Physically Adversarial Infrared Patches with Learnable Shapes and Locations Supplemental Material

Xingxing Wei*, Jie Yu, Yao Huang
Institute of Artificial Intelligence, Beihang University, Beijing, 100191, China
xxwei@buaa.edu.cn, sy2106137@buaa.edu.cn, y_huang@buaa.edu.cn

1. Details of hyperparameters in experiments

In our experiment, hyparameters are set as: $V_{thre} = 0.6$ in Eq.(8), $\alpha = 2$ in Eq.(9), $\lambda_1 = 0.25, \lambda_2 = 0.5$ in Eq.(10), $\epsilon = 0.1$ in Eq.(13), $\mu = 0.85$ in Eq.(14), $T = 100$, $\epsilon_{max} = 200$. All the hyparameters are tuned in the validation set. Besides, the number of test images is 250 (All the images can be detected before our attack).

2. Weighted factors in Eq.(10)

We tune λ_1, λ_2 versus ASR and aggregation as follows, where we see when $\lambda_1 = 0.25, \lambda_2 = 0.5$, ASR and aggragation meet a balance. Therefore, we set$\lambda_1 = 0.25, \lambda_2 = 0.5$ in Eq.(10). Because aggregation term in Eq.(10) is charge of adjusting the compactness. We can assign a large λ_2 to obtain a compact shape.

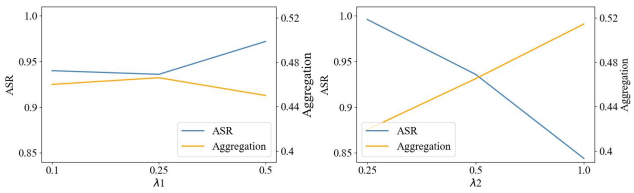

Figure 1. Parameter tuning for weighted factors

3. Effects of cover image values

To simulate the attack effects of cover image values, we conduct the attack under different gray values of $\hat{\boldsymbol{x}}$ in the digital world. The visual examples are given in Figure 3. For each gray value of $\hat{\boldsymbol{x}}$, we attack 250 images and calculate the corresponding attack success rate (ASR). The changing trend of ASR with different gray values of $\hat{\boldsymbol{x}}$ is shown in Figure 2. We can see that (1) In general, the ASR with a low gray value is significantly higher than the ASR with a high gray value, which indicates that the method to reduce thermal radiation is better than the method to increase thermal radiation if we want to prevent the object detector from detecting the target object. (2) When the gray value $\hat{\boldsymbol{x}}$ is far from the gray value of the pedestrian in the infrared image, the attack is more likely to be successful and when the gray value $\hat{\boldsymbol{x}}$ is closed to the gray value of the pedestrian in the infrared image, the attack is relatively harder to be successful. (3) The best interval of gray value for attacks is at [0, 0.2], which ASR can achieve an acceptable performance rangeing from 0.936 to 0.88 (when the patch size is 200 pixels). In the real world, because the value of cover image $\hat{\boldsymbol{x}}$ depends on the thermal properties of used insulation materials rather than a given fixed value, this interval can provide the guidance for selecting thermal materials in the physical implementation.

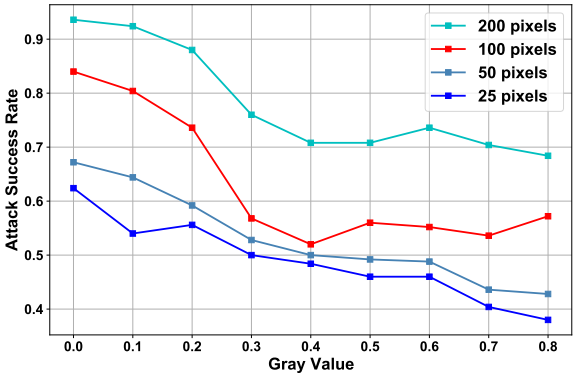

Figure 2. The changes of ASR with different gray values under the different patch sizes.

4. Effects of patch sizes

In our method, we use ϵ_{max} to determine the upper bound of the generated infrared patch (see Algorithm 1). To explore the impact of patch sizes on the attack effect, we set 25, 50, 100, and 200 pixels for ϵ_{max}. The corresponding quantitative results are shown in Figure 2, where we can see that for a fixed gray value $\hat{\boldsymbol{x}}$, increasing the patch's size can effectively improve the attack performance. This is reasonable because a large patch can cover more areas of the pedestrian, and thus the ASR will increase. An extreme case is that the infrared

*Corresponding author

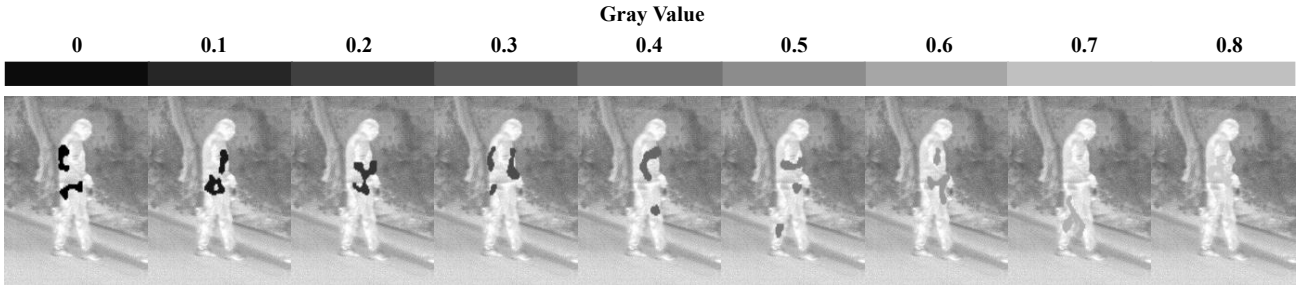

Figure 3. Adversarial examples with different gray values of cover image in infrared patches. The detailed gray values are listed above the pictures, and the corresponding ASR using different gray values are listed under the pictures.

