Invisibility Spell: Adversarial Patch Attack Against Object Detectors

Jianyi Zhang(✉), Ronglin Guan, Zhangchi Zhao, Xiuying Li, and Zezheng Sun

Beijing Electronic Science and Technology Institute, Beijing 100070, China
zjy@besti.edu.cn

Abstract. Since image recognition technology was confirmed to be vulnerable to attack, research on adversarial attack methods has emerged one after another. There are also many studies on the adversarial patch attack methods of the mainstream object detector YOLO (You Only Look Once) series. However, with the emergence of more advanced object detectors such as YOLOv5 [1], these existing attack methods have lost their effectiveness in both digital and physical attacks. To solve this problem, in this work, we propose a new adversarial attack method, InviSpell, for the new network structure, which designs a new loss function to generate adversarial patches based on the network structure of the YOLOv5 model. In this method, a new optimization strategy is proposed that uses the target confidence to adjust the optimization weights. The experiments show that the adversarial patch generated by our method can reduce the mean average precision of YOLOv5 from 71.24% to 2.86%. Our method has good attack effects in both the digital and physical worlds on YOLO v2 to v5 and Faster R-CNN. Moreover, the posters and T-shirts printed with the adversarial patch have good attack effects on the object detector and good transferability between different detectors and training datasets.

Keywords: YOLO · Adversarial attacks · Adversarial patch · Object detection

1 Introduction

In recent years, as the current object detection technology has been proven vulnerable to attack [2–7], research on adversarial methods has emerged one after another. The YOLO (You Only Look Once) series of object detection models stands as one of the most extensively employed within the current landscape. Despite numerous efforts towards adversarial approaches for YOLO, most of the current adversarial methods cannot achieve a qualified attack effect [8,9] after the appearance of the YOLOv5 [1] object detector. That is because YOLOv5 object has a more robust structure which makes the traditional attack method converge on a not ideal adversarial performance. In addition, prior research often tends to overlook the feasibility of adversarial patches in real-world scenarios.

H. Duan et al. (Eds.): SecureComm 2023, LNICST 567, pp. 88–109, 2025.
https://doi.org/10.1007/978-3-031-64948-6_5

The image recognition adversarial attacks that have appeared are roughly divided into two categories. The first category, like [10], is to deceive the deep learning model by adding targeted noise that cannot be detected by the human eye to the image to be recognized, and generate incorrect prediction results. This approach is often used in digital attacks, where noise-injected digital images are used as inputs to deep learning models. Although this attack method has good concealment, it usually needs to add noise to the whole image, which makes it difficult to achieve physical attacks in the real world. The second category is to make the deep learning model unable to recognize the key objects in the image by pasting an adversarial patch in the part of the image [11,12]. This attack method is generated with the goal of attacking the real world, and the input of the model is the images obtained from the real world by devices such as cameras [13,14].

In this work, we propose a new adversarial attack method, InviSpell, for the new object detectors' network structure. We focus on the second category of physical attacks. With cameras now abounding both indoors and outdoors, the existence of an adversarial patch can have considerable implications for public security. At present, there are many studies on such attack methods, but most of them are attacked against the old versions of detectors (*i.e.* YOLOv1 [15], YOLOv2 [16], and Faster R-CNN [17]). The patches generated by this target have no attack effect on the newer models (i.e., YOLOv4 and YOLOv5), shown in Fig. 1. The reason is that the new models, such as YOLOv4 [18] and YOLOv5, have great changes in the network structure compared to the old models.

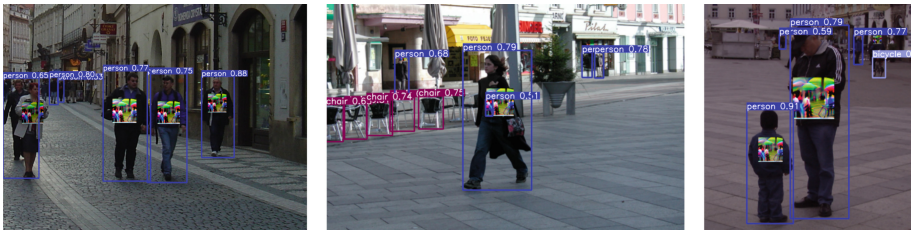


Fig. 1. The previous adversarial patch cannot attack the new object detection algorithm.

The key point of our solution to this problem is to propose a new adversarial attack method for the new network structure, which designs a new loss function according to the network structure of the YOLOv5 model and the feasibility of adversarial patches in real-world scenarios. At the same time, we also propose a new optimization strategy. This strategy can adjust the adversarial patch's optimization weight according to the target's confidence in different images. Moreover, the strategy can reduce the confidence of the bounding box to 0 as much as possible through the gradient descent method so that the target "disappears" in the camera loaded with the new object detector model.

This paper is structured as follows: Sect. 2 presents the existing research on adversarial image attacks. Section 3 discusses how we generate new adversarial patches for new model changes. In Sect. 4, we train on the INRIA dataset and evaluate digital attacks on the generated patches, in addition to capturing and evaluating real-world physical attacks using a camera. In Sect. 5, we summarize the experimental results and conclusions.

The main contributions of this work can be summarized as follows:

- To the best of our knowledge, our work is the first research to generate the adversarial patch to attack the new YOLO version. The experimental results demonstrate that InviSpell has the best attack effect and achieves the purpose of misleading the object detector from YOLO v2 to v5.
- We propose a new optimization strategy for the training process. We use the target confidence to adjust the optimization weight of the adversarial patch in each batch training set. The new strategy shows better attack performance than the state-of-the-art methods.
- We evaluate the effectiveness of our patches not only in the digital world, but also in the physical world under different circumstances. We get satisfactory results and reduce the detection recall rate with our adversarial poster and T-shirt.
- We analyze and explore how the initial parameter, transformation and patch size of the adversarial patch affect the attack performance. We present experiments to show the differences and evaluate their robustness for future designs.

