

A Single Set of Adversarial Clothes Breaks Multiple Defense Methods in the Physical World

Wei Zhang^a, Zhanhao Hu^b, Xiao Li^a, Xiaopei Zhu^a, Xiaolin Hu^{a,c,*}

^aDepartment of Computer Science and Technology, Institute for Artificial Intelligence, THBI, BNRist, Tsinghua University, Beijing 100084, China

^bBerkeley Institute for Data Science, Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720, USA

^cChinese Institute for Brain Research (CIBR), Beijing 100010, China

ABSTRACT

In recent years, adversarial attacks against deep learning-based object detectors in the physical world have attracted much attention. To defend against these attacks, researchers have proposed various defense methods against adversarial patches, a typical form of physically-realizable attack. However, our experiments showed that simply enlarging the patch size could make these defense methods fail. Motivated by this, we evaluated various defense methods against adversarial clothes which have large coverage over the human body. Adversarial clothes provide a good test case for adversarial defense against patch-based attacks because they not only have large sizes but also look more natural than a large patch on humans. Experiments show that all the defense methods had poor performance against adversarial clothes in both the digital world and the physical world. In addition, we crafted a single set of clothes that broke multiple defense methods on Faster R-CNN. The set achieved an Attack Success Rate (ASR) of 96.06 % against the undefended detector and over 64.84 % ASRs against nine defended models in the physical world, unveiling the common vulnerability of existing adversarial defense methods against adversarial clothes. Code is available at: <https://github.com/weiz0823/adv-clothes-break-multiple-defenses>.

© 2025 Elsevier Ltd. All rights reserved.

Corresponding author:

e-mail: xlu@tsinghua.edu.cn (Xiaolin Hu)

1. Introduction

Deep Neural Networks (DNNs) are known to be vulnerable to adversarial examples not only in the digital world (Goodfellow et al., 2014; Madry et al., 2018; Karmon et al., 2018), but also in the physical world (Brown et al., 2017; Thys et al., 2019; Xu et al., 2020; Wu et al., 2020b; Hu et al., 2021, 2022, 2023). Physical adversarial examples raise serious security concerns since they can be deployed in the real world. Given the widespread deployment of object detection models in various

applications, researchers have focused on fooling object detection models in the physical world in recent years, especially person detection models (Thys et al., 2019; Xu et al., 2020; Wu et al., 2020b; Hu et al., 2021, 2022, 2023).

To defend against physically realizable attacks, various defense methods (Naseer et al., 2019; Zhou et al., 2020; Yu et al., 2022; Mu and Wagner, 2021; Ji et al., 2021; Yu et al., 2021; Kim et al., 2022; Liu et al., 2022; Rossolini et al., 2023; Xu et al., 2023; Tarchoun et al., 2023; Rao et al., 2020; Metzen et al., 2021) have been proposed. They are usually designed

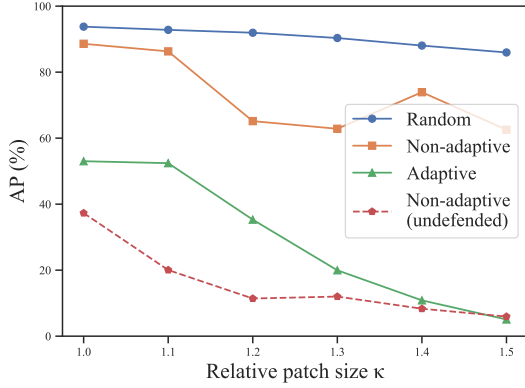


Fig. 1. APs of *person* class of the FNC-defended Faster R-CNN (Ren et al., 2015) detector against patch attack of different patch sizes on the Inria person dataset (Dalal and Triggs, 2005). Solid lines denote the APs of the FNC-defended detector, when the input images are applied with *random patch*, *non-adaptive patch* optimized on the undefended detector, and *adaptive patch* optimized on the FNC-defended detector. Dashed line denotes the AP of the undefended detector, when the input images are applied with *non-adaptive patch*. Details of experiment setups are provided in Section 3.

for and evaluated on adversarial patches (Brown et al., 2017; Thys et al., 2019). Some of the defense methods (Yu et al., 2022; Kim et al., 2022; Zhou et al., 2020) were evaluated on physical-world patch-based attacks, and others were evaluated on digital-world patch-based attacks. They worked well in defending against patch-based attacks. However, in this paper we will show that this may have given a false sense of security of protected object detectors.

Our key observation is that existing defense methods against patch-based attacks have not given enough attention to the patch size. Intuitively, larger patch size corresponds to larger optimization space, and the attack becomes stronger and harder to defend against. Motivated by this, we first investigated whether patch size influences the performance of the defense. Fig. 1 provides a clue, where we tested FNC (Yu et al., 2021), a patch defense method that was independent of the patch size. Note that the defense performance of FNC against adaptive attack (the green line) vanished as the patch became larger. However,

simply enlarging patch is not a natural way for physical implementation, since holding a board much larger than body size is not very much natural in real-world scenarios. Instead, texture-based attacks (Hu et al., 2022, 2023), also known as adversarial clothes, provide intuitively similar implementation as the enlarged patch size. It is therefore necessary to re-evaluate the state-of-the-art (SOTA) defense methods to demonstrate their capability of defending against adversarial clothes.

In this study, we evaluated various SOTA adversarial defense methods against adversarial clothes. To craft the adversarial clothes in the physical world, we need to design the texture by gradient-based optimization in the digital world first. We adopted a 3D rendering pipeline (Hu et al., 2023) to optimize adversarial textures on clothes. We tested diverse defense methods (Li et al., 2023a; Yu et al., 2021; Zhou et al., 2020; Liu et al., 2022; Kim et al., 2022; Naseer et al., 2019; Yu et al., 2022; Tarchoun et al., 2023) and found that all of them had poor performance against the digital-world adversarial texture. An implementation of the adversarial clothes in the physical world by printing the texture on real-world clothes unveiled that the vulnerability of defense methods could also be exploited in real-world scenarios.

