Learning Coated Adversarial Camouflages for Object Detectors

Yexin Duan[1,2], **Jialin Chen**[3], **Xingyu Zhou**[4], **Junhua Zou**[2], **Zhengyun He**[2,5], **Jin Zhang**[1], **Wu Zhang**[2] and **Zhisong Pan**[2*]

[1]Department of Watercraft Power, Army Military Transportation University of PLA, Zhenjiang, China
[2]College of Command and Control Engineering, Army Engineering University of PLA, Nanjing, China
[3]The 28th Research Institute of China Electronics Technology Group Corporation, Nanjing, China
[4]Communication Engineering College, Army Engineering University of PLA, Nanjing, China
[5]Railway Transportation College, Hunan University of Technology, Zhuzhou, China
duanyexin0713@163.com, hotpzs@hotmail.com

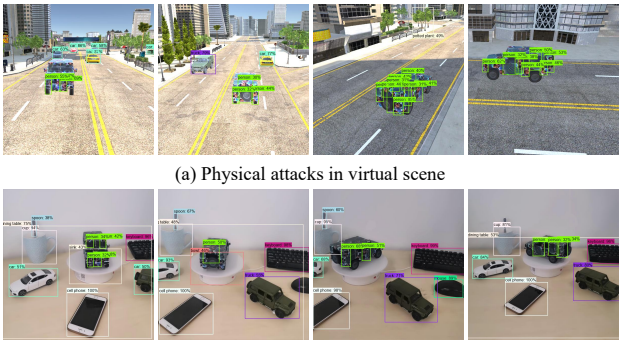
Abstract

An adversary can fool deep neural network object detectors by generating adversarial noises. Most of the existing works focus on learning local visible noises in an adversarial "patch" fashion. However, the 2D patch attached to a 3D object tends to suffer from an inevitable reduction in attack performance as the viewpoint changes. To remedy this issue, this work proposes the **Coated** Adversarial Camouflage (CAC) to attack the detectors **in arbitrary viewpoints**. Unlike the patch trained in the 2D space, our camouflage generated by a conceptually different training framework consists of 3D rendering and dense proposals attack. Specifically, we make the camouflage perform 3D spatial transformations according to the pose changes of the object. Based on the multi-view rendering results, the top-*n* proposals of the region proposal network are fixed, and all the classifications in the fixed dense proposals are attacked simultaneously to output errors. In addition, we build a virtual 3D scene to fairly and reproducibly evaluate different attacks. Extensive experiments demonstrate the superiority of CAC over the existing attacks, and it shows impressive performance both in the virtual scene and the real world. This poses a potential threat to the security-critical computer vision systems.

1 Introduction

Despite deep neural networks have achieved remarkable performance on various visual recognition tasks [Szegedy *et al.*, 2016; Redmon and Farhadi, 2018], they are found to be vulnerable to adversarial examples [Szegedy *et al.*, 2014], inputs with adversarial noises, which can fool the deep models without impeding human recognition. The adversarial attacks pose serious concerns in security-critical areas, such as medical diagnosis [Zhou *et al.*, 2019] and autonomous driving [Sitawarin *et al.*, 2018].

Attacks can be classified by the type of outcome the adversary desires: (1) non-targeted attack, the adversary's goal is to cause the deep model to predict any incorrect label; (2) targeted attack, the adversary aims to change the deep model's prediction to some specific target class, which is more challenging. In application domains, the attacks can be divided into digital attacks and physical attacks, with the latter posing a greater threat to real-world systems [Brown *et al.*, 2017]. Compared with previous works which focus on generating adversarial objects for image classifiers [Athalye *et al.*, 2018], attacking object detectors is a more realistic computer vision scenario, and is significantly harder as it requires misleading the classification results in multiple bounding boxes with different scales [Huang *et al.*, 2020].



(a) Physical attacks in virtual scene

(b) Physical attacks in the real world

Figure 1: Physical attacks (CAC) against the Faster R-CNN detector in (a) 3D virtual scene and (b) real world in free viewpoints under different brightness conditions. The adversarial vehicles are detected as the target label (*e.g.*, person). Zoom in for more details.

Most of the existing works generate local visible adversarial patches to conduct physical attacks for object detectors. However, there are several limitations: (1) the adversarial patches are trained in two-dimensional space [Thys *et al.*, 2019; Huang *et al.*, 2020], that is, during the training process, they do not carry out corresponding spatial transformations according to the pose changes of the objects, but only perform 2D plane transformations on the training images, resulting in the significant decline of the attack effectiveness when the viewpoint changes in the 3D physical world; (2) the adversarial noises are only for the planar objects, such as stop sign [Song *et al.*, 2018; Chen *et al.*, 2018], and they also

*Corresponding author

would be less effective for an object with arbitrary view angles; (3) misidentifying the adversarial object as any incorrect class, not as a specific target class [Zhang *et al.*, 2019] (*e.g.*, for autonomous driving, misidentifying an object as a designated person or stop sign is more threatening than randomly misidentifying it as a cake); (4) the missing of a unified physical evaluation environment makes it difficult to fairly evaluate the results of different attacks [Huang *et al.*, 2020]. These limitations make attacks in the 3D physical world less effective and difficult to evaluate accurately.