2 Related Work

2.1 Image Adversarial Attack

Bigio et al. [19] first discovered the existence of adversarial attacks, and Szegedy et al. [20] were the first to implement adversarial image attacks. They used disturbances that cannot be detected by the human eye to add to the images to be classified, deceiving the image classification network at that time. After that, different adversarial perturbation generation methods continued to appear, such as Goodfellow et al. [21] with the Fast Gradient Symbol Algorithm (FGSM) and Madry et al. [22] with the Projected Gradient Descent (PGD). Today, adversarial attacks can be roughly divided into two categories: digital adversarial attacks against the digital world and physical adversarial attacks against the real world. Our work focuses on achieving simultaneous successful attacks on the digital and physical worlds on more advanced object detectors.

2.2 Digital Adversarial Attack

A general attack on digital images is to inject perturbed pixels before they are fed into an object detector. Earlier works, such as the Fast Gradient Symbolic Method (FGSM) and Projected Gradient Descent (PGD), are attacked by perturbing the whole graph. This kind of attack is very suitable for the field of

image classification. In the field of image recognition, attacks on object detectors require not only labels that mislead object classification, but also misleading locations where objects exist [23]. Liu et al. [24] specially design DPatch for the object detector. This method adds a rectangular patch filled with training pixels to the image, so that the object detector cannot detect all the targets in the image, which has a good attack effect on YOLOv2 and Faster R-CNN detectors. Lee and Kolter [25] improve DPatch. They changed the training strategy of adversarial patches and made the pixels of the patch closer to the image. In addition, they made changes in the rotation, position, brightness, etc. of the patch to evaluate the robustness of the patch. Lu et al. [26] proposed an adversarial attack method utilizing adaptively sized rectangular patches for the problem of aircraft detection in remote sensing images (RSIs). The method can adapt to the change of aircraft scale in RSIs, and successfully attack the YOLOv3 [27] object detector. However, the attack effect on YOLOv5 is not ideal. These digital attacks require digitally injecting attacked pixels into images, which is often not feasible in real-world vision systems such as webcams and self-driving cars.

2.3 Physical Adversarial Attack

Adversarial examples captured directly by the camera as input and have attack effects are called physical adversarial examples because they can be effective in the real world [28]. The attacker cannot control the parameters of the camera, etc., and the adversarial samples must have the robustness to successfully attack after different transformations. Hence, the physical adversarial attack is more challenging. Evtimov et al. [29] proposed a real-world image classification attack. They generate a patch for the classification task of stop signs, making the car unrecognizable. This can be challenging as stop signs may appear in the camera at different angles and orientations. Simen Thy et al. [30] applied an adversarial attack to the scene of person detection. They placed the patch in the middle of the target instead of the upper left corner like DPatch. Similarly, they successfully attacked the YOLOv2 object detector. Xu et al. [31] proposed a TPS-based method, applied a patch to a T-shirt and achieved good results in attacking YOLOv2 and Faster R-CNN object detectors, which greatly increased the stealth of the attack. Similar attacks include eyeglass frames [32–34], car license plates [35], posters [14,36], etc.

3 Proposed Method

Our goal is to design a new patch generation strategy so that the generated patches can better deceive current popular object detectors in both digital and physical attack methods. This paper introduces the mainstream object detector YOLO as a unified object detection model. We redefine object detection as a single regression problem, going directly from image pixels to bounding box coordinates and class probabilities. YOLO has the advantages of fast detection and high mean average precision. YOLOv5 in the YOLO series is the main

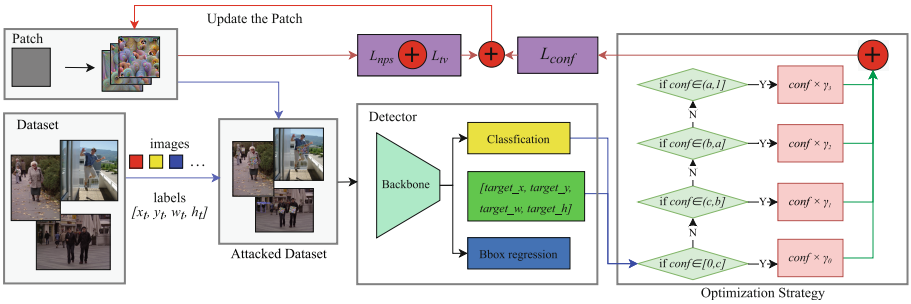


Fig. 2. Overview of generating new patches. First, we prepare the dataset and the initial patch, and render the initial patch into the image in the dataset to obtain the attack dataset. Then, we take the attack dataset as the input to the object detector and obtain the loss score. Finally, the adversarial patch is updated according to the minimization of total loss score. The optimization process is performed as a closed loop through backpropagation.

attack model in this method. We assume that the network structure used by the target model is known, and its network weight is YOLOv5n6 trained from public datasets on the network. Our research found that the difference between the YOLOv5 object detector and the previous work is that although it also sets the anchor in advance, it can adaptively calculate the best anchor in different training sets during training, so as to update the anchor value. At the same time, a new data augmentation module is added, and a new loss function is adopted, which makes the previous patch unable to deceive YOLOv5, making the attack challenge. To cope with this improvement, we adopt a loss function adapted to the new network to generate adversarial patches, so that the object detector YOLOv5 is successfully attacked. At the same time, we have studied the previous adversarial patch generation methods and found that, in order to make the target at the edge of the image also attacked with a high success rate, only object confidence is used as the main loss to be optimized. But in fact, our optimization focus on loss should be different. Therefore, we also propose a new training strategy, which can adjust the optimization weight of the patch according to the object confidence of different images in the dataset, increasing the attack's success rate. In this work, we focus on attacks on the person class. Figure 2 shows an overview of our framework.

3.1 Generating Patch

First, we prepare an initial rectangular patch used as the starting point for optimization. We obtain the ground truth box information of the object from the labels of the dataset, which includes five parameters $[Class, x_t, y_t, w_t, h_t]$. The $Class$ represents the category of the object in the ground truth box, x_t and y_t are the position of the upper left corner of the ground truth box, and w_t and h_t are the width and height of the ground truth box. According to these

parameters and the initial prepared patch, adjust the patch size according to a certain proportional relationship and attach it to the center of the object to obtain the image. Next, we input the attacked image set into the YOLOv5 object detector for detection and select different optimization weights combined with other relevant losses to iteratively update the attack patch according to the object loss in each round.

