Physical Adversarial Attack Meets Computer Vision: A Decade Survey

Hui Wei, Hao Tang, Xuemei Jia, Zhixiang Wang, Hanxun Yu, Zhubo Li, Shin’ichi Satoh, Luc Van Gool, and Zheng Wang, Senior Member, IEEE

(Survey Paper)

Abstract—Despite the impressive achievements of Deep Neural Networks (DNNs) in computer vision, their vulnerability to adversarial attacks remains a critical concern. Extensive research has demonstrated that incorporating sophisticated perturbations into input images can lead to a catastrophic degradation in DNNs’ performance. This perplexing phenomenon not only exists in the digital space but also in the physical world. Consequently, it becomes imperative to evaluate the security of DNNs-based systems to ensure their safe deployment in real-world scenarios, particularly in security-sensitive applications. To facilitate a profound understanding of this topic, this paper presents a comprehensive overview of physical adversarial attacks. First, we distill four general steps for launching physical adversarial attacks. Building upon this foundation, we uncover the pervasive role of artifacts carrying adversarial perturbations in the physical world. These artifacts influence each step. To denote them, we introduce a new term: adversarial medium. Then, we take the first step to systematically evaluate the performance of physical adversarial attacks, taking the adversarial medium as a first attempt. Our proposed evaluation metric, hitPAI, comprises six perspectives: Effectiveness, Stealthiness, Robustness, Practicability, Aesthetics, and Economics. We also provide comparative results across task categories, together with insightful observations and suggestions for future research directions.

Index Terms—Adversarial attack, adversarial medium, computer vision, physical world, survey.

I. INTRODUCTION

D EEP Neural Networks (DNNs) have achieved impressive results in a variety of fields: from computer vision [1] to natural language processing [2] to speech processing [3], and it is increasingly empowering many aspects of modern society. Nonetheless, Szegedy et al. [4] discovered in 2014 that adversarial samples can cause DNNs-based models to produce incorrect predictions, leading to a significant degradation in performance. This is a groundbreaking work that exposes the vulnerability of DNNs, casting a shadow over their reliability and security. Since then, researchers have conducted extensive explorations into adversarial samples, revealing their pervasive existence across a wide range of DNNs-based computer vision tasks [5], [6], [7], [8], [9], [10], [11], [12]. They designed adversarial clothing [13] to evade person detectors, adversarial eyeglasses [14] to deceive face recognizers, etc. Increasingly, these methods use a class of techniques called adversarial attacks.

Generally, adversarial attacks occur by adding perturbations to input data (e.g., image, video) and fooling the DNNs-based models in the inference stage. Regarding the various domains, adversarial attacks can be categorized into two distinct classes: i) Digital Adversarial Attack, which occurs in the digital space through the addition of subtle perturbations (e.g., style perturbations [19] and context-aware perturbations [20]). ii) Physical Adversarial Attack, which occurs in the real world using tangible artifacts that contain adversarial perturbations (e.g., adversarial patches [21], [22] and adversarial stickers [23], [24]). Compared to the former, the latter poses an augmented threat to social security, raising significant apprehensions, particularly in safety-critical domains like autonomous driving, video surveillance, and facial biometric systems. To facilitate a profound understanding and provide in-depth insights into this topic, we present a comprehensive review of articles on physical adversarial attacks in computer vision tasks.

Though some existing surveys have also summarized the physical adversarial attack methods [15], [16], [17], [18], they primarily focus on listing and categorization, ignoring the evaluation and comparison (see Table I). A unified evaluation criterion
is still absent. This motivates us to take the first step to evaluate the performance of physical adversarial attacks systematically.

To this end, we outline the four general steps (as shown in Fig. 1) required to build a physical adversarial attack:

1) **Step 1. Adversarial perturbation generation:** Generating perturbations in the digital domain based on given DNNs-based models, constrained by different attack forms and attack objectives.

2) **Step 2. Adversarial medium manufacturing:** Designing appropriate physical medium for carrying the perturbations in alignment with attack forms, and subsequently manufacturing them using suitable materials.

3) **Step 3. Threat image capturing:** Applying the adversarial medium in real-world scenarios to be captured by an imaging sensor, thereby generating threat images.

4) **Step 4. Attacking:** The captured threat images are fed to the DNNs-based model to initiate attacks.

Note that we introduce the concept of an “adversarial medium” to denote the tangible artifact responsible for carrying the adversarial perturbation in the real world. According to the four steps mentioned above, we discern the significant role of adversarial mediums in building a physical adversarial attack. They determine the form of perturbations (Step 1), impact manufacturing processes (Step 2), and hold relevance for real-world applications (Step 3). Therefore, we embrace an approach centered on adversarial mediums to examine the existing methods, systematically quantifying and evaluating them in the following six perspectives: Effectiveness, Stealthiness, Robustness, Practicability, Aesthetics, and Economics. Meanwhile, we introduce a comprehensive metric, the hexagonal indicator of Physical Adversarial Attack (hiPAA), and provide comparative results across task categories, along with insightful observations and suggestions for future research directions.

Our contributions can be summarized as follows:

1) Through a comprehensive review of existing methodologies (see Tables III, IV, and V), we abstract and summarize a general workflow for launching a physical adversarial attack, comprising four distinct steps (see Fig. 1).

2) Leveraging this general workflow, we discover that tangible artifacts carrying adversarial perturbations exert substantial influence over attacks, prompting the introduction of the new concept of adversarial medium to represent them.

3) As opposed to existing reviews [15], [16], [17], [18], we take the first step to systematically evaluate the performance of physical adversarial attack, taking adversarial medium as a first attempt. Our proposed metric, hiPAA, comprises six perspectives: Effectiveness, Stealthiness, Robustness, Practicability, Aesthetics, and Economics.

4) We conduct comprehensive comparisons of existing attack methods, and discuss limitations, challenges, and potential directions from the standpoint of real-world applications to facilitate future research.

The remaining article is organized as follows. We first provide a brief introduction to the preliminaries in Section II, which cover essential topics and concepts for the proper understanding of this work. Section III discusses the concept of adversarial mediums. Section IV introduces the evaluation metric for conducting a comprehensive comparison. Then we present the recent advancements in physical adversarial attacks according to mainstream visual tasks and systematically evaluate them for each task in Section V. Moreover, we provide discussions and future research opportunities in Section VI. Finally, we conclude this review in Section VII. An overview of the survey scope is shown in Fig. 2. We also provide a regularly updated project page on https://github.com/weihui1308/PAA.

**II. PRELIMINARIES**

In this section, we provide a concise introduction to problem formulation, key topics, and concepts, aiming to enhance comprehension of our work.

**A. Computer Vision**

Computer Vision (CV) aims to enable machines to perceive, observe, and understand the physical world like human eyes. An important milestone was reached when Krizhevsky et al. [25] proposed AlexNet, which secured victory in the ILSVRC, thus promoting the application of DNNs to address a wide variety of tasks. Up to the present, DNNs-based models have achieved

---

**TABLE I**

**COMPARATIVE ANALYSIS OF CURRENT SURVEYS ON PHYSICAL ADVERSARIAL ATTACK IN COMPUTER VISION**

| Survey | Physical Attack | Adversarial Medium | Evaluation | Number of Methods | Number of Tasks | Year |
|--------|-----------------|--------------------|------------|-------------------|-----------------|------|
| [35]   | ✓               | ✗                  | ✗          | 5                 | 3               | 2016 |
| [16]   | ✓               | ✗                  | ✗          | 47                | 11              | 2022 |
| [37]   | ✓               | ✓                  | ✗          | 69                | 7               | 2022 |
| [18]   | ✓               | ✓                  | ✗          | 22                | 4               | 2023 |
| Ours   | ✓               | ✓                  | ✓          | 86                | 15              | 2024 |

---

**Fig. 1.** The flow of designing a physical adversarial attack, including four main steps: 1) Adversarial perturbation generation in the digital space, 2) Adversarial medium manufacturing in the physical space, 3) Threat image capturing, and 4) Attacking.
impressive and competitive performances in CV tasks, including classification [26], [27], segmentation [28], [29], detection [30], [31], and re-identification [32], [33], [34]. This survey concentrates on three prominent tasks: classification, detection, and re-identification, which are extensively utilized and encompass numerous attack scenarios. Given a DNNs-based model $f : \mathbb{X} \rightarrow \mathbb{Y}$ with pre-trained weights $\theta$, for the sake of conciseness, we formulate the DNNs-based CV models as follows:

$$\hat{y} = f_{\theta}(x), \quad x \in \mathbb{X}, y \in \mathbb{Y},$$

where for any input data $x \in \mathbb{X}$, the well-trained model $f_{\theta}(\cdot)$ is able to predict a $\hat{y}$ that closely approximates the corresponding ground truth $y \in \mathbb{Y}$.

### B. Adversarial Attacks

1) **Problem Formulation**: Adversarial attacks involve introducing perturbations to input data, causing the model $f_{\theta}(\cdot)$ to produce incorrect predictions $\hat{y}$. Note that the attacker’s modification is limited to the input data. Exactly, the sample $x$, following the addition of perturbations $\delta$, is denoted as the

| Chronological overview | Adversarial medium | Definition in this survey |
|------------------------|---------------------|---------------------------|
| 2023                   | Sticker             | Sticker is a piece of paper or plastic, with adversarial perturbation patterns on one side, and needs to be affixed to a host object. There are no size restrictions, and it can cover the entire surface if necessary. |
| 2022                   | Patch               | Patch is a material with limited area that carries perturbations, which can be in the form of paper or cardboard. Its area is smaller than the object to which it is applied. |
| 2021                   | Clothing            | Clothing as an adversarial medium refers to the attacker imprinting or crafting perturbed textures onto garments, constituting a wearable design. Examples include t-shirts, caps, or pants. |
| 2020                   | Light               | Light as the adversarial medium refers to the attacker using light to construct adversarial perturbations in the physical space and then projecting them onto the target object. |
| 2019                   | Camera              | Camera as an adversarial medium refers to attackers exploiting the inherent imaging characteristics or parameter settings of the camera to construct adversarial perturbations. This approach involves introducing adversarial perturbations during the capture of physical scenes, thereby affecting downstream tasks. |
| 2018                   | TC materials        | TC materials as an adversarial medium, involve attackers leveraging specialized physical materials to influence the imaging of thermal infrared cameras. This primarily affects the distribution of thermal radiation, enabling the construction of adversarial textures. |
| 2017                   | Makeup              | Makeup as an adversarial medium refers to attackers using cosmetic techniques to alter facial appearance by applying pre-generated adversarial perturbations to the face. |
| 2017                   | 3D-printed artifact | 3D-printed artifact as an adversarial medium refers to attackers employing 3D printing technology to fabricate adversarial objects with a 3D spatial structure. Such adversarial objects are typically designed to execute real-world attacks from multiple angles. |

