PhyCamo: A Robust Physical Camouflage via Contrastive Learning for Multi-View Physical Adversarial Attack

Ximin Zhang[1], Jinyin Chen[1*], Haibin Zheng[1], Zhenguang Liu[2*]

[1]Institute of Cyberspace Security and the College of Information Engineering, Zhejiang University of Technology
[2] School of Computer Science and Technology, Zhejiang University
{111123030020, chenjinyin}@zjut.edu.cn, haibinzheng320@gmail.com, liuzhenguang2008@gmail.com

Abstract

Deep neural networks (DNNs) have achieved remarkable success in widespread applications. Meanwhile, its vulnerability towards carefully crafted adversarial attacks captures special attention. Not only adversarial perturbations in digital space will fool the target DNNs-based detectors making a wrong decision, but also actually printed patches can be camouflaged to defeat detectors in physical space. In particular, multi-view physical adversarial attacks pose a more serious threat to practical scenarios. The existing attacks are still challenged in three aspects, i.e., high-cost data augmentation, attack performance gap between digital and physical space, and low attack transferability across DNNs. To overcome the challenges, we introduce *PhyCamo*, a robust physical camouflage framework based on contrastive learning that distinguishes itself from prior research in various critical ways: (1) *data augmentation* - it utilizes the diffusion model for data augmentation to efficiently simulate sophisticated physical dynamics in real-world; (2) *robustness* - it leverages contrastive learning to optimize physical camouflage against encoders with the state-of-the-art (SOTA) attack performance; (3) *transferability* - it mitigates the model-specific noise in the optimization by adopting diverse input methods, thereby amplifying the transferability between models. Extensive experiments are carried out on a car dataset, a tank dataset, and a pedestrian dataset, comparing with 6 classic multi-view physical adversarial attacks in both digital and physical spaces. The results demonstrate PhyCamo's superior performance. For instance, it generates more effective physical camouflage (with higher attack success rate $\sim$ $\times 1.26$ and reduce the model's average precision by 55%). PhyCamo can also help to improve the robustness of detectors through adversarial training, which contributes to the application of deep neural networks in the field of security sensitivity.

Introduction

Deep neural networks (DNNs) have achieved extraordinary performance in wide range fields, such as face recognition (Ding and Tao 2015; Raghavendra, Raja, and Busch 2015; Guo et al. 2021; Arashloo 2021), autonomous driving (Ma et al. 2022; Sun et al. 2021; Chen et al. 2019), and motion prediction (Liu et al. 2023; Zhenguang et al. 2019). Unfortunately, carefully crafted adversarial examples reveal the inherent vulnerability of DNNs by adding imperceptible noise to the clean examples, which has captured extensive attention in recent works (Athalye et al. 2018; Carlini and Wagner 2017; Kurakin, Goodfellow, and Bengio 2016; Laidlaw and Feizi 2019; Moosavi-Dezfooli, Fawzi, and Frossard 2016; Poursaeed et al. 2018). Taking DNNs-based detectors (Chung, Annaswamy, and Prabhakaran 2023) as an example, adversarial examples, altered directly in digital space, are fed into the detector to fool it making an expected wrong decision, such as making incorrect classification predictions, or even cause the predicted bounding box (bbox) to disappear or offset. Most researches (Goel et al. 2018; Chow et al. 2020a,b) has focused on this type of attack, named digital adversarial attack. More practice, adversarial attacks optimize the localized perturbations with unrestricted amplitude, thus they can fool the detector under various transformations (e.g., lightness changes, camera viewpoint changes, etc.), named physical adversarial attack (Dong et al. 2022; Duan et al. 2020; Athalye et al. 2018; Wang et al. 2021a, 2022; Suryanto et al. 2022; Zhang et al. 2023). It pose a more serious threat to DNNs-related scenarios since physical adversarial camouflage can be directly deployed in real world.

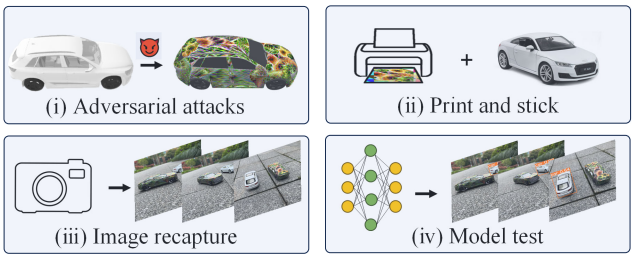

Figure 1: The pipeline of a typical multi-view physical adversarial attack against DNNs-based detector.

Intuitively, the physical adversarial attack is much more complicated than the digital one, due to uncontrollable physical transformations. In specific, as shown in Fig.1, a multi-view physical adversarial attack is generally launched following four steps, i.e., (i) adversarial camouflage in dig-

*Corresponding author.

ital space is carefully crafted; (ii) adversarial camouflage is printed and stuck on the objective model to conduct the multi-view physical adversarial attack; (iii) the physical adversarial object is recaptured by camera, which is significantly affected by transformations in the physical space (e.g., lightness changes, camera viewpoint changes, etc.); and (iv) the recaptured images are fed to the target DNNs-based detectors to affect its output.

It is noteworthy that the sequential operations of steps (ii) and (iii) are defined as digital-to-physical (D2P) transformation (Jan et al. 2019). It causes great color and shape distortions due to transformations in the physical space and characteristics of printing equipment, which brings main challenges for the multi-view physical adversarial attack.

To address the issue that physical adversarial examples should be robust to D2P transformation, recent works (Wang et al. 2022, 2021a; Suryanto et al. 2023; Zhang et al. 2023; Suryanto et al. 2022) improve the attack by simulating the possible D2P transformations in digital space to extend the training dataset and optimizing a well-designed loss function. Despite the improvement of performance, there are still three limitations shown below.