3.2 Detector

Next, we will perform a complete object detection process. It computes a loss based on its own output and ground truth to update every pixel value of the adversarial patch through backpropagation. YOLOv5 is a one-stage detection algorithm that reconstructs object detection as a single regression problem. This detector obtains the bounding box coordinates and class probabilities of objects in one step. YOLOv5 divides the input image into grids, and each grid in the image predicts B bounding boxes and confidence scores for these boxes. These confidence scores reflect the confidence that the box contains an object and the accuracy of the box. If the confidence score is below a certain threshold, the bounding box of the grid prediction is judged not to contain the true object. And these confidence-related data will be output to the next module.

3.3 Loss Function

The detection process of YOLOv5 corresponds to whether its bounding box contains the real object or not is determined by the total object confidence. Therefore, to attack the object detector and disguise the person, we simply filter all bounding boxes containing the person by reducing the object confidence to zero as much as possible. We denote the total object loss of the YOLOv5 object detector by L_{conf}. Given by:

$$L_{conf} = a \times Loss_{obj} + b \times Loss_{rect} + c \times Loss_{clc} \tag{1}$$

where a, b, and c are weight factors, $Loss_{rect}$ is the loss of the rectangular box, $Loss_{obj}$ is the loss of object confidence, and $Loss_{clc}$ is the classification loss. Our training aims to minimize this score, and further optimization for it will be discussed in depth later in this section.

Related to the other two-part losses, we adopt the method [30] mentioned in the literature to make the patch look smoother in the image and easier to obtain by printing. Among them, L_{nps} represents the non-printability score, which represents how difficult the patch is to be printed by a common printer. L_{nps} is as shown in Eq. (2):

$$L_{nps} = \sum_{p_{\text{patch}} \in p} \min_{c_{\text{print}} \in C} |p_{\text{patch}} - c_{\text{print}}| \tag{2}$$

where P_{patch} is a pixel in patch P and c_{print} is a color in a set of printable colors C. Lowering it ensures that the patch can be printed without much distortion.

L_{tv} represents the total variation in the image. This loss score represents the size of the color difference between neighboring pixels in the patch, that is, the smoothness of the image. It affects how difficult the optimizer is to optimize the patch and how natural it is. L_{tv} is as shown in Eq. (3):

$$L_{tv} = \sum_{i,j} \sqrt{(p_{i,j} - p_{i+1,j})^2 + (p_{i,j} - p_{i,j+1})^2} \tag{3}$$

where the subindices i and j refer to the pixel coordinate of the patch P. If the neighboring pixels are more similar in color, the loss score is lower; otherwise, the loss score is higher.

Adding the three losses gives the total objective loss L_{sum} we need to optimize, as shown in Eq. (4):

$$L_{sum} = \alpha L_{nps} + \beta L_{tv} + L_{conf} \tag{4}$$

where α and β are the empirically determined scaling factors and optimise using the *Adam* algorithm.

Because our optimization target is only L_{sum}, other parameters need to be kept unchanged during the training process to ensure that only the pixels of the patch have changed.

3.4 Optimization Strategy

According to the previous chapters, in the total loss L_{sum}, L_{conf} plays a decisive role in object detection, representing the score obtained by a certain category in the detection. The higher the score, the higher the probability of proving that there is a target here. We optimize L_{conf} so that it can fool the object detector by reaching a certain threshold. In order to cope with the improvement of YOLOv5 object detector in anchor and data enhancement, we propose a corresponding optimization strategy for calculating L_{conf} loss. We all know that each image's L_{conf} score of the detected target is also different. We take it for granted that after being attacked, it still has a high L_{conf} score, which means that the current patch has a poor attack effect on the target in this image, so we can increase this round of changes to patch pixels during training. On the contrary, the attack effect is better for the image rounds with low L_{conf} scores. The patch's optimization strength in this round of training can be appropriately reduced because it does not require many changes.

In order to realize this idea, we set different loss weights according to the confidence of the images in each batch during the training process. For rounds with object confidence loss greater than 0.5, we should assign larger weights to make its confidence drop faster, and for rounds with object confidence less than 0.5, we should give smaller weights. This strategy can optimize the direction of the gradient, thereby increasing the success rate of the attack. Finally, the improved L_{conf} is designed as follows Eqs. (5) and (6):

$$L_{conf} = \sum_{i} L_{conf-i} \tag{5}$$

$$L_{conf-i} = \begin{cases} \gamma_0 \times conf & \text{if} \quad conf \leq c \\ \gamma_1 \times conf & \text{if} \quad c < conf \leq b \\ \gamma_2 \times conf & \text{if} \quad b < conf \leq a \\ \gamma_3 \times conf & \text{if} \quad a < conf \end{cases} \tag{6}$$

where i is the index of the batch in an epoch, L_{conf-i} is the L_{conf} score of the i-th batch, and $conf$ is the $Loss_{obj}$ in the current batch. $[\gamma_1, \gamma_2, \gamma_3, \gamma_4]$ is the weight combination of different $conf$ cases in this method. After experiments, we selected the set of weights [0.01, 0.1, 1, 10] value.

3.5 Robustness and Transferability

In order to ensure the physical attack effect of the patch, we make some transformations of the patch before it is pasted to the target, including rotation, scale, noise, and brightness contrast, to simulate the patches' possible transformation in the real world due to sunlight, wrinkle, and so on, then we observe the attack effect after performing various transformations on the final generated patch to verify and ensure that the patch has strong robustness.

A well-performing patch also needs the ability to attack different object detectors, so that it will have better generality. In this work, we set up two scenarios to evaluate the transferability of different patches, dataset-to-dataset and model-to-model. For the first scenario, we use the YOLOv5 object detector to train on one dataset to get the adversarial patch, and apply it to the other two datasets for the attack. In the second scenario, we use the adversarial patch generated by the YOLOv5 object detector on the INRIA [37] dataset to attack other object detectors of the YOLO series trained on the INRIA dataset to verify that it is still valid.

4 Experiment and Results

4.1 Implementation Details

Dataset. In this paper, we mainly use the pedestrian dataset from INRIA to train the adversarial patch. It consists of 614 training images and 288 testing images, including pedestrians with different poses, angles, and sizes. Because the sizes of the images in the dataset are different, we uniformly resize the images to 1280 × 1280 as input, and set the size of the generated adversarial patch to 300 × 300. In addition, to achieve cross-dataset evaluation, we also selected more than 1000 images in the MS COCO [38] dataset and more than 500 images containing people in the Pascal VOC 2007 [39] dataset. Similarly, we resize them to 1280 × 1280.