We provide detailed definitions for each adversarial medium in this survey. Colors indicate the category of adversarial mediums used. Denotes milestone papers.
| Adversarial Medium | Description | Method | Victim Task | Venue | Year |
|-------------------|-------------|--------|-------------|-------|------|
| **Sticker**       | print, paste | eyeglasses: impersonation, dodging | AdvEyeGlass [14] | Face Recognition | ACM CCS | 2016 |
|                   | cover       | car body: hiding | CAMOU [23] | Vehicle Detection | ICLR | 2018 |
|                   | display     | screen: hiding | InvisibleCloak [50] | Person Detection | UEMCON | 2018 |
|                   | paste       | camera lens: misclassification | ACS [51] | General Classification | PMLR | 2019 |
|                   | print, paste | eyeglasses: impersonation, dodging | AdvEyeGlass+ [52] | Face Recognition | TOPS | 2019 |
|                   | print, paste | hat: impersonation | AdvHat [24] | Face Recognition | ICPR | 2020 |
|                   | print, paste | eyeglasses: misrecognition | CLBAAttack [53] | Face Recognition | BIOSIG | 2021 |
|                   | affix       | car body: hiding | DAS [54] | Vehicle Detection | CVPR | 2021 |
|                   | affix       | road marking: misdirection | AdvMarkings [55] | Lane Detection | USENIX | 2021 |
|                   | full cover  | car body: hiding | FCA [56] | Vehicle Detection | AAAI | 2022 |
|                   | full cover  | car body: hiding | DTA [57] | Vehicle Detection | CVPR | 2022 |
|                   | imprint     | face mask: dodging | AdvMask [58] | Face Recognition | ICMI PKDD | 2022 |
|                   | print, paste | face: impersonation | AdvSticker [59] | Face Recognition | TPAMI | 2022 |
|                   | cover       | car body: hiding | CAC [60] | Vehicle Detection | IJCAI | 2022 |
|                   | print, paste | face: impersonation | DOPatch [61] | Face Recognition | arXiv | 2023 |
|                   | print, paste | car body: false estimation | 3D2Fool [62] | Depth Estimation | CVPR | 2024 |
| **Patch**         | print, paste | image patch: misclassification | AdvPatch [21] | General Classification | NIPS | 2017 |
|                   | print, paste | traffic sign: misclassification | RP [63] | Sign Classification | CVPR | 2018 |
|                   | print, paste | traffic sign: misrecognition | NestedAE [64] | Sign Detection | CCS | 2019 |
|                   | display     | screen: lose track | PS-GAN [65] | Sign Classification | AAAI | 2019 |
|                   | print       | image patch: hiding | PAT [66] | Object Tracking | ICCV | 2019 |
|                   | print, paste | image patch: mismatching | AdvYOLO [13] | Person Detection | CVPRW | 2019 |
|                   | print, paste | image patch: false estimation | AdvPattern [67] | Person Re-ID | ICCV | 2019 |
|                   | print, paste | image patch: misclassification | FlowAttack [68] | Flow Estimation | ICCV | 2019 |
|                   | paste       | camera lens: misclassification | AdvACO [69] | General Classification | ECCV | 2020 |
|                   | print, paste | image patch: misclassification | AdvACO+ [70] | General Classification | CVPR | 2020 |
|                   | paste       | image patch: lose track | TAP [71] | Object Tracking | AAAI | 2021 |
|                   | print, paste | image patch: impersonation | TAP [72] | Face Recognition | CVPR | 2021 |
|                   | print, paste | image patch: misclassification | AdvACO+ [73] | General Classification | TIP | 2021 |
|                   | display     | screen: misclassification | AITP [74] | Sign Detection | ACM AISec | 2022 |
|                   | print, paste | image patch: misclassification | CPAAttack [75] | General Classification | NIPS | 2022 |
|                   | print, paste | image patch: false estimation | TNATack [76] | General Classification | TIP | 2022 |
|                   | print       | image patch: false estimation | OA [77] | Depth Estimation | ECCV | 2022 |
|                   | print, paste | image patch: false segmentation | RWAEs [78] | Segmentation | WACV | 2022 |
|                   | print, paste | image patch: false estimation | PAP [79] | Crowd Counting | ACM CSS | 2022 |
|                   | print, paste | image patch: misclassification | DACPatch [80] | General Classification | ECCV | 2022 |
|                   | print, paste | face: impersonation | SOPP [81] | Face Recognition | TPAMI | 2022 |
|                   | print, paste | car: hiding | AcrilAttack [82] | Vehicle Detection | WACV | 2022 |
|                   | paste       | acrylic patch: hiding | AdvInfrared [83] | Person Detection | CVPR | 2023 |
|                   | display     | screen: hiding | T-SEA [84] | Person Detection | CVPR | 2023 |
|                   | paste       | acrylic patch: hiding | CMPatch [85] | Person Detection | ICCV | 2023 |
|                   | print       | image patch: misstatement | CPApatch [86] | Image Captioning | USENIX | 2023 |
|                   | paste       | acrylic patch: hiding | TAPatch [87] | Person Detection | ICCV | 2023 |
|                   | print, paste | image patch: mis(det., cla.) | TPatch [88] | Sign Det. & Cla. | USENIX | 2023 |

We list them by the adversarial medium and time order.

Fig. 2. A general overview of the scope in our survey.
TABLE IV
PHYSICAL ADVERSARIAL ATTACK METHODS THAT USE CLOTHING, IMAGES, AND LIGHTS AS ADVERSARIAL MEDIUMS

| Adversarial Medium | Description | Method | Victim Task | Venue | Year |
|--------------------|-------------|--------|-------------|-------|------|
| Clothing           | print T-shirt, pants, sweaters, mask misdirection | AdV-T-shirt [89] | Person Detection | ECCV | 2020 |
|                    | print sweatshirt | AdvCloak [91] | Person Detection | ICCV | 2021 |
|                    | print sweatshirt | NAP [92] | Person Detection | ACM MM | 2021 |
|                    | print T-shirt, T-shirts, shirts | LAP [92] | Person Detection | CVPR | 2022 |
|                    | crop aerogel material rendering full coverage | InvisClothing [94] | Person Detection | CVPR | 2022 |
|                    | print pants, sweatshirt | AdvCaT [96] | Person Detection | NN | 2023 |
|                    | print picture misclassification | Phys-FSM [97] | General Classification | ICLR | 2017 |
|                    | print picture misclassification | Shapeshifter [98] | Sign Detection | ECML PKDD | 2018 |
|                    | print picture misclassification | RP [99] | Sign Detection | USENIX | 2018 |
|                    | print picture misclassification | D2P [100] | General Classification | AAAI | 2019 |
|                    | print picture misclassification | ABBA [101] | General Classification | NIPS | 2020 |
|                    | print picture false prediction | PhysGAN [102] | Steering Angle Prediction | CVPR | 2020 |
|                    | print picture misclassification | MetaAttack [105] | General Classification | CVPR | 2020 |
|                    | print picture misclassification | ViewFool [106] | General Classification | NIPS | 2020 |
|                    | print picture misclassification | Meta-GAN [107] | General Classification | TIFS | 2020 |
| Image              | project projector misclassification | PTAttack [108] | General Classification | AAAI | 2018 |
|                    | project projector misclassification | Postor [109] | General Classification | IEEE S&P | 2019 |
|                    | project projector impersonation | ALPA [110] | Face Recognition | CVPR | 2020 |
|                    | project projector hiding | SLAP [111] | Sign Detection | USENIX | 2021 |
|                    | project projector misclassification | OPAD [112] | General Classification | ICCV | 2021 |
|                    | project laser pointer misclassification | AdvLBP [113] | General Classification | CVPR | 2021 |
|                    | project shadow misclassification | Adv-Shadow [114] | Sign Classification | CVPR | 2021 |
|                    | project projector misclassification | SPAA [115] | General Classification | IEEE VR | 2022 |
|                    | project projector hiding | AdvLight [116] | Vehicle Detection | ICASSP | 2023 |
|                    | project laser pointer hiding | AdvLS [117] | Sign Detection | PMLR | 2023 |
|                    | project projector impersonation | SLAttack [118] | Face Recognition | CVPR | 2023 |
|                    | reflect mirror misclassification | RFLA [119] | Sign Classification | ICCV | 2023 |

We list them by the adversarial medium and time order.

TABLE V
PHYSICAL ADVERSARIAL ATTACK METHODS THAT USE NICHE ADVERSARIAL MEDIUMS

| Adversarial Medium | Description | Method | Victim Task | Venue | Year |
|--------------------|-------------|--------|-------------|-------|------|
| Camera             | capture rolling shutter effect | InvisPerturb [120] | General Classification | CVPR | 2021 |
| TC material        | heating cardboard heating | AdvISP [121] | General Classification | CVPR | 2021 |
| Makeup             | paste warm/cool paste | HOTCOLD Block [123] | General Classification | CVPR | 2021 |
| 3D-print artifacts | rendering renderer | AdvMakeup [124] | Face Recognition | IJCAI | 2021 |
|                    | 3D-print tangible turtle | EOT [125] | General Classification | PMLR | 2018 |
|                    | 3D-print tangible mesh | 3DAttack [126] | General Classification | CVPR | 2019 |
|                    | 3D-print adversarial metal | AT3D [127] | Face Recognition | CVPR | 2023 |
|                    | 3D-print tangible object | X-Adv [128] | X-ray Item Detection | USENIX | 2023 |
|                    | | T3D [129] | Clr., Det., Cap. | CVPR | 2024 |

We list them by the adversarial medium and time order.

adversarial sample \(x'\). Mathematically, the joint representation is expressed as

\[
\begin{aligned}
    x' &= x + \delta \\
    y' &= f_\theta(x') \\
    \text{s.t. } y' &\neq \hat{y},
\end{aligned}
\]

(2)

where perturbations \(\delta\) are typically constrained by factors such as intensity, size, and the adversarial medium.

2) Distinguishing Adversarial Attacks From Backdoor Attacks and Poisoning Attacks: Apart from adversarial attacks, there are two other widely used attack types: backdoor (a.k.a. trojan) attacks [35, 36, 37, 38] and poisoning attacks [39], [40], [41]. While these three attack categories aim to mislead the model into incorrect predictions, their implementation methods differ fundamentally. We elucidate these distinctions by introducing the DNNs-based model lifecycle, depicted in Fig. 3, which comprises six phases: data collection and preparation, model selection, model training, model testing, model deployment, and model updating.

Poisoning Attack: DNN-based models require substantial data to achieve high performance. Developers often use publicly available datasets or scrape data from online sources, creating opportunities for poisoning attacks. These attacks involve...
In a gray-box setting, the attack lies between black-box and white-box attacks, where the attacker has partial knowledge of the model, such as the model’s output probability distribution or model structure. Considering the adversarial goals, adversarial attacks can be divided into two categories:

**Targeted Attacks:** Targeted attacks aim to mislead DNN-based models to specified labels, e.g., misidentifying a cat image as that of a dog. The specified labels make targeted attacks more challenging to achieve with high success rates.

**Untargeted Attacks:** The untargeted attacks mislead DNNs-based models to any wrong label. Designing an effective untargeted attack only requires making the model’s predictions incorrect, without necessarily concerning what the incorrect predictions are.

### C. The Taxonomy of Adversarial Attacks

Adversarial attacks can generally be categorized into three types based on adversarial knowledge:

**White-Box Attacks:** In a white-box attack, adversaries have full access to data and model details, including network architecture, parameters, and weights. They exploit this information to assess vulnerabilities and adapt their strategies, making this attack method relatively straightforward to execute.