Based on these findings, we conjectured that these defense methods share common vulnerabilities that could be exploited by a single set of adversarial clothes. We optimized the texture of a set of clothes and included an ensemble of defended models during the optimization. We printed the clothing texture on a piece of cloth and tailored it into a set of adversarial clothes including a shirt and a pair of trousers. Experiments showed

that the set of clothes bypassed nine defense methods in both the digital world and the physical world.

The main contributions are: (1) We evaluated SOTA adversarial defense methods against adversarial clothes, and found that adversarial clothes impaired the performance of the existing defense methods in real-world scenarios. (2) We successfully broke nine defense methods in the physical world with a single set of clothes, achieving over 64.84 % ASRs against all nine defense methods. (3) The results unveil that SOTA defense methods are still vulnerable to physical-world adversarial examples when confronted with texture-based attacks.

The rest of this paper is organized as follows. Section 2 briefly reviews the physically realizable threats to person detectors. Section 3 shows the impact of patch size on adversarial defense, which provides the motivation to evaluate the defense methods against adversarial clothes. Section 4 describes the attack settings of adversarial clothes and evaluation metrics used in this study. Section 5 shows the evaluation results of various adversarial defense methods against adversarial clothes with both the digital-world results and the physical-world results. Finally, the conclusion is given in Section 6.

2. Related work

In this section, we introduce the threats including patch-based attacks and texture-based attacks to fool person detectors. Then, we briefly review the defense methods against physically realizable attacks.

2.1. Threats

Adversarial Patches. The first work for generating physically realizable adversarial patches to fool person detectors was pro-

posed by Thys et al. (2019), and the similar pipeline has been followed up by several works (Xu et al., 2020; Wu et al., 2020b; Hu et al., 2021). We denote their method as *AdvPatch*. The main process is to transform and apply an image patch onto each person in an image from the training dataset according to the person’s bounding box, and then optimize all the patches by minimizing the detection scores outputted by the target detector. The loss is defined as the maximum detection score among all bounding boxes in each image, which we minimize.

To make the patches smoother, the AdvPatch method adds a total variation (TV) loss. TV loss is lower when neighboring pixel values are closer, and the patch has smoother appearance. There is also a non-printability score (NPS) loss term to restrict pixels in the patches within a set of printable colors. In addition, Expectation over Transformations (EoT) (Athalye et al., 2018) is applied to the patches to make them more robust to physical transformations, including randomizing locations, rotations, brightness, contrast, and pixel noises.

Hu et al. (2022) extended the adversarial patches to tileable patches in order to make the attack effective in multiple viewing angles. The physical-world implementation is made by printing tileable patch repeatedly as texture on a piece of cloth, then tailoring the cloth into clothes covered with adversarial patterns. We denote the method as *AdvTexture*. Despite the notation, the optimization pipeline of the attack is still based on the adversarial patches, while employing a toroidal cropping technique to randomly crop a unit of the patch, and expandable generation technique to generate the tileable patch with generation model.

Texture-based attacks. Hu et al. (2023) proposed to use a 3D rendering pipeline to obtain adversarial camouflage texture (AdvCaT) for clothes. The AdvCaT resembles typical camouflage patterns, making the clothes natural-looking. During optimization, a 3D person mesh model is rendered from different viewing angles and the rendered images are synthesized with background images. In addition to EoT used for adversarial patches, Thin Plate Spline (TPS) (Bookstein, 1989; Donato and Belongie, 2002) deformation is also incorporated to enhance the robustness of the attack in the physical world.

2.2. Defenses

The defense methods against physically realizable attacks can be roughly divided into four categories, we briefly review them as follows.

Model-independent input preprocess. This kind of defense methods (Naseer et al., 2019; Zhou et al., 2020; Tarchoun et al., 2023; Liu et al., 2022; Xu et al., 2023; Jing et al., 2024) either mask out or suppress the regions on the input images that are suspected to contain adversarial patches, before the input images are fed into the detector. Specifically, Local Gradient Smoothing (LGS) (Naseer et al., 2019) computes the gradients of the pixels in an image with respect to pixel position, then suppresses large-gradient regions by a factor proportional to that gradient computed. Entropy-based methods (Zhou et al., 2020; Tarchoun et al., 2023) compute the entropy (Gray, 2011) of the pixels within a sliding window across the input image. High-entropy regions are suspected as adversarial regions. Information Distribution Based Defense (IDBD) (Zhou et al., 2020) incorporates entropy-based proposal with gradient-based filtering

which computes the sum of the pixel gradients within the sliding window. Jedi (Tarchoun et al., 2023) localizes adversarial patches with entropy heatmap, then completes the patch mask with an autoencoder, and finally inpaints the detected adversarial patch region. PAD (Jing et al., 2024) localizes adversarial patches with mutual information score and compression difference. However, because the computational cost of mutual information is squared to the cost of computing entropy on a sliding window, the processing speed of PAD is very slow. Segment and Complete (SAC) (Liu et al., 2022) trains a patch segmentation model that outputs a raw mask indicating the regions of the patches. A shape completion algorithm is then applied to the raw mask, generating a completed patch mask. Finally, the patch region is removed based on the completed patch mask. Similarly, NAPGuard (Wu et al., 2024) trains a patch detection model that localizes the bounding box of patch.

