Adversarial Camera Patch: An Effective and Robust Physical-World Attack on Object Detectors

Kalibinuer Tiliwalidi[1], Bei Hui[2], Chengyin Hu[1] and Jingjing Ge[3]
[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan, China
[2]School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan, China
[3]Unit 31153

202011081727@std.uestc.edu.cn
bhui@uestc.edu.cn
cyhuuestc@gmail.com
G_jingjing2023@hotmail.com

Abstract: Physical adversarial attacks present a novel and growing challenge in cybersecurity, especially for systems reliant on physical inputs for Deep Neural Networks (DNNs), such as those found in Internet of Things (IoT) devices. They are vulnerable to physical adversarial attacks where real-world objects or environments are manipulated to mislead DNNs, thereby threatening the operational integrity and security of IoT devices. The camera-based attacks are one of the most practical adversarial attacks, which are easy to implement and more robust than all the other attack methods, and pose a big threat to the security of IoT. This paper proposes Adversarial Camera Patch (ADCP), a novel approach that employs a single-camera patch to launch robust physical adversarial attacks against object detectors. ADCP optimizes the physical parameters of the camera patch using Particle Swarm Optimization (PSO) to identify the most adversarial configuration. The optimized camera patch is then attached to the lens to generate stealthy and robust adversarial samples physically. The effectiveness of the proposed approach is validated through ablation experiments in a digital environment, with experimental results demonstrating its effectiveness even under worst-case scenarios (minimal width, maximum transparency). Notably, ADCP exhibits higher robustness in both digital and physical domains compared to the baseline. Given the simplicity, robustness, and stealthiness of ADCP, we advocate for attention towards the ADCP framework as it offers a means to achieve streamlined, robust, and stealthy physical attacks. Our adversarial attacks pose new challenges and requirements for cybersecurity.

Keywords: Deep neural network, Camera-based physical attack, Object detector, Effectiveness, Robustness

1. Introduction

Deep Neural Networks (DNNs) are among the most important technologies that assist in accomplishing the goals of computer vision. They excel in tasks such as image classification, object detection, and segmentation, as they can automatically learn and extract meaningful features from images. Despite the superior performance of DNNs in real-world applications, there are serious shortcomings in their robustness in adversarial scenarios. Studies have found that DNNs are easily deceived by adversarial samples maliciously constructed by attackers, resulting in inaccuracies in model prediction (Szegedy et al 2013; Goodfellow et al 2014; Guesmi et al 2023).

(a) AdvCS (b) TTP (c) ADCP (Ours)

Figure 1: Demonstration of ADCP and other camera-based attacks

At present, a significant portion of physical attacks (Hu et al 2022; Wang et al 2021; Suryanto et al 2022) utilizes adversarial patches as perturbations to execute physical attacks against advanced object detectors. Adversarial patches commonly cover a substantial fraction of the target object's area, bolstering the robustness of physical

attacks at the expense of reduced stealthiness. While patch-based physical attacks maintain the semantic integrity of the target object, the conspicuousness of the perturbation remains a challenge. To address this issue, some studies have introduced light-based physical attacks (e.g., lasers (Hu et al 2023), projectors (Gnanasambandam et al 2021), shadow(Zhong et al 2022)). These leverage the transitory nature of illumination to instantaneously project an optimized beam onto the target object's surface, facilitating immediate physical attacks. Differing from patch-based counterparts, light-based methods can manipulate the light source's on-off state, projecting the beam during attacks and extinguishing the source when dormant. This results in superior stealthiness for light-based methods, as the physical perturbation isn't persistently affixed to the target object's surface. While light-based physical attacks enhance stealthiness, they often come at the cost of reduced robustness. Similarly, both light-based and patch-based physical attacks share a commonality: they entail modifications to the target object. To tackle this limitation, certain studies have proposed camera-based physical attacks, such as Adversarial camera stickers (AdvCS) proposed by Li et al (2019) and Translucent Patch (TTP) proposed by Zolfi et al (2021), these attacks are all based on white-box attacks that assume complete knowledge of the target model's architecture and parameters, which is often unrealistic in real-world scenarios. In this study, we investigate black-box camera-based adversarial attacks in a more extreme yet practical physical setting. The focus is on generating robust adversarial examples to attack advanced object detectors while maintaining superior robustness, stealthiness, and affordability for real-world applications. A camera patch perturbation is crafted to fool the model for all instances of a specific object class, while maintaining the detection of untargeted objects. Figure 1 vividly illustrates the deployment contrasts between the proposed approach and the baseline strategy. It is clearly that the model correctly recognizes other objects. The contributions of the study can be outlined as follows:

- Introduce a novel camera-based physical attack (ADCP) based on a black-box for robustly generating adversarial samples to execute attacks on advanced object detectors. ADCP utilized PSO to generate robust adversarial perturbations which assist in surpassing baselines in terms of robustness, stealthiness, and deployment simplicity.
- Provide a model with inexpensive and reachable attack material with effortless implementation, where the expected budget is not exceeding $7, to render readily applicable in real-world scenarios.
- Conducting a meticulous comparison of existing physical attack techniques to determine the distinctive benefits of the approach in relation to its equivalents. The comparisons confirm that ADCP is valuable and outperforms recent camera-based physical attacks where ADCP preserves its stealthy nature.
- The study extends to a comprehensive investigation of the proposed ADCP method, including ablation experiments that confirm the method's consistent performance across diverse color settings. These analyses contribute to a comprehensive understanding of the method's performance and its practical viability.