**Black-Box Attacks:** The attackers lack access to the target model’s structure and parameters and can only interact with the model to obtain predictions for specific samples.

**Gray-Box Attacks:** In a gray-box setting, the attack lies between black-box and white-box attacks, where the attacker has partial knowledge of the model, such as the model’s output probabilities inside the carrier. Thus, we introduce the novel concept of adversarial medium for the first time to represent these entities. The medium is commonly used in physics. When a substance exists inside another substance, the latter is the medium of the former. In most cases, properties such as form, density, and shape of the former would influence the properties of the latter. Analogously, for launching attacks in the physical space, the adversarial perturbation must have a carrier, i.e., the adversarial perturbation exists inside the carrier. Meanwhile, the carrier has an effect on the perturbation. Thus we define the carrier as the adversarial medium.

### III. ADVERSARIAL MEDIUMS

We observe a commonality across all physical adversarial attack methods: the necessity of a physical entity to carry the specially designed perturbations. Therefore, we introduce the novel concept of adversarial medium for the first time to represent these entities. The medium is commonly used in physics. When a substance exists inside another substance, the latter is the medium of the former. In most cases, properties such as form, density, and shape of the former would influence the properties of the latter. Analogously, for launching attacks in the physical space, the adversarial perturbation must have a carrier, i.e., the adversarial perturbation exists inside the carrier. Meanwhile, the carrier has an effect on the perturbation. Thus we define the carrier as the adversarial medium.
An adversarial medium is indispensable for physical adversarial attacks. In Tables III, IV, and V, we have compiled a comprehensive list of all the physical adversarial attack methods examined in this paper, organized by medium and chronological order. Additionally, Table II enumerates nine adversarial mediums, which, to our knowledge, encompass all current types. We provide specific definitions for each adversarial medium in this review. Here, we specifically clarify the similarities and differences between stickers and patches, as they can be easily confused. Specifically, they share commonalities, such as 1) the general use of materials like paper or plastic, 2) application to the host object, and 3) limited area coverage. The distinctions lie in 1) the method of application, with stickers emphasizing attachment to the host object, while patches can involve sticking, hanging, or embedding, and 2) the relative size of the covered area, as stickers can cover the entire host object, whereas patches only occupy a small portion.

IV. Evaluation Metric

Despite the widespread attention given to physical adversarial attacks, their performance has been subject to inconsistent and case-by-case evaluations. This phenomenon primarily stems from two factors:

1) Challenges in Replication: The creation of adversarial mediums, these physical entities, often incur high production costs and face various inherent factors during the manufacturing process. Factors such as the variability in printer models used for pattern printing and the difficulties in standardizing aspects like clothing style and color can influence the process. Furthermore, maintaining consistent real-world experimental conditions proves to be a daunting task. These challenges collectively hinder the replication of experiments conducted by others, making comparisons arduous.

2) Varied Objectives: Physical adversarial attack methods pursue diverse dimensions of enhancement. Some methods prioritize increasing attack effectiveness [13], [21], [85], while others place emphasis on improving attack stealthiness [22], [54], [65]. A separate category strives to enhance robustness [90], [93], [102]. Consequently, a single metric struggles to impartially assess multiple methods.

These problems raise our question: how does the actual performance of the physical adversarial attacks? In this section, we take the first step in conducting a comprehensive assessment of physical adversarial attack methods, encompassing all the work in this field.

Since the diverse objectives across various works, we summarize six perspectives from existing literature, namely: Effectiveness, Stealthiness, Robustness, Practicability, Aesthetics, and Economics. Hingun et al. [130] introduced a large-scale realistic adversarial patch benchmark to assess the effectiveness of attacks. Parallel to this work, Wang et al. [131] evaluated the system-level effect of attack methods in autonomous driving. Li et al. [132] proposed an evaluation of visual naturalness. These evaluation were conducted in settings with a single perspective and a single task. Towards a unified evaluation, we introduce the hexagonal indicator of Physical Adversarial Attack (hiPAA) to systematically quantify and compare attack methods across the aforementioned six perspectives. We define the hiPAA as the weighted sum of six components:

$$hiPAA = \lambda_1 \cdot Eff + \lambda_2 \cdot Rub. + \lambda_3 \cdot Ste.$$  

$$+ \lambda_4 \cdot Aes. + \lambda_5 \cdot Pra. + \lambda_6 \cdot Eco.,$$  

(3)

where the weights of \(\lambda_1\) to \(\lambda_6\) are assigned based on the importance of each component. Specifically, we categorized the six components into three tiers based on their importance, assigning different weights to each tier. Effectiveness is allocated the highest tier weight of 0.3, followed by Robustness and Stealthiness with a weight of 0.2 each. The remaining three components are assigned the lowest tier weight of 0.1 each. Overall, the weights for the six dimensions are \(0.3, 0.2, 0.2, 0.1, 0.1, 0.1\). Their sum equals 1.

A. Six Perspectives of hiPAA

Effectiveness: Attack effectiveness is employed to evaluate the influence a method can exert on the victim model. To evaluate the effectiveness of physical adversarial attacks, our primary concern lies in quantifying the degree of performance degradation induced by these attack methods. Note that while the attack success rate (ASR) serves as a prevalent metric, its applicability across all tasks is not universal. For instance, defining attack success in segmentation tasks can pose inherent challenges. Furthermore, to address variations across different tasks, we compute the percentage of performance degradation:

$$Eff. = 1 - \frac{Acc}{Acc'},$$  

(4)

where \(Acc\) and \(Acc'\) represent the model’s accuracy without and with attacks, respectively.

Robustness: Attack robustness is utilized to evaluate the method’s stability in dynamic environments. To evaluate the robustness of physical adversarial attack, we consider three evaluation scenarios: 1) Cross-model: whether the attack method remains effective when the model changes, 2) Cross-scenario: whether the attack method can consistently perform in various real-world scenarios, and 3) Transformation-resistant: whether the attack method can withstand various real-world transformations, including rotation, camera-to-object distances, view angles, etc. As shown in the “Robustness” section of the Fig. 4, we present six candidate variables for assessing robustness. Table IX serves as an example demonstration of robustness evaluation.

Stealthiness: Stealthiness is used to measure the invisibility of adversarial textures or attack methods. In real-world attacks, stealthiness is crucial, as conspicuous attacks are easily detectable and can be thwarted by humans. Since stealthiness is targeted at human observers, we conduct a human study in which participants rate the images using a 5-point Absolute Category Rating (ACR) scale [132].

Aesthetics: Aesthetics evaluates the social acceptability of the adversarial textures or attack methods. This metric is crucial, especially for wearable adversarial mediums [89], [92], [93].
Unusual patterns or appearances may lead users to reject them, whereas solutions that prioritize aesthetics are more likely to be accepted. Similar to stealthiness, we assess aesthetics through human ratings.

Li et al. [132] introduced the Dual Prior Alignment (DPA) network to assess the visual naturalness of physical adversarial attacks. In this review, we adopt two perspectives, namely stealthiness and aesthetics, to provide a more nuanced evaluation of visual naturalness. Stealthiness and aesthetics are closely tied to naturalness. The distinction between stealthiness and aesthetics lies in the former measuring the invisibility of adversarial textures or attack methods, while the latter evaluates the alignment with human aesthetics under the premise of visible adversarial textures.

Practicability: Practicability evaluates the practicality and feasibility of using the attack method in real-world scenarios. It considers factors such as the ease of implementation, availability of resources, and compatibility with existing systems. The higher the practicability, the more feasible and convenient it is to employ the adversarial medium in practical applications. Similar to stealthiness, we assess practicability through human ratings.

Economics: Economics evaluates the resource requirements and expenses associated with implementing and deploying the attack methods. Existing methods [122], [123] often provide descriptions of the costs associated with manufacturing adversarial mediums. We assess affordability based on these costs. For research that does not explicitly detail these expenses, we make estimations based on experience.

### B. Evaluation Details

For all evaluation perspectives, Table VI lists the ranges of values, along with explanations of the score meanings for each dimension. To ensure fairness, we normalized the scores to fall between 0 and 1. In Fig. 4, we present the questions and labels utilized in the human evaluation. We recruited 30 participants, all of whom were over 18 years old, and provided consent to potentially encounter offensive content during the task. Before scoring, we provided them with prerequisite knowledge about physical adversarial attacks and evaluation details. Each participant is required to rate the stealthiness, aesthetics, and practicability of each method according to the provided questions and labels.

### C. Discussion

While the hiPPA offers a valuable means of aggregating multiple performance aspects, it has limitations concerning scenario-specific weighting and cross-task comparison. First,
the hiPPA was designed with a specific set of weights to balance various aspects. However, the importance of these aspects varies across different scenarios. For instance, effectiveness may be prioritized in laboratory settings, while stealthiness and economics might be critical in real-world applications. This variability means that the fixed weights may not always reflect the optimal balance for all use cases. Future work should explore adaptive weighting schemes that dynamically adjust based on scenario-specific requirements or user-defined priorities.

Secondly, the hiPPA score’s effectiveness component is limited by the diversity of CV tasks, which use different datasets, evaluation protocols, and metrics. Our current approach involves extracting accuracy scores from each paper, leading to inconsistencies due to varying metrics and experimental setups. We mitigate this by normalizing the decrease in accuracy before and after the attack. However, this score should still be interpreted cautiously. Future research should focus on developing task-specific evaluation protocols to facilitate more consistent and comparable assessments of physical adversarial attacks on CV tasks.

V. VICTIM TASKS

In this section, we methodically review physical adversarial attacks, categorizing them based on their respective task types. As shown in Fig. 2, to our knowledge, there are presently 15 sub-tasks associated with physical adversarial attacks. We organize these tasks into 4 main groups: classification, detection, re-identification, and others. The chronological overview in Table II marks the milestone works in this field.

A. Attacks on Classification Tasks

Considering the fundamental nature and significant downstream impact, substantial research is concentrated on attacking classification tasks, primarily focusing on general classification and traffic sign classification. Please refer to Figs. 5 and 6 for visual representations.

1) General Classification: Brown et al. [21] proposed the adversarial patch, a milestone in physical adversarial attack methods. The attacker places a printed circular image patch next to the “banana”, effectively deceiving the classifier into recognizing the banana as a “toaster” with high confidence (see Fig. 5). Due to its simplicity in production, ease of deployment, and high attack potency, the adversarial patch immediately garnered widespread attention. Liu et al. [69], [73] developed class-agnostic universal adversarial patches with robust generalization capabilities for attacking classifiers in Automatic Check-out scenarios. Casper et al. [75] designed the “copy/paste” attacks, employing adversarial patches to investigate the reliability and interpretability of models. To unleash the patch-based attack potential, Chen et al. [80] designed a Deformable Adversarial Patch (DAPatch) to explore the impact of patch shapes. They simultaneously optimized shape and texture to enhance the attack performance. While these methods exhibit respectable attack performance, they are prone to visual abnormalities that render them conspicuously noticeable. To generate a natural-looking patch, Doan et al. [76] proposed searching for naturalistic patches with adversarial effects within the latent space z of Generative Adversarial Networks (GANs) [133].

