Computers & Security 122 (2022) 102876

Contents lists available at ScienceDirect

Computers & Security

journal homepage: www.elsevier.com/locate/cose

Misleading attention and classification: An adversarial attack to fool object detection models in the real world

Haotian Zhang[a], Xu Ma[a,b,*]

[a] School of Cyber Science and Engineering, Qufu Normal University, Qufu 273165, China
[b] State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an, China

ARTICLE INFO

Article history:
Received 21 March 2022
Revised 25 July 2022
Accepted 12 August 2022
Available online 17 August 2022

Keywords:
Computer vision
Model security
Adversarial attack
Object detection

ABSTRACT

Object detection is a hot topic in computer vision (CV), and it has many applications in various security fields. However, many works have demonstrated that neural network-based object detection is vulnerable to adversarial attacks. In this paper, we study adversarial attacks on object detectors in the real world and propose a new adversarial attack called Misleading Attention and Classification Attack (MACA), which can generate adversarial patches to mislead the object detectors. Specifically, we propose a new scheme to generate adversarial patches to fool the object detector. Our scheme restricts the noise of the adversarial patches and aims to ensure that the generated adversarial patches are visually similar to natural images. The attack simulates the complex external physical environment and the 3D transformations of non-rigid objects to increase the robustness of adversarial patches. We attack the up-to-date object detectors (e.g., Yolo-V5), and we prove that our technique has strong transferability among different detectors. Extensive experiments show that it is feasible to transfer the digital adversarial patches to the real world while maintaining the transferability of adversarial patches among different models and the success rate of adversarial attacks.

1. Introduction

Deep Neural Networks (DNNs) has been widely used in various fields, such as medicine (Boveiri et al., 2020), intelligent cities (Chen et al., 2020; Zhang et al., 2020), natural language processing (Chowdhary, 2020), and computer vision (Girshick et al., 2014), etc. As one of the essential tasks of DNNs, computer vision has been used extensively in face recognition (Adjabi et al., 2020), autonomous driving (Grigorescu et al., 2020), and gesture recognition (Cheng et al., 2019). Moreover, it also has been applied to various safety-critical systems. Nonetheless, a series of researches suggest that no matter in the digital world or the real world, DNNs are vulnerable to various attacks, such as model extraction attacks (Gao et al., 2021), data poisoning attacks (Chen et al., 2017), and backdoor attacks (Liu et al., 2020). Especially in the field of CV, DNNs models are vulnerable to deception by the adversarial examples (Ren et al., 2021; Szegedy et al., 2014).

Generally, adversarial attacks can be divided into two categories: adversarial digital attacks in the digital world and adversarial physical attacks in the real world. Most of the previous studies focused on adversarial digital attacks. These attacks add a minor amount of disturbances to the data in the digital world t fools DNNs. These disturbances are usually very slight and they only produce pixel-level noise. Hence it is difficult to transfer such attacks to the real world. In contrast, adversarial physical attacks can change the physical characteristics (such as texture and shape) of the actual target in the real world, which makes the object detector capture the wrong target recognition. These attacks generally need to be collected by cameras, such as maliciously trained images affixed to traffic signs which cause DNNs models to make incorrect predictions (Li et al., 2021; Morgulis et al., 2019).

Our proposed MACA is an adversarial physical attack that can hide any class of the target (e.g., person, cars) in the real world, and the adversarial patches generated by MACA can be affixed to the target to fool the object detection models. In this paper, we propose a physical adversarial attack on a ǣpersong class in the real world. We mainly consider the ǣperson” as our target class because it is one of the targets with the highest security risk. In several domains including autonomous driving and video surveillance, the misclassification of a ǣpersong will lead to a huge security threat. Meanwhile, adversarial attacks against the “person” class are more complex for the following reasons: (1) People can move, which means the attacks need to be robust against complex physical environments (such as illumination, background, viewpoints, and distance). (2) The clothing worn by the person should

[*] Corresponding author.
E-mail address: xma@qfnu.edu.cn (X. Ma).

https://doi.org/10.1016/j.cose.2022.102876
0167-4048/© 2022 Elsevier Ltd. All rights reserved.

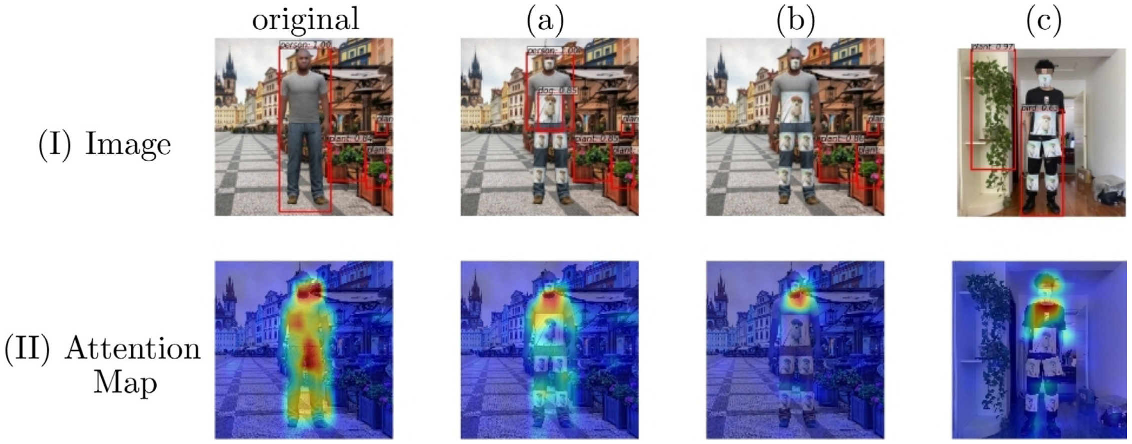


Fig. 1. (II) is the attention map generated by Grad-CAM++ (Chattopadhyay et al., 2017) according to (I). We use the original image as a baseline. (a) is the 3D model with the untrained natural patches in virtual scenes. (b) Physical attacks (MACA) in virtual scenes and (c) Physical attacks (MACA) in real world. The attention map of the images with adversarial patches is more distracted than others.

be inconspicuous. Hence, the visual naturalness of the adversarial patches should be semantically constrained. (3) Because different postures, movements, and body shapes will distort the adversarial patches, so the patches need to tolerate those changes in a way.

To address the mentioned problems, this paper propose an efficient adversarial attack by fooling both the models' attention and classification called MACA. Specifically, we use the idea of iterative optimization to optimize the loss function to find effective adversarial patches. The complex physical environments are simulated by applying various geometric transformations (e.g., rotation, cropping, brightness). We impose additional optimization constraints to ensure that the generated patches are visually meaningful, which are called ”semantic constraints”. The generated patches are added to a specific object, making the object detection model blind to the object in the real world. Since there are shared attention patterns among different networks, and different networks have similar attention to the same image, we attack and divert the detection box attention of a network to achieve the transferability of the adversarial patches in black-box networks. The “persong class is different from other targets, it is impossible to be covered by a large area of the patches (such as cars), we also fool the classifier of the object detection models to misclassify each detection box with high confidence to improve the performance of the adversarial patches.

As shown in Fig. 1, we attacked the common object detection networks, particularly the Faster R-CNN (Ren et al., 2017) and Yolo-V3 (Redmon and Farhadi, 2018). The adversarial patches are generated in digital images firstly and then printed in physical copies. The physical copies are stuck to particular positions on clothes worn by different people. These patches have well transferability and can mislead various object detectors. The contributions of our work are three folds:

1) We propose a novel joint adversarial attack called MACA, which enables a person escapes from several state-of-the-art object detection networks by wearing the trained visually natural adversarial patches. MACA is the first attack that combines attention patterns with a fooling ground-truth label to improve the effectiveness of the patches on specific targets. Meanwhile, MACA is effective among different networks.
2) Considering different physical environments, we introduce simulated disturbances (e.g., rotation, internal deformations) for the trained target while constraining patches with complex images of semantic backgrounds rather than simple vector shapes to maintain the visual naturalness of the patches. We then use 3D human models to conduct simulation experiments under different external backgrounds and camera viewpoints to demonstrate the effectiveness of our scheme.
3) We accomplish real-world adversarial attacks through trained patches affixed on clothing and fool different object detection networks successfully. Experiments show that our scheme not only achieves better results for attacking object detectors in different environments, but also has excellent generalization and transfer-ability among different networks.