Outlier feature filter. Defense methods in this category (Yu et al., 2021; Kim et al., 2022; Rossolini et al., 2023; Mu and Wagner, 2021) extract inner features from the target DNN model. They usually filter or clip the feature vectors according to their distributions. Feature Norm Clip (FNC) (Yu et al., 2021) is motivated by the observation that the l_2 norms of the convolutional neural network (CNN) feature vectors at the regions containing adversarial patches are usually larger than those of the benign regions. All feature maps of the CNN models are filtered with a clip operation, making an upper bound on the norm of feature vectors. Adversarial Patch-feature Energy (APE) (Kim et al., 2022) combines adversarial region detection with feature filtering. Adversarial regions are detected based on

multi-level outlier features. Then the first-layer outlier features are clipped within the detected adversarial regions.

Adversarial training. Adversarial Training (AT) (Madry et al., 2018) and its variants (Zhang et al., 2019; Wu et al., 2020a; Li et al., 2023b) are usually recognized as the most effective methods in defending classification models against noise-bounded adversarial attacks (Croce et al., 2020). Recently, AT has been applied to object detectors (Zhang and Wang, 2019; Chen et al., 2021; Dong et al., 2022; Li et al., 2023a), under the setting that the adversarial noise is bounded by l_p norms. However, as far as we know, no AT method has been proposed specifically for object detectors against physical attacks. In experiments, we utilized the checkpoints of the AT models from Li et al. (2023a), which is one of the SOTA AT methods for object detection against l_∞ norm bounded attack. The AT models were trained with the l_∞ norm bound $\epsilon = 4/255$.

Defensive frame. Yu et al. (2022) proposed to train a Universal Defensive Frame (UDF) on adversarial examples. The method involves training the UDF in conjunction with the attacking patch, following a pipeline similar to AT. Mao et al. (2024) proposed a similar method that optimized the defense filter in conjunction with the attacking patch. The defense filter was defined as a noise image that was linearly interpolated with the input image to robustify the detection.

3. Impact of patch size on adversarial defense

Previous adversarial patch attacks (Thys et al., 2019; Hu et al., 2021, 2022) mainly focused on patches with a fixed size and scaled them proportional to the size of target bounding

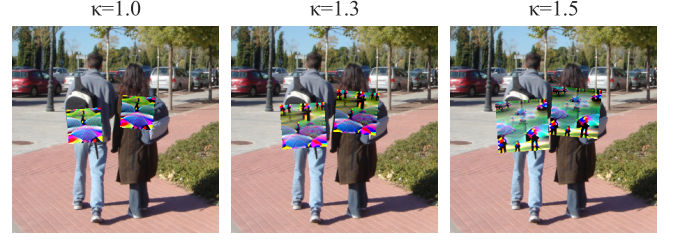


Fig. 2. Visualization of the patches of different sizes when applied onto persons in the Inria (Dalal and Triggs, 2005) dataset. The images have been cropped and zoomed in for a better view.

box. This is the setting that previous defense methods defended against. Intuitively, larger patches on the target should have better adversarial effect. To the best of our knowledge, no prior studies have quantitatively assessed the extent of adversarial effects that larger patches can achieve, particularly in the context of defended models. To show that the patch size does have impact on the performance of defense methods, we took FNC (Yu et al., 2021), a typical size-independent defense method, as an example, and used AdvPatch (Thys et al., 2019) as the attack method. Faster R-CNN (Ren et al., 2015) was chosen as the target detector. We mainly studied the detection performance of the detector defended by the FNC method against the AdvPatch attack with both non-adaptive patches and FNC-adaptive patches.

For the patch-based attack, the size of the patch applied onto a person was determined by the diagonal length of the person's ground truth bounding box. Suppose we have a bounding box of the target object with diagonal length d , the edge length l of the patch can be calculated by $l = cd$, where c is a constant controlling patch size. We used $c_0 = 0.2$ as the constant c for the baseline patch, consistent with the previous study (Thys et al., 2019). To scale up the patch, we adjusted the value of c ,

and denote $\kappa = c/c_0$ as *relative patch size*. We tested the defense performance of FNC under both non-adaptive and adaptive attacks with $\kappa \in [1.0, 1.5]$. An visualization of patches in different sizes in the digital world is provided in Fig. 2. We used the same patch resolution, 300×300 pixels, for different patch sizes, in order to keep the dimension of the attack solution space the same. As shown in Fig. 1, the FNC-defended Faster R-CNN detector had APs larger than 53.01 % against all types of patches with sizes of $\kappa = 1.0$. In comparison, the undefended Faster R-CNN detector had an AP of only 37.30 %. As κ increased, the APs of the FNC-defended detector against non-adaptive patches remained high (over 60 %), while the APs of the detector against adaptive patches dropped quickly. When κ reached 1.5, there was no significant difference between the AP of undefended detector and that of FNC-defended detector against the strongest patch (*i.e.* the adaptive patch). The results showed that the defense performance of FNC vanished as the patch became larger.

4. Attack settings of adversarial clothes

As shown in Section 3, expanding patch size does have an impact on defense performance. $\kappa = 1.5$ made the defense performance of FNC vanish. However, simply enlarging patch is not a natural way for physical implementation, since holding a board much larger than body size is not natural in real-world scenarios. Instead, texture-based attacks (Hu et al., 2022, 2023), also known as adversarial clothes, provide intuitively similar implementation as the enlarged patch size. It is therefore necessary to re-evaluate the SOTA defense methods to demonstrate their capability of defending against adversarial

clothes.