In contrast to patch-based methods that only add perturbations in a limited region, another category of methods adds perturbations across the entire image. Phy-FGSM [97] adds carefully designed perturbations to images from the ImageNet dataset [134], prints them out, and then captures them using a cell phone camera. The results demonstrate that the captured images can still successfully attack Inception v3 [135]. With the generation of digital images and their application in the physical realm, a gap arises between these domains. D2P [100] was the first to address this domain gap, utilizing conditional GANs [136] to model the digital-to-physical transformation. Its objective is to mitigate the impact of this gap on attack performance. MetaAttack [109] and Meta-GAN [107] generate robust adversarial examples to maintain attacks in physical dynamics. These perturbations are subtle and constrained within a certain range $\varepsilon$, i.e., $||\delta|| < \varepsilon$, but they are visible to the human eye, unlike digital adversarial attacks [46] that are imperceptible. To make these perturbations appear natural, ABBA [101] disguises the perturbations as motion blur, and AdvCam [103] conceals them within natural styles. In addition to adding perturbations for attacks, ViewFool [106] claims that changes in the viewpoint of the target object can also affect the classifier’s predictions. It
utilizes adversarial viewpoints to launch attacks and assess the classifier’s robustness.

Inspired by One Pixel Attack [46], Nichols et al. [108] pioneered light-based adversarial attacks on classifiers. This method enables attackers to initiate an attack by directing a specially designed light onto the target object. However, this method has only been validated on low-resolution image datasets CIFAR-10. Man et al. [109] manipulated the imaging of printed figures using the “flare effect” and the “blooming effect” to achieve attacks. OPAD [112] modeled the light emitted by a physical projector to cast adversarial patterns onto target objects, resulting in robust attacks. Duan et al. [113] proposed AdvLB, a method that enables the manipulation of the physical parameters of laser beams to execute adversarial attacks, which can be accomplished using handheld laser pointers in real-world scenarios. To enhance the stealthiness of light-based attacks, Huang et al. [115] devised an optimization algorithm to balance adversarial loss and stealthiness loss.

As a key component of physical adversarial attack workflow, cameras are responsible for the transformation from the physical scene to a digital image. In contrast to altering objects in the physical scene, Li et al. [51] innovatively proposed applying translucent stickers directly onto the camera lens to achieve the attack goal. These specially crafted stickers introduce perturbations into the captured images. Subsequently, Sayles et al. [120] also focused on the camera, leveraging the Rolling Shutter Effect inherent in cameras, combined with adversarially illuminating a scene, to introduce an attack signal into the captured images. Phan et al. [121] designed adversarial attacks that are effective only under specific camera ISP parameter settings. These methods provide an alternative approach to implementing physical adversarial attacks by modifying certain camera attributes or settings.

Apart from utilizing 2D images, methods involving 3D objects have also been explored. Athalye et al. [125] fabricated the first 3D adversarial object—a turtle misclassified as a rifle from multiple viewpoints. Their proposed Expectation Over Transformation (EOT) technique enhances the robustness of adversarial examples, later widely applied in physical adversarial attack domains. Zeng et al. [126] perturbed 3D physical properties under a renderer to deceive 3D object classification and visual question-answering models. Subsequently, aiming to enhance cross-task transferability of 3D adversarial examples, Huang et al. [129] introduced TT3D, enabling attacks across classification, detection, and captioning tasks.

The evaluation results of attacks on general classification tasks are summarized in Table VII. These results indicate that the performance of these attack methods has not continuously improved over the years. The lack of a comprehensive evaluation framework has contributed to this issue. For example, while CPAttack [75] performs well in terms of effectiveness and economics, it neglects robustness.

2) Traffic Sign Classification: The classification of traffic signs plays a pivotal role in aiding autonomous driving systems to comprehend scenes and make informed decisions [137]. Consequently, safety assessments in this domain have garnered considerable attention (see Fig. 6). Eykholt et al. [63] were the first to investigate the robustness of traffic sign classifiers and introduced Robust Physical Perturbations (RP₂), a method that misguides these classifiers by affixing black and white blocks onto signs. Liu et al. [65] proposed a perceptual-sensitive generative adversarial network (PS-GAN), which generates adversarial patches that visually resemble the scrawls typically found on signs in the real world, enhancing their stealthiness. Subsequently, building on potential real-world interferences, Zhong et al. [114] subsequently introduced Adv-Shadow to evaluate the influence of shadows on traffic sign classifiers. Similarly, Wang et al. [119] proposed using reflections from mirrors to launch attacks. Experimental results show that optimized shadows and reflections can effectively deceive target models. Fig. 7 presents a comparison of RP₂, PS-GAN, and Adv-Shadow in terms of the six perspectives of the hiPAA metric. Results show that such attacks demonstrate excellence in effectiveness and economics, but they lack aesthetics and practicality. These observations provide potential directions for further improvement.

B. Attacks on Detection Tasks

Physical adversarial attacks on detectors primarily focus on evading detection. As shown in Fig. 8, we display examples of attacking the vehicle, person, and sign detection tasks.
1) Vehicle Detection: Physical adversarial attacks on vehicle detection focus on applying distinctive patterns to a vehicle’s exterior to evade detection models.

Zhang et al. [23] were the first to effectively camouflage a car in a 3D simulation environment. They employed the photorealistic Unreal Engine 4\(^1\) to simulate the complex transformations induced by the physical environment comprehensively. This engine offers a comprehensive set of configuration parameters, including camouflage resolution, patterns, 3D vehicle models, camera settings, and environmental variables. Similar to this work, Wu et al. [138] employed the open-source simulator CARLA \([139]\). They introduced an Enlarge-and-Repeat process and a discrete search method to craft physically adversarial textures. In addition, Duan et al. [60] utilized the Unity\(^2\) and introduced a method called Coated Adversarial Camouflage (CAC).

The neural renderer is commonly used in 2D-to-3D transformation. One of the applications is to wrap the texture image to the 3D model, which then is rendered to the 2D image \([140]\), \([141]\), \([142]\). Thus utilizing the neural renderer to paint the adversarial stickers onto the vehicle surface is pervasively used. Full-coverage Camouflage Attack (FCA) \([56]\) tries rendering the non-planar texture over the full vehicle surface to overcome the partial occluded and long-distance issues. It bridges the gap between digital attacks and physical attacks via a differentiable neural renderer. Then, FCA introduces a transformation function to transfer the rendered camouflaged vehicle into a photo-realistic scenario. It outperforms other advanced attacks and achieves higher attack performance on both digital and physical attacks.

However, existing neural renderers cannot fully represent various real-world transformations due to a lack of control of scene parameters compared to legacy photo-realistic renderers. Motivated by the challenge, Suryanto et al. [57] presented the Differentiable Transformation Attack (DTA) \([56]\), a framework for generating effective physical adversarial camouflage on 3D objects. It combines the advantages of a photorealistic rendering engine with the differentiability of a novel rendering technique, outperforming previous works in both effectiveness and transferability.

The evaluation results of attack methods on vehicle detection tasks are summarized in Fig. 9. From these results, the following insights can be drawn. First, all attack methods do not perform ideally in terms of stealthiness. This is due to the fact that vehicles are large entities with a single color, and any external alterations are likely to be noticeable. Second, no method excels in all six aspects; each has its own strengths. For instance, AdvLight \([116]\) demonstrates strong robustness but fares poorly in terms of economics.

2) Person Detection: The objective of physical adversarial attacks on person detection is to conceal a person from detection models in the real world. Refer to Table VIII for the hiPPA metric among attack methods for both the person detection task and traffic sign detection task.

Table VIII: Comparison of the hiPPA metric among attack methods for both the person detection task and traffic sign detection task.

| Methods                  | Eff. | Rob. | Sens. | Acc. | PRA | Eco | hiPPA |
|--------------------------|------|------|-------|------|-----|-----|-------|
| InvisibleCloak [50]       | 0.72 | 0.67 | 0.25  | 0.25 | 0.61| 0.29| 0.52  |
| AdvYOLO [13] CVPR2009    | 0.75 | 0.50 | 0.25  | 0.24 | 0.62| 0.96| 0.56  |
| AdvT-shirt [89] ICCV2013 | 0.43 | 0.67 | 0.67  | 0.68 | 0.81| 0.95| 0.64  |
| UPC [90] CVPR2015        | 0.93 | 0.93 | 0.26  | 0.26 | 0.81| 0.91| 0.66  |
| AdvCloak [91] ICCV2013   | 0.50 | 0.83 | 0.61  | 0.60 | 0.81| 0.95| 0.67  |
| NAP [92] CVPR2015        | 0.66 | 0.50 | 0.85  | 0.88 | 0.81| 0.95| 0.73  |
| LAF [22] ACM MM2015      | 0.92 | 0.67 | 0.88  | 0.81 | 0.95| 0.95| 0.73  |
| AdvBulls [122] AAA2015   | 0.65 | 0.83 | 0.23  | 0.43 | 0.64| 0.98| 0.62  |
| TC-EGA [93] CVPR2015     | 0.65 | 1.00 | 0.41  | 0.28 | 0.81| 0.95| 0.68  |
| InvisClothing [84] CVPR2015 | 0.88 | 1.00 | 0.27  | 0.27 | 0.65| 0.92| 0.57  |
| AdvInfrared [83] CVPR2015 | 0.85 | 0.50 | 1.00  | 0.25 | 0.65| 0.95| 0.74  |
| T-SEPA [84] CVPR2015     | 0.99 | 0.67 | 0.26  | 0.24 | 0.69| 0.99| 0.60  |
| DAC [95] AAA2015         | 0.75 | 0.67 | 0.63  | 0.37 | 0.00| 0.93| 0.61  |
| AdvCas [96] CVPR2015     | 0.85 | 1.00 | 0.86  | 0.87 | 0.91| 0.91| 0.89  |
| INOTCOLD Block [123] AAA2015 | 0.57 | 0.67 | 1.00  | 0.27 | 0.47| 0.95| 0.70  |
| CMPatch [80] ICCV2015    | 0.62 | 0.67 | 1.00  | 0.26 | 0.81| 0.95| 0.72  |
| TAPatch [87] ICCV2015    | 0.88 | 0.50 | 1.00  | 0.26 | 0.81| 0.95| 0.77  |
| ShapeShifter [98] IFPA   | 0.93 | 1.00 | 0.26  | 0.24 | 0.63| 0.99| 0.72  |
| RP+ [99] [89] UDNDA      | 0.65 | 0.67 | 0.81  | 0.41 | 0.67| 0.99| 0.76  |
| NortedEal [64] CCEY       | 0.92 | 0.67 | 0.26  | 0.25 | 0.66| 0.99| 0.65  |
| LPAttack [104] AAA2015   | 0.80 | 1.00 | 0.27  | 0.40 | 0.47| 0.99| 0.68  |
| TransPatch [70] CVPR2015 | 0.42 | 0.33 | 0.88  | 0.64| 0.55| 0.99| 0.64  |
| AdvMarkings [35] UDNDA   | 1.00 | 0.83 | 1.00  | 0.64 | 0.89| 0.99| 0.92  |
| SLAP [111] UDNDA21        | 0.99 | 1.00 | 0.25  | 0.26 | 0.65| 0.92| 0.73  |
| ATFP [74] [86] [82]      | 0.90 | 0.83 | 0.27  | 0.46| 0.61| 0.99| 0.70  |
| AdvLS [117] [86] [82]    | 0.66 | 0.50 | 0.86  | 0.62| 0.80| 0.99| 0.69  |
| TPatch [88] UDNDA21       | 1.00 | 0.83 | 0.77  | 0.83| 0.51| 0.57| 0.81  |

---
\(^1\)https://www.unrealengine.com/
\(^2\)https://unity.com/
(40 cm × 40 cm) adversarial patch that, when held by an attacker, can deceive the YOLOv2 [144]. Given an input image, mainstream detectors predict the bounding boxes $V_{pos}$, the object probability $V_{obj}$, and the class score $V_{cls}$. During the training phase, they minimize $V_{obj}$ and $V_{cls}$ to cause the detector to ignore persons. Meanwhile, the INRIAPerson dataset [145] provides person instances that facilitate the generation of effective adversarial patches. Subsequently, T-SEA [84] achieves high attack transferability through a series of strategies that involve self-ensembling the input data, the victim model, and the adversarial patch.