2. Related work

Adversarial examples attack DNNs seriously threaten security systems such as traffic safety and property safety. In this section, we review recent developments in object detectors, attention patterns of computer vision, and adversarial attacks.

2.1. Object detection

Briefly speaking, the tasks of object detection are object recognition and object positioning (Girshick, 2015). There are two main categories of object detection algorithms: the R-CNN (Girshick et al., 2014) series of two-shot detection algorithms and the Yolo series (Redmon et al., 2016) of one-shot regression detection algorithms. Other object detection algorithms include the Single Shot MultiBox Detector (SSD) (Liu et al., 2016).

Faster R-CNN (Ren et al., 2017) is the most commonly used R-CNN algorithm. This algorithm integrates feature extraction, region proposal extraction, and classification into one model. The performance and speed are greatly improved in Faster R-CNN, but it still contains two stages: first, it extracts region proposals, and then it performs classification. The Yolo model is a convolutional neural network that can predict multiple detection frame positions and categories at one time. The most famous model of Yolo is Yolo-V3 (Redmon and Farhadi, 2018). It divides the image into multiple groups of feature maps, and each group containing three prediction frames. Each prediction frame comprises four coordinates of the prediction frame and the classification probabilities of identifying different objects. The SSD proposed by Liu et al. (2016) is one of the state-of-the-art detection frameworks. Belonging to a one-step regression detection model, it combines the advantages and characteristics of Yolo and Faster R-CNN. It is faster than Faster R-CNN, and it has advantages on apparent mean Average Precision (mAP) compared to Yolo.

2.2. Attention patterns of computer vision

Deep Learning (DL) is a “black-box system” with less interpretability. People use the Interpretability of DL algorithm to clearly outline a model’s task and link it to defined principles in the real world (Zhang and Zhu, 2018). Attention patterns are often used to visualise neural networks in computer vision to let one know what the network relies on to predict labels. Although most object detection algorithms have different backbone networks, they have similar attention patterns when predicting the same target. When making predictions, a model always pays more attention to the target objects rather than the meaningless parts.

Many studies have been conducted on network visualization. ZFnet (Zeiler and Fergus, 2014), first proposed by David et al., improves AlexNet and partially solves the visualization problem of convolutional neural networks. CAM (Zhou et al., 2016) proposes class activation mapping to visualize neural networks, but it requires modifying the structure of the original model which will retrain the model. Grad-CAM (Selvaraju et al., 2017) uses the global average of gradients to calculate weights without modifying the network structure or retraining. Grad-CAM++ (Chattopadhyay et al., 2017) improves on Grad-CAM for more comprehensive visualization of multiple objects in one image. Compared to CAM, CALM (Kim et al., 2021) can identify the discriminative properties of image classifiers more accurately.

2.3. Adversarial attack against DNNs in the digital world

The research on adversarial examples was firstly proposed for digital objects (Goodfellow et al., 2015; Szegedy et al., 2014). The adversarial examples, added some designed noise elaborately, can fool DNNs with high confidence, but it is imperceptible to humans. Many studies have suggested various methods on how to generate adversarial examples. In general, adversarial digital attacks can be divided into white-box attacks and black-box attacks. For the white-box attacks, the attacker needs to obtain the complete parameters of the models and fully control the input of the models, whereas the model parameters are agnostic to the attacker in the black-box attack. White-box attacks include the fast gradient method (FGSM) (Goodfellow et al., 2015), DeepFool (Moosavi-Dezfooli et al., 2016b), Jacobian-based saliency map attack (JSMA) (Papernot et al., 2016), CarliniWagner (C&W) Attack (Carlini and Wagner, 2017), etc. Classical black-box attacks include single-pixel attacks (Narodytska and Kasiviswanathan, 2016), local search attacks (Narodytska and Kasiviswanathan, 2016), transferable adversarial example attacks (Liu et al., 2017), universal adversarial attacks (Moosavi-Dezfooli et al., 2016a), etc. Yuan et al. (2019) proposed a method to generate adversarial pictures that can escape the content detection networks by adding noise and affine transformation to the original images. These attacks superimpose carefully constructed disturbances on the original images that are imperceptible to humans. However, adversarial digital attacks require the attacker to change all pixels of the input image, which are difficult to be migrated to the real world.

2.4. Adversarial attack against DNNs in the real world

Compared to the adversarial attacks in the digital world, adversarial attacks in the real world are more complex. There are many tricky problems that real-world attacks have to solve, such as the position, angles, brightness, and how many adversarial patches can be captured. Given an input clean image I with its groundtruth label y, the definition of real-world adversarial attacks are defined as:

$$\mathrm{F}(\mathrm{I}_{\mathrm{adv}}) \neq \mathrm{y} \tag{1}$$

where $\mathrm{F}(\cdot)$ represents a deep neural network, and I_{adv} is an adversarial example. The object detector will not be able to identify the true label of the adversarial example.

Morgulis et al. (2019) proposed a method to generate adversarial patches on the traffic signs, and the adversarial signs can fool a commercial car perception system in real-world driving conditions. Evtimov et al. (2017) propose a new attack algorithm that generates perturbations by taking images under different conditions into account, which can reduce the likelihood of detection by a casual observer. Duan et al. (2021) propose a real-world attack based on the laser beam, which leverages light as adversarial perturbation with high flexibility for attackers. Wang et al. (2021) proposes the dual attention suppression attack to generate visually natural real-world adversarial camouflages with transferability by suppressing model and human attention. Pedraza et al. (2021) shown that adversarial examples also appear in the real world without any attacker or maliciously selected noise involved. Compare to artificially generated adversarial examples, the natural adversarial examples have a higher distance from the originals.

The latest target of physical adversarial attacks has become the class “person” because this class has a more complex semantic environment. Thys et al. (2019) first proposed adversarial attacks on the human body. People hold a specific cardboard that has been trained to make them “disappear” in front of the object detector. Wu et al. (2019) designed an adversarial patch printed on clothing, which could fool the white-box Yolo-v2 successfully. They also noticed the non-rigid surface of the characters and the distortion of the clothing. Huang et al. (2019) improved the universal camouflage pattern to attack the classifier and regressor of the object detection models separately, at the same time, they paid attention to the scenario context of the adversarial patches. Compared with the existing works, our method not only simulates external environment conditions and TPS (Donato and Belongie, 2002) to keep the robustness of patches in non-rigid/non-planar (in Section 3.2) but also constructs semantic constraints of patches to ensure the generated patches are visually natural (in Section 3.5). We attack and divert the shared detection box attention of a network to realise a universal attack among different networks (in Section 3.3). We fool the classifier of the object detection models to improve the performance of the adversarial patches for “person” class and reduce the number of patches (in Section 3.4). A comparison with former methods is summarized in Table 1.

3. Methodology

In this section, we first explain the framework of MACA. Secondly, we show the real-world simulation to improve the robustness to make patches minor subjects to the external physical interferences, then propose the attacks against model attention and model classifier, respectively. Finally, we elaborate on the other parts of the total loss function and the total construction of the MACA algorithm.

3.1. Overview

Our adversarial attack aims to generate adversarial patches that can make the target undetectable by object detection models after putting the patches on it. Meanwhile, the adversarial patches should look natural and match the background well without attracting much attention. Our attack can hide any category of the target. Without loss of generality, we use the “persong category as an example to illustrate our method. We generate adversarial patches P_{adv} as follows:

$$P_{adv} = P + (P_r^t + \Delta P_r) \tag{2}$$

Table 1
Comparison with existing methods.