2. Related Works

Patch-based attacks involve the strategic use of meticulously crafted patches affixed to the target object's surface to deceive advanced DNNs (Chen et al 2019). Recent research has focused on achieving a balance between the patch's stealthiness and its impact on attack efficacy. Thys et al. (2019) used total variation loss to generate smoother adversarial patches that compromise pedestrian target detection systems. Building upon this, Wu et al. (2020) printed optimized adversarial patches onto clothing to diminish their conspicuity to human observers, showcasing effectiveness across white-box and black-box settings. Similarly, Tan et al. (2021) introduced a method involving cartoon-like patches that deceive pedestrian detectors, employing a two-stage training approach to generate natural and rational adversarial patches.

On the other hand, camera-based attacks involve applying an adversarial patch onto a camera lens, capturing a physical sample of the target object and subsequently deceiving DNNs. Distinguished by its nonintrusive character, camera-based attacks confer a stealth advantage compared to patch-based, which necessitate target object modifications. Li et al. (2019) introduced the AdvCS camera-based attack, which utilized lens modification to affix small non-transparent patches, effectively deceiving classifiers. It is an uncommon scenario in practical applications. Addressing these limitations, Zolfi et al. (2021) proposed TTP, a refined camera-based physical attack using translucent patches for improved sample concealment, achieving a 42.27% attack success rate against advanced detectors. Yet, the subtlety of TTP came at the expense of attack robustness. In essence, prevailing camera-based methods confer spatial invisibility without affecting target objects. Nevertheless, AdvCS and TTP's reliance on multiple small patches for lens coverage culminated in operational complexity and

substantial experimental errors, limiting attack performance and feasibility. Contrasting these approaches, ADCP circumvents these challenges through the application of a solitary translucent patch as perturbation.

3. Method

In this study, similar to the explanation provided by Zolfi et al (2021), it is suggested that the attacker possesses direct access to the camera lens, enabling the deployment of meticulously optimized perturbations upon its surface.

3.1 Problem Definition

Considering dataset D comprising clean samples, it is divided into two sets, X and Y, representing the collection of clean samples and the collection of ground truth labels, correspondingly. With f representing the pre-trained model of the object detector, for every X belonging to the dataset D, the object detector $f: X \rightarrow Y$ adeptly prognosticates the label y for the clean sample. Within y, there are three important elements: V_{pos}, denoting the positional intel of the bounding box; V_{obj} , encapsulating the confidence level of the target object; and V_{cls}, conveying the category of the prognosticated entity:

$$y = [V_{pos}, V_{obj}, V_{cls}] = f(X) \quad (1)$$

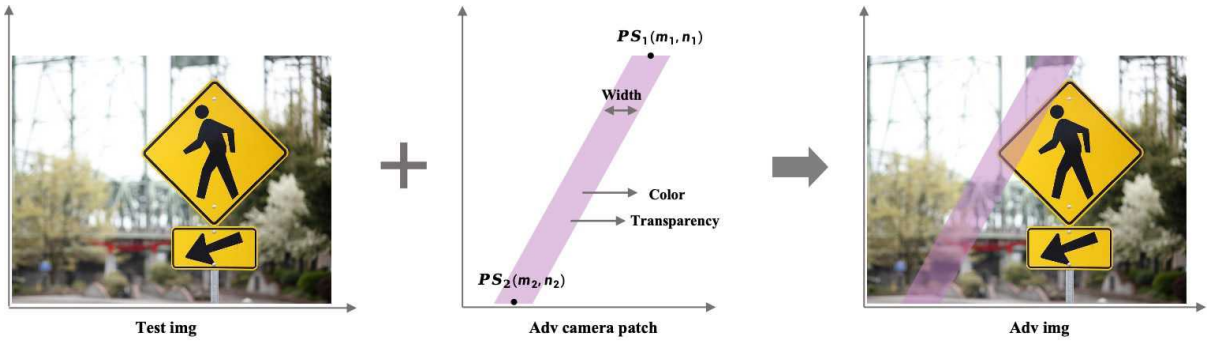

Figure 2: Camera patch modeling

3.2 Camera Patch Modeling

The approach uses a translucent patch affixed to the camera lens to conduct a black-box physical attack. The camera patch is depicted conceptually in Figure 2 and is characterized by four distinct physical parameters: position (PS), color (C), width (W), and transparency (TS).

Position PS: The position parameter is denoted as PS and determined by the coordinates of the two endpoints of a line, pinpointing the precise location of the camera patch within the image. In the experimental setup, we keep n_1 and n_2 as constant values to vary the camera patch's position solely in the horizontal plane of the image.

Color C: In the investigation, color is utilized to encapsulate the visual attributes of the camera patch, with its representation denoted as $\mathrm{C}(\mathrm{r}, \mathrm{g}, \mathrm{b})$, where r, g, and b signify the red, green, and blue channel values of the color, respectively. In digital attack experiments, the choice of color is unconstrained. However, for physical attacks, we exclusively use pink, blue, and green camera patches due to practical limitations. Subsequent experiments demonstrate a minimal correlation between color and attack success rate.

Width W: To effectively encapsulate the scale of the camera patch in the horizontal direction, the parameter width (W) is introduced. As depicted in Figure 2, the width W is designed to measure the horizontal extent of the camera patch. In prescribing the value of this parameter, it is aligned with the width of the image to ensure its adaptability across diverse image dimensions. Specifically, the width W *is defined* within a range spanning from 0.1 to 0.9, with intervals set at 0.2. The chosen value of W signifies the proportion of the camera patch's width to that of the overall image.

