Dual Attention Suppression Attack: Generate Adversarial Camouflage in Physical World

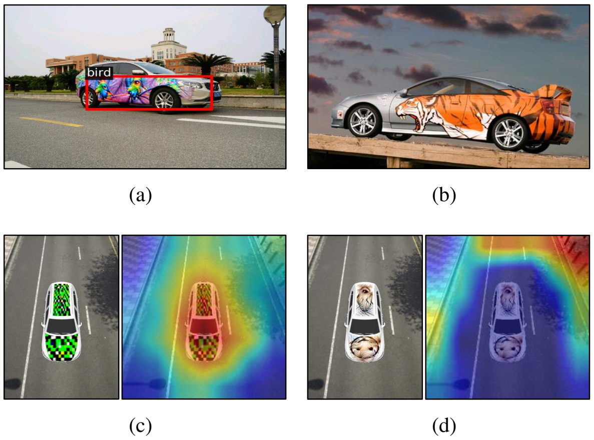
Jiakai Wang , Aishan Liu, Zixin Yin, Shunchang Liu ,
Shiyu Tang, and Xianglong Liu[*]
State Key Lab of Software Development Environment,
Beihang University, Beijing, China
{jk_buaa_scse, liuaishan, yzx835, liusc, sytang, xlliu}@buaa.edu.cn

Abstract

Deep learning models are vulnerable to adversarial examples. As a more threatening type for practical deep learning systems, physical adversarial examples have received extensive research attention in recent years. However, without exploiting the intrinsic characteristics such as model-agnostic and human-specific patterns, existing works generate weak adversarial perturbations in the physical world, which fall short of attacking across different models and show visually suspicious appearance. Motivated by the viewpoint that attention reflects the intrinsic characteristics of the recognition process, this paper proposes the Dual Attention Suppression (DAS) attack to generate visually-natural physical adversarial camouflages with strong transferability by suppressing both model and human attention. As for attacking, we generate transferable adversarial camouflages by distracting the model-shared similar attention patterns from the target to non-target regions. Meanwhile, based on the fact that human visual attention always focuses on salient items (e.g., suspicious distortions), we evade the human-specific bottom-up attention to generate visually-natural camouflages which are correlated to the scenario context. We conduct extensive experiments in both the digital and physical world for classification and detection tasks on up-to-date models (e.g., Yolo-V5) and demonstrate that our method outperforms state-of-the-art methods.[1]

1. Introduction

Deep neural networks (DNNs) have achieved remarkable performance across a wide areas of applications, *e.g.*, computer vision [24, 35], natural language [42], and acoustics [34], *etc*, but they are vulnerable to *adversarial examples* [44, 36]. These elaborately designed perturbations are imperceptible to humans but can easily lead DNNs to wrong predictions, which pose a strong security challenge to deep learning applications in both the digital and physical world [22, 13, 31, 37, 51].



(a) (b)
(c) (d)

Figure 1. (a) shows the suspicious appearance of camouflages generated by previous work (*i.e.*, UPC [19]). (b) is the painted car that commonly exists in the physical world. (c) shows the adversarial example (classified as `pop bottle`) generated by existing work (*i.e.*, CAMOU [52]) and its corresponding attention map. (d) shows the adversarial example (classified as `Shih-Tzu`) generated by our DAS and its distracted attention map.

In the past years, a long line of work has been proposed to perform adversarial attacks in different scenarios under different settings [26, 7, 2]. Though challenging deep learning, adversarial examples are also valuable for understanding the behaviors of DNNs, which could provide insights into the blind-spots and help to build robust models [20, 45, 28, 50]. Generally, adversarial attacks can be divided into two categories: *digital attacks*, which attack DNNs by perturbing the input data in the digital space; and *physical attacks*, which attack DNNs by modifying the vi-

[*] Corresponding author
[1] Our code can be found in https://github.com/nlsde-safety-team/DualAttentionAttack.

sual characteristics of the real object in the physical world. In contrast to the attacks in the digital world [23, 48, 21, 52], adversarial attacks in the physical world are more challenging due to the complex physical constraints and conditions (*e.g.*, lighting, distance, camera, *etc.*), which will impair the attacking ability of generated adversarial perturbations [12]. In this paper, we mainly focus on the more challenging physical world attack task, which is also more meaningful to the deployed deep learning applications in practice.

Though several attempts have been adopted to perform physical attacks [31, 19, 30], existing works always ignore the intrinsic characteristics such as model-agnostic and human-specific patterns so that their attacking abilities are still far from satisfactory. In particular, the limitations can be summarized as (1) the existing methods ignore the common patterns among models and generate adversarial perturbations using model-specific clues (*e.g.*, gradients and weights of a specific model), which fails to attack across different target models. In other words, the transferability of adversarial perturbations is weak, which impairs their attacking abilities in the physical world; (2) current methods generate adversarial perturbations with a visual suspicious appearance which is poorly aligned with human perception and even attracts the human attention. For example, painted on the adversarial camouflage [19], the classifier misclassifies the car into a bird. However, as shown in Figure 1(a), the camouflage apparently contains un-natural and suspicious bird-related features (*e.g.*, bird head), which attracts human attention.

To address the mentioned problems, this paper proposes the Dual Attention Suppression (DAS) attack by suppressing both the model and human attention. Regarding the **transferability for attacks**, inspired by the biological observation that cerebral activities between different individuals share similar patterns when stimulus features are encountered [49] (*i.e.*, selected attention [27]), we perform adversarial attacks by suppressing the attention patterns shared among different models. Specifically, we distract the model-shared similar attention from target to non-target regions via connected graphs. Thus, target models will be misclassified by not paying attention to the objects in the target region. Since our generated adversarial camouflage captures model-agnostic structures, it can transfer among different models, which improves the transferability.