Methods	Adversarial	Non-rigid Non-planar	Universal	Natural	Advanced	Fewer Patches
Thys et al. (2019)	✔					✔
Wu et al. (2019)	✔	✔	✔			
Huang et al. (2019)	✔	✔	✔	✔		
Wang et al. (2021)	✔		✔	✔	✔	
Ours	✔	✔	✔	✔	✔	✔

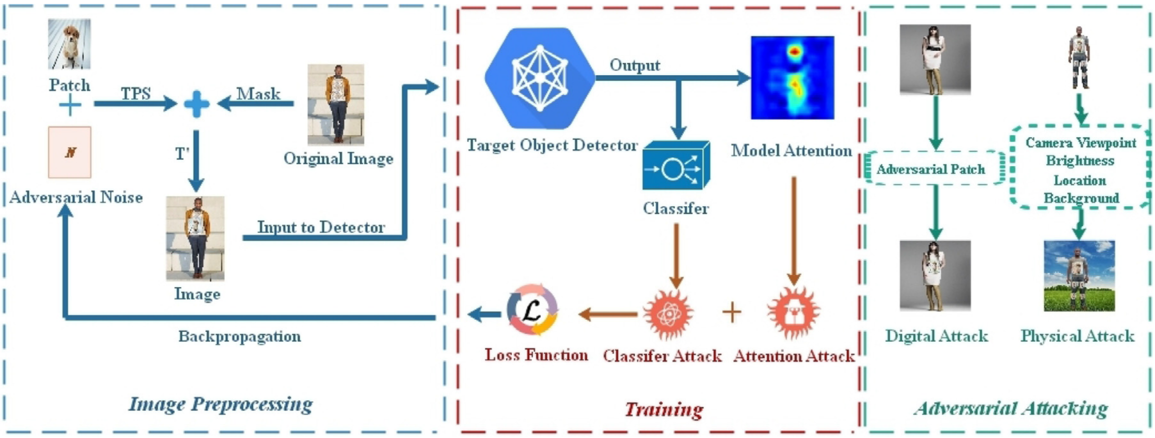


Fig. 2. The framework of the adversarial patches generation. We add a patch with adversarial noise on the object of the original images. The left dashed box represents image processing, T′ is external physical environment simulation. The mid dashed box represents our adversarial attack. The noise is continuously optimized according to the detection model's attention map and probability matrix. The right dashed box represents the adversarial attacking test in the digital world and real world.

where P is original natural picture, P_r is random noise, and we optimize the updated vector ΔP_r by training at iteration t. The adversarial patches should be effective in both the digital and the real world and do not have to be regenerated when the target changes. The framework we propose is called MACA. The steps are depicted in Fig. 2 and described below:

- **Step 1.** We first randomly generate noise on a selected natural image (e.g., dog, cat, flower) to get the target patch. For an efficient attack, we perform TPS transformation (Section 3.2) on the target patch to simulate the distortion of non-rigid and non-planar objects. Then we paste the target patch on the target images and simulate the external physical conditions (e.g., viewpoint, brightness, scale) of the target images for preprocessing.
- **Step 2.** The shared attention pattern among different models is attacked to fool target models by not paying attention to the objects in the target region. We put the preprocessed images into the object detection models to obtain its class attention mapping and minimize loss function $\mathcal{L}_{\text{CAM}}$ (Section 3.3) to reduce the object detector's attention.
- **Step 3.** To enhance the attacking strength further, we jointly attack model classification and optimize the adversarial patches by minimizing loss function $\mathcal{L}_{\text{Class}}$ (Section 3.4). Meanwhile, an additional semantic constraint is imposed to improve the naturalness and semantically meaningfulness of the adversarial patches (Section 3.5).

3.2. Physical simulation of the real world

Motion and changes in the external environment may weaken the effect of adversarial patches. Hence, in order to improve the robustness to make generated patches minor subject to the physical interferences, we simulate the external environments.

Simulation of physical environment: In each iteration of the training adversarial patches, to increase the robustness of adversarial patches, we simulate the physical interferences by randomly transforming the target images in terms of brightness, angle, occlusion, distance, and random noise. We use the virtual 3D human models (Fig. 5) rendered by 3D-Max to conduct the experiment under different environments, distance, lighting, and camera viewpoints to obtain results that are closer to the real environment.

Simulation of 3D distortion: Since our adversarial patches are attached to the clothing, the patches will be distorted randomly for many reasons, such as the wrinkles of the clothing and the distortion of the edges of the object. These distortions will affect the effectiveness of the patches. To handle non-rigid or non-planar objects effectively, we introduce Thin Plate Spline (TPS) (Donato and Belongie, 2002) transformation to model their internal deformations.

3.3. Attacks against the model attention

Although most object detection algorithms have different backbone networks, they always have similar attention patterns when predicting the same images. The attention of a network is attacked and diverted to make the adversarial patch more transferable when it is input into different black-box detection networks, as shown in existing research (Wang et al., 2021). In this paper, the models' attention is generated by Grad-CAM++. Our attack distracts the model's attention from the target in the image to other areas of the image.

To attack against model attention, we generate the attention map My of the groundtruth label y of the target image. For the image I, as the attacking target, we paste the adversarial patches P_{adv} generated by Eq. (2) on it to obtain training image I_{adv}. I_{adv} is put into the Grad-CAM++ generator of the target network, and obtain the attention map M_y of the label y of I_{adv}:

$$
\begin{aligned}
I_{adv} &= \mathrm{G}(I, \mathrm{TPS}(P_{adv}), \epsilon) \\
M_y &= \Phi(\mathrm{I}_{\mathrm{adv}}, y)
\end{aligned}
\tag{3}
$$

where ϵ denotes simulated physical conditions (e.g., rotate, brightness, scale), TPS$(\cdot)$ Donato and Belongie (2002) is the thin plate spline transformation, $\mathrm{G}(\cdot)$ is the image synthesis function, and $\Phi(\cdot)$ is the Grad-CAM++ generator proposed by Chattopadhyay et al. (2017) to obtain the image attention map.

Based on Wang et al. (2021), we improve the attention attack algorithm by restricting the algorithm to concentrate on the object prediction box, which can improve the effectiveness of the attack. After obtaining the attention map of the groundtruth label y of I_{adv}, the M_y is input to $\mathcal{L}_{\mathrm{CAM}}$ to reduce the network's attention of the I_{adv} as well as transferring attention from the target. The objective function for attacking the model attention is as follows:

$$\mathcal{L}_{\mathrm{CAM}} = \left(\frac{ReLU\,(M_y)}{N-n}\right)_{Obj} \tag{4}$$

where $(\cdot)_{Obj}$ means that we concentrate on the information in the object region generated by the model prediction box. N is the total number of pixels in the object region of attention map M_y, and n is the number of pixels in the object region of M_y whose value is greater than 0. $\mathrm{ReLU}(\cdot)$ is the ReLU function that only focuses on the area of the target, which is the part of the attention map whose value is greater than 0. $\mathrm{ReLU}(M_y)$ is the total pixel value greater than 0 in the object region of M_y. By minimizing $\mathcal{L}_{\mathrm{CAM}}$, the adversarial patches P_{adv} are optimized to make the total pixel value in the salient regions of the M_y as low as possible, which can reduce the attention of the target in the image. We make the pixels that have values larger than 0 in the M_y as few as possible to divert the attention of the target regions in the network.

3.4. Attacks against the model classification

After attacking the model attention, the detector's classifier is further fooled by decreasing the confidence of the groundtruth label y and increasing the confidence of the target label y' for attacking. Especially for the class such as “person”, the detector pays more attention to it with all different attention regions. Being different from cars, humans cannot cover all the attention regions by using the adversarial patches as much as possible like other objects. We attack an object by either decreasing the confidence of the groundtruth label and increasing the confidence of the attacking target label. We use a two-shot regression detection model as our attacking task. We define the objective function for attacking the network classifier as follows:

$$\mathcal{L}_{\mathrm{Class}} = \mathrm{l}(y', \mathrm{C}(I_{adv})) + \overline{\mathrm{C}(I_{adv}, y)} \tag{5}$$

where C is the classifier of the model, $\mathrm{C}(I_{adv})$ is the prediction output of I_{adv}, and $\mathrm{C}(I_{adv}, y)$ is the prediction output of the true label y of I_{adv}. y' is the target label for attacking, and $\mathrm{l}(\cdot)$ is the cross-entropy loss function. By minimizing $\mathcal{L}_{\mathrm{Class}}$, the adversarial patches P_{adv} are optimized to reduce the probability of the target's true label and increase the probability of incorrect labels fooling the classifier.