Pretrained Models. We downloaded several commonly used models in the YOLO series, YOLOv2, YOLOv3, YOLOv4, and YOLOv5 from the official

website, which are all the one-stage detectors. The network weights used by YOLOv3, YOLOv4, and YOLOv5 are YOLOv3tiny, YOLOv4tiny, and YOLOv5n6 respectively. To test the attack effect of the adversarial patch in two-stage detectors, we also attack Faster R-CNN with ResNet101 [40] object detector. For all these detectors, except for YOLOv5, which we trained ourselves from scratch, standard models pre-trained on COCO dataset were adopted to test the transferability of the attack on the network weights.

Testing Parameters. We use the parameters in the Table 1 to generate adversarial patches. For the initial patch before optimization, we selected pure gray, pure white, pure black, random noise, and TV image for training and compared the similarities and differences in the training process and the difference in attack effect. The initial patch size is still set to 300 × 300.

Table 1. Hyperparameters used for training adversarial patch, * is the initial learning rate. The learning rate will change with the rate of loss during training

Hyperparameter	Value
$[a, b, c]$	[0.5, 0.3, 0.2]
α	0.01
β	2
$[\gamma_1, \gamma_2, \gamma_3, \gamma_4]$	[0.01, 0.1, 1, 10]
learning rate	0.01*
batchsize	4
epochs	1000

If the drop rate of the loss still exceeds $1e^{-5}$ after 1000 epochs, then the epochs can be increased moderately to continue the training to achieve the best convergence value.

Evaluation Metrics. To evaluate the effectiveness of the attack method, we use two evaluation methods: Average Precision (AP) and Recall. Precision and recall are calculated as shown in Eqs. (7) and (8). In these two equations, TP denotes the bounding box whose Intersection over Union (IoU) with ground truth is greater than the threshold. FP' denotes two types of bounding boxes, one is the bounding box whose IoU with ground truth is less than the threshold, and the other is the redundant bounding box whose IoU with ground truth is greater than the threshold, but the confidence is not the highest. FN' denotes the objects that is not detected, and it plus TP equals to the number of ground truth.

$$Precision = \frac{TP}{TP + FP'} \tag{7}$$

$$Recall = \frac{TP}{TP + FN'} \tag{8}$$

4.2 Attack in the Digital World

Evaluation of Attack Effects. In this part of experiments, we use several patches to attack on different object detectors. These patches are (1) ADV-OBJ, the patch generated using the best-performing scheme in [30], which is currently the state-of-the-art adversarial patch for YOLOv2 object detector performance, (2) ADV-OBJv5, using the method in [30] and adjusting the loss function to train the generated patches on YOLOv5 without the optimization strategy, (3) InviSpell, an adversarial patch generated using our method, (4) UPC, the universal physical camouflage attack which generated adversarial patch designed to efficiently attack all instances belonging to the same object class [9], (5) TPS, the generated adversarial patch designed for non-rigid objects [31], (6) MIC, A patch generated by training an ensemble of detectors [8], (7) 3D inviCloak, an adversarial patch that can be applied to 3D environments [41], (8) NOISE, patch initialized with random noise, (9) GRAY, an initial patch of pure gray. They are respectively shown in Fig. 3. When generating adversarial patches, we use the pre-prepared INRIA training set to train a generic adversarial patch with an initial size of 300 × 300, and apply the adversarial patch on the test set to evaluate the attack performance.

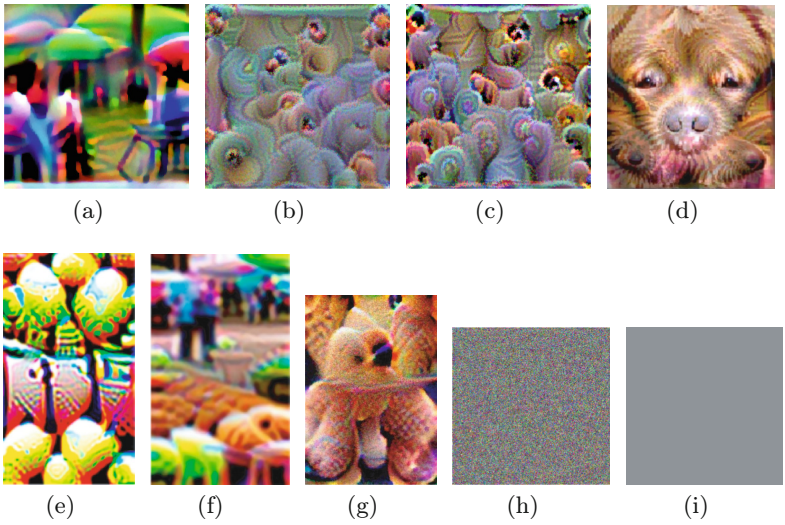


Fig. 3. Seven different adversarial patches and two standard patches used in control experiments.(a) is ADV-OBJ, (b) is ADV-OBJv5, (c) is InviSpell, (d) is UPS, (e) is TPS, (f) is MIC, (g) is 3D InviCloak, (h) is NOISE and (i) is GRAY (Color figure online)

