UPC: Learning Universal Physical Camouflage Attacks on Object Detectors

Lifeng Huang\(^1\), Chengying Gao\(^1\), Yuyin Zhou\(^2\), Changqing Zou\(^3\), Cihang Xie\(^2\), Alan Yuille\(^2\), Ning Liu\(^1\)

\(^1\)School of Data and Computer Science, Sun Yat-sen University
\(^2\)Department of Computer Science, The Johns Hopkins University
\(^3\)University of Maryland, College Park

Abstract

In this paper, we study physical adversarial attacks on object detectors in the wild. Prior arts on this matter mostly craft instance-dependent perturbations only for rigid and planar objects. To this end, we propose to learn an adversarial pattern to effectively attack all instances belonging to the same object category (e.g., person, car), referred to as Universal Physical Camouflage Attack (UPC). Concretely, UPC crafts camouflage by jointly fooling the region proposal network, as well as misleading the classifier and the regressor to output errors. In order to make UPC effective for articulated non-rigid or non-planar objects, we introduce a set of transformations for the generated camouflage patterns to mimic their deformable properties. We additionally impose optimization constraint to make generated patterns look natural for human observers. To fairly evaluate the effectiveness of different physical-world attacks on object detectors, we present the first standardized virtual database, AttackScenes, which simulates the real 3D world in a controllable and reproducible environment. Extensive experiments suggest the superiority of our proposed UPC compared with existing physical adversarial attackers not only in virtual environments (AttackScenes), but also in real-world physical environments.

1. Introduction

Deep neural networks (DNNs) have achieved outstanding performances on many computer vision tasks \([34, 9, 11]\). Nonetheless, DNNs have been demonstrated to be vulnerable to adversarial examples \([35]\) — maliciously crafted inputs that mislead DNNs to make incorrect predictions, which present potential threats for the deployment of DNN-based systems in the real world.

Different adversarial attacks have been proposed recently \([25, 3]\), which can be divided into the following categories: 1) digital attacks, which mislead DNNs by modifying the input data directly in the digital space (e.g., pixel value \([25, 12, 22]\), text content \([16, 29]\), voice signal \([4, 1]\)); 2) physical attacks, which attack DNNs by altering the characteristics of an object (e.g., color \([30]\), appearance \([8]\)) in the physical world. Examples of both digital and physical attacks are shown in Fig. 1. Current mainstream works of adversarial machine learning focus on the digital domain, which can be hardly transferred to the real world due to the lack of considering physical constraints (e.g., invariant to different environmental conditions such as viewpoint, lighting) \([8]\). In this paper, we study adversarial attacks in the physical world, which are more threatening to real-world computer vision systems \([15]\). While previous works in this category \([13, 2]\) mostly focus on attacking image classification systems, we consider the far more realistic computer vision scenario, i.e., object detection.

Though prior works have revealed the vulnerability of object detectors to adversarial perturbations in the real world, e.g., \([5, 33]\) generate perturbed stop signs, there are several limitations: (1) these algorithms focus on only at-
tacking a specific object (e.g., a stop sign [8, 5] or a commercial logo [32]); (2) existing methods generate perturbations only for rigid and planar objects (e.g., traffic sign, wall, board), which can be less effective for complex objects (articulated non-rigid or non-planar objects, e.g., human); (3) the generated perturbations lack semantics [5], which can be unnatural for human observers; and (4) a unified evaluation environment is missing, which makes it difficult to make fair comparisons between different physical attacks.

To address these issues, we present Universal Physical Camouflage Attack (UPC), which constructs a universal camouflage pattern to hide objects from being detected or to misdetect objects as the target label. Unlike former works which generate instance-level perturbations, UPC constructs a universal pattern to attack all instances that belong to the same object category (e.g., person, cars) via jointly attacking the region proposal network, as well as misleading the classifier and the regressor to output errors. To efficiently handle the deformations of complex objects in the physical world, we propose to model their internal deformable characteristics as well as external physical environments in our framework. Specifically, the internal properties are simulated by applying different geometric transformation functions (e.g., cropping, resizing, affine homography). We impose additional optimization constraint (Eq. 1) to encourage the visual resemblance between generated patterns and natural images, which we refer to as semantic constraint. As shown in Fig. 1, these generated camouflage patterns are visually similar to natural image patterns and thus can be regarded as texture patterns on object surfaces such as human accessories/car paintings. The overall pipeline is illustrated in Fig. 2. Extensive experiments suggest that the camouflage patterns generated by UPC can be effectively applied to both virtual scenes and the real physical world.