3.5. Loss functions

In this part, we present our total loss function and introduce in detail all parts of the function that need to be optimized.

Our adversarial patches will be affixed to clothing to achieve real-world disturbances. To smooth the adversarial patches, we use the small total variation (TV) constraint (Chambolle et al., 2010), which can reduce the difference square between adjacent pixels:

$$\mathcal{L}_{\mathrm{TV}} = \sum_{i,j}\left((p_{i,j} - p_{i+1,j})^2 + (p_{i,j} - p_{i,j+1})^2\right)^{\frac{1}{2}} \tag{6}$$

where $p_{i,j}$ is the pixel value of the adversarial patch p_{adv} at position (i, j). Decreasing the value of TV will make the patch smoother.

All the information on the adversarial patch we generate needs to be printed entirely in the real world. However, the printer in the real world cannot print all the colours in the digital world, which will decrease the interference ability of the patch. Therefore, we add the constraint item of the non-printable score (NPS) (Sharif et al., 2016) proposed by Mahmood to ensure the printability of the generated adversarial patches:

$$\mathcal{L}_{\mathrm{nps}} = \prod_{\substack{p' \in P_{adv} \\ p'' \in P_{color}}} |p' - p''| \tag{7}$$

where $P_{color} \subset [0,1]^3$ is a set of standard RGB color matrices (Sharif et al., 2016) that can be printed by real-world printers, p' is the RGB value of P_{adv}, and p'' is the RGB value of P_{color}. The lower the value of NPS is, the more information on the patch that can be printed.

To ensure the generated adversarial patches are semantically meaningful, we add a semantic constraint:

$$\mathcal{L}_{\mathrm{Con}} = \lambda \bullet \|P_{adv} - P_0\|^2 \tag{8}$$

where λ is the weight tensor, and P_0 is the original image of the patch. We use the euclidean metric to limit the degree of change of the patch. The larger the value of $\mathcal{L}_{\mathrm{Con}}$ is, the smaller the difference from the original image.

In summary, as shown in Algorithm 1, our total loss function is expressed as:

$$\mathcal{L} = \mathcal{L}_{\mathrm{CAM}} + \theta\mathcal{L}_{\mathrm{Class}} + \alpha\mathcal{L}_{\mathrm{TV}} + \beta\mathcal{L}_{\mathrm{nps}} + \mathcal{L}_{\mathrm{Con}} \tag{9}$$

where $\mathcal{L}_{\mathrm{CAM}}$ is the attention map loss, $\mathcal{L}_{\mathrm{Class}}$ is the classifier loss, $\mathcal{L}_{\mathrm{TV}}$ is the total variation loss, $\mathcal{L}_{\mathrm{nps}}$ is the nonprintable score loss, and $\mathcal{L}_{\mathrm{Con}}$ is the content loss. α and β are weight hyperparameters determined based on experience to measure the relative importance of each optimization item. Finally, θ is set to 0 at the beginning of training and changed to a constant after the training iteration reaches a certain number of rounds.

3.6. Overall attacking process

The proposed total loss function $\mathcal{L}$ is optimized iteratively. The ultimate goal of our optimization is to make the total loss function $\mathcal{L}$ as small as possible. Our overall algorithm for adversarial attacks can be described as Algorithm 1, where we iteratively update the random adversarial noise P_r until the attack iteration reaches the maximum. We first attack against model attention by getting the image's class attention mapping to reduce the model's attention. After iteration reaches the ep_θ, then additionally fools the classification to reduce further the detection accuracy of the model for the true label y. By minimizing Eq. (9) to optimize P_r, the adversarial patches P_{adv} can be generated by Eq. (2) to fool the detection model substantially.

4. Experiments

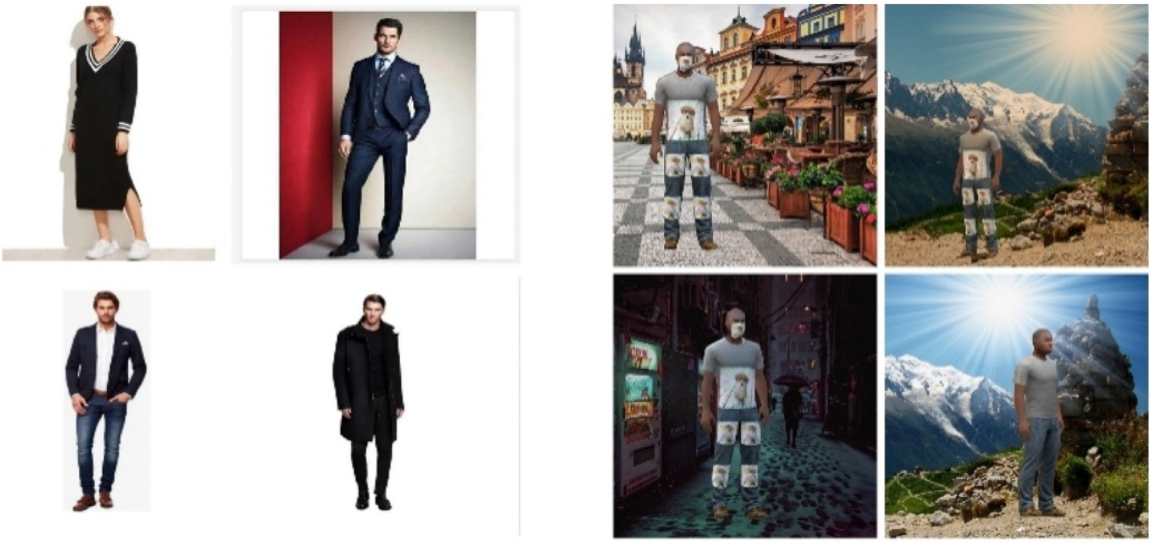
In this section, we first introduce our experiment environments. We then test the proposed method in the digital world and compare our framework with other proposed adversarial attack methods in the virtual 3D environment. Finally, we test in the real world to demonstrate the effectiveness of our proposed attacking framework in real physical environments.

4.1. Setup

Datasets. We mainly evaluate the effectiveness of our method on “persong category. We use the “person attack datasetg collected by Huang et al. (2019) as the training dataset to be attacked. We expand it to 300 real human images with various attributes (such

Algorithm 1 Misleading Attention and Classification Attack.

```
Input: $\chi$: Original image datesets; $y$: True label; $y'$: Target label; $P$: Original patch image; $P_r$: Adversarial noise; $ep_{\max}$, $ep_{\theta}$: Iteration parameters; $\theta$: Control parameter; $\mathrm{N}(\cdot)$: Target network; $\mathrm{TPS}(\cdot)$: Thin Plate Spline transformation;
Output: adversarial patch $P_{adv}$;
  $P_r \leftarrow random$, $e \leftarrow 0$, $i \leftarrow 0$, $\Delta P_r \leftarrow 0$, $\theta \leftarrow 0$, $P_{adv} \leftarrow (P + P_r)$
  while $e < ep_{\max}$ do
      $e \leftarrow e + 1$
      for all $x_i \in \chi$ do
          $P'_{adv} = \mathrm{TPS}(P_{adv})$
          Combine the patch with the object image
          $x_{i,adv} = x_i \bigoplus P'_{adv}$, choose random physical transformation $\mathrm{T_p}()$
          $x^T_{i,adv} = \mathrm{T_p}(x_{i,adv})$
          Getting attention map $M_y$ by Grad-CAM++ algorithm:
          $M_y = \mathrm{CAM}(\mathrm{N}(x^T_{i,adv}), y)$
          if $e > ep_{\theta}$ then$\theta = 1$
          end if
          $\arg\min_{\Delta P_r} \mathcal{L} = \mathcal{L}_{\mathrm{CAM}}(M_y) + \theta \mathcal{L}_{\mathrm{Class}}(\mathrm{N}(x^T_{i,adv}), y, y') + \mathcal{L}_{\mathrm{nps}}(P_{adv}) +$
          $\mathcal{L}_{\mathrm{Con}}(P, \mathcal{L}_{\mathrm{TV}}(P_{adv}))$
          $P_{adv} = P + (P_r + \Delta P_r)$
      end for
  end while
```


(a) Real human images

(b) Virtual 3D human images

Fig. 3. Different real human images and synthetic 3D human model images.

as clothing, feature, hair colour, and body size) as our training set to generate our adversarial patches. We also use virtual 3D human models made by 3D-Max as real-world simulations. We attach the generated adversarial patch to the virtual 3D models and select the position and number of the adversarial patch to simulate 3D distortion. It can control different camera viewpoints, heights, environmental backgrounds and light intensities to simulate complex external physical environment transformations, such as in Fig. 3.

