Supplemental Materials for
Naturalistic Physical Adversarial Patch for Object Detectors

1. Training and Testing Details

In the following, we summarize the details of generating the proposed adversarial patches. We use the output of $1,000th$ epoch as the final patch. The weight of total variation loss is $\lambda_{tv} = 0.1$, and the batch size of training is 8 for all detectors except for YOLOv4. The batch size of YOLOv4 is 4 because the weight consumes more memory than other detectors. As for the physical evaluation, our generated adversarial patches are printed on Cotton T-shirt using commercial heat transfer. In addition, we also print the patch on the paper. The printer is RICOH MP C2004ex. We use Logitech V-u0015 Webcam Camera (720p/30fps) to record the videos.

2. Real-world Evaluation

Figure 1 shows the T-shirt we printed for real-world physical evaluation. We resize the patch into $30 \times 40\ cm^2$ and print the patch onto a T-shirt.

As Figure 2 shown, the T-shirt attacks the object detector. People with this adversarial T-shirt would not be detected by YOLOv4tiny.

In addition, Figure 3 shows another adversarial patch. The patch attacks the detector effectively. In addition, if we rotate or translate the patch, it still can attack the detector.

Figure 4 shows some other examples. The penguin patch and the dog patch can make the people disappear.

Table 1 shows additional physical evaluations with small step movements and body rotations for both indoor and outdoor settings. The results are measured in terms of detection recall and included in parenthesis are the average detection probabilities. Sample images of the experiment are also shown in Figure 5 and Figure 6. Our adversarial patch can successfully reduce the detection probabilities and in many cases push it below the 0.5 threshold to render it "invisible" to the detector. We can observe that our adversarial patch performs well on small shifts and scale variations but sensitive to rotation.

3. Details of FasterRCNN

In Table 1 of the main paper, we utilize FasterRCNN as the victim detector to train the adversarial patch. We use the PyTorch pretrained FasterRCNN, and use ResNet50

Figure 1: Sample T-shirts with the proposed adversarial patch.

Figure 2: The adversarial patches worn by people in the real-world scenarios.

Setting	No Adversarial Patch	Front-Facing	Body Rotation	Small steps
Indoor	1.00 (0.896)	0.389 (0.467)	0.482 (0.513)	0.421 (0.504)
Outdoor	1.00 (0.892)	0.505 (0.501)	0.824 (0.631)	0.551 (0.535)

Table 1: Additional physical evaluations on indoor and outdoor settings measured in terms of recall. Numbers in parenthesis indicate the average detection probability.

Figure 3: Physical evaluations with different transformations where the patch is printed on the cloth.

as the backbone. Different from the one-stage YOLO series detectors, FasterRCNN is a two-stage detector. That is, YOLO detectors simultaneously output the class and location, but FasterRCNN first output a region proposal that indicates whether there are objects and then use a classifier to recognize the object. When we train the patch, we first filter the classifier output and only select the "person" label. Next, since each person has a score that is the output of the region proposal, we choose the person that has the highest score to calculate the objective function.

4. Transformations

We adopt several transformations to train the adversarial patch. They are in-plane rotation, out-of-plane rotation, random translation, random occlusion, and crease. Figure 8 illustrates the transformations. In addition, Figure 9 illustrates the crease we used. For real-world evaluation and application, there are many variations (*i.e.,* patch distortion) that we typically have no control over because the patch is printed on clothes. Thus, we utilize the transformations to simulate the real-world variation during training, which makes adversarial patches robust against real-world variation.

To investigate the effectiveness of each transformation, we perform an ablation study for those transformations in different settings, with or without using each transformation during the patch generation versus with or without using the corresponding transformation during evaluation, as Table 2 illustrated. In addition, for each transformation, we generate three patches with different initial starting points, and show the corresponding evaluation results with their mean and standard deviation. We find that each transformation shows its effectiveness for the attack performance to some extent and is very subjective to the initial starting point during the generation stage. Figure 7 shows the patches used in Table 2.

5. Dataset

We use three kinds of datasets to train or evaluate the proposed methods. They are INRIA, MPII, and the video that was collected by ourselves. In the main paper, we mainly focus on INRIA. Here, we demo the INRIA and MPII datasets.

Figure 4: Physical evaluations with different patch patterns where the adversarial patches are printed on the paper.

Trans.($\mathcal{T}$) | Trained / Test	w/ $\mathcal{T}$ / w/o any $\mathcal{T}$	w/ $\mathcal{T}$ / w/ $\mathcal{T}$	w/o any $\mathcal{T}$ / w/ $\mathcal{T}$
No trans.	14.95± 4.44	14.95± 4.44	14.95± 4.44
In-plane rotation	18.40± 2.63	17.46± 5.53	19.16± 4.29
Random translation	17.30± 4.67	34.48± 2.27	34.75± 3.95
Crease	13.19± 3.38	17.83± 2.16	17.22± 3.15
Out-of-plane rotation	14.42± 2.87	30.75± 2.39	32.50± 4.77
Random occlusion	17.80± 1.72	30.51± 5.43	38.57± 2.64
Blur	16.64± 3.52	18.04± 3.18	19.95± 4.99

Table 2: As Table 5 in the main paper, we generate three patches for each transformation using different initial starting points. This table shows the corresponding evaluation results with their mean and standard deviation under different training and test settings.