Transparency TS: The transparency parameter ranges from 0.1 to 0.9, encapsulating the perceptual visibility of the camera patch. A lower value implies heightened transparency, rendering the camera patch less discernible, while a higher value accentuates its presence, increasing the likelihood of capturing the observer's attention.

3.3 Camera Patch Attack

The vector $\theta = \{PS, C, W, TS\}$ is adopted to succinctly represent the camera patch generated in this approach. The size of θ adheres to the pre-set boundaries defined by the vectors θ_{min} and θ_{max}, where θ_{min} and θ_{max} can be adjusted. Hence, the process of crafting adversarial examples within a digital context by employing predetermined physical parameters can be elucidated as follows:

$$X_{adv} = S(X, \theta) \quad \theta \in (\theta_{min}, \theta_{max}) \tag{2}$$

Where, *S* represents a simple linear fusion of the generated simulation camera patch and the clean sample to obtain the digital adversarial sample X_{adv}.

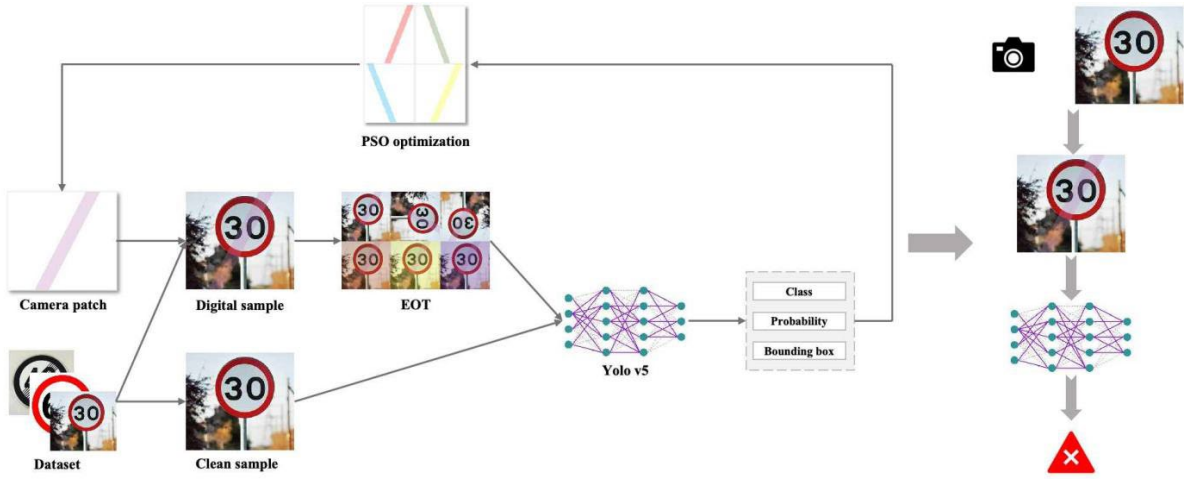

Figure 3: Overview of ADCP attack. The left side shows the optimization process in a digital environment, and the right side indicates that the method generates physical samples in a physical environment. After the robust digital adversarial samples are optimized in the digital environment, perturbations will be deployed in the physical world to generate physical samples

Figure 3 illustrates the attack strategy employed in the study. In the digital field, camera patches with an expansive spectrum of colors are simulated for the execution of the attack. Meanwhile, in the physical scenarios, fixed-color camera patches (pink, green, blue) are opted for the attack. To bridge the experimental gap between simulated and physical samples, the technique of Expectation Over Transformation (EOT) (Athalye et al., 2018) is introduced. Within the context of EOT, a transformation T encompasses the distribution of all possible transformations, serving as the conduit for simulating the transition from the digital domain to physical scenarios. For 2D adversarial examples, the transformation set T includes rotation, variance, dimming, Gaussian noise, image translation, and other relevant alterations. By integrating these transformations, the optimization of adversarial examples that remain adversarial under transformation T is achieved, thus bolstering the efficacy of physical attacks and mitigating the influence of experimental deviations. Through the incorporation of EOT, the impact of experimental discrepancies on physical attacks is successfully mitigated. Consequently, the ultimate representation of the physical sample is encapsulated as follows:

$$X_{adv} = EOT_{t\sim T}(t(S(X, \theta), \theta)) \quad \theta \in (\theta_{min}, \theta_{max}) \tag{3}$$

Within the context of this research, the primary objective centers around the simulation and optimization of the physical parameter θ, a precursor to generating the most potent adversarial camera patch. The pivotal criterion is to produce an adversarial sample X_{adv}, engendered by the physical parameter, that yields successful deception of the object detector. Consequently, the object detector would either be unable to correctly identify the object or be prompted to misidentify it. A distinguishing aspect of this study is its departure from preceding camera-based endeavors (Li et al 2019; Zolfi et al 2021), where the aim is to enhance methodological realism. Specifically, this work operates within the framework of a black-box attack scenario, signifying that access to intricate details such as the target model's network structure is precluded. Instead, only the target class output (V_{cls}) and its associated confidence score (V_{obj}) are accessible. This context precipitates the formulation of a novel approach, wherein V_{obj} is harnessed as the adversarial loss. In light of this, the objective is framed as the optimization of the physical parameters characterizing the camera patch, with the intent of minimizing the adversarial loss. This intricate process seeks to generate adversarial samples that successfully fool the object detector. The crux of this optimization endeavor can be succinctly encapsulated as follows:

$$argmin_{\theta} EOT_{t\sim T}(V_{obj} \leftarrow t(f(X_{adv}), \theta)) \quad (4)$$

To achieve global optimization, the PSO algorithm (Kennedy and Eberhart, 1995), inspired by bird predation behaviors and aimed for rapid convergence and optimal solution discovery, is employed. In the PSO algorithm, each optimization problem's solution is metaphorically represented as a "particle" within a search space. These particles possess distinct fitness values determined by the optimization function and velocity vectors that dictate their navigational trajectories. All particles collectively explore the solution space, following the trajectory of the best particle, known as the "best global particle."