Target models. To show the transferability of adversarial patches, we use common models with different structures. Specifically, we use ResNet-50 (He et al., 2016) for the classifier training task and Faster R-CNN (Ren et al., 2017) for the detector training task. We then use SqueezeNet (Iandola et al., 2016), VGG16 (Simonyan and Zisserman, 2015), ResNet-50, ResNet-101 (He et al., 2016), and DenseNet (Huang et al., 2016) as our classifier test tasks and Yolo-V2 (Redmon and Farhadi, 2017), Yolo-V3 (Redmon and Farhadi, 2018), Yolo-V5, and Faster R-CNN as our detector test tasks. We only obtain the output vector of the networks without any modification. These networks use the Pascal VOC2007 and ImageNet datasets for pretraining.

Evaluation Metric. We use accuracy to measure the impact of adversarial patches on classifier tasks. We adopt precision $p_{0.5}$ as the metric for the detector task following Chen et al. (2018) and Huang et al. (2019), which can measure the probability of whether the detector can hit the true category. We set the default threshold of non-maximum suppression (NMS) as 0.3 and the confidence threshold as 0.5.

Implementation setting. We use the Adam (Kingma and Ba, 2015) as our optimizer, and the initial learning rate is set to 0.05. We set $\lambda = 1 \times 10^{-3}$, $\alpha = 1 \times 10^{-5}$ and $\beta = 1 \times 10^{-5}$. The train epoch is set to 1000. The size of the adversarial patch is set

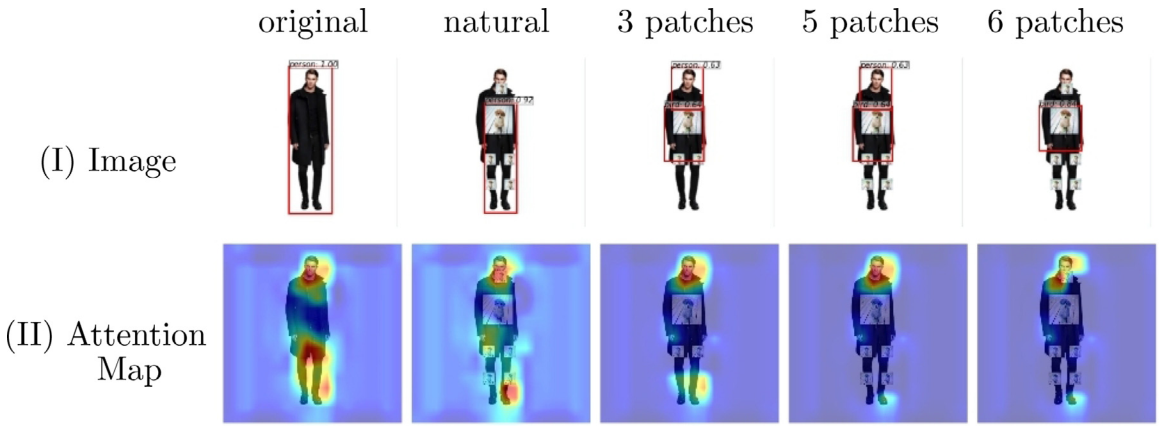


Fig. 4. The human images with different numbers of patches in virtual scene. (I) is detection results of images. (II) displays results with its class attention map. Column 1 displays results without adversarial patches. Column 2 shows detection results with natural patches. Column 3–5 display results with different numbers of adversarial patches.

Table 2
Using raw clean images and images with adversarial patches, we get the accuracy of ResNet-50, ResNet-101, VGG16, SqueezeNet, and DenseNet networks.

Network	ResNet-50	ResNet-101	VGG16	SqueezeNet	DenseNet
Raw	73.4%	89.6%	88.2%	60.3%	70.5%
Adversarial	0.6%	46.9%	25.7%	41.3%	58.8%

Table 3
The $p_{0.5}$ of fooling the Faster R-CNN, Yolo-V2, Yolo-V3 and Yolo-V5 detector networks with different number of patches.

Network	Faster R-CNN	Yolo-V2	Yolo-V3	Yolo-V5
3 patches	32.7%	53.1%	55.7%	74.1%
6 patches	8.8%	26.4%	39.1%	48.6%

Table 4
This table shows the average precision $p_{0.5}$ when we fool the Faster R-CNN and Yolo-V3 networks. With different angles, brightness, and environments, we compare the different numbers of adversarial patches in images with original and natural images.

$p_{0.5}(\%)$							
Network		Faster R-CNN			Yolo-V3		
l_0	angle	a_0	a_{15}	a_{30}	a_0	a_{15}	a_{30}
	Original	100.0	99.9	100.0	100.0	100.0	99.98
	Nature	99.8	99.2	99.7	100.0	99.9	99.1
	3 patches	75.5	75.1	69.7	72.5	76.6	68.8
	5 patches	15.1	22.4	20.2	50.2	51.3	50.1
	6 patches	7.2	8.1	12.1	25.0	27.1	26.2
l_1	angle	a_0	a_{15}	a_{30}	a_0	a_{15}	a_{30}
	Original	100.0	99.9	100.0	100.0	98.9	99.7
	Natural	99.7	100.0	100.0	99.8	99.8	99.4
	3 patches	72.1	72.7	69.8	63.1	70.1	67.09
	5 patches	22.0	24.4	19.2	45.8	40.3	41.8
	6 patches	10.8	9.1	13.7	24.3	35.1	28.8

to 100 * 100 pixels, and the batch size is set to 8. All of our codes are implemented in PyTorch. Training and testing are conducted on an NVIDIA RTX2060-6GB GPU.

4.2. Adversarial digital attacks in the digital world

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

For attacking classification tasks in the digital world, as shown in Table 2, adversarial patches are generated on ResNet50 and tested on SqueezeNet, VGG16, ResNet-101, ResNet-50 and DenseNet classifier networks. Experiments showed that our scheme can fool different classification networks. The best result is achieved on ResNet50, and it has excellent transferability on other networks.

For attacking detection tasks in the digital world, as shown in Table 3, adversarial patches are generated on Faster R-CNN and tested on the Faster R-CNN, Yolo-V2, Yolo-V3 and Yolo-V5 detector networks. And then we tested different numbers and positions of adversarial patches. The confidence threshold of all models is set as 0.5 for evaluation. We found that the greater the number of patches, the better the result. It is reasonable because the number of patches affects the coverage area of the patches directly. Meanwhile, the adversarial patches for the face and the torso have the best interference to the network. We believe that models focus more on locations with prominent features, such as human faces.

Through the attention maps of the models in Fig. 4, it is found that the number and position of the patches will affect the position of the model's attention. The more patches there were, the more distracted the attention and the better the interference effect. We can observe this in the attention maps of different models. When we put an adversarial patch in a specific location (such as faces), the value of that region in the attention map will drop, which will affect the model's attention and fool the model.

4.3. 3D Models simulation experiment

We use the virtual 3D human models rendered by 3D-Max to simulate the complex real world and render the generated adversarial patches to the 3D models. Then we put 3D models in different backgrounds. We modify the brightness of the environment by using two levels of brightness, l_0 and l_1, where l_0 means normal brightness and l_1 means low brightness. We also set three different horizontal camera angles, a_0, a_1, and a_2, where a_0=0°, a_1=15°, and a_2=30°. We generate 200 pictures with different backgrounds, brightness, and viewing angles. Meanwhile, the position and number of patches are different, and comparative tests are conducted under different detector models, as shown in Fig. 5.