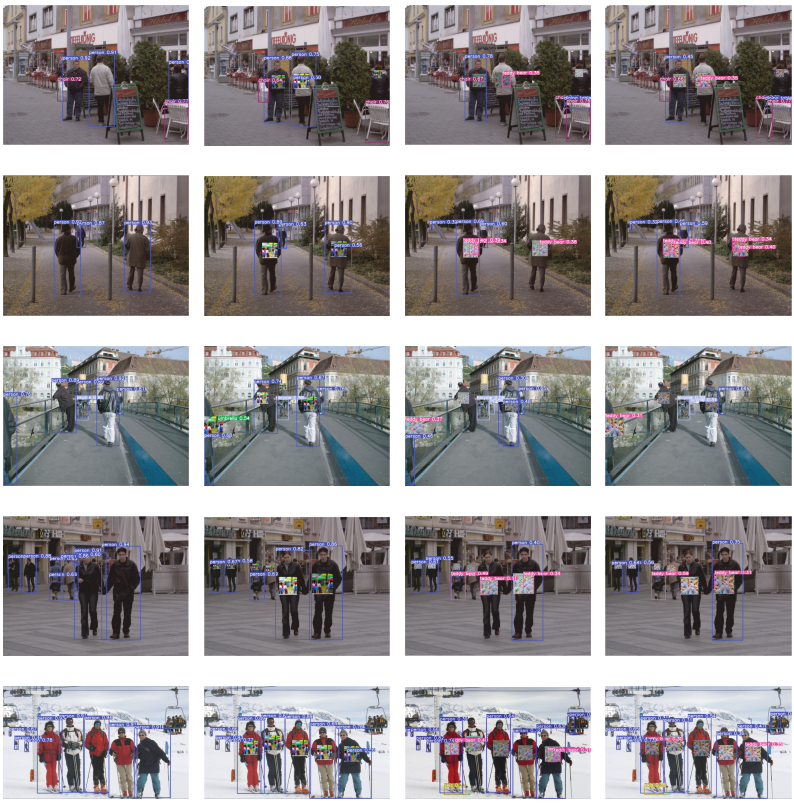
The data in the Table 2 shows that the adversarial patch ADV-OBJv5 and InviSpell generated using our method can achieve better attack results on the new object detectors YOLOv3-v5, and even in the old version of the object

Table 2. Comparison of the attack effects of adversarial patches and standard patches on different target detectors, measured using average precision (AP%).

Patch	Victim			
	YOLOv2	YOLOv3	YOLOv4	YOLOv5
Clean	65.12	70.55	71.51	71.24
Standard Patch				
NOISE	52.15	58.85	67.61	64.15
GRAY	49.53	54.14	56.11	53.94
Adversarial Patch				
ADV-OBJ [30]	**3.15**	22.19	48.16	49.56
ADV-OBJv5 [30]	10.54	15.17	12.64	9.41
TPS [31]	7.15	19.56	52.58	67.23
MIC [8]	10.70	13.98	46.67	61.77
3D InviCloak [41]	8.41	25.25	64.12	63.26
UPC [9]	34.12	53.98	49.76	59.24
InviSpell	7.51	**12.33**	**10.87**	**2.86**

detector YOLOv2, the attack effect is not far from the current best-performing ADV-OBJ for YOLOv2. Compared with the unattacked image (CLEAN) detection AP (71.24%), the best AP obtained by ADV-OBJv5 after attacking the YOLOv5 object detector is 9.41%, a decrease of 61.83%, which is significantly lower than the threshold. While the InviSpell generated by our method achieves the best AP of 2.86% after attacking the YOLOv5 object detector, a decrease of 68.38% and 6.55% lower than that of ADV-OBJv5. At the same time, we can see that the adversarial patches generated by the three different methods of ADV-OBJv5, TPS, MIC and 3D InviCloak have a significant decrease in attack effect on other target detectors except for YOLOv2, and even in the YOLOv5 model, they are even comparable to the standard adversarial patches NOISE and GRAY. In addition, UPC sacrifices more attack effects in order to maintain a more natural appearance. However, InviSpell has good attack performance in previous YOLO models. The experimental results show that different adversarial patches will have different performances for a single object detector. For different object detectors, the same adversarial patch will also have different performances. This is due to the inconsistent network structure of each model, and the different adversarial patches are trained under different network structures. Compared with ADV-OBJv5, InviSpell has a great improvement in different target detectors, and the improvement is about 25–30%, which shows that our method can not only deal with the improvement of the YOLOv5 model compared with other models, but also improve the attack effect of the adversarial patch in other models. Figure 4 shows the attack effects of three different adversarial patches, ADV-OBJ, ADV-OBJv5, and InviSpell, on the INRIA dataset and YOLOv5. The left of each picture is the rendering of CLEAN, and the right

is the detection rendering after adding different adversarial patch attacks. As can be seen from the figure, no matter what kind of adversarial patch can make the person in the image evade the detection of the object detector, InviSpell is more effective and can make almost all the person disappear from the image.



(a) CLEAN (b) ADV-OBJ (c) ADV-OBJv5 (d) InviSpell

Fig. 4. Examples of attack results of adversarial patches generated by three different methods on the INRIA dataset (In order to show the attack effect of different patches more clearly, ADV-OBJ uses 50% confidence as the display threshold of the bounding box, and 30% for ADV-OBJv5 and InviSpell since the detection scores of these attacks are much lower than 0.5).

Transferability of InviSpell Across Datasets. We train on MS COCO, Pascal VOC 2007, INRIA, and a mixture of the three to obtain the corresponding adversarial patches, and then use these four adversarial patches to attack these datasets separately to prove that our adversarial patches can be transferred and effective in different datasets. The experimental data are shown in Table 3.

From the experimental results, we can see that using the adversarial patch trained by INRIA to attack this dataset can achieve the best attack effect, with

Table 3. Average Precision (AP%) obtained by attacking different datasets with adversarial patches trained on corresponding datasets using YOLOv5.

	INRIA	COCO	VOC	Mix
INRIA-InviSpell	2.86	10.84	9.73	10.51
COCO-InviSpell	5.64	6.89	8.31	9.63
VOC-InviSpell	10.16	15.29	3.64	14.77
Mix-InviSpell	9.64	12.48	13.54	8.46
CLEAN	71.24	67.61	69.84	65.17

only 2.86% AP, which means that the target is almost undetectable in the image. The worst attack effect is the adversarial patch attack mixed dataset obtained by VOC training. The AP after the attack is 14.77%, which is 50.4% lower than the clean image (65.17%). It indicates that most of the targets can be hidden. Experimental results demonstrate that the adversarial patches generated by our method have good transferability between different datasets.