As for the **visual naturalness**, psychologists have found that the bottom-up attention of human vision will alert people to salient objects (*e.g.*, distortion) [6]. Existing methods generate physical adversarial examples with visually suspicious appearance, which shows salient features to human perception. Thus, we try to evade this human-specific visual attention by generating adversarial camouflage which contains high semantic correlation to scenario context. As a result, the generated camouflage is more unsuspicious and natural in terms of human perception. Figure 1(c) is the adversarial camouflage generated by CAMOU [52] which is suspicious to human vision. By contrast, our generated adversarial camouflage yields a more natural appearance as shown in Figure 1(d).

To the best of our knowledge, we are the first to exploit the shared attention characteristics among models and generate adversarial camouflages world by suppressing both the model and human attention. Extensive experiments in both the digital and physical world on both classification and detection tasks are conducted which demonstrate that our method outperforms other state-of-the-art methods.

2. Related Works

Adversarial examples are elaborately designed perturbations which are imperceptible to human but could mislead DNNs [44, 22]. In the past years, a long line of work has been proposed to develop adversarial attack strategies [25, 13, 30, 46, 11, 29, 52, 19]. In general, there are several different ways to categorize adversarial attack methods, *e.g.*, targeted or untargeted attacks, white-box or black-box attacks, *etc.* Based on the domain in which the adversarial perturbations are generated, adversarial attacks can be divided into digital attacks and physical attacks.

Digital attacks generate adversarial perturbations for input data in the digital pixel domain. Szegedy *et al.* [44] first introduced adversarial examples and used the L-BFGS method to generate them. By leveraging the gradients of target models, Goodfellow *et al.* proposed the Fast Gradient Sign Method (FGSM) [22] which could generate adversarial examples quickly. Moreover, Madry *et al.* [1] proposed Projected Gradient Decent (PGD), which is currently the strongest first-order attack. Based on the gradient information, a series of attack approaches have been proposed [25, 8, 48, 9]. Although these attacks achieve substantial results in the digital world, their attacking abilities degenerate significantly when introduced into the physical world.

On the other hand, physical attacks aim to generate adversarial perturbations by modifying the visual characteristics of the real object in the physical world. To achieve the goal, several works first generate adversarial perturbations in the digital world, then perform physical attacks by painting the adversarial camouflage on the real object or directly create the perturbed objects. By constructing a rendering function, Athalye *et al.* [2] generated 3D adversarial objects in the physical world to attack classifiers. Eykholt *et al.* [13] introduced NPS [33] into the loss function which considers the fabrication error so that they can generate strong adversarial attacks for traffic sign recognition. Recently, Huang *et al.* [19] proposed the Universal Physical Camouflage Attack (UPC), which crafts camouflage by jointly fooling the region proposal network and the classifier. Another line of work tries to perform physical adversarial attacks by generating adversarial patches [3], which confine the noise to a small and localized patch without perturbation constraint [30, 31].

3. Approach

In this section, we first provide the definition of the problem and then elaborate on our proposed framework.

3.1. Problem Definitions

Given a deep neural network $\mathbb{F}_\theta$ and an input clean image $\mathbf{I}$ with the ground truth label y, an adversarial example $\mathbf{I}_{adv}$ in the **digital world** can make the model conduct wrong predictions as follows:

$$\mathbb{F}_\theta(\mathbf{I}_{adv}) \neq y \quad s.t. \quad \|\mathbf{I} - \mathbf{I}_{adv}\| < \epsilon, \tag{1}$$

where $||\cdot||$ is a distance metric to quantify the distance between the two inputs $\mathbf{I}$ and $\mathbf{I}_{adv}$ sufficiently small.

In the **physical world**, let $(\mathbf{M}, \mathbf{T})$ denote a 3D real object with a mesh tensor $\mathbf{M}$, a texture tensor $\mathbf{T}$, and ground truth y. The input image $\mathbf{I}$ for a deep learning system is the rendered result of the real object $(\mathbf{M}, \mathbf{T})$ with environmental condition $c \in \mathbf{C}$ (*e.g.*, camera views, distance, illumination, *etc.*) from a renderer $\mathcal{R}$ by $\mathbf{I} = \mathcal{R}((\mathbf{M}, \mathbf{T}), c)$. To perform physical attacks, we generate $\mathbf{I}_{adv} = \mathcal{R}((\mathbf{M}, \mathbf{T}_{adv}), c)$ through replacing the original $\mathbf{T}$ with an adversarial texture tensor $\mathbf{T}_{adv}$, which has different physical properties (*e.g.*, color, shape). Thus the definition of our problem can be depicted as:

$$\mathbb{F}_\theta(\mathbf{I}_{adv}) \neq y \quad s.t. \quad \|\mathbf{T} - \mathbf{T}_{adv}\| < \epsilon, \tag{2}$$

where we ensure the naturalness of the generated adversarial camouflage in the physical world by ϵ.

In this paper, we mainly discuss adversarial attacks in the physical world and generate an adversarial camouflage (*i.e.*, texture), which is able to fool the real deep learning systems when it is painted or overlaid on a real object.

3.2. Framework Overview

To generate visually-natural physical adversarial camouflage with strong transferability, we propose the Dual Attention Suppression (DAS) framework which suppresses both the model and human attention. The overall framework can be found in Figure 2.

Regarding the **transferability for attack**, inspired by the biological observation, we suppress the similar attention patterns shared among models. Specifically, we generate adversarial camouflage by distracting the model attention from target to non-target regions (*e.g.*, background) via connected graphs. Since different deep models yield similar attention patterns towards the same object, our generated adversarial camouflage could capture the model-agnostic structures and transfer to different models.