(a) Step front (b) Step back

(c) Rotation (d) Step side

Figure 5: Indoors physical evaluations.

(a) Step front (b) Step back

(c) Rotation (d) Step side

Figure 6: Outdoors physical evaluations.

5.1. INRIA

INRIA is a pedestrian dataset. Figure 10 shows some example images in this dataset.

5.2. MPII

MPII is a human pose dataset. As shown in Figure 11, the dataset collects images about people in different activities, including running, dancing, walking, swimming, and so on. In our experiments, we select images in categories running, dancing, and walking.

6. Different-class Patches

Because BigGAN is a conditional GAN, it can control the class of the generated patch. This is, if we set the patch

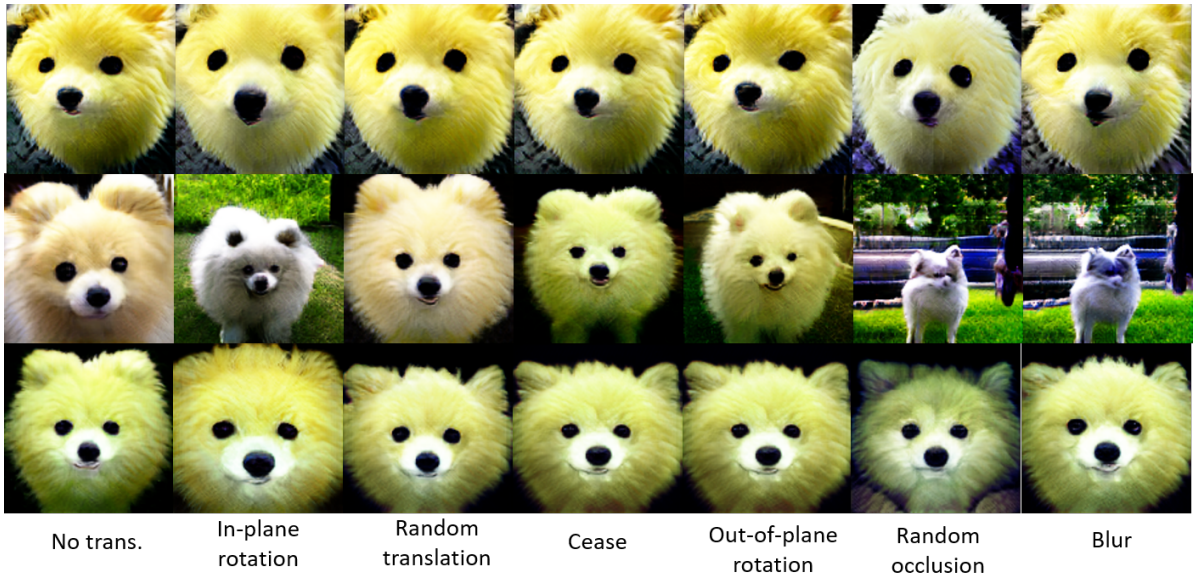


Figure 7: The corresponding patches used in Table 2, where the first row shows the corresponding patches used in Table 5 of the main paper.

as class dog, BigGAN can guarantee that generated examples is a dog. Therefore, we can manipulate the class of generated patch. Figure 12 shows some examples generated by BigGAN. These patches are trained by YOLOv4 tiny using BigGAN.

7. The Size of Patch

In our experiment, we use 20% of the bounding box to be the size of the adversarial patch. If we use the larger patch, the attack performance will become better. However, the adversarial patch should not be too large because we want to print them on the T-shirt. In addition, the size of the patch can be seen as an attack budget; thus, 20% is the suitable selection. Figure 13 and Figure 14 show the different sizes of adversarial patches. Figure 5 in the main paper shows more details for the different patch sizes. In our experiment, 20% of the bounding box is approximately 30×40 cm^2 for a regular person. Therefore, we print the patch in this size for the physical evaluation.

8. Subjective Test

Figure 15 illustrates the pictures used in the first subjective survey. The patches generated by the proposed method are high-lined with orange lines. Figure 16 illustrates the pictures used in the second subjective survey.

(a) Original

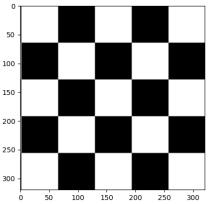
(b) Random In-plane Rotation (Roll)

(c) Random Out-of-Plane Rotation (Yaw)

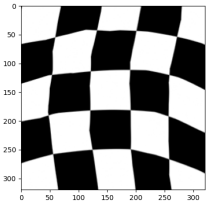
(d) Random Translation

(e) Random Occlusion

Figure 8: Various transformations of adversarial patches.



(a) Original Checkerboard Pattern



(b) Checkerboard Pattern with Crease Simulation

Figure 9: Crease Simulation.

Figure 10: Sample images of the INRIA person dataset.

Figure 11: Sample images of the MPII person dataset.