Across Models. To verify that the adversarial patches generated using our attack method also have good transferability between different models, we use the adversarial patches generated by training on the YOLOv5 model to attack other models. The experimental data can be obtained from Table 2. In the process of attacking the YOLO series, the worst performance is the attack on the YOLOv3 object detector. The AP after the attack is 15.31%, which is 59.25% lower than that of the clean image (74.56%). After attacking the Faster R-CNN object detector, the AP is still 25.61%, which is a decrease of 47.4% compared to the clean image (73.01%). We use the adversarial patches on different datasets we just obtained to attack Faster R-CNN, and the obtained data are shown in Table 4. It can be seen that the adversarial patch generated by our method also has a good attack effect on the two-stage target detector, and the detection AP after the attack is all below 50%, which further confirms that our method has strong transferability. This finding suggests that the more similar the model structures of object detectors are, the stronger the adversarial patches generated based on these models are in terms of attack transferability among them. Overall experimental results show that adversarial patches can successfully attack these object detectors even if the corresponding models do not train them. Meanwhile, we can improve the transferability of adversarial examples by the researches like [42–44].

4.3 Attack in the Physical World

Next, we will apply the adversarial patch generated by our method to attack the physical world and evaluate the effect of the attack. We choose InviSpell with the best effect in digital attack as the adversarial patch in this experiment, use the camera of HUAWEI Mate 30 mobile phone to obtain photos of the physical

Table 4. Average Precision (AP%) obtained by attacking Faster R-CNN by using InviSpell obtained by training with different data sets

Datasets	Clean AP	Patch AP	Decrease (↑)
INRIA	73.01	25.61	47.40
COCO	70.25	29.76	40.49
VOC	76.25	26.25	50.00
Mix	68.26	30.25	38.01

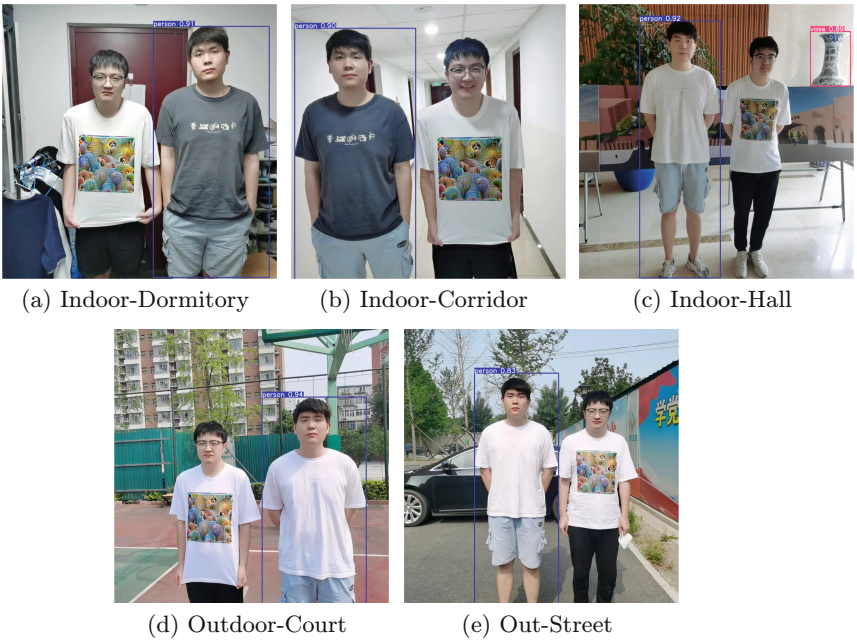
world, and use the object detector with YOLOv5 model to evaluate the detection. We define the target with higher than 50% confidence as the attack failed, and the target lower than 50% as the attack success, and use the recall rate as the evaluation index. First, we use a common printer to print the adversarial patch in a poster with a size of 40 cm × 40 cm. Then we set up the experimenter who held the poster to cover part of the body and the experimenter who was not covered. The two participants stood side by side, and the attacking person's shooting angles were front, side and back. As shown in Fig. 5. From the data statistics obtained from the experiment, it can be known that our adversarial poster can reduce the detection recall rate from 100% to 35.94%, proving that an adversarial patch in the form of the poster can achieve good attack results in the physical world.

(a) Front (b) Side (c) Back

Fig. 5. Examples of adversarial posters attacking persons in different camera angles.

To improve the practicality of the adversarial patch in the physical world, we also customize the adversarial patch as a pattern on the T-shirt. The rest of the experimental conditions are the same as the previous poster experiment. In addition, we chose to acquire photographs of experimenters wearing an adversarial T-shirt in multiple indoor and outdoor scenes. Indoor environments include dormitories, corridors, and halls; outdoor environments include courts and streets. Outdoors can be brighter and more complex compared to indoor environments. The purpose of using different scenes is to be closer to the normal life mode

of people, as shown in Fig. 6. From Table 5, we can get that the adversarial T-shirt can reduce the detection recall rate of indoor targets to 26.54%, and the detection recall rate of outdoor targets to 40.98%.



(a) Indoor-Dormitory (b) Indoor-Corridor (c) Indoor-Hall

(d) Outdoor-Court (e) Out-Street

Fig. 6. Effects of adversarial T-shirt attack indoors and outdoors.

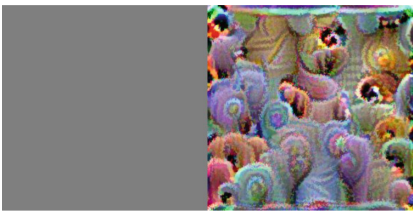
5 Discussions

5.1 Effects of Different Initial Patches

In order to verify whether training from different initial patches will affect the attack performance of the final patch, we set 4 different initial patches as the starting point for training, namely pure gray, pure white, pure black, and TV image. As shown in Fig. 7a d, the left side is the initial patch image before training, and the right side is the effect after training reaches the convergence limit. Table 6 presents the attack effects of these four adversarial patches on the INRIA dataset. Pure gray, pure white, and pure black can reduce AP to less than 5% after training, but the pure white patch is not as good as the previous ones. From this, we can conclude that most initial patches can achieve similar attack effects after training, but there are exceptions. We speculate that white has less influence on key pixels and cannot mislead the object detector to a large extent.

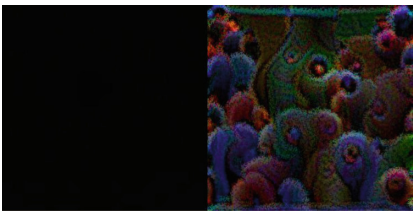
Table 5. Attack effect of adversarial T-shirts on YOLOv5 object detectors in different environments(using average precision AP% as the metrics).