Our patches are generated on Faster R-CNN. As shown in Table 4, with the different numbers of patches and the 3D models with untrained natural patches, we compared the $p_{0.5}$ of Faster R-CNN and Yolo-V3 networks while using the 3D models without any patches as the benchmark. When the brightness is l_1, the interference effect is poor because lower brightness will affect the capture of patches details, thereby reducing the effect of patching. The angle will also affect the effectiveness of the patch. We found that the result is best at angle a_0. In other words, the larger the

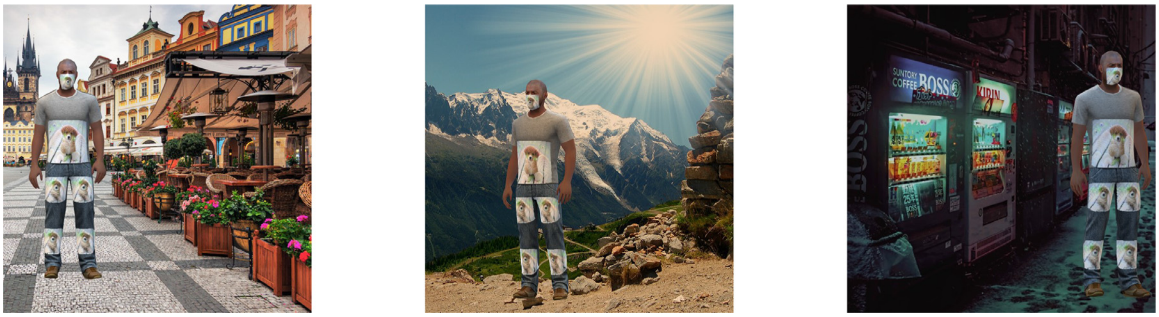
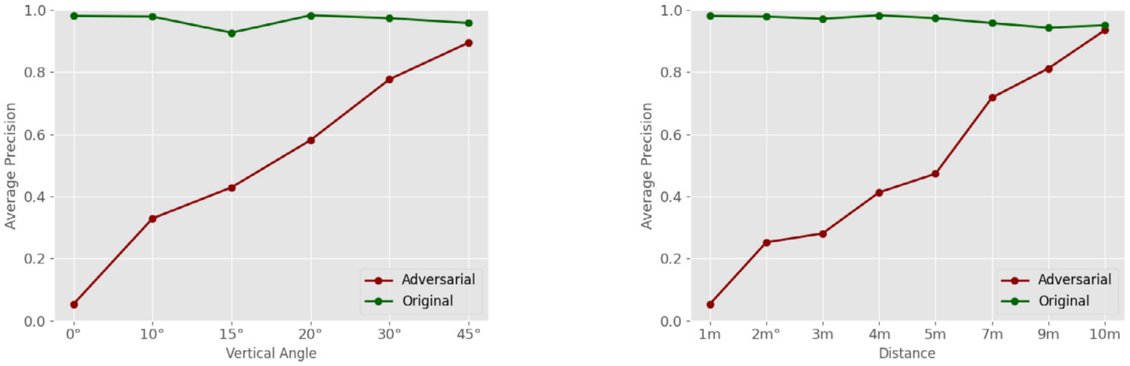


Fig. 5. We used different viewing angles, brightness, and backgrounds in different virtual environments to test the robustness of generated adversarial patches.



(a) Vertical Angle

(b) Distance

Fig. 6. The average precision $p_{0.5}$ of detectors under different vertical angle and distance conditions.

Table 5
Performance comparison with prior arts of physical attacks under the same settings in Faster R-CNN and Yolo-V5 networks to get average precision $p_{0.5}$.

Scheme	Original	Natural	Fasc (Thys et al., 2019)	UPC (Huang et al., 2019)	DAS (Wang et al., 2021)	Ours
Faster R-CNN	100.0%	99.8%	90.6%	15.0%	11.3%	7.2%
Yolo-V5	100.0%	99.7%	97.1%	84.6%	68.6%	56.4%

angle is, the worse the effect because different angles will have different occlusions on the patches, which induces low-quality attacks. We observe that the average precision almost stays at the same level via attacking “Natural patches” scheme which indicates that simply using natural images as adversarial patches is invalid for physical attacks. By contrast, our scheme yields a distinct drop of $p_{0.5}$, among which “6 patches” incur the highest performance decrease. This is no surprise because using more generated patches leads to a higher fooling rate.

We also plot the relationship between the average precision $p_{0.5}$ and vertical viewing angle/distance as in Fig. 6. At a distance of about 9m is the maximum range of effect of our adversarial patches, further away our patches will no longer have a realistic effect. It can be concluded that when the absolute value of the vertical viewing angle and distance between the person and the camera becomes larger, camouflage patterns are captured with lower quality, thus weakening the attacks.

As shown in Table 5, with the 3D models without any patches as the benchmark, we use a “6 patches” scheme to compare our scheme with natural patches and other proposed adversarial attack methods (such as Fasc Thys et al., 2019, UPC Huang et al., 2019, and DAS Wang et al., 2021) on the same virtual 3D models in simulated environments. We use the adversarial patches provided in Fasc, UPC and reconstructed DAS for human images. The average precision shows that our scheme outperforms existing methods in the same environments and can attack the up-to-date detector networks (such as Yolo-V5). This is due to we attack and divert the shared detection box attention of a network to realize a universal attack and we fool the classifier of the object detection models to improve the performance of the adversarial patches in fewer patches. We can observe through the model’s attention map shown in Fig. 7. that the number of patches, location, brightness, and angle will all affect the model’s attention map. The smaller the angle and the larger the number of patches, the more distracted the detection networks’ attention will be.

4.4. Adversarial attacks in the real world

To verify the effectiveness of our adversarial attacks in the real world, we enlarge the generated adversarial patches to about 20 cm to 10 cm and printed them using Konica Minolta Bizhub C554e printers. The patches are attached to clothing and masks to transfer to the real world. We use IPhone11 to take photos and send the obtained photos to the Faster R-CNN network.

As shown in Fig. 8, the attacker successfully fooled the object detector. Our adversarial patches have different degrees of distortion, occlusion, and angular interference, which proves that our adversarial attack has strong robustness in the real world. Meanwhile, adversarial attacks are still effective against targets at different distances. Through Faster R-CNN’s attention maps, we found that the model’s attention has been partially dis-

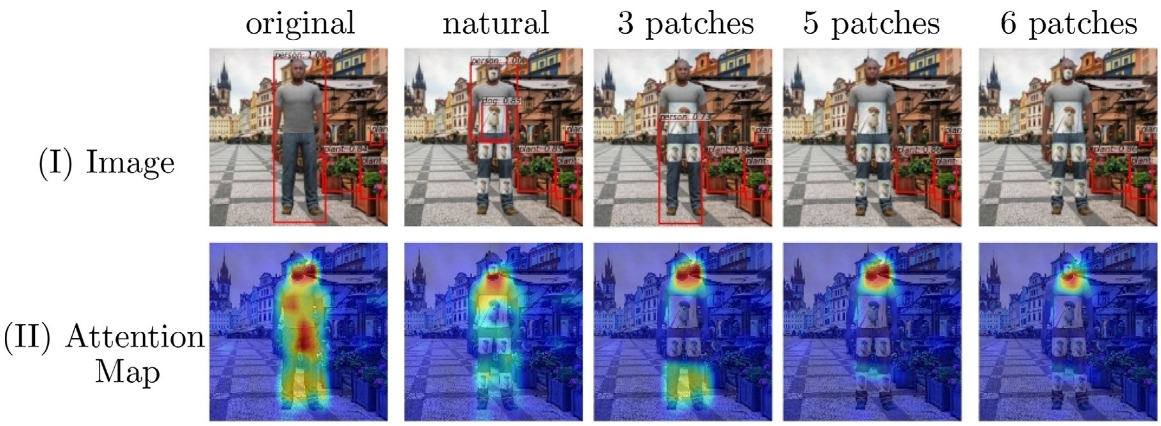


Fig. 7. The virtual 3D models with different numbers of patches in virtual scene. (I) is detection results of images. (II) displays results with its class attention map. Column 1 displays results without adversarial patches. Column 2 shows detection results with natural patches. Column 3–5 display results with different numbers of adversarial patches.