(a) bird

(b) strawberry

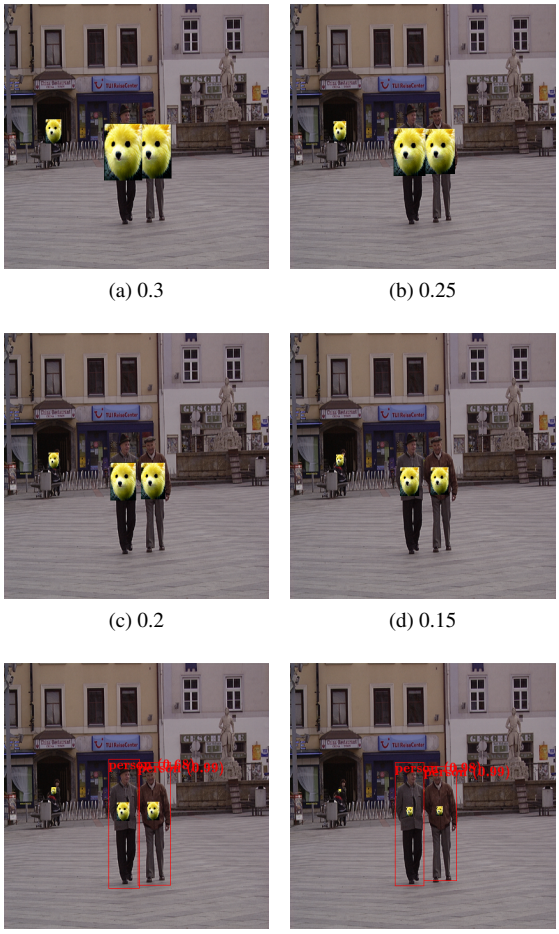
(c) persian cat

(d) elephant

(e) castle

(f) balloon

Figure 12: Adversarial patches for different classes.

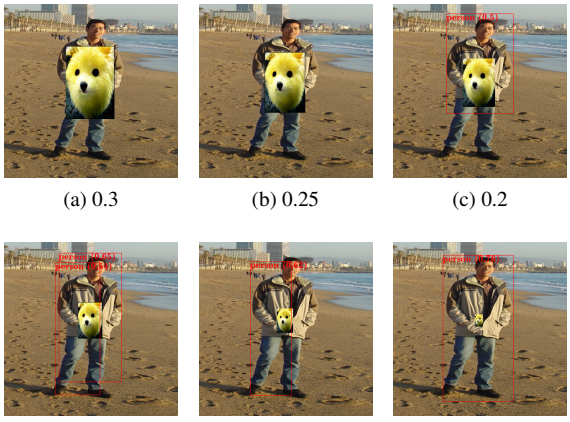


(a) 0.3 (b) 0.25

(c) 0.2 (d) 0.15

(e) 0.1 (f) 0.05

Figure 13: The effectiveness of adversarial patches under different sizes for the small-size pedestrian.



(a) 0.3 (b) 0.25 (c) 0.2

(d) 0.15 (e) 0.1 (f) 0.05

Figure 14: The illustration of effectiveness of adversarial patches under different ratios of patch size with respect to the size of pedestrian.

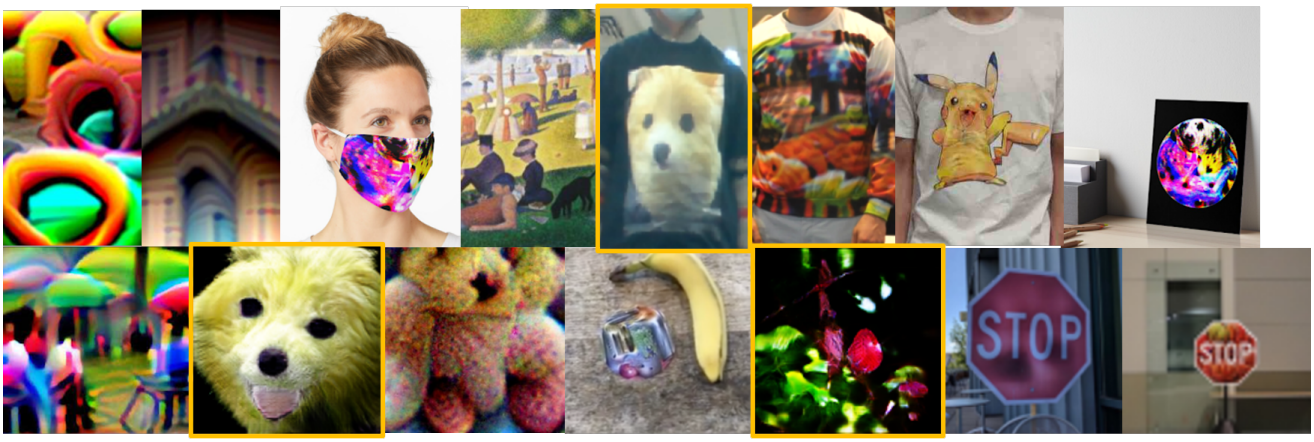


Figure 15: The pictures used in the first subjective survey.

Figure 16: The pictures used in the second subjective survey.