First of all, multi-view physical adversarial attacks could achieve better performance by emulating more D2P transformation (Feng et al. 2024). However, previous works (Suryanto et al. 2023; Zhang et al. 2023) struggle to strike a balance between effectiveness and efficiency in simulating D2P transformation in digital space. Suryanto et al. (Suryanto et al. 2023) performed several transformations (e.g., image flipping, translation, scaling, etc.) to simulate changes happening in physical space, which is easy to implement, but only provides a rough simulation. Zhang et al. (Zhang et al. 2023) combined a 3D model and virtual background images captured from CARLA (Dosovitskiy et al. 2017) to make a dataset, which offers effective physical dynamics but with extensive manual efforts.

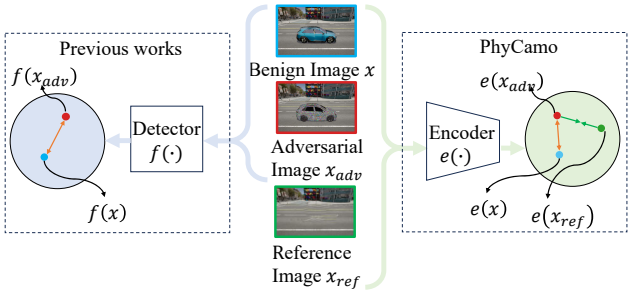

Figure 2: Previous works with detectors vs. PhyCamo with encoders.

Secondly, as shown in Fig.2, previous works (Wang et al. 2022, 2021a; Suryanto et al. 2023; Zhang et al. 2023) mainly focus on the difference between prediction of detectors and ground-truth. However, they overlooked the interaction between adversarial examples and reference examples (i.e., purely background examples) since different embeddings are beneficial to each other as they can act as anchors to better locate each other (Sha et al. 2023). For example, several works (Wang et al. 2022; Suryanto et al. 2022, 2023) design a stealthy loss function that aims to reduce the intersection over union (IoU) between the predicted bbox and the ground-truth, suppressing the region of the target predicted bbox. However, the stealthy loss only focuses on maximizing the difference between adversarial examples and benign ones while overlooking interacting across pairs. This means that they treat each image pair individually without interacting across pairs, which leads to finite effectiveness.

Thirdly, the existing multi-view physical adversarial attacks are usually designed for a targeted DNNs-based detector, while their attack capacity is limited when transferable to other models. It can be caused by overlooking the model-specific noise during the optimization process, since this noise can cause adversarial examples to overfit the source model, thus impeding their transferability. Relevant studies (Wang et al. 2021b; Zhang et al. 2023) have demonstrated that overfitting to model-specific noise was the culprit of low transferability theoretically and experimentally.

To overcome above challenges, our design goals are as follows: 1) we intend to mimic effective D2P transformation based on the diffusion model, so as to save laborious and time-consuming manual efforts; 2) we aim to reduce the embedding distance between adversarial examples and reference examples, designated as positive pairs for contrastive learning, while increasing the embedding distance for negative pairs consisting of adversarial examples and benign examples; 3) we adopted diverse input methods (Xie et al. 2019) to suppress model-specific noise. To mimic more diverse D2P transformation in physical space without expensive efforts, we proposed randomized input augmentation (RIA) that adopt a diffusion model-based image generation tool, named RePaint (Lugmayr et al. 2022) to emulate richer D2P transformation in the real world. Besides, with the advent of large models such as Whee, an image processing large model, we believe data augmentation has more high-quality solutions. In supplementary material, we demonstrate the data augmentation effectiveness of both the RIA and Whee.

Since the responses of encoders are the image representations in higher dimensions and with rich information compared to detectors' posteriors or labels, we thus assume that the encoders are more vulnerable to multi-view physical adversarial attacks. Accordingly, we quantitatively interpret the robustness of PhyCamo through an attack-independent robustness evaluation metric, namely Roby (Chen et al. 2020a), which quantifies the robustness of the adversarial examples based on the model's decision boundaries. Not surprisingly, we gain an important insight that the attack against the encoder achieves the lowest Roby value compared with benchmark methods, which implies that these adversarial examples provided by PhyCamo are substantially distant from the model's decision boundaries, rendering them more potent in deceiving detectors. Consequently, we speculate that encoders are more vulnerable to multi-view physical adversarial attacks. Besides, it should be noted that both pairs (positive pairs and negative pairs) are beneficial to each other as they can serve as anchors to better locate the position of the other embeddings in their space (Sha

et al. 2023). Hence, we consider both pairs during the optimization can improve the robustness of the attack. It has been proved that contrastive learning (Chen et al. 2020b) is a suitable approach to achieve this goal. PhyCamo is the first attempt of multi-view physical adversarial attack based on contrastive learning, which allows the embedding of the adversarial examples further away from the benign ones but closer to the reference ones.

The contributions of PhyCamo are summarized as follow.

(1) We devise RIA, which utilizes a diffusion model-based image generation tool for data augmentation, simulating diverse D2P transformation efficiently.

(2) We design positive and negative pairs for contrastive learning against encoders, which generate more effective physical camouflage and provides a new perspective on multi-view physical adversarial attacks.

(3) Extensive experiments are carried out on 3 datasets. The results show that PhyCamo outperforms SOTA baselines in both digital and physical spaces.

Related Work

In this section, we review related works on physical adversarial attacks and interpretable robustness evaluation metrics.

Physical Adversarial Attacks

Expectation over transformation (EoT) (Athalye et al. 2018) has become a leading approach for physical adversarial attacks (Sharif et al. 2019; Shen et al. 2019), considering factors like lighting and camera angles,which can be categorized into adversarial patch and multi-view physical attacks. Adversarial patch attacks are generally used in face recognition, disrupting detection by placing patches on the target. Methods like stickers on eyeglasses (Sharif et al. 2019), light-based attack (Shen et al. 2019) and FaceAdv (Shen et al. 2021). To overcome this limitation, multi-view physical adversarial attacks have been developed. Athalye et al. (Athalye et al. 2018) synthesized robust examples with EoT. Wang et al. (Wang et al. 2022) proposed FCA that works under various conditions. Suryanto et al. (Suryanto et al. 2022) introduced DTA that offers photo-realistic rendering. Suryanto et al. (Suryanto et al. 2023) presented ACTIVE that enhances vehicle camouflage against backgrounds.