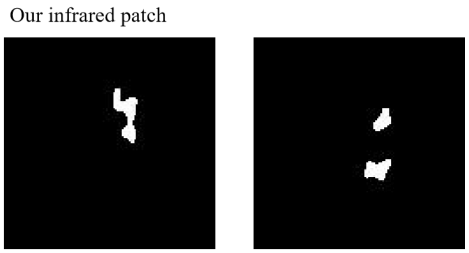

Figure 4. Two generated adversarial infrared patches in our method.

patch covers all the pedestrian area, and the corresponding ASR will be the highest. However, the implementation under this case in the physical world will become hard. Therefore, we should choose an available patch size for the practice usage. In our method, we set $\epsilon_{max} = 200$ in the experiments.

5. Number of patches

In our method, the number of generated patches can be single one or multiple ones (see Figure 4), which is automatically determined in the optimization process according to the attack goal.

In Figure 3, we see that the number of generated patches will change with different gray values of cover image. But thanks to the aggregation regularization, the number of patches will not exceed three patches according to the empirical results in our extensive experiments.

6. Experiments on latest object detectors

For fair comparisons, we choose YOLOv3 used in adversarial blubs and clothing in the paper. We have attacked the latest YOLOv5 and YOLOv7 on the 250 test images, the corresponding ASR is 93.2% and 74%, respectively.

Table 1. ASR at different object detectors

YOLOv3	YOLOv5	YOLOv7
93.6%	93.2%	74%

7. Sensibility to location and shape errors

When pasting our infrared patches on target object from digital world to physical world, it is inevitable to generate errors versus patches' location and shape. This section analyse the sensibility to them. Sec.3.3 shows that the insulation material has almost the same gray values, which means the patch value is not necessary to exactly align to its location computed in the digital world. Therefore, our infrared patch is robust to the translation error, rotation error, and incompleteness. We conduct experiments as follows, which verifies that ASR does not significantly drop.

Table 2. Sensibility to location and shape errors

	Translation		Rotation		Incompleteness	
	3pix	5pix	10°	20°	15%	30%
ASR	88.40% (↓5.20%)	80.40% (↓13.20%)	85.20% (↓8.40%)	78.80% (↓14.80%)	82.80% (↓10.80%)	72.40% (↓21.20%)

8. Convergence of the optimization process

This section discuss the convergence of our method. We illustrate the change of average total loss in Eq.(10) across 250 test images, where when epoch meets 40, the optimization achieves the convergence, costing 2∼3 seconds per image. This shows the high optimization efficiency.

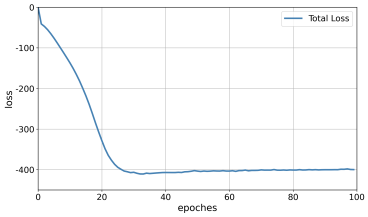

Figure 5. The convergence of our method

9. Implementation in the real world

To implement our method in the real world, we need to first determine the thermal insulation material, and then crop it according to the learned shape and location. We give the details in this section.

Thermal insulation material: As mentioned above, we find the best interval of cover image $\hat{\boldsymbol{x}}$ for attacks is at [0, 0.2], and the gray value of cover image $\hat{\boldsymbol{x}}$ depends on the thermal properties of insulation materials. Based on this phenomenon, we decide to choose aerogel material for its great thermal insulation effect. We use the aerogel material at the normal temperatures in our life and measure the gray value in infrared images. Finally, we find that its gray value can reach around 0.1, which is located in the interval [0, 0.2]. Therefore, we set the gray value of cover image $\hat{\boldsymbol{x}}$ as 0.1 in our method, and then learn the shapes and locations of the adversarial infrared patches in the digital world.

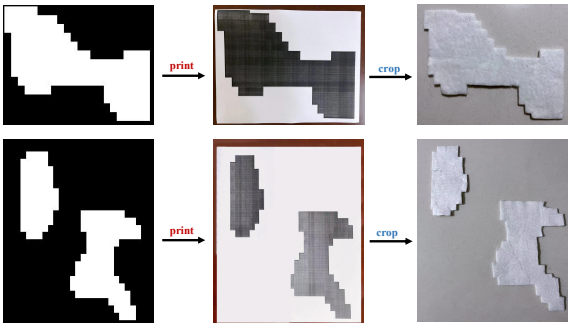

Figure 6. Implementation details of infrared patches from the digital world to the physical world. The first column denotes the infrared patch in the digital world, the second column denotes the shape of infrared patches in the printing paper, and the third column denotes the infrared patch with aerogel material used in real world.

Restore the infrared patches in real world: After obtaining the infrared patch in the digital world, we should implement the adversarial infrared patch in the real world. To achieve this goal, we design the following steps. Firstly, we print the learned mask of infrared patches on the printing paper to draw the shapes. Secondly, we cut the shape out from the printing paper. Thirdly, we crop the aerogel material along the edges of the cropped printing paper to obtain the adversarial infrared patch, which is finally attached on the learned location of the object to finish the physical attack. The process of manufacturing infrared patches is illustrated as Figure 6. We can see that the physical implementation of our method is simple in the real world, which greatly improves the accessibility. In the practice, the implementation process costs 0.5 hours at most for a new hand.