To address these issues, we propose the Coated Adversarial Camouflage (CAC) attack, which generates an adversarial camouflage that covers the entire object. A combination of spatial transformations is utilized to render various poses of the object as well as lighting and other natural variations to eliminate the adversarial blind spots, and enable the adversarial camouflage to mislead the detector to recognize the object as a specific target class from any viewpoint in different environments. In addition, inspired by the diverse input strategy, which optimizes an adversarial example with a set of transformed (*e.g.*, translated, resized) images, and has been proven effective to prevent the adversarial examples from overfitting to the white-box model being attacked [Xie *et al.*, 2019; Dong *et al.*, 2019], CAC fixes the top-*n* proposals of the region proposal network [Ren *et al.*, 2015] and attacks all the classifications in the dense proposals, which significantly improve the transferability of the adversarial objects. Moreover, CAC can generate camouflages for arbitrary objects.

Further, to fairly evaluate the effectiveness of different attacks, we use the Unity simulation engine to build a photo-realistic 3D urban scene with high fidelity streets, buildings, plants, *etc*. The simulation engine enables us to conduct experiments under a variety of environmental conditions: lighting, backgrounds, camera-to-object distances, view angles, *etc*. Experimental results demonstrate that the 3D coated camouflage can consistently mislead the detectors from any viewpoint, which is superior to the piecing together patches.

Figure 1 shows a sample of the generated adversarial vehicles in the physical world. We can fabricate the adversarial object by 3D printing, or simply print the adversarial texture with a color printer and paste it onto the original object. In summary, our main contributions are as follows:

- To the best of our knowledge, our work is the first to learn coated adversarial camouflages, which are trained in 3D space and can cause object detectors to misidentify objects as designated target labels from arbitrary viewpoints.
- We propose the dense proposals attack strategy to guarantee the attacking ability of the generated adversarial camouflages, especially improving the transferability of the adversarial objects under the black-box setting.
- We build a Unity simulation scene to fairly and reproducibly evaluate the effectiveness of different attacks, and extensive experiments show that CAC achieves state-of-the-art results.
- The proposed CAC can generate camouflages for any object and exhibits good generalization to the real world. In particular, an adversarial object can be fabricated by 3D printing directly, or obtained by pasting the camouflage to the surface of the original object.

2 Related Works

2.1 Digital Attacks

Digital attacks generate adversarial noises for inputs in the digital pixel domain. A series of attack methods [Szegedy *et al.*, 2014; Xie *et al.*, 2019; Dong *et al.*, 2019] have been proposed to generate adversarial examples to attack the image classifiers. Xie *et al.* [2017] extended adversarial examples from image classification to object detection, and proposed the Dense Adversary Generation (DAG) method to generate visually imperceptible perturbations to fool detectors. DAG makes the proposals very dense by increasing the original threshold of non-maximal suppression (NMS) (*e.g.* from 0.7 to 0.9) in the first stage of Faster R-CNN [Ren *et al.*, 2015], thus the proposals on each image increase from 300 to around 3000. In contrast, CAC fixes the original top-n ranked proposals after NMS, and attacks the dense classifications simultaneously in the second stage of Faster R-CNN. The noises of digital attacks are usually too subtle to be effective in the physical world due to the destructive environmental noises and input transformations [Lu *et al.*, 2017].

2.2 Physical Attacks

Physical attacks usually add visible local noises to the input so that the generated adversarial examples remain adversarial in the physical world. Several works have studied attacks on detectors. Chen *et al.* [2018] generated adversarial perturbed stop signs to fool detectors. Song *et al.* [2018] utilized synthetic transformations to attack object detection models, causing the object detectors to ignore the stop sign with sticker perturbations. Thys *et al.* [2019] and Huang *et al.* [2020] learned adversarial patterns to attack instances belonging to the same object category. However, because these local adversarial patterns are trained in two-dimensional space, they would become less effective as the viewpoint changes in the 3D physical world. Zhang *et al.* [2019] generated mosaic-like full-coverage camouflages to make a vehicle randomly misidentified as other classes. In contrast, our goal is to make vehicles consistently misidentified as specific target classes closely related to traffic safety, such as person and stop sign, which would be more threatening and challenging.