To fairly evaluate the effectiveness of different physical attacks, we provide the first standardized synthetic dataset, i.e., AttackScenes, which includes 20 virtual scenes. All experimental data is generated under strict parametric-controlled physical conditions (e.g., lighting or viewpoints) to ensure that the evaluation comparison is reliable and reproducible under virtual settings. Our proposed UPC achieves state-of-the-art results among all existing methods.

The contributions of our work are four-fold:

- Unlike prior works which craft instance-level adversarial patterns, UPC constructs a universal camouflage pattern for effectively attacking object detectors based on the fact that the generated pattern can be naturally camouflaged as texture patterns on object surfaces such as human accessories/car paintings (Fig. 1).
- We present the first standardized dataset, AttackScenes, to ensure that all experiments are conducted under fair comparisons. AttackScenes simulates the real 3D world under controllable and reproducible settings, which enables fair evaluation standards for future research in this domain.
- To make UPC effective for articulated non-rigid or non-planar objects, we introduce additional transformations for the camouflage patterns to simulate their internal deformations.
- Extensive experiments suggest that UPC not only generates effective camouflage patterns for attacking object detectors in the wild, but also exhibits well generalization and transferability among different models as well as different datasets.

2. Related Works

**Universal Adversarial Attack.** Unlike generating image-dependent adversarial perturbations [25, 12, 3], the image-agnostic attack, i.e., universal adversarial attack [24, 14], is defined as an attack which is able to fool different images with a single global pattern in the digital domain. Here we extend this definition to the physical domain and define instance-agnostic perturbations as universal physical attacks for object detectors. Unlike former physical attack methodologies which craft instance-level patterns, our goal is to generate a single camouflage pattern to effectively attack all instances of the same object category given different physical scenes.

**Physical Attacks.** Stem from the recent observation that printed adversarial examples can fool image classifiers in the physical world [15, 13], efforts have been investigated to study how to construct “robust” adversarial examples in the real physical world. For instance, Athalye et al. [2] propose to construct 3D adversarial objects by attacking an ensemble of different image transformations; Sharif et al. [30] successfully attack facial recognition systems by printing textures on eyeglasses; Evtimov et al. [8] use poster, sticker and graffiti as perturbations to attack stop signs in the physical world. Zeng et al. [37] apply computer graphics rendering methods to perform attacks in the 3D physical world.

These aforementioned works focus on fooling image classifiers in physical environments. Recently, physical attacks have also been studied for the more challenging scenario of object detection. Song et al. [33] propose a disappearance and creation attack to fool YoloV2 [26] in traffic scenes. Chen et al. [5] adopt the expectation over transformation method [2] to create more robust adversarial stop signs, which mislead faster r-cnn [28] to output errors. However they cannot be effectively applied to non-rigid or non-planar objects since they only focus on simulating external environment conditions, e.g., distances or viewpoints, for attacking object detectors. In addition, these approaches generate instance-dependent patterns which exhibit less semantics and therefore the perturbed images are usually un-
natural and noisy. Different from these works, our method constructs a universal semantic pattern which makes the perturbed images visually similar to natural images. Meanwhile, we introduce additional transformations to simulate the deformable properties of articulated non-rigid or non-planar objects. A detailed comparison with former methods is summarized in Table 1.

| Methods | Rigid | Non-Rigid | Planar | Non-Planar | Universal | Semantic |
|---------|-------|-----------|--------|------------|-----------|----------|
| [5]     | ✓     | ✓         | ✓      |            | ✓         | ✓        |
| [33]    | ✓     | ✓         | ✓      | ✓          | ✓         | ✓        |
| Ours    | ✓     | ✓         | ✓      | ✓          | ✓         | ✓        |

3. Methodology

3.1. Overview

Our goal is to attack object detectors by either hiding the object from being detected, or fooling detectors to output the targeted label. Without loss of generality, we use “person” category as an example to illustrate our method.

Training framework of UPC in Digital Space. We attack faster-rcnn [28], a two-stage detector, in this paper. In the first stage, the region proposal network is employed to generate object proposals. In the second stage, the detector selects top-scored proposals to predict labels. We propose to craft a universal pattern for faster-rcnn by jointly fooling the region proposal network to generate low-quality proposals, i.e., reduce the number of valid proposals, as well as misleading the classifier and the regressor to output errors. Simply misleading predictions of the classification head cannot produce satisfying results (discussed in Sec. 5.2) because it can be impractical to attack enormous candidate proposals simultaneously. Extensive experimental results also validate that the joint attack paradigm demonstrates stronger attacking strength than simply attacking the classifier as in prior methods [5, 8] (Table 3). Furthermore, to deal with complex objects, we propose to simultaneously model both internal deformable properties of complex objects and external physical environments. The internal attributes of objects, i.e., deformations, are simulated by a series of geometric transformations. As illustrated in Fig. 2(a), UPC consists of 3 steps:

- **Step 1.** A set of perturbed images are synthesized by simulating external physical conditions (e.g., viewpoint) as well as internal deformations of complex objects. An additional optimization constraint is imposed to make the generated patterns semantically meaningful. (Sec. 3.2)

- **Step 2.** Initial adversarial patterns are generated by attacking the RPN, which results in a significant drop of high-quality proposals. (Sec. 3.3)

- **Step 3.** To enhance the attacking strength further, UPC then jointly attacks RPN as well as the classification and the bounding box regression head by lowering the detection scores and distorting the bounding box. (Sec. 3.4)

We perform these steps in an iterative manner until the termination criterion is satisfied, i.e., fooling rate is larger than the threshold or the attacking iteration reaches the maximum.