As for the **visual naturalness**, we aim to evade the human-specific bottom-up attention in human vision [6] by generating visually-natural camouflage. By introducing a seed content patch $\mathbf{P}_0$, which has a strong perceptual correlation to the scenario context, the generated adversarial camouflage in this case can be more unsuspicious and natural to human perception. Since humans pay more attention to object shapes when making predictions [29], we further preserve the shape information of the seed content patch to improve the human attention correlations. Thus, the human-specific attention mechanism is evaded, leading to more natural camouflage.

3.3. Model Attention Distraction

Biologists have found that the same stimulus features (*i.e.*, selected attention) yield similar patterns of cerebral activities among different individuals [49] (*i.e.*, similar characteristics of the neuron hyper-perception). Since artificial neural networks are implemented from the human central nervous system [16], it is also reasonable for us to assume that DNNs may have the same characteristics, *i.e.*, different models have similar attention patterns towards the same objects when making the same predictions. Based on the above observations, we consider improving the transferability of adversarial camouflages by capturing the model-agnostic attention structures.

Visual attention techniques [53] have been long studied to improve the explanation and understanding of deep learning behaviors, such as CAM [53], Grad-CAM [40], and Grad-CAM++ [5]. When making predictions, a model pays most of its attention to the target objects rather than meaningless parts. Intuitively, to successfully attack a model, we directly distract the model attention from the salient objects. In other words, we distract the model-shared similar attention map on the salient area to other regions and force the attention weights to distribute uniformly through the entire image. Thus, the model may fail to focus on the target object and make the wrong predictions.

Specifically, given an object $(\mathbf{M}, \mathbf{T})$, an adversarial texture tensor $\mathbf{T}_{adv}$ to be optimized, and a certain label y, we get $\mathbf{I}_{adv}$ by $\mathcal{R}$ and then compute the attention map $\mathbf{S}^y$ with an attention module $\mathcal{A}$ as

$$\mathbf{S}^y = \mathcal{A}(\mathbf{I}_{adv}, y). \tag{3}$$

More precisely, the attention module $\mathcal{A}$ is

$$\mathcal{A}(\mathbf{I}, y) = relu(\sum_k \sum_i \sum_j \alpha_{ij}^{ky} \cdot relu(\frac{\partial p^y}{\partial A_{ij}^k}) \cdot A_{ij}^k), \tag{4}$$

where α_{ij}^{ky} is the gradient weights for a particular class y and activation map k, p^y is the score of the class y, A_{ij}^k

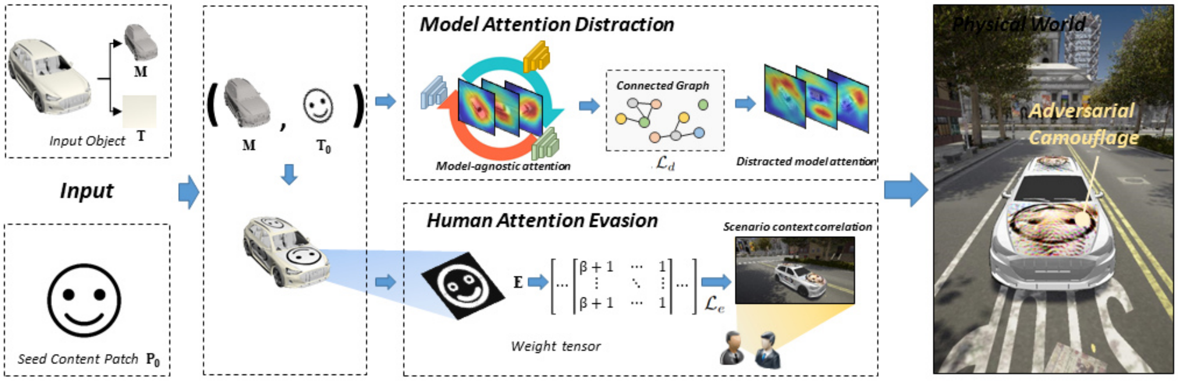


Figure 2. The framework of our DAS method. We first distract the intrinsic attention characteristic through fully exploiting the similar attention patterns of models and forcing the "heat" regions away from the target object with loss function $\mathcal{L}_d$. Then we evade the human-specific visual attention mechanism by correlating the appearance of adversaries to the context scenario and preserving the shape information of seed content image to generate visually-natural adversarial camouflage.

is the pixel value in position (i, j) of the k-th feature map, and *relu*$(\cdot)$ denotes the *relu* function. Note that the attention module can be an arbitrary deep learning model rather than the target model.

Given the attention map $\mathbf{S}^y$ calculated by Eqn 3, we aim to distract the attention region and force the model to focus on non-target regions. Intuitively, the pixel value of the attention map represents to what extent the region contributes to model predictions. To decrease the attention weights of the salient object and disperse these attention regions, we exploit the *connected graph*, which contains a path between any pair of nodes within the graph. In an image, a region with attention weights for each pixel higher than a specific threshold can be deemed as a connected region. To distract the model attention using the connected graph, we consider the following two tasks: (1) decrease the overall connectivity by separating connected graphs into multiple sub-graphs; (2) reduce the weight of each node within a connected sub-graph. To achieve these goals, we propose attention distraction loss as

$$\mathcal{L}_d = \frac{1}{K}\sum_k \frac{G_k}{N - N_k}, \quad s.t. \quad G_k \subseteq \mathbf{S}^y, \tag{5}$$

where G_k is the sum of pixel values in the region corresponding to k-th connected graph in $\mathbf{S}^y$, N is the total pixel number of the $\mathbf{S}^y$, and N_k is the total pixel number of G_k. By minimizing $\mathcal{L}_d$, the salient region in the attention map becomes smaller (*i.e.*, distracted) and the pixel values of the salient regions become lower (*i.e.*, no longer "heated"), leading to the "distracted" attention map.