To systematically evaluate the defense models, we targeted nine typical defense methods, including AT (Li et al., 2023a), FNC (Yu et al., 2021), LGS (Naseer et al., 2019), IDBD (Zhou et al., 2020), SAC (Liu et al., 2022), APE (Kim et al., 2022), UDF (Yu et al., 2022), Jedi (Tarchoun et al., 2023), and NAP-Guard (Wu et al., 2024). These defense methods cover all four categories of adversarial patch defenses as detailed in Section 2.2. The target detectors were Faster R-CNN (Ren et al., 2015) and FCOS (Tian et al., 2019), which represent typical two-stage and single-stage detectors, respectively. Both detectors were pretrained on MS-COCO (Lin et al., 2014) dataset. Input images were cropped or padded to equal width and height and resized to 416×416 , then normalized before being fed into the detector.

4.1. Optimization in the digital world

We utilized the 3D rendering pipeline proposed by Hu et al. (2023). Since the naturalness of clothing textures was not a concern in this study, we excluded the module for camouflage patterns generation in AdvCaT (Hu et al., 2023), and optimized the texture map of the 3D mesh model pixel-wise. This form of attack is denoted as *Texture3D*. To improve the robustness of physical implementation of the clothing textures, we incorporated TV loss computed on the texture map during texture optimization.

We used the same background image dataset as used by AdvCaT (Hu et al., 2023), consisting of 506 background images varying in the scene. The background images were split into 376 images for training and 130 images for testing. Textures

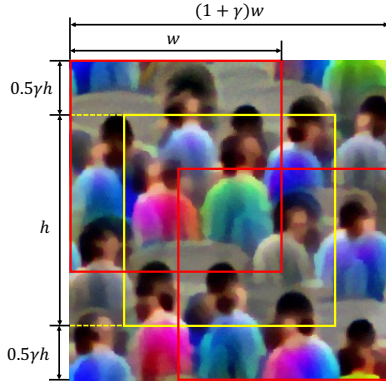


Fig. 3. Illustration of position jittering crops to enhance the physical world robustness against texture-based attacks. The whole texture represents the latent feature map. The red boxes are possible texture crops used for optimization, and the yellow box is the center texture crop used for evaluation and physical implementation.

were optimized for 100 epochs with Adam (Kingma and Ba, 2015) optimizer. The initial learning rate for Texture3D was set to 0.01.

4.2. Enhancing physical world robustness of the attack

Besides incorporating the EoTs to migrate the gap between digital-space optimization and physical-world implementation as described in the original 3D rendering attack pipeline (Hu et al., 2023), we introduce *position jittering* to the texture map to further simulate physical-world transformations, as shown in Fig. 3. Suppose the size of the texture map is (w, h) , and the jittering intensity is $0 < \gamma < 1$. Instead of optimizing the original texture map, we pad the texture map and optimize on the padded *latent texture map* with size $(\lfloor (1 + \gamma)w \rfloor, \lfloor (1 + \gamma)h \rfloor)$, where $\lfloor \cdot \rfloor$ denotes the rounding down operation. For each optimization iteration, we crop a new texture map with fixed size (w, h) and random top-left corner coordinate ranging from $(0, 0)$ to $(\lfloor \gamma w \rfloor, \lfloor \gamma h \rfloor)$ from the latent texture map for rendering (see Fig. 3, red boxes). During evaluation and for physical implementation, the texture map with size (w, h) and top-left coordi-

nate $(\lfloor 0.5\gamma w \rfloor, \lfloor 0.5\gamma h \rfloor)$ is used (see Fig. 3, yellow box).

4.3. Physical world implementation

Following Hu et al. (2022, 2023), we printed the texture map optimized by the texture-based attack on pieces of cloth and tailored them into shirts and trousers to produce sets of adversarial clothes.

4.4. Evaluation

Digital world evaluation. For the evaluation metric in the digital world, we used Average Precision (AP) on the *person* class. The Intersection over Union (IoU) threshold was set to 0.5 for both patch-based and texture-based attacks, consistent with previous works (Thys et al., 2019; Hu et al., 2021, 2022, 2023).

Physical world evaluation. To evaluate the effectiveness of the clothes in the physical world, we used the metric of Attack Success Rates (ASRs) on a set of images. ASR calculates the percentage of successfully attacked images in all test images. The attack was considered as successful if no box of *person* class had an IoU over 0.5. In addition, only boxes with confidence scores larger than 0.5 were taken into account.

Two actors (age mean: 25; age range: 22 to 28; height range: 178 cm to 188 cm) were recruited to collect physical test data. The physical test results were all averages of the results on the two subjects. The recruitment and study procedures were approved by the Department of Psychology Ethics Committee, Tsinghua University, Beijing, China.

To evaluate the average attack performance from multiple viewing angles, we recorded a video of person turning circles, and evenly extracted frames from the video to form our test

Table 1. APs (% , $\uparrow$) of Faster R-CNN equipped with different defenses (including the undefended detector) against texture-based attacks. AT and FNC defense methods were evaluated using adaptive attack (marked with $\dagger$).

Model	Random	Adversarial
Undefended	88.52	0.03
AT (Li et al., 2023a)	93.72	5.97 $\dagger$
FNC (Yu et al., 2021)	89.05	0.91 $\dagger$
LGS (Naseer et al., 2019)	89.21	0.87
IDBD (Zhou et al., 2020)	88.31	1.72
SAC (Liu et al., 2022)	48.46	0.43
APE (Kim et al., 2022)	88.31	0.05
UDF (Yu et al., 2022)	73.25	1.51
Jedi (Tarchoun et al., 2023)	88.65	5.32
NAPGuard (Wu et al., 2024)	87.35	0.08

Table 2. APs (% , $\uparrow$) of FCOS (Tian et al., 2019) equipped with different defenses (including the undefended detector) against texture-based attacks. AT and FNC defense methods are evaluated using adaptive attack (marked with $\dagger$).