The PSO algorithm begins with the initialization of a set of random particles, representing stochastic solutions. Through iterative cycles, the algorithm navigates the solution space to uncover the best solution. In each iteration, every particle updates its position based on two key values: its individual best solution (denoted as "individual extrema") and the best solution found by all particles (referred to as "global best"). These updates are performed using a formula designed to optimize the particle's movement through the solution space, aligning with the iterative dynamics of the PSO algorithm. The velocity and position of the particles are updated according to the following formula:

$$v_i^{j+1} = \omega v_i^j + c_1 r_1 (P_{i,best}^j - \theta_i^j) + c_2 r_2 (G_{best}^j - \theta_i^j) \quad (5)$$

$$\theta_i^{j+1} = \theta_i^j + v_i^{j+1} \quad (6)$$

where $i \in (1, k)$, k denotes the number of populations. j denotes the current iteration number and ω, c_1, r_1, c_2, r_2 denote the hyperparameters of the PSO algorithm. v is the velocity of the particle and θ is the position of the particle (i.e., the current solution).

Algorithm 1: Pseudocode of ADCP

Input: Input $\boldsymbol{X}$, object detector $\boldsymbol{f}$, Ground truth label $\boldsymbol{Y}$, Max step *Step*, $\boldsymbol{\omega}, \boldsymbol{c_1}, \boldsymbol{r_1}, \boldsymbol{c_2}, \boldsymbol{r_2}$;

Output: A vector of parameters $\boldsymbol{\theta}$;

For each particle $\boldsymbol{i}$:

    Initialize position $\boldsymbol{\theta_i}$ randomly;

    Initialize velocity $\boldsymbol{v_i}$ randomly;

End For

For $\boldsymbol{j} \leftarrow 0$ to *Step*:

  For each particle $\boldsymbol{i}$:

    $\boldsymbol{X_i^j = S(X, \theta_i^j(PS, C, W, TS))}$;

    $[\boldsymbol{V_{pos}, V_{obj}, V_{cls}}] \leftarrow \boldsymbol{f}(\boldsymbol{X_i^j})$;

    Obtain the individual optimal value $\boldsymbol{P_{i,best}^j}$;

    Obtain the global optimal value $\boldsymbol{G_{best}^j}$;

    If $\boldsymbol{f(X_i^j) \rightarrow \emptyset \ or \ V_{cls} \neq Y}$:

      Output $\boldsymbol{\theta = \theta_i^j(PS, C, W, TS)}$;

      Exit ();

    End If

  End For

  For each particle $\boldsymbol{i}$:

    $\boldsymbol{v_i^{j+1} = \omega v_i^j + c_1 r_1 (P_{i,best}^j - \theta_i^j) + c_2 r_2 (G_{best}^j - \theta_i^j)}$;

    $\boldsymbol{\theta_i^{j+1} = \theta_i^j + v_i^{j+1}}$;

  End For

End For

Algorithm 1 outlines the process of the proposed ADCP employing PSO for optimization. ADCP takes several inputs: a clean sample X, the target detector f, ground truth label Y, maximum number of iterations Step, and PSO hyperparameters ω, c_1, r_1, c_2, and r_2, which can be determined by the attacker. The pseudocode details are elucidated in Algorithm 1. Initially, it initializes the initial velocity and position of each particle within the swarm. Subsequently, in each iteration, the camera patch represented by each particle is combined with the clean sample to generate an adversarial sample. The confidence score of the particle, which serves as its fitness value, is then obtained. It's worth noting that a lower confidence score indicates a higher fitness level for that particle. For each particle, if its corresponding adversarial sample induces the model to misidentify or fail to recognize the target, the position information of the particle is recorded as the required physical parameters for the camera patch. Ultimately, in every iteration, the historical optimal solution $\mathrm{P_{best}}$ for each particle and the historical optimal solution $\mathrm{G_{best}}$ for the entire swarm are determined. These two optimal solutions are employed to update the velocity and position information of each particle. The program ultimately outputs the physical parameter θ of the camera patch, which is utilized for subsequent physical attacks.

4. Evaluation

4.1 Experimental Setting

Dataset: In this study, the TT100K dataset (Zhu et al 2016) is chosen as the fundamental source for both model training and attack experiments. This detaset encompasses over 30,000 traffic sign instances extracted from approximately 100,000 images, characterized by a resolution of 2048x2048 pixels and a diverse array of lighting and meteorological conditions. The original dataset is carefully curated to retain only those traffic signs containing no fewer than 100 samples. Consequently, a new dataset emerges, denominated as TT100K-CameraPatch (TT100K-CP), encompassing a total of 10,427 images. Throughout the experiments, the TT100K-CP dataset is divided into training, testing, and validation sets, comprising 7,568, 1,889, and 970 images, respectively. Moreover, for the digital attack experiment, the segregated validation set is designated as the attack dataset, enabling a comprehensive evaluation and validation of the proposed method.

