehicles," *arXiv preprint arXiv:1707.03501*, 2017.

[11] Chengyin Hu and Weiwen Shi, "Adversarial color projection: A projector-based physical attack to dnns," *arXiv preprint arXiv:2209.09652*, 2022.

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

[13] 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*. IEEE, 2020.

[14] Ben Nassi, Yisroel Mirsky, Dudi Nassi, Raz Ben-Netanel, Oleg Drokin, and Yuval Elovici, "Phantom of the adas: Securing advanced driver-assistance systems from split-second phantom attacks," in *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, 2020.

[15] Yulong Cao, Ningfei Wang, Chaowei Xiao, Dawei Yang, Jin Fang, Ruigang Yang, Qi Alfred Chen, Mingyan Liu, and Bo Li, "Invisible for both camera and lidar: Security of multi-sensor fusion based perception in autonomous driving under physical-world attacks," in *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2021.

[16] "Technical specifications for safety of power-driven vehicles operating on roads," *https://openstd.samr.gov.cn/*.

[17] Joseph Redmon and Ali Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[18] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[19] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016.

[20] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[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 *Proceedings of the 2016 acm sigsac conference on computer and communications security*, 2016.

[22] 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.