3.4. Human Attention Evasion

To overcome the problem brought by the complex environmental conditions in the physical world, most physical attacks generate adversarial perturbations with a comparatively huge magnitude [11]. Since the bottom-up human attention mechanism always alerts people to salient objects (*e.g.*, distortion) [6], adversarial examples in this case can always attract human attention due to the salient perturbations, showing suspicious appearance and lower stealthiness in the physical world.

In this paper, we aim to generate more visually-natural camouflage by suppressing the human visual mechanism, which will evade human-specific attention. Intuitively, we expect the generated camouflage to share similar visual semantics with the context to be attacked (*e.g.*, beautiful paintings on vehicles are more perceptually acceptable for humans than meaningless distortions). Thus, the generated adversarial camouflage can be highly correlated to human perception, which is unsuspicious to human perception.

In particular, we first incorporate a seed content patch $\mathbf{P}_0$ which contains a strong semantic association with the scenario context. We then paint the seed content patch on the vehicle $(\mathbf{M}, \mathbf{T})$ by $\mathbf{T}_0 = \Psi(\mathbf{P}_0, \mathbf{T})$. Specifically, $\Psi(\cdot)$ is a transformation operator which first transfers the 2D seed content patch into a 3D tensor, and then paint the car through tensor addition.

Since humans pay more attention to shapes when focusing on objects and making predictions [29], we aim to further improve the human attention correlation by better preserving the shape of the seed content patch. Specifically, we obtain the edge patch $\mathbf{P}_{edge} = \mathbf{\Phi}(\mathbf{P}_0)$ using an edge extractor $\mathbf{\Phi}$ [4] from the seed content patch. It should be noticed that $\mathbf{P}_{edge}$ has 0-1 value in each pixel. After that, we simply transform the edge patch $\mathbf{P}_{edge}$ to a mask tensor $\mathbf{E}$ which has the same dimension with $\mathbf{T}_0$.

With mask tensor $\mathbf{E}$, we can distinguish the edge and non-edge regions and limit the perturbations added to the edge regions. Thus, the attention evasion loss $\mathcal{L}_e$ can be formulated as

$$\mathcal{L}_e = \|(\beta \cdot \mathbf{E} + \mathbf{1}) \odot (\mathbf{T}_{adv} - \mathbf{T}_0)\|_2^2, \tag{6}$$

where the $\beta \cdot \mathbf{E} + \mathbf{1}$ is the weight tensor, the $\mathbf{1}$ is a tensor in which each element is 1 and its dimension is same with $\mathbf{E}$ and $\odot$ denotes the element-wise multiplication.

To further improve the naturalness of the camouflage, we introduce the smooth loss [13] by reducing the difference square between adjacent pixels. For a rendered adversarial image $\mathbf{I}_{adv}$, the smooth loss can be formulated as:

$$\mathcal{L}_s = \sum (x_{i,j} - x_{i+1,j})^2 + (x_{i,j} - x_{i,j+1})^2, \tag{7}$$

where $x_{i,j}$ is the pixel value of $\mathbf{I}_{adv}$ at coordinate (i, j).

To sum up, the generated camouflage in this case will be visually correlated to the scenario context in both the pixel and perceptual level, leading to evade the human perceptual attention.

3.5. Overall Optimization Process

Overall, we generate the adversarial camouflage by jointly optimizing the model attention distraction loss $\mathcal{L}_d$, human attention evasion loss $\mathcal{L}_e$, and smooth loss $\mathcal{L}_s$.

Specifically, we first distract the target model from the salient objects to the meaningless part (*e.g.*, background); we then evade the human-specific attention mechanism by enhancing the strong perceptual correlation to the scenario context. Thus, we can generate transferable and visually-natural adversarial camouflages by minimizing the following formulation as

$$\min \mathcal{L}_d + \lambda \mathcal{L}_e + \mathcal{L}_s, \tag{8}$$

where λ controls the contribution of the term $\mathcal{L}_e$.

To balance the attacking ability and appearance naturalness, we set λ as 10^{-5} in the classification task and 10^{-3} in the detection task, and set β as 8 according to our experimental results. The overall training algorithm can be described as Algorithm 1.

Algorithm 1 Dual Attention Suppression (DAS) Attack

```
Input: environmental parameter set $C = \{c_1, c_2, ...c_r\}$ ,
    3D real object $(\mathbf{M}, \mathbf{T})$, seed content patch $\mathbf{P}_0$, neural
    renderer $\mathcal{R}$, attention model $\mathcal{A}$, and a class label $y$
Output: adversarial texture tensor $\mathbf{T}_{adv}$
  $\mathbf{T}_0 \leftarrow \Psi(\mathbf{P}_0, \mathbf{T})$
  $\mathbf{P}_{edge} \leftarrow \Phi(\mathbf{P}_0)$
  transform $\mathbf{P}_{edge}$ to $\mathbf{E}$
  initial $\mathbf{T}_{adv}$ as $\mathbf{T}_0$
  for the number of epochs do
    select $minibatch$ environmental conditions from $C$
    for $m = r/minibatch$ steps do
      $\mathbf{I}_{adv} \leftarrow \mathcal{R}((\mathbf{M}, \mathbf{T}_{adv}), c_m)$
      $\mathbf{S}^y \leftarrow \mathcal{A}(\mathbf{I}_{adv}, y)$
      calculate the $\mathcal{L}_d$, $\mathcal{L}_e$ and $\mathcal{L}_s$ by Eqn (5, 6, 7)
      optimize the $\mathbf{T}_{adv}$ by Eqn (8)
    end for
  end for
```

4. Experiments

In this section, we first outline the experimental settings, we then illustrate the effectiveness of our proposed attacking framework by thorough evaluations in both the digital and physical world.