Object detector: In this study, YOLOv5 (Jocher 2020) was selected as the object detector for training the traffic sign detection model for Chinese road signs due to its speed, efficiency, and wide acceptance in the field. To expedite convergence, pre-trained weights from YOLOv5 were used and the model was fine-tuned with the TT100K-CP dataset. This approach accelerated the training process and facilitated model adaptation to the specific dataset, resulting in significant results on the test set, achieving an impressive average accuracy of up to 80%.

Experimental devices: In physical experiments, the experimental devices used included a support frame, camera patches, and a Redmi K40. The support frame was securely installed on the phone to ensure experiment stability and repeatability, enabling the conduct of physical attack experiments in a real-world environment to verify the effectiveness of the proposed approach in practical scenarios. After verification, it was found that the proposed method maintained robustness and feasibility on different camera devices, strengthening the practical application value of the study and making the method potentially applicable to a variety of camera devices.

Evaluation criteria: We embrace the Attack Success Rate (ASR) as the pivotal metric to gauge the efficacy of our proposed methodology. The ASR is defined as follows:

$$\mathrm{ASR(X)} = 1 - \frac{1}{\mathrm{N}}\sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{F(label_i)}$$

$$\mathrm{F(label_i)} = \begin{cases} 1 & \mathrm{label_i} \in \mathrm{L_{pre}} \\ 0 & \text{otherwise} \end{cases}$$

where N represents the number of true positive samples detected by the target detector f in the data set D in the case of no attack, and $\mathrm{L_{pre}}$ represents the set of all labels detected in the case of attack. A higher ASR indicates a more effective attack.

Other details: The hyperparameters of PSO as follows: $\omega = 0.9,\ \mathrm{c_1} = 1.6,\ \mathrm{r_1} = 1,\ \mathrm{c_2} = 2.0,\ \mathrm{r_2} = 1$. For all attack experiments, we execute on a single NVIDIA GeForce RTX 2080 Ti GPU.

4.2 Evaluation of Effectiveness

Digital test: In this study, a series of digital experiments is conducted to validate the proposed approach in a simulated environment. Meticulous ablation experiments are carried out to evaluate the performance of the

method in a digital context. These ablation experiments served two main purposes: determining the attack configuration for subsequent physical experiments and analyzing the impact of ADCP under different configurations. Two pivotal factors influencing the effectiveness of the method are identified, namely, the width and transparency of the camera patch. It is observed that excessively small camera patches may fail to yield robust adversarial effects, while overly wide patches could compromise the attack's subtlety. Similarly, highly transparent camera patches may not ensure robust adversarial effects, while those with low transparency might undermine stealth. Thus, the goal is to find a delicate balance between optimizing attack effectiveness and maintaining stealth. The width parameter of the camera patch spans from 0.1 to 0.9 with a step size of 0.2, and the transparency parameter traversed the range of 0.1 to 0.9 with an increment of 0.1. Ablation experiments are conducted across various combinations of width and transparency values to unravel the intricate interplay between these parameters and their impact on attack effectiveness.

Table 1: Results of ablation experiments

TS	W=0.1		W=0.3		W=0.5		W=0.7		W=0.9	
	ASR(%)	Query	ASR(%)	Query	ASR(%)	Query	ASR(%)	Query	ASR(%)	Query
0.1	40.07	342.86	51.47	279.15	56.62	254.54	55.88	252.63	55.88	238.96
0.2	74.63	180.78	82.35	119.05	86.03	95.29	88.24	82.90	87.13	84.38
0.3	93.75	82.39	96.69	41.27	97.79	27.61	98.53	23.08	98.90	15.62
0.4	96.69	46.17	99.63	15.72	100.00	8.55	100.00	4.25	100.00	4.15
0.5	98.53	35.68	99.26	13.68	100.00	4.72	100.00	2.93	100.00	2.25
0.6	98.53	29.50	100.00	10.78	100.00	3.50	100.00	2.38	100.00	1.64
0.7	98.53	25.11	99.63	10.04	100.00	3.02	100.00	1.88	100.00	1.36
0.8	98.53	23.85	99.26	9.44	100.00	2.79	100.00	1.62	100.00	1.28
0.9	98.16	25.06	99.63	6.81	100.00	2.15	100.00	1.67	100.00	1.23

Figure 4: Digital samples

Table 1 presents a comprehensive overview of the outcomes derived from the digital ablation experiments. A thorough scrutiny of these experimental findings allows the formulation of the following four pivotal conclusions: Firstly, it is evident that ADCP consistently demonstrates proficient digital attack outcomes across a spectrum of experimental configurations. Remarkably, even in the most challenging scenario (W=0.1, TS=0.1), ADCP still achieves a commendable attack success rate of 40.07%. Secondly, there exists a discernible correlation between the augmentation of both the width and transparency of the camera patch and the amplification of ADCP's attack success rate. This alignment with expectations is mostly upheld, except for the isolated case (W=0.1, TS=0.9). It is pertinent to mention that, in the experimental context, a higher value of the transparency parameter TS actually signifies a diminished transparency of the camera patch. Thirdly, a pinnacle attack performance is realized by ADCP when the camera patch's width is set to 0.5 and the transparency to 0.4. This signifies that further incremental adjustments in either width or transparency do not significantly elevate the attack success rate, as ADCP has already attained a 100% success rate in this configuration. Finally, an in-depth

examination of the outcomes delineated in Table 1 unveils that ADCP attains robust attack performance when the transparency TS reaches 0.3.