Interpretable Robustness Evaluations

Interpretable robustness evaluation methods are crucial for assessing the effectiveness of adversarial attacks and defenses in DNNs (Zheng et al. 2022a,b). These methods establish a verifiable baseline for DNN robustness.

Moosavi-Dezfooli et al. (Moosavi-Dezfooli, Fawzi, and Frossard 2016) introduced empirical robustness to identify the smallest adversarial perturbation needed, though it lacks interpretability. Szegedy et al. (Szegedy et al. 2013) calculated global Lipschitz constants to explain robustness but resulted in loose bounds. Hein et al. (Hein and Andriushchenko 2017) used local Lipschitz conditions for a tighter lower bound on robustness but faced challenges with multi-layer networks. Chen et al. (Chen et al. 2020a) developed Roby, an efficient, structure-agnostic metric using feature distribution to verify DNN robustness.

PhyCamo

In this section, we first describe the definition of the problem. Then we elaborate on the proposed physical camouflage for multi-view physical adversarial attack.

Problem Definition

Attackers aims at defeating detectors when they got access to predictions of detectors. To further study the problem, we describe the multi-view physical adversarial attack as an optimization problem as follows.

For a 2D image x_0, the detector $f(\cdot)$ satisfies $f(x_0) = y = (b_x, b_y, b_w, b_h, b_{obj}, b_{cls})$, where y is predition of detector. b_x and b_y represent the center coordinates of the bbox, b_w and b_h represent the height and width of the bbox, b_{obj} indicates a confidence score that the bbox contains an object, b_{cls} is the class probability distribution of the object in the bbox. The purpose of the adversarial attack is to optimize the perturbations by solving the following formula:

$$\arg\min_{x_{adv}} L\left(f\left(x_{adv}\right), y_0\right) \tag{1}$$

where x_{adv} represents the adversarial example, y_0 denotes the ground-truth label, $L(\cdot)$ is a well-designed loss function.

Considering the presence of D2P transformation, multi-view physical adversarial attacks need to render the 3D vehicle model with mesh M and adversarial texture T_{adv} from 3D space to 2D plane to obtain 2D adversarial rendered image I_{adv} during each optimization iteration. On this basis, the optimization iteration of T_{adv} is carried out.

This work utilizes a modular differentiable renderer provided by Wang et al. (Wang et al. 2021a). Specifically, given the renderer R and camera parameters θ_c, I_{adv} is generated by:

$$I_{adv} = R\left(M, T_{adv}; \theta_c\right) \tag{2}$$

PhyCamo introduces an effective function $\varphi(\cdot)$ for combining I_{adv} with x_0. Based on the information provided by y_0, it is straightforward to uniformly trim the border of I_{adv}. The fusion of x_{adv} occurs as follows:

$$\begin{aligned} x_{adv}(i,j) &= \varphi\left(R\left(M, T_{adv}; \theta_c\right), y_0\right) \\ &= \varphi\left(I_{adv}(i,j), x_0(i,j), y_0\right) \\ &= \begin{cases} I_{adv}(i,j), i \in U_1, j \in U_2 \\ x_0(i,j), \text{ otherwise} \end{cases} \end{aligned} \tag{3}$$

where U_1 and U_2 represent the sets of x-coordinates and y-coordinates of the target, respectively, as derived from the data in y_0. $U_1 = \left(b_x - \frac{b_w}{2}, b_x + \frac{b_w}{2}\right)$, $U_2 = \left(b_y - \frac{b_h}{2}, b_y + \frac{b_h}{2}\right)$.

PhyCamo Framework

PhyCamo consists of three steps, (i) data augmentation and image fusion, (ii) predicted bbox reconstruction, (iii) embedding migration and enhancement, as shown in Fig.3. In step

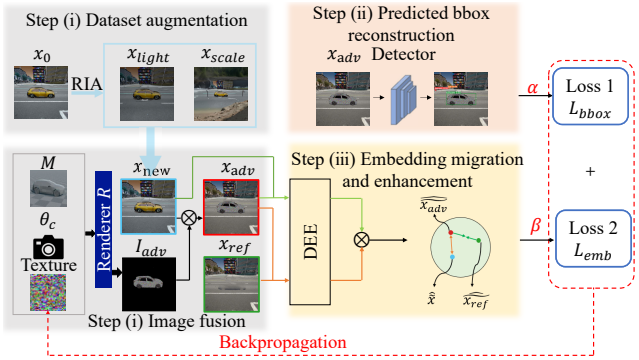

Figure 3: Framework of PhyCamo.

(i), the data augmentation part simulates more D2P transformation, and enhances training dataset $\mathcal{X}$ to obtain a new training dataset $\mathcal{X}_{new}$. The image fusion part is responsible for generating adversarial examples with the input of a 3D vehicle model with mesh M, texture T, and training dataset $\mathcal{X}_{new}$. Step (ii) is adopted to reduce IoU between the predicted bbox and ground-truth to suppress the region of ground-truth. As we know, step (ii) only targets detectors given their predicted labels, which leaves the vulnerability of encoders unexplored. To better leverage the rich representation of encoders, step (iii) is proposed. The basic idea of step (iii) is divided into the following two aspects: (1) maximizing the difference between the embedding of the adversarial image x_{adv} and that of the corresponding image x_{new} ($x_{new} \in \mathcal{X}_{new}$), and (2) minimizing the difference between the embedding of the adversarial image x_{adv} and the reference image x_{ref}. The reference image x_{ref} describes what the attacker desired: the scene has no target object on the road. Nowadays, an attacker can easily collect the reference image x_{ref} using various types of large AI models (e.g., Whee) or advanced image restoration (Lugmayr et al. 2022). By adopting step (iii), we weaken the target class-related embeddings and strengthen the target class-unrelated embeddings.