4.1. Experimental Settings

Virtual environment. To perform a physical world attack, we choose CARLA [10] as our 3D virtual simulated environment, which is the commonly used open-source simulator for autonomous driving research. Based on Unreal Engine 4, CARLA provides many high-resolution open digital assets, *e.g.*, urban layouts, buildings, and vehicles to simulate a digital world that is nearly the same as the real world.

Evaluation metrics. To evaluate the performance of our proposed method, we select the widely used Accuracy as the metric for the classification task; as for the detection task, we adopt the P@0.5 following [52], which reflects both the IoU and precision information.

Compared methods. We choose several state-of-the-art works in the 3D attack and physical attack literature, including UPC [19], CAMOU [52], and MeshAdv [47]. We use ResNet-50 as its base-model for the classification and Yolo-V4 for detection. We provide more information about these methods in Supplementary Material.

Target models. We select commonly used model architectures for experiments. Specifically, Inception-V3 [43], VGG-19 [41], ResNet-152 [15], and DenseNet [18] are employed for the classification task; Yolo-V5 [38], SSD [32], Faster R-CNN [39], and Mask R-CNN [14] are employed for the detection task. For all the models, we use the pre-trained version on ImageNet and COCO.

Implementation details. We empirically set $\lambda = 10^{-5}$ for classification task, $\lambda = 5 \times 10^{-3}$ for detection task and we set $\beta = 8$. We adopt an Adam optimizer with a learning rate of 0.01, a weight decay of 10^{-4}, and a maximum of 5 epochs. We employ a seed content patch (*e.g.*, a stick smile face image) as the appearance of the 3D object in the training process. All of our codes are implemented in PyTorch. We conduct the training and testing processes on an NVIDIA Tesla V100-SXM2-16GB GPU cluster. In the physical world attack scenario, adversaries only have limited knowledge and access to the deployed models (*i.e.*, architectures, weights, *etc*.). Considering this, we mainly focuses on attacks in the black-box settings, which is more meaningful and applicable for physical world attacks.

4.2. Digital World Attack

In this section, we evaluate the performance of our generated adversarial camouflages on the vehicle classification and detection task in the digital world under black-box settings.

We randomly select 155 points in the simulation environment to place the vehicle and use a virtual camera to capture 100 images at each point using different settings (*i.e.*, angles, and distances). Specifically, we use different distance values (5, 10, 15, and 20), four camera pitch angle values (22.5°, 45°, 67.5°, and 90°), and eight camera yaw angle values (south, north, east, west and southeast, southwest, northeast, northwest). We then collect 15,500 simulation images with different setting combinations, and we choose 12,500 images as the training set and 3,000 images as the test set. To conduct fair comparisons, we use the backbone of ResNet-50 (for classification) and Yolo-V4 (for detection) as attention modules in training. As illustrated in Table 1 and Table 2, we can draw several conclusions as follows:

(1) Our adversarial camouflage achieves significantly better performance for both classification and detection tasks on different models (a maximum drop by **41.02%** on ResNet-152 and a maximum drop by **23.93%** on Faster R-CNN).

(2) We found that UPC works comparatively worse than other baselines for detection task. We conjecture the reason might be that UPC is primarily designed for physical attacks therefore showing worse attacking ability in the digital world. By contrast, our DAS attack exploits the intrinsic characteristics, which still achieves good attacking ability in the digital world.

(3) SSD shows evidently better robustness compared to other backbone models (*i.e.*, lower accuracy decline). The reason might be that some modules in SSD are less vulnerable to adversarial attacks, which could be used to further improve model robustness. We put it as future work.

Method	Accuracy (%)			
	Inception-V3	VGG-19	ResNet-152	DenseNet
Raw	74.36	40.62	73.51	71.91
MeshAdv	42.31	32.44	35.33	58.04
CAMOU	47.51	31.46	48.93	57.56
UPC	42.40	38.00	48.18	65.87
Ours	**39.86**	**30.18**	**32.49**	**55.42**

Table 1. The results in the digital world on the classification task.

Method	P@0.5 (%)			
	Yolo-V5	SSD	Faster R-CNN	Mask R-CNN
Raw	92.07	81.54	86.04	89.24
MeshAdv	**72.45**	66.44	71.84	80.84
CAMOU	74.01	73.81	69.64	76.44
UPC	82.41	74.58	76.94	81.97
Ours	**72.58**	**65.81**	**62.11**	**70.21**

Table 2. The results in the digital world on the detection task.

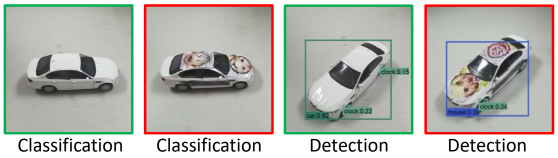


Figure 3. The results of attacking toy cars. They are respectively recognized as `car`, `sandal`, `car`, `mouse`.

4.3. Physical World Attack

As for the physical world attack, we conduct several experiments to validate the practical effectiveness of our generated adversarial camouflages. Due to the limitation of funds and conditions, we print our adversarial camouflages by an HP Color LaserJet Pro MFP M281fdw printer and stick them on a toy car model with different backgrounds to simulate the real vehicle painting. To conduct fair comparisons, we take 144 pictures of the car model on various environmental conditions (*i.e.*, 8 directions {left, right, front, back and their corresponding intersection directions}, 3 angles {0°, 45°, 90°}, 2 distances {long and short distances} and 3 different surroundings) using a Huawei P40 phone. The visualization of our generated adversarial camouflages can be found in Figure 3.

The evaluation results can be witnessed in Table 3 and Table 4. Compared with other methods, the DAS shows competitive transferable attacking ability, which is significantly better than the compared baselines (*e.g.*, **31.94%** on Inception-V3, **27.78%** on VGG-19, **29.86%** on ResNet-152, and **34.03%** on DenseNet, respectively). Moreover, the evaluation result of UPC appears a distinct improvement than that in the digital world, which is consistent with our analysis. However, the SSD shows lower robustness in the physical world which is worth further study. Besides, the Yolo-V5 shows stunning P@0.5 values, which probably because that Yolo-V5 is specially designed for applications in the physical world. Though facing this strong model, our DAS method still shows a certain attacking ability compared with others.