Based on the aforementioned insight, the configuration range for physical attacks has been delimited as follows: $0.1 \leq \mathrm{W} \leq 0.3, 0.3 \leq \mathrm{TS} \leq 0.5$. Furthermore, in other digital attack experiments, the width has consistently been set to 0.1 and the transparency to 0.5. Figure 4 graphically presents the adversarial samples engendered by the approach with a width of 0.1 and a transparency of 0.5. The top row showcases the detection outcomes for clean samples, while the subsequent row showcases the detection outcomes for adversarial samples. This visual confirmation substantiates the efficacy and adversarial potency of the methodology. These salient conclusions yield profound insights into the research, fostering an enhanced understanding of ADCP's attack characteristics, and furnishing pragmatic guidance for the configuration of physical attacks in real-world applications.

Physical test: To comprehensively assess the approach's efficacy in a physical context, the physical experiments are divided into two phases: indoor testing and outdoor testing.

Table 2: Results of indoor test experiments (ASR)

Distance	4.6m	5.2m	5.8m	6.4m	7.0m	7.6m	8.2m	8.8m
pl60+pl60	77.11%	88.03%	92.89%	83.07%	45.45%	55.96%	89.19%	83.00%
pl60+w57	98.02%	90.12%	92.13%	98.58%	98.53%	97.82%	99.50%	97.57%

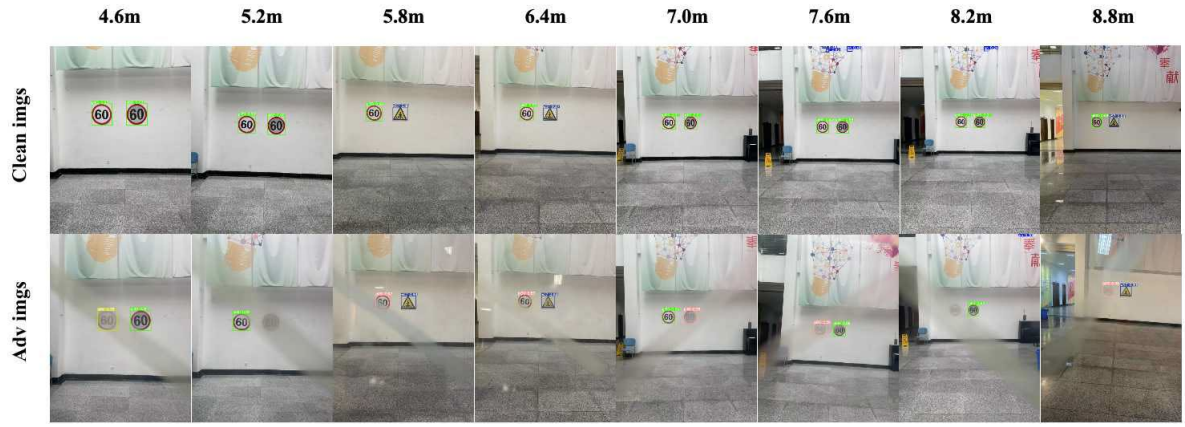

Figure 5: Indoor test

Indoor Testing Phase: To comprehensively assess the adversarial impact of the method across varied conditions, two distinct sets of experiments are conducted during the indoor testing phase. In the initial set, identical road signs are employed, while the subsequent set involved different road signs. These experiments are video-recorded, with the resultant footage being divided into frames to generate physical samples. Subsequently, the ASR is calculated based on these samples. The findings reveal that in the two sets of experiments, we accumulate 1630 and 4772 physical samples, respectively, achieving attack success rates of 78.16% and 96.31%. To gain a deeper understanding of ADCP's adversarial effect across different distances, a distance-based subdivision of the attacks is conducted, and statistical analysis is performed. The results of this analysis are presented in Table 2. These outcomes unequivocally underscore the efficacy and resilience of the method across the various tested distances within the indoor testing context. To present these results in a more intuitive manner, Figure 5 showcases the physical samples employed in indoor testing phase. This visual representation emphasizes that the method effectively executes physical attacks irrespective of the distance involved. During this testing stage, a camera patch width of W=0.1 and a transparency of TS=0.3 are opted for. Of notable importance is the fact that the presence of perturbations remains nearly imperceptible to the human eye without meticulous observation. This underscores the inherent imperceptibility of the attack method. Through this series of indoor testing experiments, confidence in the method's effectiveness within a controlled environment has been substantially fortified, laying a robust foundation for the research.

Table 3: Experimental results of outdoor testing(ASR)

Distance	4m	5m	6m	7m	8m	9m	10	11m	12m
0°	56.88%	100.00%	60.28%	61.26%	98.11%	100.00%	84.34%	87.50%	100.00%
30°	100.00%	100.00%	85.85%	87.16%	100.00%	72.90%	100.00%	100.00%	59.05%
45°	79.61%	63.03%	78.39%	73.58%	93.41%	100.00%	76.36%	87.96%	55.56%