3 Methodology

3.1 Overview

We aim to generate camouflages that can fool the object detectors to misidentify the object as the target class or hide the object from being detected. We use the vehicle as an example to illustrate our method.

We simultaneously model both the object perspectives and the physical environment variations, so as to generate an adversarial object that is robust in the physical world. The transformation functions map the camouflage to a rendering of the vehicle, simulating functions including rotation, translation, perspective projection as well as lighting, background, printing errors and environmental noise changes.

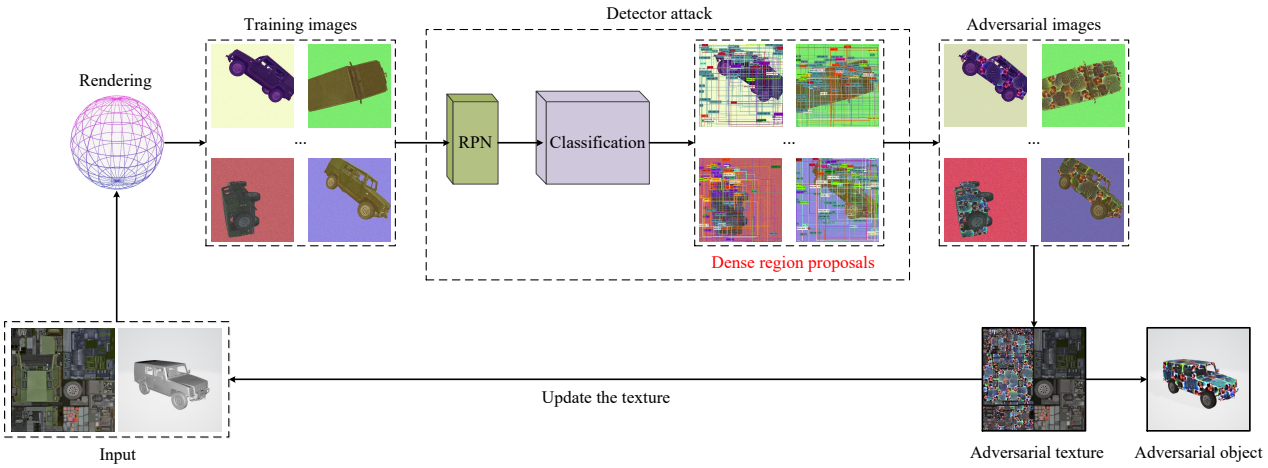


Figure 2: Overview of the pipeline to generate the adversarial camouflage texture. (1) Input transformation. The camouflage is mapped to a rendering of the vehicle, and the training images are obtained in real time by transforming viewpoints, lighting, backgrounds, printing errors, *etc.* according to a given distribution. (2) Detector attack. We simultaneously attack all the classifications in the fixed dense region proposals, and update the camouflage perturbations by minimizing the cross-entropy losses. Zoom in for more details.

Next, we attack the Faster R-CNN model [Ren *et al.*, 2015], a two-stage detector, under the white-box setting. We first run the region proposal network (RPN), and prune the region proposals by non-maximum suppression (NMS). Then we fix the top-n pruned region proposals and feed them to the second stage classification, and attack all the classifications in these proposals simultaneously to generate robust camouflages [Ren *et al.*, 2015; Chen *et al.*, 2018]. As illustrated in Figure 2, CAC mainly consists of two steps:

- Step 1. Obtaining the training images on-the-fly by simulating the 3D geometric transformations as well as the physical environment variations.
- Step 2. Attacking all the classifications in the fixed dense region proposals simultaneously.

3.2 3D Rendering

Let $(\mathbf{m}, \mathbf{c})$ be a 3D object with a mesh tensor $\mathbf{m}$ and a texture tensor $\mathbf{c}$. The training image x with ground-truth label y is the rendered result of the 3D object $(\mathbf{m}, \mathbf{c})$ with different environment conditions $e \in \mathbf{E}$ (*e.g.*, view angles, distances, backgrounds, printing errors, *etc*., the distribution detailed in Table 3 in the Appendix) from a renderer $\mathcal{R}$ by

$$x = \mathcal{R}((\mathbf{m}, \mathbf{c}), e) \tag{1}$$

We obtain the adversarial camouflage $\mathbf{c}_{adv}$ by adding perturbations to $\mathbf{c}$ and generate the adversarial example as

$$x_{adv} = \mathcal{R}((\mathbf{m}, \mathbf{c}_{adv}), e) \tag{2}$$

where $(\mathbf{m}, \mathbf{c}_{adv})$ is the obtained 3D adversarial object. Different from the 2D local adversarial patches, $\mathbf{c}_{adv}$ performs corresponding spatial transformations according to the pose changes of the object during the training process.