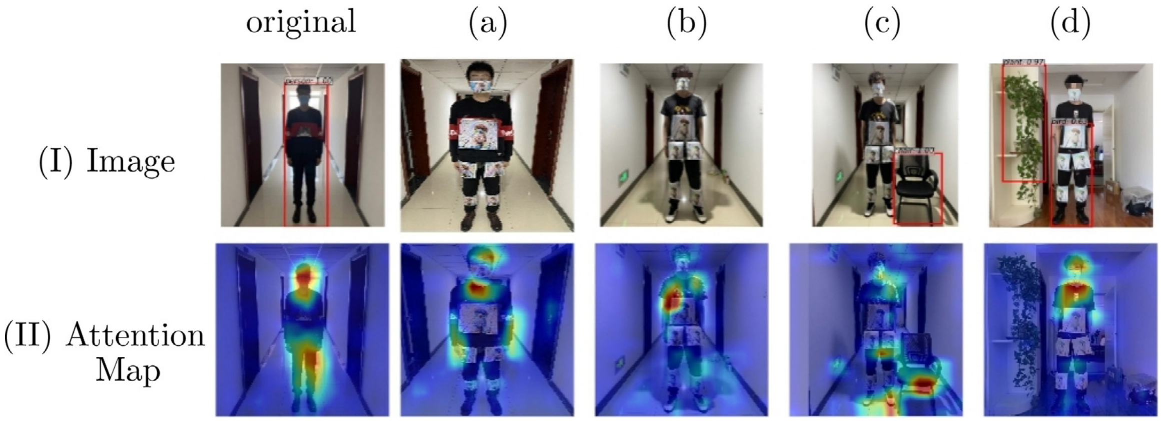
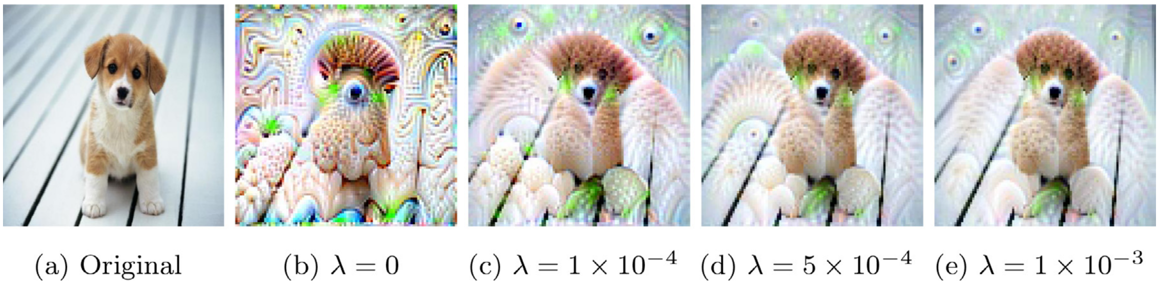


Fig. 8. Fooling the object detector, Faster R-CNN, in the physical space. Physical attacks in real world. (I) shows results with patches under different viewing conditions. (II) displays results with its class attention map. Column (a)-(b) display results with adversarial patches under different environments.



(a) Original (b) $\lambda = 0$ (c) $\lambda = 1 \times 10^{-4}$ (d) $\lambda = 5 \times 10^{-4}$ (e) $\lambda = 1 \times 10^{-3}$

Fig. 9. The adversarial patches with different value of λ.

tracted. The result indicates that our adversarial patches pose a significant threat to the detector's security system in the real world.

4.5. Visual recognition experiments

To reduce the visual attention of the human eyes, we use semantic constraints. For the different coefficients λ of semantic constraints, the value of λ directly affects the distortion of the picture. As shown in Fig. 9, we found that the more minor λ, the more distorted the corresponding adversarial patches, and the more unnatural the resulting picture. To keep balance with effectiveness of patches and naturalness, in this paper, we set $\lambda = 1 \times 10^{-3}$ to ensure the generated adversarial patches are semantically meaningful.

Table 6
The average precision $p_{0.5}$ for adversarial patches with different original image in 6 patches.

Original Image	Natural	Dog	Cat	Bird
$p_{0.5}$	100.0%	7.2%	9.2%	9.8%

As shown in Table 6, even for different original images (such as cat, and dog), the generated adversarial patches can similarly reduce the accuracy of detection models.

4.6. Ablation study

To verify the effectiveness of the adversarial patches we propose, we perform ablation studies with the "6 patches". We use

Table 7
Ablation Study.

$\mathcal{L}$	Original	Random noise	$\mathcal{L}_{\text{Class}}$	$\mathcal{L}_{\text{CAM}}$	$\mathcal{L}_{\text{CAM}} + \mathcal{L}_{\text{Class}}$ (Ours)
$p_{0.5}$	100.0%	99.8%	41.4%	61.0%	7.2%

average precision $p_{0.5}$ to measure the attack effect of the different loss functions and random noise as a baseline. Experimental results show that the attack performance drops significantly after removing some components, proving the effectiveness of our attack framework. The specific results are shown in Table 7.

5. Conclusion and future work

In this paper, we propose a misleading classifier and attention attack to generate adversarial patches to achieve an adversarial attack on a human in the real world. Specifically, we successfully fooled Faster R-CNN and Yolo networks by misleading deep neural networks' attention mechanisms and classifiers. We stick the generated adversarial patches on clothing and masks, and humans can avoid the capture of object detectors by wearing such clothing. At the same time, to ensure the robustness of the patch, we simulated semantic constraints and real environmental disturbances and simulated the non-rigid surface distortion of clothing. We experimented with 2D images and 3D models to prove their effectiveness in the digital world. We experimented in the real world and successfully transferred the digital adversarial patch to the real world with a good attack result. These studies have proved the vulnerability of existing object detectors in the face of adversarial attacks.

With the rapid development of CV, adversarial offensive and defensive problems will also be prominently presented. In the future, we hope to attack more advanced object detection networks and conduct research on adversarial defences to promote the progress and development of the security field of computer vision.

Funding

This work is supported by National Cryptography Development Fund (No. MMJJ20180110), National Natural Science Foundation of China (No. 62102300) and Shandong Social Science Planning Project (No. 21CSDJ30).

Declaration of Competing Interest

The authors declare that they have no competing interests.

Data availability

Data will be made available on request.

References

Adjabi, I., Ouahabi, A., Benzaoui, A., Taleb-Ahmed, A., 2020. Past, present, and future of face recognition: areview. Electronics 9 (8).

Boveiri, H.R., Khayami, R., Javidan, R., Mehdizadeh, A., 2020. Medical image registration using deep neural networks: a comprehensive review. Comput. Electr. Eng. 87, 106767.

Carlini, N., Wagner, D.A., 2017. Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22–26, 2017. IEEE Computer Society, pp. 39–57.

Chambolle, A., Caselles, V., Cremers, D., Novaga, M., Pock, T., 2010. An Introduction to Total Variation for Image Analysis. De Gruyter, pp. 263–340.

Chattopadhyay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N., 2017. Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks. CoRR. abs/1710.11063.

Chen, R., Li, Y., Yu, Y., Li, H., Chen, X., Susilo, W., 2020. Blockchain-based dynamic provable data possession for smart cities. IEEE Internet Things J. 7 (5), 4143–4154.

Chen, S.-T., Cornelius, C., Martin, J., Chau, D.H.P., 2018. ShapeShifter: robust physical adversarial attack on faster R-CNN object detector. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, pp. 52–68.

Chen, X., Liu, C., Li, B., Lu, K., Song, D., 2017. Targeted backdoor attacks on deep learning systems using data poisoning. CoRR. abs/1712.05526.