Data Augmentation and Image Fusion

PhyCamo have developed RIA to emulate more D2P transformations. Concretely, RIA employs RePaint (Lugmayr et al. 2022) to emulate alterations in camera position, and it replicates the variation in brightness within the physical environment by adjusting the value of the brightness channel in HSV color space. For instance, RIA diminishes the image scale [1] and employs RePaint (Lugmayr et al. 2022) to replenish the background, ensuring new image's size remain unaltered and mitigating the impact of size differences on texture optimization. The specific operations are as follows.

Initially, the original data x_0 is diminished through recent sampling to derive a smaller data x_{small}, along with a corresponding mask m:

$$m_{i,j} = \begin{cases} 0, i < w_{small}, j < h_{small} \\ 1, i \geq w_{small}, j \geq h_{small} \end{cases} \quad (4)$$

where w_{small} and h_{small} are the length and width of image x_{small}, respectively.

$$x_{scale} = D(x_{small}, m) \quad (5)$$

where $D(\cdot)$ is the diffusion model-based inpainting approach, and x_{scale} represents the image after data augmentation, $x_{scale} \in \mathcal{X}_{new}$. Subsequently, the brightness component of x_0 within the HSV color space will be randomly modified:

$$x_{light} = H(x_0, \delta) \quad (6)$$

where x_{light} represents the image after data augmentation, $x_{light} \in \mathcal{X}_{new}$. $H(\cdot)$ denotes the function responsible for altering the brightness value, δ is the amount of change in brightness.

Predicted Bbox Reconstruction

We propose the loss of detection box L_{bbox} for improving the effectiveness of T_{adv}. L_{bbox} takes into account the positional information of the detection box along with the length and width information and can be expressed as:

$$L_{\text{bbox}} = \sum_{i}^{N} \left[\text{IoU}\left(y^i, y_0^i\right) + \exp\left(\text{sim}\left(y^i, y_0^i\right)/\tau\right) \right] \quad (7)$$

where N denotes the multiscale predicted result output by the detector, y^i and y_0^i denote the predicted label and the real label of x_{adv}, respectively. $\text{sim}(u, v) = u^T v / \|u\| \|v\|$ denotes the cosine similarity between u and v see, and τ is a parameter to control the temperature.

Embedding Migration and Enhancement

To better utilize the rich information in embedding, we design the loss of embedding L_{emb}, a contrastive-learning-based loss for image embedding. Specifically, L_{emb} aims to force the embedding of x_{adv} to get close to that of x_{ref}, and further away from its corresponding original image x_{new}, where $x_{new} \in \mathcal{X}_{new}$. For the adversarial example x_{adv}, we have the corresponding two images original image x_{new}, and reference image x_{ref}. We consider (x_{adv}, x_{ref}) as a positive pair and (x_{adv}, x_{new}) as a negative pair.

A significant advantage in the computation of loss L_{emb} is its high computational efficiency since the embedding of the image can be obtained with a single forward propagation. However, the acquisition of the embedding inevitably results in model-specific noise. The presence of this noise is fatal for the transferability of physical camouflage attacks. To suppress this noise, inspired by Zhang et al. (Zhang et al. 2023), we combine the embedding migration with embedding enhancement and introduce dynamic embedding enhancement (DEE) in the optimization process. DEE mainly includes two parts: random transformations on input images and inverse transformations on the embedding. Specifically,

[1] Note that the scale is defined as the ratio of the target to the picture level. See supplementary material for the specific definition.

$q+1$ random transformations are applied to the images before they are fed into the encoder:

$$\tilde{x} = t_q \circ t_{q-1} \circ \cdots \circ t_0(x) \tag{8}$$

where $\widetilde{x}$ indicates the q th transformed image, obtained with a total of q individual random transformations $t_q \circ t_{q-1} \cdots t_0(x)$. Then, we input $\tilde{x}$ into the encoder e to get $e(\tilde{x})$.

Since $e(\widetilde{x})$ incorporates a random transformation of the image in its computation, $e(\widetilde{x})$ itself would be in a random manner. On which the inverse transformation operation is performed:

$$\widehat{\widetilde{x}} = \widehat{t_0} \circ \cdots \circ \widehat{t_{q-1}} \circ \widehat{t_q}(e(\tilde{x})) \tag{9}$$

where $\widehat{t_0} \cdots \widehat{t_q}$ denotes the inverse transformation corresponding to $t_q \cdots t_0(x)$, respectively. Therefore, the positive and negative pairs of L_{emb} are modified as follows:

$$D^+ = \exp\left(\mathrm{sim}\left(\widehat{\widetilde{x_{adv}}}, \widehat{\widetilde{x_{ref}}}\right)/\tau\right) \tag{10}$$

$$D^- = \exp\left(\mathrm{sim}\left(\widehat{\widetilde{x_{adv}}}, \widehat{\widetilde{x_{new}}}\right)/\tau\right) \tag{11}$$

In summary, $L_{emb} = D^+ - D^-$.

Consequently, our optimization objective can be summarized as

$$L_{total} = \alpha L_{bbox} + \beta L_{emb} \tag{12}$$

where α and β are the weights to control the contribution of each loss function. The determination of the values for α and β is contingent upon empirical experience.

Experiments Setting

In experiments, we conducted experiments on three datasets, comparing with six benchmark methods against six detectors across three evaluation metrics.

Data Preparation

We perform PhyCamo on three datasets, including a car dataset, a tank dataset, and a pedestrian dataset.

Car dataset. The dataset provided by Wang et al. (Wang et al. 2021a) consists of 155 points where target vehicles are placed in CARLA (Dosovitskiy et al. 2017), which is the commonly used open-source simulator for autonomous driving research (Teng et al. 2023; Tancik and Casser 2022).

Tank dataset. We amassed a self-constructed dataset with a tank model as the target. This dataset collected 900 images uniformly at 5 scales (0.1 - 0.5), 6 azimuths, and 5 pitch ranges.

Pedestrian dataset. The dataset refer to the work of Hu et al. (Hu et al. 2023), rendering the foreground photos of the pedestrian model.