Model	Random	Adversarial
Undefended	63.67	0.05
AT (Li et al., 2023a)	91.56	5.24 $\dagger$
FNC (Yu et al., 2021)	74.40	0.15 $\dagger$
LGS (Naseer et al., 2019)	61.90	0.10
IDBD (Zhou et al., 2020)	64.11	0.05
SAC (Liu et al., 2022)	29.60	0.04
APE (Kim et al., 2022)	63.01	0.03
UDF (Yu et al., 2022)	76.15	6.74
Jedi (Tarchoun et al., 2023)	61.38	0.06
NAPGuard (Wu et al., 2024)	62.46	0.02

set. The ground truth bounding boxes were manually annotated. These evaluation metrics are consistent with previous works (Hu et al., 2022, 2023).

5. Results

We evaluated the performance of detectors equipped with various defense methods against the texture-based attack Texture3D.

5.1. Digital-world evaluation

We tested the performance of various adversarial patch defenses against texture-based attacks, and the results are shown in Table 1. While random texture did not lower AP much, the adversarial texture broke all defense methods including the undefended model with APs all lower than 5.32 %, except AT

and FNC. Since AT and FNC defense methods are hard to attack ($AP > 7\%$) with the non-adaptive adversarial pattern optimized on the undefended Faster R-CNN model, we utilized the straightforward adaptive attack, *i.e.*, optimizing the adversarial pattern on the detection model equipped with the defense method. Even with adaptive attack, the strongest defense method is AT, with an AP of 5.97 %.

The results of FCOS (Tian et al., 2019) were similar as those of Faster R-CNN, as shown in Table 2. The texture-based attack broke all defenses with APs all lower than 6.74 %. The strongest defense method is UDF with the highest AP. The conclusion is similar to that of Faster R-CNN detector.

We then evaluated the performance of adversarial defense methods against transfer attacks with adversarial textures optimized on the undefended Faster R-CNN detector, the detectors with AT or FNC defense, and an ensemble of the above three models, and the results are shown in Table 3. The results show that the adversarially trained detector achieved an AP of 62.65 % on the texture optimized on the undefended model, and an AP of 69.12 % on the texture optimized on FNC model, indicating that AT succeeded in defending against the texture optimized on the undefended detector and the texture optimized on the FNC-defended detector. Adaptively attacking AT model got an AP of 5.97 %, but on the other hand diminished the adversarial effect against FNC defense, increasing AP to 22.30 %.

Targeting an ensemble of defenses. The 2nd to 4th columns of Table 3 show that the adversarial textures crafted against the undefended Faster R-CNN, the AT model and the FNC-defended model were all defended by some defense method, with the APs

Table 3. APs (% , $\uparrow$) of Faster R-CNN equipped with different defense methods against digital world transfer attacks. Each column corresponds to a texture crafted against the undefended detector or the corresponding defended detector. Ensemble denotes the ensemble attack, where the texture was jointly optimized against the undefended detector, the AT-defended detector and the FNC-defended detector. Each row corresponds to the detection performance of the undefended detector or the detector with a defense method on four adversarial textures. The best defense performance against each adversarial texture is marked in bold.

Model	Undefended	AT (Li et al., 2023a)	FNC (Yu et al., 2021)	Ensemble
Undefended	0.03	4.40	1.94	0.19
AT (Li et al., 2023a)	62.65	5.97	69.12	11.08
FNC (Yu et al., 2021)	7.59	22.30	0.91	3.71
LGS (Naseer et al., 2019)	0.87	7.09	4.41	1.19
IDBD (Zhou et al., 2020)	1.72	5.05	4.20	0.76
SAC (Liu et al., 2022)	0.43	4.67	2.23	0.60
APE (Kim et al., 2022)	0.05	4.99	2.03	0.36
UDF (Yu et al., 2022)	1.51	8.01	6.31	1.65
Jedi (Tarchoun et al., 2023)	5.32	24.18	8.04	17.24
NAPGuard (Wu et al., 2024)	0.02	5.72	2.26	0.60

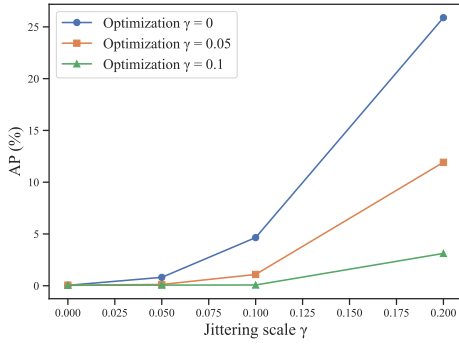


Fig. 4. APs of Faster R-CNN versus jittering scale in the digital world for different jittering scales during texture optimization. $\gamma = 0$ denotes no position jittering.

of the respective best-performing defense on the texture above 24.18 %. Therefore, we optimized the texture on an ensemble of the undefended model, AT model and FNC-defended model, denoted as *Ensemble* in Table 3. Not only did the *Ensemble* texture get good adversarial effect against all models that were optimized on, but it also transferred well to other defenses. Overall, optimizing adversarial texture on an ensemble of defended models decreased the performances of all defended models to APs lower than 17.24 %.