Environments	Indoor			Outdoor	
	Dormitory	Corridor	Hall	Court	Street
T-shirt	100%	100%	100%	100%	100%
Adversarial T-shirt	26.54%	29.17%	34.57%	40.98%	41.66%



(a) Pure Gray

(b) Pure White



(c) Pure Black

(d) TV

Fig. 7. Examples of different initial patches and their convergence limit.

In addition, we found that in the training process, the training epochs required for different initial patches to reach the convergence limit are also quite different in the training process. When the initial patch is TV, it only takes about 400 epochs to achieve the best attack effect state, while black and gray require 1100 epochs. Since TV is a recognizable class in the INRIA dataset, we tried to train with other class patterns as the initial patch. We got similar results, such as cattle (500 epochs), cars (450 epochs), etc. These results show that different initial patches can greatly affect the training speed of reaching the best convergence value, but cannot improve the attack performance against patches from the results.

5.2 Effects of Different Transformations

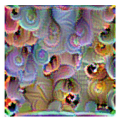
We explored various transformation methods for the patch, and tested whether it still has an attacking effect on the target after transformation, which reflects the robustness. For this, We made the following 5 transformations for the adversarial patch: (1) recombination after quartering, (2) recombining after nine equals, (3) left-right mirroring, (4) up-down mirroring, and (5) 180° rotation. Table 7 shows

Table 6. Attack effects on YOLOv5 after different initial patch training.

Initial patch	Pure Gray	Pure White	Pure Black	TV
Average Precision	2.86	20.77	4.49	3.19

what the adversarial patch looks like and its attack effect on the INRIA dataset after using different variations.

Table 7. Attack effects of adversarial patches on YOLOv5 with different transformations.

	$\frac{1}{4}$	$\frac{1}{9}$	left-right	up-down	180°
Patch					
Average Precision	7.47	10.31	4.91	15.34	17.91

We observed that using various transformations will reduce the attack's effectiveness. The worst case is that the AP of the attack after 180° rotation transformation reduces from 2.86% to 17.91%. The experimental results show that after undergoing various transformations, the attack effect does not greatly decrease or disappear completely, which proves that the adversarial patches generated by this method have good robustness.

5.3 Effects of Different Patch Sizes

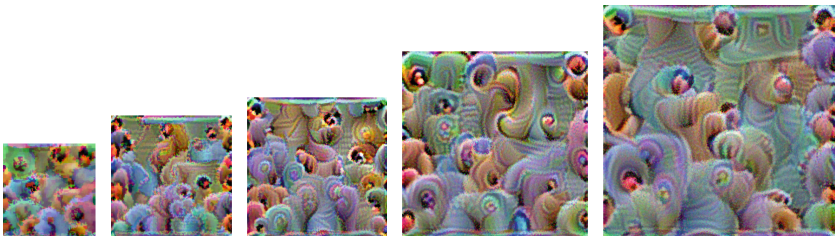
The size of the adversarial patch will affect the attack performance. For a fixed-size adversarial patch, the larger the size, the better the attack performance. That is because the larger the target area it covers, the more interference to the target features. In this work, since the adversarial patch is scaled according to the detected target size, it cannot be determined how large the initial patch should be set to have the best attack performance. In this part of the experiment, we explore the impact of adversarial patches of different sizes on the attack performance of our method. We set up five initial patches of different sizes to train and attack the INRIA dataset. The experimental results are shown in Table 8. The experimental results show that the smaller the adversarial patch, the worse the attack performance; the larger the adversarial patch, the stronger the attack performance. The overall attack effect is the best when a 300×300 adversarial patch is used. When the size of the adversarial patch is less than 300

× 300, the attack performance increases gradually with the increase of the patch; when the size of the adversarial patch exceeds 300 × 300, the attack performance gradually decreases with the increase of the patch size. This means that when using our method to generate adversarial patches, up to a certain threshold, the larger the patch, the better the attack effect.

Table 8. Attack effect (AP% and Recall%) of adversarial patches of different sizes on YOLOv5.

Size	AP	Recall
100 × 100	10.84	15.69
200 × 200	4.24	8.45
300 × 300	2.86	4.52
400 × 400	2.94	4.89
500 × 500	3.49	6.54

We show five adversarial patches of different sizes in Fig. 8. As the size of the adversarial patch gradually increases, there are more and more pixel details in it, and the visually perceived content also gradually increases, and finally achieves a more saturated effect when the size is above 300 × 300. Small adversarial patches are relatively single in the continuous change of pixels, which may explain the poor attack performance. Larger adversarial patches have larger pixel color differences, which can interfere with the characteristics of the attacked object more effectively, so the attack performance is better.



(a) 100 (b) 200 (c) 300 (d) 400 (e) 500

Fig. 8. Different sizes of adversarial patches.

6 Conclusions

In this paper, we propose a novel approach to attack advanced object detectors. This method generates a new adversarial patch against the network structure of current advanced object detectors, which can successfully attack the new

advanced object detectors while maintaining the deception of the old ones. We also propose a new optimization strategy, which can adjust the optimization weights according to the confidence of each batch of images, so as to improve the attack success rate against patches. Experiments show that the adversarial patch generated by this method can achieve better attack results in both the digital and physical worlds. For the YOLOv5 object detector, the AP can be reduced to 2.86%. In contrast, the ADV-OBJ method can only reduce to 61.56%. As shown in Fig. 3 and Table 7, adversarial patches exhibit a striking contrast with the real world, making them less suitable for conducting covert attacks on object detection systems. In our future work, we aim to explore and design adversarial patches that coexist with the attack's effects and realism. Specifically, our goal is for adversarial patches to incorporate a degree of semantic information, rendering them less conspicuous in everyday settings.

Acknowledgement. This work is supported by the Fundamental Research Funds for the Central Universities (328202204). And for the reproducibility of the proposed method, we have published our source code online at https://github.com/BESTICSP/InviSpell.

Appendix

See Table 9 for explanations of abbreviations.

Table 9. Explanations of abbreviations.