Evaluation Mmetrics

The goal of PhyCamo is to generate effective adversarial camouflage for multi-view physical adversarial attacks. We select the attack success rate (ASR) and average precision (AP) to evaluate the effectiveness of PhyCamo and baselinses in both digital and physical spaces. We have employed the metric AP@0.5:0.95, calculated by averaging AP values with IoU thresholds ranging from 0.5 to 0.95 at an interval of 0.05. Better performance is indicated by an increased ASR and a decreased AP.

In addition, Roby is used to measure the robustness of PhyCamo and baselines. Roby integrates aggregation within the feature subspace (FSA) and separation between different classes in the feature subspace distance (FSD) into a single metric. A higher Roby value signifies a greater overlap among various feature subspaces, resulting in a more robust attack.

Target Models and Benchmark Methods

For easy reproducibility, we choose Retinanet, Yolo-V3, Yolo-V5, and Faster R-CNN(denoted as FrRCNN) as the target detectors. In addition, to evaluate the transferability of different physical adversarial attack methods, we use Retinanet, Yolo-V3, Yolo-V5, FrRCNN, MaskRCNN and D_Detr as test detectors.

We compare PhyCamo with previous works: DAS (Wang et al. 2021a), FCA (Wang et al. 2022), AT_2 (Wang et al. 2024), DTA (Suryanto et al. 2022), ACTIVE (Suryanto et al. 2023), and Boosting (Zhang et al. 2023). They are all typical 3D physical attacks, but they employ different methods. DAS generates adversarial examples by sabotaging heatmap salient regions. AT_2 introduces triplet attention suppression to manage distracting attention, converge incorrect class attention, and preserve content patch shape. FCA conducts attacks by suppressing target predicted region and reducing objectness confidence. ACTIVE renders the target undetectable by minimizing detection scores across valid classes. Boosting suppresses the target object's attention while reinforcing background attention.

Implementation Details

PhyCamo works differently from other baselines, thus we follow the comparison strategy of Zhang et al. (Zhang et al. 2023), i.e., evaluating ASR and AP of PhyCamo and baselines. We do not always replicate baseline approaches, but only adopt their primary losses under controlled conditions, including epoch numbers and training datasets. In particular, we reimplemented Boosting (Zhang et al. 2023) with full-coverage camouflage. In addition, when reproducing ACTIVE (Suryanto et al. 2023), we did not use the neural texture renderer (NTR) (Suryanto et al. 2023), but the neural mesh renderer (NMR) (Wang et al. 2022), which is employed by DAS (Wang et al. 2021a), FCA (Wang et al. 2022) since we needed to exclude the interference of the renderer and focus on comparing the effectiveness of the loss function.

Experiments Results and Analysis

To demonstrate the performance of PhyCamo, we implemented extensive experiments including six research questions (RQs): (1) how effective is PhyCamo in digital space compared to benchmark methods; (2) how effective is PhyCamo in physical space compared to benchmark methods; (3) do RIA and embedding loss improve the effectiveness of benchmark method; (4) how robust is PhyCamo compared

		Retinanet		Yolo-V3		Yolo-V5		FrRCNN		MaskRCNN		D_Detr	
		ASR↑	AP↓	ASR↑	AP↓	ASR↑	AP↓	ASR↑	AP↓	ASR↑	AP↓	ASR↑	AP↓
RAW	-	5.0	81.6	6.7	86.3	0.0	83.3	14.4	80.5	11.2	80.7	10.0	90.0
Retinanet	DAS	68.1	38.6	46.3	74.2	3.3	79.6	69.2	45.3	60.0	30.1	48.1	68.3
	AT_2	70.9	35.7	45.7	70.4	2.7	79.1	75.1	40.8	62.8	28.4	47.1	69.4
	DTA	75.4	32.1	43.7	65.4	4.1	75.4	75.8	40.2	65.4	27.6	50.1	70.9
	FCA	86.6	29.3	40.7	56.5	12.6	73.3	77.1	32.0	70.3	25.0	52.1	75.1
	ACTIVE	82.7	23.4	47.3	56.4	10.3	73.2	86.9	33.1	77.6	26.2	88.3	5.5
	Boosting	86.7	26.1	30.7	63.4	14.5	72.8	86.3	10.8	79.3	23.8	**96.9**	**3.6**
	PhyCamo	**90.3**	**19.5**	**47.7**	**50.7**	**20.2**	**63.1**	**88.9**	**9.1**	**87.6**	**20.0**	94.1	3.9
Yolo-V3	DAS	26.7	78.1	68.1	38.6	3.3	82.3	54.9	41.8	49.3	69.8	30.0	70.0
	AT_2	30.2	77.6	66.7	37.2	4.7	80.7	57.0	39.8	51.2	70.4	32.7	72.1
	DTA	31.3	75.4	65.1	32.6	4.9	80.4	60.7	37.4	52.7	38.9	33.4	75.4
	FCA	35.3	69.0	63.3	39.0	6.7	82.3	67.0	31.8	53.6	68.3	46.0	47.3
	ACTIVE	68.7	47.2	66.7	3.2	55.3	31.3	74.5	20.1	68.7	36.4	75.1	31.5
	Boosting	66.0	44.1	86.7	5.7	56.7	38.6	74.9	21.8	66.7	30.1	**88.3**	**6.7**
	PhyCamo	**71.0**	**38.9**	**96.7**	**3.2**	**77.6**	**15.3**	**77.8**	**17.4**	**69.1**	**27.2**	83.8	11.6
Yolo-V5	DAS	58.4	50.7	76.7	19.7	76.7	24.0	69.7	35.3	58.6	43.2	62.3	17.1
	AT_2	60.7	49.1	74.2	15.7	77.9	19.7	68.7	37.1	58.9	42.1	62.7	19.7
	DTA	62.6	47.2	73.0	25.4	76.2	21.3	66.7	38.4	59.0	40.7	63.0	21.5
	FCA	65.0	60.9	70.0	30.4	76.7	22.0	67.6	30.2	59.3	39.8	63.4	24.8
	ACTIVE	68.3	47.5	**96.7**	**1.2**	86.7	13.9	65.7	32.4	69.3	30.4	84.5	13.5
	Boosting	69.0	46.1	80.0	22.4	66.8	32.2	73.8	23.1	**79.5**	20.0	52.7	44.9
	PhyCamo	**72.0**	**21.3**	86.7	7.5	**96.7**	**2.4**	**76.2**	**20.9**	78.4	**15.3**	**88.3**	**4.8**
FrRCNN	DAS	60.0	30.4	18.4	70.7	6.7	74.0	89.2	25.3	63.6	38.3	32.1	59.0
	AT_2	63.7	28.7	21.1	68.7	7.9	73.7	88.9	26.7	65.7	37.4	34.9	59.6
	DTA	68.4	28.0	26.7	65.4	8.2	73.0	88.3	15.9	67.1	37.0	36.4	55.4
	FCA	80.0	27.3	47.6	60.1	6.7	72.0	87.4	11.0	69.7	36.5	38.0	56.4
	ACTIVE	80.0	22.2	38.3	56.4	18.7	63.9	87.3	10.7	79.7	25.7	81.4	13.4
	Boosting	83.4	14.4	42.3	62.9	16.8	62.2	75.9	10.9	76.6	29.7	99.0	4.6
	PhyCamo	**96.7**	**6.5**	**48.7**	**50.7**	**36.7**	**57.5**	**89.9**	**8.3**	**83.1**	**20.8**	**99.9**	**3.6**