Not being satisfied with attacking detectors with printed cardboard, Xu et al. [89] crafted T-shirts with the generated adversarial patches. We termed their approach AdvT-shirt. Due to the non-rigid deformation of a T-shirt caused by pose changes of a moving person, AdvT-shirt employs a TPS-based transformer to ensure the attack’s efficacy in the physical world. Parallel to this work, Wu et al. [91] made a wearable invisibility cloak that, when placed over an object either digitally or physically, makes that object invisible to detectors. To fairly evaluate the effectiveness of different physical attacks, Huang et al. [90] presented the first standardized dataset, AttackScenes, which simulates the real 3D world under controllable and reproducible settings to ensure that all experiments are conducted under fair comparisons for future research in this domain. In addition, they proposed the Universal Physical Camouflage (UPC) attack, which crafts adversarial patterns by jointly fooling the region proposal network, as well as misleading the classifier and the regressor to output errors.

Legitimate Adversarial Patches (LAP) [22] and DAC [95] aim to evade both human perception and detection models in real-world settings. LAP generates cartoon-style adversarial patches, while DAC integrates background styles into adversarial textures to balance effectiveness and stealthiness. Generative adversarial networks (GANs) have the ability to efficiently generate desired samples [146]. Considering this, NAP [92] crafts adversarial patches by leveraging the learned image manifold of BigGAN [147] and StyleGAN [148] pretrained on real-world images. Moreover, in this work, the MPII Human Pose dataset [149] is introduced to provide the diversity of training data.

To address the segment-changing problem and support multi-angle attacks, Hu et al. [93] introduced a novel generative method, the Toroidal-Cropping-based Expandable Generative Attack (TC-EGA). This approach generates adversarial textures characterized by repetitive structures. Unlike the prior studies, in the first stage, TC-EGA trains a fully convolutional network (FCN) [150], [151] as the generator to produce textures by sampling random latent variables as input. In the second stage, TC-EGA searches for the best local pattern of the latent variable with a cropping technique: Toroidal Cropping [152]. They have produced a variety of clothing items, such as T-shirts, skirts, and dresses, that have been printed with generated adversarial textures. These clothing items are effective for carrying out attacks when the wearer turns around or changes their posture. In the subsequent year, Hu et al. [96] improved the attack’s stealthiness by refining the clothing texture based on the TC-EGA method. Specifically, they introduced camouflage textures with a high naturalness score, resulting in the production of natural-looking adversarial clothing.

The security of DNNs-based models has received substantial attention in the context of visible light, but its exploration in thermal infrared imaging remains incomplete. Additionally, thermal infrared detection systems hold significant relevance in various security-related domains, such as autonomous driving [153] and night surveillance [154]. Thus, Zhu et al. [122] presented a groundbreaking method, named AdvBulbs, which is the first to enable physical attacks on thermal infrared person detectors. Their design is classified as a patch-based attack, with experiments conducted on the Teledyne FLIR ADAS Thermal dataset [155]. The following year, Zhu et al. [94] designed infrared invisible clothing based on a new material aerogel. Compared with the AdvBulbs, infrared invisible clothing hid from infrared detectors from multiple angles. Since then, this field has garnered significant attention. Wei et al. [83] also used aerogel material, with the difference that they designed irregularly shaped patches for the attack. Parallel to this work, Wei et al. [123] ingeniously employed anti-fever stickers and heating pads to create adversarial patches. Instead of optimizing the texture and structural characteristics of the patches, their emphasis lies in studying the impact of patch size, shape, and placement on the attacks. Wei et al. [85] introduced the concept of cross-modal physical adversarial attacks, which remain effective under both thermal infrared and visible light imaging modalities.

3) Traffic Sign Detection: Traffic sign detection involves identifying road signs (e.g., stop signs and lane lines) in driving scenarios, essential for autonomous driving. Due to its critical role in safety and decision-making, adversarial attacks on traffic sign detection have evolved to improve the resilience of detection algorithms. The comparative results of these methods are shown at the bottom of Table VIII.

| Cross-model | Cross-scenario | Transformation | Method | Rob.
|-------------|----------------|---------------|--------|------
| ☒ | ☒ | ☒ | ☒ | ☒ | InvisibleCloak [50] | 0.67 |
| ☒ | ☒ | ☒ | ☒ | ☒ | AdvYOLO [13] | 0.50 |
| ☒ | ☒ | ☒ | ☒ | ☒ | AdvT-shirt [89] | 0.67 |
| ☒ | ☒ | ☒ | ☒ | ☒ | UPC [90] | 0.67 |
| ☒ | ☒ | ☒ | ☒ | ☒ | AdvCloak [91] | 0.88 |
| ☒ | ☒ | ☒ | ☒ | ☒ | NAP [92] | 0.50 |
| ☒ | ☒ | ☒ | ☒ | ☒ | LAP [22] | 0.67 |
| ☒ | ☒ | ☒ | ☒ | ☒ | AdvBulbs [122] | 0.63 |
| ☒ | ☒ | ☒ | ☒ | ☒ | TC-EGA [93] | 1.00 |
| ☒ | ☒ | ☒ | ☒ | ☒ | InvisClothing [94] | 1.00 |
| ☒ | ☒ | ☒ | ☒ | ☒ | AdvInfrared [63] | 0.50 |
| ☒ | ☒ | ☒ | ☒ | ☒ | T-SEA [84] | 0.67 |
| ☒ | ☒ | ☒ | ☒ | ☒ | DAC [95] | 0.67 |
| ☒ | ☒ | ☒ | ☒ | ☒ | AdvCtT [96] | 1.00 |
| ☒ | ☒ | ☒ | ☒ | ☒ | HOTCOLD Block [123] | 0.67 |
| ☒ | ☒ | ☒ | ☒ | ☒ | CMPatch [65] | 0.67 |

* Lig., Bac., Dis., Ang., and Rot. represent lighting, background, camera-to-object distances, view angles, and rotation respectively.
The mainstream detectors prune the region proposals by using heuristics like non-maximum suppression (NMS) [156], [157]. The pruning operations are usually non-differentiable. However, generating adversarial perturbations pervasively requires calculating a backward gradient end to end. The non-differentiable operations make it hard to optimize the objective loss. To tackle this problem, Chen et al. [98] carefully studied the Faster R-CNN object detector [156] and successfully performed optimization-based attacks using gradient descent and backpropagation. Concretely, they ran the forward pass of the region proposal network and fixed the pruned region proposals as fixed constants to the second stage classification in each iteration. The RP2 algorithm of [63] only focuses on attacking the traffic sign classification task. Following this line, Song et al. [99] extended the RP2 to provide proof-of-concept attacks for object detection networks. They experimented with the YOLOv2 [144] and achieved 85.6% ASR in an indoor environment and 72.5% ASR in an outdoor environment.

Lane detection is important for autonomous driving because it supports steering decisions. Jing et al. [55] pioneered the investigation into the security of lane detection modules in real vehicles, focusing on the Tesla Model S. We term their method "Adversarial Markings" as it utilizes small markings on the road surface to mislead the vehicle’s visual system. Extensive experiments show that Tesla Autopilot is vulnerable to Adversarial Markings in the physical world and follows the fake lane into oncoming traffic.

Zolfi et al. [70] designed a universal perturbation, called TransPatch, to fool the detector for all instances of a specific object class while maintaining the detection of other objects. TransPatch is a type of colored translucent sticker, which performs attacks by attaching this special sticker to the camera lens, resulting in disturbing the camera’s imaging. In addition, Giulio et al. [111] introduced SLAP, a light-based technique enabling physical attacks in self-driving scenarios. Using a projector, SLAP projects a specific pattern onto a Stop Sign, causing YOLOv3 [30] and Mask-RCNN [28] to misdetect the targeted object in moving vehicle environments. However, such methods may raise suspicion among human drivers. To address this, Zhu et al. [88] proposed TPatch, innovatively designing a triggered physical adversarial patch. TPatch exhibits malicious behavior only when triggered by acoustic signals; otherwise, it behaves benignly. This approach offers a novel perspective on enhancing the stealthiness of adversarial patches.

### C. Attacks on Re-Identification Tasks

In this section, we review physical adversarial attacks on re-identification (Re-ID) tasks (see Fig. 10). Table X presents the comparative results of these methods based on the hiPAA metric.

#### 1) Face Recognition

Face Recognition Systems (FRS) are widely used in surveillance and access control [158], [159]. It is valuable to explore the potential risks of FRS. Sharif et al. [14] developed a groundbreaking method to attack the face recognition algorithm by printing a pair of eyeglass frames. The person who wears the adversarial eyeglasses is able to evade being recognized or impersonate another individual. The non-printability score (NPS) is defined to ensure the perturbations are printable, and it is widely used as a loss function during adversarial perturbation optimization. They demonstrate how an attacker that is unaware of the system’s internals is able to achieve inconspicuous impersonation under a commercial FRS [160].

Pautov et al. [161] explored physical attacks on ArcFace [162] using adversarial patches. They designed a cosine similarity loss to minimize the similarity between the patched photo and the ground truth. The generated gray patch is easily printable. They tested the patch in three forms: eyeglasses, and stickers on the nose and forehead. Numerical experiments demonstrated effective real-world attacks on ArcFace. Light-based attacks have shown feasibility in classification tasks [112]. For FRS, Nguyen et al. [110] designed a real-time adversarial

#### Table X Comparison of the hiPAA Metric Among Attack Methods for Both the Face Recognition Task and Person Re-ID Task

| Methods                  | Hexagonal Score | hiPAA |
|--------------------------|----------------|-------|
| AdvEyeglass [14] CCA     | 1.00           | 0.69  |
| AdvEyeGlass+ [52] TOPs3  | 1.00           | 0.69  |
| AdvVeh [24] ICPR10       | 1.00           | 0.69  |
| ALPA [110] CVPR19        | 1.00           | 0.69  |
| CLBAAttack [53] RON021   | 0.95           | 0.69  |
| AdvMask [58] EPI1        | 0.96           | 0.69  |
| AdvMask [124] ICASS1     | 0.40           | 0.69  |
| TAP [72] CVPR18          | 1.00           | 0.69  |
| AdvSticker [59] TAMI12   | 0.98           | 0.69  |
| SOPP [81] TAMDI2         | 0.96           | 0.69  |
| SLAttack [118] CVPR18    | 0.65           | 0.69  |
| AT3D [127] CVPR13        | 0.48           | 0.69  |
| DOPatch [61] Adv03       | 0.96           | 0.69  |
| AdvPattern [67] CCA     | 0.96           | 0.69  |

We highlight the minimum and maximum values using **blue** and **red**, respectively.
light projection attack using an off-the-shelf camera-projector setup, targeting state-of-the-art FRS such as FaceNet [163] and SphereFace [164].