3.3 Dense Proposals Attack

The Faster R-CNN is a two-stage model. The region proposal network (RPN) in the first stage is a fully convolutional network that simultaneously predicts object bounds and scores at each position. The classifier in the second stage performs classification in each region proposal. Each detection includes a probability distribution over K pre-defined classes as well as the location of the detected object.

For the targeted attack, it aims to fool the deep model $f(\cdot)$ into outputting a specific target label y^*, which can be express as $f(x_{adv}) = y^*$, and $y^* \neq y$. The objective is to minimize the cross-entropy loss function $J(f(x_{adv}), y^*)$ of the classifier. We do not constrain the distance between the $\mathbf{c}_{adv}$ and the original $\mathbf{c}$, because for three-dimensional objects, the textures hardly impede human recognition. Therefore, the optimization problem for attacking the classification of an object in one proposal can be written as

$$\arg\min_{\mathbf{c}_{adv}} J(f(x_{adv}), y^*) \tag{3}$$

Some RPN proposals highly overlap with each other. To reduce redundancy, most models adopt non-maximum suppression (NMS) to prune the proposal regions based on their confidence scores [Ren *et al.*, 2015]. To improve the attacks, in each iteration, we first run the region proposal network, then fix the top-n pruned proposals. The label of each proposal is defined as the corresponding confident class. In the second classification stage, we minimize the cross-entropy losses between the target class and the predicted classes in all the fixed dense proposals. Similar to data augmentation, the objects in the fixed region proposals can be regarded as a set of transformed (*e.g.*, translated, cropped) sub-images, which can alleviate the overfitting phenomena and improve the transferability of the generated adversarial objects.

Let n be the number of the fixed proposals, and the output proposals of each image is $\mathcal{P} = \{p_i | p_i = (s_i, b_i); i = 1, 2, 3...n\}$, where s_i is the confidence score and b_i represents the location of the i-th region proposal [Huang *et al.*, 2020]. In order to enhance the attack, rather than optimize the objective function at a single point as Eq. (3), CAC simultaneously attacks the classifications of all the fixed dense

Algorithm 1 Algorithm of CAC

Input: 3D object $(\mathbf{m}, \mathbf{c})$, environment condition parameter $e \in \mathbf{E}$, neural renderer $\mathcal{R}$, target class label y^*, detector f, maximal iteration number N.
Output: Adversarial camouflage tensor $\mathbf{c}_{adv}$.

$\mathbf{c}_{adv}^0 \leftarrow \mathbf{c}$;
for $t = 0$ to $N - 1$ **do**
  Generate training images in each iteration:
  $x_{adv}^t \leftarrow \mathcal{R}((\mathbf{m}, \mathbf{c}_{adv}^t), e)$
  Obtain the dense region proposals:
  $\mathcal{P} = \{p_i | p_i = (s_i, b_i); i = 1, 2, 3...n\}$;
  Update $\mathbf{c}_{adv}^t$ via attacking all the classifications of the proposals :
  $\arg\min_{\mathbf{c}_{adv}^t} \mathbb{E}_{x\sim\mathbf{X}, e\sim\mathbf{E}}[\frac{1}{n} \sum_{p_i \in \mathcal{P}} J(f(x_{adv}^t, p_i), y^*)]$,
  $\mathbf{c}_{adv}^t = \text{Clip}(\mathbf{c}_{adv}^t, 0, 1)$;
end for
return: $\mathbf{c}_{adv} = \mathbf{c}_{adv}^N$.

region proposals to optimize the adversarial camouflage as

$$\arg\min_{\mathbf{c}_{adv}} [\frac{1}{n} \sum_{p_i \in \mathcal{P}} J(f(x_{adv}, p_i), y^*)] \tag{4}$$

Therefore, the camouflage is trained to optimize the object function

$$\arg\min_{\mathbf{c}_{adv}} \mathbb{E}_{x\sim\mathbf{X}, e\sim\mathbf{E}}[\frac{1}{n} \sum_{p_i \in \mathcal{P}} J(f(x_{adv}, p_i), y^*)] \tag{5}$$

where $\mathbf{X}$ is the training set of images generated in real time by the renderer, $\mathbf{E}$ is the distribution of the environment conditions simulated by the renderer, and $\mathbb{E}$ is the Expectation over Transformation (EOT) technique [Athalye *et al.*, 2018] which models the adversarial perturbations within the optimization procedure. The "true" input $\mathcal{R}((\mathbf{m}, \mathbf{c}_{adv}), e)$ perceived by the detector $f(\cdot)$ is the input object $(\mathbf{m}, \mathbf{c}_{adv})$ with environment condition e after the render process. It optimizes the losses between the expected detection results and the target class y^*. The resultant camouflage pixel value is clipped to the valid range (*i.e.*, [0,1] for images). The overall procedure of CAC is summarized in Algorithm 1.