Table 1: Effectiveness evaluation in digital space for multi-view physical adversarial attacks.

to benchmark methods; (5) how efficient is RIA compared with self-constructed dataset; (6) does adversarial training with PhyCamo improve the model's defense performance?

Due to page limits, experimental results are confined to the car dataset in the main content, while **details of tank dataset and pedestrian dataset are provided in the supplementary material** [2]. And the specifics of RQ3, RQ4, RQ5 and RQ6 are detailed in supplementary material, too.

Adversarial Effectiveness in Digital Space: Analysis From the Car Dataset

RQ1: How effective is PhyCamo in digital space compared to benchmark methods?

Car dataset. The evaluation results are shown in Table 1, including four target detectors listed in the first column and six detectors in the first row. Fig. 4 provide adversarial examples of PhyCamo.

Here we have the following observation: (1) In Table 1, PhyCamo generates more effective physical camouflage (with higher ASR $\sim \times 1.26$ and lower AP$\sim \times 0.55$ on average). As we can see, PhyCamo leads in AP and ASR when transferring from FrRCNN. It is because 1) RIA provides dataset with more D2P transformation, 2) PhyCamo considers both positive and negative pairs, which helps to locate more effective adversarial examples in embedding space, 3) PhyCamo mitigates the model-specific noise. (2) When targeted at Yolo-V5, PhyCamo didn't achieve the best performance detected by Yolo-V3. As we can see, adversarial ex-

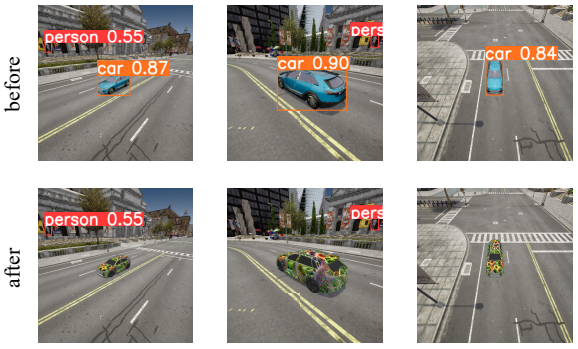

Figure 4: The detection result of the vehicle under different view angles before and after PhyCamo attack.

[2]https://github.com/zxm2020/PhyCamo

amples targeted the best detector cannot maintain the highest transferability. The inferior attack performance may stem from the fact that Yolo-V5 employs a more complex network with additional convolutional layers compared to Yolo-V3. This results in more refined embedding that is richer in model-specific noise, which can hinder the effectiveness of adversarial attacks.

Answer to RQ1: *PhyCamo outperforms the SOTA methods in digital space (with higher ASR* $\sim \times 1.26$ *and lower AP* $\sim \times 0.55$ *on average) and maintains high effectiveness with more subtle D2P transformations (with average ASR of 90.58%).*

Adversarial Effectiveness in Physical Space: Analysis From the Car Dataset

RQ2: How effective is PhyCamo in physical space compared to benchmark methods?

When answering this question, we experiment with the car datasets, wrapping textures onto models (1:64 scaled Audi model for car dataset), capturing test images, and testing with detectors as in (Wang et al. 2022). Due to cost and effort, we limit physical tests to 5 adversarial textures target at Yolo-V3.

attacks \ models	Yolo-V3	Yolo-V5	FrRCNN	MaskRCNN
DAS	37.3	10.4	30.7	49.3
AT_2	40.5	12.1	32.7	51.6
DTA	45.2	20.6	38.4	52.9
FCA	50.6	27.5	48.8	55.7
ACTIVE	52.5	47.4	60.5	65.4
Boosting	50.8	55.1	68.4	72.2
PhyCamo	**75.2**	**68.5**	**75.1**	**80.1**

Table 2: Effectiveness evaluation (ASR↑) in physical space for multi-view physical adversarial attacks.

Figure 5: The detection result of the vehicle with PhyCamo at different scales.

Car dataset. The evaluation results are shown in Table 2. Fig.6 shows the summarized performance of each camera position and pitch. Fig.5 shows the detection result of PhyCamo. Here we have the following observation: (1) As shown in Table 2, PhyCamo outperforms benchmark methods in physical space. This is primarily attributed to RIA's imitation of more D2P transformations, with concurrent suppression of model-specific noise while benchmark methods only provide rough D2P transformations. (2) As shown in Fig.6, PhyCamo is relatively stable under various distances compared to other methods. This is because RIA provided more D2P transformations in the optimization process than those of baselines.