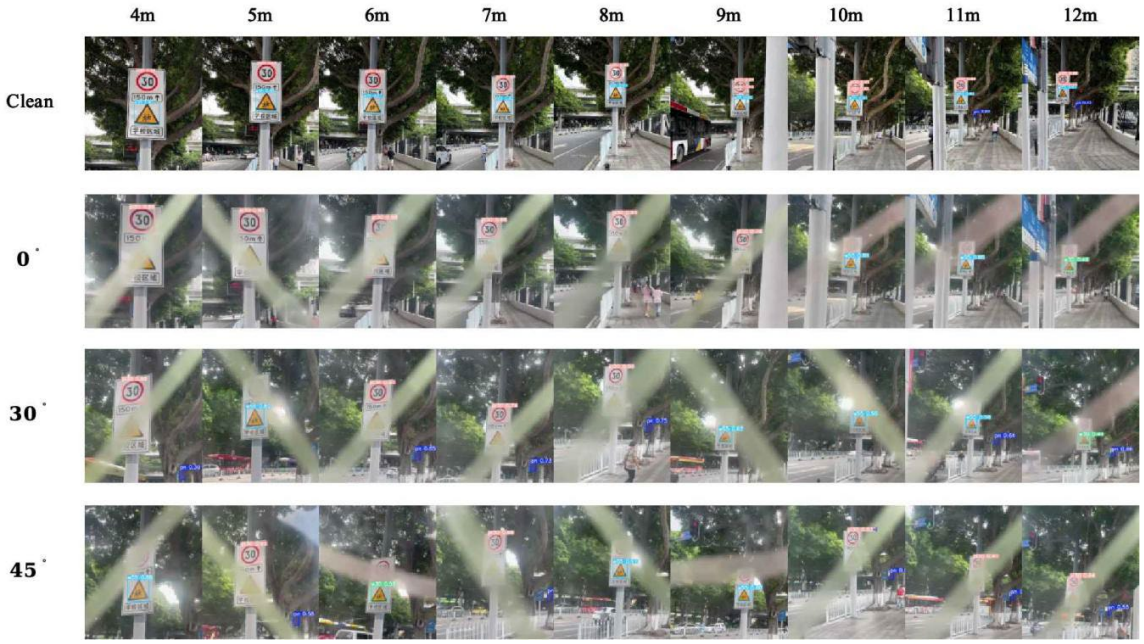

Figure 6: Outdoor test

Outdoor Testing Phase: In the outdoor testing phase, a transition is made to utilizing genuine road signs to subject the method to realistic attack scenarios. This shift allowed for a more comprehensive assessment of the method's robustness in real-world conditions. The objective is to evaluate the effectiveness of the method across various distances and angles, mirroring the unpredictability of genuine outdoor environments. During the outdoor testing, the method adhered to an attack configuration of $0.1 \leq \mathrm{W} \leq 0.3$ and $0.4 \leq \mathrm{TS} \leq 0.5$, thereby striking a balance between attack effectiveness and stealthiness. Impressively, outdoor testing yielded a total of 2624 genuine physical samples(pl30+w55), culminating in a remarkable attack success rate of 83.31%. This notable outcome underlines the practical viability and adversarial prowess of the approach within outdoor settings. The detailed experimental findings are elaborated in Table 3, which unequivocally illustrates that the method consistently manifests effective and resilient physical attack outcomes across diverse distances and angles. Remarkably, one third of the distance-angle combinations resulted in a 100% attack success rate. Even in the worst scenario (12m, 45°), a noteworthy attack success rate of 55.56% is achieved. Figure 6 visually captures the adversarial samples generated by ADCP during outdoor testing, further underscoring the method's capacity to execute impactful physical attacks across varying distances and angles. This comprehensive validation confirms the robustness and applicability of the method in complex outdoor scenarios. The series of outdoor experiments has yielded highly favorable outcomes, enhancing confidence in the practical applicability of the approach within real-world contexts.

Table 4: Comparison of experimental results between the approach and baselines

Method	Scenario	Digital attack		Physical attack	
		ASR	Query	Indoor ASR	Outdoor ASR
AdvCS(Li et al, 2019)	White-box	49.60%	ø	ø	73.26%
TTP(Zolfi et al,2021)	White-box	42.47%	ø	ø	42.27%
ADCP（Ours）	Black-box	93.75%	83.39	78.16/96.31%	88.31%

Table 4 presents a comprehensive comparison of experimental outcomes between ADCP method and the baseline approach. By scrutinizing these comparative findings, ADCP method consistently exhibits a more robust adversarial effect in both digital and physical settings when contrasted with the baseline method. This outcome underscores the inherent superiority of the approach. It is imperative to acknowledge that ADCP method operates in a black-box scenario, aligning closely with real-world practical applications. Moreover, in contrast to the AdvCS and TTP methodologies that employ multiple camera patches, which consequently yield larger experimental errors and inadequate stealthiness, ADCP approach's utilization of a solitary camera patch yields a higher attack effectiveness coupled with superior stealthiness. Drawing from the comprehensive analysis of Tables 1 and 4, it can be confidently assert that ADCP method bears a more formidable realistic threat compared to AdvCS and TTP methods. Its impressive experimental performance not only attests to its efficacy but also underscores its potential to pose significant challenges to computer vision systems within authentic environments. ADCP, thus, stands as a significant augmentation to the field of camera-based physical attacks, deserving attention for its multifaceted capabilities.

4.3 Ablation Study

In this section, a detailed exploration of another important influencing factor is conducted the color of the camera patch. The focus specifically encompasses the RGB channel of the color, encapsulating the *r, g,* and *b* values that span the range of (0, 255). Nevertheless, it is evidently impractical to execute ablation experiments for every conceivable color scenario, given the potential explosion in the number of experimental combinations. To strike a balance in the experimental design, a judicious trade-off was opted for. Ablation experiments with three distinct values were performed for each of the RGB channels: 0, 127, and 255. This approach resulted in 27 experimental configurations, carefully designed to thoroughly explore the role of color in the effectiveness of ADCP attacks. Figure 7 illustrates the outcomes of these color ablation experiments.