4 Experiments

4.1 Experimental Settings

Source Model. We generate adversarial camouflage on the Faster R-CNN with Inception-v2 [Szegedy *et al.*, 2016] as the backbone network. The model is trained on the COCO2014 dataset [Lin *et al.*, 2014]. We denote this model as FR-Incv2-14. Faster R-CNN adopts a two-stage detection strategy, the first stage generates many region proposals that may contain objects, and the second stage outputs the classification results and the refined bounding box coordinates.

Target Models. To evaluate the cross-model and cross-training transferability of the adversarial camouflage generated by the source model, we test the results on two-stage and one-stage detectors with different network structures and training datasets. For two-stage detectors, we evaluate the performance of the Faster R-CNN model, in addition to Inception-v2, the backbone networks include VGG16 [Simonyan and Zisserman, 2015] and ResNet101 [He *et al.*, 2016], which are either trained on the PascalVOC-2007 trainval set or the combined PascalVOC-2007 and PascalVOC-2012 trainval set [Everingham *et al.*, 2015], or on COCO2014. We denote these models as FR-VGG16-07, FR-RES101-07, FR-VGG16-0712, FR-RES101-0712, FR-VGG16-14 and FR-RES101-14, respectively. For one-stage detectors, we select YOLOv3 [Redmon and Farhadi, 2018] and YOLOv5 [1], whose network structures are quite different from Faster R-CNN, making it more challenging to conduct transfer attacks, and we denote these models trained on COCO2014 as YOLOv3-14 and YOLOv5-14, respectively.

Evaluation Metrics. The principal quantitative measure of the detection task is the average precision. Detections are considered true or false positives based on the area Intersection over Union (IoU) between the predicted box B_p and the ground truth box B_{gt}, which is defined as $IoU = \frac{B_p \cap B_{gt}}{B_p \cup B_{gt}}$. The threshold of IoU is set to 0.5 as in the PASCAL VOC detection challenge [Everingham *et al.*, 2015] to determine whether the detector hits or misses the true category. This metrics is denoted as P@0.5 [Zhang *et al.*, 2019; Huang *et al.*, 2020]. The confidence score threshold of all models is set to 0.3 for evaluation. (Since the jeep vehicle with the original texture is detected as a truck or a car, we consider both classes to be "true" classes).

Baselines. We generate three simple textures for comparison, *Natural*, *Naive* and *Random*. In addition, we compare our method to several state-of-the-art physical attacks. Adversarial patch (*AdvPat*) [Thys *et al.*, 2019] generates adversarial patches to fool a detector by minimizing different probabilities related to the appearance of a object. Shapeshifter (*Shape*) [Chen *et al.*, 2018] obtains adversarial objects that mislead a detector via EOT technique [Athalye *et al.*, 2018]. *UPC* [Huang *et al.*, 2020] crafts adversarial camouflages for non-rigid or non-planar objects. *Camou* [Zhang *et al.*, 2019] learns camouflages that can hide a vehicle from detectors. For fair comparison, the adversarial patterns generated by *AdvPat*, *Shape* and *UPC* are pasted on all sides of the vehicles. We restrict the CAC camouflage to the body of the vehicle, leaving wheels, windows, lightings, *etc*. unaltered as the discriminative visual cues for the detectors. Nine different textures are shown in Figure 7 in the Appendix.

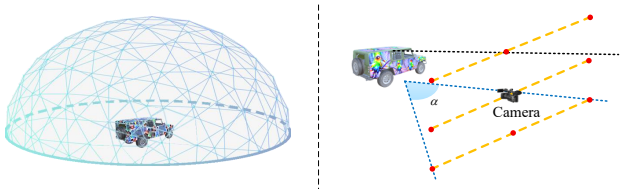
Physical Simulation. To fairly and reproducibly evaluate different attacks, we build a photo-realistic 3D simulation scene (see Figure 8 in the Appendix).

4.2 Virtual Scene Experiments

The area light and directional light sources are used to get 2 levels of brightness (*i.e.*, bright:L_b and dark:L_d) simulation scenes (illustrated in Figure 1 (a)).

As illustrated in Figure 3 (a), the view angles and distances of the cameras distribute freely as *Free Viewpoint (FV)* within a semi-ellipsoid, so that the attack effectiveness can be evaluated more accurately and comprehensively.

[1] https://github.com/ultralytics/yolov5



(a) Free Viewpoint (FV) (b) Restricted Viewpoint (RV)

Figure 3: The camera settings. (a) The camera is placed arbitrarily within a semi-ellipsoid as free viewpoint; (b) the camera is restricted to a certain range of view angles (*e.g.*, *UPC* with restricted view angle range α in its original paper).