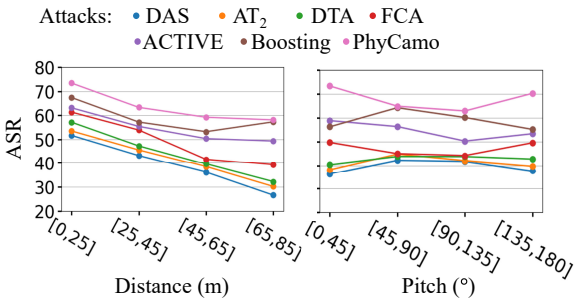

Figure 6: Attack comparison on different camera positions.

Here we have the following observation: PhyCamo remains effective in transitioning from the digital to the physical space and continues to function when transferring from Yolo_V3 to FrRCNN. PhyCamo attains an average ASR of 90.3% when employed with Yolo_V3 and an average ASR of 80.7% when employed with FrRCNN. This is because PhyCamo suppresses model-specific noise and uses RIA for data augmentation.

Answer to RQ2: *PhyCamo outperforms the SOTA methods in physical space. It generates physical camouflage with more than* $\sim \times 1.63$ *ASR on average compared to baselines.*

Conclusions

We propose a robust physical camouflage via contrastive learning, PhyCamo, to efficiently and effectively generate adversarial camouflage for DNNs-based detectors. PhyCamo devises RIA, which utilizes a diffusion model-based image generation tool for data augmentation, simulating diverse D2P transformation efficiently. Besides, PhyCamo designs positive and negative pairs for contrastive learning against encoders. We compare PhyCamo with six SOTA methods in three datasets against six DNNs-based detectors, the results show that PhyCamo has significantly better performance in terms of effectiveness and robustness. However, PhyCamo can't handle the texture blur caused by the change in camera distance. In future work, we will overcome this problem by studying the impact of D2P transformations on adversarial effectiveness.

Acknowledgments

This research was supported by the Zhejiang Provincial Natural Science Foundation (No. LDQ23F020001), National Natural Science Foundation of China (Nos. 62072406, 62406286), Key R&D Projects in Zhejiang Province (No. 2022C01018), National Key R&D Projects of China (No.

2018AAA0100801), and the Key R&D Program of Zhejiang Province (No. 2023C01217).

References

Arashloo, S. R. 2021. Matrix-regularized one-class multiple kernel learning for unseen face presentation attack detection. *IEEE Transactions on Information Forensics and Security*, 16: 4635–4647.

Athalye, A.; Engstrom, L.; Ilyas, A.; and Kwok, K. 2018. Synthesizing robust adversarial examples. In *International conference on machine learning*, 284–293. PMLR.

Carlini, N.; and Wagner, D. 2017. In *2017 ieee symposium on security and privacy (sp)*, 39–57. Ieee.

Chen, J.; Wang, Z.; Zheng, H.; Xiao, J.; and Ming, Z. 2020a. Roby: Evaluating the robustness of a deep model by its decision boundaries. *arXiv preprint arXiv:2012.10282*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020b. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Chen, Y.; Dong, C.; Palanisamy, P.; Mudalige, P.; Muelling, K.; and Dolan, J. M. 2019. Attention-based hierarchical deep reinforcement learning for lane change behaviors in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 0–0.

Chow, K.-H.; Liu, L.; Gursoy, M. E.; Truex, S.; Wei, W.; and Wu, Y. 2020a. Understanding object detection through an adversarial lens. In *Computer Security–ESORICS 2020: 25th European Symposium on Research in Computer Security, ESORICS 2020, Guildford, UK, September 14–18, 2020, Proceedings, Part II 25*, 460–481. Springer.

Chow, K.-H.; Liu, L.; Loper, M.; Bae, J.; Gursoy, M. E.; Truex, S.; Wei, W.; and Wu, Y. 2020b. Adversarial objectness gradient attacks in real-time object detection systems. In *2020 Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, 263–272. IEEE.

Chung, Y.-Y.; Annaswamy, T. M.; and Prabhakaran, B. 2023. Performance and User Experience Studies of HILLES: Home-based Immersive Lower Limb Exergame System. In *Proceedings of the 14th Conference on ACM Multimedia Systems*, 62–73.

Ding, C.; and Tao, D. 2015. Robust face recognition via multimodal deep face representation. *IEEE transactions on Multimedia*, 17(11): 2049–2058.

Dong, Y.; Zhu, J.; Gao, X.-S.; et al. 2022. Isometric 3d adversarial examples in the physical world. *Advances in Neural Information Processing Systems*, 35: 19716–19731.

Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; and Koltun, V. 2017. CARLA: An open urban driving simulator. In *Conference on robot learning*, 1–16. PMLR.

Duan, R.; Ma, X.; Wang, Y.; Bailey, J.; Qin, A. K.; and Yang, Y. 2020. Adversarial camouflage: Hiding physical-world attacks with natural styles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1000–1008.

Feng, W.; Xu, N.; Zhang, T.; Wu, B.; and Zhang, Y. 2024. Robust and Generalized Physical Adversarial Attacks via Meta-GAN. *IEEE Transactions on Information Forensics and Security*, 19: 1112–1125.

Goel, A.; Singh, A.; Agarwal, A.; Vatsa, M.; and Singh, R. 2018. Smartbox: Benchmarking adversarial detection and mitigation algorithms for face recognition. In *2018 IEEE 9th international conference on biometrics theory, applications and systems (BTAS)*, 1–7. IEEE.

Guo, J.; Zhu, X.; Lei, Z.; and Li, S. Z. 2021. Decomposed meta batch normalization for fast domain adaptation in face recognition. *IEEE Transactions on Information Forensics and Security*, 16: 3082–3095.

Hein, M.; and Andriushchenko, M. 2017. Formal guarantees on the robustness of a classifier against adversarial manipulation. *Advances in neural information processing systems*, 30.