To sum up, the experimental results demonstrate the strong transferable attacking ability of our adversarial camouflages in the physical world.

4.4. Model Attention Analysis

In this part, we conduct a detailed analysis on model attention through both qualitative and quantitative studies to validate the effectiveness the model attention distraction in our DAS attack.

Firstly, we conduct a qualitative study by visualizing the attention regions of different models towards the same image. As shown in Figure 4(a), different DNNs show similar attention patterns towards the same image. In other words, different models pay their attention to similar regions, indicating that the attention is shared among models and can be deemed as a model-agnostic characteristic.

We then conduct a quantitative study by calculating the structural similarity index measure (SSIM) [54], which is a well-known quality metric used to measure the similarity between two images [17]. Specifically, we generate the attention maps of a specific image (*i.e.*, panda) on different models and calculate the SSIM values between each pair of the attention maps on different models. As shown in Figure 4(b), different models demonstrate comparatively high similarities of the attention maps.

Finally, we visualize the attention differences before and after attacks as shown in the Figure 5, indicating that the model attention is distracted away from the salient regions.

Method	Accuracy (%)			
	Inception-V3	VGG-19	ResNet-152	DenseNet
Raw	58.33	40.28	41.67	46.53
MeshAdv	40.28	34.03	38.89	36.11
CAMOU	40.28	29.17	31.25	45.14
UPC	35.41	33.33	33.33	41.67
Ours	**31.94**	**27.78**	**29.86**	**34.03**

Table 3. The results in the physical world on the classification task.

Method	P@0.5 (%)			
	Yolo-V5	SSD	Faster R-CNN	Mask R-CNN
Raw	100.00	90.28	68.06	93.75
MeshAdv	100.00	61.11	56.25	63.19
CAMOU	99.31	61.11	61.81	63.19
UPC	100.00	63.19	52.08	61.81
Ours	**92.36**	**56.25**	**44.44**	**54.86**

Table 4. The results in the physical world on the detection task.

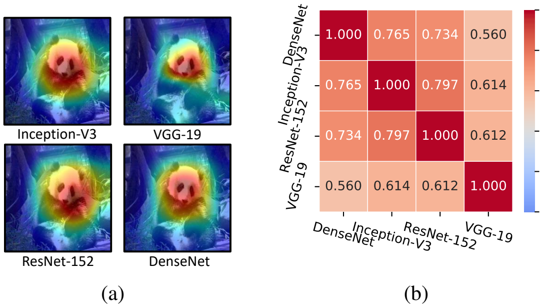


Figure 4. (a) is the attention maps on 4 different models to a particular image. (b) is a heat map drawn according to the SSIM values.

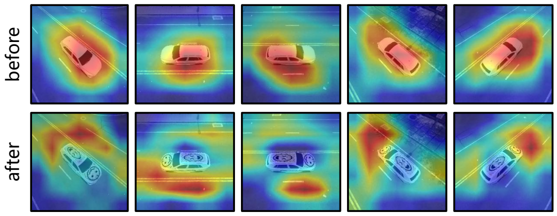


Figure 5. The attention maps before and after our DAS attack. After our DAS attack, the model attention is distracted.

In summary, we can draw several conclusions as follows: (1) different DNNs show similar attention patterns towards the same class in a specified image; (2) we can adversarially attack a DNN to wrong predictions by distracting its attention. More experimental results can be found in the Supplementary Material.

4.5. Human Perception Study

To evaluate the naturalness of our generated adversarial camouflage, we conduct a human perception study on one of the most commonly used crowdsourcing platform. We adversarially perturb our 3D car object using different methods (*i.e.*, MeshAdv, CAMOU, UPC, and Ours) and get the adversarial textures. Then we paint the car using these camouflages and get the rendered images for human perception studies as follows: (1) Recognition. The participants are asked to assign each of the camouflages generated by the methods above to one of the 8 classes (the ground-truth class, 6 classes similar to the ground-truth, and "I cannot tell what it is"). As for CAMOU, given it lacks semantic information, we do not consider it for the recognition task; (2) Naturalness. The participants are asked to score the naturalness of the camouflages from 1 to 10. In particular, we collect all responses from 106 participants.

Question	Percent (%)			
	MeshAdv	CAMOU	UPC	Ours
Recognition	36.6	–	27.4	**49.6**
Naturalness	43.4	39.6	40.6	**60.4**

Table 5. The results of human perception study.

As shown in Table 5, **49.6%** of the participants can recognize the ground-truth label for our camouflages, which are far better than those generated by other methods. As for the naturalness task, up to **60.4%** of the participants believe that our adversarial camouflage is natural-looking, which outperforms others by large margins (17%+). Thus, we can conclude that our adversarial camouflage is most visually natural and perceptually consistent to human perception.

4.6. Ablation Studies

In this section, we conduct several ablation studies to further investigate the contributions of our two main loss terms, *i.e.*, the model attention distraction loss and the human attention evasion loss. Due to the fact that the smooth loss is fully studied in [13], we set it as a fixed term.