The effect of position jittering. We optimized and evaluated the texture-based attack with various position jittering scales γ . The results are shown in Fig. 4, with $\gamma = 0$ indicating no position

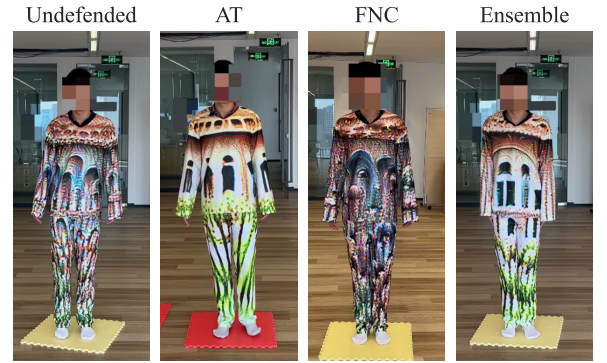


Fig. 5. Visualization of different adversarial clothes produced in the physical world.

jittering. When tested with jittering, the APs of the detector increased, especially when the textures optimized without jittering was applied. This indicates the loss of adversarial effect when the position of the adversarial texture is applied with some error, which is inevitable in physical-world experiments because the shape of person varies. The robustness of the attack to position jittering improved with higher values of γ during optimization. Therefore, applying position jittering let the attack resist to physical implementation errors and different body shapes of different individuals. In this work, we fixed $\gamma = 0.1$ when optimizing 3D textures.

Table 4. ASRs (% , ↓) evaluated on Faster R-CNN equipped with different defense methods against physical world transfer attacks. Each column corresponds to a texture optimized on the detector equipped with the corresponding defense method. Ensemble denotes an ensemble of the undefended detector, the AT-defended detector and the FNC-defended detector. The best defense performance against each set of adversarial clothes is marked in bold.

Model	Undefended	AT (Li et al., 2023a)	FNC (Yu et al., 2021)	Ensemble
Undefended	96.09	67.97	78.12	96.09
AT (Li et al., 2023a)	5.47	76.56	7.81	64.84
FNC (Yu et al., 2021)	95.31	46.88	99.22	98.44
LGS (Naseer et al., 2019)	28.12	78.91	13.28	82.81
IDBD (Zhou et al., 2020)	25.00	50.78	56.25	69.53
SAC (Liu et al., 2022)	96.09	67.97	78.12	96.09
APE (Kim et al., 2022)	92.19	65.62	75.00	94.53
UDF (Yu et al., 2022)	72.66	71.09	33.59	81.25
Jedi (Tarchoun et al., 2023)	78.91	83.59	96.88	87.50
NAPGuard (Wu et al., 2024)	96.09	67.97	78.12	96.09

5.2. Physical-world evaluation

For physical implementation, the textures optimized for different defenses were printed on fabric and subsequently tailored into shirts and trousers. See Fig. 5 for the visualization of four sets of clothes, whose textures have been optimized on different defended models. To assess the real-world effectiveness, we measured the ASRs from various viewing angles by capturing a video of a person turning circles. Frames of different angles were evenly extracted and ASRs on these frames were computed. Table 4 shows the physical world ASRs evaluated on Faster R-CNN equipped with different defense methods, against clothing textures optimized on the undefended detector, AT-defended detector, FNC-defended detector, and an ensemble of the above three models. The results are mostly consistent with those in the digital world (Table 3). The adversarially trained detector successfully defended non-adaptive textures and FNC-adaptive textures, with only 5.47 % and 7.81 % ASR, respectively. However, AT was defeated with AT-adaptive texture and the texture optimized on an ensemble of defended models, with ASRs of 76.56 % and 64.84 %, respectively. LGS was also defeated by this two kinds of textures, with ASRs

of 78.91 % and 82.81 %. IDBD performed well in defending against all three textures optimized on a single model, but the ASR increased to 69.53 % when defending against the texture optimized on an ensemble of models. FNC worked well in clipping adversarial features on the AT-adaptive texture, but failed to defend against other textures, with ASRs all above 95.31 %. Jedi, although performed well in the digital world in defending against the adversarial textures with an AP of 17.24 % on the Ensemble one, the corresponding physical world ASR evaluated on Jedi was 87.50 %, ranked only fifth among the nine defense methods evaluated, behind AT, IDBD, UDF and LGS. The results indicate that the good performance of Jedi in the digital world relies on the less realistic 3D rendering result.

When evaluated against the texture optimized on an ensemble of three defense models, all nine defenses failed to defend against the adversarial texture, with ASRs against them all above 64.84 %. The ASR against the undefended model was 96.09 %.

Different viewing angles. Fig. 6 presents ASRs evaluated on the defended detectors against the Ensemble adversarial clothes in the physical world under different viewing angles from

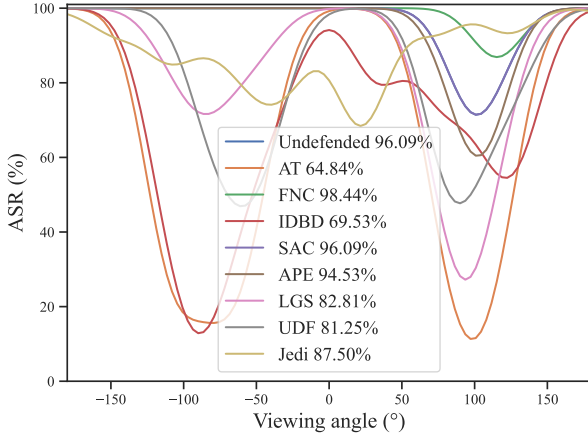


Fig. 6. ASRs evaluated on the defended detectors at a distance of 4 m from different viewing angles on the Ensemble adversarial clothes in the physical world. The person faces the camera when the viewing angle equals 0° . In the legend, we show the ASRs averaged over viewing angles.

-180° to 180° . We observed that viewing angles of $\pm 90^\circ$ (one side of person is facing the camera) were where the defenses performed best, as most defense methods got lower ASRs. This is probably because the side of person has smaller area, and only has a small area of adversarial pattern captured in the camera.

It is worth mentioning that in the physical world, most patch-based attacks (Thys et al., 2019; Hu et al., 2021; Xu et al., 2020) evaluated ASRs from the front view of the person. This is primarily because the patch needs to be fully facing the camera to effectively execute the attack. Therefore, in addition to examining ASRs from all viewing angles, we specifically focused on the viewing angle of 0° . Remarkably, regardless of the defense employed, the front view consistently yielded ASRs surpassing 80 %.