Cheng, W., Sun, Y., Li, G., Jiang, G., Liu, H., 2019. Jointly network: a network based on CNN and RBM for gesture recognition. Neural Comput. Appl. 31 (S-1), 309–323.

Chowdhary, K., 2020. Natural language processing. Fundam. Artif. Intell. 603–649.

Donato, G., Belongie, S.J., 2002. Approximate thin plate spline mappings. In: Computer Vision - ECCV 2002, 7th European Conference on Computer Vision, Copenhagen, Denmark, May 28–31, 2002, Proceedings, Part III. Springer, pp. 21–31.

Duan, R., Mao, X., Qin, A.K., Chen, Y., Ye, S., He, Y., Yang, Y., 2021. Adversarial laser beam: Effective physical-world attack to DNNs in a blink. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19–25, 2021. Computer Vision Foundation / IEEE, pp. 16062–16071.

Evtimov, I., Eykholt, K., Fernandes, E., Kohno, T., Li, B., Prakash, A., Rahmati, A., Song, D., 2017. Robust physical-world attacks on machine learning models. arXiv preprint arXiv:1707.08945 2 (3), 4.

Gao, C., Li, B., Wang, Y., Chen, W., Zhang, L., 2021. Tenet: A neural network model extraction attack in multi-core architecture. In: GLSVLSI '21: Great Lakes Symposium on VLSI 2021, Virtual Event, USA, June 22–25, 2021. ACM, pp. 21–26.

Girshick, R., 2015. Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV).

Girshick, R.B., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23–28, 2014. IEEE Computer Society, pp. 580–587.

Goodfellow, I.J., Shlens, J., Szegedy, C., 2015. Explaining and harnessing adversarial examples. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings.

Grigorescu, S.M., Trasnea, B., Cocias, T.T., Macesanu, G., 2020. A survey of deep learning techniques for autonomous driving. J. Field Rob. 37 (3), 362–386.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016. IEEE Computer Society, pp. 770–778.

Huang, G., Liu, Z., Weinberger, K.Q., 2016. Densely connected convolutional networks. CoRR. abs/1608.06993.

Huang, L., Gao, C., Zhou, Y., Zou, C., Xie, C., Yuille, A.L., Liu, N., 2019. UPC: learning universal physical camouflage attacks on object detectors. CoRR. abs/1909.04326.

Iandola, F.N., Moskewicz, M.W., Ashraf, K., Han, S., Dally, W.J., Keutzer, K., 2016. SqueezeNet: alexnet-level accuracy with 50x fewer parameters and <1mb model size. CoRR. abs/1602.07360.

Kim, J.M., Choe, J., Akata, Z., Oh, S.J., 2021. Keep CALM and improve visual feature attribution. CoRR. abs/2106.07861.

Kingma, D.P., Ba, J., 2015. Adam: a method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings.

Li, Y., Xu, X., Xiao, J., Li, S., Shen, H.T., 2021. Adaptive square attack: fooling autonomous cars with adversarial traffic signs. IEEE Internet Things J. 8 (8), 6337–6347.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S.E., Fu, C., Berg, A.C., 2016. SSD: single shot multibox detector. In: Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I. Springer, pp. 21–37.

Liu, Y., Chen, X., Liu, C., Song, D., 2017. Delving into transferable adversarial examples and black-box attacks. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings. OpenReview.net.

Liu, Y., Ma, X., Bailey, J., Lu, F., 2020. Reflection backdoor: a natural backdoor attack on deep neural networks. In: Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X. Springer, pp. 182–199.

Moosavi-Dezfooli, S., Fawzi, A., Fawzi, O., Frossard, P., 2016. Universal adversarial perturbations. CoRR. abs/1610.08401.

Moosavi-Dezfooli, S., Fawzi, A., Frossard, P., 2016. DeepFool: a simple and accurate method to fool deep neural networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016. IEEE Computer Society, pp. 2574–2582.

Morgulis, N., Kreines, A., Mendelowitz, S., Weisglass, Y., 2019. Fooling a real car with adversarial traffic signs. CoRR. abs/1907.00374.

Narodytska, N., Kasiviswanathan, S.P., 2016. Simple black-box adversarial perturbations for deep networks. CoRR. abs/1612.06299.

Papernot, N., McDaniel, P.D., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A., 2016. The limitations of deep learning in adversarial settings. In: IEEE European Symposium on Security and Privacy, EuroS&P 2016, Saarbrücken, Germany, March 21–24, 2016. IEEE, pp. 372–387.

Pedraza, A., Deniz, O., Bueno, G., 2021. Really natural adversarial examples. Int. J. Mach. Learn. Cybern. 1–13.

Redmon, J., Divvala, S.K., Girshick, R.B., Farhadi, A., 2016. You only look once: unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016. IEEE Computer Society, pp. 779–788.

Redmon, J., Farhadi, A., 2017. YOLO9000: better, faster, stronger. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017. IEEE Computer Society, pp. 6517–6525.

Redmon, J., Farhadi, A., 2018. YOLOv3: an incremental improvement. CoRR. abs/1804.02767.

Ren, H., Huang, T., Yan, H., 2021. Adversarial examples: attacks and defenses in the physical world. Int. J. Mach. Learn. Cybern. 12 (11), 3325–3336.

Ren, S., He, K., Girshick, R.B., Sun, J., 2017. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 39 (6), 1137–1149.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad–CAM: visual explanations from deep networks via gradient-based localization. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22–29, 2017. IEEE Computer Society, pp. 618–626.

Sharif, M., Bhagavatula, S., Bauer, L., Reiter, M.K., 2016. Accessorize 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, Vienna, Austria, October 24–28, 2016. ACM, pp. 1528–1540.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I.J., Fergus, R., 2014. Intriguing properties of neural networks. In: 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14–16, 2014, Conference Track Proceedings.

Thys, S., Ranst, W.V., Goedemé, T., 2019. Fooling automated surveillance cameras: adversarial patches to attack person detection. CoRR. abs/1904.08653.

Wang, J., Liu, A., Yin, Z., Liu, S., Tang, S., Liu, X., 2021. Dual attention suppression attack: generate adversarial camouflage in physical world. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19–25, 2021. Computer Vision Foundation/IEEE, pp. 8565–8574.

Wu, Z., Lim, S., Davis, L., Goldstein, T., 2019. Making an invisibility cloak: real world adversarial attacks on object detectors. CoRR. abs/1904.08653.

Yuan, K., Tang, D., Liao, X., Wang, X., Feng, X., Chen, Y., Sun, M., Lu, H., Zhang, K., 2019. Stealthy porn: understanding real-world adversarial images for illicit online promotion. In: 2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19–23, 2019. IEEE, pp. 952–966.

Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks. In: Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part I. Springer, pp. 818–833.

Zhang, Q., Zhu, S., 2018. Visual interpretability for deep learning: a survey. Frontiers Inf. Technol. Electron. Eng. 19 (1), 27–39.

Zhang, Y., Huang, X., Chen, X., Zhang, L.Y., Zhang, J., Xiang, Y., 2020. A hybrid key agreement scheme for smart homes using the Merkle puzzle. IEEE Internet Things J. 7 (2), 1061–1071.

Zhou, B., Khosla, A., Lapedriza, À., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016. IEEE Computer Society, pp. 2921–2929.

Haotian Zhang received the bachelor's degree of software engineering from Qingdao University of Science and Technology, Shandong, China, in 2019. He is currently working toward the master's degree with the School of Cyber Science and Engineering, Qufu Normal University, Shandong, China. His research interests include privacy-preserving machine learning, federal learning and adversarial attacking.

Xu Ma received the bachelor's degree of computer science from Ludong University, China, in 2008, and the master's and PhD degrees of information security and cryptography from Sun Yat-sen University, China, in 2010 and 2013, respectively. He is also a postdoctor of the Department of Cyberspace Security, Xidian University, China. He is currently an associate professor with the School of Cyber Science and Engineering, Qufu Normal University, China. His research focuses on applied cryptography, outsourcing computation, and privacy-preserving machine learning.