Hu, Z.; Chu, W.; Zhu, X.; Zhang, H.; Zhang, B.; and Hu, X. 2023. Physically Realizable Natural-Looking Clothing Textures Evade Person Detectors via 3D Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16975–16984.

Jan, S. T.; Messou, J.; Lin, Y.-C.; Huang, J.-B.; and Wang, G. 2019. Connecting the digital and physical world: Improving the robustness of adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 962–969.

Kurakin, A.; Goodfellow, I.; and Bengio, S. 2016. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*.

Laidlaw, C.; and Feizi, S. 2019. Functional adversarial attacks. *Advances in neural information processing systems*, 32.

Liu, Z.; Wu, S.; Jin, S.; Ji, S.; Liu, Q.; Lu, S.; and Cheng, L. 2023. Investigating Pose Representations and Motion Contexts Modeling for 3D Motion Prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1): 681–697.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11461–11471.

Ma, Z.; Yang, Y.; Wang, G.; Xu, X.; Shen, H. T.; and Zhang, M. 2022. Rethinking open-world object detection in autonomous driving scenarios. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1279–1288.

Moosavi-Dezfooli, S.-M.; Fawzi, A.; and Frossard, P. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2574–2582.

Poursaeed, O.; Katsman, I.; Gao, B.; and Belongie, S. 2018. Generative adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4422–4431.

Raghavendra, R.; Raja, K. B.; and Busch, C. 2015. Presentation attack detection for face recognition using light field camera. *IEEE Transactions on Image Processing*, 24(3): 1060–1075.

Sha, Z.; He, X.; Yu, N.; Backes, M.; and Zhang, Y. 2023. Can't steal? Cont-steal! Contrastive stealing attacks against image encoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16373–16383.

Sharif, M.; Bhagavatula, S.; Bauer, L.; and Reiter, M. K. 2019. A general framework for adversarial examples with objectives. *ACM Transactions on Privacy and Security (TOPS)*, 22(3): 1–30.

Shen, M.; Liao, Z.; Zhu, L.; Xu, K.; and Du, X. 2019. Vla: A practical visible light-based attack on face recognition systems in physical world. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(3): 1–19.

Shen, M.; Yu, H.; Zhu, L.; Xu, K.; Li, Q.; and Hu, J. 2021. Effective and robust physical-world attacks on deep learning face recognition systems. *IEEE Transactions on Information Forensics and Security*, 16: 4063–4077.

Sun, Z.; Balakrishnan, S.; Su, L.; Bhuyan, A.; Wang, P.; and Qiao, C. 2021. Who is in control? Practical physical layer attack and defense for mmWave-based sensing in autonomous vehicles. *IEEE Transactions on Information Forensics and Security*, 16: 3199–3214.

Suryanto, N.; Kim, Y.; Kang, H.; Larasati, H. T.; Yun, Y.; Le, T.-T.-H.; Yang, H.; Oh, S.-Y.; and Kim, H. 2022. Dta: Physical camouflage attacks using differentiable transformation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15305–15314.

Suryanto, N.; Kim, Y.; Larasati, H. T.; Kang, H.; Le, T.-T.-H.; Hong, Y.; Yang, H.; Oh, S.-Y.; and Kim, H. 2023. Active: Towards highly transferable 3d physical camouflage for universal and robust vehicle evasion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4305–4314.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Tancik, M.; and Casser. 2022. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8248–8258.

Teng, S.; Hu, X.; Deng, P.; Li, B.; Li, Y.; Ai, Y.; Yang, D.; Li, L.; Xuanyuan, Z.; Zhu, F.; et al. 2023. Motion planning for autonomous driving: The state of the art and future perspectives. *IEEE Transactions on Intelligent Vehicles*.

Wang, D.; Jiang, T.; Sun, J.; Zhou, W.; Gong, Z.; Zhang, X.; Yao, W.; and Chen, X. 2022. Fca: Learning a 3d full-coverage vehicle camouflage for multi-view physical adversarial attack. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 2414–2422.

Wang, J.; Liu, A.; Yin, Z.; Liu, S.; Tang, S.; and Liu, X. 2021a. Dual attention suppression attack: Generate adversarial camouflage in physical world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8565–8574.

Wang, J.; Liu, X.; Yin, Z.; Wang, Y.; Guo, J.; Qin, H.; Wu, Q.; and Liu, A. 2024. Generate Transferable Adversarial Physical Camouflages via Triplet Attention Suppression. *Int. J. Comput. Vision*, 132(11): 5084–5100.

Wang, Z.; Guo, H.; Zhang, Z.; Liu, W.; Qin, Z.; and Ren, K. 2021b. Feature Importance-aware Transferable Adversarial Attacks. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 7619–7628.

Xie, C.; Zhang, Z.; Zhou, Y.; Bai, S.; Wang, J.; Ren, Z.; and Yuille, A. L. 2019. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2730–2739.

Zhang, Y.; Gong, Z.; Zhang, Y.; Bin, K.; Li, Y.; Qi, J.; Wen, H.; and Zhong, P. 2023. Boosting transferability of physical attack against detectors by redistributing separable attention. *Pattern Recognition*, 138: 109435.

Zheng, H.; Chen, J.; Du, H.; Zhu, W.; Ji, S.; and Zhang, X. 2022a. GRIP-GAN: An Attack-Free Defense Through General Robust Inverse Perturbation. *IEEE Transactions on Dependable and Secure Computing*, 19(6): 4204–4224.

Zheng, H.; Chen, Z.; Du, T.; Zhang, X.; Cheng, Y.; Ji, S.; Wang, J.; Yu, Y.; and Chen, J. 2022b. NeuronFair: interpretable white-box fairness testing through biased neuron identification. In *Proceedings of the 44th International Conference on Software Engineering*, ICSE '22, 1519–1531. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392211.

Zhenguang, L.; Wu, S.; Jin, S.; Liu, Q.; Lu, S.; Zimmermann, R.; and Cheng, L. 2019. Towards Natural and Accurate Future Motion Prediction of Humans and Animals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.