Attacking in Physical Space. By imposing the semantic constraint (Sec. 3.2), the generated camouflage patterns by UPC look natural for human observers and thus can be regarded as texture patterns on human accessories. Concretely, we pre-define several regions of human accessories (e.g., garment, mask) to paint on the generated camouflage patterns (Fig. 4) for attacking, and the corresponding physical scenes are captured under different viewing conditions (e.g., illumination, viewpoints, distances, etc.) for testing (Fig. 2(b)).

3.2. Physical Simulation

Material Constraint. To keep generated adversarial patterns less noticeable, the perturbations are camouflaged as texture patterns on human accessories (e.g., garments, masks). External environments are simulated via controlling factors such as lighting, viewpoint, location and angle [5, 8].
To effectively handle non-rigid or non-planar objects, we also introduce addition transformation functions to model their internal deformations (Eq. 2).

**Semantic Constraint.** Inspired by the imperceptibility constraint in digital attacks, we use the projection function (Eq. 1) to enforce the generated adversarial patterns to be visually similar to natural images during optimization. Empirical results show that optimizing with this constraint yields high-quality semantic patterns, which can be naturally treated as camouflages on human clothing (Fig. 8).

**Training Data.** To obtain universal patterns, images with different human attributes (body sizes, postures, etc.) are sampled as the training set $X$. In summary, the perturbed images are generated by:

$$\delta^t = \text{Proj}_{\infty}(\delta^{t-1} + \Delta \delta, I, \epsilon),$$

$$\tilde{X} = \{ \tilde{x}_i | \tilde{x}_i = T_r(x_i + T_c(\delta^t), x_i \sim X) \}. \tag{2}$$

Eq. 1 is the semantic constraint, where $\delta^t$ and $\Delta \delta$ denote the adversarial pattern and its updated vector at iteration $t$, respectively. $\text{Proj}_{\infty}$ projects generated pattern onto the surface of $L_{\infty}$ norm-balls with radius $\epsilon$ and centered at $I$. Here we choose $I$ as natural images to ensure the generated camouflage patterns are semantically meaningful. Eq. 2 is the physical simulation we applied during the attack, where $T_r$ is used for the environmental simulation (e.g., brightness adjustment), $T_c$ is used for modeling the material constraint (e.g., deformations induced by stretching), and $\hat{x}$ is the generated perturbed image (marked as blue in Fig. 2(a)).

### 3.3. Region Proposal Network (RPN) Attack

For an input image with height $H$ and width $W$, the RPN extracts $M = O(HW)$ proposals across all anchors. We denote the output proposals of each image $\hat{x}$ as $P = \{ p_i | p_i = (s_i, \vec{d}_i); i = 1, 2, 3, \ldots, M \}$, where $s_i$ is the confidence score of $i$-th bounding box and $\vec{d}_i$ represents the coordinates of $i$-th bounding box. We define the objective function for attacking the RPN as following:

$$L_{\text{rpn}} = \mathbb{E}_{p_i \sim P} (L(s_i, y^t) + s_i \parallel \vec{d}_i - \Delta \vec{d}_i \parallel_p), \tag{3}$$

where $y^t$ is the target score, and we set $y^1$ for background and $y^0$ for foreground; $L$ is the Euclidean distance loss; $\Delta \vec{d}_i$ is used for attacking proposals by shifting the center coordinate and corrupting the shape of original proposals; $p$ is the norm constant and we set $p = 1$ in the experiment.

By minimizing $L_{\text{rpn}}$, our goal is to generate adversarial patterns for RPN which results in a substantial reduction of foreground proposals and severely distorted candidate boxes (marked as red in Fig. 2(a)).