We rendered 400 (200×2) images with free viewpoints in different locations of the simulation scene for each of the nine vehicles with different textures. The target class of CAC attack is "person".

Denseness of Proposals

We first study the impact of the proposal denseness on attacks. Different number of proposals are fixed after the RPN, ranging from 50 to 300. As can be seen in Figure 4, the average precision goes down as the top-n number increases, which indicates that denser proposals obtain stronger adversarial camouflages. Therefore, we choose a large number (300) that performs better.

Comparison with Simple Textures

The average precision results for models trained on the COCO2014 dataset are shown in Table 1, and for models trained on PascalVOC datasets are shown in Table 4 in the Appendix due to the space limitations. We can find that for the Natural/Naive/Random textures, the average precisions of almost all models drop little, indicating that simple textures hardly affect the detectors, *i.e.*, the detectors are robust to the simulation data, even for the noisy random texture. In contrast, the camouflage generated by CAC significantly reduces the average precisions under both the bright and dark brightness conditions, which verifies the effectiveness of the attack.

In addition, we can observe that for the simple textures and the adversarial textures, most of the attacks in the dark environment are stronger than those in the bright environment, which may be attributed to the fact that the robustness of the detector is relatively weak when the brightness level is low.

Comparison with Existing Attacks

Except for the *Camou*, the adversarial patterns of *AdvPat*, *Shape* and *UPC* are local patches on different sides of the vehicles (see Figure 7 in the Appendix). The patches are trained in 2D space and evaluated from a restricted view (*e.g.*, side view) in their original papers. We denote the side view as *Restricted Viewpoint (RV)* (shown in Figure 3 (b), α is set to 120°, *i.e.*, the left and right view angles are within 60°). However, the cameras in the wild do not film the vehicle object only from the side, so we evaluate the attacks under both the free viewpoint and restricted viewpoint settings.

Table 2 shows the physical attack results for models trained on the COCO2014 dataset, and the results for models trained on PascalVOC datasets are shown in Table 5 in the Appendix. It can be found that the local adversarial patterns can also

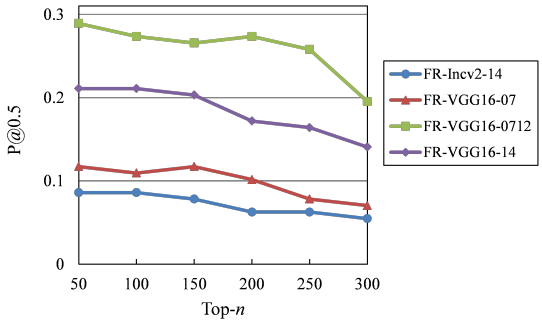


Figure 4: The average precision P@0.5 of using adversarial camouflages generated on FR-Incv2-14 to attack FR-VGG16-07, FR-VGG16-0712 and FR-VGG16-14, with respect to the top-n number.

reduce the average precision when pasted on other sides of the vehicles. However, the adversarial camouflage generated by CAC exhibits significantly better performance, and has much higher drop rates both in the restricted and free viewpoints than the baseline methods. The results demonstrate that the attack effectiveness of our coated camouflage outperforms the piecing-together patches. Besides, CAC is also superior to the full-coverage *Camou* texture, which further demonstrates the cross-model transferability of CAC. It worth to note that *Camou* conducts non-targeted attacks, while CAC conducts targeted attacks, which are more challenging.

We find that the specific target class label is well transferred, *i.e.*, the target models also misidentify the vehicles with our adversarial camouflage trained on the source model as the person class label. This is because different models have similar decision boundaries due to the similar or same training dataset [Xie *et al.*, 2019]. See Figures 9 to 12 in the Appendix for more details.

Moreover, it can be seen that the average precision in restricted viewpoint is usually higher than that in free viewpoint (*i.e.*, lower drop rates), indicating that the attacks are more difficult to succeed from the side view.

4.3 Physical Environment Experiments

We evaluate physical attacks in the real world over a 3D-printed vehicle with the adversarial camouflage generated by CAC. We take photos and record videos with an HONOR 20 cell phone. Similar to the virtual scene experiments, the 3D-printed vehicle is filmed in free viewpoints. We put the vehicle on a rotating turntable to eliminate blind spots. Figure 14 shows some qualitative results, and a demo video can be found at: https://www.bilibili.com/video/BV1zL411J73r.

In Figure 5 we plot the relationship between the camera-to-object distance and average precision of FR-Incv2-14 under the 360° free viewpoint setting. The abscissa is the ratio of distance to the vehicle length. It can be seen that the average precision goes up as the distance increases, this is probably because the camouflage is captured with lower quality from a distance. The results show low detection precisions, which demonstrate that the 3D-printed adversarial vehicle is strongly adversarial over a variety of viewpoints, and the camouflage generated by the virtual scene experiments exhibits high generalization to the real world.