Adhat [24] implements an easily reproducible physical adversarial attack on the state-of-the-art public Face ID system [162], [165]. In the digital space, Advhat uses Spatial Transformer Layer (STL) [166] to project the obtained sticker on the image of the face. In the physical space, Advhat launches attacks by wearing a hat with a special sticker on the forehead area, which significantly reduces the similarity to the ground truth class.

Yin et al. [124] proposed AdvMakeup, a unified adversarial face generation method that addresses a common and practical scenario: applying makeup to eye regions to deceive FRS while maintaining a visually inconspicuous appearance, resembling natural makeup. Concretely, AdvMakeup first introduces a makeup generation module, which can add natural eye shadow over the orbital region. Then, a task-driven fine-grained meta-learning adversarial attack strategy guarantees the attacking effectiveness of the generated makeup. Experimental results show that the AdvMakeup attack effectiveness is substantially higher than Advhat [24] and AdvEyeglass [14].

2) Person Re-Identification: Person Re-ID is the task of identifying and tracking an individual of interest across multiple non-overlapping cameras [167]. This task plays an important role in surveillance and security applications. Wang et al. [67] were the first and only ones to propose a physical attack on the Re-ID model, known as AdvPattern. They accomplished evasion and impersonation attacks by formulating distinct optimization objectives. As shown in Fig. 10(b), AdvPattern employs adversarial patches featuring specially crafted patterns as the adversarial medium, which are affixed to a person’s chest. The method degrades the rank-1 accuracy of person Re-ID models from 87.9% to 27.1% and under impersonation attack. This easily implementable approach exposes the vulnerability of the DNNs-based Re-ID system.

D. Attacks on Other Tasks

In addition to the three aforementioned mainstream tasks, physical adversarial attacks extend to eight niche tasks, encompassing optical flow estimation [68], steering angle prediction [102], crowd counting [79], segmentation [78], object tracking [66], [71], monocular depth estimation [77], image captioning [86], and X-ray detection [128]. Table XI presents the comparative results based on the hiPAA metric.

Optical Flow Estimation (OFE): aims to measure the pixel 2D motion of an image sequence [168]. As shown in Fig. 11(a), Ranjan et al. [68] proposed FlowAttack to perturb the OFE models. FlowAttack utilizes the gradients from pre-trained optical flow networks to update adversarial patches. Experimental results show that FlowAttack can cause large errors for encoder-decoder networks but not strongly affect spatial pyramid networks. This phenomenon demonstrates the correlation between network structure and vulnerability.

Crowd Counting: aims to estimate the number of individuals within images or videos, with significant applications in public safety and traffic management [169]. Liu et al. [79] proposed a Perceptual Adversarial Patch (PAP) for attacking crowd-counting systems in the real world. PAP generates an adversarial patch by maximizing the model loss, leading the target victim model to overestimate the count by up to 100 on 80% of the samples (see Fig. 11(b)).

Monocular Depth Estimation (MDE): aims to estimate the distance between the camera and a target object, which is crucial for autonomous driving [170]. Recently, Cheng et al. [77] introduced an attack named OAP against MDE models (see Fig. 11(c)). OAP employs a rectangular patch region optimization method to find the optimal patch-pasting region, resulting in over 6 meters mean depth estimation error and 93% ASR in downstream tasks. Despite its effectiveness, OAP relies on 2D image patches and cannot achieve multi-viewpoint attacks. To address this limitation, Zheng et al. [62] proposed 3D²Fool, which integrates UV mapping into adversarial texture optimization, creating robust 3D camouflage textures capable of making the car vanish.
Semantic Segmentation: aims to classify each pixel into predefined categories without distinguishing between individual object instances [28]. Nesti et al. [78] crafted adversarial patches to perturb the semantic segmentation models. As shown in Fig. 11(d), they created a large adversarial patch, measuring 1 m × 2 m, which disrupts the predictions of segmentation models in the physical world. The adversarial patch is optimized using pixel-wise cross-entropy loss on the pre-trained ICNet [171]. Meanwhile, they built abundant and diverse scenes by the CARLA Simulator [139] for scene-specific attacks. Experimental results show that their attack method can reduce the model’s accuracy in the digital space, but in the real world, the attack is greatly downgraded.

Steering Angle Prediction: assists autonomous driving systems in making informed decisions [172], [173], [174]. To evaluate the safety and robustness of this task, Kong et al. [102] introduced PhysGAN, a method that generates physically resilient adversarial examples to deceive autonomous steering systems. As shown in Fig. 12(a), by utilizing the discriminator within the GAN framework to assess the visual disparities between adversarial roadside signs and their original counterparts, PhysGAN can generate realistic adversarial examples. Meanwhile, it can maintain attack effectiveness continuously across all frames throughout the entire trajectory.

Object Tracking: aims to detect moving objects and track them from frame to frame [175]. Wiyatno et al. [66] proposed the first physical adversarial attack on this task. Specifically, they perform optimization to create a distinctive pattern, which is then presented on a large monitor as a background. When a person moves in front of the monitor, the tracker tends to prioritize locking onto the background and disregards the person. Subsequently, Ding et al. [71] proposed a patch-based attack method to launch universal physical attacks on single object tracking. As shown in Fig. 12(b), in the presence of the patch, the tracker neglects the originally tracked object. These explorations raise security concerns for real-world visual tracking.

Image Captioning: focuses on generating a description of an image, which requires recognizing the important objects, their attributes, and their relationships in an image [176]. Inspired by patch-based attacks, Zhang et al. [86] designed CAPatch, a method capable of inducing errors in final captions within real-world scenarios (see Fig. 12(c)). CAPatch deceives image captioning systems, causing them to produce a specified caption or conceal certain keywords. In contrast to existing attack methods, this study represents the initial endeavor to employ an adversarial patch against multi-modal artificial intelligence systems.

X-ray Detection: is widely used in security screening to identify prohibited items in safety-critical scenarios [177]. Liu et al. [128] pioneered the exploration of physical adversarial attacks in X-ray imaging. They introduced X-Adv, a technique for creating physically plausible adversarial metal objects. When positioned near the targeted prohibited item, these objects enable the item to evade detection. X-Adv exposes vulnerabilities in X-ray detection systems, emphasizing the necessity for enhanced robustness.

VI. DISCUSSION

During the development of this paper, we have observed the diversity and broad scope of physical adversarial attacks. Despite the growth in published works in recent years, there are still gaps to be explored. Here, we discuss the current challenges and opportunities in this field.

A. Current Challenges

1) Existing Domain Gaps: The workflow of physical adversarial attacks (see Fig. 1) reveals a process where attackers first design in the digital space, deploy in the physical space, and ultimately execute attacks in the digital domain. This workflow involves the transformation between the digital and physical domains. Some researchers have recognized this; for instance, Jan et al. [100] designed a D2P network to model the transformation of images from the digital domain to the physical domain. However, the current research has paid limited attention to addressing domain gaps, resulting in the unstable attack performance of many methods. The present practices in physical adversarial attacks face significant challenges in reliability and reproducibility.

2) Uncontrollable Evaluation Settings: Most existing works evaluate their physical adversarial attack methods in the real world using the adversarial mediums they manufacture. The real-world environment is dynamic, and the process of crafting adversarial mediums involves subjective factors, e.g., the material of the clothing, the quality of the printing, etc, all of which are uncontrollable. Future work with reliable and controllable evaluation setups is anticipated.

B. Future Work

1) New Adversarial Medium: In this survey, we define the adversarial medium as the object that carries the adversarial perturbations in the physical world. The adversarial medium plays a significant role in performing a physical attack. From the above discussion, we see that different tasks have different requirements for the adversarial medium, and the suitable adversarial medium can improve the performance of the attack in solving the trilemma, i.e., effectiveness, stealthiness, and stealthiness. The attacks using patches [21], light [110], camera ISP [121], makeup [124], 3D-printed object [125], clothing [89],...
etc., emerged in turn. Recently, the Laser Beam [113] and small lighting bulbs have been used to deceive the DNNs-based models, which inspire novel attack methods and expose the potential risks of DNNs-based applications.

2) Cross-Domain Physical Adversarial Attacks: To address the challenges mentioned in Section VI-A1, attackers need to consider two domain gaps.

Digital-to-physical domain gap: arises in Step 2 (see Fig. 1), specifically denoting the procedure in which attackers manufacture physical perturbations based on digital perturbations. A typical example is the printing loss proposed by Sharif et al. [14], which specifically refers to the inability to accurately and reliably reproduce colors due to the smaller color space of printing devices compared to the RGB color space. They introduced the non-printability score (NPS) to address this issue. Additionally, Jan et al. [100] employed an image-to-image translation network to model the digital-to-physical transformation. Moreover, the adversarial medium, materials, and certain physical constraints all influence this domain gap. A more detailed consideration will facilitate effective cross-domain attacks.

Physical-to-digital domain gap: arises in Step 3 (see Fig. 1). Adversarial perturbations carried by the adversarial medium are captured by cameras in the real world, converted into digital images, and then used to attack DNNs-based models. Throughout this process, there exists a domain gap between the transformations from physical perturbations to digital images. Phan et al. [121] have studied physical adversarial attacks under specific ISP conditions, but they did not explore the performance of attacks across different ISPs. Differentiable ISP simulation or camera simulation is an ideal solution. Combining these simulators will enable the generated adversarial perturbations to maintain attack stability across various hardware imaging devices.

3) Reproducible Evaluation: To address the challenges mentioned in Section VI-A2, researchers should be required to disclose more experimental details. On one hand, this includes production details such as materials, size, and manufacturing processes of the adversarial medium. On the other hand, it involves the environmental conditions of physical experiments, such as lighting, background, and shooting distance. While these details may seem trivial and easily overlooked, they are indispensable for a comprehensive evaluation. Otani et al. [178] have proposed a template for researchers to reference, thereby promoting fairness and reproducibility in evaluation. In the field of physical adversarial attacks, a template for disclosing experimental details is anticipated in the future.

4) Physical World Simulation: The key characteristic of physical adversarial attacks is their real-world feasibility. Precisely simulating the physical environment enhances attacks’ robustness in dynamic settings. Simulation engines like Unreal Engine and Unity offer various conditions for attacks, such as lighting, backgrounds, camera distances, and view angles. Most existing methods use these simulators to assess attack efficacy [23], [60]. However, due to non-differentiability, they cannot be used in end-to-end optimization for adversarial perturbations. Beyond basic operations like rotation, noise addition, affine transformations, and occlusions [13], [22], [93], integrating advanced physical scene simulation methods into the attack pipeline is crucial for considering dynamic settings during adversarial perturbation design.

