Supplementary Material: Fooling thermal infrared pedestrian detectors in real world using small bulbs

Submission ID: 3500

Details of the pixel-level adversarial patch attack in the digital world

The pixel-level adversarial patch used pixels as the basic unit. Each pixel value was an optimization variable. We followed the work of Thys et al. (2019) to build a square patch with the pixel size of 300×300. The difference was that our patch was a grayscale image instead of an RGB image. We first initialized a 300×300 pixel-level patch. There were two options: random initialization and uniform initialization. We found that when each pixel was initialized to 0.5, the network convergence effect was better in our experiment, so we adopted the uniform initialization method.

To make the patch more robust, we designed a variety of transformations including random noise on the patch, random rotation of the patch (clockwise or counterclockwise within 20 degrees), random translation of the patch, and random changes in the brightness and contrast of the patch. These transformations simulate the perturbation of the physical world to a certain extent, which effectively improves the robustness of the patch. Then we used the training set of *FLIR_person_select* and placed the patch on the upper body of the pedestrians according to the position of the bounding box. The size of the patch was 1/5 of the height of the bounding box. Next, we used the patched image as input and ran the YOLOv3 pedestrian detector we had trained. We used a stochastic gradient descent optimizer with momentum, and the size of each batch was 8. The optimizer used the backpropagation algorithm to update the pixel values by minimizing the loss function. Through this process, we obtained a series of patches. Figure S1 is an example after 65801 iterations.

Next, we applied the optimized patch shown in Figure S1 to the test set, using the same process we used during training, including various transformations. We used random noise patches with maximum amplitude value 1 and constant pixel value patches (blank patches) for control experiments. The pixel values of blank patch in our experiment were 0.75 (Other values had similar influence to detectors; see the next section). We applied these different patches to the *FLIR_person_select* test set, and then input the patched images to the same detection network to test its detection performance. We adopted the IOU method to calculate the accuracy of the detection. The precision-recall (PR) curves are shown in Figure S2. We defined the output of the clean image input as the ground truth, then the pixel-level adversarial patch made the average precision (AP, the area under the PR curve) of the target detector drop by 74.57%. An example is shown in Figure S3. In contrast, the AP of the target detector dropped by 25.30% and 29.27% using random noise patch and blank patch, respectively.

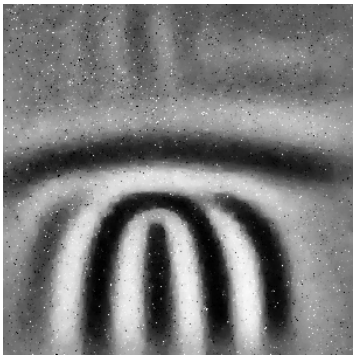

Figure S1: An example of the pixel-level adversarial patch.

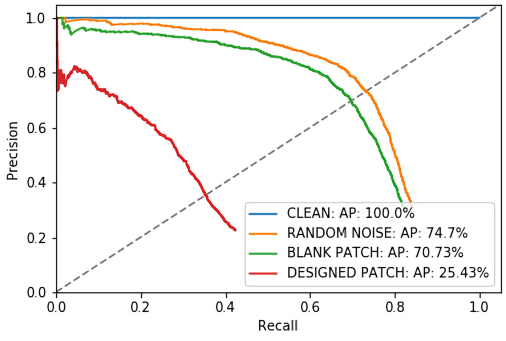

Figure S2: Evaluation of the pixel-level adversarial patch attack.

Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

(a) Clean (b) Blank patch

(c) Random noise (d) Designed patch

Figure S3: The pixel-level adversarial patch attack and control experiments.

Influence of the pixel value of the blank patch to the detection performance

All pixels of the blank patch had the same value. So we studied the influence of the pixel value to the detection performance. The pixel value varied from 0 to 1. We chose five values (0.1, 0.25, 0.5, 0.75 and 0.9). The PR curves are shown in Figure S4. The blank patches with different pixel values caused the AP of YOLOv3 to drop by 30% $\pm$ 5%. Therefore the influence to the detector did not vary significantly with different pixel values. We chose a typical value of 0.75 in other experiments.

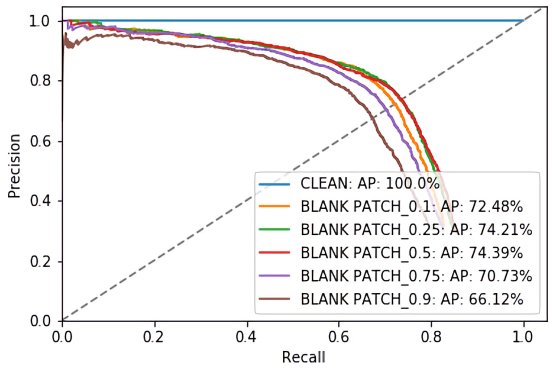

Figure S4: Evaluation of blank patches with different pixel values.

Attack the visible light and infrared object detection systems at the same time

An interesting question is whether we can design a physical board that can evade the person detectors working on both visible light images and infrared images. Based on our method, the solution turned out to be simple. We printed an adversarial patch on a paper, which was crafted according to the a previous work (Thys, Ranst, and Goedemé 2019) by using YOLOv3 as the target detector. The size of the adversarial patch was 29.8cm $\times$ 29.8cm. Then we put the paper on the physical board with small bulbs we designed before. The digital patch and the physical board is shown in Figure S5(a) and Figure S5(b).

(a) The digital adversarial patch (b) The physical adversarial board

Figure S5: The adversarial digital patch and physical board. Note that the small bulbs are covered by the printed paper in (b).

We invited several persons to participate in the test. They could hold the adversarial board, or a blank board, or nothing. We used visible light camera and thermal infrared camera to shoot these people under the same conditions. Then we input the images to the target detector YOLOv3. The result showed that we could successfully attack the visible light and infrared object detection systems at the same time. Several examples are shown in Figure S6.

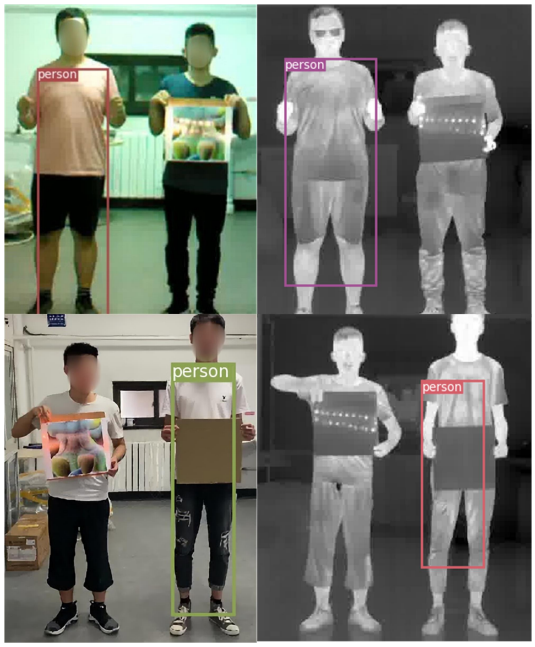

Figure S6: An example of visible light and infrared physical board attacks. For privacy reasons, we blurred the facial area on visible light images.

References

Thys, S.; Ranst, W. V.; and Goedemé, T. 2019. Fooling Automated Surveillance Cameras: Adversarial Patches to Attack Person Detection. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, 49–55. Computer Vision Foundation / IEEE. doi:10.1109/CVPRW.2019.00012. URL http://arxiv.org/abs/1904.08653.