Different distances. We tested the performance of defenses against the adversarial clothes when the person is of different distances away from the camera. We captured a video of a person moving close to the camera facing it, and moving away

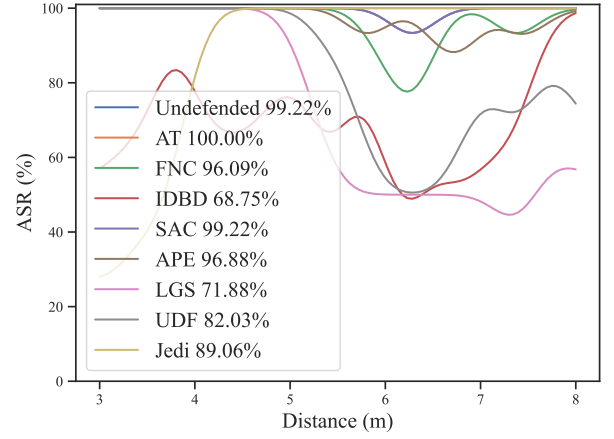


Fig. 7. ASRs evaluated on the defended detectors at different distances with the Ensemble adversarial clothes in the physical world. In the legend, we show the ASRs averaged over distances.

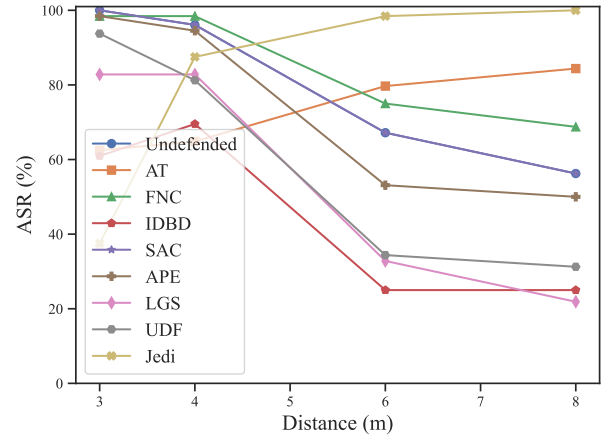


Fig. 8. ASRs evaluated on the defended detectors at different distances averaged over all viewing angles from -180° to 180° with the Ensemble adversarial clothes in the physical world. Distance of 4 m is where prior physical experiments were conducted.

from the camera facing opposite to it, thus fixing the viewing angle to 0° and 180° . ASRs at different distances are shown in Fig. 7. The ASRs evaluated on the undefended model were steadily high as distance changed, with average ASRs across distances achieving 99.22 %. A significant drop of ASR when distance increased appeared on LGS and UDF defenses, especially when the distance exceeded 6 m. IDBD and Jedi, the two entropy-based defenses, performed better when distance decreased.

Different distances for all viewing angles. Fig. 8 presents ASRs averaged over all viewing angles (-180° to 180°) at different distances. With the distance increasing, the ASRs of all defenses declined except AT and Jedi. The worse performance of AT model when distance increased was likely attributed to the diminished performance of the AT model in detecting smaller objects, as detailed by Li et al. (2023a). The performance of Jedi defense when distance changed was consistent with that evaluated on only two view angles (see Fig. 7). Nevertheless, the ASRs against all defenses were still over 20 % across all distances.

6. Conclusion

In this study, we show that a single set of adversarial clothes broke nine defense methods in real-world scenario. Motivated by the finding that enlarged patch broke a typical size-independent defense method, we evaluated nine different defense methods against adversarial clothes. All defense methods had poor performance against adversarial clothes in both the digital world and the physical world. Moreover, we created a single set of adversarial clothes by optimizing the adversarial texture on an ensemble of three defended models. The adversarial clothes achieved an ASR of 96.06 % on the undefended model, and broke nine defended models with ASRs over 64.84 % in the physical world. More detailed analyses on viewing angles show that the defense methods worked better with either side of the person facing the camera. Furthermore, different defense methods performed differently as distance between the person and the camera varied, but the overall defense performance was still sub-optimal.

This paper reveals that SOTA defense methods are still commonly vulnerable to physical-world adversarial examples when confronted with texture-based adversarial attacks. Therefore, there is an urgent need for future adversarial defenses to consider a broader range of attacks, at least including adversarial clothes.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under grant U2341228.

References

- Athalye, A., Engstrom, L., Ilyas, A., Kwok, K., 2018. Synthesizing robust adversarial examples, in: International Conference on Machine Learning, PMLR. pp. 284–293.
- Bookstein, F.L., 1989. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11, 567–585.
- Brown, T.B., Mané, D., Roy, A., Abadi, M., Gilmer, J., 2017. Adversarial patch. *arXiv preprint arXiv:1712.09665*.
- Chen, P.C., Kung, B.H., Chen, J.C., 2021. Class-aware robust adversarial training for object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10420–10429.
- Croce, F., Andriushchenko, M., Sehwal, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P., Hein, M., 2020. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*.
- Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE. pp. 886–893.
- Donato, G., Belongie, S., 2002. Approximate thin plate spline mappings, in: European Conference on Computer Vision, Springer. pp. 21–31.
- Dong, Z., Wei, P., Lin, L., 2022. Adversarially-aware robust object detector, in: European Conference on Computer Vision, Springer. pp. 297–313.
- Goodfellow, I.J., Shlens, J., Szegedy, C., 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.
- Gray, R.M., 2011. Entropy and information theory. Springer Science & Business Media.
- Hu, Y.C.T., Kung, B.H., Tan, D.S., Chen, J.C., Hua, K.L., Cheng, W.H., 2021. Naturalistic physical adversarial patch for object detectors, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7848–7857.
- Hu, Z., Chu, W., Zhu, X., Zhang, H., Zhang, B., Hu, X., 2023. Physically realizable natural-looking clothing textures evade person detectors via 3d modeling, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16975–16984.
- Hu, Z., Huang, S., Zhu, X., Sun, F., Zhang, B., Hu, X., 2022. Adversarial texture for fooling person detectors in the physical world, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13307–13316.
- Ji, N., Feng, Y., Xie, H., Xiang, X., Liu, N., 2021. Adversarial yolo: Defense human detection patch attacks via detecting adversarial patches. *arXiv preprint arXiv:2103.08860*.
- Jing, L., Wang, R., Ren, W., Dong, X., Zou, C., 2024. Pad: Patch-agnostic defense against adversarial patch attacks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 24472–24481.
- Karmon, D., Zoran, D., Goldberg, Y., 2018. Lavan: Localized and visible adversarial noise, in: International Conference on Machine Learning, PMLR. pp. 2507–2515.