5) Hexagonal Physical Adversarial Attacks: The proposed evaluation metric, hiPAA, evaluates physical adversarial attacks from six perspectives. However, during the evaluation, we observed that most existing methods lack a comprehensive consideration [68], [94]. They tend to focus on individual perspectives while neglecting others. Some approaches involve tradeoffs between individual dimensions, such as effectiveness and stealthiness [22], [92]. In real-world applications, physical adversarial attacks often need to excel from all six perspectives. Future work with holistic considerations is expected, advancing and evaluating their methods across various perspectives.

6) Physical Adversarial Attacks on New Tasks: As described in this survey, the current mainstream physical adversarial attack methods are oriented to tasks such as person detection [22], [90], traffic sign detection [99], [111], face recognition [14], [161], etc. Although many fields have been covered, some tasks have not yet been explored. For example, Cheng et al. [77] recently proposed an adversarial patch attack against monocular depth estimation (MDE), which is a critical vision task in real-world driving scenarios. It is the first time to propose an attack on MDE. Besides, we consider that domains with the following two characteristics can be explored for physical adversarial attacks: 1) using the DNNs techniques, and 2) applying in the physical world. Such as trajectory prediction [179], pose estimation [180], action recognition [181], etc.

VII. CONCLUSION

Physical adversarial attacks have cast a shadow over the reliability of deep neural networks, raising security concerns. Consequently, extensive research has proposed various methods for real-world attacks across multiple tasks. We have provided an overview of the field of physical adversarial attacks on computer vision tasks, covering classification, detection, re-identification, and some niche tasks, with a focus on the adversarial mediums and a comprehensive evaluation. We first propose a general workflow for launching a physical adversarial attack, underlining the important role of the adversarial medium. Additionally, we have devised a new metric termed hiPAA, systematically quantifying and assessing attack methods from six distinct perspectives. Correspondingly, we present comparative results for existing methods, offering valuable insights for future improvements. Many challenges remain ahead, and we hope that this paper can motivate further discussion in this field and provides important guidance for future research, ultimately advancing the safety and reliability of machine vision systems.

REFERENCES

[1] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, “Deep learning for computer vision: A brief review,” Comput. Intell. Neurosci., vol. 2018, 2018, Art. no. 7068349.
[2] D. W. Otter, J. R. Medina, and J. K. Kalita, “A survey of the usages of deep learning for natural language processing,” IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 2, pp. 604–624, Feb. 2021.
[38] X. Qi, T. Xie, R. Pun, J. Zhu, Y. Yang, and K. Bu, “Towards practical deployment-stage backdoor attack on deep neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2022, pp. 6206–6215.

[39] Z. Cai et al., “Zero-query transfer attacks on context-aware object detectors,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2022, pp. 1502–15034.

[40] W. Sun, M. Tan, and Z. Zhou, “A survey of practical adversarial example attacks,” Cybersecurity, vol. 1, pp. 1–9, 2018.

[41] Z. Liu et al., “Swin transformer: Hierarchical vision transformer using shifted windows,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 10012–10022.

[42] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, “DINO: DETR with improved denoising anchor boxes for end-to-end object detection,” 2022, arXiv:2203.03605.

[43] H. Zhang et al., “DINO: DETR with improved denoising anchor boxes for end-to-end object detection,” 2018, arXiv:1804.02767.

[44] M. Ye, Z. Wang, X. Lan, and P. C. Yuen, “Visible thermal person re-identification via dual-constrained top-ranking,” in Proc. Int. Joint Conf. Artif. Intell., 2018, Art. no. 2.

[45] Z. Wang, M. Ye, F. Yang, X. Bai, and S. S. 000 1, “Cascaded SR-GAN for scale-adaptive low resolution person re-identification,” in Proc. Int. Joint Conf. Artif. Intell., 2018, Art. no. 4.

[46] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, “MagFace: A universal representation for face recognition and quality assessment,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 14225–14234.

[47] T. Gu, B. Dolan-Gavitt, and S. Garg, “BadNets: Identifying vulnerabilities in the machine learning model supply chain,” 2017, arXiv:1708.06733.

[48] Y. Liu et al., “Trojaning attack on neural networks,” in 25th Annu. Netw. Distrib. Syst. Secur. Symp. (NDSS 2018), pp. 1–15, 2018.

[49] E. Wenger, J. Passananti, A. N. Bhagoji, Y. Yao, H. Zheng, and B. Y. Zhao, “Backdoor attacks against deep learning systems in the physical world,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 6206–6215.

[50] Z. Cai et al., “Zero-query transfer attacks on context-aware object detectors,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2022, pp. 1502–15034.

[51] W. Sun, M. Tan, and Z. Zhou, “A survey of practical adversarial example attacks,” Cybersecurity, vol. 1, pp. 1–9, 2018.

[52] Z. Liu et al., “Swin transformer: Hierarchical vision transformer using shifted windows,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 10012–10022.

[53] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, “DINO: DETR with improved denoising anchor boxes for end-to-end object detection,” 2022, arXiv:2203.03605.

[54] M. Ye, Z. Wang, X. Lan, and P. C. Yuen, “Visible thermal person re-identification via dual-constrained top-ranking,” in Proc. Int. Joint Conf. Artif. Intell., 2018, Art. no. 2.

[55] Z. Wang, M. Ye, F. Yang, X. Bai, and S. S. 000 1, “Cascaded SR-GAN for scale-adaptive low resolution person re-identification,” in Proc. Int. Joint Conf. Artif. Intell., 2018, Art. no. 4.

[56] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, “MagFace: A universal representation for face recognition and quality assessment,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 14225–14234.

[57] T. Gu, B. Dolan-Gavitt, and S. Garg, “BadNets: Identifying vulnerabilities in the machine learning model supply chain,” 2017, arXiv:1708.06733.

[58] Y. Liu et al., “Trojaning attack on neural networks,” in 25th Annu. Netw. Distrib. Syst. Secur. Symp. (NDSS 2018), pp. 1–15, 2018.

[59] E. Wenger, J. Passananti, A. N. Bhagoji, Y. Yao, H. Zheng, and B. Y. Zhao, “Backdoor attacks against deep learning systems in the physical world,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 6206–6215.
Y. Zhao, H. Zhu, R. Liang, S. Zhang, and K. Chen, “Seeing isn’t believing: Towards more robust adversarial attack against real world object detectors,” in Proc. ACM SIGSAC Conf. Comput. Secur. Commun., 2019, pp. 1999–2004.

A. Liu et al., “Perceptual-sensitive GAN for generating adversarial patches,” in Proc. AAAI Conf. Artif. Intell., 2019, pp. 1028–1035.

R. R. Wiyatno and A. Xu, “Physical adversarial textures that fool visual object tracking,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 882–883.

Z. Wang, S. Zheng, M. Song, Q. Wang, A. Rahimpour, and H. Qi, “advpattern: Physical-world attacks on deep person re-identification via adversarially transformable patterns,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 8341–8350.

A. Ranjan, J. Janai, A. Geiger, and M. J. Black, “Attacking optical flow,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 2404–2413.

A. Liu, J. Wang, X. Liu, B. Cao, C. Zhang, and H. Yu, “Bias-based universal adversarial patch attack for automatic check-out,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 395–410.

A. Zolfi, M. Kravchik, Y. Elovici, and A. Shabtai, “The translu-cent patch: A physical and universal attack on object detectors,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 15232–15241.

L. Ding et al., “Towards universal physical attacks on single object tracking,” in Proc. AAAI Conf. Artif. Intell., 2021, pp. 1236–1245.

Z. Xiao et al., “Improving transferability of adversarial patches on face recognition with generative models,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 11845–11854.

J. Wang, A. Liu, X. Bai, and X. Liu, “Universal adversarial patch attack for automatic checkout using perceptual and attentional bias,” IEEE Trans. Image Process., vol. 31, pp. 598–611, 2021.

P. A. Sava, J.-P. Schulze, P. Sperl, and K. Böttiger, “Assessing the impact of transformations on physical adversarial attacks,” in Proc. 15th ACM Workshop Artif. Intell. Secur., 2022, pp. 79–90.

S. Casper, M. Nadeau, D. Hadfield-Menell, and G. Kreiman, “Robust feature-level adversaries are interpretability tools,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2022, pp. 33093–33106.

B. G. Doan, M. Xue, S. Ma, E. Abbasnejad, and D. C. Ranasinghe, “TnT attacks! Universal naturalistic adversarial patches against deep neural network systems,” IEEE Trans. Inf. Forensics Secur., vol. 17, pp. 3816–3830, 2022.

Z. Cheng et al., “Physical attack on monocular depth estimation with optimal adversarial patches,” in Proc. Eur. Conf. Comput. Vis., 2022, pp. 514–532.

F. Nesti, G. Rossolino, S. Nair, A. Bontdi, and G. Buttazzo, “Evaluating the robustness of semantic segmentation for autonomous driving against real-world adversarial patch attacks,” in Proc. IEEE Winter Conf. Appl. Comput. Vis., 2022, pp. 2280–2289.

S. Liu et al., “Harnessing perceptual adversarial patches for crowd counting,” in Proc. ACM SIGSAC Conf., New York, NY, USA: Association for Computing Machinery, 2022, pp. 2055–2060.

Z. Chen, B. Li, S. Wu, J. Xu, S. Ding, and W. Zhang, “Shape matters: deformable patch attack,” in Proc. Eur. Conf. Comput. Vis., 2022, pp. 529–548.

X. Wei, J. Yu, and Y. Huang, “Infrared adversarial patches with learnable shapes and locations,” in Proc. ACM SIGSAC Conf. Comput. Secur., 2020, pp. 679–696.

X. Wei, J. Yu, and Y. Huang, “Infrared adversarial patches with learnable shapes and locations in the physical world,” Int. J. Comput. Vis., vol. 132, pp. 1928–1944, 2023.

W. Zhu, X. Ji, Y. Cheng, S. Zhang, and W. Xu, “Tpatch: A triggered universal adversarial patch,” 2023, arXiv:2301.00448.

K. Xu et al., “Adversarial t-shirt! evading person detectors in a physical world,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 665–681.

L. Huang et al., “Universal physical camouflage attacks on object de-tectors,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 720–729.

Z. Wu, S.-N. Lim, L. S. Davis, and T. Goldstein, “Making an invisibility cloak: Real world adversarial attacks on object detectors,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 1–17.

Y.-C.-T. Hu, B.-H. Kung, D. S. Tan, J.-C. Chen, K.-L. Hua, and W.-H. Cheng, “Naturalistic physical adversarial patch for object detectors,” in Proc. IEEE Int. Conf. Comput. Vis., 2021, pp. 7848–7857.

Z. Hu, S. Huang, X. Zhu, F. Sun, B. Zhang, and X. Hu, “Adversarial texture for fooling person detectors in the physical world,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2022, pp. 13307–13316.

X. Zhu, Z. Hu, S. Huang, J. Li, and X. Hu, “Infrared invisible clothing: Hiding from infrared detectors at multiple angles in real world,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2022, pp. 13317–13326.