Model	FR-Incv2-14			FR-VGG16-14			FR-RES101-14			YOLOv3-14			YOLOv5-14		
Scheme	L_b	L_d	Avg (Drop)	L_b	L_d	Avg (Drop)	L_b	L_d	Avg (Drop)	L_b	L_d	Avg (Drop)	L_b	L_d	Avg (Drop)
Original	1.00	1.00	1.00 (-)	0.96	0.98	0.97 (-)	0.98	0.99	0.98 (-)	1.00	1.00	1.00 (-)	1.00	1.00	1.00 (-)
Natural	1.00	1.00	1.00 (0)	0.90	0.92	0.91 (0.06)	0.91	0.93	0.92 (0.06)	0.93	0.90	0.92 (0.08)	0.93	0.96	0.95 (0.05)
Naive	1.00	0.94	0.97 (0.03)	0.93	0.83	0.88 (0.09)	0.92	0.94	0.93 (0.05)	0.93	0.86	0.90 (0.10)	0.92	0.95	0.94 (0.06)
Random	0.90	0.93	0.91 (0.09)	0.86	0.84	0.85 (0.12)	0.92	0.90	0.91 (0.07)	0.84	0.76	0.80 (0.20)	0.90	0.88	0.89 (0.11)
CAC	0.08*	0.04*	0.06 (**0.94**)	0.18	0.10	0.14 (**0.83**)	0.59	0.47	0.53 (**0.45**)	0.21	0.25	0.23 (**0.77**)	0.54	0.44	0.49 (**0.51**)

Table 1: Average precision P@0.5 and drop rates in virtual scene experiments under two illumination level settings. Each P@0.5 is averaged over free viewpoints. * indicates the white-box attacks. The best results are in bold.

Model	FR-Incv2-14		FR-VGG16-14		FR-RES101-14		YOLOv3-14		YOLOv5-14	
Scheme	RV (Drop)	FV (Drop)	RV (Drop)	FV (Drop)	RV (Drop)	FV (Drop)	RV (Drop)	FV (Drop)	RV (Drop)	FV (Drop)
Original	1.00 (-)	1.00 (-)	1.00 (-)	0.97 (-)	1.00 (-)	0.98 (-)	1.00 (-)	1.00 (-)	1.00 (-)	1.00 (-)
AdvPat	0.92 (0.08)	0.82 (0.18)	0.97 (0.03)	0.73 (0.24)	0.99 (0.01)	0.77 (0.21)	0.33* (0.67)	0.30* (0.70)	0.70 (0.30)	0.59 (0.41)
Shape	0.99* (0.01)	0.85* (0.15)	0.99 (0.01)	0.86 (0.11)	0.99 (0.01)	0.85 (0.13)	0.95 (0.05)	0.74 (0.26)	0.96 (0.04)	0.78 (0.22)
UPC	0.95 (0.05)	0.74 (0.26)	0.79 (0.21)	0.56 (0.41)	0.93 (0.07)	0.82 (0.16)	0.32 (0.68)	0.34 (0.66)	0.69 (0.31)	0.57 (0.43)
Camou	0.97 (0.03)	0.55 (0.45)	0.94 (0.06)	0.52 (0.45)	0.99 (0.01)	0.69 (0.29)	0.51* (0.49)	0.28* (0.72)	0.94 (0.06)	0.64 (0.36)
CAC	0.08* (**0.92**)	0.06* (**0.94**)	0.22 (**0.78**)	0.14 (**0.83**)	0.85 (**0.15**)	0.53 (**0.45**)	0.29 (**0.71**)	0.23 (**0.77**)	0.67 (**0.33**)	0.49 (**0.51**)

Table 2: Average precision P@0.5 and drop rates of different attacks in restricted and free viewpoints. Each P@0.5 is averaged over two brightness levels. * indicates the white-box attacks. The best results are in bold.

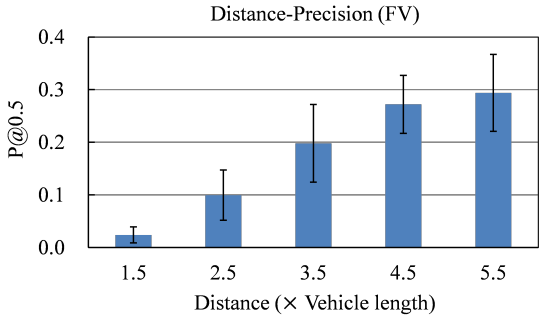


Figure 5: The P@0.5 (with $\pm std$ over 5 tests) under different distance conditions in the real world.