### 3.4. Classifier and Regressor Attack

After applying non-maximum suppression (NMS) on the outputs of RPN, top-$k$ proposals are ordered by their confidence scores and selected as a subset $P'$. These top-scored proposals $P'$ are then fed to the classification and the regression head for generating final outputs. We note that if only a subset of proposed bounding boxes are perturbed, the detection result of the attacked image may still be correct if a new set of candidate boxes is picked in the next iteration, which results in great challenges for attackers. To overcome this issue, we instead extract proposals densely as in [36]. Specifically, we attack an object by either decreasing the confidence of the groundtruth label or increasing the confidence of the target label. We further enhance the attacking strength by distorting the aspect ratio of proposals and shifting the center coordinate simultaneously [18]. In summary, we attack the classification and the regression head by:

$$L_{\text{cls}} = \mathbb{E}_{p \sim P'} (C(p)_o + \mathbb{L}_{C(p)_{\max} \in o} (C(p), y^t), \tag{4}$$

$$L_{\text{reg}} = \mathbb{E}_{C(p)_{\max} \in o} \parallel R(p)_o - \Delta \vec{d}_i \parallel_p, \tag{5}$$

where $\mathbb{L}$ is the cross-entropy loss, $C$ and $R$ are the prediction output of the classifier and the regressor, $o$ is the true label, $t$ is the target label for attacking, and $\Delta \vec{d}$ denotes the distortion offset. We select $\ell_2$ norm, i.e., $p = 2$ in Eq. 5. Eq. 4 and Eq. 5 are designed for fooling the classifier and the regressor, respectively, and are referred to as C&R attack (marked as green in Fig. 2(a)). For **untargeted attack**, we set $t = o$ for maximizing (instead of minimizing) Eq. 4.
3.5. Two-Stage Attacking Procedure

In summary, UPC generates the physical universal adversarial perturbations by considering all the factors above:

$$\arg\min_{\delta} \mathbb{E}_{\tilde{x} \sim \tilde{X}} \left( L_{rpn} + \lambda_1 L_{cls} + \lambda_2 L_{reg} + L_{tv}(\delta) \right), \quad (6)$$

where $\delta$ and $\tilde{X}$ denote the universal pattern and the set of perturbed images, respectively. $L_{tv}$ stands for the total variation loss [23] with $l_2$ norm constraint applied. We note that $L_{tv}$ is important for reducing noise and producing more natural patterns.

The overall procedure of UPC is illustrated in Algorithm 1, where we alternately update the universal perturbation pattern $\delta$ and the perturbed images $\tilde{x} \sim \tilde{X}$ until the fooling rate becomes larger than a certain threshold or the attack iteration reaches the maximum. $\delta$ is updated using a two-stage strategy. During the first stage, we exclusively attack the RPN to reduce the number of valid proposals, i.e., set $\lambda_1 = 0$ and $\lambda_2 = 0$ in Eq. 6. After significantly reducing the number of high-quality proposals, our attack then additionally fools the classification and bounding box regression head in the second stage. By minimizing Eq. 6, the generated perturbation $\delta$ substantially lowers the quality of proposals and thereby achieves a high fooling rate.

4. AttackScenes Dataset

Due to the lack of a standardized benchmark dataset, earlier works measure the performance under irreproducible physical environments, which makes it difficult to make fair comparisons between different attacks. To this end, we build the first standardized dataset, named AttackScenes, for fair and reproducible evaluation.

Environments. AttackScenes includes 20 virtual scenes under various physical conditions (Fig. 3). Specifically, there are 10 indoors scenes (e.g., bathroom, living room) and 10 outdoors scenes (e.g., bridge, public road, market) in total.

Camera Setting. For each virtual scene, 18 cameras are placed for capturing images from different viewpoints. To ensure the diversity of images, these cameras are located at different angles, heights and distances (Fig. 2(b)).

Illumination Control. To the best of our knowledge, the experiments in earlier studies usually conduct tests in bright environments. However, this simulated condition is quite limited since there exist many dark scenes in the real world. Accordingly, we extend the testing environment to different levels of lighting conditions to better simulate different daily times like evening, dusk and dawn. For indoor scenes, area and sphere lights are used to simulate interior illumination. Meanwhile, we use directional light sources to simulate sunlight outdoors. The illumination varies from dark to bright at 3 levels by controlling the strength of light sources (i.e., L1-L3).

5. Experiments

In this section, we empirically show the effectiveness of the proposed UPC by providing thorough evaluations in both virtual and physical environments. Our experiments show that UPC can effectively attack proposal-based object detectors in both virtual environments (AttackScenes) and the physical world, and also exhibits well transferability to other unknown models.
5.1. Implementation Details

We mainly evaluate the effectiveness of our method on “person” category due to its importance in video surveillance and person tracking [17]. We collect 200 human images with various attributes (e.g., hair color, body size) as our training set to generate universal adversarial patterns. Following [36], we evaluate the performance of faster r-cnn using 2 network architectures (i.e., VGG-16 [31] and ResNet-101[9]) which are either trained on the PascalVOC-2007 trainval, or on the combined set of PascalVOC-2007 trainval and PascalVOC-2012 trainval. We denote these models as FR-VGG16-0712, FR-RES101-0712, FR-VGG16-0712 and FR-RES101-0712.