The effect of different loss terms. Different loss terms play different roles, we conduct an ablation study to further investigate the effect of loss terms. We argue that the model attention distraction loss $\mathcal{L}_d$ mainly provides a transferable attacking ability in our DAS method and the human attention evasion provides the natural appearance. To prove these views, we conduct an experiment by calculating different loss term combinations. Specifically, we optimize the adversarial camouflage using function $\mathcal{L}_d$, $\mathcal{L}_e$, and $\mathcal{L}_d+\lambda\mathcal{L}_e$ respectively (with $\mathcal{L}_s$ fixed). As shown in Table 6, the accuracy shows a significant drop (*i.e.*, **36.53%** under $\mathcal{L}_d$ setting to 59.87% under $\mathcal{L}_e$ setting, 39.86% under $\mathcal{L}_d + \mathcal{L}_e$ setting). And the corresponding SSIM values generated with a benign image are 0.6905, 0.9987, and 0.7551 respectively, demonstrating our viewpoints. Besides, an interesting result can be observed in our experiments. When training under $\mathcal{L}_e$ setting, the accuracy appears an evident improvement on VGG-19 and DenseNet but drop on Inception-V3 and ResNet-152, which means that common textures may cause agnostic impact to DNNs, further demonstrating their vulnerability.

Method	Accuracy (%)			
	Inception-V3	VGG-19	ResNet-152	DenseNet
Raw	74.36	40.62	73.51	71.91
$\mathcal{L}_d$	36.53	25.87	31.20	51.73
$\mathcal{L}_e$	59.87	50.00	47.87	75.07
$\mathcal{L}_d + \lambda\mathcal{L}_e$	39.86	30.18	32.49	55.42

Table 6. The ablation study on attention distraction portion. We set λ as 10^{-5}.

The effect of hyper-parameter λ. Regarding the hyper-parameter λ, we argue that it controls the level of the strong semantic correlation with the scenario context. We evaluate the effectiveness of λ on a ResNet-50 model using Accuracy and SSIM. Specifically, we set the λ as 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, and 10^{-1}, respectively. As illustrated in Figure 6, the model accuracy first increases and then keeps a stable value as λ increases. We calculate the SSIM values between each pair of the clean and corresponding adversarial example, which shows the similar tendency (*i.e.*, 0.7034, 0.7551, 0.8750, 0.9982, 0.9991, and 0.9998, the closer to 1 the SSIM value is, the more similar the images are). According to the results, we can draw the conclusion that λ balances the attacking ability and appearance. When λ gets bigger, the accuracy and SSIM value get bigger, which means lower attacking ability and better appearance. And finally, the SSIM achieves its upper bound, leading to the loss of additional attacking ability.

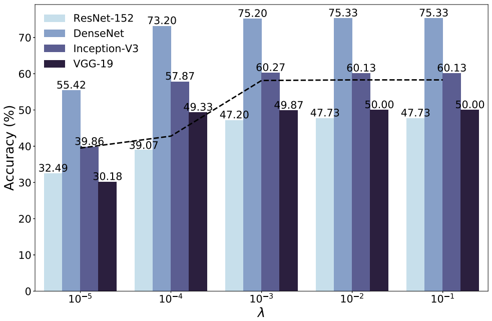


Figure 6. Ablation on studying the effectiveness of λ. The dotted line represents the trend of accuracy change, and the corresponding value of each λ is the average accuracy of the four models.

5. Conclusion

In this paper, we propose the Dual Attention Suppression (DAS) attack to generate adversarial camouflage in the physical world by suppressing both model and human attention. To improve the transferability of adversarial camouflages, we suppress the model attention by distracting the model-shared similar attention from target to non-target regions. Since our generated camouflage captures the model-agnostic structures, it can transfer among different models. To generate more visually-natural camouflage, we suppress the human attention by evading the human-specific bottom-up attention. By preserving the shape of a seed content patch which has strong semantic association to the scenario context, the generated camouflage can be highly correlated to human perception, which is more natural and unsuspicious to human attention. We conduct extensive experiments for both classification and detection tasks in both the digital and physical world under black-box setting, and our DAS outperforms state-of-the-art baselines.

In the future, we are interested in investigating the attack abilities of our adversarial camouflage using a real vehicle in the real-world scenario. Using projection or 3D printing, we could simply paint our camouflage on a real-world vehicle. Further, we would also like to investigate the effectiveness of our generated camouflage to improving model robustness against different noises.

6. Acknowledge

This work was supported by The National Key Research and Development Plan of China (2020AAA0103502), National Natural Science Foundation of China (62022009, 61872021), Beijing Nova Program of Science and Technology (Z191100001119050), and State Key Lab of Software Development Environment (SKLSDE-2020ZX-06).

References

[1] Madry A., Makelov A., L. Schmidt, Tsipras D., and Vladu A. Towards deep learning models resistant to adversarial attacks. In *arXiv preprint arXiv:1706.06083*, June 2017. 2

[2] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing Robust Adversarial Examples. *arXiv e-prints*, page arXiv:1707.07397, July 2017. 1, 2

[3] Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *CoRR*, December 2017. 3

[4] J. Canny. A computational approach to edge detection. *PAMI*, PAMI-8, November 1986. 4

[5] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *WACV*, March 2018. 3

[6] Charles E. Connor, Howard E. Egeth, and Steven Yantis. Visual attention: Bottom-up versus top-down. *Current Biology*, 14(19), October 2004. 2, 3, 4

[7] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, and Hang Su. Boosting adversarial attacks with momentum. In *CVPR*, June 2018. 1

[8] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *CVPR*, June 2018. 2

[9] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *CVPR*, June 2019. 2

[10] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *CoRL*, November 2017. 5

[11] Ranjie Duan, Xingjun Ma, Yisen Wang, James Bailey, A. K. Qin, and Yun Yang. Adversarial camouflage: Hiding physical-world attacks with natural styles. In *CVPR*, June 2020. 2, 4

[12] Gamaleldin Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alexey Kurakin, Ian Goodfellow, and Jascha Sohl-Dickstein. Adversarial examples that fool both computer vision and time-limited humans. In *NeurIPS*, December 2018. 2

[13] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *CVPR*, June 2018. 1, 2, 5, 8

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, March 2017. 5

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, June 2016. 5

[16] Michael Hentrich. Methodology and coronary artery disease cure. *Available at SSRN 2645417*, August 2015. 3