J. Sun, W. Yao, T. Jiang, D. Wang, and X. Chen, “Differential evolution based dual adversarial camouflage: Fooling human eyes and object detectors,” Neural Netw., vol. 163, pp. 256–271, 2023.

Z. Hu, W. Chu, X. Zhu, H. Zhang, B. Zhang, and X. Hu, “Physically realizable natural-looking clothing textures evade person detectors via 3D modeling,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2023, pp. 16975–16984.

A. Kurakin, I. J. Goodfellow, and S. Bengio, “Adversarial examples in the physical world,” 2017. [Online]. Available: https://openreview.net/forum?id=HJGU3Rodl

S.-T. Chen, C. Cornelius, J. Martin, and D. H. P. Chau, “Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector,” in Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discov. Databases, 2018, pp. 52–68.

D. Song et al., “Physical adversarial examples for object detec-tors,” in Proc. 12th USENIX Workshop Offensive Technol., 2018, pp. 1–10.

S. T. Jan, J. Messou, Y.-C. Lin, J.-B. Huang, and G. Wang, “Connecting the digital and physical world: Improving the robustness of adversarial attacks,” in Proc. AAAI Conf. Artif. Intell., 2019, pp. 962–969.

Q. Guo et al., “Watch out! Motion is blurring the vision of your deep neural networks,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2020, pp. 975–985.

Z. Kong, J. Guo, A. Li, and C. Liu, “PhysiGAN: Generating physiological-world-resilient adversarial examples for autonomous driving,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 14254–14263.

R. Duan, X. Ma, Y. Wang, J. Bailey, A. K. Qin, and Y. Yang, “Adversarial camouflage: Hiding physical-world attacks with natural styles,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 1000–1008.

K. Yang, T. Tsai, H. Yu, T.-Y. Ho, and Y. Jin, “Beyond digital domain: Fooling deep learning based recognition system in physical world,” in Proc. AAAI Conf. Artif. Intell., 2020, pp. 1098–1095.
[105] W. Feng, B. Wu, T. Zhang, Y. Zhang, and Y. Zhang, “Meta-attack: Class-agnostic and model-agnostic physical adversarial attack,” in Proc. IEEE Int. Conf. Comput. Vis., 2021, pp. 7787–7796.

[106] Y. Dong, S. Ruan, H. Su, C. Kang, X. Wei, and J. Zhu, “View-forall robustvul: The robustness of visual recognition to adversarial viewpoints,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2022, pp. 36789–36803.

[107] W. Feng, N. Xu, T. Zhang, B. Wu, and Y. Zhang, “Robust and generalized physical adversarial attacks via meta-GAN,” IEEE Trans. Inf. Forensics Security, vol. 19, pp. 1112–1225, 2023.

[108] N. Nichols and R. Jasper, “Projecting trouble: Light based adversarial attacks on deep learning classifiers,” in AAAI Full Symp.: ALEC, pp. 44–49, 2018.

[109] Y. Man, M. Li, and R. Gerdets, “POSTER: Perceived adversarial examples,” in Proc. IEEE Secur. Symp. Privacy, 2019, pp. 1–2.

[110] D.-L. Nguyen, S. S. Arora, Y. Wu, and H. Yang, “Adversarial light projection attacks on face recognition systems: A feasibility study,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, 2020, pp. 814–815.

[111] G. Llovisotto, H. Turner, I. Sluganovic, M. Strohmeier, and I. Martinovic, “SLAP: Improving physical adversarial examples with short-lived adversarial perturbations,” in Proc. SENIX Secur. Symp., 2021, pp. 1865–1882.

[112] A. Guanasaarudam, A. M. Sherman, and S. H. Chan, “Optical adversarial attack,” in Proc. IEEE Int. Conf. Comput. Vis., 2021, pp. 92–101.

[113] R. Duan et al., “Adversarial laser beam: Effective physical-world attack to DNNs in a blink,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 16062–16071.

[114] Y. Zhong, X. Liu, D. Zhai, J. Jiang, and X. Ji, “Shadows can be dangerous: Stealthy and effective physical-world adversarial attack by natural phenomenon,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2022, pp. 15345–15354.

[115] B. Huang and H. Ling, “SPAA: Stealthy projector-based adversarial attacks on deep image classifiers,” in Proc. IEEE Conf. Virtual Reality 3D User Interfaces, 2022, pp. 534–542.

[116] H. Wen, S. Chang, and L. Zhou, “Light projection-based physical-world vanishing attack against car detection,” in Proc. IEEE Int. Conf. Acoust. Speech Signal Process., 2023, pp. 1–5.

[117] C. Hu, Y. Wang, K. Tilivilaadi, and W. Li, “Adversarial laser spot: Robust and covert physical-world attack to DNNs,” in Proc. Asian Conf. Mach. Learn., 2023, pp. 483–498.

[118] Y. Li, Y. Li, X. Dai, S. Guo, and B. Xiao, “Physical-world optical adversarial attacks on 3D face recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2023, pp. 24699–24708.

[119] D. Wang, W. Yao, T. Jiang, C. Li, and X. Chen, “RFLA: A stealthy and effective physical-world adversarial attack,” in Proc. IEEE Int. Conf. Comput. Vis., 2023, pp. 4412–4423.

[120] S. Li et al., “Towards benchmarking and assessing visual naturalness of physical world adversarial attacks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2023, pp. 12324–12333.

[121] I. Goodfellow et al., “Generative adversarial networks,” Commun. ACM, vol. 63, no. 11, pp. 139–144, 2020.

[122] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2009, pp. 248–255.

[123] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 2818–2826.

[124] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 2223–2232.

[125] J. Wang et al., “Defensive patches for robust recognition in the physical world,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2022, pp. 2456–2465.

[126] T. Wu, X. Ning, W. Li, R. Huang, H. Yang, and Y. Wang, “Physical adversarial attack on vehicle detector in the CARLA simulator,” 2020, arXiv: 2007.16118.

[127] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, “CARLA: An open urban driving simulator,” in Proc. Conf. Robot Learn., 2017, pp. 1–16.

[128] H. Kato, Y. Ushiku, and T. Harada, “Neural 3D mesh renderer,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2023, pp. 3907–3916.

[129] J. Thies, M. Zollhöfer, and M. Nießner, “Deferred neural rendering: Image synthesis using neural textures,” ACM Trans. Graph., vol. 38, no. 4, pp. 1–12, 2019.

[130] K. Rematas and V. Ferrari, “Neural voxel renderer: Learning an accurate and controllable rendering tool,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 5417–5427.

[131] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 779–788.

[132] J. Redmon and A. Farhadi, “YOLO9000: better, faster, stronger,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 7263–7271.

[133] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2005, pp. 886–893.

[134] Z. Wang, Q. She, and T. E. Ward, “Generative adversarial networks in computer vision: A survey and taxonomy,” ACM Comput. Surv., vol. 54, no. 2, pp. 1–38, 2021.

[135] A. Brock, J. Donahue, and K. Simonyan, “Large scale gan training for high fidelity natural image synthesis,” 2018, arXiv: 1809.11096.

[136] T. Karras et al., “Alias-free generative adversarial networks,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2021, pp. 852–863.

[137] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, “2D human pose estimation: New benchmark and state of the art analysis,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2014, pp. 3686–3693.

[138] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 3431–3440.

[139] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, “Solving for simplicity: The all convolutional net,” 2014, arXiv:1412.6086.

[140] A. Hatcher, Algebraic Topology, Cambridge, UK: Cambridge Univ. Press, 2002.

[141] M. Martin et al., “DriveKact: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 2801–2810.

[142] M. Ye, C. Chen, J. Shen, and L. Shao, “Dynamic tri-level relation mining with attentive graph for visible infrared re-identification,” IEEE Trans. Inf. Forensics Security, vol. 17, pp. 386–398, 2021.
Hanxun Yu received the BS degree from the school of Computer Science, Wuhan University, China. He is currently working toward the master’s degree with Zhejiang University, China. His research interests include designing physical adversarial attacks to enhance the reliability and trustworthiness of machine learning in real-world applications.

Hui Wei is currently working toward the PhD degree with the School of Computer Science, Wuhan University, China. His research interests include designing physical adversarial attacks to enhance the reliability and trustworthiness of machine learning in real-world applications.

Hao Tang received the master's degree from Peking University, China, and the PhD degree from the University of Trento, Italy. He is an assistant professor with Peking University, China. Previously, he held postdoctoral positions with both CMU, USA, and ETH Zürich, Switzerland. He was a visiting Ph.D. student with the University of Oxford, U.K., and an intern with BAI, UAE. His research interests include AI/ML, machine learning, computer vision, embodied agents, and their applications to scientific domains.

Xuemei Jia received the BS and MS degrees from the Wuhan University of Technology, Wuhan, China, in 2019 and 2022, respectively. She is currently working toward the PhD degree with Wuhan University, Wuhan. Her research interests include multimedia content analysis and retrieval, and adversarial attack.

Zhixiang Wang received the MPhil degree from National Taiwan University, in 2020. He is currently working toward the PhD degree with the University of Tokyo, Japan. He was a research intern with Snap Research, USA, in 2022, a research associate with Nanyang Technological University, Singapore, in 2020, and a research intern with the National Institute of Informatics, Japan, in 2018. His research interests include center around computational imaging and computational photography.
Zhubo Li received the BS degree from the School of Cyber Science and Engineering, Wuhan University, China. He is currently a postgraduate student with the School of Cyber Science and Engineering, Wuhan University. His research interests include adversarial attacks and robustness of computer vision models.

Shin’ichi Satoh received the BE degree in electronics engineering, the ME and PhD degrees in information engineering from the University of Tokyo, in 1987, 1989, and 1992, respectively. He joined National Center for Science Information Systems (NACSIS), Tokyo, in 1992. He is a full professor with National Institute of Informatics (NII), Tokyo, since 2004. He was a visiting scientist with the Robotics Institute, Carnegie Mellon University, from 1995 to 1997. His research interests include image processing, video content analysis and multimedia database. Currently he is leading the video processing project at NII addressing video analysis, indexing, retrieval, and mining for broadcasted video archives.

Luc Van Gool received the degree in electromechanical engineering from the Katholieke Universiteit Leuven, in 1981. Currently, he is a professor with the Katholieke Universiteit Leuven in Belgium and the ETH Zurich in Switzerland. He leads computer vision research with both places, where he also teaches computer vision. His main interests include 3D reconstruction and modeling, object recognition, tracking, and gesture analysis, and the combination of those.

Zheng Wang (Senior Member, IEEE) received the BS and MS degrees from Wuhan University, Wuhan, China, in 2006 and 2008, respectively, and the PhD degree from the National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, in 2017. He was a JSPS fellowship researcher with the Shin’ichi Satoh’s Laboratory, National Institute of Informatics, Japan, and a project assistant professor with the University of Tokyo, Japan. He is currently a professor with the National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University. His research interests include person re-identification and instance search. He received the Best Paper Award from the 15th Pacific-Rim Conference on Multimedia (PCM 2014) and the 2017 ACM Wuhan Doctoral Dissertation Award.