Parameters setting. We set $\lambda_1 = 1.0$ and $\lambda_2 = 0.01$ in Eq. 6 as initialization. For generating universal adversarial patterns, we set $\text{iter}_s = 100$ and the maximum iteration $\text{iter}_{\text{max}} = 2000$ in Algorithm 1.

Evaluation Metric. For faster r-cnn, we set the threshold $\text{iter}_{\text{max}} = 2007$ trainval. For faster r-cnn, we set the threshold $\text{iter}_{\text{max}} = 2007$ trainval. For faster r-cnn, we set the threshold $\text{iter}_{\text{max}} = 2007$ trainval.

5.2. Virtual Scene Experiment

Human Model and Pattern Schemes. We select human models in AttackScenes with different poses (i.e., standing, walking and sitting) as the attacking target. 6 different schemes (Fig. 4) are used under the material constraint (Sec. 3.2) for experimental comparison.

In the virtual scene experiment, 1080 $(20 \times 3 \times 18)$ images are rendered for each pattern scheme. Without loss of generality, we choose “dog” and “bird” as target labels to fool detectors in our experiment. We use 6 different pattern schemes (i.e., Original/Naive/Natural/3-Patterns/7-Patterns/8-Patterns schemes as illustrated in Fig. 4) for validating the efficacy of the propose UPC.

As shown in Table 2, we find that the attack strength is generally weaker in darker environments. This can be attributed to the fact that the adversarial patterns are badly captured when the level of brightness is low, which induces low-quality attacks. Additionally, we observe that for different human poses the average precision almost stays the same level (i.e., Standing: $p_{0.5}$ drops from 0.98 to 0.97/0.96; Walking: $p_{0.5}$ drops from 0.95 to 0.94/0.93; Sitting: $p_{0.5}$ drops from 0.98 to 0.94/0.95 using FR-VGG16) via attacking Naive/Natural pattern scheme which indicates that simply using naive camouflage or natural images as adversarial patterns is invalid for physical attacks.

By contrast, our method yields a distinct drop rate of $p_{0.5}$ for all 3 pattern schemes (i.e., 3/7/8-Pattern schemes), among which 8-Pattern scheme observes the highest performance drop (i.e., Standing: $p_{0.5}$ drops from 0.98 to 0.97; Walking: $p_{0.5}$ drops from 0.95 to 0.94; Sitting: $p_{0.5}$ drops from 0.98 to 0.96 using FR-VGG16). It is no surprise to observe such a phenomenon since using more generated patterns for physical attack results in more surface area occluded, which naturally leads to a higher fooling rate. The detection result further shows our attack is invariant to different viewing conditions (e.g., viewpoints, brightness).

Table 2. Average precision $p_{0.5}$ in virtual scene experiments after attacking faster r-cnn. Note that $p_{0.5}$ is averaged over all viewpoints of each pattern scheme under 3 brightness conditions.

| Network | 0712 | 0712 |
|---------|------|------|
| Pose    | Standing | Walking | Sitting |
| UPC     | 0.07 (0.91) | 0.04 (0.91) | 0.46 (0.52) |
| UPC     | 0.63 (0.32) | 0.33 (0.82) | 0.76 (0.22) |
| CLS     | 0.17 (0.80) | 0.06 (0.89) | 0.58 (0.84) |
| SS      | 0.69 (0.28) | 0.39 (0.56) | 0.78 (0.20) |
| ERP     | 0.84 (0.13) | 0.48 (0.47) | 0.87 (0.11) |

Table 3. Performance comparison with prior arts of physical attacks under different settings. Note that $p_{0.5}$ is averaged over all viewpoints of 8-pattern scheme.

| Network | 0712 |
|---------|------|
| Pose    | Standing | Walking | Sitting |
| UPC     | 0.11 (0.88) | 0.06 (0.93) | 0.56 (0.43) |
| UPC     | 0.73 (0.26) | 0.42 (0.57) | 0.86 (0.13) |
| CLS     | 0.30 (0.69) | 0.16 (0.83) | 0.65 (0.34) |
| SS      | 0.83 (0.16) | 0.47 (0.52) | 0.88 (0.11) |
| ERP     | 0.79 (0.20) | 0.44 (0.55) | 0.91 (0.08) |
Additionally, we also find that among these 3 poses “Sitting” is the most difficult to attack since some patterns (e.g., pants or cloth patterns) are partially occluded or deformed (see sampled images from Fig. 1 and Fig. 3).