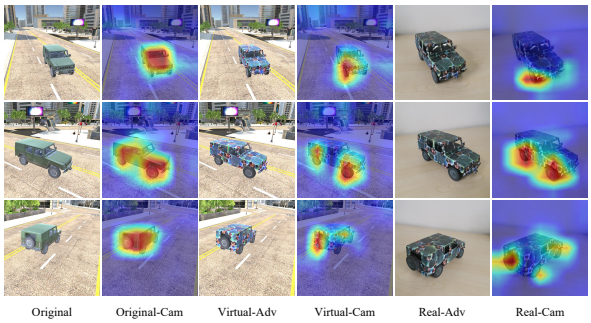


Figure 6: Visualization of discriminative regions for "jeep" class label in the virtual scene and the real world. For the vehicles with the adversarial camouflage, the model places its attention predominantly on the regions with no adversarial perturbations (*e.g.*, wheels).

4.4 Model Attention Analysis

The regions that the models pay attention to can be deemed as the discriminative regions. We generate the attention maps of the vehicle with different viewpoints on VGG16 model by the model-agnostic Grad-CAM [Selvaraju *et al.*, 2017] technique. Figure 6 shows the original vehicle, virtual adversarial vehicle (Virtual-Adv), 3D printed adversarial vehicle (Real-Adv) and their attention maps for the "jeep" class label, respectively. We can observe the CAC attack distracts the attention maps from the vehicle body and focuses them in the wheel areas where there is no adversarial noise. This also explains why restricted side-view attacks are more difficult than free-view attacks in Table 2, as the noiseless wheels are a very salient vehicle distinguishing feature in the side view.

4.5 Attacks for Other Objects and Classes

To demonstrate the generalization of CAC, we generate adversarial camouflages to fool other objects. As shown in Figure 15 in the Appendix, the barrels and containers with adversarial camouflages can be misidentified as the target label in different viewpoints, indicating that CAC is generic for any object beyond vehicles. In addition, we generate other target class (*e.g.*, stop sign) camouflages closely related to traffic safety for the vehicle, the qualitative results in Figure 16 in the Appendix further prove that CAC can fool the object detector into outputting arbitrary specific class labels.

5 Conclusion

In this paper, we investigate the problem of generating robust 3D adversarial camouflages in the physical world for object detectors. By modeling the 3D rendering, and using a set of dense proposals to optimize the adversarial camouflage in each iteration, the vehicle with the resultant adversarial camouflage generated by the proposed CAC can fool object detectors from any viewpoint, and exhibits significantly better performance than the state-of-the-art baseline methods. In addition, we build a 3D scene to fairly and reproducibly evaluate different attacks. With the 3D printing technology, we successfully fabricate the first physical adversarial object that is detected as a specific target class under arbitrary viewpoints and different lighting conditions. For future work, we plan to make the camouflages visually more natural.

Acknowledgments

This work was supported by National Natural Science Foundation of China under Grant 62076251.

References

[Athalye *et al.*, 2018] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pages 284–293. PMLR, 2018.

[Brown *et al.*, 2017] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.

[Chen *et al.*, 2018] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Polo 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.

[Dong *et al.*, 2019] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4312–4321, 2019.

[Everingham *et al.*, 2015] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Huang *et al.*, 2020] Lifeng Huang, Chengying Gao, Yuyin Zhou, Cihang Xie, Alan L Yuille, Changqing Zou, and Ning Liu. Universal physical camouflage attacks on object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 720–729, 2020.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

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

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

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015.

[Selvaraju *et al.*, 2017] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[Simonyan and Zisserman, 2015] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR*, 2015.

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

[Song *et al.*, 2018] Dawn Song, Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Florian Tramer, Atul Prakash, and Tadayoshi Kohno. Physical adversarial examples for object detectors. In *12th {USENIX} Workshop on Offensive Technologies ({WOOT} 18)*, 2018.

[Szegedy *et al.*, 2014] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR*, 2014.

[Szegedy *et al.*, 2016] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[Thys *et al.*, 2019] Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Fooling automated surveillance cameras: adversarial patches to attack person detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[Xie *et al.*, 2017] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1369–1378, 2017.

[Xie *et al.*, 2019] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2730–2739, 2019.

[Zhang *et al.*, 2019] Yang Zhang, Hassan Foroosh, Philip David, and Boqing Gong. Camou: Learning physical vehicle camouflages to adversarially attack detectors in the wild. In *International Conference on Learning Representations*, 2019.

[Zhou *et al.*, 2019] Yi Zhou, Xiaodong He, Lei Huang, Li Liu, Fan Zhu, Shanshan Cui, and Ling Shao. Collaborative learning of semi-supervised segmentation and classification for medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2079–2088, 2019.