[17] A. Horé and D. Ziou. Image quality metrics: Psnr vs. ssim. In *ICPR*, August 2010. 7

[18] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *CoRR*, August 2016. 5

[19] Lifeng Huang, Chengying Gao, Yuyin Zhou, Cihang Xie, Alan L. Yuille, Changqing Zou, and Ning Liu. Universal physical camouflage attacks on object detectors. In *CVPR*, June 2020. 1, 2, 5

[20] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *NeurIPS*, May 2019. 1

[21] Nathan Inkawhich, Wei Wen, Hai (Helen) Li, and Yiran Chen. Feature space perturbations yield more transferable adversarial examples. In *CVPR*, June 2019. 2

[22] Goodfellow I J, Shlens J, and Szegedy C. Explaining and harnessing adversarial examples. In *arXiv preprint arXiv:1412.6572, 2014*, December 2014. 1, 2

[23] Yunhan Jia, Yantao Lu, Senem Velipasalar, Zhenyu Zhong, and Tao Wei. Enhancing cross-task transferability of adversarial examples with dispersion reduction. *CoRR*, May 2019. 2

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, May 2012. 1

[25] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *CoRR*, July 2016. 2

[26] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *ICLR Workshop*, July 2017. 1

[27] Tricoche L, Ferrand-Verdejo J, Pélisson D, and Meunier M. Peer presence effects on eye movements and attentional performance. In *Front Behav Neurosci*, Jan 2020. 2

[28] Tianlin Li, Aishan Liu, Xianglong Liu, Yitao Xu, Chongzhi Zhang, and Xiaofei Xie. Understanding adversarial robustness via critical attacking route. *Information Sciences*, February 2021. 1

[29] Aishan Liu, Tairan Huang, Xianglong Liu, Yitao Xu, Yuqing Ma, Xinyun Chen, Stephen J. Maybank, and Dacheng Tao. Spatiotemporal attacks for embodied agents. In *ECCV*, August 2020. 2, 3, 4

[30] Aishan Liu, Xianglong Liu, Jiaxin Fan, Yuqing Ma, Anlan Zhang, Huiyuan Xie, and Dacheng Tao. Perceptual-sensitive GAN for generating adversarial patches. In *AAAI*, January 2019. 2, 3

[31] Aishan Liu, Jiakai Wang, Xianglong Liu, Chongzhi Zhang, Bowen Cao, and Hang Yu. Patch attack for automatic checkout. In *ECCV*, May 2020. 1, 2, 3

[32] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, and Scott Reed. Ssd: Single shot multibox detector. In *ECCV*, March 2016. 5

[33] Sharif M, Bhagavatula S, Bauer L, and Michael K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *CCS*, October 2016. 2

[34] A. Mohamed, G. E. Dahl, and G. Hinton. Acoustic modeling using deep belief networks. *IEEE-ACM T AUDIO SPE*, 20(1), January 2012. 1

[35] Haotong Qin, Zhongang Cai, Mingyuan Zhang, Yifu Ding, Haiyu Zhao, Shuai Yi, Xianglong Liu, and Hao Su. Bipointnet: Binary neural network for point clouds. In *International Conference on Learning Representations*, 2021. 1

[36] Haotong Qin, Ruihao Gong, Xianglong Liu, Xiao Bai, Jingkuan Song, and Nicu Sebe. Binary neural networks: A survey. *Pattern Recognition*, 105:107281, 2020. 1

[37] Haotong Qin, Ruihao Gong, Xianglong Liu, Mingzhu Shen, Ziran Wei, Fengwei Yu, and Jingkuan Song. Forward and backward information retention for accurate binary neural networks. In *IEEE CVPR*, 2020. 1

[38] Redmon and Joseph et al. You only look once: Unified, real-time object detection. In *CVPR*, September 2016. 5

[39] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *TPAMI*, 39, June 2015. 5

[40] 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 *ICCV*, Oct 2017. 3

[41] Zisserman A Simonyan K. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, September 2014. 5

[42] I Sutskever, O Vinyals, and QV Le. Sequence to sequence learning with neural networks. *NeurIPS*, December 2014. 1

[43] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, December 2015. 5

[44] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, December 2013. 1, 2

[45] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *ICLR*, May 2019. 1

[46] Xiu-Shen Wei, Quan Cui, Lei Yang, Peng Wang, and Lingqiao Liu. RPC: A large-scale retail product checkout dataset. *CoRR*, January 2019. 2

[47] Chaowei Xiao, Dawei Yang, Bo Li, Jia Deng, and Mingyan Liu. Meshadv: Adversarial meshes for visual recognition. In *CVPR*, June 2019. 5

[48] 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 *CVPR*, June 2019. 2

[49] Evans A C Zatorre R J, Mondor T A. Auditory attention to space and frequency activates similar cerebral systems. In *Neuroimage*, November 1999. 2, 3

[50] Chongzhi Zhang, Aishan Liu, Xianglong Liu, Yitao Xu, Hang Yu, Yuqing Ma, and Tianlin Li. Interpreting and improving adversarial robustness with neuron sensitivity. *IEEE Transactions on Image Processing*, 2020. 1

[51] Xiangguo Zhang, Haotong Qin, Yifu Ding, Ruihao Gong, Qinghua Yan, Renshuai Tao, Yuhang Li, Fengwei Yu, and Xianglong Liu. Diversifying sample generation for data-free quantization. In *IEEE CVPR*, 2021. 1

[52] Yang Zhang, Hassan Foroosh, Philip David, and Boqing Gong. CAMOU: Learning physical vehicle camouflages to adversarially attack detectors in the wild. In *ICLR*, May 2019. 1, 2, 5

[53] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, June 2016. 3

[54] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), April 2004. 7