We compare UPC with different existing physical attacks under the following settings (Table 3): (1) both internal deformations and external physical environments are simulated, i.e., \( T_s \) and \( T_r \) are used in Eq. 2, denoted as \( \text{UPC} \); (2) only external physical environments are modeled, i.e., \( T_r \) is used in Eq. 2, denoted as \( \text{UPC}_r \); (3) only attack the classification head, i.e., \( L_{cls} \) is used to generate patterns, denoted as \( \text{CLS} \); (4) ShapeShifter [5], i.e., only use \( T_r \) in Eq. 2 and attack against the classifier, denoted as \( SS \), and (5) we follow [33] by extending \( R^2 \) for attacking faster r-cnn, denoted as \( \text{ERP}^2 \). FR-RES101-0712/FR-VGG16-0712 are used to generate the universal camouflage patterns for these five scenarios.

From Table 3, the implications are two-fold. First, we can see the drop rates of \( \text{UPC} \) and \( \text{CLS} \) are significantly higher than those of \( \text{UPC}_r \), \( SS \) and \( \text{ERP}^2 \) (i.e., Standing: the average drop rate of \( p_{0.5} \), is 0.90/0.80 in \( \text{UPC} / \text{CLS} \)), while \( p_{0.5} \) drops 0.32/0.28/0.10 in \( \text{UPC}_r / \text{SS} / \text{ERP}^2 \) using FR-VGG16-0712. These quantitative results indicate that the proposed transformation function \( T_e \) can effectively mimic the deformations (e.g., stretching) of complex objects. Second, \( \text{UPC}_r \) and \( \text{UPC}_r \) outperform \( \text{CLS} \) and \( SS \), which suggest that the joint attack paradigm (i.e., RPN and C&R attack) generally shows stronger attacking strength than only attacking the classification head [5]. In conclusion, all these experimental results demonstrate the efficacy of the proposed transformation term \( T_e \) as well as the joint attack paradigm for fooling object detectors in the wild. Moreover, our proposed UPC significantly outperforms existing methods [5, 8] by a large margin, and thereby establish state-of-the-art for physical adversarial attack on proposal-based object detectors.

5.3. Physical Environment Experiment

Following the setup of virtual scene experiments, we stick the same camouflage pattern on different volunteers with diverse body sizes and garment styles. During the physical experiment, we use Sonya7r camera to take photos and record videos. Our physical experiments include two parts: stationary testing and walking testing.

Stationary Testing. In the physical world, we choose 5 scenes including indoors and outdoors scenes under different lighting conditions. Similar to virtual scene experiments, we take 18 photos of the attacked person for each pattern scheme. To evaluate the robustness of our method under different deformations, the person is required to switch from 6 different poses (i.e., standing, sitting, leg lifting, waving hands, fork waist, shaking head) during photographing. We record the average precision \( p_{0.5} \) and drop rates of FR-VGG16-0712 and FR-RES101-0712 under three brightening conditions in Table 4. Similar to our findings in Sec. 5.2, UPC expresses its superior attacking capability (e.g., For VGG16, Standing: \( p_{0.5} \) drops from 1.00 to 0.17, Sitting: \( p_{0.5} \) drops from 1.00 to 0.22, etc.) in the real physical world compared to natural image patterns which results in nearly zero drop rate in every posture.

As can be seen from Table 2 and Table 4, the behaviors of detectors exhibit similar trends under different physical conditions such as lighting conditions in both virtual scenes and physical environments. Another noteworthy comment is that the generated patterns from virtual scene experiments demonstrate high transferability to the real physical world (see Table 4). These facts indicate that our AttackScenes is a suitable dataset to study physical attacks.

Motion Testing. To further demonstrate the efficacy of UPC, we also test our algorithm on human motions. The list of posture classes is the same as described in Sec. 5.2. The video clips were obtained under different physical conditions (e.g., different lighting conditions, scenes) while the volunteers are walking towards the camera. Meanwhile, they are randomly changing postures from the 6 classes as mentioned above. A total of 3693 frames where 583, 377, 219, 713, 804 and 997 frames are collected under 5 different physical scenes so as to make this dataset diverse and representative. And the detection precisions are 26% (150/583), 21% (80/377), 17% (37/219), 34% (240/713), 15% (118/804) and 24% (240/997), respectively. Experiments in all physical scenes have observed low detection
5.4. Transferability Experiment

We generate camouflage patterns from one architecture to attack other models. In our experiment, FR-VGG16-0712 and FR-RES101-0712 are used to compute camouflage patterns. We introduce ResNet-50, ResNet-152 and MobileNet [10] based faster r-cnn which are trained on MS-COCO2014 [20] dataset as black-box attack models. Other architecture models including R-FCN (ResNet-101) [6], SSD (VGG-16) [21], Yolov2 [26], Yolov3 [27] and RetinaNet [19] are considered in our transferability experiments. Eight models are publicly available, and we denote them as FR-RES50-14, FR-RES152-14, FR-MN-14, RFCN-RES101-07, SSD-VGG16-0712, Yolov2-14, Yolov3-14 and Retina-14. The confidence threshold of all black-box models is set as 0.5 for evaluation.