- Kim, T., Yu, Y., Ro, Y.M., 2022. Defending physical adversarial attack on object detection via adversarial patch-feature energy, in: Proceedings of the 30th ACM International Conference on Multimedia, pp. 1905–1913.
- Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization, in: International Conference on Learning Representations.
- Li, X., Chen, H., Hu, X., 2023a. On the importance of backbone to the adversarial robustness of object detectors. arXiv preprint arXiv:2305.17438.
- Li, X., Wang, Z., Zhang, B., Sun, F., Hu, X., 2023b. Recognizing object by components with human prior knowledge enhances adversarial robustness of deep neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 8861–8873.
- Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context, in: European Conference on Computer Vision, Springer. pp. 740–755.
- Liu, J., Levine, A., Lau, C.P., Chellappa, R., Feizi, S., 2022. Segment and complete: Defending object detectors against adversarial patch attacks with robust patch detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14973–14982.
- Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A., 2018. Towards deep learning models resistant to adversarial attacks, in: International Conference on Learning Representations.
- Mao, Z., Chen, S., Miao, Z., Li, H., Xia, B., Cai, J., Yuan, W., You, X., 2024. Enhancing robustness of person detection: A universal defense filter against adversarial patch attacks. *Computers & Security* 146, 104066.
- Metzen, J.H., Finnie, N., Huttmacher, R., 2021. Meta adversarial training against universal patches. arXiv preprint arXiv:2101.11453.
- Mu, N., Wagner, D., 2021. Defending against adversarial patches with robust self-attention, in: ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning.
- Naseer, M., Khan, S., Porikli, F., 2019. Local gradients smoothing: Defense against localized adversarial attacks, in: IEEE Winter Conference on Applications of Computer Vision, pp. 1300–1307.
- Rao, S., Stutz, D., Schiele, B., 2020. Adversarial training against location-optimized adversarial patches, in: European Conference on Computer Vision, Springer. pp. 429–448.
- Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems* 28.
- Rossolini, G., Nesti, F., Brau, F., Biondi, A., Buttazzo, G., 2023. Defending from physically-realizable adversarial attacks through internal over-activation analysis, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 15064–15072.
- Tarchoun, B., Ben Khalifa, A., Mahjoub, M.A., Abu-Ghazaleh, N., Alouani, I., 2023. Jedi: Entropy-based localization and removal of adversarial patches, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4087–4095.
- Thys, S., Van Ranst, W., Goedemé, T., 2019. Fooling automated surveillance cameras: adversarial patches to attack person detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–7.
- Tian, Z., Shen, C., Chen, H., He, T., 2019. Fcos: Fully convolutional one-stage object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9627–9636.
- Wu, D., Xia, S.T., Wang, Y., 2020a. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems* 33, 2958–2969.
- Wu, S., Wang, J., Zhao, J., Wang, Y., Liu, X., 2024. Napguard: Towards detecting naturalistic adversarial patches, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 24367–24376.
- Wu, Z., Lim, S.N., Davis, L.S., Goldstein, T., 2020b. Making an invisibility cloak: Real world adversarial attacks on object detectors, in: European Conference on Computer Vision, Springer. pp. 1–17.
- Xu, K., Xiao, Y., Zheng, Z., Cai, K., Nevatia, R., 2023. Patchzero: Defending against adversarial patch attacks by detecting and zeroing the patch, in: IEEE Winter Conference on Applications of Computer Vision, pp. 4632–4641.
- Xu, K., Zhang, G., Liu, S., Fan, Q., Sun, M., Chen, H., Chen, P.Y., Wang, Y., Lin, X., 2020. Adversarial t-shirt! evading person detectors in a physical world, in: European Conference on Computer Vision, Springer. pp. 665–681.
- Yu, C., Chen, J., Xue, Y., Liu, Y., Wan, W., Bao, J., Ma, H., 2021. Defending against universal adversarial patches by clipping feature norms, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 16434–16442.
- Yu, Y., Lee, H.J., Lee, H., Ro, Y.M., 2022. Defending person detection against adversarial patch attack by using universal defensive frame. *IEEE Transactions on Image Processing* 31, 6976–6990.
- Zhang, H., Wang, J., 2019. Towards adversarially robust object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 421–430.
- Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., Jordan, M., 2019. Theoretically principled trade-off between robustness and accuracy, in: International Conference on Machine Learning, PMLR. pp. 7472–7482.
- Zhou, G., Gao, H., Chen, P., Liu, J., Dai, J., Han, J., Li, R., 2020. Information distribution based defense against physical attacks on object detection, in: IEEE International Conference on Multimedia and Expo Workshops, pp. 1–6.