YOLO	You only look once
R-CNN	Region-based Convolutional Neural Network
INRIA	the Inria Aerial Image Labeling Dataset
FGSM	Fast Gradient Symbolic Method
PGD	Projected Gradient Descent
RSIs	remote sensing images
AP	Attack performance

References

1. Jocher, G.: YOLOv5 (2020). https://github.com/ultralytics/yolov5/
2. Morgulis, N., Kreines, A., Mendelowitz, S., Weisglass, Y.: Fooling a real car with adversarial traffic signs. arXiv preprint arXiv:1907.00374 (2019)
3. Chen, S.-T., Cornelius, C., Martin, J., Chau, D.H.P.: ShapeShifter: robust physical adversarial attack on faster R-CNN object detector. In: Berlingerio, M., Bonchi, F., Gärtner, T., Hurley, N., Ifrim, G. (eds.) ECML PKDD 2018. LNCS (LNAI), vol. 11051, pp. 52–68. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-10925-7_4

4. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
5. Liu, A., Liu, X., Fan, J., Ma, Y., Zhang, A., Xie, H., Tao, D.: Perceptual-sensitive GAN for generating adversarial patches. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 1028–1035 (2019)
6. Husnoo, M.A., Anwar, A.: Do not get fooled: defense against the one-pixel attack to protect IoT-enabled deep learning systems. Ad Hoc Netw. **122**, 102627 (2021)
7. Liu, A., Wang, J., Liu, X., Cao, B., Zhang, C., Yu, H.: Bias-based universal adversarial patch attack for automatic check-out. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12358, pp. 395–410. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58601-0_24
8. Wu, Z., Lim, S.-N., Davis, L.S., Goldstein, T.: Making an invisibility cloak: real world adversarial attacks on object detectors. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12349, pp. 1–17. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58548-8_1
9. Huang, L., et al.: Universal physical camouflage attacks on object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 720–729 (2020)
10. Dong, Y., et al.: Boosting adversarial attacks with momentum. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 9185–9193 (2018)
11. Kang, H., Kim, H., et al.: Robust adversarial attack against explainable deep classification models based on adversarial images with different patch sizes and perturbation ratios. IEEE Access **9**, 133 049–133 061 (2021)
12. Brown, T.B., Mané, D., Roy, A., Abadi, M., Gilmer, J.: Adversarial patch. arXiv preprint arXiv:1712.09665 (2017)
13. Li, J., Schmidt, F., Kolter, Z.: Adversarial camera stickers: a physical camera-based attack on deep learning systems. In: International Conference on Machine Learning, pp. 3896–3904. PMLR (2019)
14. Komkov, S., Petiushko, A.: AdvHat: real-world adversarial attack on ArcFace face id system. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 819–826. IEEE (2021)
15. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7263–7271 (2017)
16. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
17. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2017)
18. Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y.M.: YOLOv4: optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
19. Biggio, B., et al.: Evasion attacks against machine learning at test time. In: Blockeel, H., Kersting, K., Nijssen, S., Železný, F. (eds.) ECML PKDD 2013. LNCS (LNAI), vol. 8190, pp. 387–402. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40994-3_25
20. Szegedy, C., et al.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
21. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)

22. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
23. Song, D., et al.: Physical adversarial examples for object detectors. In: 12th USENIX workshop on offensive technologies (WOOT 2018) (2018)
24. Liu, X., Yang, H., Liu, Z., Song, L., Li, H., Chen, Y.: DPatch: an adversarial patch attack on object detectors. arXiv preprint arXiv:1806.02299 (2018)
25. Lee, M., Kolter, Z.: On physical adversarial patches for object detection. arXiv preprint arXiv:1906.11897 (2019)
26. Lu, M., Li, Q., Chen, L., Li, H.: Scale-adaptive adversarial patch attack for remote sensing image aircraft detection. Remote Sens. **13**(20), 4078 (2021)
27. Redmon, J., Farhadi, A.: YOLOv3: an incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
28. Kurakin, A., Goodfellow, I.J., Bengio, S.: Adversarial examples in the physical world. In: Artificial Intelligence Safety and Security, pp. 99–112. Chapman and Hall/CRC (2018)
29. Evtimov, I., et al.: Robust physical-world attacks on machine learning models. arXiv preprint arXiv:1707.08945, vol. 2, no. 3, p. 4 (2017)
30. Thys, S., Van Ranst, W., Goedemé, T.: Fooling automated surveillance cameras: adversarial patches to attack person detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2019)
31. Xu, K., et al.: Adversarial T-shirt! Evading person detectors in a physical world. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12350, pp. 665–681. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58558-7_39
32. Pautov, M., Melnikov, G., Kaziakhmedov, E., Kireev, K., Petiushko, A.: On adversarial patches: real-world attack on ArcFace-100 face recognition system. In: 2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON), pp. 0391–0396. IEEE (2019)
33. Sharif, M., Bhagavatula, S., Bauer, L., Reiter, M.K.: 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, pp. 1528–1540 (2016)
34. Sharif, M., Bhagavatula, S., Bauer, L., Reiter, M.: A general framework for adversarial examples with objectives. ACM Trans. Priv. Secur. (TOPS) **22**(3), 1–30 (2019)
35. Yang, K., Tsai, T., Yu, H., Ho, T.-Y., Jin, Y.: Beyond digital domain: fooling deep learning based recognition system in physical world. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 01, pp. 1088–1095 (2020)
36. Kong, Z., Guo, J., Li, A., Liu, C.: PhysGAN: generating physical-world-resilient adversarial examples for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14 254–14 263 (2020)
37. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), vol. 1, pp. 886–893. IEEE (2005)
38. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
39. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (VOC) challenge. Int. J. Comput. Vision **88**(2), 303–338 (2010)

40. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
41. Xue, M., He, C., Wu, Z., Wang, J., Liu, Z., Liu, W.: 3D invisible cloak. arXiv preprint arXiv:2011.13705 (2020)
42. Xie, C., et al.: Improving transferability of adversarial examples with input diversity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2730–2739 (2019)
43. Wu, D., Wang, Y., Xia, S.-T., Bailey, J., Ma, X.: Skip connections matter: on the transferability of adversarial examples generated with ResNets. arXiv preprint arXiv:2002.05990 (2020)
44. Dong, Y., Pang, T., Su, H., Zhu, J.: Evading defenses to transferable adversarial examples by translation-invariant attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4312–4321 (2019)