The following 2 transfer experiments are conducted: (1) Cross-Training Transfer. The transferability between source and attacked models have the same architecture but are trained on different datasets (e.g., using the pattern generated from FR-VGG16-0712 to attack FR-VGG16-07); (2) Cross-Network Transfer. The transferability through different network structures (e.g., using the pattern computed from FR-VGG16-0712 to attack Yolov3).

During the transfer experiments, virtual walking humans with 8-Patterns scheme (see Fig. 4) are used to evaluate the transferability under black-box attacks. The transfer performances are illustrated in Table 5. The original pattern scheme is used to calculate the baseline precision of each network.
model (denoted as “Original” in Table 5). We observe that
the precisions of all detectors have dropped, which means
the generated patterns exhibits well transferability and gen-
erality across different models and datasets.

5.5. Generalization to Other Categories

To demonstrate the generalization of UPC, we construct
camouflage patterns by untargeted attacks to fool the “car”
category (i.e., rigid but non-planar object). We use Volvo
XC60 (champagne) and Volkswagen Tiguan (white) as the
attacking target in the physical world. The pattern will be
regarded as car paintings by human observers. In order to
not affect driving, we restrict the camouflage coverage re-
gions to exclude windows, lightings, and tires. We collect
120 photos which includes different distances (8 ∼ 12m)
and angles (0 ∼ 45°) in 4 different environments. The video
is recorded simultaneously at various angles (0 ∼ 45°).
Our results show 24% (29/120) images and 31% (140/453)
frames are detected as “car” correctly, which verifies the ef-
ficacy of UPC. Qualitative examples are shown in Fig. 7.

6. Discussion

Abstract Semantic Patterns. A side finding is that the
generated patterns without semantic constraint also exhibit
semantic meanings. For instance, the camouflage pattern
can be imagined as an abstract object of the target label by
human observers (Fig. 8). This observation suggest that
human and machine classification of adversarial images are
robustly related as suggested in [38].

Defense method evaluation. Our collected dataset, At-
tackScenes, can also be used for accessing the effectiveness of
defense methods against physical attacks. We hope this
data set can benefit future research in this direction.

7. Conclusion

In this paper, we study the problem of physical attacks
on object detectors. Specifically, we propose UPC to gen-
erate universal camouflage patterns which hide a category
of objects from being detected or to misdetect objects as
the target label by state-of-the-art object detectors. In ad-
dition, we present the first standardized benchmark dataset,
AttackScenes, to simulate the real 3D world in controllable
and reproducible environments. This dataset can be used
for accessing the performance of physical-world attacks at
a fair standard. Our study shows that the learned universal
camouflage patterns not only mislead object detectors in
the virtual environment, i.e., AttackScenes, but also attack
detectors successfully in the real world.

References

[1] M. Alzantot, B. Balaji, and M. Srivastava. Did you hear that?
adversarial examples against automatic speech recognition.
arXiv preprint arXiv:1801.00554, 2018. 1
[2] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok. Syn-
thesizing robust adversarial examples. arXiv preprint
arXiv:1707.07397, 2017. 1, 2
[3] A. N. Bhagoji, W. He, B. Li, and D. Song. Practical black-
box attacks on deep neural networks using efficient query
mechanisms. In European Conference on Computer Vision,
pages 158–174. Springer, 2018. 1, 2
[4] N. Carlini and D. Wagner. Audio adversarial examples: Target-
ged attacks on speech-to-text. In 2018 IEEE Security and
Privacy Workshops (SPW), pages 1–7. IEEE, 2018. 1
[5] S.-T. Chen, C. Cornelius, J. Martin, and D. H. P. Chau.
Shapeshifter: Robust physical adversarial attack on faster
r-cnn object detector. In Joint European Conference on
Machine Learning and Knowledge Discovery in Databases,
pages 52–68. Springer, 2018. 1, 2, 3, 6, 7
[6] J. Dai, L. Yi, K. He, and S. Jian. R-fcn: Object detection via
region-based fully convolutional networks. 2016. 8
[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei.
Imagenet: A large-scale hierarchical image database. 2009. 5
[8] I. Evtimov, K. Eykholt, E. Fernandes, T. Kohno, B. Li,
A. Prakash, A. Rahmati, and D. Song. Robust physical-
world attacks on deep learning models. arXiv preprint
arXiv:1705.07945, 1:1, 2017. 1, 2, 3, 6, 7
[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learn-
ing for image recognition. In Proceedings of the IEEE con-
ference on computer vision and pattern recognition, pages
770–778, 2016. 1, 6
[10] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang,
T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient
convolutional neural networks for mobile vision appli-
cations. arXiv preprint arXiv:1704.04861, 2017. 8
[11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Wein-
berger. Densely connected convolutional networks. In Pro-
cedings of the IEEE conference on computer vision and pat-
tern recognition, pages 4700–4708, 2017. 1
[12] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin. Black-box ad-
versarial attacks with limited queries and information. arXiv
preprint arXiv:1804.08598, 2018. 1, 2
[13] 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 The Thirty-Third AAAI
Conference on Artificial Intelligence (AAAI’19), 2019. 1, 2
[14] V. Khurulkov and I. Oseledets. Art of singular vectors and
universal adversarial perturbations. In Proceedings of the
IEEE Conference on Computer Vision and Pattern Recogni-
tion, pages 8562–8570, 2018. 2