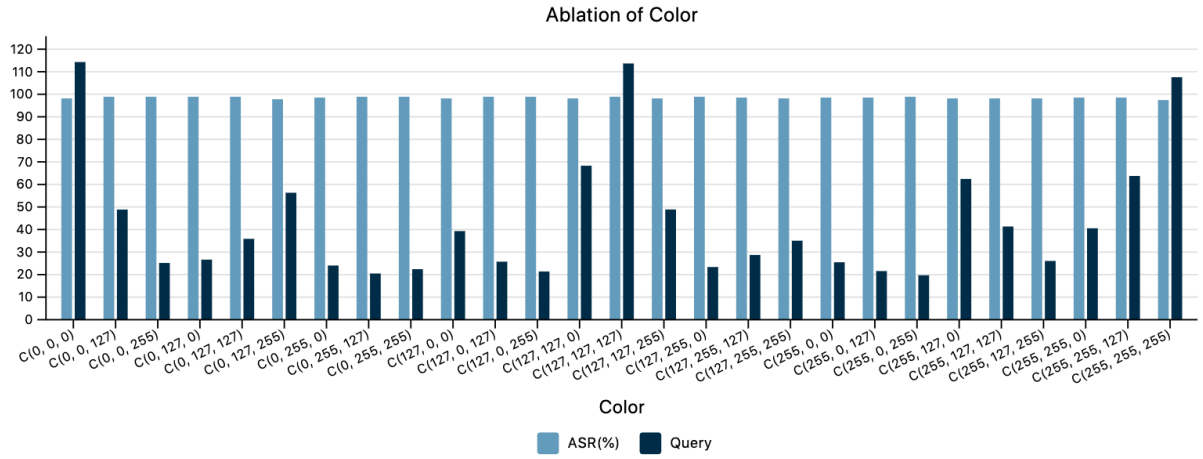

Figure 7: Ablation of C

The most important observation from these results is that color has minimal impact on the adversarial effectiveness of ADCP. In essence, camera patches sporting diverse colors seemingly exert relatively minor sway over the potency of the attack. This discovery tangibly bolsters the robustness and universality of the method. The method's ability to maintain a relatively steadfast attack effectiveness across different color variations reinforces its utility and adaptability, thus underscoring its reliability in contexts.

5. Conclusion

In this study, we introduce ADCP, a novel camera-based physical attack, optimized using the PSO algorithm. ADCP's key advantage over traditional methods lies in its ability to achieve successful attacks without modifying the target object, enhancing stealthiness and providing an advantage over existing patch-based and light-based attacks. Additionally, ADCP demonstrates superiority in operational simplicity and robustness, using only one camera patch to achieve reliable and resilient black-box attacks. The comprehensive experimental framework and results demonstrate the effectiveness and resilience of ADCP in real-world scenarios, evidencing its substantial adversarial impact in both digital and physical realms, as well as its high success rates in physical testing. The method's invisibility stems from perturbing the camera itself, enhancing stealth. The research highlights ADCP's security threats in the physical domain, with implications for enhancing comprehension and

application of physical attacks. With the increasing prevalence of physical adversarial attacks, the search for effective countermeasures continues to be a pivotal research focus.

Acknowledgements

This research work was partly Sponsored by Natural Science Foundation of Xinjiang Uygur Autonomous Region(No.2022D01B185).

References

Athalye et al.(2018) Synthesizing robust adversarial examples, in: International conference on machine learning, PMLR. pp. 28

Chen et al.(2019) Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector, in: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18, Springer. pp. 52–68.

Goodfellow et al.(2014) Explaining and harnessing adversarial examples. arXiv:1412.6572 .

Guesmi et al.(2023) Advrain: Adversarial raindrops to attack camera-based smart vision systems. arXiv preprint arXiv:2303.01338.

Gnanasambandam, A., Sherman, A.M., Chan, S.H.(2021) Optical adversarial attack, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 92–101.

Hu et al.(2023) Adversarial laser spot: Robust and covert physical-world attack to dnns, in: Asian Conference on Machine Learning, PMLR. pp. 483–498.

Hu et al.(2022) Adversarial texture for fooling person detectors in the physical world, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 13307–13316.

Huang et al.(2021) An improved shapeshifter method of generating adversarial examples for physical attacks on stop signs against faster r-cnns. Computers & Security 104, 102120.

Jocher(2020) Ultralytics yolov5. URL: https://github.com/ultralytics/ yolov5, doi:10.5281/zenodo.3908559.

Kennedy and Eberhart(1995) Particle swarm optimization, in: Proceedings of ICNN'95-international conference on neural networks, IEEE. pp. 1942–1948.

Li et al.(2019) Adversarial camera stickers: A physical camera-based attack on deep learning systems, in: International Conference on Machine Learning, PMLR. pp. 3896–3904.

Suryanto et al.(2022) Dta: Physical camouflage attacks using differentiable transformation network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15305–15314.

Szegedy et al.(2013) Intriguing properties of neural networks. arXiv:1312.6199 .

Tan et al.(2021) Legitimate adversarial patches: Evading human eyes and detection models in the physical world, in: Proceedings of the 29th ACM international conference on multimedia, pp. 5307–5315.

Thys et al.(2019) Fooling automated surveillance cameras: adversarial patches to attack person detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp. 0–0.

Wang et al.(2021) Dual attention suppression attack: Generate adversarial camouflage in physical world, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8565–8574.

Wu et al.(2020) Making an invisibility cloak: Real world adversarial attacks on object detectors, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16, Springer. pp. 1–17.

Zhong et al.(2022) Shadows can be dangerous: Stealthy and effective physical-world adversarial attack by natural phenomenon, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15345–15354.

Zhu et al.(2016) Traffic-sign detection and classification in the wild, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2110–2118.

Zolfi et al.(2021) The translucent patch: A physical and universal attack on object detectors, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15232–15241.