Figure 8. Generated camouflage patterns are semantically
meaningful. Even for unconstrained patterns, human observer can
relate the generated camouflage patterns to the targeted label.
A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533, 2016. 1, 2

J. Li, S. Ji, T. Du, B. Li, and T. Wang. Textbugger: Generating adversarial text against real-world applications. arXiv preprint arXiv:1812.05271, 2018. 1

J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan. Scale-aware fast r-cnn for pedestrian detection. IEEE transactions on Multimedia, 20(4):985–996, 2018. 6

Y. Li, D. Tian, X. Bian, S. Lyu, et al. Robust adversarial perturbation on deep proposal-based models. arXiv preprint arXiv:1809.05962, 2018. 4

T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. Focal loss for dense object detection. In The IEEE International Conference on Computer Vision (ICCV), Oct 2017. 8

T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In European conference on computer vision, pages 740–755. Springer, 2014. 8

W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In European Conference on Computer Vision, 2016. 8

X. Liu, H. Yang, Z. Liu, L. Song, H. Li, and Y. Chen. Dpatch: An adversarial patch attack on object detectors. arXiv preprint arXiv:1806.02299, 2018. 1

A. Mahendran and A. Vedaldi. Understanding deep image representations by inverting them. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5188–5196, 2015. 5

S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1765–1773, 2017. 2

S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2574–2582, 2016. 1, 2

J. Redmon and A. Farhadi. Yolo9000: better, faster, stronger. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7263–7271, 2017. 2, 8

J. Redmon and A. Farhadi. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018. 8

S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems, pages 91–99, 2015. 2, 3

S. Samanta and S. Mehta. Towards crafting text adversarial samples. arXiv preprint arXiv:1707.02812, 2017. 1

M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter. Ac-cessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pages 1528–1540. ACM, 2016. 1, 2

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014. 6
A. Experiments on Virtual Scenes

In this section, we provide more qualitative results of FR-VGG16-0712 and FR-RES101-0712 in the synthesized virtual environments. These results are shown in Figure 10. We also show the results of targeting other categories in Figure 11.

Figure 9. More qualitative results under different attack settings in virtual experiments. Each column uses same physical conditions (i.e., lighting, viewpoints, environment, etc.). The camouflage patterns generated from $UPC_{rc}$ achieve the most superior performance and visually similar to natural image, which can be regarded as pattern designs on human accessories.
Figure 10. More qualitative results of FR-VGG16-0712 and FR-RES101-0712 on virtual environment. Each row set different virtual environments with the same viewpoint of camera, and each column uses different lighting condition.
Figure 11. **More qualitative results of targeting other categories.** Each row applies different patterns (i.e., boat/car/cat/horse), and captured in different viewpoints and background environments.
B. Experiments in Physical Environments

In Figure 12, we provide more qualitative results of FR-VGG16-0712 and FR-RES101-0712 in physical environment. In addition, videos of this experiment is available in the supplemental files.

Figure 12. More qualitative results of FR-VGG16-0712 and FR-RES101-0712 on in physical environment. These universal camouflage patterns are generated using FR-VGG16-0712 and FR-RES101-0712, respectively. Each row applies different pattern schemes (i.e., 3/7/8-Pattern schemes), and captured in different viewpoints and background environments.
C. Generalization to Other Categories

In this section, we show some qualitative results of UPC on fooling the “car” category in both virtual scenes and physical world. Video of this experiment is available in the supplemental files.

Figure 13. More qualitative results of attacking the “car” category in virtual scenes. We use two different car models (red car in top three rows and white car in bottom three rows) to evaluate the generalizability of UPC.
Figure 14. More experimental results of fooling the “car” category in physical world. We attack two different cars, i.e., Volvo XC60 and Volkswagen Tiguan.